IMPROVED ANALYSIS OF STEIN VARIATIONAL GRADIENT DESCENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Stein Variational Gradient Descent (SVGD) is a particle-based variational inference method that transports a collection of particles toward a target distribution through deterministic interacting updates. Existing convergence analyses of SVGD can broadly be divided into two regimes: Mean-Field SVGD, which studies the population (infinite-particle) limit of the algorithm, and Finite-Particle SVGD, which directly analyzes the practical algorithm with finitely many interacting particles. Despite substantial recent progress, existing theories in both regimes typically rely on restrictive assumptions that significantly limit their applicability. For instance, available convergence guarantees for Mean-Field SVGD may fail to cover even commonly used target distributions such as Gaussian mixtures, while existing finite-particle analyses can impose conditions that exclude even Gaussian target distributions. In this work, we develop an improved convergence analysis of SVGD that addresses these limitations in both the mean-field and finite-particle settings. Our results establish convergence under weaker assumptions than those required in prior work, thereby covering a broader class of practically relevant target distributions. Moreover, our analysis yields improved convergence guarantees over existing results. Together, these results provide a more general and sharper theoretical understanding of SVGD and help narrow the gap between its theoretical guarantees and its practical applicability.

1 INTRODUCTION

Sampling from probability distributions is a fundamental task in Bayesian inference, machine learning, computational statistics, and related fields. Classical approaches such as Markov chain Monte Carlo (MCMC) (Metropolis et al., 1953; Hastings, 1970; Robert & Casella, 2011) generate samples through stochastic transitions, while variational inference (Blei et al., 2017) typically approximates the target distribution within a prescribed parametric family. Stein Variational Gradient Descent (SVGD), introduced by Liu & Wang (2016), provides an alternative that combines features of particle-based sampling and variational inference. Starting from a collection of particles, SVGD deterministically transports them to match the target distribution through iterative updates designed to decrease the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) as rapidly as possible at each step.

Since its introduction, SVGD has attracted considerable attention and has been successfully applied to a wide range of problems (Haarnoja et al., 2017; Liu et al., 2017; Pu et al., 2017; Zhang & Curtis, 2020; Barcelos et al., 2021; Lambert et al., 2021; Jia et al., 2022; Lee et al., 2025; Corrales et al., 2025), owing to its strong empirical performance and broad applicability. Its particle-based formulation is particularly appealing, as it provides a flexible nonparametric approximation to the target distribution, while its deterministic updates are straightforward to implement. This combination of flexibility, algorithmic simplicity, and favorable empirical performance has made SVGD an appealing alternative to both conventional MCMC methods and parametric variational inference.

Despite its empirical success, establishing rigorous convergence guarantees for SVGD has remained challenging. Existing theoretical analyses of SVGD have largely developed along two directions. The first is the mean-field regime, in which the number of particles is taken to infinity and the algorithm is described by the evolution of the underlying probability measure. The second is the finite-particle regime, which directly studies the particle-based update implemented in practice.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In the mean-field regime, a series of works (Korba et al., 2020; Salim et al., 2022; Sun et al., 2023; Fujisawa & Futami, 2025) has established sublinear convergence rates for SVGD. However, existing results typically rely on two types of assumptions: one concerns the smoothness of the potential function V of the target distribution (see (1) for the definition of V), and the other involves trajectory-dependent bounds, transport inequalities, or functional inequalities on the target distribution (see Subsection 1.2 for details). While such assumptions provide useful tools for establishing convergence, they can be restrictive for general non-log-concave and multimodal targets. In particular, even setting aside the additional conditions imposed on the target distribution, Gaussian mixture distributions can fail to satisfy the existing (generalized) smoothness conditions on V (see Example 1 in Appendix A). Therefore, the scope of existing convergence guarantees remains limited. Motivated by this limitation, we are led to the following research question:

Q1: Can we establish convergence guarantees for Mean-Field SVGD under more general and less restrictive assumptions that cover at least Gaussian mixtures as target distributions?

The finite-particle regime has fewer known theoretical results. Shi & Mackey (2023) established the first finite-particle convergence rate for SVGD, showing convergence in Kernelized Stein Discrepancy (KSD) (Chwialkowski et al., 2016; Liu et al., 2016; Gorham & Mackey, 2017) under a decay condition on the kernel derivatives, Talagrand's T_1 inequality (Villani, 2009, Definition 22.1), and two other standard assumptions, but with a relatively slow dependence on the number of particles. More recently, Banerjee et al. (2025) substantially improved the rate under additional growth and regularity conditions. However, the additional growth condition excludes even Gaussian targets (see Subsection 1.2 for details). In a complementary direction, Liu et al. (2023) obtained sharper convergence guarantees for centered Gaussian targets, but their analysis is specialized to SVGD with bilinear kernels and requires additional conditions on the initialization. Thus, existing finite-particle analyses either provide general guarantees under nontrivial structural assumptions or obtain sharper results in specialized settings. These limitations naturally motivate the following research question:

Q2: Can we establish sharp convergence guarantees for standard Finite-Particle SVGD under broadly applicable conditions that cover at least Gaussian target distributions?

1.1 Our Contributions

This work answers the two questions above affirmatively by providing an improved analysis of SVGD in both mean-field and finite-particle regimes. Specifically, we establish the following results:

1. To address *Q1*, we prove in Theorem 1 that Mean-Field SVGD converges for β-upper smooth V (see Assumption 1.1) without requiring any extra structural assumptions on the target distribution, while retaining the best-known $\mathcal{O}(1/T)$ convergence rate, where T denotes the number of iterations. Notably, β-upper smoothness is a strict relaxation of the β-smoothness assumption adopted in many prior works and, importantly, is satisfied by all Gaussian mixture distributions.
2. Furthermore, we propose a new adaptive stepsize strategy in Theorem 2, yielding an improved rate with a better higher-order term than that in Theorem 1.
3. As a byproduct, Theorem 7 establishes a linear convergence rate under the same β-upper smoothness condition when the target distribution additionally satisfies the Stein-log-Sobolev inequality (Duncan et al., 2023; Carrillo et al., 2024).
4. To address *Q2*, we prove in Theorem 3 that Finite-Particle SVGD achieves an $\mathcal{O}(1/\sqrt{N})$ convergence rate, where N denotes the number of particles, matching the best-known bound in the literature, while avoiding the additional growth condition and thereby covering Gaussian targets. Moreover, we highlight that our bound is attainable with $T = N^2$ iterations, which substantially sharpens the known iteration complexity $T = N^{\frac{2-\alpha}{1-2\alpha}}$, since the latter can become arbitrarily large as $\alpha \to 1/2$, where $\alpha \in [0, 1/2)$ parameterizes the growth condition. Thus, our result simultaneously weakens the assumptions and improves the iteration complexity.
5. Moreover, under an even weaker set of assumptions, we show in Theorem 4 that Finite-Particle SVGD converges at a rate of $\mathcal{O}(1/\sqrt[3]{N})$ with $T = N^{\frac{3}{2}}$ iterations. Although this rate is slower than that in Theorem 3, it holds under strictly weaker assumptions and therefore provides a favorable convergence guarantee in a more general setting.

For both regimes, our proofs contain several novel ideas that may be of independent interest and potentially useful for future research. For details, we refer readers to Appendices B, C (especially Subsection C.4), and D (especially Subsection D.3).

1.2 Related Work

The SVGD algorithm was first proposed by Liu & Wang (2016) as a particle-based method for approximate Bayesian inference. Since then, a large body of work has investigated its theoretical properties, including its limiting behavior and convergence guarantees.

SVGD in discrete time. Existing discrete-time analyses of SVGD have mainly proceeded in two directions: the mean-field regime, which studies the evolution of the probability distribution arising in the population limit of the algorithm, and the finite-particle setting, which directly analyzes the interacting particle system implemented in practice. These two regimes capture complementary aspects of SVGD and have led to fruitful theoretical results in the literature.

In the mean-field regime, assuming that V is twice differentiable and β-smooth for some $\beta > 0$ (i.e., $\left\|\nabla^2 V\right\|_{\mathrm{op}} \leq \beta$), Korba et al. (2020) established the first $\mathcal{O}(1/T)$ convergence rate for SVGD. However, their proof critically relies on the assumption that the trajectory has a uniformly bounded KSD and requires a known upper bound on the KSD to determine the stepsize. Later, Salim et al. (2022) obtained the same $\mathcal{O}(1/T)$ rate under the same differentiability and smoothness conditions on V, while replacing the restrictive condition in Korba et al. (2020) with Talagrand's T_1 inequality. Subsequently, Sun et al. (2023) extended the approach of Salim et al. (2022) to twice differentiable (β_0, β_1)-smooth V (i.e., $\left\|\nabla^2 V\right\|_{\mathrm{op}} \leq \beta_0 + \beta_1 \left\|\nabla V\right\|$ for some $\beta_0, \beta_1 \geq 0$) (Zhang et al., 2020b) and to the generalized T_p inequality (Sun et al., 2023, Definition 1) for some $p \geq 1$. However, this extension requires a new, additional assumption on the target distribution (see their Assumption (poly, Q)). Recently, for twice differentiable and β-smooth V, Fujisawa & Futami (2025) proved a sublinear convergence rate for the KL divergence under the log-Sobolev inequality (Gross, 1975), together with other technical requirements on the target distribution. However, as mentioned previously, these results do not apply to general Gaussian mixture distributions because such targets can fail to satisfy the required (generalized) smoothness conditions on V. See Example 1 in Appendix A for a concrete illustration.

In the finite-particle regime, the convergence analysis is known to be more difficult, and the first finite-time convergence rate was derived by Shi & Mackey (2023), showing a convergence rate $\mathcal{O}(1/\sqrt{\log\log N})$ for KSD after $T = \log\log N$ iterations. Their analysis assumes a decay condition on the kernel derivatives (see their Assumption 3), Talagrand's T_1 inequality (equivalently, the target distribution is sub-Gaussian), and two other standard assumptions. More recently, Banerjee et al. (2025) proved a sharper $\mathcal{O}(1/\sqrt{N})$ rate under additional growth and regularity conditions, after $T = N^{\frac{2-\alpha}{1-2\alpha}}$ iterations[1], where $\alpha \in [0, 1/2)$ and the potential V is required to satisfy $\|\nabla V\| \leq A\,(V - \inf V)^{\alpha}$ for some constant $A > 0$ (i.e., the growth condition; see their Assumption 2.(b)). However, this growth condition excludes even the standard Gaussian distribution as a target, since the Gaussian potential requires $\alpha = 1/2$. In another line of work, Liu et al. (2023) focused specifically on centered Gaussian targets and established linear convergence of the empirical covariance matrix, but their analysis is specialized to SVGD with bilinear kernels and requires additional conditions on the initialization.

SVGD in continuous time. Another line of work studies SVGD through its continuous-time limit, obtained by letting the stepsize tend to zero. This perspective, first investigated by Liu (2017), describes the continuous-time behavior of the particle distribution and provides an alternative framework for analyzing the dynamics and convergence behavior of SVGD.

Similar to the discrete-time setting, existing theoretical studies of SVGD in continuous time can also be broadly divided into the mean-field regime, often formulated as a gradient flow, and the finite-particle regime. Since continuous-time SVGD is not the main focus of this paper, and to keep the discussion concise, we do not review these two lines of work separately; instead, we refer readers to

[1]This statement is based on the updated arXiv version of Banerjee et al. (2025). The number of iterations in the conference version contains typos, as noted by the authors of the paper.

the relevant literature (Lu et al., 2019; Chewi et al., 2020; Nüsken & Renger, 2023; Karimi Jaghargh et al., 2023; Duncan et al., 2023; Carrillo et al., 2024; Carrillo & Skrzeczkowski, 2025; Carrillo et al., 2026; Chizat et al., 2026; Balasubramanian et al., 2026; Teolis & de Hoop, 2026).

Variants of SVGD. In addition to the original formulation of SVGD, a broad range of variants have been developed in the literature to adapt the method to different computational settings, target geometries, and model structures. Examples include Amortized SVGD (Feng et al., 2017), Riemannian SVGD (Liu & Zhu, 2018), Gradient-Free SVGD (Han & Liu, 2018), Graphical SVGD (Wang et al., 2018), Message Passing SVGD (Zhuo et al., 2018), Parallel SGLD plus repulsion (Gallego & Insua, 2018), Matrix SVGD (Wang et al., 2019), Random Batch Method SVGD (Li et al., 2020), Stochastic Particle-Optimization Sampling (Zhang et al., 2020a; Hong et al., 2026), Projected SVGD (Chen & Ghattas, 2020), Stochastic SVGD (Gorham et al., 2020), Sliced SVGD (Gong et al., 2021), Annealed SVGD (D'Angelo & Fortuin, 2021), Hybrid-kernel SVGD (D'Angelo et al., 2021; MacDonald et al., 2025), Grassmann SVGD (Liu et al., 2022), Mirrored SVGD (Shi et al., 2022), Normalized SVGD (Xu et al., 2022), Multilevel SVGD (Alsup et al., 2022), Virtual Particle SVGD and Global Batch SVGD (Das & Nagaraj, 2023), Regularized SVGD (He et al., 2025; 2026), Noisy SVGD (Priser et al., 2025), SVGD with adaptive kernel selection (Melcher et al., 2025), and Accelerated SVGD (Stein & Li, 2026).

2 PRELIMINARIES

Notation. $\mathbb{N}$ denotes the set of natural numbers (excluding 0). $[n] \triangleq \{1, \dots, n\}, \forall n \in \mathbb{N}$. $\langle \cdot, \cdot \rangle$ is the Euclidean inner product and $\|\cdot\| \triangleq \sqrt{\langle \cdot, \cdot \rangle}$ is the standard 2-norm. $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_{\mathrm{F}}$ denote the operator norm and the Frobenius norm, respectively. Tr is the trace operator. $\boldsymbol{I}$ denotes the identity map and $\boldsymbol{J}$ denotes the Jacobian matrix. Δ represents the Laplacian operator. Given $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $\nabla_2 f(\boldsymbol{x}, \boldsymbol{y})$ and $\nabla_2^2 f(\boldsymbol{x}, \boldsymbol{y})$ denote the gradient and Hessian w.r.t. $\boldsymbol{y}$, respectively (whenever well-defined). Given $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^d$, $\langle \nabla, \boldsymbol{f} \rangle \triangleq \sum_{i=1}^d \partial_i f_i$ denotes the divergence of $\boldsymbol{f}$. Leb_d is the Lebesgue measure on $\mathbb{R}^d$. Given two probability measures P and Q on a common measurable space $\mathbb{X}$ equipped with σ-algebra $\Sigma_{\mathbb{X}}$, $\mathsf{P} \ll \mathsf{Q}$ means that P is absolutely continuous w.r.t. Q. The Kullback-Leibler (KL) divergence between P and Q is defined as $\mathsf{KL}(\mathsf{P} \| \mathsf{Q}) \triangleq \int \log\left(\frac{\mathrm{d}\mathsf{P}}{\mathrm{d}\mathsf{Q}}\right) \mathrm{d}\mathsf{P}$ when $\mathsf{P} \ll \mathsf{Q}$, and $\mathsf{KL}(\mathsf{P} \| \mathsf{Q}) \triangleq +\infty$ otherwise. For $n \in \mathbb{N}$, $\mathsf{P}^{\otimes n}$ denotes the n-fold product measure of P on $(\mathbb{X}^n, \Sigma_{\mathbb{X}}^{\otimes n})$. Given another measurable space $(\mathbb{Y}, \Sigma_{\mathbb{Y}})$ and a measurable function $f : \mathbb{X} \to \mathbb{Y}$, the pushforward measure $f_{\#}\mathsf{P}(\cdot)$ on $(\mathbb{Y}, \Sigma_{\mathbb{Y}})$ is defined by $f_{\#}\mathsf{P}(A) \triangleq \mathsf{P}(f^{-1}(A)), \forall A \in \Sigma_{\mathbb{Y}}$.

Objective. Our goal is to sample from a probability measure Q on $\mathbb{R}^d$ with density (w.r.t. Leb_d)

$$q(\boldsymbol{x}) = \frac{\exp(-V(\boldsymbol{x}))}{Z} \quad \text{where} \quad V : \mathbb{R}^d \to \mathbb{R} \text{ and } Z \triangleq \int_{\mathbb{R}^d} \exp(-V(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} < +\infty. \tag{1}$$

Remark 1. Unless otherwise specified, Q hereinafter denotes the target distribution defined in (1).

Reproducing kernel Hilbert space. We provide basic background on reproducing kernel Hilbert spaces. For more details on the basic properties and applications, we refer readers to, e.g., Saitoh & Sawano (2016). In what follows, $\mathbb{X}$ denotes an arbitrary set, which can be abstract and not necessarily a subset of $\mathbb{R}^d$.

Definition 1 (Reproducing kernel Hilbert space)**.** A Hilbert space $\mathcal{H} \subseteq \mathbb{R}^{\mathbb{X}}$, equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the induced norm $\|\cdot\|_{\mathcal{H}} \triangleq \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$, is called a Reproducing Kernel Hilbert Space (RKHS) if $\sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} |f(\boldsymbol{x})| < +\infty$ for every $\boldsymbol{x} \in \mathbb{X}$.

By the Moore-Aronszajn theorem (Aronszajn, 1950), every RKHS is associated with a unique symmetric positive definite kernel[2] $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ called the reproducing kernel, such that $K(\boldsymbol{x}, \cdot) \in \mathcal{H}, \forall \boldsymbol{x} \in \mathbb{X}$ and $f(\boldsymbol{x}) = \langle f, K(\boldsymbol{x}, \cdot) \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}, \boldsymbol{x} \in \mathbb{X}$, where the latter identity is known as the reproducing property. Conversely, given a symmetric positive definite kernel $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, there exists a unique RKHS $\mathcal{H} \subseteq \mathbb{R}^{\mathbb{X}}$ whose reproducing kernel is K.

[2]This means that $K(\boldsymbol{x}, \boldsymbol{y}) = K(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}$ and $\sum_{i=1}^n \sum_{j=1}^n a_i a_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$ for every $n \in \mathbb{N}$, $\{a_i \in \mathbb{R}\}_{i=1}^n$, and $\{\boldsymbol{x}_i \in \mathbb{X}\}_{i=1}^n$.

Accordingly, throughout the remainder of the paper, we use the tuple $(\mathcal{H}, K, \mathbb{X})$ to denote an RKHS $\mathcal{H}$ on $\mathbb{X}$ with reproducing kernel K.

Definition 2 (Product RKHS)**.** Given an RKHS $(\mathcal{H}, K, \mathbb{X})$ and $d \in \mathbb{N}$, we define the product RKHS $\mathcal{H}^d \triangleq \left\{\boldsymbol{f} = (f_1, f_2, \ldots, f_d)^\top : f_i \in \mathcal{H}, \forall i \in [d]\right\} \subseteq (\mathbb{R}^d)^{\mathbb{X}}$. The space $\mathcal{H}^d$ is a Hilbert space equipped with inner product $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{H}^d} \triangleq \sum_{i=1}^d \langle f_i, g_i \rangle_{\mathcal{H}}, \forall \boldsymbol{f}, \boldsymbol{g} \in \mathcal{H}^d$.

Langevin Stein operator. We provide the definition of the Langevin Stein operator (Stein, 1972; Gorham & Mackey, 2015). Here, Q need not denote the target distribution introduced earlier.

Definition 3 (Langevin Stein operator)**.** Given a probability measure Q on $\mathbb{R}^d$ with a differentiable density $q > 0$, the Langevin Stein operator $\mathcal{T}_{\mathsf{Q}} : (\mathbb{R}^d)^{(\mathbb{R}^d)} \to \mathbb{R}^{(\mathbb{R}^d)}$ is defined as

$$\mathcal{T}_{\mathsf{Q}}[\boldsymbol{f}] \triangleq \langle \nabla \log q, \boldsymbol{f} \rangle + \langle \nabla, \boldsymbol{f} \rangle .$$

Kernelized Stein discrepancy. We provide the definition of the kernelized Stein discrepancy (Chwialkowski et al., 2016; Liu et al., 2016; Gorham & Mackey, 2017). Here, Q need not denote the target distribution introduced earlier.

Definition 4 (Kernelized Stein discrepancy)**.** Given an RKHS $(\mathcal{H}, K, \mathbb{R}^d)$, let P be a probability measure on $\mathbb{R}^d$ and $\mathcal{T}_{\mathsf{Q}}$ be the Langevin Stein operator associated with a probability measure Q on $\mathbb{R}^d$ with a differentiable density $q > 0$ (cf. Definition 3), then the Kernelized Stein Discrepancy (KSD) between P and Q is defined as

$$\mathsf{KSD}_{\mathcal{H}}(\mathsf{P} \| \mathsf{Q}) \triangleq \sup_{\boldsymbol{f} \in \mathcal{H}^d, \|\boldsymbol{f}\|_{\mathcal{H}^d} \leq 1} \mathbb{E}_{\boldsymbol{x} \sim \mathsf{P}} \left[\mathcal{T}_{\mathsf{Q}}[\boldsymbol{f}](\boldsymbol{x})\right].$$

3 Stein Variational Gradient Descent

In this section, we present our new convergence results for SVGD. In Subsection 3.1, we focus on Mean-Field SVGD, i.e., the infinite-particle limit of the practical algorithm. In Subsection 3.2, we turn our attention to Finite-Particle SVGD, i.e., the algorithm actually implemented in practice.

3.1 Mean-Field SVGD

Algorithm 1 Mean-Field Stein Variational Gradient Descent (Liu & Wang, 2016)

1: **Input:** initial probability measure P_1 on $\mathbb{R}^d$, symmetric positive definite kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, stepsize $\{\eta_t > 0\}_{t=1}^T$
2: **for** $t = 1$ **to** T **do**
3: $\quad \mathsf{P}_{t+1} = \left(\boldsymbol{I} - \eta_t \boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}_t]\right)_{\#} \mathsf{P}_t$ where $\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}](\cdot) = \mathbb{E}_{\boldsymbol{x} \sim \mathsf{P}}[K(\cdot, \boldsymbol{x}) \nabla V(\boldsymbol{x}) - \nabla_2 K(\cdot, \boldsymbol{x})]$
4: **end for**

We present Mean-Field SVGD in Algorithm 1. At each iteration, Mean-Field SVGD directly updates the probability distribution P_t by pushing it forward through the transport map $\boldsymbol{I} - \eta_t \boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}_t]$, rather than evolving a finite collection of particles. This update can be viewed as the population limit of the practical SVGD update and produces a sequence of probability measures that aims to approach the target distribution Q. Although Mean-Field SVGD is generally not directly implementable in practice, it provides a natural first step toward understanding the convergence behavior of SVGD and has therefore been extensively studied.

Next, we list the assumptions on which our analysis for Mean-Field SVGD relies.

Assumption 1. *We make the following assumptions:*

1. *V is differentiable and β-upper smooth for some $\beta > 0$, i.e.,*
$$V(\boldsymbol{x}) \leq V(\boldsymbol{y}) + \langle \nabla V(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{\beta}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$
2. *The initial probability measure P_1 satisfies $\mathsf{KL}(\mathsf{P}_1 \| \mathsf{Q}) < +\infty$.*
3. *The kernel function K satisfies the following conditions: (a) $K \in C^2(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$; (b) $K_0 \triangleq \sup_{\boldsymbol{x} \in \mathbb{R}^d} \|K(\cdot, \boldsymbol{x})\|_{\mathcal{H}} < +\infty$; and (c) $K_1 \triangleq \sup_{\boldsymbol{x} \in \mathbb{R}^d} \|\nabla_2 K(\cdot, \boldsymbol{x})\|_{\mathcal{H}^d} < +\infty$.*

Discussion of Assumption 1. As far as we are aware, Assumption 1 captures a minimal set of conditions for the convergence analysis of Mean-Field SVGD. As discussed in Subsection 1.2, prior convergence analyses of Mean-Field SVGD typically impose additional conditions on the algorithm trajectory or structural assumptions on the target distribution, such as Talagrand's T_1 inequality. In contrast, our analysis proceeds under Assumption 1 alone, without requiring such additional transport or functional inequalities. Consequently, our result applies under a weaker and more broadly applicable set of assumptions, thereby covering a broader range of settings.

In particular, Assumption 1.1 requires only that the potential V be β-upper smooth instead of being β-smooth (Korba et al., 2020; Salim et al., 2022; Fujisawa & Futami, 2025) or (β_0, β_1)-smooth (Sun et al., 2023). Since β-smoothness implies β-upper smoothness but not vice versa, Assumption 1.1 is a strict relaxation of the smoothness assumption in these prior works. Although β-upper smoothness is not directly comparable to (β_0, β_1)-smoothness, we highlight an important distinction that Gaussian mixture distributions are always β-upper smooth for some $\beta > 0$, but may fail to satisfy any (β_0, β_1)-smoothness condition (and hence any β-smoothness condition) required in existing analyses. See Example 1 in Appendix A for details.

Lastly, we note that Assumptions 1.2 and 1.3 are standard in the existing literature. For example, the widely used Radial Basis Function (RBF) kernel, also known as the Gaussian kernel, satisfies Assumption 1.3.

Convergence rates. Our first new result for Mean-Field SVGD is stated in Theorem 1.

Theorem 1 (Simplified version of Theorem 5)**.** *Under Assumption 1, let the stepsize* $\{\eta_t\}_{t=1}^T$ *be*

$$\eta_t = \min\left\{\frac{1}{K_0^2\beta + 2K_1^2}, \frac{1}{2K_1\left[\sqrt{2(K_0^2\beta + 2K_1^2)\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})} + 4K_1\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})\right]}\right\}, \forall t \in [T],$$

then Mean-Field SVGD (Algorithm 1) guarantees that

$$\mathsf{KSD}_{\mathcal{H}}^2(\bar{\mathsf{P}}_T\|\mathsf{Q}) \lesssim \frac{K_1^2\mathsf{KL}^2(\mathsf{P}_1\|\mathsf{Q}) + \left(K_0^2\beta + K_1^2\right)\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{T} \quad \textit{where} \quad \bar{\mathsf{P}}_T \triangleq \frac{\sum_{t=1}^T \mathsf{P}_t}{T}.$$

Theorem 1 establishes an $\mathcal{O}(1/T)$ convergence rate in terms of the standard KSD criterion, matching the rates obtained in prior work (Korba et al., 2020; Salim et al., 2022; Sun et al., 2023), while requiring only the weaker Assumption 1 and no additional structural assumptions on Q. To the best of our knowledge, this is the first result that applies to arbitrary Gaussian mixture targets.

Next, Theorem 2 provides another new result, showing that an adaptive stepsize can further improve the dependence on $\mathsf{KL}^2(\mathsf{P}_1\|\mathsf{Q})$, which may be large under poor initialization.

Theorem 2 (Simplified version of Theorem 6)**.** *Under Assumption 1, let the stepsize* $\{\eta_t\}_{t=1}^T$ *be*

$$\eta_t = \frac{\gamma_t}{\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d}}, \gamma_t = \min\left\{\frac{1}{2K_1}, \sqrt{\frac{2\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{(K_0^2\beta + 2K_1^2)T}}\right\}, \forall t \in [T],$$

then Mean-Field SVGD (Algorithm 1) guarantees that

$$\mathsf{KSD}_{\mathcal{H}}^2(\bar{\mathsf{P}}_T\|\mathsf{Q}) \lesssim \frac{K_1^2\mathsf{KL}^2(\mathsf{P}_1\|\mathsf{Q})}{T^2} + \frac{\left(K_0^2\beta + K_1^2\right)\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{T} \quad \textit{where} \quad \bar{\mathsf{P}}_T \triangleq \frac{\sum_{t=1}^T \mathsf{P}_t}{T}.$$

Remark 2. The time horizon T in $\{\gamma_t\}_{t=1}^T$ can be replaced by t if T is unknown in advance, at the cost of changing the second $\mathcal{O}(1/T)$ term to $\mathcal{O}(\log^2 T/T)$.

Compared to Theorem 1, the rate in Theorem 2 has a smaller higher-order term, which is due to the adaptive normalization factor $\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d}$. The stepsize design is inspired by Normalized Gradient Descent (Nesterov, 1984). Specifically, the term quadratic in the initial KL divergence improves from $\mathcal{O}(K_1^2\mathsf{KL}^2(\mathsf{P}_1\|\mathsf{Q})/T)$ to $\mathcal{O}(K_1^2\mathsf{KL}^2(\mathsf{P}_1\|\mathsf{Q})/T^2)$, while the leading $\mathcal{O}(1/T)$ term remains unchanged. Thus, the adaptive stepsize retains the best-known $\mathcal{O}(1/T)$ dependence on the number of iterations while becoming less sensitive to a large initial KL divergence $\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})$.

Lastly, one may wonder whether convergence can be established in KL divergence rather than KSD, thereby providing a stronger notion of convergence. In Theorem 7, under the Stein-log-Sobolev inequality (Duncan et al., 2023; Carrillo et al., 2024), we show that Mean-Field SVGD converges linearly in KL divergence. For details, we refer interested readers to Appendix C.

Proofs. Due to limited space, we defer all proofs to Appendix C. In particular, in Subsection C.4, we provide a detailed discussion of the proof roadmap and the main novel ideas in our analysis.

3.2 FINITE-PARTICLE SVGD

Algorithm 2 Finite-Particle Stein Variational Gradient Descent (Liu & Wang, 2016)

1: **Input:** initial points $\{\boldsymbol{x}_i(1) \in \mathbb{R}^d\}_{i=1}^N$, symmetric positive definite kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, stepsize $\{\eta_t > 0\}_{t=1}^T$
2: **for** $t = 1$ **to** T **do**
3: $\quad \boldsymbol{x}_i(t+1) = \boldsymbol{x}_i(t) - \frac{\eta_t}{N} \sum_{j=1}^N K(\boldsymbol{x}_i(t), \boldsymbol{x}_j(t)) \nabla V(\boldsymbol{x}_j(t)) - \nabla_2 K(\boldsymbol{x}_i(t), \boldsymbol{x}_j(t)), \forall i \in [N]$
4: **end for**

We present Finite-Particle SVGD in Algorithm 2. The algorithm maintains N particles and, at each iteration, updates the particles simultaneously through kernel-based interactions that guide them toward the target distribution. Equivalently, the algorithm evolves the empirical measure $\hat{\mathsf{P}}_t \triangleq \frac{1}{N} \sum_{i=1}^N \delta_{\boldsymbol{x}_i(t)}$ according to $\boldsymbol{x}_i(t+1) = \boldsymbol{x}_i(t) - \eta_t \boldsymbol{g}_{\mathsf{Q}}[\hat{\mathsf{P}}_t](\boldsymbol{x}_i(t)), \forall i \in [N]$, which serves as a finite-particle approximation to the target distribution Q. Unlike Mean-Field SVGD, Finite-Particle SVGD is directly implementable in practice and corresponds to the algorithm actually used in applications. However, despite its practical relevance, fewer theoretical guarantees are known due to the difficulty of finite-particle analysis.

