Provably Fast Finite Particle Variants of SVGD via Virtual Particle Stochastic Approximation

Anonymous Author(s)
Affiliation
Address
email

Abstract

SVGD is a popular particle-based variational inference algorithm with well studied mean-field dynamics. However, its finite-particle behavior is far less understood. Our work introduces the notion of *virtual particles* to develop novel stochastic approximations of mean-field SVGD dynamics in the space of probability measures, that are exactly realizable using finite particles. As a result, we design two computationally efficient variants of SVGD (VP-SVGD and GB-SVGD) with provably fast finite-particle convergence rates. Our algorithms are specific random-batch approximations of SVGD which are computationally more efficient than ordinary SVGD. We show that the n output particles of VP-SVGD and GB-SVGD, run for T steps with batchsize K, are as good as i.i.d samples from a measure whose Kernel Stein Discrepancy to the target is at most $O(d^{1/3}/(KT)^{1/6})$ under standard assumptions. We prove similar results under a mild growth condition on the score function, which is weaker than the assumptions of prior works. Our convergence rates for the empirical measure (of the particles output by VP-SVGD and GB-SVGD) to the target distribution enjoys a *double exponential improvement* over the best known finite-particle analysis of SVGD. Furthermore, our results give the *first known polynomial oracle complexity in dimension*, completely eliminating the curse of dimensionality exhibited by previously known finite-particle rates.

1 Introduction

Sampling from a distribution over $\mathbb{R}^d$ whose density $\pi^\star(\mathbf{x}) \propto \exp(-F(\mathbf{x}))$ is known only upto a normalizing constant, is a fundamental problem in mahine learning [44, 19, 25] and statistics [35, 31, 15]. Stein Variational Gradient Descent (SVGD) by Liu and Wang [27] is a popular algorithm for this problem. It uses a positive definite kernel k to evolve n interacting particles $(\mathbf{x}_t^{(i)})_{i\in[n],t\in\mathbb{N}}$:

$$\mathbf{x}_{t+1}^{(i)} \leftarrow \mathbf{x}_t^{(i)} - \frac{\gamma}{n}\sum_{j=1}^{n}\left[k(\mathbf{x}_t^{(i)},\mathbf{x}_t^{(j)})\nabla F(\mathbf{x}_t^{(j)}) - \nabla_2 k(\mathbf{x}_t^{(i)},\mathbf{x}_t^{(j)})\right] \tag{1}$$

SVGD exhibits remarkable empirical performance in various Bayesian inference, generative modeling and reinforcement learning tasks [27, 43, 21, 29] and usually converges rapidly to the target density while using only a few particles, often outperforming Markov Chain Monte Carlo methods. However, in contrast to its wide practical applicability, theoretical analysis of its behavior is relatively unexplored. Prior works on the analysis of SVGD [23, 14, 26, 36, 7] mainly consider the mean-field limit (or population limit), where the number of particles $n \to \infty$. These works assume that the initial distribution of the (infinite number of) particles has a finite KL divergence to the target $\pi^\star$ and subsequently, interpret mean-field SVGD dynamics as 'Projected' Gradient Descent (GD) of KL divergence on the space of probability measures, equipped with the Wasserstein geometry. Under suitable assumptions on the target density, these works use the theory of Wasserstein Gradient Flows

Submitted to the Optimal Transport and Machine Learning Workshop at NeurIPS 2023. Do not distribute.

