Sampling from the Mean-Field Stationary Distribution

Yunbum Kook* Matthew S. Zhang†
Sinho Chewi‡ Murat A. Erdogdu§ Mufan (Bill) Li¶

February 11, 2024

Abstract

We study the complexity of sampling from the stationary distribution of a mean-field SDE, or equivalently, the complexity of minimizing a functional over the space of probability measures which includes an interaction term. Our main insight is to *decouple* the two key aspects of this problem: (1) approximation of the mean-field SDE via a finite-particle system, via uniform-in-time propagation of chaos, and (2) sampling from the finite-particle stationary distribution, via standard log-concave samplers. Our approach is conceptually simpler and its flexibility allows for incorporating the state-of-the-art for both algorithms and theory. This leads to improved guarantees in numerous settings, including better guarantees for optimizing certain two-layer neural networks in the mean-field regime.

1 Introduction

The minimization of energy functionals $\mathcal{E}$ over the Wasserstein space $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ of probability measures has attracted substantial research activity in recent years, encompassing numerous application domains, including distributionally robust optimization [Kuh+19; YKW22], sampling [JKO98; Wib18; Che23], and variational inference [LW16; Lam+22; Dia+23; JCP23; Lac23; YY23].

A canonical example of such a functional is $\mathcal{E}(\mu) = \int V \,\mathrm{d}\mu + \int \log \mu \,\mathrm{d}\mu$, where $V : \mathbb{R}^d \to \mathbb{R}$ is called the potential. Up to an additive constant, which is irrelevant for the optimization, this energy functional equals the KL divergence $\mathsf{KL}(\mu \,\|\, \pi)$ with respect to the density $\pi \propto \exp(-V)$, and the celebrated result of [JKO98] identifies the Wasserstein gradient flow of $\mathcal{E}$ with the Langevin diffusion. This link has inspired a well-developed theory for log-concave sampling, with applications to Bayesian inference and randomized algorithms; see [Che23] for an exposition.

The energy functional above contains two terms, corresponding to two of the fundamental examples of functionals considered in Villani's well-known treatise on optimal transport [Vil03]. Namely, they are the "potential energy" and the entropy, the latter being a special case of the "internal energy." However, Villani identifies a third fundamental functional—the "interaction energy"—with the *pairwise* form given by

$$\mathcal{E}(\mu) \coloneqq \int V(x)\,\mu(\mathrm{d}x) + \iint W(x-y)\,\mu(\mathrm{d}x)\,\mu(\mathrm{d}y) + \frac{\sigma^2}{2}\int \log \mu(x)\,\mu(\mathrm{d}x)\,. \tag{pE}$$

More generally, in this work we consider minimizing the *generic* entropy-regularized energy

$$\mathcal{E}(\mu) \coloneqq \mathcal{F}(\mu) + \frac{\sigma^2}{2}\int \log \mu \,\mathrm{d}\mu \tag{gE}$$

where $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R}$ is a known functional. The minimization of the energy (gE) has recently been of interest due to its role in analysing neural network training dynamics in the mean-field regime,

*School of Computer Science at Georgia Institute of Technology, yb.kook@gatech.edu
†Department of Computer Science at University of Toronto, and Vector Institute, matthew.zhang@mail.utoronto.ca
‡School of Mathematics at Institute for Advanced Study, schewi@ias.edu
§Department of Computer Science at University of Toronto, and Vector Institute, erdogdu@cs.toronto.edu
¶Department of Operations Research and Financial Engineering at Princeton University mufan.li@princeton.edu

including with [SNW22] and without [CB18; MMN18] entropic regularization, as well as with Fisher regularization [Cla+23].

For the sake of exposition, let us first focus on minimizing the pairwise energy (pE). A priori, this question is more difficult than log-concave sampling; for instance, π does not admit a closed form but rather is the solution to a non-linear equation

$$\pi(x) \propto \exp\Big(-\frac{2}{\sigma^2}\, V(x) - \frac{2}{\sigma^2} \int W(x - \cdot)\, \mathrm{d}\pi\Big)\,. \tag{1.1}$$

However, here too there is a well-developed mathematical theory which suggests a principled algorithmic approach. Just as the Wasserstein gradient flow of (pE) in the case when $W = 0$ can be identified with the Langevin diffusion, the Wasserstein gradient flow of (pE) in the case when $W \neq 0$ corresponds to a (pairwise) McKean–Vlasov SDE, i.e., an SDE whose coefficients depend on the marginal law of the process, given as

$$\mathrm{d}X_t = -\Big(\nabla V(X_t) + \int \nabla W(X_t - \cdot)\, \mathrm{d}\pi_t\Big)\, \mathrm{d}t + \sigma\, \mathrm{d}B_t\,, \tag{pMV}$$

where $\pi_t = \mathsf{law}(X_t)$, W is even, and $\{B_t\}_{t\geq 0}$ is a standard Brownian motion on $\mathbb{R}^d$. Since the McKean–Vlasov SDE is the so-called *mean-field limit* of interacting particle systems, we can approximately sample from the minimizer π by numerically discretizing a system of SDEs, which describe the evolution of N particles$\{X_t^{1:N}\}_{t\geq 0} \coloneqq \{(X_t^1, \dots, X_t^N)\}_{t\geq 0}$ as:

$$\mathrm{d}X_t^i = -\Big(\nabla V(X_t^i) + \frac{1}{N-1} \sum_{j\in[N]\setminus i} \nabla W(X_t^i - X_t^j)\Big)\, \mathrm{d}t + \sigma\, \mathrm{d}B_t^i\,, \quad \forall\, i \in [N]\,, \tag{pMV$_N$}$$

where $\{B^i : i \in [N]\}$ is a collection of independent Brownian motions. Moreover, the error from approximating the mean-field limit via this finite particle system has been studied in the literature on *propagation of chaos* [Szn91]. Similarly, the Wasserstein gradient flow for (gE) corresponds to the *mean-field Langevin dynamics* and admits an analogous particle approximation.

The bounds for propagation of chaos have been refined over time, with [LL23] recently establishing a tight error dependence $\mathcal{O}(1/N)$ on the total number of particles N. These bounds, however, do not translate immediately into algorithmic guarantees. Existing sampling analyses study the propagation of chaos and discretization as a single **entangled** problem, which thus far have only been able to use weaker $\mathcal{O}(\sqrt{1/N})$ rates for the former. Furthermore, there has been recent interest in using more sophisticated particle-based algorithms, e.g., "non-linear" Hamiltonian Monte Carlo [BS23] and the mean-field underdamped Langevin dynamics [FW23] to reduce the discretization error. Currently, this requires repeatedly carrying out the propagation of chaos and time discretization analyses from the ground up for each instance.

This motivates us to pose the following questions: **(1)** Can we incorporate improvements in the propagation of chaos literature, such as the $\mathcal{O}(1/N)$ error dependence shown in [LL23], to improve existing theoretical guarantees? **(2)** Can we leverage recent advances in the theory of log-concave sampling to design better algorithms?

Our main proposal in this work is to **decouple** the error into two terms, representing the propagation of chaos and discretization errors respectively. This simple and *modular* approach immediately allows us to answer both questions in the affirmative. Namely, we show how to combine established propagation of chaos bounds in various settings [including the sharp rate of LL23] with a large class of sophisticated off-the-shelf log-concave samplers, such as interacting versions of the randomized midpoint discretization of the underdamped Langevin dynamics [SL19; HBE20], Metropolis-adjusted algorithms [Che+21; WSC22; AC23], and the proximal sampler [LST21; Che+22a; FYC23]. Our framework yields improvements upon prior state-of-the-art, such as [BS23; FW23], and provides a clear path for future ones.

1.1 Contributions and Organization

Propagation of chaos at stationarity. We provide three propagation of chaos results which hold in the $\mathcal{W}_2$, $\sqrt{\mathsf{KL}}$, and $\sqrt{\mathsf{FI}}$ "metrics"; the rates reflect the distance of the k-particle marginal of the finite-particle system from $\pi^{\otimes k}$: (1) In the setting of (pE), under strong displacement convexity, we obtain a $\mathcal{O}(\sqrt{k/N})$ rate by adapting techniques from [Szn91; Mal01]; (2) without assuming displacement convexity, but assuming a

weaker interaction, we obtain the sharp rate of $\widetilde{\mathcal{O}}(k/N)$ following [LL23]; (3) finally, in the general setting of (gE), and assuming $\mathcal{F}$ is convex along linear interpolations, we obtain a $\mathcal{O}(\sqrt{k/N})$ rate following [CRW22].

Unlike prior works, our proofs are carried out at stationarity; thus, our proofs are *self-contained*, streamlined, and include various improvements (e.g., weaker assumptions and explicit bounds). As a result, our work also serves as a helpful exposition to the mathematics of propagation of chaos.

Discretization. Once the error due to particle approximation is controlled, we then obtain improved complexity guarantees by applying recent advances in the theory of log-concave sampling to the finite-particle stationary distribution. See Table 1 for a summary of our results, and the discussion in §4 for comparisons with prior works and an application to neural network training.

Once again, the importance of our framework is its *modularity*, which allows for any combination of uniform-in-time propagation of chaos bounds and log-concave sampler, provided that the finite-particle stationary distribution satisfies certain isoperimetric properties needed for the sampling guarantees. Toward this end, we also provide tools for verifying these isoperimetric properties with constants that hold independently of the number of particles (see §3.2.1).

1.2 Related Work

Mean-field equations. The McKean–Vlasov SDE was first formulated in the works [McK66; Fun84; Mél96], with origins dating to much earlier [Bol72]. It has applications in many domains, from fluid dynamics [Vil02] to game theory [LL07; CD18]; see [CD22a; CD22b] for a comprehensive survey. The kinetic version of this equation is known as the Boltzmann equation, and propagation of chaos has similarly been studied under a variety of assumptions [BGM10; Mon17; GM21; GLM22]. One prominent application within machine learning is the study of infinitely wide two-layer neural networks in the mean-field regime (see §4.2).

Propagation of chaos and sampling for (pE). The original propagation of chaos arguments of [Szn91] were first made uniform in time in [Mal01; Mal03] in both entropy and $\mathcal{W}_2$. The aforementioned works all achieve an error of order $\widetilde{\mathcal{O}}(\sqrt{k/N})$, and require a strong convexity assumption on V and W. These were later adapted for non-smooth potentials [JW17; JW18; BJW23]. Finally, [CRW22] obtained an entropic propagation of chaos bound under a higher-order smoothness assumption. See [CD22a] for a more complete bibliography.

The breakthrough result of [Lac21] obtained the sharp bound of $\widetilde{\mathcal{O}}(k/N)$ when the interaction is sufficiently weak, and this bound was made uniform in time in [LL23]. Their approach differs significantly from previous proofs by considering a local analysis based on the recursive BBGKY hierarchy. These results have been extended to other divergences, e.g., the χ^2 divergence, but without a uniform-in-time guarantee [HR23].

The question of sampling from minimizers of (pE) was first studied in [Tal96; BT97; AK02]. These works focused on the Euler–Maruyama discretization of the finite-particle system (pMV_N), under L^∞-boundedness of the gradients. Subsequently, the convergence of the Euler–Maruyama scheme has been studied in many works, including but not limited to [BH22; RES22; Li+23]. Finally, [BS23] considered a non-linear version of Hamiltonian Monte Carlo; we give a detailed comparison with their work in §4.

Propagation of chaos and sampling for (gE). The mean-field (underdamped) Langevin algorithm for minimizing (gE) was proposed and studied in [CRW22; Che+23]. Under alternative assumptions (see §3.1.2), they established propagation of chaos with a $\mathcal{O}(\sqrt{k/N})$ rate, for both the overdamped and the underdamped finite-particle approximations. Recent works from the machine learning community [NWS22; SNW22; FW23; SWN23] studied the application of these algorithms for optimizing two-layer neural networks and obtained sampling guarantees. We provide a detailed comparison with their works in §4.2.

2 Background and Notation

Let $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ be the set of probability measures on $\mathbb{R}^d$ that admit a density with respect to the Lebesgue measure and have finite second moment. We will also abuse notation and use the same symbol for a measure and its density when there is no confusion. We use superscripts for the particle index, and subscripts for the time variable. We will use $\mathcal{O}, \widetilde{\mathcal{O}}$ to signify upper bounds up to numeric constants and polylogarithms respectively. We recall the definitions of convexity and smoothness:

Definition 1. *A function $U : \mathbb{R}^d \to \mathbb{R}$ is α-uniformly convex (allowing for $\alpha \le 0$) and β-smooth if the following hold respectively*

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \ge \alpha \|x - y\|^2 \qquad \text{for all } x, y \in \mathbb{R}^d,$$
$$\|\nabla U(x) - \nabla U(y)\| \le \beta \|x - y\| \qquad \text{for all } x, y \in \mathbb{R}^d.$$

For two probability measures $\mu, \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, we define the KL divergence and the (relative) Fisher information by

$$\mathsf{KL}(\mu \,\|\, \nu) \coloneqq \mathbb{E}_\mu\big[\log \frac{\mu}{\nu}\big] \qquad \text{and} \qquad \mathsf{FI}(\mu \,\|\, \nu) \coloneqq \mathbb{E}_\mu\big[\big\|\nabla \log \frac{\mu}{\nu}\big\|^2\big],$$

with the convention $\mathsf{KL}(\mu \,\|\, \nu) = \mathsf{FI}(\mu \,\|\, \nu) = \infty$ whenever $\mu \not\ll \nu$.

We recall the definition of the log-Sobolev inequality, which is used both for propagation of chaos arguments as well as mixing time bounds.

Definition 2 (Log-Sobolev Inequality)**.** *A measure π satisfies a log-Sobolev inequality with parameter C_{LSI} if for all $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$,*

$$\mathsf{KL}(\mu \,\|\, \pi) \le \frac{C_{\mathsf{LSI}}}{2} \mathsf{FI}(\mu \,\|\, \pi). \tag{LSI}$$

When $\log(1/\pi)$ is α-uniformly convex for $\alpha > 0$, it follows from the Bakry–Émery condition that π satisfies (LSI) with constant $C_{\mathsf{LSI}} \le 1/\alpha$ [BGL14, Proposition 5.7.1].

We can also define the p-Wasserstein distance $\mathcal{W}_p(\mu, \pi)$, $p \ge 1$, between μ, π as

$$\mathcal{W}_p^p(\mu, \pi) = \inf_{\gamma \in \Gamma(\mu,\pi)} \int \|x - y\|^p \, \gamma(\mathrm{d}x, \mathrm{d}y),$$

where $\Gamma(\mu, \pi)$ is the set of all joint probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ, π respectively.

Lastly, we recall that the celebrated Otto calculus interprets the space $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, equipped with the $\mathcal{W}_2$ metric, as a formal Riemannian manifold [Ott01]. In particular, the Wasserstein gradient of a functional $\mathcal{L} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ is given as $\nabla_{\mathcal{W}_2} \mathcal{L} = \nabla \delta \mathcal{L}$. Here, $\delta \mathcal{L}$ is the first variation defined as follows: for all $\nu_0, \nu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, $\delta \mathcal{L}(\nu_0) : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$\lim_{t \searrow 0} \frac{\mathcal{L}((1-t)\,\nu_0 + t\,\nu_1) - \mathcal{L}(\nu_0)}{t} = \langle \delta \mathcal{L}(\nu_0), \nu_1 - \nu_0 \rangle \coloneqq \int \delta \mathcal{L}(\nu_0) \, \mathrm{d}(\nu_1 - \nu_0).$$

The first variation is defined up to an additive constant, but the Wasserstein gradient is unambiguous. See [AGS08] for a rigorous development. As a shorthand, we will write $\delta \mathcal{L}(\nu_0, x) \coloneqq \delta \mathcal{L}(\nu_0)(x)$ and similarly $\nabla_{\mathcal{W}_2} \mathcal{L}(\nu_0, x) \coloneqq \nabla_{\mathcal{W}_2} \mathcal{L}(\nu_0)(x)$.

2.1 SDE Systems and Their Stationary Distributions

2.1.1 The Pairwise McKean–Vlasov Setting

In the formalism introduced in the previous section, we note that (pMV) can be interpreted as Wasserstein gradient flow for (pE). In this paper, we refer to (pMV) as the *pairwise McKean–Vlasov* process. As noted in the introduction, it has the stationary distribution (1.1) which minimizes (pE). Recall also that the equation (pMV) is the mean-field limit of the finite-particle system (pMV_N). This N-particle system has the following stationary distribution: for $x^{1:N} = [x^1, \dots, x^N] \in \mathbb{R}^{d \times N}$,

$$\mu^{1:N}(x^{1:N}) \propto \exp\Big(-\frac{2}{\sigma^2} \sum_{i \in [N]} V(x^i) - \frac{1}{\sigma^2 (N-1)} \sum_{i \in [N]} \sum_{j \in [N] \setminus i} W(x^i - x^j)\Big). \tag{2.1}$$

The system (pMV_N) can be viewed as an approximation to (pMV), with the expectation term in the drift replaced by an empirical average. Note that the measure $\mu^{1:N}$ is exchangeable.[1] While the standard approach is to apply an Euler–Maruyama discretization to (pMV_N) in order to sample from (pMV), our perspective is to write more sophisticated samplers for $\mu^{1:N}$ directly. Indeed, unlike (1.1), the finite-particle stationary distribution (2.1) is explicit and amenable to sampling methods.

[1] Exchangeability refers to the property that the law of $[x^1, \dots, x^N]$ equals the law of $[x^{\sigma(1)}, \dots, x^{\sigma(N)}]$ for any permutation σ of $\{1, \dots, N\}$.