Next, we list the assumptions on which our analysis for Finite-Particle SVGD relies.

Assumption 2. *We make the following assumptions:*

1. *V is twice differentiable, β-smooth for some $\beta > 0$, and lower bounded, i.e.,*

$$\sup_{\boldsymbol{x} \in \mathbb{R}^d} \left\| \nabla^2 V(\boldsymbol{x}) \right\|_{\mathrm{op}} \leq \beta \quad \textit{and} \quad V_\star \triangleq \inf_{\boldsymbol{x} \in \mathbb{R}^d} V(\boldsymbol{x}) > -\infty.$$

2. *The initial point $\vec{\boldsymbol{x}}(1) \triangleq (\boldsymbol{x}_1^\top(1), \ldots, \boldsymbol{x}_N^\top(1))^\top \in (\mathbb{R}^d)^N$ is random and satisfies the following two conditions:*
   (a) *$\vec{\boldsymbol{x}}(1) \in \mathrm{S}_1 \triangleq \left\{ (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_N^\top)^\top \in (\mathbb{R}^d)^N : \frac{1}{N} \sum_{i=1}^N V(\boldsymbol{x}_i) < V^\star \right\}$ for some constant $V^\star$;*
   (b) *$\mathsf{KL}(\mathrm{Law}(\vec{\boldsymbol{x}}(1)) \| \mathsf{Q}^{\otimes N}) \leq C_{\mathsf{KL}} dN$ for some constant $C_{\mathsf{KL}} > 0$.*
3. *The kernel function K satisfies the following conditions: (a) $K \in C^2(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$; (b) $K_0 \triangleq \sup_{\boldsymbol{x} \in \mathbb{R}^d} \|K(\cdot, \boldsymbol{x})\|_{\mathcal{H}} < +\infty$; (c) $K_1 \triangleq \sup_{\boldsymbol{x} \in \mathbb{R}^d} \|\nabla_2 K(\cdot, \boldsymbol{x})\|_{\mathcal{H}^d} < +\infty$; and (d) $K_2 \triangleq \sup_{\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d} \left\| \nabla_2^2 K(\boldsymbol{x}, \boldsymbol{y}) \right\|_{\mathrm{F}} < +\infty$.*
4. *(**Optional**) There exists a constant $C_\star > 0$ such that*

$$\sup_{\boldsymbol{x} \in \mathbb{R}^d} \langle \nabla_2 K(\boldsymbol{x}, \boldsymbol{x}), \nabla V(\boldsymbol{x}) \rangle + K(\boldsymbol{x}, \boldsymbol{x}) \Delta V(\boldsymbol{x}) - \Delta_2 K(\boldsymbol{x}, \boldsymbol{x}) \leq C_\star d,$$

   where Δ_2 denotes the Laplacian operator w.r.t. the second argument.

Remark 3. In fact, the specific form of (1), together with β-smoothness (or even the weaker β-upper smoothness), already implies that V is uniformly bounded from below (see Lemma 11 in Appendix B), and hence $V_\star$ is well-defined without explicitly assuming its existence. However, for simplicity, we still state lower boundedness explicitly as an assumption, but do not regard it as an additional requirement when comparing with the existing literature.

Discussion of Assumption 2. Compared with existing finite-particle analyses, we drop several restrictive conditions imposed in prior work. In particular, we do not require the decay condition on the kernel derivatives (Shi & Mackey, 2023, Assumption 3) or Talagrand's T_1 inequality (Shi & Mackey, 2023, Assumption 4), nor the additional growth condition imposed by Banerjee et al. (2025, Assumption 2.(b)), which, as discussed in Subsection 1.2, excludes even Gaussian distributions as possible targets. As a consequence, our theory applies to a broader range of settings.

First, Assumption 2.1 is a standard smoothness condition and is also adopted in Shi & Mackey (2023, Assumption 1)[3] and Banerjee et al. (2025, Assumption 2.(c)). We note, however, that it is stronger than the β-upper smoothness condition in Assumption 1.1. Extending the finite-particle analysis to the weaker upper-smoothness setting remains open and appears technically challenging, and we leave this question for future work.

Next, Assumption 2.2 is inspired by Banerjee et al. (2025, Theorem 3 and Assumption 2.(e)). In fact, condition (b) in Assumption 2.2 is stronger than what is required by our analysis, and it suffices to assume only $\mathsf{KL}(\mathrm{Law}(\vec{\boldsymbol{x}}(1))\|\mathsf{Q}^{\otimes N}) < +\infty$. The current condition is imposed only to align with Banerjee et al. (2025) for a fair comparison.

Moreover, Assumption 2.3 consists of standard regularity conditions on the kernel K and is similar to the corresponding conditions used in Shi & Mackey (2023, Assumption 2)[4] and Banerjee et al. (2025, Assumption 2.(a)).

Finally, Assumption 2.4 was considered in Banerjee et al. (2025, Assumption 2.(d)) and holds in many settings (including the commonly used RBF kernel), as noted by Banerjee et al. (2025, Remark 1). However, we observe that this assumption is actually optional: removing it yields a slower convergence rate but allows our analysis to cover a broader range of situations. For details, we refer readers to Theorems 3 and 4 below.

Convergence rates. Before stating our new results, we recall that $\hat{\mathsf{P}}_t = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{x}_i(t)}$ denotes the empirical distribution of the particles at time $t \in [T]$. Moreover, in later comparisons with the existing literature, we report convergence rates in terms of KSD^2. Accordingly, the bounds of Shi & Mackey (2023) and Banerjee et al. (2025) correspond to $\mathcal{O}(1/\log\log N)$ and $\mathcal{O}(1/N)$, respectively.

Remark 4. When comparing our results with those of Shi & Mackey (2023) and Banerjee et al. (2025), we omit the dependence on the dimension d, even though ours is more favorable. This is because the regularity conditions on K differ slightly across the three works (although they are essentially equivalent up to polynomial factors in d), making a direct comparison potentially unfair.

Theorem 3 (Simplified version of Theorem 8)**.** *Under Assumption 2 (with Assumption 2.4), let the stepsize* $\{\eta_t\}_{t=1}^T$ *be*

$$\eta_t = \min\left\{\frac{1}{4\sqrt{5K_0^4\beta^2 + 5K_1^4 + 5K_2^2 + 20K_0^2K_1^2\beta\,(V^\star - V_\star)\,N}}, \frac{1}{2\sqrt[3]{20K_0^2K_1^4\beta NT}}, \right.$$
$$\left. \sqrt{\frac{C_{\mathsf{KL}}dN}{(dK_0^4\beta^2 + K_2^2 + 2K_0^2K_1^2\beta(V^\star - V_\star))\,T}}, \sqrt[3]{\frac{C_{\mathsf{KL}}dN}{K_0^2K_1^4\beta T^2}}\right\}, \forall t \in [T]\,,$$

then Finite-Particle SVGD (Algorithm 2) guarantees that

$$\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}\left(\frac{\sum_{t=1}^T \hat{\mathsf{P}}_t}{T}\middle\|\mathsf{Q}\right)\right] \lesssim \frac{C_{\mathrm{FP},1}}{N^2} + \frac{C_{\mathrm{FP},2}}{N^{\frac{3}{2}}} + \frac{C_{\mathrm{FP},3}}{N} \quad \textit{when} \quad T = N^2,$$

where

$$C_{\mathrm{FP},1} \triangleq C_{\mathsf{KL}}d(K_0^2\beta + K_1^2 + K_2),\ C_{\mathrm{FP},3} \triangleq C_{\mathsf{KL}}d\sqrt[3]{K_0^2K_1^4\beta} + \sqrt[3]{C_{\mathsf{KL}}^2d^2K_0^2K_1^4\beta} + C_\star d,$$
$$C_{\mathrm{FP},2} \triangleq C_{\mathsf{KL}}dK_0K_1\sqrt{\beta(V^\star - V_\star)} + \sqrt{C_{\mathsf{KL}}d\,(dK_0^4\beta^2 + K_2^2 + K_0^2K_1^2\beta(V^\star - V_\star))}.$$

To better understand Theorem 3, let us compare it with Banerjee et al. (2025, Theorem 3). Our bound yields the same $\mathcal{O}(1/N)$ dependence on N in terms of squared KSD, both substantially better than the $\mathcal{O}(1/\log\log N)$ bound of Shi & Mackey (2023, Corollary 2). However, Banerjee et al. (2025) require the additional growth condition, i.e., $\|\nabla V\| \le A(V - \inf V)^\alpha$ for some $A > 0$ and $\alpha \in [0, 1/2)$, and their $\mathcal{O}(1/N)$ rate is attained after $T = N^{\frac{2-\alpha}{1-2\alpha}}$ iterations. In contrast, Theorem 3 does not require this growth condition, and the same rate is attainable with only $T = N^2$ iterations. Thus, our result matches the best-known rate while achieving a sharper iteration

[3] We clarify that V is only required to be differentiable in Shi & Mackey (2023).

[4] In fact, their assumption is stronger due to an additional condition on the fourth-order derivatives of K.

complexity under less restrictive assumptions. Moreover, removing the growth condition allows our result to cover a broader class of targets, including Gaussian distributions, which are excluded by the growth condition of Banerjee et al. (2025).

We note that $T = N^2$ is chosen to balance different terms. In fact, our analysis applies to arbitrary $T \in \mathbb{N}$. See Theorem 8 for details.

Theorem 4 (Simplified version of Theorem 9)**.** *Under Assumption 2 (without Assumption 2.4), let the stepsize* $\{\eta_t\}_{t=1}^T$ *be*

$$\eta_t = \min\left\{\frac{1}{4\sqrt{5K_0^4\beta^2 + 5K_1^4 + 5K_2^2 + 20NK_0^2K_1^2\beta\,(V^\star - V_\star)}}, \frac{1}{2\sqrt[3]{20NK_0^2K_1^4\beta T}},\right.$$
$$\left.\sqrt{\frac{C_{\mathsf{KL}}dN}{(dK_0^4\beta^2 + K_2^2 + 2K_0^2K_1^2\beta(V^\star - V_\star))\,T}}, \sqrt[3]{\frac{C_{\mathsf{KL}}dN}{K_0^2K_1^4\beta T^2}}, \sqrt[3]{\frac{C_{\mathsf{KL}}^2d^2N^2}{K_0^2K_1^4\beta T^3}}\right\}, \forall t \in [T],$$

then Finite-Particle SVGD (Algorithm 2) guarantees that

$$\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}_{\mathcal{H}}^2\left(\frac{\sum_{t=1}^T \hat{\mathsf{P}}_t}{T}\middle\|\mathsf{Q}\right)\right] \lesssim \frac{C_{\mathrm{FP},1}}{N^{\frac{3}{2}}} + \frac{C_{\mathrm{FP},4}}{N^{\frac{5}{4}}} + \frac{C_{\mathrm{FP},5}}{N} + \frac{C_{\mathrm{FP},6}}{N^{\frac{5}{6}}} + \frac{C_{\mathrm{FP},7}}{N^{\frac{2}{3}}} \quad \textit{when} \quad T = N^{\frac{3}{2}},$$

where

$$C_{\mathrm{FP},4} \triangleq \sqrt{C_{\mathsf{KL}}d\,(dK_0^4\beta^2 + K_2^2 + K_0^2K_1^2\beta(V^\star - V_\star))},$$
$$C_{\mathrm{FP},5} \triangleq K_0K_1(1 + C_{\mathsf{KL}}d)\sqrt{\beta(V^\star - V_\star)} + dK_0^2\beta + \sqrt{d}K_2,$$
$$C_{\mathrm{FP},6} \triangleq \sqrt[3]{C_{\mathsf{KL}}^2d^2K_0^2K_1^4\beta},\ C_{\mathrm{FP},7} \triangleq \sqrt[3]{C_{\mathsf{KL}}dK_0^2K_1^4\beta} + C_{\mathsf{KL}}d\sqrt[3]{K_0^2K_1^4\beta}.$$

In Theorem 4, we further remove Assumption 2.4, which may fail for certain kernel choices. As a result, the rate slows to $\mathcal{O}(1/N^{\frac{2}{3}})$, but can be achieved with $T = N^{\frac{3}{2}}$ iterations. As far as we are aware, this is the first finite-particle convergence result for SVGD under such a minimal set of assumptions, considerably weaker than those in Shi & Mackey (2023) and Banerjee et al. (2025). However, whether, and if so, how this rate can be improved to match that of Theorem 3 remains unclear, and we leave this question for future work.

Similar to Theorem 3, the number of iterations $T = N^{\frac{3}{2}}$ is made to balance different terms and the full version, Theorem 9, applies to all $T \in \mathbb{N}$.

Proofs. Due to space limitations, we place the complete proofs in Appendix D. Subsection D.3 further summarizes the overall proof strategy and highlights the key new ingredients in our analysis.

4 Conclusion

In this work, we develop an improved convergence analysis of SVGD in both the mean-field and finite-particle regimes under weaker assumptions than those commonly imposed in the literature. Using squared KSD as the convergence metric, for Mean-Field SVGD, we retain the best-known $\mathcal{O}(1/T)$ rate while allowing β-upper smooth targets, including Gaussian mixtures. For Finite-Particle SVGD, we obtain an $\mathcal{O}(1/N)$ rate with only $T = N^2$ iterations, yielding a sharper iteration complexity than the best-known one under weaker conditions, and an $\mathcal{O}(1/N^{\frac{2}{3}})$ rate with only $T = N^{\frac{3}{2}}$ iterations under the weakest assumptions currently available.

Several questions remain open, and we highlight a few of them here. For the mean-field analysis, one can similarly extend the notion of β-upper smoothness to (β_0, β_1)-upper smoothness, and it would be interesting to study convergence under this even weaker condition, which could cover a broader range of settings. Another interesting but challenging question is whether Theorem 2 is tight. Moreover, extending the finite-particle analysis from β-smooth potentials to the weaker β-upper smooth setting remains technically challenging. It is also unclear whether the $\mathcal{O}(1/N^{\frac{2}{3}})$ rate obtained without Assumption 2.4 can be improved to match the $\mathcal{O}(1/N)$ rate. More broadly, it would be interesting to determine the weakest conditions under which sharp finite-particle convergence guarantees can be established and whether the proof techniques developed here can be extended to other practical variants of SVGD.

AI USE STATEMENT

In this work, we used generative AI tools for editing and polishing the text. We have not used generative AI tools to generate mathematical proofs. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All missing proofs are deferred to the appendix.

REFERENCES

Terrence Alsup, Luca Venturi, and Benjamin Peherstorfer. Multilevel stein variational gradient descent with applications to bayesian inverse problems. In Joan Bruna, Jan Hesthaven, and Lenka Zdeborova (eds.), *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pp. 93–117. PMLR, 16–19 Aug 2022. URL https://proceedings.mlr.press/v145/alsup22a.html.

Jason M. Altschuler and Sinho Chewi. Shifted composition ii: Shift harnack inequalities and curvature upper bounds. *IEEE Transactions on Information Theory*, 72(1):42–57, Jan 2026. ISSN 1557-9654. doi: 10.1109/TIT.2025.3631770.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. ISSN 00029947, 10886850. URL http://www.jstor.org/stable/1990404.

Krishnakumar Balasubramanian, Sayan Banerjee, and Anna Korba. Uniform-in-time propagation-of-chaos for stein variational gradient descent. *arXiv preprint arXiv:2607.00149*, 2026.

Sayan Banerjee, Krishna Balasubramanian, and PROMIT GHOSAL. Improved finite-particle convergence rates for stein variational gradient descent. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 67985–68006, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a9f61e75161ceb9350890677b3aa0f0c-Paper-Conference.pdf.

Lucas Barcelos, Alexander Lambert, Rafael Oliveira, Paulo Borges, Byron Boots, and Fabio Ramos. Dual Online Stein Variational Inference for Control and Dynamics. In *Proceedings of Robotics: Science and Systems*, Virtual, July 2021. doi: 10.15607/RSS.2021.XVII.068.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

José Carrillo and Jakub Skrzeczkowski. Convergence and stability results for the particle system in the stein gradient descent method. *Mathematics of Computation*, 94(354):1793–1814, 2025.

José A Carrillo, Jakub Skrzeczkowski, and Jethro Warnett. The stein-log-sobolev inequality and the exponential rate of convergence for the continuous stein variational gradient descent method. *arXiv preprint arXiv:2412.10295*, 2024.

José A Carrillo, Jakub Skrzeczkowski, and Jethro Warnett. Stein variational gradient descent dynamics for highly concentrated kernels. *arXiv preprint arXiv:2605.03627*, 2026.

Peng Chen and Omar Ghattas. Projected stein variational gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1947–1958. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/14faf969228fc18fcd4fcf59437b0c97-Paper.pdf.

Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, and Philippe Rigollet. Svgd as a kernelized wasserstein gradient flow of the chi-squared divergence. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2098–2109. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/16f8e136ee5693823268874e58795216-Paper.pdf.

Lénaïc Chizat, Maria Colombo, Roberto Colombo, and Xavier Fernández-Real. Quantitative local convergence of mean-field stein variational gradient flow. *arXiv preprint arXiv:2605.09456*, 2026.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2606–2615, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/chwialkowski16.html.

Miguel Corrales, Sean Berti, Bertrand Denel, Paul Williamson, Mattia Aleardi, and Matteo Ravasi. Annealed stein variational gradient descent for improved uncertainty estimation in full-waveform inversion. *Geophysical Journal International*, 241(2):1088–1113, 05 2025. ISSN 1365-246X. doi: 10.1093/gji/ggaf096. URL https://doi.org/10.1093/gji/ggaf096.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In Satyen Kale and Ohad Shamir (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 678–689. PMLR, 07–10 Jul 2017. URL https://proceedings.mlr.press/v65/dalalyan17a.html.

Francesco D'Angelo and Vincent Fortuin. Annealed stein variational gradient descent. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021. URL https://openreview.net/forum?id=pw2v8HFJIYg.

Francesco D'Angelo, Vincent Fortuin, and Florian Wenzel. On stein variational neural network ensembles. *arXiv preprint arXiv:2106.10760*, 2021.

Aniket Das and Dheeraj Nagaraj. Provably fast finite particle variants of svgd via virtual particle stochastic approximation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49748–49760. Curran Associates, Inc., 2023. doi: 10.52202/075280-2165. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9bf1962c5b65a243ee243bb03ff2c506-Paper-Conference.pdf.

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983. doi: https://doi.org/10.1002/cpa.3160360204. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160360204.

Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39, 2023. URL http://jmlr.org/papers/v24/20-602.html.

Lawrence C Evans. *Measure theory and fine properties of functions*. Chapman and Hall/CRC, 2025.

Yihao Feng, Dilin Wang, and Qiang Liu. Learning to draw samples with amortized stein variational gradient descent. *arXiv preprint arXiv:1707.06626*, 2017.

Masahiro Fujisawa and Futoshi Futami. On the convergence of SVGD in KL divergence via approximate gradient flow. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=AG1zXt5aoA.

Victor Gallego and David Rios Insua. Stochastic gradient mcmc with repulsive forces. *arXiv preprint arXiv:1812.00071*, 2018.

Wenbo Gong, Yingzhen Li, and José Miguel Hernández-Lobato. Sliced kernelized stein discrepancy. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=t0TaKv0Gx6Z.

Jackson Gorham and Lester Mackey. Measuring sample quality with stein's method. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/698d51a19d8a121ce581499d7b701668-Paper.pdf.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1292–1301. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/gorham17a.html.

Jackson Gorham, Anant Raj, and Lester Mackey. Stochastic stein discrepancies. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17931–17942. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d03a857a23b5285736c4d55e0bb067c8-Paper.pdf.

Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4): 1061–1083, 1975. ISSN 00029327, 10806377. URL http://www.jstor.org/stable/2373688.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1352–1361. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/haarnoja17a.html.

Jun Han and Qiang Liu. Stein variational gradient descent without gradient. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1900–1908. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/han18b.html.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2334940.

Ye He, Krishnakumar Balasubramanian, Bharath K Sriperumbudur, and Jianfeng Lu. Regularized stein variational gradient flow. *Foundations of Computational Mathematics*, 25(4):1199–1257, 2025.

Ye He, Krishnakumar Balasubramanian, Sayan Banerjee, and Promit Ghosal. Finite-particle rates for regularized stein variational gradient descent. *arXiv preprint arXiv:2602.05172*, 2026.

Wei Hong, Wei Liu, and Luhan Yang. Mean-field stochastic pdes: Well-posedness and quantitative dimension-free propagation of chaos. *arXiv preprint arXiv:2607.17195*, 2026.

Junxiong Jia, Peijun Li, and Deyu Meng. Stein variational gradient descent on infinite-dimensional space and applications to statistical inverse problems. *SIAM Journal on Numerical Analysis*, 60(4):2225–2252, 2022. doi: 10.1137/21M1440773. URL https://doi.org/10.1137/21M1440773.

Mohammad Reza Karimi Jaghargh, Ya-Ping Hsieh, and Andreas Krause. Stochastic approximation algorithms for systems of interacting particles. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 55826–55847. Curran Associates, Inc., 2023. doi: 10.52202/075280-2436. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/aebbbfa9680eafefd43a0edc85c101f9-Paper-Conference.pdf.

Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for stein variational gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4672–4682. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3202111cf90e7c816a472aaceb72b0df-Paper.pdf.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214/aoms/1177729694.

Alexander Lambert, Fabio Ramos, Byron Boots, Dieter Fox, and Adam Fishman. Stein variational model predictive control. In Jens Kober, Fabio Ramos, and Claire Tomlin (eds.), *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 1278–1297. PMLR, 16–18 Nov 2021. URL https://proceedings.mlr.press/v155/lambert21a.html.

Yewon Lee, Andrew Z. Li, Philip Huang, Eric Heiden, Krishna Murthy Jatavallabhula, Fabian Damken, Kevin Smith, Derek Nowrouzezahrai, Fabio Ramos, and Florian Shkurti. Stamp: Differentiable task and motion planning via stein variational gradient descent. *IEEE Robotics and Automation Letters*, 10(6):6007–6014, June 2025. ISSN 2377-3766. doi: 10.1109/LRA.2025.3561575.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Logsmooth gradient concentration and tighter runtimes for metropolized hamiltonian monte carlo. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2565–2597. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/lee20b.html.

Lei Li, Yingzhou Li, Jian-Guo Liu, Zibu Liu, and Jianfeng Lu. A stochastic version of Stein variational gradient descent for efficient sampling. *Communications in Applied Mathematics and Computational Science*, 15(1):37 – 63, 2020. doi: 10.2140/camcos.2020.15.37. URL https://doi.org/10.2140/camcos.2020.15.37.

Chang Liu and Jun Zhu. Riemannian stein variational gradient descent for bayesian inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Qiang Liu. Stein variational gradient descent as gradient flow. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/17ed8abedc255908be746d245e50263a-Paper.pdf.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 276–284, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/liub16.html.

Tianle Liu, Promit Ghosal, Krishnakumar Balasubramanian, and Natesh Pillai. Towards understanding the dynamics of gaussian-stein variational gradient descent. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 61234–61291. Curran Associates, Inc., 2023. doi: 10.52202/075280-2676. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c0ae487420ebc8d0ed7c541b4e3f09d4-Paper-Conference.pdf.

Xing Liu, Harrison Zhu, Jean-Francois Ton, George Wynne, and Andrew Duncan. Grassmann stein variational gradient descent. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2002–2021. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/liu22a.html.

Yang Liu, Prajit Ramachandran, Qiang Liu, and Jian Peng. Stein variational policy gradient. *arXiv preprint arXiv:1704.02399*, 2017.

Jianfeng Lu, Yulong Lu, and James Nolen. Scaling limit of the stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671, 2019. doi: 10.1137/18M1187611. URL https://doi.org/10.1137/18M1187611.

Anson MacDonald, Scott A Sisson, and Sahani Pathiraja. Convergence aspects of hybrid kernel SVGD. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=JZkbMSQDmD.

Moritz Melcher, Simon Weissmann, Ashia C Wilson, and Jakob Zech. Adaptive kernel selection for stein variational gradient descent. *arXiv preprint arXiv:2510.02067*, 2025.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 06 1953. ISSN 0021-9606. doi: 10.1063/1.1699114. URL https://doi.org/10.1063/1.1699114.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013.

Yurii E Nesterov. Minimization methods for nonsmooth convex and quasiconvex functions. *Matekon*, 29(3):519–531, 1984.

Nikolas Nüsken and D. R. Michiel Renger. Stein variational gradient descent: Many-particle and long-time asymptotics, 2023. URL https://www.aimsciences.org/article/id/63b40317aa1db67769a1644a.

Victor Priser, PASCAL BIANCHI, and Adil Salim. Long-time asymptotics of noisy svgd outside the population limit. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 25933–25964, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/410bbba8388369d8bb5875544d1d4428-Paper-Conference.pdf.

Yuchen Pu, Zhe Gan, Ricardo Henao, Chunyuan Li, Shaobo Han, and Lawrence Carin. Vae learning via stein variational gradient descent. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/443dec3062d0286986e21dc0631734c9-Paper.pdf.

Christian Robert and George Casella. A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science*, 26(1):102 – 115, 2011. doi: 10.1214/10-STS351. URL https://doi.org/10.1214/10-STS351.

Saburou Saitoh and Yoshihiro Sawano. *Theory of reproducing kernels and applications*, volume 44. Springer, 2016.

Adil Salim, Lukang Sun, and Peter Richtarik. A convergence theory for SVGD in the population limit under talagrand's inequality t1. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19139–19152. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/salim22a.html.

Jiaxin Shi and Lester Mackey. A finite-particle convergence rate for stein variational gradient descent. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 26831–26844. Curran Associates, Inc., 2023. doi: 10.52202/075280-1166. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/54e5d7af6250ccab796ad7fe75663ba5-Paper-Conference.pdf.

Jiaxin Shi, Chang Liu, and Lester Mackey. Sampling with mirrored stein operators. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=eMudnJsb1T5.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, volume 6, pp. 583–603. University of California Press, 1972.

Viktor Stein and Wuchen Li. Accelerated stein variational gradient flow. In Frank Nielsen and Frédéric Barbaresco (eds.), *Geometric Science of Information*, pp. 80–90, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-03921-7.

Lukang Sun, Avetik Karagulyan, and Peter Richtarik. Convergence of stein variational gradient descent under a weaker smoothness condition. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 3693–3717. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/sun23d.html.

Trevor Teolis and Maarten V de Hoop. Riesz-kernel stein variational gradient descent: Renormalized entropy and long-time particle limits. *arXiv preprint arXiv:2607.14527*, 2026.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/65a99bb7a3115fdede20da98b08a370f-Paper.pdf.

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

Dilin Wang, Zhe Zeng, and Qiang Liu. Stein variational message passing for continuous graphical models. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5219–5227. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/wang18l.html.

Dilin Wang, Ziyang Tang, Chandrajit Bajaj, and Qiang Liu. Stein variational gradient descent with matrix-valued kernels. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5dcd0ddd3d918c70d380d32bce4e733a-Paper.pdf.

Lantian Xu, Anna Korba, and Dejan Slepcev. Accurate quantization of measures via interacting particle-based optimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24576–24595. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/xu22d.html.

Jianyi Zhang, Ruiyi Zhang, Lawrence Carin, and Changyou Chen. Stochastic particle-optimization sampling and the non-asymptotic convergence theory. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1877–1887. PMLR, 26–28 Aug 2020a. URL https://proceedings.mlr.press/v108/zhang20d.html.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=BJgnXpVYwS.

Xin Zhang and Andrew Curtis. Seismic tomography using variational inference methods. *Journal of Geophysical Research: Solid Earth*, 125(4):e2019JB018589, 2020. doi: https://doi.org/10.1029/2019JB018589. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JB018589. e2019JB018589 10.1029/2019JB018589.

Jingwei Zhuo, Chang Liu, Jiaxin Shi, Jun Zhu, Ning Chen, and Bo Zhang. Message passing stein variational gradient descent. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 6018–6027. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/zhuo18a.html.

A GAUSSIAN MIXTURES ARE UPPER SMOOTH BUT NEED NOT BE SMOOTH

Example 1. Let $\mathsf{Q} \triangleq \sum_{i=1}^n w_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ be a Gaussian mixture distribution with $2 \leq n \in \mathbb{N}$ components, where $w_i > 0$, $\boldsymbol{\mu}_i \in \mathbb{R}^d$, and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{d\times d}$ is symmetric positive definite for every $i \in [n]$, with $\sum_{i=1}^n w_i = 1$. Moreover, we denote by $V \triangleq -\log q$, where q is the density of Q.

- V is $\left[\max_{i\in[n]} \lambda_{\max}(\boldsymbol{\Sigma}_i^{-1})\right]$-upper smooth, where $\lambda_{\max}(\boldsymbol{A})$ denotes the largest eigenvalue of $\boldsymbol{A}$.
- V is not (β_0, β_1)-smooth for any $\beta_0, \beta_1 \geq 0$ if the following conditions hold:
  1. There exists $i_\star \neq j_\star$ in $[n]$ such that $\boldsymbol{\mu}_{i_\star} = \boldsymbol{\mu}_{j_\star} = \boldsymbol{\mu}_\star$.
  2. There exists $\boldsymbol{y}_\star \in \mathbb{R}^d$ such that $\boldsymbol{\Sigma}_{i_\star}^{-1}\boldsymbol{y}_\star \neq \boldsymbol{\Sigma}_{j_\star}^{-1}\boldsymbol{y}_\star$ and $\{i_\star, j_\star\} = \arg\min_{k\in[n]} \boldsymbol{y}_\star^\top \boldsymbol{\Sigma}_k^{-1}\boldsymbol{y}_\star$.

Remark 5. Recall that, for twice differentiable V, (β_0, β_1)-smoothness is equivalent to $\left\|\nabla^2 V\right\|_{\mathrm{op}} \leq \beta_0 + \beta_1 \|\nabla V\|$. Note that every twice differentiable β-smooth function is $(\beta_0 = \beta, \beta_1 = 0)$-smooth. Hence, under the specified conditions, V is also not β-smooth for any $\beta > 0$.