Result	**Algorithm**	**Assumption**	**Rate**	**Oracle Complexity**
Korba et al. [23]	Population Limit SVGD	Uniformly Bounded $\mathsf{KSD}_{\pi^\star}(\bar{\mu}_t\|\|\pi^\star)$	$\frac{\mathsf{poly}(d)}{\sqrt{T}}$	Not Implementable
Salim et al. [36]	Population Limit SVGD	Sub-gaussian $\pi^\star$	$\frac{d^{3/2}}{\sqrt{T}}$	Not Implementable
Shi and Mackey [37]	SVGD	Sub-gaussian $\pi^\star$	$\frac{\mathsf{poly}(d)}{\sqrt{\log\log n^{\Theta(1/d)}}}$	$\frac{\mathsf{poly}(d)}{\epsilon^2} e^{\Theta(d e^{\mathsf{poly}(d)/\epsilon^2})}$
Ours, Corollary 1	**VP-SVGD**	**Sub-gaussian** $\pi^\star$	$(d/n)^{1/4} + (d/n)^{1/2}$	d^4/ϵ^{12}
Ours, Corollary 1	**GB-SVGD**	**Sub-gaussian** $\pi^\star$	$d^{1/3}/n^{1/12} + (d/n)^{1/2}$	d^6/ϵ^{18}
Ours, Corollary 1	**VP-SVGD**	**Sub-exponential** $\pi^\star$	$\frac{d^{1/3}}{n^{1/6}} + \frac{d}{n^{1/2}}$	d^6/ϵ^{16}
Ours, Corollary 1	**GB-SVGD**	**Sub-exponential** $\pi^\star$	$\frac{d^{3/8}}{n^{1/16}} + \frac{d}{n^{1/2}}$	d^9/ϵ^{24}

Table 1: Comparison of our results with prior works. d, T, and n denote the dimension, no. of iterations and no. of output particles respectively. Oracle Complexity denotes number of evaluations of ∇F needed to achieve $\mathsf{KSD}_{\pi^\star}(\cdot||\pi^\star) \leq \epsilon$ (with n and T appropriately optimized), and Rate denotes convergence rate w.r.t KSD metric. Note that: 1. Population Limit SVGD is not implementable as it requires infinite particles 2. The uniformly bounded $\mathsf{KSD}_{\pi^\star}(\bar{\mu}_t||\pi^\star)$ assumption is much stronger than subgaussianity and cannot be verified apriori (see Salim et al. [36] Section 1.2.1)

[1] to establish non-asymptotic (in time) convergence of mean-field SVGD to $\pi^\star$ in the Kernel Stein Discrepancy (KSD) metric. While this framework suffices to explain the behavior of SVGD in the mean-field limit, the same techniques are insufficient for analyzing finite-particle regime. This is mainly due to the fact that the empirical measure $\hat{\mu}^{(n)}$ of a finite number of particles does not admit a density (w.r.t Lebesgue Measure), and thus, its KL divergence to the target is always infinite. Moreover, a direct analysis of the dynamics of finite-particle SVGD becomes prohibitively difficult due to complex inter-particle dependencies. To the best of our knowledge, Shi and Mackey [37] is the only result that obtains an explicit convergence rate for finite-particle SVGD by tracking the deviation between the law of n-particle SVGD and mean-field SVGD. The authors show that for subgaussian $\pi^\star$, the empirical measure of n-particle SVGD converges to $\pi^\star$ at $O(\sqrt{\frac{\mathsf{poly}(d)}{\log\log n^{\Theta(1/d)}}})$ rate in KSD (we explicate the d dependence in Shi and Mackey [37] by closely following their analysis). The obtained rate (which suffers from curse of dimensionality) is quite slow and fails to adequately explain the practical performance of SVGD.

Our work deliberately avoids computing the deviation between mean-field SVGD and finite-particle SVGD. Instead, we directly analyze the dynamics of KL divergence along a carefully constructed trajectory in the space of distributions. Our proposed algorithm, Virtual Particle SVGD (VP-SVGD) devises an *unbiased stochastic approximation (in the space of measures) to mean-field SVGD*. We achieve this by considering additional particles called *virtual particles* which evolve in time but aren't part of the output (i.e. *real particles*). These virtual particles are used only to compute information about the current population-level distribution of the real particles, and enable exact implementation of our stochastic approximation to mean-field SVGD, while using only a finite number of particles. Our analysis is similar in spirit to non-asymptotic analyses of Stochastic Gradient Descent (SGD) that do not attempt to track GD (analogous to mean-field SVGD in this case), but instead track the evolution of the objective function along the SGD trajectory using appropriate descent lemmas [20, 11]. The key feature of our proposed stochastic approximation is the fact that it can be exactly implemented using only a finite number of particles. This allows us to design faster variants of SVGD with provably fast finite-particle convergence.