2.1.2 The General McKean–Vlasov Setting

More generally, we consider the functional (gE) where $\mathcal{F}$ is of the form $\mathcal{F}(\mu) = \mathcal{F}_0(\mu) + \frac{\lambda}{2}\int \|\cdot\|^2 \,\mathrm{d}\mu$ with $\lambda \geq 0$. The second term acts as regularization and is common in the literature [FW23; SWN23]. We can describe its Wasserstein gradient flow as the marginal law of a particle trajectory satisfying the following SDE, which we call the *general McKean–Vlasov equation*:

$$\mathrm{d}X_t = \{-\nabla_{\mathcal{W}_2}\mathcal{F}_0(\pi_t, X_t) - \lambda X_t\}\,\mathrm{d}t + \sigma\,\mathrm{d}B_t\,, \tag{gMV}$$

where $\pi_t = \mathsf{law}(X_t)$, and $\{B_t\}_{t\geq 0}$ is a standard Brownian motion on $\mathbb{R}^d$. The stationary distribution π of (gMV), and its linearization π_μ around a measure $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, satisfy the following equations:

$$\pi(x) \propto \exp\Big(-\frac{2}{\sigma^2}\,\delta\mathcal{F}_0(\pi, x) - \frac{\lambda\,\|x\|^2}{\sigma^2}\Big) \;\text{ and }\; \pi_\mu(x) \propto \exp\Big(-\frac{2}{\sigma^2}\,\delta\mathcal{F}_0(\mu, x) - \frac{\lambda\,\|x\|^2}{\sigma^2}\Big)\,. \tag{2.2}$$

The latter is called the *proximal Gibbs distribution* with respect to μ. The general dynamics corresponds to the mean-field limit of the following finite-particle system described by an N-tuple of stochastic processes $\{X_t^{1:N}\}_{t\geq 0} \coloneqq \{(X_t^1, \ldots, X_t^N)\}_{t\geq 0}$:

$$\mathrm{d}X_t^i = \{-\nabla_{\mathcal{W}_2}\mathcal{F}_0(\rho_{X_t^{1:N}}, X_t^i) - \lambda X_t^i\}\,\mathrm{d}t + \sigma\,\mathrm{d}B_t^i\,, \tag{gMV$_N$}$$

and $\rho_{x^{1:N}} = \frac{1}{N}\sum_{i=1}^N \delta_{x^i}$ is the empirical measure of the particle system. The stationary distribution for (gMV$_N$) is given as follows [CRW22, (2.16)]: for $x^{1:N} = [x^1, \ldots, x^N] \in \mathbb{R}^{d\times N}$,

$$\mu^{1:N}(x^{1:N}) \propto \exp\Big(-\frac{2N}{\sigma^2}\,\mathcal{F}_0(\rho_{x^{1:N}}) - \frac{\lambda}{\sigma^2}\,\|x^{1:N}\|^2\Big)\,. \tag{2.3}$$

One can show that $\nabla_{x^i}\mathcal{F}_0(\rho_{x^{1:N}}) = \frac{1}{N}\nabla_{\mathcal{W}_2}\mathcal{F}_0(\rho_{x^{1:N}}, x^i)$, and hence (gMV$_N$) is simply the Langevin diffusion corresponding to stationary measure (2.3). Moreover, when $\lambda = 0$ and choosing $\mathcal{F}_0(\mu) = \int V(x)\,\mu(\mathrm{d}x) + \iint W(x-y)\,\mu(\mathrm{d}x)\,\mu(\mathrm{d}y)$, then the equations (gMV), (2.2), (gMV$_N$), and (2.3) reduce to (pMV), (1.1), (pMV$_N$), and (2.1), respectively.

3 Technical Ingredients

Our general approach for sampling from the stationary distribution π in either (1.1) or (2.2) is to directly apply an off-the-shelf sampler for the finite-particle stationary distribution $\mu^{1:N}$. The theoretical guarantees for this procedure require two main ingredients: (1) control of the "bias"—i.e., the error incurred by approximating π by the 1-particle marginal of $\mu^{1:N}$—and (2) verification of isoperimetric properties which allow for fast sampling from the measure $\mu^{1:N}$.

3.1 Bias Control via Uniform-in-Time Propagation of Chaos

In this section, we focus on the first ingredient, namely, obtaining control of the bias via uniform-in-time propagation of chaos results. Proofs for this section are given in §A.

3.1.1 Pairwise McKean–Vlasov Setting

We first consider the pairwise McKean–Vlasov setting described in §2.1.1. Our first propagation of chaos result uses the following three assumptions.

Assumption 1. *The potentials V, W are β_V, β_W-smooth respectively.*

Assumption 2. *The distribution π satisfies (LSI) with parameter $C_{\mathsf{LSI}}(\pi)$.*

Assumption 3. *The ratio $\rho \coloneqq \sigma^4/8\beta_W^2 C_{\mathsf{LSI}}^2(\pi)$ is at least 3.*

Remark. Note that from (1.1), we typically would expect $C_{\mathsf{LSI}}^2(\pi)$ to also scale as σ^4 (e.g., in the case when V and W are α-uniformly convex for $\alpha > 0$). Therefore, Assumption 3 is typically invariant to the scaling of σ and can be satisfied even for $\sigma \searrow 0$. Under these assumptions, we obtain the following result via a modification of the argument of [LL23]. Compared to their work, we remove the assumption of boundedness of $\|\nabla W\|$, and our argument is simpler since we apply it directly to the stationary distribution.

Theorem 3 (Sharp Propagation of Chaos)**.** *Under Assumptions 1, 2 and 3, for any $N \geq 100$ and $k \in [N]$, it holds that $\mathsf{KL}(\mu^{1:k} \parallel \pi^{\otimes k}) = \widetilde{\mathcal{O}}(dk^2/N^2)$. Thus, $\mathsf{KL}(\mu^{1:k} \parallel \pi^{\otimes k}) < \varepsilon^2$ if*

$$N \geq 100 \vee \widetilde{\Omega}\big(k\sqrt{d}\,\varepsilon^{-1}\big)\,. \tag{3.1}$$

We note that the rate in Theorem 3 is sharp; see the Gaussian case in Example 10. A condition such as Assumption 3 is in general necessary, since otherwise the minimizer of (pE) may not even be unique [see the example and discussion in LL23]. However, it can be restrictive, as it requires the interaction to be sufficiently weak. With the following convexity assumption, we can obtain a propagation of chaos result without Assumption 3.

Assumption 4. *The potentials V, W are α_V, α_W-uniformly convex with $\alpha_V + \alpha_W^- > 0$. Here, $\alpha_W^- \coloneqq \alpha_W \wedge 0$ denotes the negative part of α_W.*

The following weaker result consists of two parts. The first, a Wasserstein propagation of chaos result, is based on [Szn91]. The second, building on the first, is a uniform-in-time entropic propagation of chaos bound following from a Fisher information bound. The arguments are similar to those in [Mal01; Mal03], albeit simplified (since we work at stationarity) and presented here with explicit constants.

Theorem 4 (Weak Propagation of Chaos)**.** *Under Assumptions 1 and 4, for any $N \geq \frac{\alpha_V - \alpha_W^-}{\alpha_V + \alpha_W^-} \vee 2$, if we denote $\alpha \coloneqq \alpha_V + \alpha_W^-$, then*

$$\mathcal{W}_2^2(\mu^{1:k}, \pi^{\otimes k}) \leq \frac{4\beta_W^2 \sigma^2 d}{\alpha^3}\,\frac{k}{N}\,, \tag{3.2}$$

$$\mathsf{KL}(\mu^{1:k} \parallel \pi^{\otimes k}) \leq \frac{\sigma^2}{4\alpha}\,\mathsf{FI}(\mu^{1:k} \parallel \pi^{\otimes k}) \leq \frac{132\beta_W^2\,(\beta_V + \beta_W)^2\,d}{\alpha^4}\,\frac{k}{N}\,. \tag{3.3}$$

3.1.2 General McKean–Vlasov Setting

In the more general case where we aim to minimize (gE) for a generic functional $\mathcal{F}$ of the form $\mathcal{F}(\mu) = \mathcal{F}_0(\mu) + \frac{\lambda}{2}\int \|\cdot\|^2\,\mathrm{d}\mu$, we impose the following assumptions. They can be largely seen as generalizations of the conditions for the pairwise case, and they are inherited from [CRW22; SWN23]. There is an additional convexity condition (Assumption 5), which in the pairwise McKean–Vlasov setting amounts to positive semidefiniteness of the kernel $(x, y) \mapsto W(x - y)$ on $\mathbb{R}^d \times \mathbb{R}^d$; thus, in general, the following assumptions are incomparable with the ones in §3.1.1.

Assumption 5. *The functional $\mathcal{F}_0$ is convex in the usual sense. For all $\nu_0, \nu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, $t \in [0, 1]$,*

$$\mathcal{F}_0((1-t)\,\nu_0 + t\,\nu_1) \leq (1-t)\,\mathcal{F}_0(\nu_0) + t\,\mathcal{F}_0(\nu_1)\,.$$

Assumption 6. *The functional $\mathcal{F}$ is smooth in the sense that for all $x, y \in \mathbb{R}^d$, $\nu, \nu' \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, there is a uniform constant β such that*

$$\|\nabla_{\mathcal{W}_2}\mathcal{F}(\nu, x) - \nabla_{\mathcal{W}_2}\mathcal{F}(\nu', y)\| \leq \beta\,(\|x - y\| + \mathcal{W}_1(\nu, \nu'))\,.$$

Assumption 7. *The proximal Gibbs measures satisfy (LSI) with a uniform constant: namely, it holds that $C_{\mathsf{LSI}}(\pi) \vee \sup_{\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)} C_{\mathsf{LSI}}(\pi_\mu) \leq \overline{C}_{\mathsf{LSI}}$.*

Remark. These assumptions taken together cover settings not covered in the preceding sections, including optimization of two-layer neural networks. See [CRW22, Remark 3.1] and §4.2.

Under these assumptions, we can derive an entropic propagation of chaos bound by following the proof of [CRW22]. Through a tighter analysis, we are able to reduce the dependence on the condition number $\kappa \coloneqq \overline{C}_{\mathsf{LSI}}\beta/\sigma^2$ from κ^2 to κ.

Theorem 5 (Propagation of Chaos for General Functionals)**.** *Under Assumptions 5, 6, and 7, for* $N \geq {}^{160\beta\overline{C}_{\mathsf{LSI}}}/_{\sigma^2}$*, we have*

$$\frac{1}{2\overline{C}_{\mathsf{LSI}}}\mathcal{W}_2^2(\mu^{1:k}, \pi^{\otimes k}) \leq \mathsf{KL}(\mu^{1:k} \,\|\, \pi^{\otimes k}) \leq \frac{33\beta\overline{C}_{\mathsf{LSI}}dk}{\sigma^2 N}\,.$$

Among these assumptions, the hardest to verify is the uniform LSI of Assumption 7. Following [SWN23], we introduce the following sufficient condition for the validity of Assumption 7; see Lemma 21 for a more precise statement.

Assumption 8. *There exists a uniform bound on the Wasserstein gradient of the interaction term* $\mathcal{F}_0$*: for some constant* $B < \infty$ *and all* $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$*,* $x \in \mathbb{R}^d$*,* $\|\nabla_{\mathcal{W}_2}\mathcal{F}_0(\mu, x)\| \leq B$*.*

Lemma 6 (Informal)**.** *Assumptions 6 and 8 imply Assumption 7 with an explicit constant* $\overline{C}_{\mathsf{LSI}}$*, given in terms of* d*,* B*,* β*,* λ*, and* σ*.*

3.2 Isoperimetric Properties of the Stationary Distributions

In this section, we verify the isoperimetric properties of $\pi, \mu^{1:N}$, with proofs provided in §B.

3.2.1 Pairwise McKean–Vlasov Setting

If V, W satisfy Assumptions 1 and 4 (i.e., V and W have bounded Hessians), then the potential for (1.1), i.e., $\log(1/\pi)$, is $\frac{2}{\sigma^2}(\alpha_V + \alpha_W)$-convex and $\frac{2}{\sigma^2}(\beta_V + \beta_W)$-smooth. By the Bakry–Émery condition, π satisfies (LSI) with parameter $C_{\mathsf{LSI}}(\pi) \leq {}^{\sigma^2}/_{2\,(\alpha_V+\alpha_W)}$.

Similarly, for the invariant measure $\mu^{1:N}$ in (2.1), we can prove the following.

Lemma 7. *If* V *and* W *satisfy Assumption 1, then* $\log(1/\mu^{1:N})$ *is* $\frac{2}{\sigma^2}\left(\beta_V + \frac{N}{N-1}\beta_W\right)$*-smooth.*

If V *and* W *satisfy Assumption 4, then* $\log(1/\mu^{1:N})$ *is* $\frac{2}{\sigma^2}\left(\alpha_V + \frac{N}{N-1}\alpha_W^-\right)$*-convex.*[2]

We now consider the non-log-concave case. It is standard in the sampling literature that the assumption of (LSI) for the stationary distribution yields mixing time guarantees. Since our strategy is to sample from (2.1), we therefore seek an LSI for $\mu^{1:N}$, formalized as the following assumption.

Assumption 9. *The distribution* $\mu^{1:N}$ *satisfies* (LSI) *with parameter* $C_{\mathsf{LSI}}(\mu^{1:N})$*.*

In this section, we provide an easily verifiable condition, based on the Holley–Stroock condition [HS87], for this assumption to hold with an N-independent constant.

Assumption 10. *The potentials* V *and* W *can be decomposed as* $V = V_0 + V_1$ *and* $W = W_0 + W_1$ *such* V_0*,* W_0 *satisfy Assumption 4 and* $\mathsf{osc}(V_1), \mathsf{osc}(W_1) < \infty$*, where for a function* $U : \mathbb{R}^d \to \mathbb{R}$ *we define* $\mathsf{osc}(U) := \sup U - \inf U$*.*

Under this assumption, a careful application of the Holley–Stroock perturbation principle yields the following lemma.

Lemma 8. *Under Assumption 10,* $\pi, \mu^{1:N}$ *satisfy* (LSI) *with parameters*

$$C_{\mathsf{LSI}}(\pi) \leq \frac{\sigma^2}{2\,(\alpha_{V_0} + \alpha_{W_0})} \exp\Big(\frac{2}{\sigma^2}\left(\mathsf{osc}(V_1) + \mathsf{osc}(W_1)\right)\Big)$$

$$C_{\mathsf{LSI}}(\mu^{1:N}) \leq \frac{\sigma^2}{2\,(\alpha_{V_0} + \frac{N}{N-1}\alpha_{W_0}^-)} \exp\Big(\frac{2}{\sigma^2}\left(\mathsf{osc}(V_1) + \mathsf{osc}(W_1)\right)\Big)\,.$$

[2] Only the negative part of α_W contributes to the strong log-concavity of $\mu^{1:N}$. This is consistent with [Vil03, Theorem 5.15], which asserts that when $\alpha_W > 0$, the interaction energy $\mu \mapsto \iint W(x-y)\,\mu(\mathrm{d}x)\,\mu(\mathrm{d}y)$ is α_W-strongly displacement convex over the subspace of probability measures with fixed mean, but only weakly convex over the full Wasserstein space.

Algorithm	"Metric"	Assumptions	M	N
LMC	$\sqrt{\alpha}/\sigma\,\mathcal{W}_2$	1, 2, 3, 9	$\kappa^2 d/\varepsilon^2$	$d^{1/2}/\varepsilon$
MALA–PS			$\kappa d^{3/4}/\varepsilon^{1/2}$	
ULMC–PS			$\kappa^{3/2} d^{1/2}/\varepsilon$	
ULMC$^+$	$\sqrt{\alpha}/\sigma\,\mathcal{W}_2$	1, 3, 4	$\kappa d^{1/3}/\varepsilon^{2/3}$	$d^{1/2}/\varepsilon$
LMC	$\sqrt{\mathsf{KL}}$	1, 4	$\kappa^2 d/\varepsilon^2$	$\kappa^4 d/\varepsilon^2$
ULMC			$\kappa^{3/2} d^{1/2}/\varepsilon$	
LMC	$\sqrt{\alpha}/\sigma\,\mathcal{W}_2$		$\kappa d/\varepsilon^2$	$\kappa^2 d/\varepsilon^2$
ULMC$^+$			$\kappa d^{1/3}/\varepsilon^{2/3}$	
LMC	$\sqrt{\mathsf{KL}}$	5, 6, 7, 9	$\kappa^2 d/\varepsilon^2$	$\kappa d/\varepsilon^2$
ULMC–PS			$\kappa^{3/2} d^{1/2}/\varepsilon$	

Table 1: In this table, we record M, the total number of oracle queries to $\log \mu^{1:N}$ made by the log-concave sampler, and N, the number of particles.

3.2.2 General McKean–Vlasov Setting

In the setting (gE) with $\mathcal{F}(\mu) = \mathcal{F}_0(\mu) + \frac{\lambda}{2}\int \|\cdot\|^2 \,\mathrm{d}\mu$, we verify that Assumption 8 yields (LSI) for $\mu^{1:N}$ with an N-independent constant. See Lemma 22 for a more precise statement.

Lemma 9 (Informal)**.** *In the mean-field Langevin setting of §2.1.2, suppose that Assumption 8 holds. Then, Assumption 9 holds with $C_{\mathsf{LSI}}(\mu^{1:N})$ depending on d, B, λ, and σ, but not on N.*

4 Sampling from the Mean-Field Target

In this section, we present results for sampling from π. As outlined in Algorithm 1, we use off-the-shelf log-concave samplers to sample from $\mu^{1:N}$, during which we access the first-order[3] oracle for $\mu^{1:N}$ (i.e., an oracle for evaluation of $\log \mu^{1:N}$ up to an additive constant, and for evaluation of $\nabla \log \mu^{1:N}$). For N sufficiently large, the first particle given by Algorithm 1 is approximately distributed according to π: for $\hat{\mu}^{1:N}$ the law of the output of the log-concave sampler and its 1-particle marginal distribution $\hat{\mu}^1$,

$$\mathcal{W}_2(\hat{\mu}^1, \pi) \le \mathcal{W}_2(\hat{\mu}^1, \mu^1) + \mathcal{W}_2(\mu^1, \pi) \le \sqrt{\frac{1}{N}}\,\mathcal{W}_2(\hat{\mu}^{1:N}, \mu^{1:N}) + \mathcal{W}_2(\mu^1, \pi)\,,$$

where the inequality follows from exchangeability (Lemma 23). A similar decomposition also holds for KL, although the argument is more technical. We defer its presentation to §E.