Proof. Let q_i be the density of $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, i.e.,

$$q_i(\boldsymbol{x}) \triangleq \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right)}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma}_i)}}, \forall i \in [n]\,. \tag{2}$$

A direct calculation yields, for all $i \in [n]$,

$$\nabla q_i(\boldsymbol{x}) = -q_i(\boldsymbol{x})\boldsymbol{f}_i(\boldsymbol{x}) \quad \text{and} \quad \nabla^2 q_i(\boldsymbol{x}) = q_i(\boldsymbol{x})\left(\boldsymbol{f}_i(\boldsymbol{x})\boldsymbol{f}_i^\top(\boldsymbol{x}) - \boldsymbol{\Sigma}_i^{-1}\right), \tag{3}$$

where

$$\boldsymbol{f}_i(\boldsymbol{x}) \triangleq \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i). \tag{4}$$

Now, we can find that $q(\boldsymbol{x}) = \sum_{i=1}^n w_i q_i(\boldsymbol{x})$, which further implies that

$$V(\boldsymbol{x}) = -\log q(\boldsymbol{x}) = -\log\left(\sum_{i=1}^n w_i q_i(\boldsymbol{x})\right).$$

Therefore, we have

$$\nabla V(\boldsymbol{x}) = -\frac{\sum_{i=1}^n w_i \nabla q_i(\boldsymbol{x})}{q(\boldsymbol{x})}. \tag{5}$$

Differentiating once more gives

$$\nabla^2 V(\boldsymbol{x}) = \frac{\left(\sum_{i=1}^n w_i \nabla q_i(\boldsymbol{x})\right)\left(\sum_{i=1}^n w_i \nabla q_i(\boldsymbol{x})\right)^\top}{q^2(\boldsymbol{x})} - \frac{\sum_{i=1}^n w_i \nabla^2 q_i(\boldsymbol{x})}{q(\boldsymbol{x})}. \tag{6}$$

Plug (3) into (6) to have

$$\begin{aligned}\nabla^2 V(\boldsymbol{x}) = &\left(\frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x})\boldsymbol{f}_i(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\left(\frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x})\boldsymbol{f}_i(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^\top \\ &- \frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x})\left(\boldsymbol{f}_i(\boldsymbol{x})\boldsymbol{f}_i^\top(\boldsymbol{x})\right)}{q(\boldsymbol{x})} + \frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x})\boldsymbol{\Sigma}_i^{-1}}{q(\boldsymbol{x})}.\end{aligned} \tag{7}$$

- We first prove that V is $\left[\max_{i\in[n]} \lambda_{\max}(\boldsymbol{\Sigma}_i^{-1})\right]$-upper smooth. Note that

$$\left(\frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x})\boldsymbol{f}_i(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\left(\frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x})\boldsymbol{f}_i(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^\top \preceq \frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x})\left(\boldsymbol{f}_i(\boldsymbol{x})\boldsymbol{f}_i^\top(\boldsymbol{x})\right)}{q(\boldsymbol{x})}, \forall \boldsymbol{x} \in \mathbb{R}^d,$$

  which implies that, by (7),

$$\nabla^2 V(\boldsymbol{x}) \preceq \frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x})\boldsymbol{\Sigma}_i^{-1}}{q(\boldsymbol{x})} \preceq \frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x})\lambda_{\max}(\boldsymbol{\Sigma}_i^{-1})\boldsymbol{I}}{q(\boldsymbol{x})} \preceq \left[\max_{i\in[n]} \lambda_{\max}(\boldsymbol{\Sigma}_i^{-1})\right]\boldsymbol{I}, \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Therefore,

$$V(\boldsymbol{x}) - V(\boldsymbol{y}) - \langle \nabla V(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle = \int_0^1 \int_0^t (\boldsymbol{x} - \boldsymbol{y})^\top \nabla^2 V(\boldsymbol{y} + s(\boldsymbol{x} - \boldsymbol{y}))\, (\boldsymbol{x} - \boldsymbol{y})\, \mathrm{d}s\mathrm{d}t$$
$$\leq \left[\max_{i \in [n]} \lambda_{\max}(\boldsymbol{\Sigma}_i^{-1}) \right] \|\boldsymbol{x} - \boldsymbol{y}\|^2 \int_0^1 \int_0^t \mathrm{d}s\mathrm{d}t$$
$$= \frac{\max_{i \in [n]} \lambda_{\max}(\boldsymbol{\Sigma}_i^{-1})}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 .$$

- We next prove that V is not (β_0, β_1)-smooth for any $\beta_0, \beta_1 \geq 0$ under the two stated conditions. By (7), we know

$$\mathrm{Tr}(\nabla^2 V(\boldsymbol{x})) = \left\| \frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x}) \boldsymbol{f}_i(\boldsymbol{x})}{q(\boldsymbol{x})} \right\|^2 - \frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x}) \|\boldsymbol{f}_i(\boldsymbol{x})\|^2}{q(\boldsymbol{x})} + \frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x}) \mathrm{Tr}\left(\boldsymbol{\Sigma}_i^{-1}\right)}{q(\boldsymbol{x})}$$
$$= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{w_i w_j q_i(\boldsymbol{x}) q_j(\boldsymbol{x})}{q^2(\boldsymbol{x})} \left\| \boldsymbol{f}_i(\boldsymbol{x}) - \boldsymbol{f}_j(\boldsymbol{x}) \right\|^2 + \frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x}) \mathrm{Tr}\left(\boldsymbol{\Sigma}_i^{-1}\right)}{q(\boldsymbol{x})}$$
$$\leq -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{w_i w_j q_i(\boldsymbol{x}) q_j(\boldsymbol{x})}{q^2(\boldsymbol{x})} \left\| \boldsymbol{f}_i(\boldsymbol{x}) - \boldsymbol{f}_j(\boldsymbol{x}) \right\|^2 + \max_{i \in [n]} \mathrm{Tr}\left(\boldsymbol{\Sigma}_i^{-1}\right)$$
$$\leq -\frac{w_{i_\star} w_{j_\star} q_{i_\star}(\boldsymbol{x}) q_{j_\star}(\boldsymbol{x})}{q^2(\boldsymbol{x})} \left\| \boldsymbol{f}_{i_\star}(\boldsymbol{x}) - \boldsymbol{f}_{j_\star}(\boldsymbol{x}) \right\|^2 + \max_{i \in [n]} \mathrm{Tr}\left(\boldsymbol{\Sigma}_i^{-1}\right). \tag{8}$$

Now, we take

$$\boldsymbol{x}_\star(t) = t\boldsymbol{y}_\star + \boldsymbol{\mu}_\star \quad \text{where} \quad t \in \mathbb{R}, \tag{9}$$

to have (since $\boldsymbol{\mu}_{i_\star} = \boldsymbol{\mu}_{j_\star} = \boldsymbol{\mu}_\star$ and $\boldsymbol{y}_\star^\top \boldsymbol{\Sigma}_{i_\star}^{-1} \boldsymbol{y}_\star = \boldsymbol{y}_\star^\top \boldsymbol{\Sigma}_{j_\star}^{-1} \boldsymbol{y}_\star$)

$$\left\| \boldsymbol{f}_{i_\star}(\boldsymbol{x}_\star(t)) - \boldsymbol{f}_{j_\star}(\boldsymbol{x}_\star(t)) \right\|^2 \overset{(4)}{=} \left\| \left(\boldsymbol{\Sigma}_{i_\star}^{-1} - \boldsymbol{\Sigma}_{j_\star}^{-1} \right) \boldsymbol{y}_\star \right\|^2 t^2 \tag{10}$$

and

$$\frac{q_{i_\star}(\boldsymbol{x}_\star(t))}{q(\boldsymbol{x}_\star(t))} \overset{(2)}{=} \frac{\frac{1}{\sqrt{\det(\boldsymbol{\Sigma}_{i_\star})}}}{\sum_{k=1}^n \frac{w_k \exp(h_k(t))}{\sqrt{\det(\boldsymbol{\Sigma}_k)}}}, \qquad \frac{q_{j_\star}(\boldsymbol{x}_\star(t))}{q(\boldsymbol{x}_\star(t))} \overset{(2)}{=} \frac{\frac{1}{\sqrt{\det(\boldsymbol{\Sigma}_{j_\star})}}}{\sum_{k=1}^n \frac{w_k \exp(h_k(t))}{\sqrt{\det(\boldsymbol{\Sigma}_k)}}}, \tag{11}$$

where, for $Q_\star = \boldsymbol{y}_\star^\top \boldsymbol{\Sigma}_{i_\star}^{-1} \boldsymbol{y}_\star = \boldsymbol{y}_\star^\top \boldsymbol{\Sigma}_{j_\star}^{-1} \boldsymbol{y}_\star$ and $k \in [n]$,

$$h_k(t) \triangleq -\frac{(t\boldsymbol{y}_\star + \boldsymbol{\mu}_\star - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (t\boldsymbol{y}_\star + \boldsymbol{\mu}_\star - \boldsymbol{\mu}_k) - Q_\star t^2}{2}$$
$$= -\frac{(\boldsymbol{y}_\star \boldsymbol{\Sigma}_k^{-1} \boldsymbol{y}_\star - Q_\star) t^2 + 2(\boldsymbol{\mu}_\star - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{y}_\star t + (\boldsymbol{\mu}_\star - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\mu}_\star - \boldsymbol{\mu}_k)}{2}. \tag{12}$$

The condition $\boldsymbol{\Sigma}_{i_\star}^{-1} \boldsymbol{y}_\star \neq \boldsymbol{\Sigma}_{j_\star}^{-1} \boldsymbol{y}_\star$ implies that

$$\lim_{t \to \infty} \left\| \boldsymbol{f}_{i_\star}(\boldsymbol{x}_\star(t)) - \boldsymbol{f}_{j_\star}(\boldsymbol{x}_\star(t)) \right\|^2 \overset{(10)}{=} \lim_{t \to \infty} \left\| \left(\boldsymbol{\Sigma}_{i_\star}^{-1} - \boldsymbol{\Sigma}_{j_\star}^{-1} \right) \boldsymbol{y}_\star \right\|^2 t^2 = +\infty,$$

and the conditions $\boldsymbol{y}_\star^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{y}_\star > Q_\star, \forall k \notin \{i_\star, j_\star\}$ and $\boldsymbol{\mu}_{i_\star} = \boldsymbol{\mu}_{j_\star} = \boldsymbol{\mu}_\star$ imply that

$$\lim_{t \to \infty} h_k(t) \overset{(12)}{=} \begin{cases} 0 & k \in \{i_\star, j_\star\} \\ -\infty & k \notin \{i_\star, j_\star\} \end{cases} \Rightarrow \begin{array}{l} \lim_{t \to \infty} \frac{q_{i_\star}(\boldsymbol{x}_\star(t))}{q(\boldsymbol{x}_\star(t))} \overset{(11)}{=} \frac{1}{w_{i_\star} + w_{j_\star} \sqrt{\frac{\det(\boldsymbol{\Sigma}_{i_\star})}{\det(\boldsymbol{\Sigma}_{j_\star})}}} \\ \lim_{t \to \infty} \frac{q_{j_\star}(\boldsymbol{x}_\star(t))}{q(\boldsymbol{x}_\star(t))} \overset{(11)}{=} \frac{1}{w_{j_\star} + w_{i_\star} \sqrt{\frac{\det(\boldsymbol{\Sigma}_{j_\star})}{\det(\boldsymbol{\Sigma}_{i_\star})}}} \end{array}.$$

Therefore, there exists a constant $C > 0$ such that, for all sufficiently large t,

$$\frac{w_{i_\star} w_{j_\star} q_{i_\star}(\boldsymbol{x}_\star(t)) q_{j_\star}(\boldsymbol{x}_\star(t))}{q^2(\boldsymbol{x}_\star(t))} \left\| \boldsymbol{f}_{i_\star}(\boldsymbol{x}_\star(t)) - \boldsymbol{f}_{j_\star}(\boldsymbol{x}_\star(t)) \right\|^2 \geq Ct^2$$
$$\Rightarrow \mathrm{Tr}(-\nabla^2 V(\boldsymbol{x}_\star(t))) \overset{(8)}{\geq} Ct^2 - \max_{i \in [n]} \mathrm{Tr}\left(\boldsymbol{\Sigma}_i^{-1}\right).$$

We can also observe that

$$\begin{aligned}\|\nabla V(\boldsymbol{x}_\star(t))\| &\overset{(3),(5)}{=} \left\| \frac{\sum_{i=1}^n w_i q_i(\boldsymbol{x}_\star(t)) \boldsymbol{f}_i(\boldsymbol{x}_\star(t))}{q(\boldsymbol{x}_\star(t))} \right\| \leq \max_{i\in[n]} \|\boldsymbol{f}_i(\boldsymbol{x}_\star(t))\| \\ &\overset{(4),(9)}{=} \max_{i\in[n]} \left\|\boldsymbol{\Sigma}_i^{-1}(t\boldsymbol{y}_\star + \boldsymbol{\mu}_\star - \boldsymbol{\mu}_i)\right\| \\ &\leq \max_{i\in[n]} \left\|\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_\star\right\| |t| + \max_{i\in[n]} \left\|\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\mu}_\star - \boldsymbol{\mu}_i)\right\|.\end{aligned}$$

However, if V is (β_0, β_1)-smooth for some $\beta_0, \beta_1 \geq 0$, then $\nabla^2 V(\boldsymbol{x}_\star(t)) \succeq -(\beta_0 + \beta_1 \|\nabla V(\boldsymbol{x}_\star(t))\|)\boldsymbol{I}, \forall t \in \mathbb{R}$, which implies that

$$\operatorname{Tr}(-\nabla^2 V(\boldsymbol{x}_\star(t))) \leq d\left(\beta_0 + \beta_1 \|\nabla V(\boldsymbol{x}_\star(t))\|\right), \forall t \in \mathbb{R}.$$

But this cannot hold because, as shown above, $\operatorname{Tr}(-\nabla^2 V(\boldsymbol{x}_\star(t))) = \Omega(t^2)$ and $\|\nabla V(\boldsymbol{x}_\star(t))\| = \mathcal{O}(t)$ for all sufficiently large t. Therefore, V cannot be (β_0, β_1)-smooth for any $\beta_0, \beta_1 \geq 0$.

□

B General Technical Lemmas

In this section, we provide some general technical lemmas used in the later analysis, some of which (e.g., Lemmas 9 and 10) are new to the best of our knowledge. We note that Q in this section does not necessarily denote the target distribution defined earlier in (1).

B.1 Lemmas for RKHS

We provide some basic inequalities related to RKHS.

Lemma 1. *Given an RKHS $(\mathcal{H}, K, \mathbb{R}^d)$ with $K \in C^2(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$, then for every $i \in [d]$, we have:*

1. *For any $\boldsymbol{x} \in \mathbb{R}^d$, $\partial_{2,i} K(\cdot, \boldsymbol{x}) \in \mathcal{H}$.*
2. *For any $\boldsymbol{x} \in \mathbb{R}^d$ and $f \in \mathcal{H}$, $\partial_i f(\boldsymbol{x})$ is well-defined and $\partial_i f(\boldsymbol{x}) = \langle f, \partial_{2,i} K(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}}$.*

Proof. See Saitoh & Sawano (2016, Theorem 2.6). □

Lemma 2. *Given an RKHS $(\mathcal{H}, K, \mathbb{R}^d)$ with $K \in C^2(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$, then for every $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^d$, we have:*

1. $|K(\boldsymbol{z}, \boldsymbol{x}) - K(\boldsymbol{z}, \boldsymbol{y})| \leq \|K(\cdot, \boldsymbol{z})\|_{\mathcal{H}} \left(\int_0^1 \|\nabla_2 K(\cdot, \boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))\|_{\mathcal{H}^d} \,\mathrm{d}t\right) \|\boldsymbol{x} - \boldsymbol{y}\|$.
2. $\|\nabla_2 K(\boldsymbol{x}, \boldsymbol{z}) - \nabla_2 K(\boldsymbol{y}, \boldsymbol{z})\| \leq \|\nabla_2 K(\cdot, \boldsymbol{z})\|_{\mathcal{H}^d} \left(\int_0^1 \|\nabla_2 K(\cdot, \boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))\|_{\mathcal{H}^d} \,\mathrm{d}t\right) \|\boldsymbol{x} - \boldsymbol{y}\|$.
3. $\|\nabla_2 K(\boldsymbol{z}, \boldsymbol{x}) - \nabla_2 K(\boldsymbol{z}, \boldsymbol{y})\| \leq \left(\int_0^1 \left\|\nabla_2^2 K(\boldsymbol{z}, \boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))\right\|_{\mathrm{op}} \mathrm{d}t\right) \|\boldsymbol{x} - \boldsymbol{y}\|$.

Proof. We first note that, by Lemma 1, $\nabla_2 K(\cdot, \boldsymbol{x}) \in \mathcal{H}^d, \forall \boldsymbol{x} \in \mathbb{R}^d$. Hence, the R.H.S. of the first two inequalities is well-defined. Now, we are ready to prove them.

1. Observe that

$$\begin{aligned}K(\boldsymbol{z}, \boldsymbol{x}) - K(\boldsymbol{z}, \boldsymbol{y}) &= \int_0^1 \langle \nabla_2 K(\boldsymbol{z}, \boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})), \boldsymbol{x} - \boldsymbol{y}\rangle \,\mathrm{d}t \\ &= \int_0^1 \langle \nabla_2 K(\cdot, \boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})), K(\cdot, \boldsymbol{z})(\boldsymbol{x} - \boldsymbol{y})\rangle_{\mathcal{H}^d} \,\mathrm{d}t,\end{aligned}$$

which implies that

$$\begin{aligned}|K(\boldsymbol{z},\boldsymbol{x})-K(\boldsymbol{z},\boldsymbol{y})| &\le \int_0^1 |\langle \nabla_2 K(\cdot,\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y})), K(\cdot,\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{y})\rangle_{\mathcal{H}^d}|\,\mathrm{d}t \\ &\le \int_0^1 \|\nabla_2 K(\cdot,\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y}))\|_{\mathcal{H}^d}\,\|K(\cdot,\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{y})\|_{\mathcal{H}^d}\,\mathrm{d}t \\ &= \|K(\cdot,\boldsymbol{z})\|_{\mathcal{H}}\left(\int_0^1 \|\nabla_2 K(\cdot,\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y}))\|_{\mathcal{H}^d}\,\mathrm{d}t\right)\|\boldsymbol{x}-\boldsymbol{y}\|.\end{aligned}$$

2. Observe that

$$\begin{aligned}\|\nabla_2 K(\boldsymbol{x},\boldsymbol{z})-\nabla_2 K(\boldsymbol{y},\boldsymbol{z})\| &= \sqrt{\sum_{i=1}^d \left(\partial_{2,i}K(\boldsymbol{x},\boldsymbol{z})-\partial_{2,i}K(\boldsymbol{y},\boldsymbol{z})\right)^2} \\ &\overset{\text{Lemma 1}}{=} \sqrt{\sum_{i=1}^d \left(\langle K(\cdot,\boldsymbol{x})-K(\cdot,\boldsymbol{y}),\partial_{2,i}K(\cdot,\boldsymbol{z})\rangle_{\mathcal{H}}\right)^2} \le \sqrt{\sum_{i=1}^d \|K(\cdot,\boldsymbol{x})-K(\cdot,\boldsymbol{y})\|_{\mathcal{H}}^2\,\|\partial_{2,i}K(\cdot,\boldsymbol{z})\|_{\mathcal{H}}^2} \\ &= \|\nabla_2 K(\cdot,\boldsymbol{z})\|_{\mathcal{H}^d}\,\|K(\cdot,\boldsymbol{x})-K(\cdot,\boldsymbol{y})\|_{\mathcal{H}}.\end{aligned}$$

Moreover, when $K\in C^2(\mathbb{R}^d\times\mathbb{R}^d,\mathbb{R})$, we have

$$K(\cdot,\boldsymbol{x})-K(\cdot,\boldsymbol{y})=\int_0^1 \langle \nabla_2 K(\cdot,\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y})),\boldsymbol{x}-\boldsymbol{y}\rangle\,\mathrm{d}t,$$

which implies that

$$\begin{aligned}\|K(\cdot,\boldsymbol{x})-K(\cdot,\boldsymbol{y})\|_{\mathcal{H}} &\le \int_0^1 \|\langle \nabla_2 K(\cdot,\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y})),\boldsymbol{x}-\boldsymbol{y}\rangle\|_{\mathcal{H}}\,\mathrm{d}t \\ &\le \int_0^1 \sum_{i=1}^d \|\partial_{2,i}K(\cdot,\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y}))\|_{\mathcal{H}}\,|x_i-y_i|\,\mathrm{d}t \\ &\le \int_0^1 \sqrt{\sum_{i=1}^d \|\partial_{2,i}K(\cdot,\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y}))\|_{\mathcal{H}}^2}\,\|\boldsymbol{x}-\boldsymbol{y}\|\,\mathrm{d}t \\ &= \left(\int_0^1 \|\nabla_2 K(\cdot,\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y}))\|_{\mathcal{H}^d}\,\mathrm{d}t\right)\|\boldsymbol{x}-\boldsymbol{y}\|.\end{aligned}$$

Therefore, we obtain that

$$\|\nabla_2 K(\boldsymbol{x},\boldsymbol{z})-\nabla_2 K(\boldsymbol{y},\boldsymbol{z})\| \le \|\nabla_2 K(\cdot,\boldsymbol{z})\|_{\mathcal{H}^d}\left(\int_0^1 \|\nabla_2 K(\cdot,\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y}))\|_{\mathcal{H}^d}\,\mathrm{d}t\right)\|\boldsymbol{x}-\boldsymbol{y}\|.$$

3. Observe that

$$\nabla_2 K(\boldsymbol{z},\boldsymbol{x})-\nabla_2 K(\boldsymbol{z},\boldsymbol{y})=\int_0^1 \nabla_2^2 K(\boldsymbol{z},\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y}))(\boldsymbol{x}-\boldsymbol{y})\mathrm{d}t,$$

which implies that

$$\begin{aligned}\|\nabla_2 K(\boldsymbol{z},\boldsymbol{x})-\nabla_2 K(\boldsymbol{z},\boldsymbol{y})\| &\le \int_0^1 \left\|\nabla_2^2 K(\boldsymbol{z},\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y}))(\boldsymbol{x}-\boldsymbol{y})\right\|\mathrm{d}t \\ &\le \left(\int_0^1 \left\|\nabla_2^2 K(\boldsymbol{z},\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y}))\right\|_{\mathrm{op}}\mathrm{d}t\right)\|\boldsymbol{x}-\boldsymbol{y}\|.\end{aligned}$$

□

Lemma 3. *Given an RKHS $(\mathcal{H},K,\mathbb{R}^d)$ with $K\in C^2(\mathbb{R}^d\times\mathbb{R}^d,\mathbb{R})$, then for every $\boldsymbol{f}\in\mathcal{H}^d$ and $\boldsymbol{x}\in\mathbb{R}^d$, we have:*

1. $\|\boldsymbol{f}(\boldsymbol{x})\| \leq \|K(\cdot,\boldsymbol{x})\|_{\mathcal{H}} \|\boldsymbol{f}\|_{\mathcal{H}^d}$.
2. $\|\boldsymbol{J}[\boldsymbol{f}](\boldsymbol{x})\|_{\mathrm{F}} \leq \|\nabla_2 K(\cdot,\boldsymbol{x})\|_{\mathcal{H}^d} \|\boldsymbol{f}\|_{\mathcal{H}^d}$.

Proof. We first note that, by Lemma 1, $\nabla_2 K(\cdot,\boldsymbol{x}) \in \mathcal{H}^d, \forall \boldsymbol{x} \in \mathbb{R}^d$. Hence, the R.H.S. of the second inequality is well-defined. Now, we are ready to prove them.

1. Note that

$$\begin{aligned}\|\boldsymbol{f}(\boldsymbol{x})\| &= \sqrt{\sum_{i=1}^d \left(f_i(\boldsymbol{x})\right)^2} = \sqrt{\sum_{i=1}^d \left(\langle f_i, K(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}}\right)^2} \\ &\leq \sqrt{\sum_{i=1}^d \|f_i\|_{\mathcal{H}}^2 \|K(\cdot,\boldsymbol{x})\|_{\mathcal{H}}^2} = \|K(\cdot,\boldsymbol{x})\|_{\mathcal{H}} \|\boldsymbol{f}\|_{\mathcal{H}^d}.\end{aligned}$$

2. Note that, for every $i, j \in [d]$, we have

$$\left(\boldsymbol{J}[\boldsymbol{f}](\boldsymbol{x})\right)_{i,j} = \partial_j f_i(\boldsymbol{x}) \overset{\text{Lemma 1}}{=} \langle f_i, \partial_{2,j} K(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}}, \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Hence, for every $\boldsymbol{x} \in \mathbb{R}^d$,

$$\begin{aligned}\|\boldsymbol{J}[\boldsymbol{f}](\boldsymbol{x})\|_{\mathrm{F}} &= \sqrt{\sum_{i=1}^d \sum_{j=1}^d \left[\left(\boldsymbol{J}[\boldsymbol{f}](\boldsymbol{x})\right)_{i,j}\right]^2} = \sqrt{\sum_{i=1}^d \sum_{j=1}^d \left[\langle f_i, \partial_{2,j} K(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}}\right]^2} \\ &\leq \sqrt{\sum_{i=1}^d \sum_{j=1}^d \|f_i\|_{\mathcal{H}}^2 \|\partial_{2,j} K(\cdot,\boldsymbol{x})\|_{\mathcal{H}}^2} = \|\nabla_2 K(\cdot,\boldsymbol{x})\|_{\mathcal{H}^d} \|\boldsymbol{f}\|_{\mathcal{H}^d}.\end{aligned}$$

□

B.2 One Lemma for Langevin Stein Operator

We provide a representation lemma for the Langevin Stein operator (Chwialkowski et al., 2016; Liu et al., 2016; Gorham & Mackey, 2017).

Lemma 4. *Given an RKHS $(\mathcal{H}, K, \mathbb{R}^d)$ with $K \in C^2(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$, let $\boldsymbol{f} \in \mathcal{H}^d$ and $\mathcal{T}_{\mathsf{Q}}$ be the Langevin Stein operator associated with a probability measure Q on $\mathbb{R}^d$ with a differentiable density $q > 0$ (cf. Definition 3), then we have*

$$\mathcal{T}_{\mathsf{Q}}[\boldsymbol{f}](\boldsymbol{x}) = \langle \boldsymbol{f}, K(\cdot,\boldsymbol{x})\nabla \log q(\boldsymbol{x}) + \nabla_2 K(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}^d}, \forall \boldsymbol{x} \in \mathbb{R}^d,$$

where

$$K(\cdot,\boldsymbol{x})\nabla \log q(\boldsymbol{x}) + \nabla_2 K(\cdot,\boldsymbol{x}) \in \mathcal{H}^d, \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Proof. By the definition of the Langevin Stein operator $\mathcal{T}_{\mathsf{Q}}$ (cf. Definition 3), for any $\boldsymbol{x} \in \mathbb{R}^d$,

$$\begin{aligned}\mathcal{T}_{\mathsf{Q}}[\boldsymbol{f}](\boldsymbol{x}) &= \sum_{i=1}^d f_i(\boldsymbol{x})\partial_i \log q(\boldsymbol{x}) + \partial_i f_i(\boldsymbol{x}) \overset{\text{Lemma 1}}{=} \sum_{i=1}^d \langle f_i, K(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}} \partial_i \log q(\boldsymbol{x}) + \langle f_i, \partial_{2,i} K(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^d \langle f_i, K(\cdot,\boldsymbol{x})\partial_i \log q(\boldsymbol{x}) + \partial_{2,i} K(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}} = \langle \boldsymbol{f}, K(\cdot,\boldsymbol{x})\nabla \log q(\boldsymbol{x}) + \nabla_2 K(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}^d}.\end{aligned}$$

□

B.3 One Lemma for KSD

We provide a basic property of KSD (Chwialkowski et al., 2016; Liu et al., 2016; Gorham & Mackey, 2017).

Lemma 5. *Given an RKHS $(\mathcal{H}, K, \mathbb{R}^d)$ satisfying the following conditions:*

$$K \in C^2(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}), \quad K_0 \triangleq \sup_{\boldsymbol{x}\in\mathbb{R}^d} \|K(\cdot,\boldsymbol{x})\|_{\mathcal{H}} < +\infty, \quad K_1 \triangleq \sup_{\boldsymbol{x}\in\mathbb{R}^d} \|\nabla_2 K(\cdot,\boldsymbol{x})\|_{\mathcal{H}^d} < +\infty.$$

Let $\boldsymbol{f} \in \mathcal{H}^d$, $\mathcal{T}_{\mathsf{Q}}$ be the Langevin Stein operator associated with a probability measure Q on $\mathbb{R}^d$ with a differentiable density $q > 0$ (cf. Definition 3), and P be a probability measure on $\mathbb{R}^d$ satisfying $\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}[\|\nabla \log q(\boldsymbol{x})\|] < +\infty$, then we have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\mathcal{T}_{\mathsf{Q}}\left[\boldsymbol{f}\right](\boldsymbol{x})\right] = -\left\langle \boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right], \boldsymbol{f}\right\rangle_{\mathcal{H}^d},$$

where

$$\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right](\cdot) \triangleq -\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[K(\cdot,\boldsymbol{x})\nabla \log q(\boldsymbol{x}) + \nabla_2 K(\cdot,\boldsymbol{x})\right] \in \mathcal{H}^d.$$

Consequently,

$$\mathsf{KSD}_{\mathcal{H}}(\mathsf{P}\|\mathsf{Q}) = \left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right]\right\|_{\mathcal{H}^d}.$$

Remark 6. In contrast to the existing literature, Lemma 5 provides a simple sufficient condition for the well-posedness of $\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]$.

Proof. We first note that, by Lemma 1, $\nabla_2 K(\cdot,\boldsymbol{x}) \in \mathcal{H}^d, \forall \boldsymbol{x} \in \mathbb{R}^d$. Hence, $\|\nabla_2 K(\cdot,\boldsymbol{x})\|_{\mathcal{H}^d}$ is well-defined.