1.1 Contributions and Technical Chellenges

1.2 Contributions

VP-SVGD and GB-SVGD We propose two variants of SVGD that enjoy provably fast finite-particle convergence guarantees: Virtual Particle SVGD (VP-SVGD, Algorithm 1) and Global Batch SVGD (GB-SVGD, Algorithm 2). VP-SVGD is a conceptually elegant stochastic approximation (in the space of probability measures) of mean-field SVGD, and GB-SVGD is a practically efficient version of SVGD which achieves good empirical performance. Our analysis of GB-SVGD builds upon that of VP-SVGD. When the potential F is smooth and satisfies a quadratic growth condition (which holds under subgaussianity of $\pi^\star$, a common assumption in prior works [36, 37]), we show that the n particles output by T steps of our algorithms, run with batch-size K, are at least as good as i.i.d draws from a distribution whose KSD to $\pi^\star$ is at most $O(d^{1/3}/(KT)^{1/6})$. Our results also hold under a mild subquadratic growth condition for F, which is much weaker than isoperimetric (e.g. Poincare Inequality) or information-transport (e.g. Talagrand's Inequality T_1) assumptions generally considered in the sampling literature [41, 36, 37, 8, 2].

State-of-the-art Finite Particle Guarantees As corollaries of the above result, we establish that *VP-SVGD and GB-SVGD exhibit the best known finite-particle guarantees in the literature which significantly outperform that of prior works*. Our results are summarized in Table 1. In particular, under subgaussianity of $\pi^\star$, we show that the empirical measure of the n particles output by VP-SVGD converges to $\pi^\star$ in KSD at a $O((d/n)^{1/4} + (d/n)^{1/2})$ rate. Similarly, the empirical measure of the n output particles of GB-SVGD converges to $\pi^\star$ at a KSD rate of $O(d^{1/3}/n^{1/12} + (d/n)^{1/2})$. Both these results are a *double exponential improvement* over the $O(\frac{\mathsf{poly}(d)}{\sqrt{\log \log n^{\Theta(1/d)}}})$ KSD rate of n-particle SVGD obtained by Shi and Mackey [37], which, to our knowledge, is the best known finite-particle rate for SVGD so far. In terms of gradient oracle complexity (i.e., the number of ∇F evaluations required to achieve $\mathsf{KSD}_{\pi^\star}(\cdot||\pi^\star) \leq \epsilon$), we show that for subgaussian $\pi^\star$, the oracle complexity of VP-SVGD is $O(d^4/\epsilon^{12})$ while that of GB-SVGD is $O(d^6/\epsilon^{18})$. To the best of our knowledge, our result presents the *first known oracle complexity guarantee with polynomial dimension dependence*, and consequently, does not suffer from a curse of dimensionality unlike prior works. Furthermore, as discused above, the conditions under which our result holds is far weaker than subgaussianity of $\pi^\star$, and as such, includes sub-exponential targets and beyond. In particular, *our guarantees for sub-exponential target distributions are (to the best of our knowledge) the first of its kind.*

Empirical Evaluation Our experiments in Appendix 8 show that GB-SVGD obtains similar performance as SVGD *but requires fewer computations*.

Our analysis resolves the following important technical challenges of independent interest:

Stochastic Approximation in the Space of Probability Measures Stochastic approximations are widely used in optimization and and sampling [24, 44]. In sampling, such approximations are generally implemented in path space, e.g., Stochastic Gradient Langevin Dynamics [44] takes a stochastic approximation of the form $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\eta}{K}\sum_{j=0}^{K-1} \nabla f(\mathbf{x}_t, \xi_j) + \sqrt{2\eta}\epsilon_t, \ \epsilon_t \sim \mathcal{N}(0, \mathbf{I}); \ \mathbb{E}[f(\mathbf{x}_t, \xi_j)|\mathbf{x}_t] = F(\mathbf{x}_t)$. Such stochastic approximations are analyzed using the theory of stochastic processes over $\mathbb{R}^d$ [12, 34, 22]. However, when viewed in the space of probability measures (i.e, $\mu_t = \mathrm{Law}(\mathbf{x}_t)$), the time-evolution of these algorithms is deterministic. In contrast, our approach designs *stochastic approximations in the space of probability measures*. In particular, the time-evolution of the law of any particle in VP-SVGD and GB-SVGD are a stochastic approximation of the dynamics of mean-field SVGD. Careful design ensures that our stochastic approximation requires only a finite number of particles for exact implementation.