Algorithm 1 Sampling from the Mean-Field Stationary Distribution

Input: the number N of total particles, a log-concave sampler LC-Sampler
Output: k particles $\hat{X}^{1:k}$

1: Sample $\hat{X}^{1:N} \sim \hat{\mu}^{1:N}$ via LC-Sampler, so that $\hat{\mu}^{1:N} \approx \mu^{1:N}$, e.g., in $\mathcal{W}_2$ or $\sqrt{\mathsf{KL}}$.
2: Output the first k particles $\hat{X}^{1:k}$.

To bound the second term by ε, it suffices to choose N according to the propagation of chaos results in §3.1. Our results are summarized in Table 1, in which we record **the total number of oracle calls M for $\mu^{1:N}$ made by the sampler** M and **the number of particles** N needed to achieve ε error in the desired metric, hiding polylogarithmic factors. Note that in the pairwise McKean–Vlasov setting, each oracle call to $\mu^{1:N}$ requires N calls to an oracle for V, and $\binom{N}{2}$ calls to an oracle for W.

The algorithms in the table refer to: Langevin Monte Carlo (LMC); underdamped Langevin Monte Carlo (ULMC); discretizations of the underdamped Langevin diffusion via the randomized midpoint method [SL19] or the shifted ODE method [FLO21] (ULMC$^+$); and implementation of the proximal sampler [LST21; Che+22a] via the Metropolis-adjusted Langevin algorithm or via ULMC (MALA–PS and ULMC–PS respectively).

[3]For our results involving the proximal sampler, we also assume access to a proximal oracle for simplicity.

Note that LMC applied to sample from $\mu^{1:N}$ is simply the Euler–Maruyama discretization of (pMV_N), and likewise ULMC is the algorithm considered in [FW23]. We refer to §E for proofs and references.

To streamline the rates, we simplify the notation by defining $\beta = \beta_V + \beta_W$ if Assumption 1 holds, otherwise we use the value from Assumption 6. We let $\alpha = \alpha_V + \alpha_W^-$ under Assumption 4, $\alpha = \sigma^2/2\max\{C_{\mathsf{LSI}}(\mu^{1:N}), C_{\mathsf{LSI}}(\pi)\}$ under Assumptions 2 and 9, and $\alpha = \sigma^2/2\max\{C_{\mathsf{LSI}}(\mu^{1:N}), \overline{C}_{\mathsf{LSI}}\}$ in the general McKean–Vlasov setting. Finally, we let $\kappa \coloneqq \beta/\alpha$ denote the condition number. The additional assumption $\kappa \le \sqrt{d}/\varepsilon$ will be used to simplify some of the rates.

In the following subsections, we discuss some of the results in greater detail.

4.1 Pairwise McKean–Vlasov Setting

Example 10 (Gaussian Case)**.** Consider a quadratic confinement and interaction,

$$V(x) = \frac{1}{2}\, x^\mathsf{T} A x = \frac{1}{2}\, \|x\|_A^2\,, \qquad W(x) = \frac{\lambda}{2}\, \|x\|^2\,,$$

for some matrix $A \in \mathbb{R}^{d\times d}$ with $A \succ 0$, $\lambda \ge 0$. The resulting stationary distributions can be calculated explicitly to be Gaussians. We show in §C that for large N, $\mathsf{KL}(\mu^{1:k} \parallel \pi^{\otimes k}) = \widetilde{\Theta}(dk^2/N^2)$. This shows that the rate in Theorem 3 is sharp.

Example 11 (Strongly Convex Case)**.** Consider the strongly convex case where $\alpha = \alpha_V + \alpha_W^- > 0$. The prior work [BS23] also considered the problem of sampling from the mean-field stationary distribution π, with $\sigma^2 = 2$. If we count the number of calls to a gradient oracle for V, their complexity bound reads $\widetilde{\mathcal{O}}(\kappa^{5/3}d^{4/3}/\varepsilon^{8/3})$ to achieve $\sqrt{\alpha}/\sigma\, \mathcal{W}_1(\hat{\mu}^1, \pi) \le \varepsilon$. We note that their assumptions are not strictly comparable to ours. They require the interaction W to be sufficiently weak, in the sense that $\beta_W \lesssim \alpha$, which is similar[4] to our Assumption 3; on the other hand, they only assume $\alpha_V > 0$, rather than $\alpha_V + \alpha_W^- > 0$. Nevertheless, we attempt to make some comparisons with their work below.

Without Assumption 3, ULMC$^+$ achieves $\sqrt{\alpha}/\sigma\, \mathcal{W}_2(\hat{\mu}^1, \pi) \le \varepsilon$ with complexity $\widetilde{\mathcal{O}}(\kappa^3 d^{4/3}/\varepsilon^{8/3})$, which matches the guarantee of [BS23] up to the dependence on κ. We can also obtain guarantees in $\sqrt{\mathsf{KL}}$, at the cost of an extra factor of κ^2.

With Assumption 3, MALA–PS has complexity $\widetilde{\mathcal{O}}(\kappa d^{5/4}/\varepsilon^{3/2})$ and ULMC$^+$ has complexity $\widetilde{\mathcal{O}}(\kappa d^{5/6}/\varepsilon^{5/3})$, which improve substantially upon [BS23].

To summarize, in the strongly convex case, we have obtained numerous improvements: (i) we can obtain results even without the weak interaction condition (Assumption 3); (ii) when we assume the weak interaction condition, we obtain improved complexities; (iii) our results hold in stronger metrics; (iv) our approach is generic, allowing for the consideration of numerous different samplers without needing to establish new propagation of chaos results (by way of comparison, [BS23] developed a tailored propagation of chaos argument for their non-linear Hamiltonian Monte Carlo algorithm).

Example 12 (Bounded Perturbations)**.** Both the results of [BS23] as well as our own allow for non-convex potentials, albeit under different assumptions—[BS23] require strong convexity at infinity, whereas we require (LSI) for the stationary measures $\mu^{1:N}$ and π. In order to obtain sampling guarantees with low complexity, it is important for the LSI constant of $\mu^{1:N}$ to be independent of N. We have provided a sufficient condition for this to hold: V and W are bounded perturbations of V_0 and W_0 respectively, where $\alpha_{V_0} + \alpha_{W_0}^- > 0$; see Lemma 8.

We also note that in this setting, both of our works require a weak interaction condition. This is in general necessary in order to ensure uniqueness of the mean-field stationary distribution, see the discussion in §3.1.1.

4.2 General McKean–Vlasov Setting

Example 13 (General Functionals)**.** In the general setting, under Assumptions 5, 6, and 7, the work of [SWN23] provided the first discretization bounds. They impose further assumptions and their resulting complexity bound is rather complicated, but it reads roughly $MN = \widetilde{\mathcal{O}}(\mathrm{poly}(\kappa)\, d^2/\varepsilon^4)$ for the discretization

[4] See eq. (2.24) therein; note that they have a scaling factor of ε in front of their interaction term, so that our parameter β_W is equivalent to their $\varepsilon\tilde{L}$.

of (gMV_N). Subsequently, [FW23] obtained an improved complexity of $MN = \widetilde{\mathcal{O}}(\kappa^4 d^{3/2}/\varepsilon^3)$ via ULMC in the averaged TV distance. In comparison, we can improve this complexity guarantee to $\widetilde{\mathcal{O}}(\kappa^{5/2} d^{3/2}/\varepsilon^3)$, and the guarantee even holds in $\sqrt{\mathsf{KL}}$ if we combine ULMC with the proximal sampler. It appears that we gain one factor of $\sqrt{\kappa}$ through sharper discretization analysis (via [Zha+23], or via the error analysis of the proximal sampler in [AC23]), and one factor of κ via a sharper propagation of chaos result (Theorem 5).

We also note that the result of [FW23] is based on a kinetic version of the propagation of chaos argument from [Che+23], whereas our approach uses the original "non-kinetic" argument from [CRW22] in the form of Theorem 5.

Application to Two-Layer Neural Networks. Let us consider the problem of learning a two-layer neural network in the mean-field regime. Let $f_\theta : \mathbb{R}^d \to \mathbb{R}$ be a function parameterized by $\theta \in \mathbb{R}^p$, and for any probability measure μ over $\mathbb{R}^p$, let $f_\mu \coloneqq \int f_\theta \, \mu(\mathrm{d}\theta)$. For example, in a standard two-layer neural network, we take $\theta = (a, w) \in \mathbb{R} \times \mathbb{R}^d$ and $f_{a,w}(x) = a \, \mathsf{ReLU}(\langle w, x \rangle)$. When $\mu = \frac{1}{m} \sum_{j=1}^m \delta_{(a_j, w_j)}$ is an empirical measure, then f_μ is the function computed by a two-layer neural network with m hidden neurons. In this formulation, however, we can take μ to be any probability measure, corresponding to the *mean-field limit* $m \to \infty$ [CB18; MMN18; Chi22; RV22; SS20].

Given a dataset $\{(x_i, y_i)\}_{i=1}^n$ in $\mathbb{R}^d \times \mathbb{R}$ and a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we can formulate neural network training as the problem of minimizing the loss $\mu \mapsto \sum_{i=1}^n \ell(f_\mu(x_i), y_i)$. To place this within the general McKean–Vlasov framework, we add two regularization terms: (1) $\frac{\lambda}{2} \int \|\cdot\|^2 \, \mathrm{d}\mu$ corresponds to weight decay; and (2) $\frac{\sigma^2}{2} \int \log \mu \, \mathrm{d}\mu$ is entropic regularization. We are now in the setting of §2.1.2, with $\mathcal{F}_0(\mu) = \sum_{i=1}^n \ell(f_\mu(x_i), y_i)$. To minimize this energy, it is natural to consider the Euler–Maruyama discretization of (gMV_N), which corresponds to learning the neural network via noisy GD, and was considered in [SWN23]. Recent works [Che+23; FW23] also considered the underdamped version of (gMV) and its discretization. Under the assumptions common to those works as well as our own, our results yield improved algorithmic guarantees for this task (see Example 13).

Unfortunately, the assumptions used for the analysis of the general McKean–Vlasov are restrictive and limit the applicability to neural network training. For example, it suffices for ℓ to be convex in its first argument (to satisfy Assumption 5), to have two bounded derivatives (w.r.t. its first argument), and for $\theta \mapsto f_\theta(x_i)$ to have two bounded derivatives for each x_i. The last condition is satisfied, e.g., for $f_\theta(x) = \tanh(\langle \theta, x \rangle)$. For a genuinely two-layer example, we can take $f_\theta(x) = \tanh(a) \tanh(\langle w, x \rangle)$ for $\theta = (a, w) \in \mathbb{R} \times \mathbb{R}^d$. Under these conditions, Assumptions 6 and 8 hold, which in turn furnish log-Sobolev inequalities via Lemmas 6 and 9. In general, these LSI constants depend exponentially on quantities such as the dimension p of the parameter space, which is unavoidable as global non-convex optimization is intractable in the worst case. However, we note that these limitations are inherited from the prior literature and not specific to our approach.

5 Conclusion

In this work, we propose a framework for obtaining sampling guarantees for the minimizers of (pE) and (gE), based on decoupling the problem into *(i)* particle approximation via *propagation of chaos*, and *(ii)* time-discretization via *log-concave sampling theory*. Our approach leads to simpler proofs and improved guarantees compared to previous works, and our results readily benefit from any improvements in either *(i)* or *(ii)*.

We conclude by listing some future directions of study. We believe there is further room for improvement in the propagation of chaos results. For example, can the sharp rate in Theorem 3 be extended to stronger metrics such as Rényi divergences, as well as to situations when the weak interaction condition (Assumption 3) fails, e.g., in the strongly displacement convex case or in the setting of §3.1.2? For the sampling guarantees, the prior works [BS23; SWN23] considered different settings, such as potentials satisfying convexity at infinity or the use of stochastic gradients; these extensions are compatible with our approach and could possibly lead to improvements in these cases, as well as others. Finally, it would be interesting to develop further applications of sampling from the mean-field distribution, e.g., applications to neural network training under less stringent hypotheses.

Acknowledgements

YK was supported in part by NSF awards CCF-2007443 and CCF-2134105. MSZ was supported by NSERC through the CGS-D program. SC was supported by the Eric and Wendy Schmidt Fund at the Institute for Advanced Study. MAE was supported by NSERC Grant [2019-06167] and CIFAR AI Chairs program at the Vector Institute. MBL was supported by NSF grant DMS-2133806.

References

[AC23] J. M. Altschuler and S. Chewi. "Faster high-accuracy log-concave sampling via algorithmic warm starts". In: *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. 2023, pp. 2169–2176.

[AGS08] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Second. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2008, pp. x+334.

[AK02] F. Antonelli and A. Kohatsu-Higa. "Rate of convergence of a particle method to the solution of the McKean–Vlasov equation". In: *The Annals of Applied Probability* 12.2 (2002), pp. 423–476.

[BGL14] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 103. Springer, 2014.

[BH22] J. Bao and X. Huang. "Approximations of McKean–Vlasov stochastic differential equations with irregular coefficients". In: *Journal of Theoretical Probability* (2022), pp. 1–29.

[BGM10] F. Bolley, A. Guillin, and F. Malrieu. "Trend to equilibrium and particle approximation for a weakly self consistent Vlasov–Fokker–Planck equation". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 44.5 (2010), pp. 867–884.

[Bol72] L. Boltzmann. "Further studies on the heat balance among gas molecules". In: *History of Modern Physical Sciences* 1 (1872), pp. 262–349.

[BT97] M. Bossy and D. Talay. "A stochastic particle method for the McKean–Vlasov and the Burgers equation". In: *Mathematics of Computation* 66.217 (1997), pp. 157–192.

[BS23] N. Bou-Rabee and K. Schuh. "Nonlinear Hamiltonian Monte Carlo & its particle approximation". In: *arXiv preprint arXiv:2308.11491* (2023).

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013, pp. x+481.

[BJW23] D. Bresch, P.-E. Jabin, and Z. Wang. "Mean field limit and quantitative estimates with singular attractive kernels". In: *Duke Mathematical Journal* 172.13 (2023), pp. 2591–2641.

[CD18] R. Carmona and F. Delarue. *Probabilistic theory of mean field games with applications I–II*. Springer, 2018.

[CG22] P. Cattiaux and A. Guillin. "Functional inequalities for perturbed measures with applications to log-concave measures and to some Bayesian problems". In: *Bernoulli* 28.4 (2022), pp. 2294–2321.

[CD22a] L.-P. Chaintron and A. Diez. "Propagation of chaos: a review of models, methods and applications. I. Models and methods". In: *Kinet. Relat. Models* 15.6 (2022), pp. 895–1015.

[CD22b] L.-P. Chaintron and A. Diez. "Propagation of chaos: a review of models, methods and applications. II. Applications". In: *Kinet. Relat. Models* 15.6 (2022), pp. 1017–1173.

[Che+23] F. Chen, Y. Lin, Z. Ren, and S. Wang. "Uniform-in-time propagation of chaos for kinetic mean field Langevin dynamics". In: *arXiv preprint arXiv:2307.02168* (2023).

[CRW22] F. Chen, Z. Ren, and S. Wang. "Uniform-in-time propagation of chaos for mean field Langevin dynamics". In: *arXiv preprint arXiv:2212.03050* (2022).

[Che+22a] Y. Chen, S. Chewi, A. Salim, and A. Wibisono. "Improved analysis for a proximal algorithm for sampling". In: *Proceedings of Thirty Fifth Conference on Learning Theory (COLT)*. Ed. by P.-L. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 2984–3014.

[Che23] S. Chewi. "Log-concave sampling". Book draft available at https://chewisinho.github.io. 2023.

[Che+22b] S. Chewi, M. A. Erdogdu, M. Li, R. Shen, and M. Zhang. "Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev". In: *Proceedings of Thirty Fifth Conference on Learning Theory (COLT)*. Ed. by P.-L. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 1–2.

[Che+21] S. Chewi, C. Lu, K. Ahn, X. Cheng, T. Le Gouic, and P. Rigollet. "Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm". In: *Conference on Learning Theory (COLT)*. PMLR. 2021, pp. 1260–1300.

[Chi22] L. Chizat. "Mean-field Langevin dynamics: exponential convergence and annealing". In: *Transactions on Machine Learning Research* (2022).

[CB18] L. Chizat and F. Bach. "On the global convergence of gradient descent for over-parameterized models using optimal transport". In: *Advances in Neural Information Processing Systems (NeurIPS)* 31 (2018).

[Cla+23] J. Claisse, G. Conforti, Z. Ren, and S. Wang. "Mean field optimization problem regularized by Fisher information". In: *arXiv preprint 2302.05938* (2023).

[Csi84] I. Csiszár. "Sanov property, generalized I-projection and a conditional limit theorem". In: *The Annals of Probability* (1984), pp. 768–793.

[DKR22] A. S. Dalalyan, A. Karagulyan, and L. Riou-Durand. "Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets". In: *J. Mach. Learn. Res.* 23 (2022), Paper No. 235, 38.

[Dia+23] M. Z. Diao, K. Balasubramanian, S. Chewi, and A. Salim. "Forward-backward Gaussian variational inference via JKO in the Bures–Wasserstein space". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 7960–7991.

[DMM19] A. Durmus, S. Majewski, and B. Miasojedow. "Analysis of Langevin Monte Carlo via convex optimization". In: *J. Mach. Learn. Res.* 20 (2019), Paper No. 73, 46.

[FYC23] J. Fan, B. Yuan, and Y. Chen. "Improved dimension dependence of a proximal algorithm for sampling". In: *Proceedings of Thirty Sixth Conference on Learning Theory (COLT)*. Ed. by G. Neu and L. Rosasco. Vol. 195. Proceedings of Machine Learning Research. PMLR, 2023, pp. 1473–1521.