Next, for every $\boldsymbol{x} \in \mathbb{R}^d$,

$$\begin{aligned}&\|K(\cdot,\boldsymbol{x})\nabla \log q(\boldsymbol{x}) + \nabla_2 K(\cdot,\boldsymbol{x})\|_{\mathcal{H}^d} \le \|K(\cdot,\boldsymbol{x})\nabla \log q(\boldsymbol{x})\|_{\mathcal{H}^d} + \|\nabla_2 K(\cdot,\boldsymbol{x})\|_{\mathcal{H}^d}\\ =&\|K(\cdot,\boldsymbol{x})\|_{\mathcal{H}} \|\nabla \log q(\boldsymbol{x})\| + \|\nabla_2 K(\cdot,\boldsymbol{x})\|_{\mathcal{H}^d} \le K_0 \|\nabla \log q(\boldsymbol{x})\| + K_1,\end{aligned}$$

which implies that

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|K(\cdot,\boldsymbol{x})\nabla \log q(\boldsymbol{x}) + \nabla_2 K(\cdot,\boldsymbol{x})\|_{\mathcal{H}^d}\right] \le K_0 \mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla \log q(\boldsymbol{x})\|\right] + K_1 < +\infty.$$

Thus, $K(\cdot,\boldsymbol{x})\nabla \log q(\boldsymbol{x}) + \nabla_2 K(\cdot,\boldsymbol{x})$ is Bochner integrable and

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[K(\cdot,\boldsymbol{x})\nabla \log q(\boldsymbol{x}) + \nabla_2 K(\cdot,\boldsymbol{x})\right] \in \mathcal{H}^d.$$

Finally, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\mathcal{T}_{\mathsf{Q}}\left[\boldsymbol{f}\right](\boldsymbol{x})\right] &\overset{\text{Lemma 4}}{=} \mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\langle \boldsymbol{f}, K(\cdot,\boldsymbol{x})\nabla \log q(\boldsymbol{x}) + \nabla_2 K(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}^d}\right]\\ &= \langle \boldsymbol{f}, \mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[K(\cdot,\boldsymbol{x})\nabla \log q(\boldsymbol{x}) + \nabla_2 K(\cdot,\boldsymbol{x})\right]\rangle_{\mathcal{H}^d} = -\left\langle \boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right], \boldsymbol{f}\right\rangle_{\mathcal{H}^d}.\end{aligned}$$

Consequently, by the definition of KSD (cf. Definition 4), we obtain that

$$\begin{aligned}\mathsf{KSD}_{\mathcal{H}}(\mathsf{P}\|\mathsf{Q}) &= \sup_{\boldsymbol{f}\in\mathcal{H}^d, \|\boldsymbol{f}\|_{\mathcal{H}^d}\le 1} \mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\mathcal{T}_{\mathsf{Q}}\left[\boldsymbol{f}\right](\boldsymbol{x})\right]\\ &= \sup_{\boldsymbol{f}\in\mathcal{H}^d, \|\boldsymbol{f}\|_{\mathcal{H}^d}\le 1} -\left\langle \boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right], \boldsymbol{f}\right\rangle_{\mathcal{H}^d} = \left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right]\right\|_{\mathcal{H}^d}.\end{aligned}$$

□

B.4 One Lemma for KL Divergence

Lemma 6 is one of the key differences between our analysis and the existing literature. Unlike standard change-of-variables arguments that require the map to be globally invertible or bi-Lipschitz, Lemma 6 only requires the lower-Lipschitz condition to hold on a Borel set S with $\mathsf{P}(\mathrm{S}) = 1$. This weaker formulation is crucial for our finite-particle analysis, where such control is available only on the set of points reachable along the SVGD trajectory rather than on the entire space.

Lemma 6. *Given two probability measures P and Q on $\mathbb{R}^d$ with densities $p \triangleq \frac{\mathrm{d}\mathsf{P}}{\mathrm{d}\mathsf{Leb}_d}$ and $q \triangleq \frac{\mathrm{d}\mathsf{Q}}{\mathrm{d}\mathsf{Leb}_d} > 0$, respectively, and suppose that $\mathsf{KL}(\mathsf{P}\|\mathsf{Q}) < +\infty$, let $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^d$ be a differentiable function satisfying $\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\| \ge C\|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{S}$ for some $C > 0$, where $\mathrm{S} \subseteq \mathbb{R}^d$ is a Borel set with $\mathsf{P}(\mathrm{S}) = 1$, then we have*

$$\mathsf{KL}(\boldsymbol{f}_{\#}\mathsf{P}\|\mathsf{Q}) = \mathsf{KL}(\mathsf{P}\|\mathsf{Q}) + \mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[-\log q(\boldsymbol{f}(\boldsymbol{x})) + \log q(\boldsymbol{x}) - \log\left|\det\left(\boldsymbol{J}\left[\boldsymbol{f}\right](\boldsymbol{x})\right)\right|\right].$$

Proof. By the condition on $\boldsymbol{f}$, its inverse on $\boldsymbol{f}(\mathrm{S})$, i.e., $\boldsymbol{f}_{|\mathrm{S}}^{-1}(\boldsymbol{y}) : \boldsymbol{f}(\mathrm{S}) \to \mathrm{S}, \boldsymbol{y} \mapsto \boldsymbol{x} \in \mathrm{S}$ such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}$, is well-defined and $\frac{1}{C}$-Lipschitz on $\boldsymbol{f}(\mathrm{S})$. As such, given a Borel measurable set $A \subseteq \mathbb{R}^d$ satisfying $\mathsf{Leb}_d(A) = 0$, we know

$$0 \leq \mathsf{Leb}_d(\boldsymbol{f}_{|\mathrm{S}}^{-1}(A \cap \boldsymbol{f}(\mathrm{S}))) \leq \left(\frac{1}{C}\right)^d \mathsf{Leb}_d(A \cap \boldsymbol{f}(\mathrm{S})) \leq \mathsf{Leb}_d(A) = 0$$

$$\Rightarrow \mathsf{Leb}_d(\boldsymbol{f}_{|\mathrm{S}}^{-1}(A \cap \boldsymbol{f}(\mathrm{S}))) = 0 \overset{\mathsf{P} \ll \mathsf{Leb}_d}{\Rightarrow} \mathsf{P}(\boldsymbol{f}_{|\mathrm{S}}^{-1}(A \cap \boldsymbol{f}(\mathrm{S}))) = 0, \tag{13}$$

which implies that

$$\boldsymbol{f}_{\#}\mathsf{P}(A) = \mathsf{P}(\boldsymbol{f}^{-1}(A)) \overset{(a)}{=} \mathsf{P}(\boldsymbol{f}^{-1}(A) \cap \mathrm{S}) = \mathsf{P}(\boldsymbol{f}_{|\mathrm{S}}^{-1}(A \cap \boldsymbol{f}(\mathrm{S})) \overset{(13)}{=} 0 \Rightarrow \boldsymbol{f}_{\#}\mathsf{P} \ll \mathsf{Leb}_d,$$

where (a) holds because $\boldsymbol{f}^{-1}(A) \cap (\mathbb{R}^d \setminus \mathrm{S})$ is a P-null set. Therefore, $\boldsymbol{f}_{\#}\mathsf{P} \ll \mathsf{Leb}_d$ and its density $p_{\boldsymbol{f}} \triangleq \frac{\mathrm{d}\boldsymbol{f}_{\#}\mathsf{P}}{\mathrm{d}\mathsf{Leb}_d}$ exists and satisfies (by the area formula)

$$p_{\boldsymbol{f}}(\boldsymbol{f}(\boldsymbol{x})) = \frac{p(\boldsymbol{x})}{|\det\left(\boldsymbol{J}\left[\boldsymbol{f}\right](\boldsymbol{x})\right)|}, \quad \mathsf{P}\text{-almost surely.} \tag{14}$$

Finally, we have

$$\begin{aligned}
\mathsf{KL}(\boldsymbol{f}_{\#}\mathsf{P} \| \mathsf{Q}) &= \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{f}_{\#}\mathsf{P}}\left[\log \frac{p_{\boldsymbol{f}}(\boldsymbol{x})}{q(\boldsymbol{x})}\right] = \mathbb{E}_{\boldsymbol{x} \sim \mathsf{P}}\left[\log \frac{p_{\boldsymbol{f}}(\boldsymbol{f}(\boldsymbol{x}))}{q(\boldsymbol{f}(\boldsymbol{x}))}\right] \\
&= \mathbb{E}_{\boldsymbol{x} \sim \mathsf{P}}\left[\log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} + \log \frac{q(\boldsymbol{x})}{q(\boldsymbol{f}(\boldsymbol{x}))} + \log \frac{p_{\boldsymbol{f}}(\boldsymbol{f}(\boldsymbol{x}))}{p(\boldsymbol{x})}\right] \\
&\overset{(14)}{=} \mathsf{KL}(\mathsf{P} \| \mathsf{Q}) + \mathbb{E}_{\boldsymbol{x} \sim \mathsf{P}}\left[-\log q(\boldsymbol{f}(\boldsymbol{x})) + \log q(\boldsymbol{x}) - \log|\det\left(\boldsymbol{J}\left[\boldsymbol{f}\right](\boldsymbol{x})\right)|\right].
\end{aligned}$$

□

B.5 One Lemma for Matrix Inequality

We state a useful matrix inequality.

Lemma 7. *Given $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ such that $\|\boldsymbol{A}\|_{\mathrm{op}} < 1$, we have*

$$-\log|\det\left(\boldsymbol{I} - \boldsymbol{A}\right)| \leq \mathrm{Tr}\left(\boldsymbol{A}\right) + \frac{\|\boldsymbol{A}\|_{\mathrm{F}}^2}{2(1 - \|\boldsymbol{A}\|_{\mathrm{op}})}.$$

Remark 7. Using $\sum_{k=2}^{+\infty} \frac{a^{k-2}}{k} = \frac{-\log(1-a)-a}{a^2}$ for $0 \leq a < 1$, one can tighten the last term in the above bound to $\|\boldsymbol{A}\|_{\mathrm{F}}^2 \frac{-\log(1-\|\boldsymbol{A}\|_{\mathrm{op}})-\|\boldsymbol{A}\|_{\mathrm{op}}}{\|\boldsymbol{A}\|_{\mathrm{op}}^2}$. However, we retain the current one for simplicity.

Proof. Note that $\det\left(\boldsymbol{I} - \boldsymbol{A}\right)$ must be positive, we hence have

$$\begin{aligned}
-\log|\det\left(\boldsymbol{I} - \boldsymbol{A}\right)| &= -\log\det\left(\boldsymbol{I} - \boldsymbol{A}\right) \overset{(a)}{=} \mathrm{Tr}\left(-\log\left(\boldsymbol{I} - \boldsymbol{A}\right)\right) \overset{(b)}{=} \mathrm{Tr}\left(\sum_{k=1}^{+\infty} \frac{\boldsymbol{A}^k}{k}\right) \\
&= \sum_{k=1}^{+\infty} \frac{\mathrm{Tr}(\boldsymbol{A}^k)}{k} \overset{(c)}{\leq} \mathrm{Tr}(\boldsymbol{A}) + \sum_{k=2}^{+\infty} \frac{\|\boldsymbol{A}\|_{\mathrm{op}}^{k-2} \|\boldsymbol{A}\|_{\mathrm{F}}^2}{k} \\
&\leq \mathrm{Tr}(\boldsymbol{A}) + \sum_{k=2}^{+\infty} \frac{\|\boldsymbol{A}\|_{\mathrm{op}}^{k-2} \|\boldsymbol{A}\|_{\mathrm{F}}^2}{2} \leq \mathrm{Tr}(\boldsymbol{A}) + \frac{\|\boldsymbol{A}\|_{\mathrm{F}}^2}{2(1 - \|\boldsymbol{A}\|_{\mathrm{op}})},
\end{aligned}$$

where (a) and (b) hold due to $\|\boldsymbol{A}\|_{\mathrm{op}} < 1$ and (c) is by, for $2 \leq k \in \mathbb{N}$ and standard basis $\left\{\boldsymbol{e}_i \in \mathbb{R}^d\right\}_{i=1}^d$,

$$\begin{aligned}
\mathrm{Tr}(\boldsymbol{A}^k) &= \sum_{i=1}^d \left\langle \boldsymbol{A}^{k-1}\boldsymbol{e}_i, \boldsymbol{A}^\top \boldsymbol{e}_i \right\rangle \leq \sum_{i=1}^d \left\|\boldsymbol{A}^{k-1}\boldsymbol{e}_i\right\| \left\|\boldsymbol{A}^\top \boldsymbol{e}_i\right\| \leq \sum_{i=1}^d \left\|\boldsymbol{A}^{k-2}\right\|_{\mathrm{op}} \|\boldsymbol{A}\boldsymbol{e}_i\| \left\|\boldsymbol{A}^\top \boldsymbol{e}_i\right\| \\
&\leq \left\|\boldsymbol{A}^{k-2}\right\|_{\mathrm{op}} \sqrt{\left(\sum_{i=1}^d \|\boldsymbol{A}\boldsymbol{e}_i\|^2\right)\left(\sum_{i=1}^d \left\|\boldsymbol{A}^\top \boldsymbol{e}_i\right\|^2\right)} = \left\|\boldsymbol{A}^{k-2}\right\|_{\mathrm{op}} \|\boldsymbol{A}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{A}\|_{\mathrm{op}}^{k-2} \|\boldsymbol{A}\|_{\mathrm{F}}^2.
\end{aligned}$$

□

B.6 Lemmas for Upper Smooth Functions

We provide several technical results on upper smooth functions, many of which are new.

We start by recalling the definition of upper smoothness.

Definition 5 (β-upper smoothness)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called β-upper smooth for some $\beta > 0$, if

$$f(\boldsymbol{x}) \le f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{\beta}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

Remark 8. Differentiability is not essential for defining upper smoothness. More generally, a function $f : \mathbb{R}^d \to \mathbb{R}$ may be called β-upper smooth if $\boldsymbol{x} \mapsto \frac{\beta}{2} \|\boldsymbol{x}\|^2 - f(\boldsymbol{x})$ is convex. For simplicity, however, we restrict our attention to differentiable functions.

Lemma 8 gives a useful property of upper smooth functions.

Lemma 8. *Given a function $f : \mathbb{R}^d \to \mathbb{R}$ that is differentiable and β-upper smooth (cf. Definition 5), for $\eta > 0$, let*

$$\boldsymbol{G}_\eta(\boldsymbol{x}) \triangleq \boldsymbol{x} - \eta \nabla f(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathbb{R}^d, \tag{15}$$

then, when $\eta < \frac{1}{\beta}$, $\boldsymbol{G}_\eta$ is invertible on $\boldsymbol{G}_\eta(\mathbb{R}^d)$ and its inverse $\boldsymbol{G}_\eta^{-1} : \boldsymbol{G}_\eta(\mathbb{R}^d) \to \mathbb{R}^d$ is $\frac{1}{1-\beta\eta}$-Lipschitz.

Proof. By β-upper smoothness, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we have

$$f(\boldsymbol{x}) \le f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{\beta}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \quad \text{and} \quad f(\boldsymbol{y}) \le f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\beta}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2,$$

which together imply that

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \le \beta \|\boldsymbol{x} - \boldsymbol{y}\|^2, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d. \tag{16}$$

Next, we observe that, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\langle \boldsymbol{G}_\eta(\boldsymbol{x}) - \boldsymbol{G}_\eta(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \overset{(15)}{=} \|\boldsymbol{x} - \boldsymbol{y}\|^2 - \eta \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \overset{(16)}{\ge} (1 - \beta\eta) \|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Combining the above inequality with $\langle \boldsymbol{G}_\eta(\boldsymbol{x}) - \boldsymbol{G}_\eta(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \le \|\boldsymbol{G}_\eta(\boldsymbol{x}) - \boldsymbol{G}_\eta(\boldsymbol{y})\| \|\boldsymbol{x} - \boldsymbol{y}\|$ (due to Cauchy-Schwarz inequality), we obtain that

$$\|\boldsymbol{G}_\eta(\boldsymbol{x}) - \boldsymbol{G}_\eta(\boldsymbol{y})\| \ge (1 - \beta\eta) \|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

In particular, when $\eta < \frac{1}{\beta}$, the above inequality is saying that $\boldsymbol{G}_\eta$ is invertible on $\boldsymbol{G}_\eta(\mathbb{R}^d)$ and satisfies

$$\left\|\boldsymbol{G}_\eta^{-1}(\boldsymbol{x}) - \boldsymbol{G}_\eta^{-1}(\boldsymbol{y})\right\| \le \frac{1}{1 - \beta\eta} \|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{G}_\eta(\mathbb{R}^d).$$

□

We are now ready to state our first new result, Lemma 9. As far as we know, Lemma 9 provides the first exponential-moment bound for the score norm under only β-upper smoothness, without convexity, a lower Hessian bound, or additional functional inequalities as required in prior work (Lee et al., 2020; Altschuler & Chewi, 2026).

Lemma 9. *Given a function $f : \mathbb{R}^d \to \mathbb{R}$ that is differentiable and β-upper smooth (cf. Definition 5), suppose*

$$Z \triangleq \int_{\mathbb{R}^d} \exp\left(-f(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} < +\infty, \tag{17}$$

let

$$C(\lambda) \triangleq \int_{\mathbb{R}^d} \frac{\exp\left(\lambda \|\nabla f(\boldsymbol{x})\| - f(\boldsymbol{x})\right)}{Z} \mathrm{d}\boldsymbol{x}, \forall \lambda \ge 0, \tag{18}$$

then we have

$$C(\lambda) \le \exp\left(\frac{\lambda\sqrt{4d\beta + \lambda^2\beta^2} + \lambda^2\beta}{4} + d \log\left(\frac{\sqrt{4d + \lambda^2\beta} + \lambda\sqrt{\beta}}{2\sqrt{d}}\right)\right), \forall \lambda \ge 0.$$

Proof. Let $0 < \eta < \frac{1}{\beta}$ be a parameter, we define

$$\boldsymbol{G}_\eta(\boldsymbol{x}) \triangleq \boldsymbol{x} - \eta \nabla f(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathbb{R}^d. \tag{19}$$

By Lemma 8, $\boldsymbol{G}_\eta$ is invertible on $\boldsymbol{G}_\eta(\mathbb{R}^d)$ and its inverse $\boldsymbol{G}_\eta^{-1}$ is $\frac{1}{1-\beta\eta}$-Lipschitz, i.e.,

$$\left\|\boldsymbol{G}_\eta^{-1}(\boldsymbol{x}) - \boldsymbol{G}_\eta^{-1}(\boldsymbol{y})\right\| \leq \frac{1}{1-\beta\eta} \left\|\boldsymbol{x} - \boldsymbol{y}\right\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{G}_\eta(\mathbb{R}^d). \tag{20}$$

Now, by β-upper smoothness, we know

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\beta}{2} \left\|\boldsymbol{y} - \boldsymbol{x}\right\|^2, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

Plugging $\boldsymbol{y} = \boldsymbol{G}_\eta(\boldsymbol{x})$ into the above inequality, we obtain that

$$f(\boldsymbol{G}_\eta(\boldsymbol{x})) \leq f(\boldsymbol{x}) - \left(\eta - \frac{\eta^2\beta}{2}\right) \left\|\nabla f(\boldsymbol{x})\right\|^2 = f(\boldsymbol{x}) - h(\eta) \left\|\nabla f(\boldsymbol{x})\right\|^2, \forall \boldsymbol{x} \in \mathbb{R}^d, \tag{21}$$

where, for simplicity, we denote by

$$h(\eta) \triangleq \eta - \frac{\eta^2\beta}{2}, \forall \eta > 0. \tag{22}$$

Note that $h(\eta) > 0$, since we assume $\eta < \frac{1}{\beta}$.

Next, for any $\lambda > 0$, there is

$$\lambda \left\|\nabla f(\boldsymbol{x})\right\| - f(\boldsymbol{x}) \overset{(21)}{\leq} \lambda \left\|\nabla f(\boldsymbol{x})\right\| - h(\eta) \left\|\nabla f(\boldsymbol{x})\right\|^2 - f(\boldsymbol{G}_\eta(\boldsymbol{x})) \leq \frac{\lambda^2}{4h(\eta)} - f(\boldsymbol{G}_\eta(\boldsymbol{x})), \forall \boldsymbol{x} \in \mathbb{R}^d, \tag{23}$$

where the last step is due to, by AM-GM inequality,

$$\lambda \left\|\nabla f(\boldsymbol{x})\right\| - h(\eta) \left\|\nabla f(\boldsymbol{x})\right\|^2 \leq \frac{\lambda^2}{4h(\eta)}.$$

As such, we can find that

$$\begin{aligned} C(\lambda) &\overset{(18)}{=} \int_{\mathbb{R}^d} \frac{\exp\left(\lambda \left\|\nabla f(\boldsymbol{x})\right\| - f(\boldsymbol{x})\right)}{Z} \mathrm{d}\boldsymbol{x} \overset{(23)}{\leq} \exp\left(\frac{\lambda^2}{4h(\eta)}\right) \int_{\mathbb{R}^d} \frac{\exp\left(-f(\boldsymbol{G}_\eta(\boldsymbol{x}))\right)}{Z} \mathrm{d}\boldsymbol{x} \\ &\overset{(a)}{=} \exp\left(\frac{\lambda^2}{4h(\eta)}\right) \int_{\mathbb{R}^d} \frac{\exp\left(-f(\boldsymbol{x})\right)}{Z} \mathrm{d}\left(\boldsymbol{G}_\eta\right)_{\#} \mathsf{Leb}_d(\boldsymbol{x}) \overset{(b)}{\leq} \frac{\exp\left(\frac{\lambda^2}{4h(\eta)}\right)}{(1-\beta\eta)^d} \int_{\mathbb{R}^d} \frac{\exp\left(-f(\boldsymbol{x})\right)}{Z} \mathrm{d}\boldsymbol{x} \\ &\overset{(17),(22)}{=} \exp\left(\frac{\lambda^2}{4\eta - 2\eta^2\beta} - d\log(1-\beta\eta)\right), \end{aligned} \tag{24}$$

where (a) is by the basic property of the pushforward measure, i.e., given a mapping $\boldsymbol{G} : \mathbb{R}^d \to \mathbb{R}^d$, for any nonnegative function h, there is

$$\int_{\mathbb{R}^d} h(\boldsymbol{G}(\boldsymbol{x})) \mathrm{d}\mathsf{Leb}_d(\boldsymbol{x}) = \int_{\mathbb{R}^d} h(\boldsymbol{x}) \mathrm{d}\boldsymbol{G}_{\#}\mathsf{Leb}_d(\boldsymbol{x}),$$

and (b) follows from, for any Borel measurable set $A \subseteq \mathbb{R}^d$,

$$\begin{aligned} \left(\boldsymbol{G}_\eta\right)_{\#} \mathsf{Leb}_d(A) &= \mathsf{Leb}_d\left(\boldsymbol{G}_\eta^{-1}(A)\right) = \mathsf{Leb}_d\left(\boldsymbol{G}_\eta^{-1}(A \cap \boldsymbol{G}_\eta(\mathbb{R}^d))\right) \\ &\overset{(c)}{\leq} \left(\mathrm{Lip}(\boldsymbol{G}_\eta^{-1})\right)^d \mathsf{Leb}_d(A \cap \boldsymbol{G}_\eta(\mathbb{R}^d)) \leq \left(\mathrm{Lip}(\boldsymbol{G}_\eta^{-1})\right)^d \mathsf{Leb}_d(A) \overset{(20)}{=} \frac{\mathsf{Leb}_d(A)}{(1-\beta\eta)^d}, \end{aligned}$$

see, for example, Evans (2025, Theorem 2.8) for the proof of (c).

Therefore, we have

$$\begin{aligned} C(\lambda) &\overset{(24)}{\leq} \exp\left(\frac{\lambda^2}{4\eta - 2\eta^2\beta} - d\log(1-\beta\eta)\right), \forall 0 < \eta < \frac{1}{\beta} \\ \Rightarrow C(\lambda) &\leq \exp\left(\inf_{0<\eta<\frac{1}{\beta}} \left(\frac{\lambda^2}{4\eta - 2\eta^2\beta} - d\log(1-\beta\eta)\right)\right) \\ &\overset{\eta=\frac{\gamma}{\beta}}{=} \exp\left(\inf_{0<\gamma<1} \left(\frac{\lambda^2\beta}{4\gamma - 2\gamma^2} - d\log(1-\gamma)\right)\right) = \exp\left(\inf_{0<\gamma<1} k(\gamma)\right), \end{aligned}$$

where

$$k(\gamma) \triangleq \frac{\lambda^2\beta}{4\gamma - 2\gamma^2} - d\log(1-\gamma).$$

We can find that

$$k'(\gamma) = -\frac{\lambda^2\beta}{4\gamma^2} + \frac{\lambda^2\beta}{4(2-\gamma)^2} + \frac{d}{1-\gamma} = \frac{d\gamma^2(2-\gamma)^2 - \lambda^2\beta(1-\gamma)^2}{\gamma^2(2-\gamma)^2(1-\gamma)}.$$

Let $\gamma_\star \triangleq \frac{2\sqrt{d}+\lambda\sqrt{\beta}-\sqrt{4d+\lambda^2\beta}}{2\sqrt{d}} \in (0,1)$ be the only solution of $k'(\gamma)=0$ in $(0,1)$, then

$$\begin{aligned}\inf_{0<\gamma<1} k(\gamma) &= \min\left\{k(0), k(\gamma_\star), k(1)\right\} = k(\gamma_\star)\\ &= \frac{\lambda\sqrt{4d\beta+\lambda^2\beta^2}+\lambda^2\beta}{4} + d\log\left(\frac{\sqrt{4d+\lambda^2\beta}+\lambda\sqrt{\beta}}{2\sqrt{d}}\right),\end{aligned}$$

which leads us to

$$C(\lambda) \le \exp\left(\frac{\lambda\sqrt{4d\beta+\lambda^2\beta^2}+\lambda^2\beta}{4} + d\log\left(\frac{\sqrt{4d+\lambda^2\beta}+\lambda\sqrt{\beta}}{2\sqrt{d}}\right)\right).$$

□

Equipped with Lemma 9, we can establish another novel inequality in Lemma 10 below. As far as we can check, this score-moment bound under only β-upper smoothness has not previously appeared. Notably, when $\mathsf{P}=\mathsf{Q}$, it recovers the standard bound $\mathbb{E}_{\boldsymbol{x}\sim\mathsf{Q}}\left[\|\nabla f(\boldsymbol{x})\|\right] \le \sqrt{d\beta}$ (Dalalyan, 2017; Vempala & Wibisono, 2019).

We highlight that Lemma 10 plays a fundamental role in our analysis of Mean-Field SVGD. For details, we refer interested readers to Appendix C.4.

Lemma 10. *Under the same setting as in Lemma 9, let* $\mathrm{d}\mathsf{Q} \triangleq \frac{\exp(-f(\boldsymbol{x}))}{Z}\mathrm{d}\boldsymbol{x}$ *and* P *be an arbitrary probability measure on* $\mathbb{R}^d$*, then we have*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla f(\boldsymbol{x})\|\right] \le \sqrt{d\beta} + \sqrt{2\beta\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}.$$

Proof. By the Donsker-Varadhan variational formula of KL divergence (Donsker & Varadhan, 1983), for any $h:\mathbb{R}^d\to\mathbb{R}$ such that $\mathbb{E}_{\boldsymbol{x}\sim\mathsf{Q}}\left[\exp(h(\boldsymbol{x}))\right] \in (0,+\infty)$, there is

$$\begin{aligned}&\mathsf{KL}(\mathsf{P}\|\mathsf{Q}) \ge \mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[h(\boldsymbol{x})\right] - \log\mathbb{E}_{\boldsymbol{x}\sim\mathsf{Q}}\left[\exp(h(\boldsymbol{x}))\right]\\ \Rightarrow\ &\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[h(\boldsymbol{x})\right] \le \mathsf{KL}(\mathsf{P}\|\mathsf{Q}) + \log\mathbb{E}_{\boldsymbol{x}\sim\mathsf{Q}}\left[\exp(h(\boldsymbol{x}))\right].\end{aligned} \tag{25}$$

Now, note that for $h(\boldsymbol{x}) = \lambda\|\nabla f(\boldsymbol{x})\|$, where $\lambda>0$,

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{Q}}\left[\exp\left(h(\boldsymbol{x})\right)\right] = \underbrace{\int_{\mathbb{R}^d}\frac{\exp\left(\lambda\|\nabla f(\boldsymbol{x})\| - f(\boldsymbol{x})\right)}{Z}\mathrm{d}\boldsymbol{x}}_{\triangleq C(\lambda)},$$

which satisfies, by Lemma 9,

$$0 < C(\lambda) \le \exp\left(\frac{\lambda\sqrt{4d\beta+\lambda^2\beta^2}+\lambda^2\beta}{4} + d\log\left(\frac{\sqrt{4d+\lambda^2\beta}+\lambda\sqrt{\beta}}{2\sqrt{d}}\right)\right) < +\infty.$$

Plug the above inequality into (25) and divide both sides of the obtained inequality by λ to have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla f(\boldsymbol{x})\|\right] \le \frac{\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{\lambda} + \frac{\sqrt{4d\beta+\lambda^2\beta^2}+\lambda\beta}{4} + \frac{d}{\lambda}\log\left(\frac{\sqrt{4d+\lambda^2\beta}+\lambda\sqrt{\beta}}{2\sqrt{d}}\right), \forall\lambda>0. \tag{26}$$

Next, substituting $\lambda = 2\sqrt{\frac{d}{\beta}}\gamma$ (where $\gamma > 0$) in (26), we obtain that

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla f(\boldsymbol{x})\|\right] \leq \frac{\sqrt{d\beta}}{2}\left[\sqrt{1+\gamma^2}+\gamma+\frac{\log\left(\sqrt{1+\gamma^2}+\gamma\right)+\frac{\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d}}{\gamma}\right] = \frac{\sqrt{d\beta}}{2}h(\gamma), \forall\gamma>0, \tag{27}$$

where we define

$$h(\gamma) \triangleq \sqrt{1+\gamma^2}+\gamma+\frac{\log\left(\sqrt{1+\gamma^2}+\gamma\right)+\frac{\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d}}{\gamma}, \forall\gamma>0. \tag{28}$$

Note that (27) implies that

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla f(\boldsymbol{x})\|\right] \leq \frac{\sqrt{d\beta}}{2}\inf_{\gamma>0}h(\gamma). \tag{29}$$

We observe that

$$h'(\gamma) = \frac{\gamma^2+\gamma\sqrt{1+\gamma^2}-\log\left(\sqrt{1+\gamma^2}+\gamma\right)-\frac{\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d}}{\gamma^2} = \frac{k(\gamma)-\frac{\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d}}{\gamma^2}, \tag{30}$$

where

$$k(\gamma) \triangleq \gamma^2+\gamma\sqrt{1+\gamma^2}-\log\left(\sqrt{1+\gamma^2}+\gamma\right), \forall\gamma>0.$$

Note that $k'(\gamma) = 2\gamma+2\sqrt{1+\gamma^2}-\frac{2}{\sqrt{1+\gamma^2}}$ is nondecreasing in $(0,+\infty)$, $\lim_{\gamma\to0^+}k(\gamma)=0$, $\lim_{\gamma\to+\infty}k(\gamma)=+\infty$. Therefore, by (29),

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla f(\boldsymbol{x})\|\right] \leq \frac{\sqrt{d\beta}}{2}\inf_{\gamma>0}h(\gamma) = \frac{\sqrt{d\beta}}{2}h(\gamma_\star), \tag{31}$$

where $\gamma_\star > 0$ is the unique solution satisfying that

$$h'(\gamma_\star) \overset{(30)}{=} 0 \Leftrightarrow \gamma_\star^2+\gamma_\star\sqrt{1+\gamma_\star^2}-\log\left(\sqrt{1+\gamma_\star^2}+\gamma_\star\right) = \frac{\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d} \tag{32}$$

$$\Leftrightarrow \left(\sqrt{1+\gamma_\star^2}+\gamma_\star\right)^2-\log\left(\sqrt{1+\gamma_\star^2}+\gamma_\star\right)^2 = 1+\frac{2\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d} \tag{33}$$

$$\Leftrightarrow -\left(\sqrt{1+\gamma_\star^2}+\gamma_\star\right)^2\exp\left(-\left(\sqrt{1+\gamma_\star^2}+\gamma_\star\right)^2\right) = -\frac{1}{\exp\left(1+\frac{2\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d}\right)}$$

$$\Leftrightarrow \sqrt{1+\gamma_\star^2}+\gamma_\star = \sqrt{-W_{-1}\left(-\frac{1}{\exp\left(1+\frac{2\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d}\right)}\right)}, \tag{34}$$

where W_{-1} is the -1 branch of the Lambert W function.