Tracking KL Divergence in the Finite-Particle Regime The population limit ($n \to \infty$) ensures that the initial empirical distribution (μ_0) of SVGD admits a density (w.r.t the Lebesgue measure). Prior works on population-limit SVGD analyze the time-evolution of the KL divergence to $\pi^\star$. However, this approach cannot be directly used for finite-particle SVGD since the empirical distribution of a finite number of particles does not admit a density, and thus its KL divergence to $\pi^\star$ is infinite. Our analysis of VP-SVGD and GB-SVGD circumvents this obstacle by considering the dynamics of an infinite number of particles, whose empirical measure then admits a density. However, the careful design ensures that the dynamics of n of these particles can be computed exactly, using only a finite total number of (real + virtual) particles. When conditioned on the virtual particles, these particles are i.i.d. and their conditional law is close to the target distribution with high probability.

2 Notation and Problem Setup

We use $\|\cdot\|$, $\langle\cdot,\cdot\rangle$ to denote the Euclidean norm and inner product over $\mathbb{R}^d$ respectively, while other norms and inner products are subscripted with their underlying space. $\mathcal{B}(R)$ denotes the ball of radius R in $(\mathbb{R}^d, \langle\cdot,\cdot\rangle)$. $\mathcal{P}_2(\mathbb{R}^d)$ denotes the space of probability measures on $\mathbb{R}^d$ with finite second moment, with the Wasserstein-2 metric denoted as $\mathcal{W}_2(\mu,\nu)$ for $\mu,\nu \in \mathcal{P}_2(\mathbb{R}^d)$. For any two probability measures μ,ν, we denote their KL divergence as $\mathsf{KL}(\mu||\nu)$. For any function $f : X \to Y$ and any probability measure μ over X, we let $f_{\#}\mu$ denote the law of $f(\mathbf{x}) : \mathbf{x} \sim \mu$. Given a sigma algebra $\mathcal{F}$ over some space Ω, and a measurable space $\mathcal{X}$, $\mu(\cdot\,;\cdot) : \mathcal{F} \times \mathcal{X} \to \mathbb{R}^+$ is a probability kernel if $\forall\, x \in \mathcal{X}$, $\mu(\cdot\,; x)$ is a measure over $\mathcal{F}$ and $\forall\, A \in \mathcal{F}$, the map $x \to \mu(A\,; x)$ is measurable. We use probability measures $\mu(\cdot\,; \mathbf{x})$, where $\mathbf{x}$ is a random element of some appropriate space $\mathcal{X}$, resulting in random probability measures. We use $[m]$ and (m) to denote the sets $\{1, \dots, m\}$ and $\{0, \dots, m-1\}$ respectively, and $S_{(m)}$ to denote the set of all permutations of (m). We use the O notation to characterize the dependence of our rates on the number of iterations T, dimension d and batch-size K, suppressing numerical and problem-dependent constants. We use $\lesssim$ to denote $\leq$ upto universal constants. We fix a symmetric positive definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and denote the corresponding reproducing kernel Hilbert space (RKHS) [38] as $\mathcal{H}_0$. We denote the product RKHS as $\mathcal{H} = \prod_{i=1}^{d} \mathcal{H}_0$, equipped with the standard inner product for product spaces. We assume k is differentiable in both its arguments and let $\nabla_2 k(\mathbf{x}, \mathbf{y})$ denote its gradient w.r.t the second argument. For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we assume $\mathcal{H} \subset L^2(\mu)$ and the inclusion map $i_\mu : \mathcal{H} \to L^2(\mu)$ is continuous. We use $P_\mu : L^2(\mu) \to \mathcal{H}$ to denote the adjoint of i_μ, i.e., the unique operator which satisfies $\langle f, i_\mu g \rangle_{L^2(\mu)} = \langle P_\mu f, g \rangle_{\mathcal{H}}$ for any $f \in L^2(\mu), g \in \mathcal{H}$. Carmeli et al. [6] shows that P_μ can be expressed as a kernel convolution, i.e., $(P_\mu f)(\mathbf{x}) = \int k(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \mathsf{d}\mu(\mathbf{y})$. We define the function $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as $h(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{y}) \nabla F(\mathbf{y}) - \nabla_2 k(\mathbf{x}, \mathbf{y})$ and $h_\mu \in \mathcal{H}$ as $h_\mu = P_\mu(\nabla_{\mathbf{x}} \log(\frac{\mathsf{d}\mu}{\mathsf{d}\pi^\star}(\mathbf{x})))$ for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Integration by parts shows that $h_\mu(\mathbf{x}) = \int h(\mathbf{x}, \mathbf{y}) \mathsf{d}\mu(\mathbf{y})$. Similar to prior works [36, 23, 37] we use Kernel Stein Discrepancy (KSD) as a convergence metric.