[FLO21] J. Foster, T. Lyons, and H. Oberhauser. "The shifted ODE method for underdamped Langevin MCMC". In: *arXiv preprint 2101.03446* (2021).

[FW23] Q. Fu and A. Wilson. "Mean-field underdamped Langevin dynamics and its space-time discretization". In: *arXiv preprint arXiv:2312.16360* (2023).

[Fun84] T. Funaki. "A certain class of diffusion processes associated with nonlinear parabolic equations". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 67.3 (1984), pp. 331–348.

[GLM22] A. Guillin, P. Le Bris, and P. Monmarché. "Convergence rates for the Vlasov–Fokker–Planck equation and uniform in time propagation of chaos in non convex cases". In: *Electronic Journal of Probability* 27 (2022), pp. 1–44.

[GM21] A. Guillin and P. Monmarché. "Uniform long-time and propagation of chaos estimates for mean field kinetic particles in non-convex landscapes". In: *Journal of Statistical Physics* 185 (2021), pp. 1–20.

[HBE20] Y. He, K. Balasubramanian, and M. A. Erdogdu. "On the ergodicity, bias and asymptotic normality of randomized midpoint sampling method". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 7366–7376.

[HR23] E. Hess-Childs and K. Rowan. "Higher-order propagation of chaos in L^2 for interacting diffusions". In: *arXiv preprint arXiv:2310.09654* (2023).

[HS87] R. Holley and D. Stroock. "Logarithmic Sobolev inequalities and stochastic Ising models". In: *J. Statist. Phys.* 46.5-6 (1987), pp. 1159–1194.

[JW17] P.-E. Jabin and Z. Wang. "Mean field limit for stochastic particle systems". In: *Active Particles, Volume 1: Advances in Theory, Models, and Applications* (2017), pp. 379–402.

[JW18] P.-E. Jabin and Z. Wang. "Quantitative estimates of propagation of chaos for stochastic systems with $W^{-1,\infty}$ kernels". In: *Invent. Math.* 214.1 (2018), pp. 523–591.

[JCP23] Y. Jiang, S. Chewi, and A.-A. Pooladian. "Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space". In: *arXiv preprint 2312.02849* (2023).

[JKO98] R. Jordan, D. Kinderlehrer, and F. Otto. "The variational formulation of the Fokker–Planck equation". In: *SIAM J. Math. Anal.* 29.1 (1998), pp. 1–17.

[Kuh+19] D. Kuhn, P. M. Esfahani, V. Nguyen, and S. Shafieezadeh-Abadeh. "Wasserstein distributionally robust optimization: theory and applications in machine learning". In: *Operations Research & Management Science in the Age of Analytics*. 2019. Chap. 6, pp. 130–166.

[Lac21] D. Lacker. "Hierarchies, entropy, and quantitative propagation of chaos for mean field diffusions". In: *arXiv preprint arXiv:2105.02983* (2021).

[Lac23] D. Lacker. "Independent projections of diffusions: gradient flows for variational inference and optimal mean field approximations". In: *arXiv preprint 2309.13332* (2023).

[LL23] D. Lacker and L. Le Flem. "Sharp uniform-in-time propagation of chaos". In: *Probability Theory and Related Fields* (2023), pp. 1–38.

[Lam+22] M. Lambert, S. Chewi, F. Bach, S. Bonnabel, and P. Rigollet. "Variational inference via Wasserstein gradient flows". In: *Advances in Neural Information Processing Systems*. Ed. by A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho. 2022.

[LL07] J.-M. Lasry and P.-L. Lions. "Mean field games". In: *Jpn. J. Math.* 2.1 (2007), pp. 229–260.

[LST21] Y. T. Lee, R. Shen, and K. Tian. "Structured logconcave sampling with a restricted Gaussian oracle". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by M. Belkin and S. Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 2021, pp. 2993–3050.

[Li+23] Y. Li, X. Mao, Q. Song, F. Wu, and G. Yin. "Strong convergence of Euler–Maruyama schemes for McKean–Vlasov stochastic differential equations under local Lipschitz conditions of state variables". In: *IMA Journal of Numerical Analysis* 43.2 (2023), pp. 1001–1035.

[LW16] Q. Liu and D. Wang. "Stein variational gradient descent: a general purpose Bayesian inference algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc., 2016.

[Mal01] F. Malrieu. "Logarithmic Sobolev inequalities for some nonlinear PDE's". In: *Stochastic Process. Appl.* 95.1 (2001), pp. 109–132.

[Mal03] F. Malrieu. "Convergence to equilibrium for granular media equations and their Euler schemes". In: *The Annals of Applied Probability* 13.2 (2003), pp. 540–560.

[McK66] H. P. McKean Jr. "A class of Markov processes associated with nonlinear parabolic equations". In: *Proc. Nat. Acad. Sci. U.S.A.* 56 (1966), pp. 1907–1911.

[MMN18] S. Mei, A. Montanari, and P.-M. Nguyen. "A mean field view of the landscape of two-layer neural networks". In: *Proceedings of the National Academy of Sciences* 115.33 (2018), E7665–E7671.

[Mél96] S. Méléard. "Asymptotic behaviour of some interacting particle systems; McKean–Vlasov and Boltzmann models". In: *Probabilistic Models for Nonlinear Partial Differential Equations: Lectures given at the 1st Session of the Centro Internazionale Matematico Estivo (C.I.M.E.) held in Montecatini Terme, Italy, May 22–30, 1995*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 42–95.

[Mon17] P. Monmarché. "Long-time behaviour and propagation of chaos for mean field kinetic particles". In: *Stochastic Processes and their Applications* 127.6 (2017), pp. 1721–1737.

[NWS22] A. Nitanda, D. Wu, and T. Suzuki. "Convex analysis of the mean field Langevin dynamics". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR. 2022, pp. 9741–9757.

[Ott01] F. Otto. "The geometry of dissipative evolution equations: the porous medium equation". In: *Comm. Partial Differential Equations* 26.1-2 (2001), pp. 101–174.

[OV00] F. Otto and C. Villani. "Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality". In: *Journal of Functional Analysis* 173.2 (2000), pp. 361–400.

[RES22] G. dos Reis, S. Engelhardt, and G. Smith. "Simulation of McKean–Vlasov SDEs with super-linear growth". In: *IMA Journal of Numerical Analysis* 42.1 (2022), pp. 874–922.

[RV22] G. M. Rotskoff and E. Vanden-Eijnden. "Trainability and accuracy of artificial neural networks: an interacting particle system approach". In: *Comm. Pure Appl. Math.* 75.9 (2022), pp. 1889–1935.

[SL19] R. Shen and Y. T. Lee. "The randomized midpoint method for log-concave sampling". In: *Advances in Neural Information Processing Systems (NeurIPS)* 32 (2019).

[SS20] J. Sirignano and K. Spiliopoulos. "Mean field analysis of neural networks: A law of large numbers". In: *SIAM Journal on Applied Mathematics* 80.2 (2020), pp. 725–752.

[SNW22] T. Suzuki, A. Nitanda, and D. Wu. "Uniform-in-time propagation of chaos for the mean-field gradient Langevin dynamics". In: *The Eleventh International Conference on Learning Representations (ICLR)*. 2022.

[SWN23] T. Suzuki, D. Wu, and A. Nitanda. "Convergence of mean-field Langevin dynamics: time and space discretization, stochastic gradient, and variance reduction". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

[Szn91] A.-S. Sznitman. "Topics in propagation of chaos". In: *École d'Été de Probabilités de Saint-Flour XIX—1989*. Vol. 1464. Lecture Notes in Math. Springer, Berlin, 1991, pp. 165–251.

[Tal96] D. Talay. "Probabilistic numerical methods for partial differential equations: elements of analysis". In: *Probabilistic models for nonlinear partial differential equations (Montecatini Terme, 1995)*. Vol. 1627. Lecture Notes in Math. Springer, Berlin, 1996, pp. 148–196.

[VW19] S. Vempala and A. Wibisono. "Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices". In: *Advances in Neural Information Processing Systems (NeurIPS)* 32 (2019).

[Vil02] C. Villani. "A review of mathematical topics in collisional kinetic theory". In: *Handbook of mathematical fluid dynamics, Vol. I*. North-Holland, Amsterdam, 2002, pp. 71–305.

[Vil03] C. Villani. *Topics in optimal transportation*. Vol. 58. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2003, pp. xvi+370.

[Wib18] A. Wibisono. "Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem". In: *Proceedings of the 31st Conference on Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2093–3027.

[WSC22] K. Wu, S. Schmidler, and Y. Chen. "Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling". In: *The Journal of Machine Learning Research (JMLR)* 23.1 (2022), pp. 12348–12410.

[YY23] R. Yao and Y. Yang. “Mean-field variational inference via Wasserstein gradient flow”. In: *arXiv preprint 2207.08074* (2023).

[YKW22] M.-C. Yue, D. Kuhn, and W. Wiesemann. “On linear optimization over Wasserstein balls”. In: *Math. Program.* 195.1-2 (2022), pp. 1107–1122.

[Zha+23] M. S. Zhang, S. Chewi, M. Li, K. Balasubramanian, and M. A. Erdogdu. “Improved discretization analysis for underdamped Langevin Monte Carlo”. In: *Conference on Learning Theory (COLT)*. PMLR. 2023, pp. 36–71.

A Control of the Finite-Particle Error

In this section, we prove the results in §3.1 on the finite-particle error. We will make extensive use of the following transport inequality, which arises as a consequence of (LSI).

Lemma 14 (Talagrand's Transport Inequality, [OV00])**.** *If a measure π satisfies* (LSI) *with constant C_{LSI}, then for all measures $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$,*

$$\mathcal{W}_2^2(\mu,\pi) \le 2C_{\mathsf{LSI}}\,\mathsf{KL}(\mu \parallel \pi)\,. \tag{TI}$$

A.1 LSI Case

We provide the proof of Theorem 3 under the assumption of (LSI) for the invariant measures of (pMV) and (pMV_N). This relies on a BBGKY hierarchy based on the arguments of [LL23].

Recall that $\mu^{1:k}$ is the k-particle distribution of the finite-particle system. Explicitly,

$$\log \mu^{1:k}(x^{1:k}) = \log \int \exp\Big(-\frac{2}{\sigma^2}\sum_{i=1}^N V(x^i) - \frac{1}{\sigma^2\,(N-1)}\sum_{\substack{i,j=1\\ i\neq j}}^N W(x^i-x^j)\Big)\,\mathrm{d}x^{k+1:N} + \text{const.}$$

Using exchangeability, we can then compute the gradient of the potential for this measure as

$$-\frac{\sigma^2}{2}\,\nabla_{x^i}\log\mu^{1:k}(x^{1:k}) = \nabla V(x^i) + \frac{1}{N-1}\sum_{\substack{j=1\\ i\neq j}}^k \nabla W(x^i-x^j) + \frac{N-k}{N-1}\,\mathbb{E}_{\mu^{k+1|1:k}(\cdot|x^{1:k})}\,\nabla W(x^i-\cdot)\,.$$

Let $X^{1:k}\sim\mu^{1:k}$ and introduce the notation

$$\mathsf{K}_k \coloneqq \mathsf{KL}(\mu^{1:k} \parallel \pi^{\otimes k})\,.$$

Invoking (LSI) of the mean-field invariant measure (and tensorizing) leads to

$$\begin{aligned}
\mathsf{K}_k &\le \frac{C_{\mathsf{LSI}}(\pi)}{2}\,\mathsf{FI}(\mu^{1:k} \parallel \pi^{\otimes k})\\
&= \frac{2C_{\mathsf{LSI}}(\pi)}{\sigma^4}\sum_{i=1}^k \mathbb{E}\Big[\Big\|\frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^k \nabla W(X^i-X^j) - \int \nabla W(X^i-\cdot)\,\mathrm{d}\pi\\
&\qquad\qquad + \frac{N-k}{N-1}\int \nabla W(X^i-\cdot)\,\mathrm{d}\mu^{k+1|1:k}(\cdot\mid X^{1:k})\Big\|^2\Big]\\
&\le \frac{4k\,C_{\mathsf{LSI}}(\pi)}{\sigma^4\,(N-1)^2}\underbrace{\mathbb{E}\Big[\Big\|\sum_{j=2}^k\Big(\nabla W(X^1-X^j) - \int \nabla W(X^1-\cdot)\,\mathrm{d}\pi\Big)\Big\|^2\Big]}_{\mathsf{A}}\\
&\qquad + \frac{4k\,C_{\mathsf{LSI}}(\pi)\,(N-k)^2}{\sigma^4\,(N-1)^2}\underbrace{\mathbb{E}\Big[\Big\|\int \nabla W(X^1-\cdot)\,\big(\mathrm{d}\mu^{k+1|1:k}(\cdot\mid X^{1:k}) - \mathrm{d}\pi\big)\Big\|^2\Big]}_{\mathsf{B}}\,,
\end{aligned}$$

where the last line follows from exchangeability and $\|a+b\|^2 \le 2\,(\|a\|^2+\|b\|^2)$ for vectors $a,b\in\mathbb{R}^d$.

A.1.1 Bounding the Error Terms

We now handle terms A,B separately.

$$\mathsf{A} = \sum_{j=2}^k \mathbb{E}[\|\nabla W(X^1-X^j) - \mathbb{E}_\pi\,\nabla W(X^1-\cdot)\|^2]$$

$$
\begin{aligned}
&+ \sum_{\substack{i,j=2 \\ i\neq j}}^{k} \mathbb{E}\big\langle \nabla W(X^1 - X^i) - \mathbb{E}_\pi \nabla W(X^1 - \cdot),\, \nabla W(X^1 - X^j) - \mathbb{E}_\pi \nabla W(X^1 - \cdot)\big\rangle \\
&\underset{\text{(i)}}{=} (k-1)\,\mathbb{E}[\|\nabla W(X^1 - X^2) - \mathbb{E}_\pi \nabla W(X^1 - \cdot)\|^2] + \\
&\quad + (k-1)\,(k-2)\,\mathbb{E}\big\langle \nabla W(X^1 - X^2) - \mathbb{E}_\pi \nabla W(X^1 - \cdot), W(X^1 - X^3) - \mathbb{E}_\pi \nabla W(X^1 - \cdot)\big\rangle \\
&\underset{\text{(ii)}}{\leq} (k-1)\,\beta_W^2\,\mathbb{E}[\|X - Y\|^2] \\
&\quad + (k-1)^2\,\mathbb{E}\big\langle \nabla W(X^1 - X^2) - \mathbb{E}_\pi \nabla W(X^1 - \cdot),\, \nabla W(X^1 - X^3) - \mathbb{E}_\pi \nabla W(X^1 - \cdot)\big\rangle\,,
\end{aligned}
$$

where we used the exchangeability of the particles in (i) and the smoothness of W in (ii). Here, $X \sim \mu^1$ and $Y \sim \pi$ are independent.

Let us deal with these two terms separately. For the first term, let $\bar{Y} \sim \pi$ be *optimally* coupled with X. Then, by independence and sub-Gaussian concentration (implied by (LSI)),

$$
\begin{aligned}
\mathbb{E}[\|X - Y\|^2] &\leq 2\,\mathbb{E}[\|X - \bar{Y}\|^2] + 2\,\mathbb{E}[\|Y - \bar{Y}\|^2] = 2\,\mathcal{W}_2^2(\mu^1, \pi) + 4\,\mathbb{E}[\|Y - \mathbb{E}\,Y\|^2] \\
&\leq 4C_{\mathsf{LSI}}(\pi)\,\mathsf{KL}(\mu^1 \parallel \pi) + 4dC_{\mathsf{LSI}}(\pi) \leq 4C_{\mathsf{LSI}}(\pi)\,(\mathsf{K}_3 + d)\,,
\end{aligned} \tag{A.1}
$$

where the second inequality follows from (TI), and the last one follows from the data-processing inequality for the KL divergence.

For the second term, the Cauchy–Schwarz inequality leads to

$$
\begin{aligned}
&\mathbb{E}\big\langle \nabla W(X^1 - X^2) - \mathbb{E}_\pi \nabla W(X^1 - \cdot),\, \nabla W(X^1 - X^3) - \mathbb{E}_\pi \nabla W(X^1 - \cdot)\big\rangle \\
&\quad = \mathbb{E}\big\langle \nabla W(X^1 - X^2) - \mathbb{E}_\pi \nabla W(X^1 - \cdot),\, \mathbb{E}_{\mu^{3|1:2}(\cdot|X^{1:2})} \nabla W(X^1 - \cdot) - \mathbb{E}_\pi \nabla W(X^1 - \cdot)\big\rangle \\
&\quad \leq \beta_W^2 \sqrt{\mathbb{E}[\|X - Y\|^2]} \sqrt{\mathbb{E}\,\mathcal{W}_2^2\big(\mu^{3|1:2}(\cdot \mid X^{1:2}),\, \pi\big)} \\
&\quad \underset{\text{(i)}}{\leq} \beta_W^2 \sqrt{4C_{\mathsf{LSI}}(\pi)\,(\mathsf{K}_3 + d)} \sqrt{2C_{\mathsf{LSI}}(\pi)\,\mathbb{E}\,\mathsf{KL}\big(\mu^{3|1:2}(\cdot \mid X^{1:2}) \,\big\|\, \pi\big)} \\
&\quad \underset{\text{(ii)}}{\leq} 3\beta_W^2 C_{\mathsf{LSI}}(\pi) \sqrt{\mathsf{K}_3 + d} \sqrt{\mathsf{K}_3} \\
&\quad \leq 3\beta_W^2 C_{\mathsf{LSI}}(\pi)\,(\mathsf{K}_3 + d)\,,
\end{aligned} \tag{A.2}
$$

where in (i) we applied the bound (A.1) as well as (TI), and in (ii) we used the chain rule for the KL divergence.