Finally, we conclude that

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla f(\boldsymbol{x})\|\right] \overset{(28),(31)}{\leq} \frac{\sqrt{d\beta}}{2}\left[\sqrt{1+\gamma_\star^2}+\gamma_\star+\frac{\log\left(\sqrt{1+\gamma_\star^2}+\gamma_\star\right)+\frac{\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d}}{\gamma_\star}\right]$$

$$\overset{(32)}{=} \sqrt{d\beta}\left(\sqrt{1+\gamma_\star^2}+\gamma_\star\right) \overset{(34)}{=} \sqrt{-d\beta W_{-1}\left(-\frac{1}{\exp\left(1+\frac{2\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d}\right)}\right)}. \tag{35}$$

A simpler bound is that

$$\left(\sqrt{1+\gamma_\star^2}+\gamma_\star\right)^2-2\left(\sqrt{1+\gamma_\star^2}+\gamma_\star-1\right) \overset{(33)}{\leq} \left(\sqrt{1+\gamma_\star^2}+\gamma_\star\right)^2-\log\left(\sqrt{1+\gamma_\star^2}+\gamma_\star\right)^2$$

$$= 1+\frac{2\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d},$$

which implies that

$$\sqrt{1+\gamma_\star^2}+\gamma_\star \overset{(34)}{=} \sqrt{-W_{-1}\left(-\frac{1}{\exp\left(1+\frac{2\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d}\right)}\right)} \leq 1+\sqrt{\frac{2\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d}}.$$

Hence, we know

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla f(\boldsymbol{x})\|\right] \overset{(35)}{\leq} \sqrt{-d\beta W_{-1}\left(-\frac{1}{\exp\left(1+\frac{2\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}{d}\right)}\right)} \leq \sqrt{d\beta}+\sqrt{2\beta\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}.$$

□

Moreover, we show that if $\exp(-f(\boldsymbol{x}))$ is integrable, then β-upper smoothness of f implies that f is bounded from below, with an explicit lower bound. As far as we can check, this result is new.

Lemma 11. *Given a function $f:\mathbb{R}^d\to\mathbb{R}$ that is differentiable and β-upper smooth (cf. Definition 5), suppose*

$$Z \triangleq \int_{\mathbb{R}^d}\exp\left(-f(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x} < +\infty, \tag{36}$$

then we have

$$f(\boldsymbol{x}) \geq \frac{\|\nabla f(\boldsymbol{x})\|^2}{2\beta}+\frac{d}{2}\log\frac{2\pi}{\beta}-\log Z, \forall \boldsymbol{x}\in\mathbb{R}^d.$$

Consequently,

$$\inf_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \geq \frac{d}{2}\log\frac{2\pi}{\beta}-\log Z.$$

Proof. By β-upper smoothness, for any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d$, we have

$$f(\boldsymbol{x}) \leq f(\boldsymbol{y})+\langle\nabla f(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle+\frac{\beta}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2,$$

which implies that, for any $\boldsymbol{y}\in\mathbb{R}^d$,

$$\begin{aligned}
Z &= \int_{\mathbb{R}^d}\exp\left(-f(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x} \geq \int_{\mathbb{R}^d}\exp\left(-f(\boldsymbol{y})-\langle\nabla f(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle-\frac{\beta}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2\right)\mathrm{d}\boldsymbol{x} \\
&= \exp\left(-f(\boldsymbol{y})+\frac{\|\nabla f(\boldsymbol{y})\|^2}{2\beta}\right)\int_{\mathbb{R}^d}\exp\left(-\frac{\beta}{2}\left\|\boldsymbol{x}-\boldsymbol{y}+\frac{\nabla f(\boldsymbol{y})}{\beta}\right\|^2\right)\mathrm{d}\boldsymbol{x} \\
&= \exp\left(-f(\boldsymbol{y})+\frac{\|\nabla f(\boldsymbol{y})\|^2}{2\beta}\right)\sqrt{(2\pi/\beta)^d}.
\end{aligned}$$

Taking logarithms on both sides and rearranging terms completes the proof. □

Finally, we state a well-known result in the optimization literature (Nesterov, 2013).

Lemma 12. *Given a function $f:\mathbb{R}^d\to\mathbb{R}$ that is differentiable, β-upper smooth (cf. Definition 5), and lower bounded, i.e., $f_\star \triangleq \inf_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) > -\infty$, we have*

$$\|\nabla f(\boldsymbol{x})\|^2 \leq 2\beta\left(f(\boldsymbol{x})-f_\star\right), \forall \boldsymbol{x}\in\mathbb{R}^d.$$

Proof. By lower-boundedness and β-upper smoothness, we have

$$f_\star \leq f(\boldsymbol{y}) \leq f(\boldsymbol{x})+\langle\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle+\frac{\beta}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2, \forall \boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d.$$

Fix $\boldsymbol{x}\in\mathbb{R}^d$ and take $\boldsymbol{y}=\boldsymbol{x}-\frac{\nabla f(\boldsymbol{x})}{\beta}$ to obtain that

$$f_\star \leq f(\boldsymbol{x})-\frac{\|\nabla f(\boldsymbol{x})\|^2}{2\beta} \Rightarrow \|\nabla f(\boldsymbol{x})\|^2 \leq 2\beta\left(f(\boldsymbol{x})-f_\star\right).$$

□

C MISSING PROOFS FOR MEAN-FIELD SVGD

We provide the omitted proofs for Mean-Field SVGD in this section. In particular, we discuss our proof strategy in Subsection C.4.

C.1 FULL VERSION OF THEOREM 1 AND ITS PROOF.

Theorem 5 (Full version of Theorem 1)**.** *Under Assumption 1, suppose the stepsize* $\{\eta_t\}_{t=1}^T$ *satisfies*

$$\eta_t \leq \min\left\{\frac{2}{L}, \frac{1}{2K_1\left[\sqrt{2L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}+4K_1\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})\right]}\right\}, \forall t\in[T],$$

where $L = K_0^2\beta + 2K_1^2$ *is defined in Lemma 14, then Mean-Field SVGD (Algorithm 1) guarantees that*

$$\mathsf{KSD}^2_{\mathcal{H}}(\bar{\mathsf{P}}_T\|\mathsf{Q}) \leq \frac{\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{\sum_{t=1}^T \eta_t - \frac{\eta_t^2 L}{2}} \quad \textit{where} \quad \bar{\mathsf{P}}_T \triangleq \frac{\sum_{t=1}^T\left(\eta_t - \frac{\eta_t^2 L}{2}\right)\mathsf{P}_t}{\sum_{t=1}^T \eta_t - \frac{\eta_t^2 L}{2}}.$$

In particular, when

$$\eta_t = \eta_\star \triangleq \min\left\{\frac{1}{L}, \frac{1}{2K_1\left[\sqrt{2L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}+4K_1\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})\right]}\right\}, \forall t\in[T],$$

we have

$$\mathsf{KSD}^2_{\mathcal{H}}(\bar{\mathsf{P}}_T\|\mathsf{Q}) \leq \frac{24K_1^2\mathsf{KL}^2(\mathsf{P}_1\|\mathsf{Q}) + 3L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{T}.$$

Proof. By (38), for all t in $[T]$,

$$\mathsf{KL}(\mathsf{P}_{t+1}\|\mathsf{Q}) \leq \mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) - \left(\eta_t - \frac{\eta_t^2 L}{2}\right)\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|^2_{\mathcal{H}^d} = \mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) - \left(\eta_t - \frac{\eta_t^2 L}{2}\right)\mathsf{KSD}^2_{\mathcal{H}}(\mathsf{P}_t\|\mathsf{Q}),$$

which implies that

$$\sum_{t=1}^T\left(\eta_t - \frac{\eta_t^2 L}{2}\right)\mathsf{KSD}^2_{\mathcal{H}}(\mathsf{P}_t\|\mathsf{Q}) \leq \mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}) - \mathsf{KL}(\mathsf{P}_{T+1}\|\mathsf{Q}) \leq \mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}).$$

Note that $\eta_t < \frac{2}{L} \Rightarrow \eta_t - \frac{\eta_t^2 L}{2} > 0, \forall t \in [T]$. Therefore, by convexity of $\mathsf{P} \mapsto \mathsf{KSD}^2_{\mathcal{H}}(\mathsf{P}\|\mathsf{Q})$ and the definition of $\bar{\mathsf{P}}_T$, we obtain that

$$\mathsf{KSD}^2_{\mathcal{H}}(\bar{\mathsf{P}}_T\|\mathsf{Q}) \leq \frac{\sum_{t=1}^T\left(\eta_t - \frac{\eta_t^2 L}{2}\right)\mathsf{KSD}^2_{\mathcal{H}}(\mathsf{P}_t\|\mathsf{Q})}{\sum_{t=1}^T \eta_t - \frac{\eta_t^2 L}{2}} \leq \frac{\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{\sum_{t=1}^T \eta_t - \frac{\eta_t^2 L}{2}}.$$

In particular, when $\eta_t = \eta_\star, \forall t \in [T]$, we have

$$\begin{aligned}
\mathsf{KSD}^2_{\mathcal{H}}(\bar{\mathsf{P}}_T\|\mathsf{Q}) &\leq \frac{\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{\left(\eta_\star - \frac{\eta_\star^2 L}{2}\right)T} \overset{\eta_\star \leq \frac{1}{L}}{\leq} \frac{2\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{\eta_\star T} \\
&\leq \frac{2\left[L + 2K_1\sqrt{2L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})} + 8K_1^2\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})\right]\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{T} \\
&\leq \frac{24K_1^2\mathsf{KL}^2(\mathsf{P}_1\|\mathsf{Q}) + 3L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{T}.
\end{aligned}$$

□

C.2 Full Version of Theorem 2 and Its Proof.

Theorem 6 (Full version of Theorem 2)**.** *Under Assumption 1, suppose the stepsize* $\{\eta_t\}_{t=1}^T$ *satisfies*

$$\eta_t = \frac{\gamma_t}{\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d}}, \gamma_t \leq \frac{1}{2K_1}, \forall t \in [T]\,,$$

then Mean-Field SVGD (Algorithm 1) guarantees that, for $L = K_0^2\beta + 2K_1^2$ *defined in Lemma 14,*

$$\mathsf{KSD}^2_{\mathcal{H}}(\bar{\mathsf{P}}_T\|\mathsf{Q}) \leq \left(\frac{\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{\sum_{t=1}^T \gamma_t} + \frac{\sum_{t=1}^T \gamma_t^2 L}{2\sum_{t=1}^T \gamma_t}\right)^2 \quad \textit{where} \quad \bar{\mathsf{P}}_T \triangleq \frac{\sum_{t=1}^T \gamma_t \mathsf{P}_t}{\sum_{t=1}^T \gamma_t}.$$

In particular, when

$$\gamma_t = \gamma_\star \triangleq \min\left\{\frac{1}{2K_1}, \sqrt{\frac{2\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{LT}}\right\}, \forall t \in [T]\,,$$

we have

$$\mathsf{KSD}^2_{\mathcal{H}}(\bar{\mathsf{P}}_T\|\mathsf{Q}) \leq \frac{8K_1^2\mathsf{KL}^2(\mathsf{P}_1\|\mathsf{Q})}{T^2} + \frac{4L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{T}.$$

Remark 9. For the finite-particle implementation, one can approximate $\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d}$ by

$$\begin{aligned}
&\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\hat{\mathsf{P}}_t\right]\right\|_{\mathcal{H}^d} = \left\|\frac{1}{N}\sum_{i=1}^N K(\cdot, \boldsymbol{x}_i(t))\nabla V(\boldsymbol{x}_i(t)) - \nabla_2 K(\cdot, \boldsymbol{x}_i(t))\right\|_{\mathcal{H}^d} \\
=&\frac{1}{N}\sqrt{\sum_{i,j=1}^N \langle K(\cdot, \boldsymbol{x}_i(t))\nabla V(\boldsymbol{x}_i(t)) - \nabla_2 K(\cdot, \boldsymbol{x}_i(t)), K(\cdot, \boldsymbol{x}_j(t))\nabla V(\boldsymbol{x}_j(t)) - \nabla_2 K(\cdot, \boldsymbol{x}_j(t))\rangle_{\mathcal{H}^d}} \\
=&\frac{1}{N}\sqrt{\sum_{i,j=1}^N A_{i,j} - B_{i,j} - C_{i,j} + D_{i,j}},
\end{aligned}$$

where

$$\begin{aligned}
A_{i,j} &\triangleq K(\boldsymbol{x}_i(t), \boldsymbol{x}_j(t))\left\langle \nabla V(\boldsymbol{x}_i(t)), \nabla V(\boldsymbol{x}_j(t))\right\rangle, \qquad B_{i,j} \triangleq \left\langle \nabla V(\boldsymbol{x}_i(t)), \nabla_2 K(\boldsymbol{x}_i(t), \boldsymbol{x}_j(t))\right\rangle, \\
C_{i,j} &\triangleq \left\langle \nabla V(\boldsymbol{x}_j(t)), \nabla_1 K(\boldsymbol{x}_i(t), \boldsymbol{x}_j(t))\right\rangle, \qquad D_{i,j} \triangleq \mathrm{Tr}\left(\nabla_1\nabla_2 K(\boldsymbol{x}_i(t), \boldsymbol{x}_j(t))\right).
\end{aligned}$$

Proof. By (40), we have

$$\begin{aligned}
\mathsf{KL}(\mathsf{P}_{t+1}\|\mathsf{Q}) &\leq \mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) - \gamma_t\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d} + \frac{\gamma_t^2 L}{2} \\
&= \mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) - \gamma_t\mathsf{KSD}_{\mathcal{H}}(\mathsf{P}_t\|\mathsf{Q}) + \frac{\gamma_t^2 L}{2},
\end{aligned}$$

which implies that

$$\frac{\sum_{t=1}^T \gamma_t \mathsf{KSD}_{\mathcal{H}}(\mathsf{P}_t\|\mathsf{Q})}{\sum_{t=1}^T \gamma_t} \leq \frac{\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{\sum_{t=1}^T \gamma_t} + \frac{\sum_{t=1}^T \gamma_t^2 L}{2\sum_{t=1}^T \gamma_t} \Rightarrow \mathsf{KSD}^2_{\mathcal{H}}(\bar{\mathsf{P}}_T\|\mathsf{Q}) \leq \left(\frac{\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{\sum_{t=1}^T \gamma_t} + \frac{\sum_{t=1}^T \gamma_t^2 L}{2\sum_{t=1}^T \gamma_t}\right)^2.$$

In particular, when $\gamma_t = \gamma_\star, \forall t \in [T]$, we have

$$\mathsf{KSD}^2_{\mathcal{H}}(\bar{\mathsf{P}}_T\|\mathsf{Q}) \leq \left(\frac{2K_1\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{T} + \sqrt{\frac{2L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{T}}\right)^2 \leq \frac{8K_1^2\mathsf{KL}^2(\mathsf{P}_1\|\mathsf{Q})}{T^2} + \frac{4L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})}{T}.$$

□

C.3 Convergence under Stein-log-Sobolev Inequality

In this subsection, we show that Mean-Field SVGD converges in KL divergence under the Stein-log-Sobolev inequality (Duncan et al., 2023; Carrillo et al., 2024).

Definition 6 (Stein-log-Sobolev inequality)**.** Given an RKHS $(\mathcal{H}, K, \mathbb{R}^d)$ and a probability measure Q on $\mathbb{R}^d$, we say that they together satisfy the Stein-log-Sobolev Inequality (SLSI) if there exists $\mu > 0$ such that

$$\mathsf{KL}(\mathsf{P}\|\mathsf{Q}) \leq \frac{\mathsf{KSD}^2_{\mathcal{H}}(\mathsf{P}\|\mathsf{Q})}{2\mu}, \forall \mathsf{P} \in \mathcal{P}(\mathbb{R}^d) \text{ such that } \mathsf{KL}(\mathsf{P}\|\mathsf{Q}) < +\infty,$$

where $\mathcal{P}(\mathbb{R}^d)$ denotes the set of all probability measures on $\mathbb{R}^d$.

Remark 10. We omit the discussion of conditions ensuring the SLSI due to its technical complexity, and refer interested readers to, for example, Carrillo et al. (2024, Theorem 1.1) for details.

Theorem 7. *Under Assumption 1 and SLSI (cf. Definition 6), suppose the stepsize $\{\eta_t\}_{t=1}^T$ satisfies*

$$\eta_t \leq \min\left\{\frac{2}{L}, \frac{1}{2K_1\left[\sqrt{2L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})} + 4K_1\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})\right]}\right\}, \forall t \in [T],$$

where $L = K_0^2\beta + 2K_1^2$ is defined in Lemma 14, then Mean-Field SVGD (Algorithm 1) guarantees that

$$\mathsf{KL}(\mathsf{P}_{T+1}\|\mathsf{Q}) \leq \exp\left(-2\mu\sum_{t=1}^T \eta_t - \frac{\eta_t^2 L}{2}\right)\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}).$$

In particular, when

$$\eta_t = \eta_\star \triangleq \min\left\{\frac{1}{L}, \frac{1}{2K_1\left[\sqrt{2L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})} + 4K_1\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})\right]}\right\}, \forall t \in [T],$$

we have

$$\mathsf{KL}(\mathsf{P}_{T+1}\|\mathsf{Q}) \leq \exp\left(-\mu\eta_\star T\right)\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}).$$

Proof. By (38), for all t in $[T]$,

$$\mathsf{KL}(\mathsf{P}_{t+1}\|\mathsf{Q}) \leq \mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) - \left(\eta_t - \frac{\eta_t^2 L}{2}\right)\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|^2_{\mathcal{H}^d} = \mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) - \left(\eta_t - \frac{\eta_t^2 L}{2}\right)\mathsf{KSD}^2_{\mathcal{H}}(\mathsf{P}_t\|\mathsf{Q}),$$

which implies that, under SLSI (cf. Definition 6),

$$\mathsf{KL}(\mathsf{P}_{t+1}\|\mathsf{Q}) \leq \left[1 - 2\mu\left(\eta_t - \frac{\eta_t^2 L}{2}\right)\right]\mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) \leq \exp\left(-2\mu\left(\eta_t - \frac{\eta_t^2 L}{2}\right)\right)\mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}).$$

We unroll the recursion to obtain that

$$\mathsf{KL}(\mathsf{P}_{T+1}\|\mathsf{Q}) \leq \exp\left(-2\mu\sum_{t=1}^T \eta_t - \frac{\eta_t^2 L}{2}\right)\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}).$$

In particular, when $\eta_t = \eta_\star, \forall t \in [T]$, we have

$$\mathsf{KL}(\mathsf{P}_{T+1}\|\mathsf{Q}) \leq \exp\left(-2\mu\left(\eta_\star - \frac{\eta_\star^2 L}{2}\right)T\right)\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}) \overset{\eta_\star \leq \frac{1}{L}}{\leq} \exp\left(-\mu\eta_\star T\right)\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}).$$

□

C.4 Theoretical Analysis of Mean-Field SVGD

We provide the detailed theoretical analysis of Mean-Field SVGD in this subsection. Our analysis consists of three main ingredients, of which the first and third contain the main novel ideas.

First ingredient. We begin by highlighting a technical point that is often left implicit in existing analyses (Korba et al., 2020; Salim et al., 2022; Sun et al., 2023; Fujisawa & Futami, 2025): the well-posedness of the representation

$$\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right](\cdot) = \mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[K(\cdot,\boldsymbol{x})\nabla V(\boldsymbol{x}) - \nabla_2 K(\cdot,\boldsymbol{x})\right]$$

used in Mean-Field SVGD (Algorithm 1). In particular, the formal representation alone does not guarantee that $\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right] \in \mathcal{H}^d$. As the starting point, we therefore need to first establish this well-posedness rigorously under our assumptions. By Lemma 5, under Assumption 1.3, it suffices to verify $\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla V(\boldsymbol{x})\|\right] < +\infty$.

However, this integrability condition need not hold for an arbitrary pair (P, V), or equivalently, (P, Q). Our strategy is therefore to control it through the KL divergence, which motivates one of our main technical results, Lemma 10, stated previously in Appendix B. Specifically, under β-upper smooth V (cf. Assumption 1.1), Lemma 10 gives

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla V(\boldsymbol{x})\|\right] \le \sqrt{d\beta} + \sqrt{2\beta\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}.$$

In other words,

$$\mathsf{KL}(\mathsf{P}\|\mathsf{Q}) < +\infty \Rightarrow \mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla V(\boldsymbol{x})\|\right] < +\infty \Rightarrow \boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right] \in \mathcal{H}^d.$$

For readability, we state Lemma 13 below as a simplified version of Lemma 10.

Lemma 13. *Under Assumption 1.1, let* P *be a probability measure on* $\mathbb{R}^d$ *such that* $\mathsf{KL}(\mathsf{P}\|\mathsf{Q}) < +\infty$*, then we have* $\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla V(\boldsymbol{x})\|\right] < +\infty$.

Proof. See Lemma 10 for the quantitative version. □

Consequently, to establish that $\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right] \in \mathcal{H}^d$ for all $t \in [T]$, it is enough to show $\mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) < +\infty, \forall t \in [T]$.

Second ingredient. Based on the above discussion, we next study how the KL divergence changes under the pushforward map $\boldsymbol{I} - \eta\boldsymbol{f}$, where $\boldsymbol{f} \in \mathcal{H}^d$. Since SVGD is designed as an iterative descent method to decrease the KL divergence, it is natural to expect the KL divergence to decrease along the updates. To formalize this intuition, we establish the following descent inequality, noting that related inequalities have appeared in the literature:

Lemma 14. *Under Assumptions 1.1 and 1.3, let* P *be a probability measure on* $\mathbb{R}^d$ *such that* $\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla V(\boldsymbol{x})\|\right] < +\infty$*, then for any* $\boldsymbol{f} \in \mathcal{H}^d$ *and* $|\eta| \le \frac{1}{2K_1\|\boldsymbol{f}\|_{\mathcal{H}^d}}$*, we have*

$$\mathsf{KL}((\boldsymbol{I} - \eta\boldsymbol{f})_{\#}\mathsf{P}\|\mathsf{Q}) \le \mathsf{KL}(\mathsf{P}\|\mathsf{Q}) - \eta\left\langle \boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right], \boldsymbol{f}\right\rangle_{\mathcal{H}^d} + \frac{\eta^2 L}{2}\left\|\boldsymbol{f}\right\|_{\mathcal{H}^d}^2,$$

where $L \triangleq K_0^2\beta + 2K_1^2$.

Remark 11. The condition $|\eta| \le \frac{1}{2K_1\|\boldsymbol{f}\|_{\mathcal{H}^d}}$ can be relaxed to $|\eta| \le \frac{\alpha}{K_1\|\boldsymbol{f}\|_{\mathcal{H}^d}}$ for $\alpha \in (0,1)$, under which L will be changed to $K_0^2\beta + \frac{K_1^2}{1-\alpha}$.

Proof. We defer the proof to Appendix C.4.1. □

Substituting $\mathsf{P} = \mathsf{P}_t$ and $\boldsymbol{f} = \boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]$ in Lemma 14, and assuming $\mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) < +\infty$, we obtain

$$\eta_t \le \min\left\{\frac{2}{L}, \frac{1}{2K_1\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d}}\right\} \Rightarrow \mathsf{KL}(\mathsf{P}_{t+1}\|\mathsf{Q}) \le \mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}).$$

Therefore, if $\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}) < +\infty$ (i.e., Assumption 1.2), an induction gives $\mathsf{KL}(\mathsf{P}_{t+1}\|\mathsf{Q}) \le \mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) \le \cdots \le \mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}) < +\infty$, implying that $\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}_{t+1}}\left[\|\nabla V(\boldsymbol{x})\|\right] < +\infty$ and $\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_{t+1}\right] \in \mathcal{H}^d$. Hence, this argument can be applied recursively to control the KL divergence along the entire trajectory.

Third ingredient. The final barrier is to eliminate the dependence on $\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d}$ when using a nonadaptive stepsize. To this end, we establish the following inequality in Lemma 15, which, to the best of our knowledge, is new. The proof of Lemma 15 is inspired by the cocoercivity of smooth convex functions (Nesterov, 2013).

Lemma 15. *Under Assumptions 1.1 and 1.3, suppose $\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla V(\boldsymbol{x})\|\right] < +\infty$, then we have*

$$\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right]\right\|_{\mathcal{H}^d} \leq \sqrt{2L\mathsf{KL}(\mathsf{P}\|\mathsf{Q})} + 4K_1\mathsf{KL}(\mathsf{P}\|\mathsf{Q}),$$

where $L = K_0^2\beta + 2K_1^2$ is defined in Lemma 14.

Proof. We defer the proof to Appendix C.4.2. □

Therefore, we know $\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d} \leq \sqrt{2L\mathsf{KL}(\mathsf{P}_t\|\mathsf{Q})} + 4K_1\mathsf{KL}(\mathsf{P}_t\|\mathsf{Q})$, which can be further upper bounded by $\mathrm{poly}(\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}))$ by our previous discussion.

Analysis for the nonadaptive stepsize. Combining the above three ingredients, we obtain the following key lemma, which plays a central role in the proofs of Theorems 5 and 7.

Lemma 16. *Under Assumption 1, suppose the stepsize $\{\eta_t\}_{t=1}^T$ satisfies*

$$\eta_t \leq \min\left\{\frac{2}{L}, \frac{1}{2K_1\left[\sqrt{2L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})} + 4K_1\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})\right]}\right\}, \forall t \in [T],$$

where $L = K_0^2\beta + 2K_1^2$ is defined in Lemma 14, then Mean-Field SVGD (Algorithm 1) is well-defined in the following sense:

1. $\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right] \in \mathcal{H}^d, \forall t \in [T]$.
2. $\mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) < +\infty, \forall t \in [T+1]$.

Proof. We first note that $\mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) < +\infty \overset{\text{Lemma 13}}{\Rightarrow} \mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}_t}\left[\|\nabla V(\boldsymbol{x})\|\right] < +\infty$, which further implies that $\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right] \in \mathcal{H}^d$ by Lemma 5. Therefore, it suffices to prove only the second point.

To do so, we use induction to prove the following proposition:

$$\mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) \leq \cdots \leq \mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}) < +\infty, \forall t \in [T+1]. \tag{37}$$

- When $t = 1$, we have $\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}) < +\infty$ by Assumption 1.2.
- Suppose (37) holds for the index t, where $t \in [T]$.
- At time $t+1$, by the induction hypothesis, it suffices to prove

  $$\mathsf{KL}(\mathsf{P}_{t+1}\|\mathsf{Q}) \leq \mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}).$$

  By the induction hypothesis, $\mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) \leq \mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}) < +\infty$, which implies that $\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}_t}\left[\|\nabla V(\boldsymbol{x})\|\right] < +\infty$ and $\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right] \in \mathcal{H}^d$ as mentioned above. Therefore,

  $$\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d} \overset{\text{Lemma 15}}{\leq} \sqrt{2L\mathsf{KL}(\mathsf{P}_t\|\mathsf{Q})} + 4K_1\mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) \leq \sqrt{2L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})} + 4K_1\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q}).$$

  Next, by the condition on the stepsize, there is

  $$\eta_t \leq \frac{1}{2K_1\left[\sqrt{2L\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})} + 4K_1\mathsf{KL}(\mathsf{P}_1\|\mathsf{Q})\right]} \leq \frac{1}{2K_1\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d}}.$$

  Thus,

  $$\begin{aligned}\mathsf{KL}(\mathsf{P}_{t+1}\|\mathsf{Q}) &= \mathsf{KL}((\boldsymbol{I} - \eta_t\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right])_{\#}\mathsf{P}_t\|\mathsf{Q}) \\ &\overset{\text{Lemma 14}}{\leq} \mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}) - \left(\eta_t - \frac{\eta_t^2 L}{2}\right)\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d}^2 \leq \mathsf{KL}(\mathsf{P}_t\|\mathsf{Q}),\end{aligned} \tag{38}$$

  where the last step is due to $\eta_t \leq \frac{2}{L} \Rightarrow \eta_t - \frac{\eta_t^2 L}{2} \geq 0$. So we know (37) holds for time $t+1$.

By induction, we conclude the second point. □

Analysis for the adaptive stepsize. The analysis for the adaptive stepsize is relatively simpler, as we only need the first two ingredients discussed above. Analogous to Lemma 16, we establish the following Lemma 17, which plays a key role in the proof of Theorem 6.

Lemma 17. *Under Assumption 1, suppose the stepsize* $\{\eta_t\}_{t=1}^T$ *satisfies*

$$\eta_t = \frac{\gamma_t}{\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d}}, \gamma_t \leq \frac{1}{2K_1}, \forall t \in [T]\,,$$

then Mean-Field SVGD (Algorithm 1) is well-defined in the following sense:

1. $\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right] \in \mathcal{H}^d, \forall t \in [T]\,.$
2. $\mathsf{KL}(\mathsf{P}_t \| \mathsf{Q}) < +\infty, \forall t \in [T+1].$

Proof. Same as the proof of Lemma 16. It suffices to show $\mathsf{KL}(\mathsf{P}_t \| \mathsf{Q}) < +\infty, \forall t \in [T+1]$.