Definition 1 (Kernel Stein Discrepancy[28, 9])**.** *Define the Langevin Stein Operator of $\pi^\star$ acting on any differentiable $g : \mathbb{R}^d \to \mathbb{R}^d$ as $(T_{\pi^\star} g)(\mathbf{x}) = \nabla \cdot g(\mathbf{x}) - \langle \nabla F(\mathbf{x}), g(\mathbf{x}) \rangle$. Then, for any two probability measures μ, ν, the Kernel Stein Discrepancy between μ and ν w.r.t $\pi^\star$ is defined as* $\mathsf{KSD}_{\pi^\star}(\mu || \nu) = \sup_{\|g\|_{\mathcal{H}} \leq 1} \mathbb{E}_\mu[T_{\pi^\star} g] - \mathbb{E}_\nu[T_{\pi^\star} g] = \|h_\mu - h_\nu\|_{\mathcal{H}}$.

3 Background on Mean-Field SVGD

We briefly introduce the analysis of mean-field SVGD using the theory of Wasserstein Gradient Flows and refer the readers to prior work [23, 36] for a detailed treatment. The metric space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is called the Wasserstein space, which admits the following Riemannian structure : For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, the tangent space $T_\mu \mathcal{P}_2(\mathbb{R}^d)$ can be identified with the Hilbert space $L^2(\mu)$. We can then define differentiable functionals $\mathcal{L} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and compute their Wasserstein gradients $\nabla_{\mathcal{W}_2} \mathcal{L}$. Note that the target $\pi^\star$ is the unique minimizer over of the functional $\mathcal{L}[\mu] = \mathsf{KL}\,(\mu || \pi^\star)$ over $\mathcal{P}_2(\mathbb{R}^d)$, and its Wasserstein Gradient is $\nabla_{\mathcal{W}_2} \mathcal{L}[\mu] = \nabla_{\mathbf{x}} \log(\frac{\mathsf{d}\mu}{\mathsf{d}\pi^\star}(\mathbf{x}))$ [1]. This powerful machinery has served as a backbone for the analysis of algorithms such as LMC [45, 3] and mean-field SVGD [14, 23, 36]. In particular, mean-field SVGD can be viewed as 'Projected' Gradient Descent in $\mathcal{P}_2(\mathbb{R}^d)$. To infer this, let $\hat{\mu}_t^n$ denote the empirical measures of the SVGD particles $(\mathbf{x}_t^{(i)})_{i \in [n]}$ at step t and recall that $h_\mu(\mathbf{x}) = P_\mu(\nabla_{\mathbf{x}} \log(\frac{\mathsf{d}\mu}{\mathsf{d}\pi^\star}))(\mathbf{x}) = \int h(\mathbf{x}, \mathbf{y}) \mathsf{d}\mu(\mathbf{y})$ (Sec. 2). The SVGD updates in (1) can be recast as $\hat{\mu}_{t+1}^n = (I - \gamma h_{\hat{\mu}_t^n})_{\#} \hat{\mu}_n^t$. In the limit of infinite particles $n \to \infty$, suppose the empirical measure $\hat{\mu}_t^n$ converges to the population measure $\bar{\mu}_t$. In this mean-field limit, the updates can be expressed as,