We return to the analysis of the term B. In a similar way, we obtain

$$
\begin{aligned}
\mathsf{B} &= \mathbb{E}\Big[\Big\| \int \nabla W(X^1 - \cdot)\,\big(\mathrm{d}\mu^{k+1|1:k}(\cdot \mid X^{1:k}) - \mathrm{d}\pi\big)\Big\|^2\Big] \leq \beta_W^2\,\mathbb{E}\,\mathcal{W}_2^2\big(\mu^{k+1|1:k}(\cdot \mid X^{1:k}),\, \pi\big) \\
&\leq 2\beta_W^2 C_{\mathsf{LSI}}(\pi)\,(\mathsf{K}_{k+1} - \mathsf{K}_k)\,.
\end{aligned}
$$

A.1.2 Induction

Putting our bounds on A and B together, we obtain for $N \geq 30$,

$$
\mathsf{K}_k \leq \frac{30k^3\beta_W^2 C_{\mathsf{LSI}}^2(\pi)}{\sigma^4 N^2}\,(\mathsf{K}_3 + d) + \frac{8k\beta_W^2 C_{\mathsf{LSI}}^2(\pi)}{\sigma^4}\,(\mathsf{K}_{k+1} - \mathsf{K}_k)\,. \tag{A.3}
$$

In particular, the case of $k = N$ involves our bounds only on A, leading to

$$
\mathsf{K}_N \leq \frac{30N\beta_W^2 C_{\mathsf{LSI}}^2(\pi)}{\sigma^4}\,(\mathsf{K}_3 + d)\,.
$$

By grouping together the K_k terms in (A.3),

$$
\mathsf{K}_k \leq \underbrace{\frac{8k\beta_W^2 C_{\mathsf{LSI}}^2(\pi)/\sigma^4}{1 + 8k\beta_W^2 C_{\mathsf{LSI}}^2(\pi)/\sigma^4}}_{=:\mathcal{C}_k} \Big(\mathsf{K}_{k+1} + \big(\frac{2k}{N}\big)^2\,(\mathsf{K}_3 + d)\Big)\,. \tag{A.4}
$$

Iterating this inequality down to $k = 3$, for $\rho \coloneqq \sigma^4/8\beta_W^2 C_{\mathsf{LSI}}^2(\pi)$,

$$\begin{aligned}\mathsf{K}_3 &\le \Big(\prod_{k=3}^{N-1} \mathcal{C}_k\Big) \frac{30N\beta_W^2 C_{\mathsf{LSI}}^2(\pi)}{\sigma^4} (\mathsf{K}_3 + d) + \sum_{k=3}^{N-1} \Big(\prod_{\ell=3}^{k} \mathcal{C}_\ell\Big) \big(\frac{2k}{N}\big)^2 (\mathsf{K}_3 + d) \\ &\le \underbrace{\Big[\Big(\prod_{k=3}^{N-1} \mathcal{C}_k\Big) \frac{4N}{\rho} + \sum_{k=3}^{N-1} \Big(\prod_{\ell=3}^{k} \mathcal{C}_\ell\Big) \big(\frac{2k}{N}\big)^2\Big]}_{=:c_N} (\mathsf{K}_3 + d)\,.\end{aligned}$$

Now we show $c_N < 1/2$, which implies $\mathsf{K}_3 \le 2c_N d$. We require the following lemma.

Lemma 15. *For $3 \le i \le k \le N$,*

$$\prod_{\ell=i}^{k} \mathcal{C}_\ell \le \Big(\frac{i+\rho}{k+1+\rho}\Big)^\rho\,.$$

Proof. For $\mathcal{C}_\ell = \frac{\ell\rho^{-1}}{1+\ell\rho^{-1}}$, we have

$$C \coloneqq \log \prod_{\ell=i}^{k} \mathcal{C}_\ell = \sum_{\ell=i}^{k} \log\Big(1 - \frac{1}{1+\ell\rho^{-1}}\Big) \le -\sum_{\ell=i}^{k} \frac{1}{1+\ell\rho^{-1}}\,.$$

As the summand is decreasing in ℓ, it follows that

$$C \le -\sum_{\ell=i}^{k} \int_{\ell}^{\ell+1} \frac{1}{1+x\rho^{-1}}\,\mathrm{d}x = -\int_{i}^{k+1} \frac{1}{1+x\rho^{-1}}\,\mathrm{d}x = -\rho \log \frac{k+1+\rho}{i+\rho}\,.$$

Therefore,

$$\prod_{\ell=i}^{k} \mathcal{C}_\ell = \exp C \le \Big(\frac{k+1+\rho}{i+\rho}\Big)^{-\rho}\,,$$

which proves the lemma. □

Using Lemma 15, we obtain

$$c_N \le 4\,(3+\rho)^\rho \Big(\frac{N^{1-\rho}}{\rho} + \frac{1}{N^2}\sum_{k=3}^{N-1} k^{2-\rho}\Big)\,.$$

Under Assumption 3, i.e., $\rho \ge 3$, we may assume $\rho = 3$ since we can always take a worse bound on the constants β_W so that $\rho = 3$. As seen shortly, the rate does not improve even if $\rho > 3$.[5] For $\rho = 3$ and $N \ge 100$, we therefore obtain

$$c_N \le 864 \Big(\frac{1}{3N^2} + \frac{1}{N^2}\sum_{k=3}^{N-1} \frac{1}{k}\Big) \le \frac{1}{2}\,,$$

and thus

$$\mathsf{K}_3 \lesssim \frac{d \log N}{N^2}\,. \tag{A.5}$$

A.1.3 Bootstrapping

Substituting the bound (A.5) for K_3 into the recursive inequality (A.4), we end up with a suboptimal rate of $\widetilde{\mathcal{O}}(k^3/N^2)$ for K_k. To improve the bound, we substitute our established bound (A.5) into (A.2), which results in an improved recursive inequality. Indeed,

$$\mathsf{A} \lesssim k\beta_W^2 C_{\mathsf{LSI}}^2(\pi)\,(\mathsf{KL}_3 + d) + k^2\beta_W^2 C_{\mathsf{LSI}}(\pi)\sqrt{\mathsf{KL}_3 + d}\sqrt{\mathsf{KL}_3} \lesssim dk\beta_W^2 C_{\mathsf{LSI}}(\pi)\sqrt{\log N}$$

[5] Alternatively, one can show the bound in Lemma 15 decreases in ρ, so we can just substitute $\rho = 3$ therein.

and therefore

$$\mathsf{K}_k \le \widetilde{\mathcal{O}}\Big(\frac{dk^2\beta_W^2 C_{\mathsf{LSI}}^2(\pi)}{\sigma^4 N^2}\Big) + \frac{8k\beta_W^2 C_{\mathsf{LSI}}^2(\pi)}{\sigma^4}\,(\mathsf{K}_{k+1} - \mathsf{K}_k)\,.$$

For $k = N$ this yields

$$\mathsf{K}_N \le \widetilde{\mathcal{O}}\Big(\frac{d\beta_W^2 C_{\mathsf{LSI}}^2(\pi)}{\sigma^4}\Big)\,.$$

Regrouping K_k as before, we obtain

$$\mathsf{K}_k \le \mathcal{C}_k\left(\mathsf{K}_{k+1} + \widetilde{\mathcal{O}}(\frac{dk}{N^2})\right).$$

Iterating this down to $k = N$,

$$\begin{aligned}\mathsf{K}_k &\le \Big(\prod_{\ell=k}^{N-1}\mathcal{C}_\ell\Big)\,\mathsf{K}_N + \sum_{\ell=k}^{N-1}\Big(\prod_{j=k}^{\ell}\mathcal{C}_j\Big)\,\widetilde{\mathcal{O}}(\frac{d\ell}{N^2}) \\ &\underset{\text{(i)}}{\le} \widetilde{\mathcal{O}}\Big(\frac{k^3}{N^3}\,\frac{d\beta_W^2 C_{\mathsf{LSI}}^2(\pi)}{\sigma^4} + \sum_{\ell=k}^{N-1}\frac{k^3}{\ell^3}\,\frac{d\ell}{N^2}\Big) \underset{\text{(ii)}}{\le} \widetilde{\mathcal{O}}(\frac{dk^2}{N^2})\,,\end{aligned}$$

where in (i) we used Lemma 15 with $\rho = 3$, and (ii) follows from $\rho \ge 3$ and $\sum_{\ell\ge k}\ell^{-2} \le k^{-1}$. Therefore, for some fixed k it suffices to take $N = 100 \vee \widetilde{\Omega}(k\sqrt{d}/\varepsilon)$ to achieve ε^2-bias in KL, completing the proof of Theorem 3.

A.2 Strongly Convex Case

The following propagation of chaos argument for the strongly log-concave case is based on [Szn91]. Let $(X_t^{1:N})_{t\ge0}$ denote the stochastic process following the finite-particle stochastic differential equation (pMV_N). Let the corresponding semigroup be denoted $(\mathcal{T}_t)_{t\ge0}$, defined as follows. For any test function $f : \mathbb{R}^{d\times N} \to \mathbb{R}$,

$$\mathcal{T}_t f(x^{1:N}) = \mathbb{E}[f(X_t^{1:N}) \mid X_0^{1:N} = x^{1:N}]\,.$$

Then, the following simple lemma proves Wasserstein contraction for the finite-particle system.

Lemma 16. *Under Assumption 4 and for $N \ge \frac{\alpha_V - \alpha_W^-}{\alpha_V + (\alpha_W)_-}$, $(\mathcal{T}_t)_{t\ge0}$ is a contraction in the 2-Wasserstein distance with exponential rate at least $\alpha/2$, where $\alpha \coloneqq \alpha_V + \alpha_W^-$. In other words, for any measures $\mu_0^{1:N}$, $\nu_0^{1:N}$ in $\mathcal{P}_2(\mathbb{R}^{d\times N})$,*

$$\mathcal{W}_2(\mu_0^{1:N}\mathcal{T}_t,\, \nu_0^{1:N}\mathcal{T}_t) \le \exp(-\alpha t/2)\,\mathcal{W}_2(\mu_0^{1:N}, \nu_0^{1:N})\,.$$

Proof. Note that $(\mathcal{T}_t)_{t\ge0}$ corresponds to the time-scaled (by factor $\sigma^2/2$) Langevin diffusion with stationary distribution $\mu^{1:N}$, which is $\frac{2}{\sigma^2}\,(\alpha_V + \frac{N}{N-1}\,\alpha_W^-)$-strongly log-concave by Lemma 7. The condition on N ensures that this is at least α/σ^2. Consequently, it is well-known (e.g., via synchronous coupling) that the diffusion is a contraction in the Wasserstein distance with rate at least $\alpha/2$. □

We next bound the error incurred in one step from applying the finite-particle semigroup to $\pi^{\otimes N}$.

Lemma 17. *Under Assumptions 1 and 4, for any $\lambda > 0$, $\mathcal{T}_h$ induces the following error in Wasserstein distance:*

$$\mathcal{W}_2^2(\pi^{\otimes N}\mathcal{T}_h,\, \pi^{\otimes N}) \le \frac{(1+\lambda^{-1})\,\beta_W^2\sigma^2 dh^2}{\alpha}\exp\Big(\frac{(1+\lambda)\,\beta_W^2 h^2}{2}\Big)\,.$$

Proof. We resort to a coupling argument, noting that π is stationary under (pMV). Starting with $\pi^{\otimes N}$, we evolve $(X_t^{1:N})_{t\geq 0}$ and $(Y_t^{1:N})_{t\geq 0}$ according to (pMV_N) and (pMV) respectively, i.e., $X_t^{1:N} \sim \pi^{\otimes N}\mathcal{T}_t$ and $Y_t^{1:N} \sim \pi^{\otimes N}$. This argument is adapted from the original propagation of chaos proof by [Szn91].

We can compute the evolution under a synchronous coupling as:

$$\begin{aligned}
\mathrm{d}(X_t^i - Y_t^i) &= -\big(\nabla V(X_t^i) - \nabla V(Y_t^i)\big)\,\mathrm{d}t - \frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\big(\nabla W(X_t^i - X_t^j) - \mathbb{E}_\pi\,\nabla W(Y_t^i - \cdot)\big)\,\mathrm{d}t \\
&= -\big(\nabla V(X_t^i) - \nabla V(Y_t^i)\big)\,\mathrm{d}t - \frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\big(\nabla W(X_t^i - X_t^j) - \nabla W(Y_t^i - X_t^j)\big)\,\mathrm{d}t \\
&\quad - \frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\big(\nabla W(Y_t^i - X_t^j) - \nabla W(Y_t^i - Y_t^j)\big)\,\mathrm{d}t \\
&\quad - \frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\big(\nabla W(Y_t^i - Y_t^j) - \mathbb{E}_\pi\,\nabla W(Y_t^i - \cdot)\big)\,\mathrm{d}t\,.
\end{aligned}$$

Now let us denote by $\overline{\nabla W}(x,y) \coloneqq \nabla W(x-y) - \mathbb{E}_\pi\,\nabla W(x-\cdot)$ the centered gradient (with respect to π). By Itô's formula and Assumption 4,

$$\begin{aligned}
\mathrm{d}\|X_t^i - Y_t^i\|^2 &= 2\,\langle X_t^i - Y_t^i, \mathrm{d}(X_t^i - Y_t^i)\rangle \\
&\leq -2\,(\alpha_V + \alpha_W)\,\|X_t^i - Y_t^i\|^2\,\mathrm{d}t \\
&\quad - \frac{2}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\langle X_t^i - Y_t^i, \nabla W(Y_t^i - X_t^j) - \nabla W(Y_t^i - Y_t^j)\rangle\,\mathrm{d}t \\
&\quad - \frac{2}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\langle X_t^i - Y_t^i, \nabla W(Y_t^i - Y_t^j) - \mathbb{E}_\pi\,\nabla W(Y_t^i - \cdot)\rangle\,\mathrm{d}t \\
&\leq \frac{2\beta_W\,\|X_t^i - Y_t^i\|}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\|X_t^j - Y_t^j\|\,\mathrm{d}t + \frac{2\,\|X_t^i - Y_t^i\|}{N-1}\Big\|\sum_{\substack{j=1\\ j\neq i}}^{N}\overline{\nabla W}(Y_t^i, Y_t^j)\Big\|\,\mathrm{d}t
\end{aligned}$$

or

$$\mathrm{d}\|X_t^i - Y_t^i\| \leq \frac{\beta_W}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\|X_t^j - Y_t^j\|\,\mathrm{d}t + \frac{1}{N-1}\Big\|\sum_{\substack{j=1\\ j\neq i}}^{N}\overline{\nabla W}(Y_t^i, Y_t^j)\Big\|\,\mathrm{d}t\,.$$

Integrating and squaring,

$$\begin{aligned}
\|X_t^i - Y_t^i\|^2 &\leq \Big|\int_0^t\Big(\frac{\beta_W}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\|X_s^j - Y_s^j\| + \frac{1}{N-1}\Big\|\sum_{\substack{j=1\\ j\neq i}}^{N}\overline{\nabla W}(Y_s^i, Y_s^j)\Big\|\Big)\,\mathrm{d}s\Big|^2 \\
&\leq \frac{(1+\lambda)\,\beta_W^2 t}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\int_0^t\|X_s^j - Y_s^j\|^2\,\mathrm{d}s + \frac{(1+\lambda^{-1})\,t}{(N-1)^2}\int_0^t\Big\|\sum_{\substack{j=1\\ j\neq i}}^{N}\overline{\nabla W}(Y_s^i, Y_s^j)\Big\|^2\,\mathrm{d}s\,,
\end{aligned}$$

where the last line follows from Young's inequality.

Next, we take expectations. Note that $\overline{\nabla W}(\cdot,\cdot)$ is centered in its second variable, so for any $j \neq k$,

$$\mathbb{E}\langle\overline{\nabla W}(Y_t^i, Y_t^j), \overline{\nabla W}(Y_t^i, Y_t^k)\rangle = 0\,.$$

Otherwise, we can bound the terms via

$$\mathbb{E}[\|\overline{\nabla W}(Y_t^i, Y_t^j)\|^2] \le \beta_W^2 \, \mathbb{E}_{\substack{Y_t^j \sim \pi \\ Z \sim \pi}}[\|Y_t^j - Z\|^2] \le \frac{\beta_W^2 \sigma^2 d}{\alpha} \,.$$

Here, Z is an independent draw from π and so cannot be reduced via coupling. The second inequality follows from a standard bound on the centered second moment of a strongly log-concave measure, using the fact that π is $2\alpha/\sigma^2$-strongly log-concave [c.f. DKR22].

Therefore, taking expectations and summing over the particles,

$$\mathbb{E}[\|X_t^{1:N} - Y_t^{1:N}\|^2] \le (1+\lambda)\, \beta_W^2 t \int_0^t \|X_s^{1:N} - Y_s^{1:N}\|^2 \, \mathrm{d}s + \frac{(1+\lambda^{-1})\, \beta_W^2 \sigma^2 d t^2}{\alpha} \,.$$

By Grönwall's inequality below,

$$\mathbb{E}[\|X_h^{1:N} - Y_h^{1:N}\|^2] \le \frac{(1+\lambda^{-1})\, \beta_W^2 \sigma^2 d h^2}{\alpha} \exp\Big(\frac{(1+\lambda)\, \beta_W^2 h^2}{2}\Big) \,.$$

This concludes the proof. □

Lemma 18 (Grönwall's Inequality)**.** *For $T > 0$, let $f : [0,T] \to \mathbb{R}_{\ge 0}$ be bounded. Suppose that the following holds pointwise for some functions $a, b : [0,T] \to \mathbb{R}$, where a is increasing:*

$$f(t) \le a(t) + \int_0^t b(s) f(s) \, \mathrm{d}s \,.$$

Then,

$$f(t) \le a(t) \exp\Big(\int_0^t b(s) \, \mathrm{d}s\Big) \,.$$

Composing Lemmas 16 and 17, we now prove our propagation of chaos results.