To do so, we use induction to prove the following proposition:

$$\mathsf{KL}(\mathsf{P}_t \| \mathsf{Q}) < +\infty, \forall t \in [T+1]\,. \tag{39}$$

When $t = 1$, $\mathsf{KL}(\mathsf{P}_1 \| \mathsf{Q}) < +\infty$ holds due to Assumption 1.2. Now, suppose (39) holds for the index t, where $t \in [T]$. Then, we have $\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}_t}\left[\|\nabla V(\boldsymbol{x})\|\right] < +\infty$ and $\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right] \in \mathcal{H}^d$. At time $t+1$, by the condition on the stepsize, there is

$$|\eta_t| = \frac{\gamma_t}{\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d}} \leq \frac{1}{2K_1\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d}}.$$

Thus,

$$\mathsf{KL}(\mathsf{P}_{t+1} \| \mathsf{Q}) = \mathsf{KL}\left(\left(\boldsymbol{I} - \eta_t \boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right)_{\#}\mathsf{P}_t \| \mathsf{Q}\right) \overset{\text{Lemma 14}}{\leq} \mathsf{KL}(\mathsf{P}_t \| \mathsf{Q}) - \gamma_t\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}_t\right]\right\|_{\mathcal{H}^d} + \frac{\gamma_t^2 L}{2} < +\infty. \tag{40}$$

So, we know (39) holds for time $t+1$. By induction, we conclude the second point. □

C.4.1 Proof of Lemma 14

Proof. First, $\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\|\nabla V(\boldsymbol{x})\|\right] < +\infty$ implies that $\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right] \in \mathcal{H}^d$ is well-defined by Lemma 5. Moreover, w.l.o.g., we can assume $\mathsf{KL}(\mathsf{P} \| \mathsf{Q}) < +\infty$ (which also implies that $\mathsf{P} \ll \mathsf{Q}$); otherwise, the inequality holds automatically.

Under Assumption 1.3, by Lemma 3, for any $\boldsymbol{f} \in \mathcal{H}^d$ and $\boldsymbol{x} \in \mathbb{R}^d$,

$$\|\boldsymbol{f}(\boldsymbol{x})\| \leq K_0\|\boldsymbol{f}\|_{\mathcal{H}^d} \quad \text{and} \quad \|\boldsymbol{J}\left[\boldsymbol{f}\right](\boldsymbol{x})\|_{\mathrm{F}} \leq K_1\|\boldsymbol{f}\|_{\mathcal{H}^d}\,. \tag{41}$$

Therefore, the condition $|\eta| \leq \frac{1}{2K_1\|\boldsymbol{f}\|_{\mathcal{H}^d}}$ implies that

$$|\eta|\,\|\boldsymbol{J}\left[\boldsymbol{f}\right](\boldsymbol{x})\|_{\mathrm{op}} \leq |\eta|\,\|\boldsymbol{J}\left[\boldsymbol{f}\right](\boldsymbol{x})\|_{\mathrm{F}} \leq \frac{1}{2} < 1, \forall \boldsymbol{x} \in \mathbb{R}^d, \tag{42}$$

which further indicates that $\|(\boldsymbol{I} - \eta\boldsymbol{f})(\boldsymbol{x}) - (\boldsymbol{I} - \eta\boldsymbol{f})(\boldsymbol{y})\| \geq \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.

Therefore, by Lemma 6, we have

$$\begin{aligned}
&\mathsf{KL}((\boldsymbol{I} - \eta\boldsymbol{f})_{\#}\mathsf{P} \| \mathsf{Q}) - \mathsf{KL}(\mathsf{P} \| \mathsf{Q}) \\
=&\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[-\log q(\boldsymbol{x} - \eta\boldsymbol{f}(\boldsymbol{x})) + \log q(\boldsymbol{x}) - \log\left|\det\left(\boldsymbol{I} - \eta\boldsymbol{J}\left[\boldsymbol{f}\right](\boldsymbol{x})\right)\right|\right] \\
\overset{(1)}{=}&\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[V(\boldsymbol{x} - \eta\boldsymbol{f}(\boldsymbol{x})) - V(\boldsymbol{x}) - \log\left|\det\left(\boldsymbol{I} - \eta\boldsymbol{J}\left[\boldsymbol{f}\right](\boldsymbol{x})\right)\right|\right] \\
\overset{(a)}{\leq}&\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[-\eta\left\langle\nabla V(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x})\right\rangle + \frac{\eta^2\beta}{2}\|\boldsymbol{f}(\boldsymbol{x})\|^2 - \log\left|\det\left(\boldsymbol{I} - \eta\boldsymbol{J}\left[\boldsymbol{f}\right](\boldsymbol{x})\right)\right|\right] \\
\overset{(b)}{\leq}&\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[-\eta\left\langle\nabla V(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x})\right\rangle + \frac{\eta^2\beta}{2}\|\boldsymbol{f}(\boldsymbol{x})\|^2 + \eta\mathrm{Tr}\left(\boldsymbol{J}\left[\boldsymbol{f}\right](\boldsymbol{x})\right) + \eta^2\|\boldsymbol{J}\left[\boldsymbol{f}\right](\boldsymbol{x})\|_{\mathrm{F}}^2\right] \\
\overset{(c)}{=}&-\eta\left\langle\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right], \boldsymbol{f}\right\rangle + \eta^2\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}\left[\frac{\beta}{2}\|\boldsymbol{f}(\boldsymbol{x})\|^2 + \|\boldsymbol{J}\left[\boldsymbol{f}\right](\boldsymbol{x})\|_{\mathrm{F}}^2\right] \\
\overset{(41)}{\leq}&-\eta\left\langle\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right], \boldsymbol{f}\right\rangle + \frac{\eta^2\left(K_0^2\beta + 2K_1^2\right)}{2}\|\boldsymbol{f}\|_{\mathcal{H}^d}^2 = -\eta\left\langle\boldsymbol{g}_{\mathsf{Q}}\left[\mathsf{P}\right], \boldsymbol{f}\right\rangle + \frac{\eta^2 L}{2}\|\boldsymbol{f}\|_{\mathcal{H}^d}^2\,,
\end{aligned}$$

where (a) is due to β-upper smoothness of V (cf. Assumption 1.1), (b) is by, for $\boldsymbol{A}(\boldsymbol{x}) \triangleq \eta \boldsymbol{J}[\boldsymbol{f}](\boldsymbol{x})$ (note that $\|\boldsymbol{A}(\boldsymbol{x})\|_{\mathrm{op}} \overset{(42)}{\leq} \frac{1}{2}, \forall \boldsymbol{x} \in \mathbb{R}^d$),

$$-\log|\det(\boldsymbol{I}-\boldsymbol{A}(\boldsymbol{x}))| \overset{\text{Lemma 7}}{\leq} \mathrm{Tr}(\boldsymbol{A}(\boldsymbol{x})) + \frac{\|\boldsymbol{A}(\boldsymbol{x})\|_{\mathrm{F}}^2}{2(1-\|\boldsymbol{A}(\boldsymbol{x})\|_{\mathrm{op}})} \leq \mathrm{Tr}(\boldsymbol{A}(\boldsymbol{x})) + \|\boldsymbol{A}(\boldsymbol{x})\|_{\mathrm{F}}^2, \forall \boldsymbol{x} \in \mathbb{R}^d,$$

and (c) follows from

$$-\eta\langle\nabla V(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x})\rangle + \eta\mathrm{Tr}(\boldsymbol{J}[\boldsymbol{f}](\boldsymbol{x})) = \eta(\langle\nabla \log q(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x})\rangle + \langle\nabla, \boldsymbol{f}(\boldsymbol{x})\rangle) = \eta\mathcal{T}_{\mathsf{Q}}[\boldsymbol{f}](\boldsymbol{x})$$
$$\Rightarrow \mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}[-\eta\langle\nabla V(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x})\rangle + \eta\mathrm{Tr}(\boldsymbol{J}[\boldsymbol{f}](\boldsymbol{x}))] = \eta\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}[\mathcal{T}_{\mathsf{Q}}[\boldsymbol{f}](\boldsymbol{x})] \overset{\text{Lemma 5}}{=} -\eta\langle \boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}], \boldsymbol{f}\rangle_{\mathcal{H}^d}.$$

□

C.4.2 Proof of Lemma 15

Proof. First, $\mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}[\|\nabla V(\boldsymbol{x})\|] < +\infty$ implies that $\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}] \in \mathcal{H}^d$ is well-defined by Lemma 5. Moreover, w.l.o.g., we can assume $\mathsf{KL}(\mathsf{P}\|\mathsf{Q}) < +\infty$ (which also implies that $\mathsf{P} \ll \mathsf{Q}$); otherwise, the inequality holds automatically.

By Lemma 14, for any $\|\boldsymbol{f}\|_{\mathcal{H}^d} \leq \frac{1}{2K_1}$,

$$0 \leq \mathsf{KL}((\boldsymbol{I}-\boldsymbol{f})_{\#}\mathsf{P}\|\mathsf{Q}) \leq \mathsf{KL}(\mathsf{P}\|\mathsf{Q}) - \langle \boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}], \boldsymbol{f}\rangle_{\mathcal{H}^d} + \frac{L}{2}\|\boldsymbol{f}\|_{\mathcal{H}^d}^2.$$

Now, we take $\boldsymbol{f} = \frac{\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]}{L+2K_1\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d}}$ in the above inequality to have

$$\begin{aligned}
0 &\leq \mathsf{KL}(\mathsf{P}\|\mathsf{Q}) - \frac{\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d}^2}{L+2K_1\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d}} + \frac{L\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d}^2}{2\left(L+2K_1\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d}\right)^2} \\
&\leq \mathsf{KL}(\mathsf{P}\|\mathsf{Q}) - \frac{\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d}^2}{L+2K_1\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d}} + \frac{\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d}^2}{2\left(L+2K_1\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d}\right)} \\
&= \mathsf{KL}(\mathsf{P}\|\mathsf{Q}) - \frac{\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d}^2}{2\left(L+2K_1\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d}\right)}.
\end{aligned}$$

Rearrange terms to obtain that

$$\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d}^2 - 4K_1\mathsf{KL}(\mathsf{P}\|\mathsf{Q})\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d} - 2L\mathsf{KL}(\mathsf{P}\|\mathsf{Q})] \leq 0,$$

which implies that

$$\begin{aligned}
\|\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}]\|_{\mathcal{H}^d} &\leq \frac{4K_1\mathsf{KL}(\mathsf{P}\|\mathsf{Q}) + \sqrt{16K_1^2\mathsf{KL}^2(\mathsf{P}\|\mathsf{Q}) + 8L\mathsf{KL}(\mathsf{P}\|\mathsf{Q})}}{2} \\
&\leq \sqrt{2L\mathsf{KL}(\mathsf{P}\|\mathsf{Q})} + 4K_1\mathsf{KL}(\mathsf{P}\|\mathsf{Q}).
\end{aligned}$$

□

D Missing Proofs for Finite-Particle SVGD

D.1 Full Version of Theorem 3 and Its Proof.

Theorem 8 (Full version of Theorem 3)**.** *Under Assumption 2 (with Assumption 2.4), suppose the stepsize $\{\eta_t\}_{t=1}^T$ satisfies*

$$\eta_t \leq \frac{1}{2\sqrt{5K_0^4\beta^2 + 5K_1^4 + 5K_2^2 + 20NK_0^2K_1^2\beta(V^\star - V_\star) + 10NK_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s}}, \forall t \in [T],$$

then Finite-Particle SVGD (Algorithm 2) guarantees that

$$\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}\left(\frac{\sum_{t=1}^T \eta_t \hat{\mathsf{P}}_t}{\sum_{t=1}^T \eta_t}\middle\|\mathsf{Q}\right)\right]$$
$$\leq \frac{2C_{\mathsf{KL}}d}{\sum_{t=1}^T \eta_t} + \frac{2C_\star d}{N} + \frac{12\sum_{t=1}^T \eta_t^2\left(dK_0^4\beta^2 + K_2^2 + 2K_0^2K_1^2\beta(V^\star - V_\star) + K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right)}{N\sum_{t=1}^T \eta_t}.$$

In particular, when

$$\eta_t = \eta_\star \triangleq \min\left\{\frac{1}{4\sqrt{5K_0^4\beta^2 + 5K_1^4 + 5K_2^2 + 20K_0^2K_1^2\beta\,(V^\star - V_\star)\,N}}, \frac{1}{2\sqrt[3]{20K_0^2K_1^4\beta NT}},\right.$$
$$\left.\sqrt{\frac{C_{\mathsf{KL}}dN}{(dK_0^4\beta^2 + K_2^2 + 2K_0^2K_1^2\beta(V^\star - V_\star))\,T}}, \sqrt[3]{\frac{C_{\mathsf{KL}}dN}{K_0^2K_1^4\beta T^2}}\right\}, \forall t \in [T], \quad (43)$$

we have

$$\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}\left(\frac{\sum_{t=1}^T \hat{\mathsf{P}}_t}{T}\middle\|\mathsf{Q}\right)\right]$$
$$\lesssim \frac{C_{\mathsf{KL}}d\sqrt{K_0^4\beta^2 + K_1^4 + K_2^2 + K_0^2K_1^2\beta\,(V^\star - V_\star)\,N}}{T} + \frac{C_{\mathsf{KL}}d\sqrt[3]{K_0^2K_1^4\beta N}}{T^{\frac{2}{3}}}$$
$$+ \sqrt{\frac{C_{\mathsf{KL}}d\,(dK_0^4\beta^2 + K_2^2 + K_0^2K_1^2\beta(V^\star - V_\star))}{NT}} + \sqrt[3]{\frac{C_{\mathsf{KL}}^2d^2K_0^2K_1^4\beta}{NT}} + \frac{C_\star d}{N}.$$

Remark 12. To recover Theorem 3, we take $T = N^2$ to obtain the $\mathcal{O}(1/N)$ rate.

Proof. We recall that $\hat{\mathsf{P}}_t = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{x}_i(t)}$ and $\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}](\cdot) = \mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}[K(\cdot,\boldsymbol{x})\nabla V(\boldsymbol{x}) - \nabla_2 K(\cdot,\boldsymbol{x})]$, and use the following equation to have

$$\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|^2_{\mathcal{H}^d}\right] = \mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}}(t))\right\|^2_{\mathcal{H}^d}\right] \overset{(45)}{=} \mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\hat{\mathsf{P}}_t\right]\right\|^2_{\mathcal{H}^d}\right]$$
$$\overset{\text{Lemma 5}}{=} \mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}(\hat{\mathsf{P}}_t\|\mathsf{Q})\right].$$

Therefore, by Lemma 21, for any $t \in [T]$,

$$\mathsf{KL}(\vec{\mathsf{P}}_{t+1}\|\mathsf{Q}^{\otimes N}) \leq \mathsf{KL}(\vec{\mathsf{P}}_t\|\mathsf{Q}^{\otimes N}) - \frac{\eta_t N}{2}\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}(\hat{\mathsf{P}}_t\|\mathsf{Q})\right] + \eta_t\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\sum_{i=1}^N\left\langle\nabla_{2,\boldsymbol{x}_i}, \boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle\right]$$
$$+ \eta_t^2\left(6dK_0^4\beta^2 + 6K_2^2 + 12K_0^2K_1^2\beta(V^\star - V_\star) + 6K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right)$$
$$\overset{\text{Lemma 18}}{\leq} \mathsf{KL}(\vec{\mathsf{P}}_t\|\mathsf{Q}^{\otimes N}) - \frac{\eta_t N}{2}\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}(\hat{\mathsf{P}}_t\|\mathsf{Q})\right] + \eta_t C_\star d$$
$$+ \eta_t^2\left(6dK_0^4\beta^2 + 6K_2^2 + 12K_0^2K_1^2\beta(V^\star - V_\star) + 6K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right).$$

Sum the above inequality over $t = 1$ to T, rearrange terms, drop the nonnegative term $\mathsf{KL}(\vec{\mathsf{P}}_{T+1}\|\mathsf{Q}^{\otimes N})$, divide both sides by $\sum_{t=1}^T \frac{\eta_t N}{2}$, and use the convexity of $\mathsf{P} \mapsto \mathsf{KSD}^2_{\mathcal{H}}(\mathsf{P}\|\mathsf{Q})$ to have

$$\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}\left(\frac{\sum_{t=1}^T \eta_t\hat{\mathsf{P}}_t}{\sum_{t=1}^T \eta_t}\middle\|\mathsf{Q}\right)\right] \leq \frac{\sum_{t=1}^T \eta_t\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}(\hat{\mathsf{P}}_t\|\mathsf{Q})\right]}{\sum_{t=1}^T \eta_t}$$
$$\leq \frac{2\mathsf{KL}(\vec{\mathsf{P}}_1\|\mathsf{Q}^{\otimes N})}{N\sum_{t=1}^T \eta_t} + \frac{2C_\star d}{N} + \frac{12\sum_{t=1}^T \eta_t^2\left(dK_0^4\beta^2 + K_2^2 + 2K_0^2K_1^2\beta(V^\star - V_\star) + K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right)}{N\sum_{t=1}^T \eta_t}$$
$$\leq \frac{2C_{\mathsf{KL}}d}{\sum_{t=1}^T \eta_t} + \frac{2C_\star d}{N} + \frac{12\sum_{t=1}^T \eta_t^2\left(dK_0^4\beta^2 + K_2^2 + 2K_0^2K_1^2\beta(V^\star - V_\star) + K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right)}{N\sum_{t=1}^T \eta_t},$$

where the last step is by $\mathsf{KL}(\vec{\mathsf{P}}_1\|\mathsf{Q}^{\otimes N}) \overset{\text{Assumption 2.2}}{\leq} C_{\mathsf{KL}}dN$.

Now, suppose $\eta_t = \eta_\star, \forall t \in [T]$ for some $\eta_\star > 0$, we can simplify the bound into

$$\begin{aligned}&\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}\left(\frac{\sum_{t=1}^T \hat{\mathsf{P}}_t}{T}\middle\|\mathsf{Q}\right)\right]\\ \leq&\frac{2C_{\mathsf{KL}}d}{\eta_\star T}+\frac{2C_\star d}{N}+\frac{12\left(dK_0^4\beta^2+K_2^2+2K_0^2K_1^2\beta(V^\star-V_\star)\right)\eta_\star}{N}+\frac{6K_0^2K_1^4\beta\eta_\star^2 T}{N}.\end{aligned}$$

In particular, when $\eta_\star$ equals (43), we have

$$\begin{aligned}&\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}\left(\frac{\sum_{t=1}^T \hat{\mathsf{P}}_t}{T}\middle\|\mathsf{Q}\right)\right]\\ \lesssim&\frac{C_{\mathsf{KL}}d\sqrt{K_0^4\beta^2+K_1^4+K_2^2+K_0^2K_1^2\beta\left(V^\star-V_\star\right)N}}{T}+\frac{C_{\mathsf{KL}}d\sqrt[3]{K_0^2K_1^4\beta N}}{T^{\frac{2}{3}}}\\ &+\sqrt{\frac{C_{\mathsf{KL}}d\left(dK_0^4\beta^2+K_2^2+K_0^2K_1^2\beta(V^\star-V_\star)\right)}{NT}}+\sqrt[3]{\frac{C_{\mathsf{KL}}^2d^2K_0^2K_1^4\beta}{NT}}+\frac{C_\star d}{N}.\end{aligned}$$

□

D.2 Full Version of Theorem 4 and Its Proof.

Theorem 9 (Full version of Theorem 4)**.** *Under Assumption 2 (without Assumption 2.4), suppose the stepsize $\{\eta_t\}_{t=1}^T$ satisfies*

$$\eta_t \leq \frac{1}{2\sqrt{5K_0^4\beta^2+5K_1^4+5K_2^2+20NK_0^2K_1^2\beta\left(V^\star-V_\star\right)+10NK_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s}}, \forall t\in[T],$$

then Finite-Particle SVGD (Algorithm 2) guarantees that

$$\begin{aligned}&\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}\left(\frac{\sum_{t=1}^T \eta_t\hat{\mathsf{P}}_t}{\sum_{t=1}^T \eta_t}\middle\|\mathsf{Q}\right)\right]\\ \leq&\frac{2C_{\mathsf{KL}}d}{\sum_{t=1}^T\eta_t}+\frac{2(K_0K_1\sqrt{2\beta(V^\star-V_\star)}+dK_0^2\beta+\sqrt{d}K_2)}{N}+\frac{2K_0K_1^2\sqrt{\beta}\sum_{t=1}^T\eta_t\sqrt{\sum_{s=1}^{t-1}\eta_s}}{N\sum_{t=1}^T\eta_t}\\ &+\frac{12\sum_{t=1}^T\eta_t^2\left(dK_0^4\beta^2+K_2^2+2K_0^2K_1^2\beta(V^\star-V_\star)+K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right)}{N\sum_{t=1}^T\eta_t}.\end{aligned}$$

In particular, when

$$\begin{aligned}\eta_t=\eta_\star\triangleq\min\Bigg\{&\frac{1}{4\sqrt{5K_0^4\beta^2+5K_1^4+5K_2^2+20K_0^2K_1^2\beta\left(V^\star-V_\star\right)N}},\frac{1}{2\sqrt[3]{20K_0^2K_1^4\beta NT}},\\ &\sqrt{\frac{C_{\mathsf{KL}}dN}{\left(dK_0^4\beta^2+K_2^2+2K_0^2K_1^2\beta(V^\star-V_\star)\right)T}},\sqrt[3]{\frac{C_{\mathsf{KL}}dN}{K_0^2K_1^4\beta T^2}},\sqrt[3]{\frac{C_{\mathsf{KL}}^2d^2N^2}{K_0^2K_1^4\beta T^3}}\Bigg\},\forall t\in[T],\end{aligned}\tag{44}$$

we have

$$\begin{aligned}&\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}\left(\frac{\sum_{t=1}^T \hat{\mathsf{P}}_t}{T}\middle\|\mathsf{Q}\right)\right]\\ \lesssim&\frac{C_{\mathsf{KL}}d\sqrt{K_0^4\beta^2+K_1^4+K_2^2+K_0^2K_1^2\beta\left(V^\star-V_\star\right)N}}{T}+\frac{C_{\mathsf{KL}}d\sqrt[3]{K_0^2K_1^4\beta N}}{T^{\frac{2}{3}}}\\ &+\sqrt{\frac{C_{\mathsf{KL}}d\left(dK_0^4\beta^2+K_2^2+K_0^2K_1^2\beta(V^\star-V_\star)\right)}{NT}}+\sqrt[3]{\frac{C_{\mathsf{KL}}^2d^2K_0^2K_1^4\beta}{NT}}\\ &+\frac{K_0K_1\sqrt{\beta(V^\star-V_\star)}+dK_0^2\beta+\sqrt{d}K_2}{N}+\frac{\sqrt[3]{C_{\mathsf{KL}}dK_0^2K_1^4\beta}}{N^{\frac{2}{3}}}.\end{aligned}$$

Remark 13. To recover Theorem 4, we take $T = N^{\frac{3}{2}}$ to obtain the $\mathcal{O}(1/N^{\frac{2}{3}})$ rate.

Proof. We recall that $\hat{\mathsf{P}}_t = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{x}_i(t)}$ and $\boldsymbol{g}_{\mathsf{Q}}[\mathsf{P}](\cdot) = \mathbb{E}_{\boldsymbol{x}\sim\mathsf{P}}[K(\cdot,\boldsymbol{x})\nabla V(\boldsymbol{x}) - \nabla_2 K(\cdot,\boldsymbol{x})]$, and use the following equation to have

$$\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2\right] = \mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}}(t))\right\|_{\mathcal{H}^d}^2\right] \overset{(45)}{=} \mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\left\|\boldsymbol{g}_{\mathsf{Q}}\left[\hat{\mathsf{P}}_t\right]\right\|_{\mathcal{H}^d}^2\right]$$
$$\overset{\text{Lemma 5}}{=} \mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}_{\mathcal{H}}^2(\hat{\mathsf{P}}_t\|\mathsf{Q})\right].$$

Therefore, by Lemma 21, for any $t \in [T]$,

$$\mathsf{KL}(\vec{\mathsf{P}}_{t+1}\|\mathsf{Q}^{\otimes N}) \le \mathsf{KL}(\vec{\mathsf{P}}_t\|\mathsf{Q}^{\otimes N}) - \frac{\eta_t N}{2}\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}_{\mathcal{H}}^2(\hat{\mathsf{P}}_t\|\mathsf{Q})\right] + \eta_t\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\sum_{i=1}^N\left\langle\nabla_{2,\boldsymbol{x}_i}, \boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle\right]$$
$$+ \eta_t^2\left(6dK_0^4\beta^2 + 6K_2^2 + 12K_0^2K_1^2\beta(V^\star - V_\star) + 6K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right).$$

Note that

$$\sum_{i=1}^N\left\langle\nabla_{2,\boldsymbol{x}_i}, \boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle \overset{\text{Lemma 18}}{\le} K_0K_1\sqrt{\frac{2\beta}{N}\sum_{i=1}^N V(\boldsymbol{x}_i) - V_\star} + dK_0^2\beta + \sqrt{d}K_2$$

and

$$\frac{1}{N}\sum_{i=1}^N V(\boldsymbol{x}_i) - V_\star \overset{(47)}{\le} V^\star - V_\star + \frac{K_1^2}{2}\sum_{s=1}^{t-1}\eta_s, \ \ \vec{\mathsf{P}}_t\text{-almost surely.}$$

Therefore,

$$\mathsf{KL}(\vec{\mathsf{P}}_{t+1}\|\mathsf{Q}^{\otimes N}) \le \mathsf{KL}(\vec{\mathsf{P}}_t\|\mathsf{Q}^{\otimes N}) - \frac{\eta_t N}{2}\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}_{\mathcal{H}}^2(\hat{\mathsf{P}}_t\|\mathsf{Q})\right]$$
$$+ \eta_t\left(K_0K_1\sqrt{2\beta\left(V^\star - V_\star + \frac{K_1^2}{2}\sum_{s=1}^{t-1}\eta_s\right)} + dK_0^2\beta + \sqrt{d}K_2\right)$$
$$+ \eta_t^2\left(6dK_0^4\beta^2 + 6K_2^2 + 12K_0^2K_1^2\beta(V^\star - V_\star) + 6K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right).$$

Sum the above inequality over $t = 1$ to T, rearrange terms, drop the nonnegative term $\mathsf{KL}(\vec{\mathsf{P}}_{T+1}\|\mathsf{Q}^{\otimes N})$, divide both sides by $\sum_{t=1}^T \frac{\eta_t N}{2}$, and use the convexity of $\mathsf{P} \mapsto \mathsf{KSD}_{\mathcal{H}}^2(\mathsf{P}\|\mathsf{Q})$ to have

$$\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}_{\mathcal{H}}^2\left(\frac{\sum_{t=1}^T\eta_t\hat{\mathsf{P}}_t}{\sum_{t=1}^T\eta_t}\middle\|\mathsf{Q}\right)\right] \le \frac{\sum_{t=1}^T\eta_t\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}_{\mathcal{H}}^2(\hat{\mathsf{P}}_t\|\mathsf{Q})\right]}{\sum_{t=1}^T\eta_t}$$
$$\le \frac{2\mathsf{KL}(\vec{\mathsf{P}}_1\|\mathsf{Q}^{\otimes N})}{N\sum_{t=1}^T\eta_t} + \frac{2(K_0K_1\sqrt{2\beta(V^\star - V_\star)} + dK_0^2\beta + \sqrt{d}K_2)}{N} + \frac{2K_0K_1^2\sqrt{\beta}\sum_{t=1}^T\eta_t\sqrt{\sum_{s=1}^{t-1}\eta_s}}{N\sum_{t=1}^T\eta_t}$$
$$+ \frac{12\sum_{t=1}^T\eta_t^2\left(dK_0^4\beta^2 + K_2^2 + 2K_0^2K_1^2\beta(V^\star - V_\star) + K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right)}{N\sum_{t=1}^T\eta_t}$$
$$\le \frac{2C_{\mathsf{KL}}d}{\sum_{t=1}^T\eta_t} + \frac{2(K_0K_1\sqrt{2\beta(V^\star - V_\star)} + dK_0^2\beta + \sqrt{d}K_2)}{N} + \frac{2K_0K_1^2\sqrt{\beta}\sum_{t=1}^T\eta_t\sqrt{\sum_{s=1}^{t-1}\eta_s}}{N\sum_{t=1}^T\eta_t}$$
$$+ \frac{12\sum_{t=1}^T\eta_t^2\left(dK_0^4\beta^2 + K_2^2 + 2K_0^2K_1^2\beta(V^\star - V_\star) + K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right)}{N\sum_{t=1}^T\eta_t},$$

where the last step is by $\mathsf{KL}(\vec{\mathsf{P}}_1\|\mathsf{Q}^{\otimes N}) \overset{\text{Assumption 2.2}}{\leq} C_{\mathsf{KL}}dN$.

Now, suppose $\eta_t = \eta_\star, \forall t \in [T]$ for some $\eta_\star > 0$, we can simplify the bound into

$$\begin{aligned}
&\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}\left(\frac{\sum_{t=1}^T \hat{\mathsf{P}}_t}{T}\middle\|\mathsf{Q}\right)\right] \\
\leq&\frac{2C_{\mathsf{KL}}d}{\eta_\star T} + \frac{2(K_0K_1\sqrt{2\beta(V^\star - V_\star)} + dK_0^2\beta + \sqrt{d}K_2)}{N} + \frac{2K_0K_1^2\sqrt{\beta\eta_\star T}}{N} \\
&+ \frac{12\left(dK_0^4\beta^2 + K_2^2 + 2K_0^2K_1^2\beta(V^\star - V_\star)\right)\eta_\star}{N} + \frac{6K_0^2K_1^4\beta\eta_\star^2 T}{N}.
\end{aligned}$$

In particular, when $\eta_\star$ equals (44), we have

$$\begin{aligned}
&\mathbb{E}_{\vec{\boldsymbol{x}}(1)}\left[\mathsf{KSD}^2_{\mathcal{H}}\left(\frac{\sum_{t=1}^T \hat{\mathsf{P}}_t}{T}\middle\|\mathsf{Q}\right)\right] \\
\lesssim&\frac{C_{\mathsf{KL}}d\sqrt{K_0^4\beta^2 + K_1^4 + K_2^2 + K_0^2K_1^2\beta\,(V^\star - V_\star)\,N}}{T} + \frac{C_{\mathsf{KL}}d\sqrt[3]{K_0^2K_1^4\beta N}}{T^{\frac{2}{3}}} \\
&+ \sqrt{\frac{C_{\mathsf{KL}}d\left(dK_0^4\beta^2 + K_2^2 + K_0^2K_1^2\beta(V^\star - V_\star)\right)}{NT}} + \sqrt[3]{\frac{C_{\mathsf{KL}}^2d^2K_0^2K_1^4\beta}{NT}} \\
&+ \frac{K_0K_1\sqrt{\beta(V^\star - V_\star)} + dK_0^2\beta + \sqrt{d}K_2}{N} + \frac{\sqrt[3]{C_{\mathsf{KL}}dK_0^2K_1^4\beta}}{N^{\frac{2}{3}}}.
\end{aligned}$$

□

D.3 Theoretical Analysis of Finite-Particle SVGD

We provide the detailed theoretical analysis of Finite-Particle SVGD in this subsection. Our proof strategy builds on Banerjee et al. (2025), but contains several important new ideas and technical improvements, which we discuss later.