$$\bar{\mu}_{t+1} = (I - h_{\bar{\mu}_t})_{\#}\, \bar{\mu}_t = \left(I - \gamma P_{\bar{\mu}_t} \left(\nabla \log(\tfrac{d\bar{\mu}_t}{d\pi^\star})\right)\right)_{\#} \bar{\mu}_t = (I - \gamma P_{\bar{\mu}_t} \left(\nabla_{\mathcal{W}_2} \mathsf{KL}\,(\bar{\mu}_t || \pi^\star)\right))_{\#}\, \bar{\mu}_t$$

Recall from Sec. 2 that $P_{\bar{\mu}_t} : L^2(\bar{\mu}_t) \to \mathcal{H}$ is the adjoint of $i_{\bar{\mu}_t}$. Since $\mathcal{H} \subset L^2(\bar{\mu}_t)$, the updates of mean-field SVGD can be seen as 'Projected' Wasserstein Gradient Descent for $\mathcal{L}[\mu] = \mathsf{KL}\,(\mu || \pi^\star)$, with the Wasserstein Gradient at each step being projected onto the RKHS $\mathcal{H}$. Assuming $\mathsf{KL}\,(\bar{\mu}_0 || \pi^\star) < \infty$, convergence of population limit SVGD is then established by tracking the evolution of $\mathsf{KL}\,(\bar{\mu}_t || \pi^\star)$ under appropriate structural assumptions (such as subgaussianity) on $\pi^\star$.

4 Algorithm and Intuition

In this section, we derive VP-SVGD (Algorithm 1), and build upon it to obtain GB-SVGD. Consider a countably infinite collection of particles $\mathbf{x}_0^{(l)} \in \mathbb{R}^d,\ l \in \mathbb{N} \cup \{0\}$, sampled i.i.d from a measure μ_0, having a density w.r.t. the Lebesgue measure. By the strong law of large numbers, the empirical measure of $\mathbf{x}_0^{(l)}$ is almost surely equal to μ_0 [13, Theorem 11.4.1]. Let $K \in \mathbb{N}$ denote the batch size and define the filtration $\mathcal{F}_t = \sigma(\{\mathbf{x}_0^{(l)} \mid l \leq Kt - 1\}),\ \forall\, t \in \mathbb{N}$ with $\mathcal{F}_0$ being the trivial σ algebra.

Algorithm 1 Virtual Particle SVGD (VP-SVGD)

Input: Number of steps T, number of output particles n, batch size K, Initial positions $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(n+KT-1)} \overset{i.i.d.}{\sim} \mu_0$, Kernel k, step size γ.

for $t \in \{0, \dots, T-1\}$ **do**
  for $s \in \{0, \dots, KT+n-1\}$ **do**
    $\mathbf{x}_{t+1}^{(s)} = \mathbf{x}_t^{(s)} - \frac{\gamma}{K}\sum_{l=0}^{K-1}[k(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(tK+l)})\nabla F(\mathbf{x}_t^{(tK+l)}) - \nabla_2 k(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(tK+l)})]$
  end for
end for
Draw S uniformly at random from $\{0, \dots, T-1\}$
Output $(\mathbf{y}^{(0)}, \dots, \mathbf{y}^{(n-1)}) = (\mathbf{x}_S^{(TK)}, \dots, \mathbf{x}_S^{(TK+n-1)})$