Proof of Theorem 4 Indeed, we have

$$\begin{aligned} \mathcal{W}_2(\mu^{1:N}, \pi^{\otimes N}) &= \mathcal{W}_2(\mu^{1:N}\mathcal{T}_h, \, \pi^{\otimes N}) \le \mathcal{W}_2(\mu^{1:N}\mathcal{T}_h, \, \pi^{\otimes N}\mathcal{T}_h) + \mathcal{W}_2(\pi^{\otimes N}\mathcal{T}_h, \, \pi^{\otimes N}) \\ &\le \exp(-\alpha h/2)\, \mathcal{W}_2(\mu^{1:N}, \pi^{\otimes N}) + \sqrt{\frac{(1+\lambda^{-1})\, \beta_W^2 \sigma^2 d h^2}{\alpha}} \exp\Big(\frac{(1+\lambda)\, \beta_W^2 h^2}{4}\Big) \,. \end{aligned}$$

Rearranging,

$$\mathcal{W}_2(\mu^{1:N}, \pi^{\otimes N}) \le \frac{1}{1 - \exp(-\alpha h/2)} \sqrt{\frac{(1+\lambda^{-1})\, \beta_W^2 \sigma^2 d h^2}{\alpha}} \exp\Big(\frac{(1+\lambda)\, \beta_W^2 h^2}{4}\Big) \,.$$

Let $h \searrow 0$ first and then $\lambda \nearrow \infty$ to obtain

$$\mathcal{W}_2^2(\mu^{1:N}, \pi^{\otimes N}) \le \frac{4 \beta_W^2 \sigma^2 d}{\alpha^3} \,.$$

Finally, when $k < N$, we use exchangeability (see Lemma 23 below) to conclude the proof of (3.2).

For (3.3), by the Bakry–Émery condition we have $C_{\mathsf{LSI}}(\pi) \le \sigma^2/2\alpha$, and tensorization [c.f. BGL14, Proposition 5.2.7] leads to $C_{\mathsf{LSI}}(\pi^{\otimes N}) \le \sigma^2/2\alpha$. Thus, (TI) leads to

$$\mathsf{KL}(\mu^{1:N} \parallel \pi^{\otimes N}) \le \frac{\sigma^2}{4\alpha} \, \mathsf{FI}(\mu^{1:N} \parallel \pi^{\otimes N}) \,.$$

However, one notes that the density of $\mu^{1:N}$ is log-smooth with parameter $\frac{2}{\sigma^2}\,(\beta_V + \frac{N}{N-1}\,\beta_W)$ (Lemma 7). Likewise, $\pi^{\otimes N}$ is log-smooth with parameter $\frac{2}{\sigma^2}\,(\beta_V + \beta_W)$. Now consider a functional $\mathcal{F}$ on the space

of probability measures on $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^{d\times N})$ given by $\mathcal{F}:\nu\mapsto\mathbb{E}_\nu[\|\nabla\log\frac{\mu^{1:N}}{\pi^{\otimes N}}\|^2]$. Note that $\log(\mu^{1:N}/\pi^{\otimes N})$ is smooth with parameter at most $\frac{4}{\sigma^2}\left(\beta_V+\frac{N}{N-1}\beta_W\right)\le\frac{8}{\sigma^2}\left(\beta_V+\beta_W\right)$, for $N\ge 2$.

Next, note that for $Y^{1:N}\sim\pi^{\otimes N}$,

$$\begin{aligned}\mathcal{F}(\pi^{\otimes N})&=\mathbb{E}_{\pi^{\otimes N}}[\|\nabla\log\mu^{1:N}-\nabla\log\pi^{\otimes N}\|^2]\\&=\frac{4N}{\sigma^4\left(N-1\right)^2}\,\mathbb{E}\Big[\Big\|\sum_{j=2}^N\Big(\nabla W(Y^1-Y^j)-\int\nabla W(Y^1-\cdot)\,\mathrm{d}\pi\Big)\Big\|^2\Big]\\&=\frac{4N}{\sigma^4\left(N-1\right)^2}\,\mathbb{E}\Big[\Big\|\sum_{j=2}^N\overline{\nabla W}(Y^1,Y^j)\Big\|^2\Big]\,,\end{aligned}$$

by using exchangeability and the definition of $\overline{\nabla W}$.

Subsequently, one derives the following inequality using the Wasserstein distance bound:

$$\begin{aligned}\mathcal{F}(\mu^{1:N})&\le\frac{128}{\sigma^4}\left(\beta_V+\beta_W\right)^2\mathcal{W}_2^2(\mu^{1:N},\pi^{\otimes N})+2\mathcal{F}(\pi^{\otimes N})\\&\le\frac{512\beta_W^2 d}{\alpha^3\sigma^2}\left(\beta_V+\beta_W\right)^2+\frac{8N}{\sigma^4\left(N-1\right)^2}\,\mathbb{E}\Big[\Big\|\sum_{j=2}^N\overline{\nabla W}(Y^1,Y^j)\Big\|^2\Big]\\&\le\frac{512\beta_W^2 d}{\alpha^3\sigma^2}\left(\beta_V+\beta_W\right)^2+\frac{16\beta_W^2 N}{\sigma^4\left(N-1\right)}\,\mathbb{E}[\|Y^1-\mathbb{E}\,Y^1\|^2]\\&\le\frac{512\beta_W^2 d}{\alpha^3\sigma^2}\left(\beta_V+\beta_W\right)^2+\frac{16\beta_W^2 d}{\alpha\sigma^2}\,,\end{aligned}$$

by using (3.2) and the fact that $\overline{\nabla W}(\cdot,\cdot)$ is a centered random variable in its second argument. This concludes the proof for $k=N$, and as in the $\mathcal{W}_2^2$ bound, Lemma 24 will conclude the proof for $k<N$. □

A.3 General Functional Case

For any measure μ, define its entropy as $\mathsf{ent}(\mu)=\int\log\mu\,\mathrm{d}\mu$. We now provide a self-contained propagation of chaos argument in the general McKean–Vlasov setting, following [CRW22]. We begin with the following entropy toast inequality, i.e., half of the entropy sandwich inequality from [CRW22].

Lemma 19 (Entropy Toast Inequality)**.** *Define the empirical total energy for an N-finite particle system as follows. Given a measure $\nu^{1:N}\in\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^{d\times N})$,*

$$\mathcal{E}^N(\nu^{1:N})=N\int\mathcal{F}(\rho_{x^{1:N}})\,\nu^{1:N}(\mathrm{d}x^{1:N})+\frac{\sigma^2}{2}\,\mathsf{ent}(\mu)\,.$$

Under Assumptions 5, it holds for all measures $\nu^{1:N}\in\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^{d\times N})$

$$\frac{\sigma^2}{2}\,\mathsf{KL}(\nu^{1:N}\,\|\,\pi^{\otimes N})\le\mathcal{E}^N(\nu^{1:N})-N\mathcal{E}(\pi)\,,$$

where $\mathcal{E}$ is the total energy (gE) and π is the stationary measure (2.2).

Proof. By Assumption 5, we have

$$\begin{aligned}\mathcal{E}^N(\nu^{1:N})-N\mathcal{E}(\pi)&=N\,\mathbb{E}_{x^{1:N}\sim\nu^{1:N}}[\mathcal{F}(\rho_{x^{1:N}})-\mathcal{F}(\pi)]+\frac{\sigma^2}{2}\left(\mathsf{ent}(\nu^{1:N})-N\,\mathsf{ent}(\pi)\right)\\&\ge\mathbb{E}_{x^{1:N}\sim\nu^{1:N}}\Big[N\int\delta\mathcal{F}(\pi,z)\left(\rho_{x^{1:N}}(\mathrm{d}z)-\pi(\mathrm{d}z)\right)\Big]+\frac{\sigma^2}{2}\left(\mathsf{ent}(\nu^{1:N})-N\,\mathsf{ent}(\pi)\right)\\&=-\frac{\sigma^2}{2}\,\mathbb{E}_{x^{1:N}\sim\nu^{1:N}}\Big[N\int\log\pi(z)\left(\rho_{x^{1:N}}(\mathrm{d}z)-\pi(\mathrm{d}z)\right)\Big]+\frac{\sigma^2}{2}\left(\mathsf{ent}(\nu^{1:N})-N\,\mathsf{ent}(\pi)\right)\end{aligned}$$

$$= -\frac{\sigma^2}{2}\,\mathbb{E}_{x^{1:N}\sim\nu^{1:N}}\Big[N\int \log\pi(z)\,\rho_{x^{1:N}}(\mathrm{d}z)\Big] + \frac{\sigma^2}{2}\,\mathsf{ent}(\nu^{1:N})$$
$$= -\frac{\sigma^2}{2}\int\sum_{i=1}^{N}\log\pi(x^i)\,\nu^{1:N}(\mathrm{d}x^{1:N}) + \frac{\sigma^2}{2}\,\mathsf{ent}(\nu^{1:N})\,.$$

However, this is just $\frac{\sigma^2}{2}\,\mathsf{KL}(\nu^{1:N}\,\|\,\pi^{\otimes N})$, so we are done. □

Proof of Theorem 5 We bound $\mathcal{E}^N(\mu^{1:N}) - N\mathcal{E}^N(\pi)$ via the following argument. First, define the finite-particle mean-field functional as $\mathcal{F}^N(\nu^{1:N}) = N\int\mathcal{F}(\rho_{x^{1:N}})\,\nu^{1:N}(\mathrm{d}x^{1:N})$. In the sequel, we also use the following notation for conditional measures: if $x^{-i} \coloneqq (x^{1:i-1}, x^{i+1:N}) \in \mathbb{R}^{d\times(N-1)}$,

$$\mu^{1:N}(x^{1:N}) = \mu^{i|-i}(x^i \mid x^{-i}) \times \mu^{-i}(x^{-i})\,.$$

We know that

$$\mathcal{E}^N(\mu^{1:N}) - N\mathcal{E}(\pi) = \mathcal{F}^N(\mu^{1:N}) - N\mathcal{F}(\pi) + \frac{\sigma^2}{2}\,\mathsf{ent}(\mu^{1:N}) - \frac{N\sigma^2}{2}\,\mathsf{ent}(\pi)\,.$$

Furthermore, by Assumption 5,

$$\mathcal{F}^N(\mu^{1:N}) - N\mathcal{F}(\pi) \le N\,\mathbb{E}_{x^{1:N}\sim\mu^{1:N}}\int\delta\mathcal{F}(\rho_{x^{1:N}}, z)\,(\rho_{x^{1:N}}(\mathrm{d}z) - \pi(\mathrm{d}z))\,.$$

Using the subadditivity of entropy, we can therefore write

$$\mathcal{E}^N(\mu^{1:N}) - N\mathcal{E}(\pi) \le \sum_{i=1}^{N}\mathbb{E}_{x^{1:N}\sim\mu^{1:N}}\Big[\delta\mathcal{F}(\rho_{x^{1:N}}, x^i) - \int\delta\mathcal{F}(\rho_{x^{1:N}},\cdot)\,\mathrm{d}\pi + \frac{\sigma^2}{2}\left(\mathsf{ent}(\mu^{i|-i}(\cdot\mid x^{-i})) - \mathsf{ent}(\pi)\right)\Big]\,.$$

To decouple the terms, we now replace each $\delta\mathcal{F}(\rho_{x^{1:N}},\cdot)$ term with $\delta\mathcal{F}(\rho_{x^{-i}},\cdot)$:

$$\begin{aligned}\mathcal{E}^N(\mu^{1:N}) - N\mathcal{E}(\pi) &\le \underbrace{\sum_{i=1}^{N}\mathbb{E}_{x^{1:N}\sim\mu^{1:N}}\Big[\delta\mathcal{F}(\rho_{x^{-i}}, x^i) - \int\delta\mathcal{F}(\rho_{x^{-i}},\cdot)\,\mathrm{d}\pi + \frac{\sigma^2}{2}\left(\mathsf{ent}(\mu^{i|-i}(\cdot\mid x^{-i})) - \mathsf{ent}(\pi)\right)\Big]}_{\mathsf{A}}\\ &\quad + \underbrace{\sum_{i=1}^{N}\mathbb{E}_{x^{1:N}\sim\mu^{1:N}}\Big[\delta\mathcal{F}(\rho_{x^{1:N}}, x^i) - \delta\mathcal{F}(\rho_{x^{-i}}, x^i) - \int\big(\delta\mathcal{F}(\rho_{x^{1:N}},\cdot) - \delta\mathcal{F}(\rho_{x^{-i}},\cdot)\big)\,\mathrm{d}\pi\Big]}_{\mathsf{B}}\,.\end{aligned}$$

We consider the two terms in turn, beginning with the first.

Note that by Fubini's theorem,

$$\mathbb{E}_{x^{1:N}\sim\mu^{1:N}}\,\delta\mathcal{F}(\rho_{x^{-i}}, x^i) = \mathbb{E}_{x^{-i}\sim\mu^{-i}}\int\delta\mathcal{F}(\rho_{x^{-i}},\cdot)\,\mathrm{d}\mu^{i|-i}(\cdot\mid x^{-i})\,.$$

In order to relate the first term A to a KL divergence, for each $x^{-i}\in\mathbb{R}^{d\times(N-1)}$ we introduce the probability measure $\tau_{x^{-i}}\in\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ via

$$\tau_{x^{-i}} \propto \exp\Big(-\frac{2}{\sigma^2}\,\delta\mathcal{F}(\rho_{x^{-i}},\cdot)\Big)\,.$$

We can compute

$$\mathsf{KL}\big(\mu^{i|-i}(\cdot\mid x^{-i})\,\big\|\,\tau_{x^{-i}}\big)$$

$$= \int \Big(\frac{2}{\sigma^2}\, \delta\mathcal{F}(\rho_{x^{-i}}, \cdot) + \log \mu^{i|-i}(\cdot \mid x^{-i}) \Big)\, \mathrm{d}\mu^{i|-i}(\cdot \mid x^{-i}) + \log Z(\tau_{x^{-i}})\,,$$

where $Z(\tau_{x^{-i}})$ is the normalization constant for $\tau_{x^{-i}}$,

$$\begin{aligned}
\log Z(\tau_{x^{-i}}) &= \log \int \exp\Big(-\frac{2}{\sigma^2}\, \delta\mathcal{F}(\rho_{x^{-i}}, z) \Big)\, \mathrm{d}z \\
&= \log \int \exp\Big(\frac{2}{\sigma^2} \left(\delta\mathcal{F}(\pi, z) - \delta\mathcal{F}(\rho_{x^{-i}}, z) \right) \Big)\, \pi(\mathrm{d}z) + \log Z(\pi) \\
&\geq -\frac{2}{\sigma^2} \int \delta\mathcal{F}(\rho_{x^{-i}}, \cdot)\, \mathrm{d}\pi - \mathsf{ent}(\pi)\,.
\end{aligned}$$

Upon taking expectations, we obtain

$$\mathsf{A} \leq \frac{\sigma^2}{2} \sum_{i=1}^N \mathbb{E}_{x^{-i} \sim \mu^{-i}}\, \mathsf{KL}\big(\mu^{i|-i}(\cdot \mid x^{-i}) \;\big\|\; \tau_{x^{-i}} \big)\,.$$

Moreover, we can recognize that $\tau_{x^{-i}}$ is a proximal Gibbs measure. By Assumptions 6 and 7,

$$\begin{aligned}
\mathsf{A} &\leq \frac{\overline{C}_{\mathsf{LSI}}\, \sigma^2}{4} \sum_{i=1}^N \mathbb{E}_{x^{-i} \sim \mu^{-i}}\, \mathsf{FI}\big(\mu^{i|-i}(\cdot \mid x^{-i}) \;\big\|\; \tau_{x^{-i}} \big) \\
&= \frac{\overline{C}_{\mathsf{LSI}}\, \sigma^2}{4} \sum_{i=1}^N \mathbb{E}_{x^{1:N} \sim \mu^{1:N}} \Big[\Big\| \nabla_{x^i} \log \mu^{i|-i}(x^i \mid x^{-i}) + \frac{2}{\sigma^2}\, \nabla_{\mathcal{W}_2} \mathcal{F}(\rho_{x^{-i}}, x^i) \Big\|^2 \Big] \\
&= \frac{\overline{C}_{\mathsf{LSI}}\, \sigma^2}{4} \sum_{i=1}^N \mathbb{E}_{x^{1:N} \sim \mu^{1:N}} \Big[\Big\| \nabla_{x^i} \log \mu^{1:N}(x^{1:N}) + \frac{2}{\sigma^2}\, \nabla_{\mathcal{W}_2} \mathcal{F}(\rho_{x^{-i}}, x^i) \Big\|^2 \Big] \\
&= \frac{\overline{C}_{\mathsf{LSI}}}{\sigma^2} \sum_{i=1}^N \mathbb{E}_{x^{1:N} \sim \mu^{1:N}} [\| \nabla_{\mathcal{W}_2} \mathcal{F}(\rho_{x^{1:N}}, x^i) - \nabla_{\mathcal{W}_2} \mathcal{F}(\rho_{x^{-i}}, x^i) \|^2] \\
&\leq \frac{\beta^2 \overline{C}_{\mathsf{LSI}}}{\sigma^2} \sum_{i=1}^N \mathbb{E}_{x^{1:N} \sim \mu^{1:N}}\, \mathcal{W}_1^2(\rho_{x^{1:N}}, \rho_{x^{-i}})\,.
\end{aligned}$$

To transport the mass from $\rho_{x^{1:N}}$ to $\rho_{x^{-i}}$, we take the transport plan which moves $\frac{1}{N(N-1)}$ of the mass from x^i to each x^j, $j \neq i$. It yields

$$\mathcal{W}_1(\rho_{x^{1:N}}, \rho_{x^{-i}}) \leq \frac{1}{N(N-1)} \sum_{\substack{j=1 \\ j \neq i}}^N \|x^i - x^j\|\,. \tag{A.6}$$