Reformulation of Finite-Particle SVGD. Inspired by Banerjee et al. (2025), we first introduce the following $\uparrow$ operation.

Definition 7. Given a function $\boldsymbol{f} : \mathbb{R}^d \times (\mathbb{R}^d)^N \to \mathbb{R}^d$, we define $\boldsymbol{f}^{\uparrow N}$ by

$$\begin{aligned}
&\boldsymbol{f}^{\uparrow N} : (\mathbb{R}^d)^N \to (\mathbb{R}^d)^N; \\
&\vec{\boldsymbol{x}} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_N^\top)^\top \mapsto \boldsymbol{f}^{\uparrow N}(\vec{\boldsymbol{x}}) \triangleq \begin{pmatrix} \boldsymbol{f}(\boldsymbol{x}_1, \vec{\boldsymbol{x}}) \\ \vdots \\ \boldsymbol{f}(\boldsymbol{x}_N, \vec{\boldsymbol{x}}) \end{pmatrix}.
\end{aligned}$$

Remark 14. More generally, the $\uparrow$ operation can be defined with N replaced by any $n \in \mathbb{N}$. For simplicity, we restrict attention to $n = N$.

With the $\uparrow$ operation, the finite-particle update rule can be reformulated as follows.

1. First, define the stacked particle vector
$$\vec{\boldsymbol{x}}(t) \triangleq (\boldsymbol{x}_1^\top(t), \ldots, \boldsymbol{x}_N^\top(t))^\top \in (\mathbb{R}^d)^N, \forall t \in [T+1].$$
2. Then, starting from $\vec{\boldsymbol{x}}(1) \in (\mathbb{R}^d)^N$, the update rule can be written as
$$\vec{\boldsymbol{x}}(t+1) = \vec{\boldsymbol{x}}(t) - \eta_t \boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}(\vec{\boldsymbol{x}}(t)), \forall t \in [T],$$
where
$$\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}, \vec{\boldsymbol{y}}) \triangleq \frac{1}{N}\sum_{i=1}^N K(\boldsymbol{x}, \boldsymbol{y}_i)\nabla V(\boldsymbol{y}_i) - \nabla_2 K(\boldsymbol{x}, \boldsymbol{y}_i), \forall (\boldsymbol{x}, \vec{\boldsymbol{y}}) \in \mathbb{R}^d \times (\mathbb{R}^d)^N. \tag{45}$$
3. Equivalently, letting $\vec{\mathsf{P}}_t \triangleq \mathrm{Law}(\vec{\boldsymbol{x}}(t)), \forall t \in [T+1]$, we have
$$\vec{\mathsf{P}}_{t+1} = \left(\boldsymbol{I} - \eta_t \boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right)_{\#} \vec{\mathsf{P}}_t, \forall t \in [T].$$

First ingredient. The first ingredient of our analysis yields several new technical inequalities.

We begin with Lemma 18, which provides three bounds related to $\boldsymbol{g}_{\mathsf{Q}}$. The first inequality improves upon Banerjee et al. (2025, Lemma 2), whose analysis relies on a growth condition on V that we do not impose. The second inequality is new and does not appear in Banerjee et al. (2025). Importantly, it is the key to removing the optional Assumption 2.4. Concretely, one can show

$$\sum_{i=1}^{N} \left\langle \nabla_{2,\boldsymbol{x}_i}, \boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i, \vec{\boldsymbol{x}}) \right\rangle \overset{(54)}{=} \frac{1}{N} \sum_{i=1}^{N} \langle \nabla_2 K(\boldsymbol{x}_i, \boldsymbol{x}_i), \nabla V(\boldsymbol{x}_i) \rangle + K(\boldsymbol{x}_i, \boldsymbol{x}_i) \Delta V(\boldsymbol{x}_i) - \Delta_2 K(\boldsymbol{x}_i, \boldsymbol{x}_i). \tag{46}$$

In Banerjee et al. (2025), the R.H.S. of (46) is controlled directly through Assumption 2.4, corresponding to the third inequality in Lemma 18. In contrast, our novel observation is that, even without Assumption 2.4, the R.H.S. of (46) can still be properly upper bounded, leading to the second inequality in Lemma 18.

Lemma 18. *Under Assumptions 2.1 and 2.3, for any $\vec{\boldsymbol{x}} \in (\mathbb{R}^d)^N$, we have*

1. $\left\| \boldsymbol{g}_{\mathsf{Q}}(\cdot, \vec{\boldsymbol{x}}) \right\|_{\mathcal{H}^d} \le K_0 \sqrt{\frac{2\beta}{N} \sum_{i=1}^{N} [V(\boldsymbol{x}_i) - V_\star]} + K_1.$

2. $\sum_{i=1}^{N} \left\langle \nabla_{2,\boldsymbol{x}_i}, \boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i, \vec{\boldsymbol{x}}) \right\rangle \le K_0 K_1 \sqrt{\frac{2\beta}{N} \sum_{i=1}^{N} [V(\boldsymbol{x}_i) - V_\star]} + dK_0^2\beta + \sqrt{d} K_2.$

3. $\sum_{i=1}^{N} \left\langle \nabla_{2,\boldsymbol{x}_i}, \boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i, \vec{\boldsymbol{x}}) \right\rangle \le C_\star d$ *if Assumption 2.4 additionally holds.*

Remark 15. In fact, this lemma continues to hold if the β-smoothness condition in Assumption 2.1 is replaced by the weaker β-upper smoothness condition, provided that V remains twice differentiable.

Proof. We defer the proof to Appendix D.3.2. □

Next, we provide two additional inequalities involving $\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}$ in Lemma 19, which constitutes another key component of the first ingredient. The first inequality is new and is introduced to facilitate the application of Lemma 6. The second is inspired by Banerjee et al. (2025, Lemma 4), but yields a sharper bound.

Lemma 19. *Under Assumptions 2.1 and 2.3, for any $\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}} \in (\mathbb{R}^d)^N$, we have*

1. $\left\| \boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}(\vec{\boldsymbol{x}}) - \boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}(\vec{\boldsymbol{y}}) \right\| \le \sqrt{5K_0^4\beta^2 + 5K_1^4 + 5K_2^2 + 20K_0^2K_1^2\beta \sum_{i=1}^{N} [V(\boldsymbol{x}_i) - V_\star]} \, \|\vec{\boldsymbol{x}} - \vec{\boldsymbol{y}}\|.$

2. $\left\| \boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}}) \right\|_{\mathrm{F}} \le \sqrt{2NK_1^2 \left\| \boldsymbol{g}_{\mathsf{Q}}(\cdot, \vec{\boldsymbol{x}}) \right\|_{\mathcal{H}^d}^2 + 6dK_0^4\beta^2 + 6K_2^2 + \frac{12K_0^2K_1^2\beta}{N} \sum_{i=1}^{N} [V(\boldsymbol{x}_i) - V_\star]}.$

Remark 16. Currently, we can only prove this lemma under β-smoothness, which constitutes the main barrier to extending our analysis to the weaker β-upper smooth case.

Proof. We defer the proof to Appendix D.3.3. □

Second ingredient. As seen before, all the inequalities discussed above, except the third one in Lemma 18, involve the term $\frac{1}{N}\sum_{i=1}^{N} V(\boldsymbol{x}_i) - V_\star$. Therefore, our second ingredient is to control this quantity along the trajectory, i.e., $\frac{1}{N}\sum_{i=1}^{N} V(\boldsymbol{x}_i(t)) - V_\star$, which was also studied in Banerjee et al. (2025, Lemma 3).

However, their proof relies on the additional growth condition. By contrast, our new Lemma 20 establishes a desired upper bound without this restrictive assumption, while requiring only a constant threshold on the stepsize, rather than the $\mathcal{O}(1/\sqrt{T})$ threshold required in Banerjee et al. (2025, Lemma 3). Moreover, our bound depends linearly on the stepsize $\{\eta_t\}_{t=1}^{T}$, yielding a sharper dependence than Banerjee et al. (2025, Lemma 3), whose bound scales as $\max\left\{(\eta t)^{\frac{1}{1-\alpha}}, 1\right\}$ when $\eta_t = \eta, \forall t \in [T]$, where $\alpha \in [0, 1/2)$ parameterizes the growth condition.

Lemma 20. *Under Assumptions 2.1, 2.2(only point (a)), and 2.3, suppose the stepsize* $\{\eta_t\}_{t=1}^T$ *satisfies*

$$\eta_t \leq \frac{1}{K_0^2\beta}, \forall t \in [T]\,,$$

then Finite-Particle SVGD (Algorithm 2) guarantees that

$$\frac{1}{N}\sum_{i=1}^N V(\boldsymbol{x}_i(t)) - V_\star \leq V^\star - V_\star + \frac{K_1^2}{2}\sum_{s=1}^{t-1}\eta_s, \forall t \in [T+1]\,.$$

Remark 17. Like Lemma 18, Lemma 20 also holds under the weaker β-upper smoothness condition.

Proof. We defer the proof to Appendix D.3.4. □

Third ingredient. With the previous three lemmas in place, we are ready to present the third and central ingredient of our analysis, Lemma 21, which establishes a descent inequality for $\mathsf{KL}(\vec{\mathsf{P}}_t\|\mathsf{Q}^{\otimes N})$ based on Lemma 6. The key novelty is that, rather than relying solely on pointwise control of the Jacobian, we establish a lower-Lipschitz property on a set of full $\vec{\mathsf{P}}_t$-measure. This directly verifies the condition required by Lemma 6, even when the relevant trajectory-dependent set is not necessarily connected.

Lemma 21. *Under Assumptions 2.1, 2.2, and 2.3, suppose the stepsize* $\{\eta_t\}_{t=1}^T$ *satisfies*

$$\eta_t \leq \frac{1}{2\sqrt{5K_0^4\beta^2 + 5K_1^4 + 5K_2^2 + 20NK_0^2K_1^2\beta\,(V^\star - V_\star) + 10NK_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s}}, \forall t \in [T]\,,$$

then Finite-Particle SVGD (Algorithm 2) guarantees that

$$\begin{aligned}\mathsf{KL}(\vec{\mathsf{P}}_{t+1}\|\mathsf{Q}^{\otimes N}) \leq &\mathsf{KL}(\vec{\mathsf{P}}_t\|\mathsf{Q}^{\otimes N}) - \frac{\eta_t N}{2}\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2\right] + \eta_t\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\sum_{i=1}^N\left\langle\nabla_{2,\boldsymbol{x}_i}, \boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle\right] \\ &+ \eta_t^2\left(6dK_0^4\beta^2 + 6K_2^2 + 12K_0^2K_1^2\beta(V^\star - V_\star) + 6K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right), \forall t \in [T]\,.\end{aligned}$$

Proof. Let $\mathrm{S}_1 = \left\{(\boldsymbol{x}_1^\top,\ldots,\boldsymbol{x}_N^\top)^\top \in (\mathbb{R}^d)^N : \frac{1}{N}\sum_{i=1}^N V(\boldsymbol{x}_i) < V^\star\right\}$ be as specified in Assumption 2.2, and recursively define

$$\mathrm{S}_{t+1} \triangleq \left(\boldsymbol{I} - \eta_t\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right)(\mathrm{S}_t), \forall t \in [T]\,.$$

Note that, for any $t \in [T+1]$,

$$1 \geq \vec{\mathsf{P}}_t(\mathrm{S}_t) = \Pr\left[\vec{\boldsymbol{x}}(t) \in \mathrm{S}_t\right] \geq \Pr\left[\vec{\boldsymbol{x}}(1) \in \mathrm{S}_1\right] \overset{\text{Assumption 2.2}}{=} 1 \Rightarrow \vec{\mathsf{P}}_t(\mathrm{S}_t) = 1.$$

Moreover, for any $\vec{\boldsymbol{x}} \in \mathrm{S}_t$, there exists $\vec{\boldsymbol{x}}(1) \in \mathrm{S}_1$ such that $\vec{\boldsymbol{x}}$ is obtained after $t-1$ Finite-Particle SVGD updates starting from $\vec{\boldsymbol{x}}(1)$. Therefore (note that the condition on the stepsize ensures $\eta_t \leq \frac{1}{K_0^2\beta}, \forall t \in [T]$).

$$\frac{1}{N}\sum_{i=1}^N V(\boldsymbol{x}_i) - V_\star \overset{\text{Lemma 20}}{\leq} V^\star - V_\star + \frac{K_1^2}{2}\sum_{s=1}^{t-1}\eta_s, \forall \vec{\boldsymbol{x}} \in \mathrm{S}_t, t \in [T+1]\,. \tag{47}$$

Note that $K \in C^2(\mathbb{R}^d\times\mathbb{R}^d,\mathbb{R})$ implies that $\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}}) \in \mathcal{H}^d$ for any $\vec{\boldsymbol{x}} \in (\mathbb{R}^d)^N$, meaning that $\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2$ is well-defined. Moreover, for any $\vec{\boldsymbol{x}} \in \mathrm{S}_t$, we have

$$\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d} \overset{\text{Lemma 18}}{\leq} K_0\sqrt{\frac{2\beta}{N}\sum_{i=1}^N V(\boldsymbol{x}_i) - V_\star} + K_1 \overset{(47)}{\leq} K_0\sqrt{2\beta\,(V^\star - V_\star) + \frac{K_1^2}{2}\sum_{s=1}^{t-1}\eta_s} + K_1,$$

which implies that

$$\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2\right]<+\infty.$$

Therefore, w.l.o.g., we can assume $\mathsf{KL}(\vec{\mathsf{P}}_t\|\mathsf{Q}^{\otimes N})<+\infty$ (which also implies that $\vec{\mathsf{P}}_t\ll\mathsf{Q}^{\otimes N}$); otherwise, the inequality holds automatically.

Next, we observe that, for any $\vec{\boldsymbol{x}},\vec{\boldsymbol{y}}\in\mathrm{S}_t$,

$$\begin{aligned}&\left\|\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}(\vec{\boldsymbol{x}})-\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}(\vec{\boldsymbol{y}})\right\|\\ \overset{\text{Lemma 19}}{\leq}&\sqrt{5K_0^4\beta^2+5K_1^4+5K_2^2+20K_0^2K_1^2\beta\sum_{i=1}^N V(\boldsymbol{x}_i)-V_\star}\left\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{y}}\right\|\\ \overset{(47)}{\leq}&\sqrt{5K_0^4\beta^2+5K_1^4+5K_2^2+20NK_0^2K_1^2\beta\left(V^\star-V_\star\right)+10NK_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s}\left\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{y}}\right\|.\end{aligned}$$

Combining the above inequality with the condition

$$\eta_t\leq\frac{1}{2\sqrt{5K_0^4\beta^2+5K_1^4+5K_2^2+20NK_0^2K_1^2\beta\left(V^\star-V_\star\right)+10NK_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s}}$$

implies that

$$\begin{aligned}&\left\|\eta_t\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}(\vec{\boldsymbol{x}})-\eta_t\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}(\vec{\boldsymbol{y}})\right\|\leq\frac{1}{2}\left\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{y}}\right\|,\forall\vec{\boldsymbol{x}},\vec{\boldsymbol{y}}\in\mathrm{S}_t\\ \Rightarrow&\left\|(\boldsymbol{I}-\eta_t\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N})(\vec{\boldsymbol{x}})-(\boldsymbol{I}-\eta_t\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N})(\vec{\boldsymbol{y}})\right\|\geq\frac{1}{2}\left\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{y}}\right\|,\forall\vec{\boldsymbol{x}},\vec{\boldsymbol{y}}\in\mathrm{S}_t.\end{aligned}\tag{48}$$

Additionally, by invariance of domain, this also indicates that S_t is open (and hence Borel) for all $t\in[T+1]$, since S_1 is open by its definition.

Therefore, letting $q^{\otimes N}(\vec{\boldsymbol{x}})\triangleq\prod_{i=1}^N q(\boldsymbol{x}_i),\forall\vec{\boldsymbol{x}}\in(\mathbb{R}^d)^N$ be the density of $\mathsf{Q}^{\otimes N}$, we have

$$\begin{aligned}&\mathsf{KL}(\vec{\mathsf{P}}_{t+1}\|\mathsf{Q}^{\otimes N})-\mathsf{KL}(\vec{\mathsf{P}}_t\|\mathsf{Q}^{\otimes N})=\mathsf{KL}((\boldsymbol{I}-\eta_t\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N})_{\#}\vec{\mathsf{P}}_t\|\mathsf{Q}^{\otimes N})-\mathsf{KL}(\vec{\mathsf{P}}_t\|\mathsf{Q}^{\otimes N})\\ \overset{\text{Lemma 6}}{=}&\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[-\log q^{\otimes N}(\vec{\boldsymbol{x}}-\eta_t\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}(\vec{\boldsymbol{x}}))+\log q^{\otimes N}(\vec{\boldsymbol{x}})-\log\left|\det\left(\boldsymbol{I}-\eta_t\boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right)\right|\right]\\ =&\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\sum_{i=1}^N\left(-\log q(\boldsymbol{x}_i-\eta_t\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}}))+\log q(\boldsymbol{x}_i)\right)-\log\left|\det\left(\boldsymbol{I}-\eta_t\boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right)\right|\right]\\ \overset{(1)}{=}&\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\sum_{i=1}^N\left(V(\boldsymbol{x}_i-\eta_t\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}}))-V(\boldsymbol{x}_i)\right)-\log\left|\det\left(\boldsymbol{I}-\eta_t\boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right)\right|\right]\\ \overset{(a)}{\leq}&\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\sum_{i=1}^N\left(-\eta_t\left\langle\nabla V(\boldsymbol{x}_i),\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle+\frac{\eta_t^2\beta}{2}\left\|\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\|^2\right)\right]\\ &+\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[-\log\left|\det\left(\boldsymbol{I}-\eta_t\boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right)\right|\right]\\ \overset{(b)}{\leq}&\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\sum_{i=1}^N\left(-\eta_t\left\langle\nabla V(\boldsymbol{x}_i),\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle+\frac{\eta_t^2\beta}{2}\left\|\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\|^2\right)\right]\\ &+\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\eta_t\mathrm{Tr}\left(\boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right)+\eta_t^2\left\|\boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right\|_{\mathrm{F}}^2\right]\end{aligned}\tag{49}$$

where (a) is due to β-smoothness of V (cf. Assumption 2.1) and (b) is by, for $\boldsymbol{A}(\vec{\boldsymbol{x}})\triangleq\eta_t\boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}})$ (note that $\|\boldsymbol{A}(\vec{\boldsymbol{x}})\|_{\mathrm{op}}\leq\frac{1}{2},\forall\vec{\boldsymbol{x}}\in\mathrm{S}_t$ due to (48) and the openness of S_t),

$$-\log\left|\det\left(\boldsymbol{I}-\boldsymbol{A}(\vec{\boldsymbol{x}})\right)\right|\overset{\text{Lemma 7}}{\leq}\mathrm{Tr}\left(\boldsymbol{A}(\vec{\boldsymbol{x}})\right)+\frac{\|\boldsymbol{A}(\vec{\boldsymbol{x}})\|_{\mathrm{F}}^2}{2(1-\|\boldsymbol{A}(\vec{\boldsymbol{x}})\|_{\mathrm{op}})}\leq\mathrm{Tr}\left(\boldsymbol{A}(\vec{\boldsymbol{x}})\right)+\|\boldsymbol{A}(\vec{\boldsymbol{x}})\|_{\mathrm{F}}^2,\forall\vec{\boldsymbol{x}}\in\mathrm{S}_t.$$

One more step, we observe that for any differentiable $\boldsymbol{f}:\mathbb{R}^d\times(\mathbb{R}^d)^N\to\mathbb{R}^d$,

$$\mathrm{Tr}\left(\boldsymbol{J}\left[\boldsymbol{f}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right)=\sum_{i=1}^N\langle\nabla_{\boldsymbol{x}_i},\boldsymbol{f}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\rangle=\sum_{i=1}^N\langle\nabla_1,\boldsymbol{f}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\rangle+\langle\nabla_{2,\boldsymbol{x}_i},\boldsymbol{f}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\rangle\,. \tag{50}$$

By (49) and (50), we have

$$\begin{aligned}
&\mathsf{KL}(\vec{\mathsf{P}}_{t+1}\|\mathsf{Q}^{\otimes N})-\mathsf{KL}(\vec{\mathsf{P}}_t\|\mathsf{Q}^{\otimes N})\\
\le&\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[-\eta_t\sum_{i=1}^N\left(\left\langle\nabla V(\boldsymbol{x}_i),\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle-\left\langle\nabla_1,\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle\right)\right]\\
&+\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\eta_t\sum_{i=1}^N\left\langle\nabla_{2,\boldsymbol{x}_i},\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle+\frac{\eta_t^2\beta}{2}\sum_{i=1}^N\left\|\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\|^2+\eta_t^2\left\|\boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right\|_{\mathrm{F}}^2\right]\\
\overset{(c)}{=}&\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[-\eta_tN\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2+\eta_t\sum_{i=1}^N\left\langle\nabla_{2,\boldsymbol{x}_i},\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle\right]\\
&+\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\frac{\eta_t^2\beta}{2}\sum_{i=1}^N\left\|\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\|^2+\eta_t^2\left\|\boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right\|_{\mathrm{F}}^2\right]\\
\overset{(d)}{\le}&\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[-\eta_tN\left(1-\frac{\eta_t(K_0^2\beta+4K_1^2)}{2}\right)\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2\right]+\eta_t\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\sum_{i=1}^N\left\langle\nabla_{2,\boldsymbol{x}_i},\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle\right]\\
&+\eta_t^2\left(6dK_0^4\beta^2+6K_2^2+12K_0^2K_1^2\beta(V^\star-V_\star)+6K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right)\\
\overset{(e)}{\le}&-\frac{\eta_tN}{2}\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2\right]+\eta_t\mathbb{E}_{\vec{\boldsymbol{x}}\sim\vec{\mathsf{P}}_t}\left[\sum_{i=1}^N\left\langle\nabla_{2,\boldsymbol{x}_i},\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle\right]\\
&+\eta_t^2\left(6dK_0^4\beta^2+6K_2^2+12K_0^2K_1^2\beta(V^\star-V_\star)+6K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s\right),
\end{aligned}$$

where (c) is due to

$$\begin{aligned}
&\left\langle\nabla V(\boldsymbol{x}_i),\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle-\left\langle\nabla_1,\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle\\
=&\sum_{j=1}^d\partial_jV(\boldsymbol{x}_i)g_{\mathsf{Q},j}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})-\partial_{1,j}g_{\mathsf{Q},j}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\\
=&\sum_{j=1}^d\partial_jV(\boldsymbol{x}_i)\left\langle g_{\mathsf{Q},j}(\cdot,\vec{\boldsymbol{x}}),K(\cdot,\boldsymbol{x}_i)\right\rangle_{\mathcal{H}}-\left\langle g_{\mathsf{Q},j}(\cdot,\vec{\boldsymbol{x}}),\partial_{2,j}K(\cdot,\boldsymbol{x}_i)\right\rangle_{\mathcal{H}}\\
=&\left\langle\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}}),K(\cdot,\boldsymbol{x}_i)\nabla V(\boldsymbol{x}_i)-\nabla_2K(\cdot,\boldsymbol{x}_i)\right\rangle_{\mathcal{H}^d},
\end{aligned}$$

which implies that

$$\begin{aligned}
&\sum_{i=1}^N\left\langle\nabla V(\boldsymbol{x}_i),\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle-\left\langle\nabla_1,\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle\\
=&N\left\langle\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}}),\frac{1}{N}\sum_{i=1}^NK(\cdot,\boldsymbol{x}_i)\nabla V(\boldsymbol{x}_i)-\nabla_2K(\cdot,\boldsymbol{x}_i)\right\rangle_{\mathcal{H}^d}\overset{(45)}{=}N\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2,
\end{aligned}$$

(d) is by the following two bounds:

$$\left\|\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\|^2\overset{\text{Lemma 3}}{\le}K_0^2\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2\Rightarrow\frac{\eta_t^2\beta}{2}\sum_{i=1}^N\left\|\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\|^2\le\frac{\eta_t^2NK_0^2\beta}{2}\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2$$

and, for any $\vec{\boldsymbol{x}} \in \mathrm{S}_t$,

$$\left\|\boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right\|_{\mathrm{F}}^2 \overset{\text{Lemma 19}}{\leq} 2NK_1^2\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2 + 6dK_0^4\beta^2 + 6K_2^2 + \frac{12K_0^2K_1^2\beta}{N}\sum_{i=1}^N V(\boldsymbol{x}_i) - V_\star$$
$$\overset{(47)}{\leq} 2NK_1^2\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2 + 6dK_0^4\beta^2 + 6K_2^2 + 12K_0^2K_1^2\beta(V^\star - V_\star) + 6K_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s,$$

and (e) follows from

$$\eta_t \leq \frac{1}{2\sqrt{5K_0^4\beta^2 + 5K_1^4 + 5K_2^2 + 20NK_0^2K_1^2\beta\left(V^\star - V_\star\right) + 10NK_0^2K_1^4\beta\sum_{s=1}^{t-1}\eta_s}}$$
$$\Rightarrow \eta_t(K_0^2\beta + 4K_1^2) \leq 1.$$

□

D.3.1 A Helpful Intermediate Lemma

We provide a useful auxiliary lemma for the $\uparrow$ operation.

Lemma 22. *Given an RKHS $(\mathcal{H}, K, \mathbb{R}^d)$ with $K \in C^2(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$, then for any differentiable $\boldsymbol{f} : \mathbb{R}^d \times (\mathbb{R}^d)^N \to \mathbb{R}^d$ such that $\boldsymbol{f}(\cdot, \vec{\boldsymbol{x}}) \in \mathcal{H}^d, \forall \vec{\boldsymbol{x}} \in (\mathbb{R}^d)^N$, we have*

$$\left\|\boldsymbol{J}\left[\boldsymbol{f}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right\|_{\mathrm{F}} \leq \sqrt{2\sum_{i=1}^N \left\|\nabla_2 K(\cdot,\boldsymbol{x}_i)\right\|_{\mathcal{H}^d}^2 \left\|\boldsymbol{f}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2 + 2\sum_{i=1}^N\sum_{j=1}^d\sum_{k=1}^N\sum_{l=1}^d \left(\partial_{2,k,l} f_j(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right)^2}.$$

Proof. We first remark that, by Lemma 1, $\nabla_2 K(\cdot,\boldsymbol{x}) \in \mathcal{H}^d, \forall \boldsymbol{x} \in \mathbb{R}^d$. Hence, the R.H.S. of the inequality is well-defined. Note that, given $i \in [N]$, $j \in [d]$, $k \in [N]$, $l \in [d]$, there is

$$\begin{aligned}\left(\boldsymbol{J}\left[\boldsymbol{f}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right)_{ij,kl} = \frac{\partial f_j(\boldsymbol{x}_i,\vec{\boldsymbol{x}})}{\partial x_{k,l}} &= \begin{cases}\partial_{1,l} f_j(\boldsymbol{x}_i,\vec{\boldsymbol{x}}) + \partial_{2,k,l} f_j(\boldsymbol{x}_i,\vec{\boldsymbol{x}}) & k = i\\ \partial_{2,k,l} f_j(\boldsymbol{x}_i,\vec{\boldsymbol{x}}) & k \neq i\end{cases}\\ &= \mathbb{1}[k=i]\,\partial_{1,l} f_j(\boldsymbol{x}_i,\vec{\boldsymbol{x}}) + \partial_{2,k,l} f_j(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\\ &= \mathbb{1}[k=i]\left\langle f_j(\cdot,\vec{\boldsymbol{x}}), \partial_{2,l}K(\cdot,\boldsymbol{x}_i)\right\rangle_{\mathcal{H}} + \partial_{2,k,l} f_j(\boldsymbol{x}_i,\vec{\boldsymbol{x}}).\end{aligned}$$

Therefore, we know

$$\begin{aligned}\left\|\boldsymbol{J}\left[\boldsymbol{f}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right\|_{\mathrm{F}} &= \sqrt{\sum_{i=1}^N\sum_{j=1}^d\sum_{k=1}^N\sum_{l=1}^d \left(\mathbb{1}[k=i]\left\langle f_j(\cdot,\vec{\boldsymbol{x}}), \partial_{2,l}K(\cdot,\boldsymbol{x}_i)\right\rangle_{\mathcal{H}} + \partial_{2,k,l} f_j(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right)^2}\\ &\leq \sqrt{2\sum_{i=1}^N\sum_{j=1}^d\sum_{k=1}^N\sum_{l=1}^d \left[\mathbb{1}[k=i]\left\langle f_j(\cdot,\vec{\boldsymbol{x}}), \partial_{2,l}K(\cdot,\boldsymbol{x}_i)\right\rangle_{\mathcal{H}}^2 + \left(\partial_{2,k,l} f_j(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right)^2\right]}\\ &\leq \sqrt{2\sum_{i=1}^N\sum_{j=1}^d\sum_{k=1}^N\sum_{l=1}^d \left[\mathbb{1}[k=i]\left\|f_j(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}}^2\left\|\partial_{2,l}K(\cdot,\boldsymbol{x}_i)\right\|_{\mathcal{H}}^2 + \left(\partial_{2,k,l} f_j(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right)^2\right]}\\ &= \sqrt{2\sum_{i=1}^N \left\|\nabla_2 K(\cdot,\boldsymbol{x}_i)\right\|_{\mathcal{H}^d}^2 \left\|\boldsymbol{f}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2 + 2\sum_{i=1}^N\sum_{j=1}^d\sum_{k=1}^N\sum_{l=1}^d \left(\partial_{2,k,l} f_j(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right)^2}.\end{aligned}$$

□

D.3.2 Proof of Lemma 18

Proof. We first note that, under Assumption 2.1, Lemma 12 gives that

$$\left\|\nabla V(\boldsymbol{x})\right\|^2 \leq 2\beta\left(V(\boldsymbol{x}) - V_\star\right), \forall \boldsymbol{x} \in \mathbb{R}^d. \tag{51}$$

Moreover, under Assumption 2.3, there are

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d}\|K(\cdot,\boldsymbol{x})\|_{\mathcal{H}}\le K_0,\quad \sup_{\boldsymbol{x}\in\mathbb{R}^d}\|\nabla_2 K(\cdot,\boldsymbol{x})\|_{\mathcal{H}^d}\le K_1,\quad \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d}\left\|\nabla_2^2 K(\boldsymbol{x},\boldsymbol{y})\right\|_{\mathrm{F}}\le K_2 \tag{52}$$

Now, we are ready to prove these three inequalities. In the following proof, $\vec{\boldsymbol{x}}\in(\mathbb{R}^d)^N$ is fixed.