Hence,

$$\begin{aligned}
\mathsf{A} &\leq \frac{\beta^2 \overline{C}_{\mathsf{LSI}}}{\sigma^2 N^2 (N-1)^2}\, \mathbb{E}_{x^{1:N} \sim \mu^{1:N}} \sum_{i=1}^N \Big(\sum_{\substack{j=1 \\ j \neq i}}^N \|x^i - x^j\| \Big)^2 \\
&\leq \frac{\beta^2 \overline{C}_{\mathsf{LSI}}}{\sigma^2 N^2 (N-1)}\, \mathbb{E}_{x^{1:N} \sim \mu^{1:N}} \sum_{i \neq j} \|x^i - x^j\|^2 = \frac{\beta^2 \overline{C}_{\mathsf{LSI}}}{\sigma^2 N}\, \mathbb{E}_{x^{1:2} \sim \mu^{1:2}} [\|x^1 - x^2\|^2]\,.
\end{aligned}$$

We then use the inequality

$$\begin{aligned}
\frac{1}{2}\, \mathbb{E}_{x^{1:2} \sim \mu^{1:2}} [\|x^1 - x^2\|^2] &\leq 2\mathcal{W}_2^2(\mu^{1:2}, \pi^{\otimes 2}) + \mathbb{E}_{x^{1:2} \sim \pi^{\otimes 2}} [\|x^1 - x^2\|^2] \\
&\leq \frac{4}{N}\, \mathcal{W}_2^2(\mu^{1:N}, \pi^{\otimes N}) + 2\, \mathbb{E}_{x \sim \pi} [\|x - \mathbb{E}\, x\|^2]
\end{aligned}$$

$$\leq \frac{8\overline{C}_{\mathsf{LSI}}}{N}\,\mathsf{KL}(\mu^{1:N} \parallel \pi^{\otimes N}) + 2d\overline{C}_{\mathsf{LSI}}\,, \tag{A.7}$$

where we used Lemma 23 and the Poincaré inequality for π. Hence,

$$\mathsf{A} \leq \frac{2\beta^2\overline{C}_{\mathsf{LSI}}}{\sigma^2 N}\left(\frac{8\overline{C}_{\mathsf{LSI}}}{N}\,\mathsf{KL}(\mu^{1:N} \parallel \pi^{\otimes N}) + 2d\overline{C}_{\mathsf{LSI}}\right).$$

Next, we turn toward term B. First, define a function $\zeta^i_{x^{1:N}} : \mathbb{R}^d \to \mathbb{R}$ by

$$\zeta^i_{x^{1:N}}(y) \coloneqq \delta\mathcal{F}(\rho_{x^{-i}}, y) - \delta\mathcal{F}(\rho_{x^{1:N}}, y)\,.$$

It is clear from Assumption 6 that this function is Lipschitz with constant $2\beta\mathcal{W}_1(\rho_{x^{1:N}}, \rho_{x^{-i}})$. Thus, we obtain using this Lipschitzness, (A.6), and Young's inequality,

$$\begin{aligned}
\mathsf{B} &= \sum_{i=1}^N \mathbb{E}_{x^{1:N}\sim\mu^{1:N}} \int \left(\zeta^i_{x^{1:N}}(x^i) - \zeta^i_{x^{1:N}}(z)\right)\pi(\mathrm{d}z) \\
&\leq \sum_{i=1}^N \mathbb{E}_{x^{1:N}\sim\mu^{1:N}} \int \frac{2\beta}{N\,(N-1)} \sum_{\substack{j=1\\ j\neq i}}^N \|x^j - x^i\|\,\|x^i - z\|\,\pi(\mathrm{d}z) \\
&\leq \frac{\beta}{N\,(N-1)}\,\mathbb{E}_{x^{1:N}\sim\mu^{1:N}} \sum_{i\neq j} \|x^i - x^j\|^2 + \frac{\beta}{N}\sum_{i=1}^N \mathbb{E}_{(x^i,z)\sim\mu^1\otimes\pi}[\|x^i - z\|^2] \\
&= \beta\,\mathbb{E}_{x^{1:2}\sim\mu^{1:2}}[\|x^1 - x^2\|^2] + \beta\,\mathbb{E}_{(x,z)\sim\mu^1\otimes\pi}[\|x - z\|^2]\,.
\end{aligned}$$

For the first term, we can apply (A.7), and for the second term, we can apply (A.1). It yields

$$\mathsf{B} \leq \frac{20\beta\overline{C}_{\mathsf{LSI}}}{N}\,\mathsf{KL}(\mu^{1:N} \parallel \pi^{\otimes N}) + 8\beta\overline{C}_{\mathsf{LSI}}d\,.$$

Putting the bounds together with Lemma 19,

$$\mathsf{KL}(\mu^{1:N} \parallel \pi^{\otimes N}) \leq \frac{33\beta\overline{C}_{\mathsf{LSI}}d}{\sigma^2}$$

for all $N \geq 160\beta\overline{C}_{\mathsf{LSI}}/\sigma^2$. The result for $k \leq N$ follows from Lemma 24. □

B Isoperimetric Results for the Stationary Distributions

B.1 Convexity and Smoothness

Here, we verify the convexity and smoothness properties of $\mu^{1:N}$ in the pairwise McKean–Vlasov setting.

Proof of Lemma 7 For $x^{1:N} = [x^1, \ldots, x^N] \in \mathbb{R}^{d\times N}$, the Hessian of $\log(1/\mu^{1:N})$ can be explicitly computed as

$$-\frac{\sigma^2}{2}\,\nabla^2 \log \mu^{1:N}(x^{1:N}) = \begin{bmatrix} \nabla^2 V(x^1) & 0 & \cdots & 0 \\ 0 & \nabla^2 V(x^2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \nabla^2 V(x^N) \end{bmatrix}$$
$$+ \frac{1}{N-1}\underbrace{\begin{bmatrix} \sum_{j=2}^N \nabla^2 W(x^1 - x^j) & -\nabla^2 W(x^1 - x^2) & \cdots & -\nabla^2 W(x^1 - x^N) \\ -\nabla^2 W(x^2 - x^1) & \sum_{\substack{j=1\\ j\neq 2}}^N \nabla^2 W(x^2 - x^j) & \cdots & -\nabla^2 W(x^2 - x^N) \\ \vdots & \vdots & \ddots & \vdots \\ -\nabla^2 W(x^N - x^1) & -\nabla^2 W(x^N - x^2) & \cdots & \sum_{j=1}^{N-1} \nabla^2 W(x^N - x^j) \end{bmatrix}}_{=\mathsf{B}}.$$

Clearly, the first block matrix has eigenvalues between α_V and β_V. For the second block matrix B, let us denote $A_{i,j} \coloneqq \nabla^2 W(x^i - x^j)$ for $i, j \in [N]$. Note that $A_{i,j} = A_{j,i}$ since W is even, and each $A_{i,j}$ is clearly symmetric.

For $\mathsf{B} \in \mathbb{R}^{dN \times dN}$ the second matrix and $y = [y^1, \dots, y^N] \in \mathbb{R}^{dN}$, we have

$$\begin{aligned} y^\mathsf{T} \mathsf{B} y &= \sum_{i \le N} y_i^\mathsf{T} \Big(\sum_{j \in [N] \setminus i} A_{i,j} \Big) y_i - \sum_{\substack{i,j \le N \\ i \ne j}} y_i^\mathsf{T} A_{i,j} y_j \\ &= \sum_{\substack{i,j \le N \\ i < j}} \left(y_i^\mathsf{T} A_{i,j} y_i + y_j^\mathsf{T} A_{j,i} y_j - y_i^\mathsf{T} A_{i,j} y_j - y_j^\mathsf{T} A_{j,i} y_i \right) = \sum_{\substack{i,j \le N \\ i < j}} (y_i - y_j)^\mathsf{T} A_{i,j} (y_i - y_j) \,. \end{aligned}$$

Using $\alpha_W I_d \preceq A_{i,j} \preceq \beta_W I_d$ and

$$\mathsf{M} \coloneqq \nabla_y^2 \sum_{\substack{i,j \le N \\ i < j}} \|y_i - y_j\|^2 = 2 \begin{bmatrix} N-1 & -1 & \cdots & -1 \\ -1 & N-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & N-1 \end{bmatrix} \otimes I_d \,,$$

we have $\frac{1}{2}\,\alpha_W \mathsf{M} \preceq \mathsf{B} \preceq \frac{1}{2}\,\beta_W \mathsf{M}$. Since the circulant matrix in M is PSD due to diagonal dominance and its largest eigenvalue is at most N, it follows that the eigenvalues of M lie between 0 and $2N$. Hence, the eigenvalues of B lie in the interval $[\frac{N}{N-1}\,\alpha_W^-, \frac{N}{N-1}\,\beta_W]$. □

B.2 Bounded Perturbations

In this section, we prove the isoperimetric results from §3.2.1. We again introduce the conditional measure: if $x^{-i} \coloneqq (x^{1:i-1}, x^{i+1:N}) \in \mathbb{R}^{d \times (N-1)}$ we define

$$\mu^{1:N}(x^{1:N}) = \mu^{i|-i}(x^i \mid x^{-i}) \times \mu^{-i}(x^{-i})$$

for the conditional distribution of the i-th particle and the distribution of an N-particle system with the i-th particle marginalized out.

Proof of Lemma 8 We begin by proving the statement about π. The potential of the invariant measure π can also be written as

$$\begin{aligned} \log \frac{1}{\pi(x)} &= \frac{2}{\sigma^2} \Big(V_0(x) + V_1(x) + \int \big(W_0(x - \cdot) + W_1(x - \cdot) \big) \,\mathrm{d}\pi \Big) \\ &= \frac{2}{\sigma^2} \Big(V_0(x) + \int W_0(x - \cdot) \,\mathrm{d}\pi \Big) + \frac{2}{\sigma^2} \Big(V_1(x) + \int W_1(x - \cdot) \,\mathrm{d}\pi \Big) \,. \end{aligned}$$

This is the sum of a $\frac{2}{\sigma^2}\,(\alpha_{V_0} + \alpha_{W_0})$-convex function with a $\frac{2}{\sigma^2}\,(\mathsf{osc}(V_1) + \mathsf{osc}(W_1))$-bounded perturbation. Thus, π satisfies (LSI) with the claimed parameter.

We now prove the statement about $\mu^{1:N}$. By the tensorization argument in [BGL14, Proposition 5.2.7] or [BLM13, Theorem 4.10], we obtain the inequality that for any smooth test function $g^2 : \mathbb{R}^{d \times N} \to \mathbb{R}_+$ and $X^{-i} \sim \mu^{-i}$,

$$\mathsf{ent}_{\mu^{1:N}}(g^2) \le \sum_{i=1}^N \mathbb{E}\, \mathsf{ent}_{\mu^{i|-i}(\cdot \mid X^{-i})} \big(g^2(X^{1:i-1}, \cdot, X^{i+1:N}) \big) \,, \tag{B.1}$$

where each entropy term in RHS involves the distribution of the i-th particle, conditioned on the remaining particles. This conditional measure has a density of the form

$$\mu^{i|-i}(\cdot \mid x^{-i}) \propto \exp\Big(-\frac{2}{\sigma^2}\, V(\cdot) - \frac{2}{\sigma^2\,(N-1)} \sum_{j \in [N] \setminus i} W(\cdot - x^j) \Big) \,,$$

where both V and W are bounded perturbations of α_{V_0}, α_{W_0}-strongly convex functions respectively, irrespective of the conditional variables. Thus, by Holley–Stroock perturbation and the Bakry–Émery condition, each $\mu^{i|-i}(\cdot \mid x^{-i})$ satisfies (LSI) with parameter

$$C_{\mathsf{LSI}} \le \frac{\sigma^2}{2}\left(\alpha_{V_0} + \frac{N}{N-1}\,\alpha_{W_0}^{-}\right)^{-1} \exp\Big(\frac{2}{\sigma^2}\left(\mathsf{osc}(V_1) + \mathsf{osc}(V_1)\right)\Big)\,.$$

Therefore, we can further bound each entropy term in (B.1) by

$$\mathsf{ent}_{\mu^{1:N}}(g^2) \le \sum_{i=1}^{N} 2C_{\mathsf{LSI}}\,\mathbb{E}\,\mathbb{E}_{Z\sim\mu^{i|-i}(\cdot|X^{-i})}[|\partial_i g(X^{1:i-1}, Z, X^{i+1:N})|^2] = 2C_{\mathsf{LSI}}\,\mathbb{E}_{\mu^{1:N}}[\|\nabla g\|^2]\,.$$

Therefore, $\mu^{1:N}$ satisfies (LSI) with parameter C_{LSI}. □

B.3 Logarithmic Sobolev Inequalities via Perturbations

In this section, we state log-Sobolev inequalities for Lipschitz perturbations of strongly log-concave measures, which is used for the general McKean–Vlasov setting in §2.1.2.

Lemma 20 (LSI under Lipschitz Perturbations [CG22, Theorem 2.7])**.** *Let $\nu \propto \exp(-V)$ for an α_V-strongly convex and β_V-smooth function $V : \mathbb{R}^d \to \mathbb{R}$. Let $H : \mathbb{R}^d \to \mathbb{R}$ be a L-Lipschitz continuous function. Then, $\mu \propto \exp(-H)\,\nu$ satisfies (LSI) with constant $C_{\mathsf{LSI}}(\mu)$ given by*

$$C_{\mathsf{LSI}}(\mu) \le \frac{4}{\alpha_V} + \Big(\frac{L}{\alpha_V} + \sqrt{\frac{2}{\alpha_V}}\Big)^2 \Big(2 + d + \frac{d}{2}\log\frac{\beta_V}{\alpha_V} + \frac{4L^2}{\alpha_V}\Big)\exp\Big(\frac{L^2}{2\alpha_V}\Big)\,.$$

From this one derives the following log-Sobolev inequality for the proximal Gibbs measure.

Lemma 21 (Uniform LSI for the Proximal Gibbs Measure [SWN23, Theorem 1])**.** *For the proximal Gibbs measure (2.3) in the setting of §2.1.2, under Assumptions 6, and 8, we have that*

$$\sup_{\mu\in\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)} C_{\mathsf{LSI}}(\pi_\mu) \le \overline{C}_{\mathsf{LSI}}\,,$$

where α can be bounded by

$$\overline{C}_{\mathsf{LSI}} \le \frac{\sigma^2}{\beta}\exp\Big(\frac{8B^2}{\lambda\sigma^2}\sqrt{2d/\pi}\Big) \wedge \Big\{\frac{2\sigma^2}{\beta} + \big(\frac{B}{\beta} + \frac{\sigma}{\sqrt{\beta}}\big)^2\big(2 + d + \frac{8B^2}{\beta\sigma^2}\big)\exp\big(\frac{B^2}{\lambda\sigma^2}\big)\Big\}\,.$$

Finally, we provide an LSI for the finite-particle stationary distribution $\mu^{1:N}$ with a constant independent of N, using an approximate tensorization argument.

Lemma 22. *Consider the general McKean–Vlasov setting with $\mathcal{F}(\mu) = \mathcal{F}_0(\mu) + \frac{\lambda}{2}\int\|\cdot\|^2\,\mathrm{d}\mu$ and suppose that Assumption 8 holds. Then, $\mu^{1:N}$ satisfies (LSI) with parameter independent of N given by*

$$C_{\mathsf{LSI}}(\mu^{1:N}) \le \frac{4\sigma^2}{\lambda} + \Big(\frac{2B}{\lambda} + \frac{\sigma}{\sqrt{\lambda}}\Big)^2\Big(2 + d + \frac{16B^2}{\lambda\sigma^2}\Big)\exp\Big(\frac{2B^2}{\lambda\sigma^2}\Big)\,.$$

Proof. Since $\mu^{1:N}(x^{1:N}) \propto \exp(-\frac{2N}{\sigma^2}\,\mathcal{F}(\rho_{x^{1:N}}))$, where $\rho_{x^{1:N}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{x^i}$ is the empirical measure, we can check that

$$\mu^{i|-i}(x^i \mid x^{-i}) \propto \exp\Big(-\frac{2N}{\sigma^2}\,\mathcal{F}_0(\rho_{x^{1:N}}) - \frac{\lambda}{\sigma^2}\,\|x^i\|^2\Big)\,.$$

Also, note that

$$\nabla_{x^i}\mathcal{F}_0(\rho_{x^{1:N}}) = \frac{1}{N}\,\nabla_{\mathcal{W}_2}\mathcal{F}_0(\rho_{x^{1:N}}, x^i)\,.$$

Using Assumption 8, $x^i \mapsto \frac{2N}{\sigma^2}\mathcal{F}_0(\rho_{x^{1:N}})$ is $2B/\sigma^2$-Lipschitz.

Then, apply Lemma 20 with $\alpha_V = \beta_V = \lambda/\sigma^2$, to obtain

$$C_{\mathsf{LSI}}\big(\mu^{i|-i}(\cdot \mid x^{-i})\big) \le \frac{4\sigma^2}{\lambda} + \Big(\frac{2B}{\lambda} + \frac{\sigma}{\sqrt{\lambda}}\Big)^2 \Big(2 + d + \frac{16B^2}{\lambda\sigma^2}\Big) \exp\Big(\frac{2B^2}{\lambda\sigma^2}\Big)\,.$$

Applying the tensorization argument from the proof of Lemma 8 completes the proof. □

C Explicit Calculations for the Gaussian Case

Here we provide complete details for Example 10: for any $k \le N$,

$$\frac{dk^2}{N^2} \lesssim \mathsf{KL}(\mu^{1:k} \,\|\, \pi^{\otimes k}) \lesssim \frac{dk^2}{N^2}\log N\,.$$

Note that for $\mathsf{C} \in \mathbb{R}^{N\times N}$ with $\mathsf{C}_{i,i} = N-1$ and $\mathsf{C}_{i,j} = -1$ if $i \ne j$,

$$\mu^{1:N} = \mathcal{N}\Big(0, \underbrace{\frac{\sigma^2}{2}\Big(I_N \otimes A + \frac{\lambda}{N-1}\,\mathsf{C}\otimes I_d\Big)^{-1}}_{=:\Sigma_1}\Big) \qquad \text{and} \qquad \pi = \mathcal{N}\Big(0, \underbrace{\frac{\sigma^2}{2}(A+\lambda I_d)^{-1}}_{=:\Sigma_2}\Big)\,.$$

The k-particle marginal $\mu^{1:k}$ is a Gaussian with zero mean and covariance being the upper-left $(kN\times kN)$-block matrix of Σ_1, which we denote by $\Sigma_{1,k}$. Clearly, $\pi^{\otimes k}$ is also a Gaussian with zero mean and covariance $\Sigma_{2,k} \coloneqq I_k \otimes \Sigma_2$. From a well-known formula for the KL divergence between two Gaussian distributions,

$$\mathsf{KL}(\mu^{1:k} \,\|\, \pi^{\otimes k}) = \frac{1}{2}\big(-\log\det(\Sigma_{2,k}^{-1}\Sigma_{1,k}) - dk + \mathrm{tr}(\Sigma_{2,k}^{-1}\Sigma_{1,k})\big)\,. \tag{C.1}$$

Let $1_p \in \mathbb{R}^p$ be the p-dimensional vector with all entries 1. From $\mathsf{C} = NI_N - 1_N 1_N^\mathsf{T}$,

$$\begin{aligned}
\frac{2}{\sigma^2}\Sigma_1 &= \Big(I_N \otimes \underbrace{\big(A + \frac{\lambda N}{N-1}I_d\big)}_{=:A_\lambda} - \frac{\lambda}{N-1}(1_N 1_N^\mathsf{T})\otimes I_d\Big)^{-1} \\
&\underset{\text{(i)}}{=} \Big(I_N \otimes A_\lambda - \frac{\lambda}{N-1}(1_N\otimes I_d)(1_N^\mathsf{T}\otimes I_d)\Big)^{-1} \\
&\underset{\text{(ii)}}{=} (I_N\otimes A_\lambda)^{-1} \\
&\qquad - (I_N\otimes A_\lambda)^{-1}(1_N\otimes I_d)\big(I_d + (1_N^\mathsf{T}\otimes I_d)(I_N\otimes A_\lambda)^{-1}(1_N\otimes I_d)\big)^{-1}(1_N^\mathsf{T}\otimes I_d)(I_N\otimes A_\lambda)^{-1} \\
&\underset{\text{(iii)}}{=} I_N\otimes A_\lambda^{-1} - (1_N\otimes A_\lambda^{-1})\big(I_d + (1_N^\mathsf{T}1_N)\otimes A_\lambda^{-1}\big)^{-1}(1_N^\mathsf{T}\otimes A_\lambda^{-1}) \\
&= I_N\otimes A_\lambda^{-1} - (1_N\otimes A_\lambda^{-1})(I_d + NA_\lambda^{-1})^{-1}(1_N^\mathsf{T}\otimes A_\lambda^{-1}) \\
&= I_N\otimes A_\lambda^{-1} - (1_N 1_N^\mathsf{T})\otimes(A_\lambda^2 + NA_\lambda)^{-1}\,,
\end{aligned}$$

where in (i) we used $(A\otimes B)(C\otimes D) = (AC)\otimes(BD)$, (ii) follows from the Woodbury matrix identity, and (iii) used $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$. Hence it follows that

$$\frac{2}{\sigma^2}\Sigma_{1,k} = I_k\otimes A_\lambda^{-1} - (1_k 1_k^\mathsf{T})\otimes(A_\lambda^2 + NA_\lambda)^{-1}\,.$$

By the spectral decomposition of A, we can write $A = UDU^\mathsf{T}$ for a diagonal $D \in \mathbb{R}^{d\times d}$ and an orthogonal matrix $U \in \mathbb{R}^{d\times d}$ such that $\{\sigma_i \coloneqq D_{i,i}\}_{i\in[d]}$ correspond to the eigenvalues of A. Since $\log\det(\cdot)$ and $\mathrm{tr}(\cdot)$ in (C.1) are orthogonally invariant, let us look at the orthogonal conjugate of $\Sigma_{2,k}^{-1}\Sigma_{1,k}$ by $I_k\otimes U^\mathsf{T} \in \mathbb{R}^{dk\times dk}$. Using $(A\otimes B)^\mathsf{T} = A^\mathsf{T}\otimes B^\mathsf{T}$ and denoting $D_\lambda \coloneqq D + \frac{\lambda N}{N-1}I_d$,

$$(I_k\otimes U^\mathsf{T})\Sigma_{2,k}^{-1}\Sigma_{1,k}(I_k\otimes U)$$

$$
\begin{aligned}
&= (I_k \otimes U^\mathsf{T})\big(I_k \otimes (A + \lambda I_d)\big)\big(I_k \otimes A_\lambda^{-1} - (1_k 1_k^\mathsf{T}) \otimes (A_\lambda^2 + N A_\lambda)^{-1}\big)(I_k \otimes U) \\
&= \big(I_k \otimes (D + \lambda I_d)\big)\big(I_k \otimes D_\lambda^{-1} - (1_k 1_k^\mathsf{T}) \otimes (D_\lambda^2 + N D_\lambda)^{-1}\big) \\
&= I_k \otimes \big((D + \lambda I_d) D_\lambda^{-1}\big) - \underbrace{(1_k 1_k^\mathsf{T})}_{=:J_k} \otimes \underbrace{\big((D + \lambda I_d)(D_\lambda^2 + N D_\lambda)^{-1}\big)}_{=:S_\lambda} \\
&= I_{dk} - \Big(\underbrace{\frac{\lambda}{N-1} I_k \otimes D_\lambda^{-1} + J_k \otimes S_\lambda}_{=:\mathsf{M}}\Big) .
\end{aligned}
$$

For $\sigma_d \coloneqq \min_{i\in[d]} \sigma_i$, $\alpha \coloneqq \sigma_d + \lambda$, and $\varepsilon \coloneqq \lambda/(N-1)$, we have $D_\lambda^{-1} \precsim \frac{1}{\alpha} I_d$ and $S_\lambda \precsim \frac{1}{\alpha+N} I_d$ due to

$$
((D_\lambda)^{-1})_{i,i} \le \frac{1}{\alpha + \varepsilon} \qquad \text{and} \qquad (S_\lambda)_{i,i} \le \frac{\alpha}{(\alpha+\varepsilon)(\alpha+\varepsilon+N)} .
$$

Since the eigenvalues of $A \otimes B$ consist of all possible combinations arising from the product of eigenvalues, one from A and one from B, the largest eigenvalue η_1 of M is less than 1:

$$
\eta_1 \le \frac{\lambda}{N-1} \|D_\lambda^{-1}\| + k \|S_\lambda\| \le \frac{\varepsilon}{\alpha+\varepsilon} + \frac{\alpha N}{(\alpha+\varepsilon)(\alpha+\varepsilon+N)} = \frac{\varepsilon + N}{\alpha + \varepsilon + N} .
$$

Denoting the eigenvalues of M by η_i, it follows from (C.1) that

$$
\begin{aligned}
2\,\mathsf{KL}(\mu^{1:k} \,\|\, \pi^{\otimes k}) &= -\big(\log\det(I_{dk} - \mathsf{M}) + dk - \operatorname{tr}(I_{dk} - \mathsf{M})\big) = -\sum_{i=1}^{dk} \big(\log(1-\eta_i) + \eta_i\big) \\
&= \sum_{i=1}^{dk} \sum_{n\ge 2} \frac{\eta_i^n}{n} .
\end{aligned}
$$

Then, we have a trivial lower bound of $\frac{1}{2}\sum_{i=1}^{dk} \eta_i^2$, and as for the upper bound,

$$
\sum_{i=1}^{dk} \sum_{n\ge 2} \frac{\eta_i^n}{n} \le \sum_{i=1}^{dk} \Big(\eta_i^2 + \sum_{n\ge 1} \frac{\eta_i^{n+2}}{n}\Big) = \sum_{i=1}^{dk} \eta_i^2 \log\big(\frac{e}{1-\eta_i}\big) \lesssim \big(1 \vee \log\frac{N}{\alpha}\big) \operatorname{tr}(\mathsf{M}^2) ,
$$

where the last inequality follows from $(1-\eta_i)^{-1} \le (1-\eta_1)^{-1}$.

Using $\operatorname{tr}(A \otimes B) = \operatorname{tr}(A) \cdot \operatorname{tr}(B)$, and $D_\lambda^{-1} \precsim \frac{1}{\alpha} I_d$ and $S_\lambda \precsim \frac{1}{\alpha+N} I_d$, we have

$$
\begin{aligned}
\operatorname{tr}(\mathsf{M}^2) &\lesssim \frac{\lambda^2}{(N-1)^2} \operatorname{tr}(I_k) \operatorname{tr}(D_\lambda^{-2}) + \operatorname{tr}(J_k^2) \operatorname{tr}(S_\lambda^2) \\
&\lesssim \frac{\lambda^2}{\alpha^2} \frac{dk}{N^2} + \frac{dk^2}{(\alpha+N)^2} \lesssim \frac{dk^2}{N^2} .
\end{aligned}
$$

As for the lower bound, since $(S_\lambda)_{i,i} \sim \frac{1}{N}$ for large N, we have

$$
\operatorname{tr}(\mathsf{M}^2) \gtrsim \frac{dk^2}{N^2} ,
$$

which completes the proof.

D Additional Technical Lemmas

In our proofs, we used the following general lemmas on exchangeability.

Lemma 23. *Let $\mu^{1:N}$, $\nu^{1:N}$ be two exchangeable measures over $\mathbb{R}^{d\times N}$. For any $k \le N$,*

$$
\mathcal{W}_2^2(\mu^{1:k}, \nu^{1:k}) \le \frac{k}{N} \mathcal{W}_2^2(\mu^{1:N}, \nu^{1:N}) .
$$

Proof. Let $(X^{1:N}, Y^{1:N})$ be optimally coupled for $\mu^{1:N}$ and $\nu^{1:N}$. For each subset $S \subseteq [N]$ of size k, it induces a coupling (X^S, Y^S) of $\mu^{1:k}$ and $\nu^{1:k}$ (by exchangeability). In particular, the law of $(X^{\mathsf{S}}, Y^{\mathsf{S}})$, where S is an independent and uniformly random subset of size k, is also a coupling of $\mu^{1:k}$ and $\nu^{1:k}$. Hence,

$$\begin{aligned}\mathcal{W}_2^2(\mu^{1:k}, \nu^{1:k}) &\le \mathbb{E}[\|X^{\mathsf{S}} - Y^{\mathsf{S}}\|^2] = \frac{1}{\binom{N}{k}} \sum_{|S|=k} \mathbb{E}[\|X^S - Y^S\|^2] \\ &= \frac{1}{\binom{N}{k}} \sum_{i=1}^{N} \sum_{|S|=k\,:\,i\in S} \mathbb{E}[\|X^i - Y^i\|^2] = \frac{\binom{N-1}{k-1}}{\binom{N}{k}} \sum_{i=1}^{N} \mathbb{E}[\|X^i - Y^i\|^2] \\ &= \frac{k}{N}\, \mathcal{W}_2^2(\mu^{1:N}, \nu^{1:N})\,,\end{aligned}$$

which completes the proof. □

Lemma 24 (Information Inequality [Csi84])**.** *If $\mathcal{X}^1, \dots, \mathcal{X}^N$ are Polish spaces and $\mu^{1:N}$, $\nu^{1:N}$ are probability measures on $\mathcal{X}^1 \times \cdots \times \mathcal{X}^N$, where $\nu^{1:N} = \nu^1 \otimes \cdots \otimes \nu^N$ is a product measure, then for the marginals μ^i of μ, it holds that*

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

In particular when $\mu^{1:N}$, $\nu^{1:N}$ are both exchangeable, this states that $\mathsf{KL}(\mu^1 \,\|\, \nu^1) \le \frac{1}{N}\,\mathsf{KL}(\mu^{1:N} \,\|\, \nu^{1:N})$.

Note that Lemma 24 follows from the chain rule and convexity of the KL divergence.

E Sampling Guarantees

Here, we show how to obtain the claimed rates in §4. We begin with some preliminary facts.

KL divergence guarantees. To obtain our guarantees in KL divergence, we use the following lemma.

Lemma 25 ([Zha+23, Proof of Theorem 6])**.** *Let $\hat{\mu}$, μ, and π be three probability measures, and assume that μ satisfies (LSI) with constant $C_{\mathsf{LSI}}(\mu)$. Then,*

$$\mathsf{KL}(\hat{\mu} \,\|\, \pi) \le 2\,\chi^2(\hat{\mu} \,\|\, \mu) + \mathsf{KL}(\mu \,\|\, \pi) + \frac{C_{\mathsf{LSI}}(\mu)}{4}\,\mathsf{FI}(\mu \,\|\, \pi)\,.$$

We instantiate the lemma with $\hat{\mu}^{1:N}$, $\mu^{1:N}$, and $\pi^{\otimes N}$ respectively. In the setting of Theorem 4, it is seen that $\mathsf{KL}(\mu^{1:N} \,\|\, \pi^{\otimes N})$ and $C_{\mathsf{LSI}}(\mu^{1:N})\,\mathsf{FI}(\mu^{1:N} \,\|\, \pi^{\otimes N})$ are of the same order and can be made at most $N\varepsilon^2$ if we take $N \asymp \kappa^4 d/\varepsilon^2$. Thus, if we have a sampler that achieves $\chi^2(\hat{\mu}^{1:N} \,\|\, \mu^{1:N}) \le N\varepsilon^2$, it follows from exchangeability (Lemma 24) that $\mathsf{KL}(\hat{\mu}^1 \,\|\, \pi) \lesssim \varepsilon^2$.

Guarantees using the sharp propagation of chaos bound. Here, we impose Assumptions 1, 2, 3, and 9. It follows from Theorem 3 that $N = \widetilde{\Theta}(\sqrt{d}/\varepsilon)$ suffices in order to make $\sqrt{\alpha}/\sigma\, \mathcal{W}_2(\mu^1, \pi) \le \varepsilon$. For the first term, we use exchangeability (Lemma 23) to argue that

$$\mathcal{W}_2(\hat{\mu}^1, \mu^1) \le N^{-1/2}\, \mathcal{W}_2(\hat{\mu}^{1:N}, \mu^{1:N})\,,$$

and hence we invoke sampling guarantees to ensure that $\sqrt{\alpha}/\sigma\, \mathcal{W}_2(\hat{\mu}^{1:N}, \mu^{1:N}) \le N^{1/2}\varepsilon$ under (LSI).

- **LMC:** We use the guarantee for Langevin Monte Carlo from [VW19].
- **MALA–PS:** We use the guarantee for the Metropolis-adjusted Langevin algorithm together with the proximal sampler from [AC23]. Note that the iteration complexity is $\widetilde{\mathcal{O}}(\kappa d^{1/2} N^{1/2})$, and we substitute in the chosen value for N.

- **ULMC–PS:** Here, we use underdamped Langevin Monte Carlo to implement the proximal sampler. To justify the sampling guarantee, note that since $\log \mu^{1:N}$ is β-smooth, if we choose step size $h = \frac{1}{2\beta}$ for the proximal sampler, then the RGO is β-strongly log-concave and 3β-log-smooth. According to [AC23, Proof of Theorem 5.3], it suffices to implement the RGO in each iteration to accuracy $N^{1/2}\varepsilon/\kappa^{1/2}$ in $\sqrt{\mathsf{KL}}$. Then, from [Zha+23], this can be done via ULMC with complexity $\widetilde{\mathcal{O}}(\kappa^{1/2}d^{1/2}/\varepsilon)$. Finally, since the number of outer iterations of the proximal sampler is $\widetilde{\mathcal{O}}(\kappa)$, we obtain the claim.

- **ULMC$^+$:** Here, we use either the randomized midpoint discretization [SL19] or the shifted ODE discretization [FLO21] of the underdamped Langevin diffusion. We also replace the LSI assumptions (Assumptions 2 and 9) with strong convexity (Assumption 4).

Guarantees under strong displacement convexity. Here, we impose Assumptions 1 and 4. As discussed above, to obtain KL guarantees, we require log-concave samplers that can achieve $\chi^2(\hat{\mu}^{1:N} \parallel \mu^{1:N}) \leq N\varepsilon^2$. For $\mathcal{W}_2$ guarantees, by Theorem 4 we take $N \asymp \kappa^2 d/\varepsilon^2$ and we require log-concave samplers that can achieve $\sqrt{\alpha}/\sigma\, \mathcal{W}_2(\hat{\mu}^{1:N}, \mu^{1:N}) \leq N^{1/2}\varepsilon$.

- **LMC:** For Langevin Monte Carlo, we use the χ^2 guarantee from [Che+22b] and the $\mathcal{W}_2$ guarantee from [DMM19].

- **ULMC:** For underdamped Langevin Monte Carlo, we use the χ^2 guarantee from [AC23].

- **ULMC$^+$:** Here, we use the $\mathcal{W}_2$ guarantees for either the randomized midpoint discretization [SL19] or the shifted ODE discretization [FLO21] of the underdamped Langevin diffusion.

Guarantees in the general McKean–Vlasov setting. In the setting of Theorem 5, we take $N \asymp \kappa d/\varepsilon^2$. We use the same sampling guarantees under (LSI) as in the prior settings.

We also note that in order to apply the log-concave sampling guarantees, we must check that $\mu^{1:N}$ is log-smooth. This follows from Assumption 6. Indeed,

$$\begin{aligned}
\|\nabla \log \mu^{1:N}(x^{1:N}) - \nabla \log \mu^{1:N}(y^{1:N})\| &= \frac{2}{\sigma^2}\sqrt{\sum_{i=1}^N \|\nabla_{\mathcal{W}_2}\mathcal{F}(\rho_{x^{1:N}}, x^i) - \nabla_{\mathcal{W}_2}\mathcal{F}(\rho_{y^{1:N}}, y^i)\|^2} \\
&\leq \frac{2\sqrt{2}\,\beta}{\sigma^2}\sqrt{\sum_{i=1}^N \big(\|x^i - y^i\|^2 + \mathcal{W}_1^2(\rho_{x^{1:N}}, \rho_{y^{1:N}})\big)} \\
&\leq \frac{4\beta}{\sigma^2}\,\|x^{1:N} - y^{1:N}\|\,.
\end{aligned}$$