1. Note that $\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\in\mathcal{H}^d$ under Assumption 2.3. Thus, we have

$$\begin{aligned}
\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d} &\overset{(45)}{=} \frac{1}{N}\left\|\sum_{i=1}^N K(\cdot,\boldsymbol{x}_i)\nabla V(\boldsymbol{x}_i)-\nabla_2 K(\cdot,\boldsymbol{x}_i)\right\|_{\mathcal{H}^d}\\
&\le \frac{1}{N}\sum_{i=1}^N \|K(\cdot,\boldsymbol{x}_i)\nabla V(\boldsymbol{x}_i)\|_{\mathcal{H}^d}+\|\nabla_2 K(\cdot,\boldsymbol{x}_i)\|_{\mathcal{H}^d}\\
&= \frac{1}{N}\sum_{i=1}^N \|K(\cdot,\boldsymbol{x}_i)\|_{\mathcal{H}}\|\nabla V(\boldsymbol{x}_i)\|+\|\nabla_2 K(\cdot,\boldsymbol{x}_i)\|_{\mathcal{H}^d} \overset{(52)}{\le} \frac{K_0}{N}\left(\sum_{i=1}^N\|\nabla V(\boldsymbol{x}_i)\|\right)+K_1\\
&\overset{(51)}{\le} \frac{K_0}{N}\left(\sum_{i=1}^N\sqrt{2\beta(V(\boldsymbol{x}_i)-V_\star)}\right)+K_1 \le K_0\sqrt{\frac{2\beta}{N}\sum_{i=1}^N V(\boldsymbol{x}_i)-V_\star}+K_1,
\end{aligned} \tag{53}$$

   where the last step follows from concavity of $a\mapsto\sqrt{a}$.

2. Given $i\in[N]$, we have

$$\begin{aligned}
\left\langle\nabla_{2,\boldsymbol{x}_i},\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle &= \sum_{j=1}^d \partial_{2,i,j}g_{\mathsf{Q},j}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\\
&\overset{(45)}{=}\frac{1}{N}\sum_{j=1}^d \partial_{2,j}K(\boldsymbol{x}_i,\boldsymbol{x}_i)\partial_j V(\boldsymbol{x}_i)+K(\boldsymbol{x}_i,\boldsymbol{x}_i)\partial_j^2 V(\boldsymbol{x}_i)-\partial_{2,j}^2 K(\boldsymbol{x}_i,\boldsymbol{x}_i)\\
&=\frac{1}{N}\left[\langle\nabla_2 K(\boldsymbol{x}_i,\boldsymbol{x}_i),\nabla V(\boldsymbol{x}_i)\rangle+K(\boldsymbol{x}_i,\boldsymbol{x}_i)\Delta V(\boldsymbol{x}_i)-\Delta_2 K(\boldsymbol{x}_i,\boldsymbol{x}_i)\right],
\end{aligned}$$

   which implies that

$$\sum_{i=1}^N\left\langle\nabla_{2,\boldsymbol{x}_i},\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle=\frac{1}{N}\sum_{i=1}^N\langle\nabla_2 K(\boldsymbol{x}_i,\boldsymbol{x}_i),\nabla V(\boldsymbol{x}_i)\rangle+K(\boldsymbol{x}_i,\boldsymbol{x}_i)\Delta V(\boldsymbol{x}_i)-\Delta_2 K(\boldsymbol{x}_i,\boldsymbol{x}_i). \tag{54}$$

   One more step, we know

$$\begin{aligned}
\langle\nabla_2 K(\boldsymbol{x}_i,\boldsymbol{x}_i),\nabla V(\boldsymbol{x}_i)\rangle &\le \|\nabla_2 K(\boldsymbol{x}_i,\boldsymbol{x}_i)\|\,\|\nabla V(\boldsymbol{x}_i)\| \le \|K(\cdot,\boldsymbol{x}_i)\|_{\mathcal{H}}\|\nabla_2 K(\cdot,\boldsymbol{x}_i)\|_{\mathcal{H}^d}\|\nabla V(\boldsymbol{x}_i)\|\\
&\overset{(52)}{\le} K_0K_1\|\nabla V(\boldsymbol{x}_i)\| \overset{(51)}{\le} K_0K_1\sqrt{2\beta(V(\boldsymbol{x}_i)-V_\star)}.
\end{aligned} \tag{55}$$

   Moreover, we have, under Assumption 2.1,

$$K(\boldsymbol{x}_i,\boldsymbol{x}_i)\Delta V(\boldsymbol{x}_i)=\|K(\cdot,\boldsymbol{x}_i)\|_{\mathcal{H}}^2\operatorname{Tr}(\nabla^2 V(\boldsymbol{x}_i))\le\|K(\cdot,\boldsymbol{x}_i)\|_{\mathcal{H}}^2 d\left\|\nabla^2 V(\boldsymbol{x}_i)\right\|_{\mathrm{op}}\overset{(52)}{\le} dK_0^2\beta, \tag{56}$$

   and

$$-\Delta_2 K(\boldsymbol{x}_i,\boldsymbol{x}_i)=-\operatorname{Tr}(\nabla_2^2 K(\boldsymbol{x}_i,\boldsymbol{x}_i))\le\sqrt{d}\left\|\nabla_2^2 K(\boldsymbol{x}_i,\boldsymbol{x}_i)\right\|_{\mathrm{F}}\overset{(52)}{\le}\sqrt{d}K_2. \tag{57}$$

   Hence, after plugging (55), (56), and (57) back into (54), we obtain that

$$\begin{aligned}
\sum_{i=1}^N\left\langle\nabla_{2,\boldsymbol{x}_i},\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right\rangle &\le \frac{1}{N}\sum_{i=1}^N K_0K_1\sqrt{2\beta(V(\boldsymbol{x}_i)-V_\star)}+dK_0^2\beta+\sqrt{d}K_2\\
&\le K_0K_1\sqrt{\frac{2\beta}{N}\sum_{i=1}^N V(\boldsymbol{x}_i)-V_\star}+dK_0^2\beta+\sqrt{d}K_2,
\end{aligned}$$

   where the last step is due to concavity of $a\mapsto\sqrt{a}$.

3. If Assumption 2.4 holds, we can directly upper bound (54) by $C_\star d$.

□

D.3.3 Proof of Lemma 19

Proof. We first note that, under Assumption 2.1, Lemma 12 states that

$$\|\nabla V(\boldsymbol{x})\|^2 \leq 2\beta \left(V(\boldsymbol{x}) - V_\star\right), \forall \boldsymbol{x} \in \mathbb{R}^d. \tag{58}$$

Moreover, under Assumption 2.3, there are

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} \|K(\cdot,\boldsymbol{x})\|_{\mathcal{H}} \leq K_0, \quad \sup_{\boldsymbol{x}\in\mathbb{R}^d} \|\nabla_2 K(\cdot,\boldsymbol{x})\|_{\mathcal{H}^d} \leq K_1, \quad \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d} \left\|\nabla_2^2 K(\boldsymbol{x},\boldsymbol{y})\right\|_{\mathrm{F}} \leq K_2, \tag{59}$$

which implies that, by Lemma 2, for any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^d$

$$|K(\boldsymbol{z},\boldsymbol{x}) - K(\boldsymbol{z},\boldsymbol{y})| \leq K_0 K_1 \|\boldsymbol{x} - \boldsymbol{y}\|, \tag{60}$$

$$\|\nabla_2 K(\boldsymbol{x},\boldsymbol{z}) - \nabla_2 K(\boldsymbol{y},\boldsymbol{z})\| \leq K_1^2 \|\boldsymbol{x} - \boldsymbol{y}\|, \tag{61}$$

$$\|\nabla_2 K(\boldsymbol{z},\boldsymbol{x}) - \nabla_2 K(\boldsymbol{z},\boldsymbol{y})\| \leq K_2 \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{62}$$

Now, we are ready to prove these two inequalities.

1. By Definition 7, we know

$$\left\|\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}(\vec{\boldsymbol{x}}) - \boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}(\vec{\boldsymbol{y}})\right\| = \sqrt{\sum_{i=1}^N \left\|\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i,\vec{\boldsymbol{x}}) - \boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{y}_i,\vec{\boldsymbol{y}})\right\|^2}$$

$$\overset{(45)}{=} \sqrt{\sum_{i=1}^N \left\| \frac{1}{N}\sum_{j=1}^N K(\boldsymbol{x}_i,\boldsymbol{x}_j)\nabla V(\boldsymbol{x}_j) - \nabla_2 K(\boldsymbol{x}_i,\boldsymbol{x}_j) - K(\boldsymbol{y}_i,\boldsymbol{y}_j)\nabla V(\boldsymbol{y}_j) + \nabla_2 K(\boldsymbol{y}_i,\boldsymbol{y}_j)\right\|^2}$$

$$\leq \sqrt{\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N \left\|K(\boldsymbol{x}_i,\boldsymbol{x}_j)\nabla V(\boldsymbol{x}_j) - \nabla_2 K(\boldsymbol{x}_i,\boldsymbol{x}_j) - K(\boldsymbol{y}_i,\boldsymbol{y}_j)\nabla V(\boldsymbol{y}_j) + \nabla_2 K(\boldsymbol{y}_i,\boldsymbol{y}_j)\right\|^2}$$

$$\leq \sqrt{\mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV} + \mathrm{V}}, \tag{63}$$

where

$$\mathrm{I} \triangleq \frac{5}{N}\sum_{i=1}^N\sum_{j=1}^N \left(K(\boldsymbol{x}_i,\boldsymbol{x}_j) - K(\boldsymbol{y}_i,\boldsymbol{x}_j)\right)^2 \|\nabla V(\boldsymbol{x}_j)\|^2, \tag{64}$$

$$\mathrm{II} \triangleq \frac{5}{N}\sum_{i=1}^N\sum_{j=1}^N \left(K(\boldsymbol{y}_i,\boldsymbol{x}_j) - K(\boldsymbol{y}_i,\boldsymbol{y}_j)\right)^2 \|\nabla V(\boldsymbol{x}_j)\|^2, \tag{65}$$

$$\mathrm{III} \triangleq \frac{5}{N}\sum_{i=1}^N\sum_{j=1}^N K^2(\boldsymbol{y}_i,\boldsymbol{y}_j) \left\|\nabla V(\boldsymbol{x}_j) - \nabla V(\boldsymbol{y}_j)\right\|^2, \tag{66}$$

$$\mathrm{IV} \triangleq \frac{5}{N}\sum_{i=1}^N\sum_{j=1}^N \|\nabla_2 K(\boldsymbol{x}_i,\boldsymbol{x}_j) - \nabla_2 K(\boldsymbol{y}_i,\boldsymbol{x}_j)\|^2, \tag{67}$$

$$\mathrm{V} \triangleq \frac{5}{N}\sum_{i=1}^N\sum_{j=1}^N \left\|\nabla_2 K(\boldsymbol{y}_i,\boldsymbol{x}_j) - \nabla_2 K(\boldsymbol{y}_i,\boldsymbol{y}_j)\right\|^2. \tag{68}$$

Term I. Note that

$$|K(\boldsymbol{x}_i,\boldsymbol{x}_j) - K(\boldsymbol{y}_i,\boldsymbol{x}_j)| = |K(\boldsymbol{x}_j,\boldsymbol{x}_i) - K(\boldsymbol{x}_j,\boldsymbol{y}_i)| \overset{(60)}{\leq} K_0 K_1 \|\boldsymbol{x}_i - \boldsymbol{y}_i\|,$$

which implies that

$$\mathrm{I} \overset{(64)}{\leq} \frac{5K_0^2K_1^2}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{y}_i\|^2\|\nabla V(\boldsymbol{x}_j)\|^2 = \frac{5K_0^2K_1^2}{N}\left(\sum_{j=1}^{N}\|\nabla V(\boldsymbol{x}_j)\|^2\right)\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{y}}\|^2$$
$$\overset{(58)}{\leq} \frac{10K_0^2K_1^2\beta}{N}\left(\sum_{j=1}^{N}V(\boldsymbol{x}_j)-V_\star\right)\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{y}}\|^2. \tag{69}$$

Term II**.** Note that

$$\left|K(\boldsymbol{y}_i,\boldsymbol{x}_j)-K(\boldsymbol{y}_i,\boldsymbol{y}_j)\right| \overset{(60)}{\leq} K_0K_1\left\|\boldsymbol{x}_j-\boldsymbol{y}_j\right\|,$$

which implies that

$$\mathrm{II} \overset{(65)}{\leq} 5K_0^2K_1^2\sum_{j=1}^{N}\left\|\boldsymbol{x}_j-\boldsymbol{y}_j\right\|^2\|\nabla V(\boldsymbol{x}_j)\|^2 \leq 5K_0^2K_1^2\left(\sum_{j=1}^{N}\|\nabla V(\boldsymbol{x}_j)\|^2\right)\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{y}}\|^2$$
$$\overset{(58)}{\leq} 10K_0^2K_1^2\beta\left(\sum_{j=1}^{N}V(\boldsymbol{x}_j)-V_\star\right)\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{y}}\|^2. \tag{70}$$

Term III**.** Note that

$$\left|K(\boldsymbol{y}_i,\boldsymbol{y}_j)\right| = \left|\left\langle K(\cdot,\boldsymbol{y}_i),K(\cdot,\boldsymbol{y}_j)\right\rangle_{\mathcal{H}}\right| \leq \|K(\cdot,\boldsymbol{y}_i)\|_{\mathcal{H}}\left\|K(\cdot,\boldsymbol{y}_j)\right\|_{\mathcal{H}} \overset{(59)}{\leq} K_0^2,$$

which implies that

$$\mathrm{III} \overset{(66)}{\leq} 5K_0^4\sum_{j=1}^{N}\left\|\nabla V(\boldsymbol{x}_j)-\nabla V(\boldsymbol{y}_j)\right\|^2 \leq 5K_0^4\beta^2\sum_{j=1}^{N}\left\|\boldsymbol{x}_j-\boldsymbol{y}_j\right\|^2 = 5K_0^4\beta^2\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{y}}\|^2. \tag{71}$$

Term IV**.** Note that

$$\|\nabla_2K(\boldsymbol{x}_i,\boldsymbol{x}_j)-\nabla_2K(\boldsymbol{y}_i,\boldsymbol{x}_j)\| \overset{(61)}{\leq} K_1^2\|\boldsymbol{x}_i-\boldsymbol{y}_i\|,$$

which implies that

$$\mathrm{IV} \overset{(67)}{\leq} 5K_1^4\sum_{i=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{y}_i\|^2 = 5K_1^4\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{y}}\|^2. \tag{72}$$

Term V**.** Note that

$$\left\|\nabla_2K(\boldsymbol{y}_i,\boldsymbol{x}_j)-\nabla_2K(\boldsymbol{y}_i,\boldsymbol{y}_j)\right\| \overset{(62)}{\leq} K_2\left\|\boldsymbol{x}_j-\boldsymbol{y}_j\right\|,$$

which implies that

$$\mathrm{V} \overset{(68)}{\leq} 5K_2^2\sum_{j=1}^{N}\left\|\boldsymbol{x}_j-\boldsymbol{y}_j\right\|^2 = 5K_2^2\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{y}}\|^2. \tag{73}$$

Finally, we plug (69), (70), (71), (72), and (73) back into (63) to conclude that

$$\left\|\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}(\vec{\boldsymbol{x}})-\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}(\vec{\boldsymbol{y}})\right\| \leq \sqrt{5K_0^4\beta^2+5K_1^4+5K_2^2+20K_0^2K_1^2\beta\sum_{i=1}^{N}V(\boldsymbol{x}_i)-V_\star}\,\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{y}}\|.$$

2. By Lemma 22, for any $\vec{\boldsymbol{x}}\in(\mathbb{R}^d)^N$,

$$\left\|\boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right\|_{\mathrm{F}} \leq \sqrt{2\sum_{i=1}^{N}\|\nabla_2K(\cdot,\boldsymbol{x}_i)\|_{\mathcal{H}^d}^2\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2+2\sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{k=1}^{N}\sum_{l=1}^{d}\left(\partial_{2,k,l}g_{\mathsf{Q},j}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right)^2}$$
$$\overset{(59)}{\leq} \sqrt{2NK_1^2\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2+2\sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{k=1}^{N}\sum_{l=1}^{d}\left(\partial_{2,k,l}g_{\mathsf{Q},j}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right)^2}. \tag{74}$$

Given $i \in [N]$, $j \in [d]$, $k \in [N]$, $l \in [d]$, we observe that

$$\partial_{2,k,l} g_{\mathsf{Q},j}(\boldsymbol{x}_i, \vec{\boldsymbol{x}}) \overset{(45)}{=} \frac{1}{N}\left[\partial_{2,l} K(\boldsymbol{x}_i, \boldsymbol{x}_k)\partial_j V(\boldsymbol{x}_k) + K(\boldsymbol{x}_i, \boldsymbol{x}_k)\partial_l \partial_j V(\boldsymbol{x}_k) - \partial_{2,l}\partial_{2,j} K(\boldsymbol{x}_i, \boldsymbol{x}_k)\right],$$

which implies that

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{k=1}^{N}\sum_{l=1}^{d}\left(\partial_{2,k,l} g_{\mathsf{Q},j}(\boldsymbol{x}_i, \vec{\boldsymbol{x}})\right)^2 \\
=&\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{k=1}^{N}\sum_{l=1}^{d}\left(\partial_{2,l} K(\boldsymbol{x}_i, \boldsymbol{x}_k)\partial_j V(\boldsymbol{x}_k) + K(\boldsymbol{x}_i, \boldsymbol{x}_k)\partial_l \partial_j V(\boldsymbol{x}_k) - \partial_{2,l}\partial_{2,j} K(\boldsymbol{x}_i, \boldsymbol{x}_k)\right)^2 \\
\leq&\mathrm{VI} + \mathrm{VII} + \mathrm{VIII}, \qquad (75)
\end{aligned}$$

where

$$\mathrm{VI} \triangleq \frac{3}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{k=1}^{N}\sum_{l=1}^{d}\left(\partial_{2,l} K(\boldsymbol{x}_i, \boldsymbol{x}_k)\partial_j V(\boldsymbol{x}_k)\right)^2, \qquad (76)$$

$$\mathrm{VII} \triangleq \frac{3}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{k=1}^{N}\sum_{l=1}^{d}\left(K(\boldsymbol{x}_i, \boldsymbol{x}_k)\partial_l \partial_j V(\boldsymbol{x}_k)\right)^2, \qquad (77)$$

$$\mathrm{VIII} \triangleq \frac{3}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{k=1}^{N}\sum_{l=1}^{d}\left(\partial_{2,l}\partial_{2,j} K(\boldsymbol{x}_i, \boldsymbol{x}_k)\right)^2. \qquad (78)$$

Term VI**.** We have

$$\begin{aligned}
\left|\partial_{2,l} K(\boldsymbol{x}_i, \boldsymbol{x}_k)\right| &= \left|\left\langle K(\cdot, \boldsymbol{x}_i), \partial_{2,l} K(\cdot, \boldsymbol{x}_k)\right\rangle_{\mathcal{H}}\right| \\
&\leq \left\|K(\cdot, \boldsymbol{x}_i)\right\|_{\mathcal{H}} \left\|\partial_{2,l} K(\cdot, \boldsymbol{x}_k)\right\|_{\mathcal{H}} \overset{(59)}{\leq} K_0 \left\|\partial_{2,l} K(\cdot, \boldsymbol{x}_k)\right\|_{\mathcal{H}},
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathrm{VI} &\overset{(76)}{\leq} \frac{3}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{k=1}^{N}\sum_{l=1}^{d}\left(K_0 \left\|\partial_{2,l} K(\cdot, \boldsymbol{x}_k)\right\|_{\mathcal{H}} \partial_j V(\boldsymbol{x}_k)\right)^2 \\
&= \frac{3K_0^2}{N}\sum_{k=1}^{N}\left(\sum_{l=1}^{d}\left\|\partial_{2,l} K(\cdot, \boldsymbol{x}_k)\right\|_{\mathcal{H}}^2\right)\left(\sum_{j=1}^{d}\left(\partial_j V(\boldsymbol{x}_k)\right)^2\right) \\
&= \frac{3K_0^2}{N}\sum_{k=1}^{N}\left\|\nabla_2 K(\cdot, \boldsymbol{x}_k)\right\|_{\mathcal{H}^d}^2 \left\|\nabla V(\boldsymbol{x}_k)\right\|^2 \\
&\overset{(59)}{\leq} \frac{3K_0^2 K_1^2}{N}\sum_{k=1}^{N}\left\|\nabla V(\boldsymbol{x}_k)\right\|^2 \overset{(58)}{\leq} \frac{6K_0^2 K_1^2 \beta}{N}\sum_{k=1}^{N} V(\boldsymbol{x}_k) - V_\star. \qquad (79)
\end{aligned}$$

Term VII**.** We have

$$\left|K(\boldsymbol{x}_i, \boldsymbol{x}_k)\right| = \left|\left\langle K(\cdot, \boldsymbol{x}_i), K(\cdot, \boldsymbol{x}_k)\right\rangle_{\mathcal{H}}\right| \leq \left\|K(\cdot, \boldsymbol{x}_i)\right\|_{\mathcal{H}} \left\|K(\cdot, \boldsymbol{x}_k)\right\|_{\mathcal{H}} \overset{(59)}{\leq} K_0^2,$$

and, under Assumption 2.1,

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d}\left\|\nabla^2 V(\boldsymbol{x})\right\|_{\mathrm{F}} \leq \sup_{\boldsymbol{x}\in\mathbb{R}^d}\sqrt{d}\left\|\nabla^2 V(\boldsymbol{x})\right\|_{\mathrm{op}} \leq \sqrt{d}\beta,$$

which together imply that

$$\mathrm{VII} \overset{(77)}{\leq} \frac{3K_0^4}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{k=1}^{N}\sum_{l=1}^{d}\left(\partial_l \partial_j V(\boldsymbol{x}_k)\right)^2 = \frac{3K_0^4}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\left\|\nabla^2 V(\boldsymbol{x}_k)\right\|_{\mathrm{F}}^2 \leq 3dK_0^4\beta^2. \qquad (80)$$

Term VIII**.** We have

$$\text{VIII} \overset{(78)}{=} \frac{3}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\left\|\nabla_2^2 K(\boldsymbol{x}_i,\boldsymbol{x}_k)\right\|_{\mathrm{F}}^2 \overset{(59)}{\leq} 3K_2^2. \tag{81}$$

Finally, we plug (79), (80), and (81) back into (75) to obtain that

$$\sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{k=1}^{N}\sum_{l=1}^{d}\left(\partial_{2,k,l}g_{\mathsf{Q},j}(\boldsymbol{x}_i,\vec{\boldsymbol{x}})\right)^2 \leq 3dK_0^4\beta^2 + 3K_2^2 + \frac{6K_0^2K_1^2\beta}{N}\sum_{k=1}^{N}V(\boldsymbol{x}_k) - V_\star.$$

We combine the above inequality with (74) to have

$$\left\|\boldsymbol{J}\left[\boldsymbol{g}_{\mathsf{Q}}^{\uparrow N}\right](\vec{\boldsymbol{x}})\right\|_{\mathrm{F}} \leq \sqrt{2NK_1^2\left\|\boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}})\right\|_{\mathcal{H}^d}^2 + 6dK_0^4\beta^2 + 6K_2^2 + \frac{12K_0^2K_1^2\beta}{N}\sum_{k=1}^{N}V(\boldsymbol{x}_k) - V_\star}.$$

□

D.3.4 Proof of Lemma 20

Proof. Given $t \in [T]$, by β-smoothness (cf. Assumption 2.1), for all $i \in [N]$,

$$\begin{aligned}V(\boldsymbol{x}_i(t+1)) &\leq V(\boldsymbol{x}_i(t)) + \langle\nabla V(\boldsymbol{x}_i(t)), \boldsymbol{x}_i(t+1) - \boldsymbol{x}_i(t)\rangle + \frac{\beta}{2}\left\|\boldsymbol{x}_i(t+1) - \boldsymbol{x}_i(t)\right\|^2 \\ &= V(\boldsymbol{x}_i(t)) - \eta_t\left\langle\nabla V(\boldsymbol{x}_i(t)), \boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i(t),\vec{\boldsymbol{x}}(t))\right\rangle + \frac{\eta_t^2\beta}{2}\left\|\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i(t),\vec{\boldsymbol{x}}(t))\right\|^2,\end{aligned}$$

which implies that

$$\frac{1}{N}\sum_{i=1}^{N}V(\boldsymbol{x}_i(t+1)) \leq \frac{1}{N}\sum_{i=1}^{N}V(\boldsymbol{x}_i(t)) - \eta_t\left\langle\nabla V(\boldsymbol{x}_i(t)), \boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i(t),\vec{\boldsymbol{x}}(t))\right\rangle + \frac{\eta_t^2\beta}{2}\left\|\boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i(t),\vec{\boldsymbol{x}}(t))\right\|^2. \tag{82}$$

By the kernel trick, for all $i \in [N]$, we have

$$\begin{aligned}&\left\langle\nabla V(\boldsymbol{x}_i(t)), \boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i(t),\vec{\boldsymbol{x}}(t))\right\rangle = \sum_{j=1}^{d}\partial_j V(\boldsymbol{x}_i(t))g_{\mathsf{Q},j}(\boldsymbol{x}_i(t),\vec{\boldsymbol{x}}(t)) \\ =&\sum_{j=1}^{d}\partial_j V(\boldsymbol{x}_i(t))\left\langle g_{\mathsf{Q},j}(\cdot,\vec{\boldsymbol{x}}(t)), K(\cdot,\boldsymbol{x}_i(t))\right\rangle_{\mathcal{H}} = \left\langle K(\cdot,\boldsymbol{x}_i(t))\nabla V(\boldsymbol{x}_i(t)), \boldsymbol{g}_{\mathsf{Q}}(\cdot,\vec{\boldsymbol{x}}(t))\right\rangle_{\mathcal{H}^d}.\end{aligned} \tag{83}$$

Combine (82) and (83) to obtain that

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} V(\boldsymbol{x}_i(t+1)) \\
\leq &\frac{1}{N}\sum_{i=1}^{N} V(\boldsymbol{x}_i(t)) - \eta_t \left\langle \frac{1}{N}\sum_{i=1}^{N} K(\cdot, \boldsymbol{x}_i(t))\nabla V(\boldsymbol{x}_i(t)), \boldsymbol{g}_{\mathsf{Q}}(\cdot, \vec{\boldsymbol{x}}(t)) \right\rangle_{\mathcal{H}^d} + \frac{\eta_t^2 \beta}{2N}\sum_{i=1}^{N} \left\| \boldsymbol{g}_{\mathsf{Q}}(\boldsymbol{x}_i(t), \vec{\boldsymbol{x}}(t)) \right\|^2 \\
\overset{(a)}{\leq} &\frac{1}{N}\sum_{i=1}^{N} V(\boldsymbol{x}_i(t)) - \eta_t \left\langle \frac{1}{N}\sum_{i=1}^{N} K(\cdot, \boldsymbol{x}_i(t))\nabla V(\boldsymbol{x}_i(t)), \boldsymbol{g}_{\mathsf{Q}}(\cdot, \vec{\boldsymbol{x}}(t)) \right\rangle_{\mathcal{H}^d} + \frac{\eta_t^2 K_0^2 \beta}{2} \left\| \boldsymbol{g}_{\mathsf{Q}}(\cdot, \vec{\boldsymbol{x}}(t)) \right\|_{\mathcal{H}^d}^2 \\
= &\frac{1}{N}\sum_{i=1}^{N} V(\boldsymbol{x}_i(t)) - \frac{\eta_t - \eta_t^2 K_0^2 \beta}{2} \left\| \boldsymbol{g}_{\mathsf{Q}}(\cdot, \vec{\boldsymbol{x}}(t)) \right\|_{\mathcal{H}^d}^2 - \frac{\eta_t}{2} \left\| \frac{1}{N}\sum_{i=1}^{N} K(\cdot, \boldsymbol{x}_i(t))\nabla V(\boldsymbol{x}_i(t)) \right\|_{\mathcal{H}^d}^2 \\
&+ \frac{\eta_t}{2} \left\| \frac{1}{N}\sum_{i=1}^{N} K(\cdot, \boldsymbol{x}_i(t))\nabla V(\boldsymbol{x}_i(t)) - \boldsymbol{g}_{\mathsf{Q}}(\cdot, \vec{\boldsymbol{x}}(t)) \right\|_{\mathcal{H}^d}^2 \\
\overset{(45)}{=} &\frac{1}{N}\sum_{i=1}^{N} V(\boldsymbol{x}_i(t)) - \frac{\eta_t - \eta_t^2 K_0^2 \beta}{2} \left\| \boldsymbol{g}_{\mathsf{Q}}(\cdot, \vec{\boldsymbol{x}}(t)) \right\|_{\mathcal{H}^d}^2 - \frac{\eta_t}{2} \left\| \frac{1}{N}\sum_{i=1}^{N} K(\cdot, \boldsymbol{x}_i(t))\nabla V(\boldsymbol{x}_i(t)) \right\|_{\mathcal{H}^d}^2 \\
&+ \frac{\eta_t}{2} \left\| \frac{1}{N}\sum_{i=1}^{N} \nabla_2 K(\cdot, \boldsymbol{x}_i(t)) \right\|_{\mathcal{H}^d}^2 \overset{(b)}{\leq} \frac{1}{N}\sum_{i=1}^{N} V(\boldsymbol{x}_i(t)) + \frac{\eta_t K_1^2}{2},
\end{aligned} \tag{84}
$$

where (a) is due to Lemma 3 and Assumption 2.3 and (b) follows from $\eta_t \leq \frac{1}{K_0^2\beta} \Rightarrow \eta_t - \eta_t^2 K_0^2 \beta \geq 0$ and $\left\| \frac{1}{N}\sum_{i=1}^{N} \nabla_2 K(\cdot, \boldsymbol{x}_i(t)) \right\|_{\mathcal{H}^d} \leq \frac{1}{N}\sum_{i=1}^{N} \left\| \nabla_2 K(\cdot, \boldsymbol{x}_i(t)) \right\|_{\mathcal{H}^d} \leq K_1$ (cf. Assumption 2.3).

Finally, iterating (84), we have, for any $t \in [T]$,

$$
\frac{1}{N}\sum_{i=1}^{N} V(\boldsymbol{x}_i(t+1)) \leq \frac{1}{N}\sum_{i=1}^{N} V(\boldsymbol{x}_i(1)) + \frac{K_1^2}{2}\sum_{s=1}^{t} \eta_s \leq V^\star + \frac{K_1^2}{2}\sum_{s=1}^{t} \eta_s,
$$

where the last step is due to the initial condition (cf. Assumption 2.2). After subtracting $V_\star$ from both sides and changing t to $t-1$, we conclude the proof. □