For ease of exposition, we discuss the case of $K = 1$ below and present a complete derivation for arbitrary $K \geq 1$ in Section C. Recall from Section 3 that the updates of mean-field SVGD in $\mathcal{P}_2(\mathbb{R}^d)$ is as follows:

$$\bar{\mu}_{t+1} = (I - \gamma h_{\bar{\mu}_t})_{\#}\bar{\mu}_t \tag{2}$$

We aim to design a stochastic approximation in $\mathcal{P}_2(\mathbb{R}^d)$ for the updates (2), such that it admits a finite-particle realization. To this end, we propose the following dynamics in $\mathbb{R}^d$

$$\mathbf{x}_{t+1}^{(s)} = \mathbf{x}_t^{(s)} - \gamma h(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(t)}), \quad s \in \mathbb{N} \cup \{0\} \tag{3}$$

Now, for each time-step t, we focus on the time evolution of the particles $(\mathbf{x}_t^{(l)})_{l \geq t}$ (called the *lower triangular evolution*). From (3), we observe that for any $t \in \mathbb{N}$ and $l \geq t$, $\mathbf{x}_t^{(l)}$ depends only on $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(t-1)}, \mathbf{x}_0^{(l)}$. Therefore there exists a deterministic, measurable function H_t such that:

$$\mathbf{x}_t^{(l)} = H_t(\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(t-1)}, \mathbf{x}_0^{(l)})\,; \quad \text{for every } l \geq t \tag{4}$$

Since $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(t-1)}, \mathbf{x}_0^{(l)} \overset{i.i.d.}{\sim} \mu_0$, we conclude from (4) that $(\mathbf{x}_t^{(l)})_{l \geq t}$ are i.i.d when conditioned on $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(t-1)}$. To this end, we define the random measure $\mu_t|\mathcal{F}_t$ as the law of $\mathbf{x}_t^{(t)}$ conditioned on $\mathcal{F}_t$, i.e., $\mu_t|\mathcal{F}_t$ is a probability kernel $\mu_t(\cdot\,; \mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(t-1)})$, where $\mu_0|\mathcal{F}_0 := \mu_0$. By the strong law of large numbers, $\mu_t|\mathcal{F}_t$ is equal to the empirical measure of $(\mathbf{x}_t^{(l)})_{l \geq t}$ conditioned on $\mathcal{F}_t$. We will use $\mu_t|\mathcal{F}_t$ and $\mu_t(\cdot\,; \mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(t-1)})$ interchangeably.

Define the random function $g_t : \mathbb{R}^d \to \mathbb{R}^d$ as $g_t(\mathbf{x}) := h(\mathbf{x}, \mathbf{x}_t^{(t)})$. From (4), we note that g_t is $\mathcal{F}_{t+1}$ measurable. From (3), we infer that the particles satisfy the following relation:

$$\mathbf{x}_{t+1}^{(s)} = (I - \gamma g_t)(\mathbf{x}_t^{(s)}), \quad s \geq t+1$$

Recall that $\mathbf{x}_{t+1}^{(s)}|\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(t)} \sim \mu_{t+1}|\mathcal{F}_{t+1}$ for any $s \geq t+1$. Furthermore, from Equation (4), we note that for $s \geq t+1$, $\mathbf{x}_t^{(s)}$ depends only on $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(t-1)}$ and $\mathbf{x}_0^{(s)}$. Hence, we conclude that $\mathrm{Law}(\mathbf{x}_t^{(s)}|\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(t)}) = \mathrm{Law}(\mathbf{x}_t^{(s)}|\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(t-1)}) = \mu_t|\mathcal{F}_t \;\forall\; s \geq t+1$. With this insight, the dynamics of the lower-triangular evolution in $\mathcal{P}_2(\mathbb{R}^d)$ that the following holds almost surely:

$$\mu_{t+1}|\mathcal{F}_{t+1} = (I - \gamma g_t)_{\#}\mu_t|\mathcal{F}_t \tag{5}$$

$\mathbf{x}_t^{(t)}|\mathcal{F}_t \sim \mu_t|\mathcal{F}_t$ implies $\mathbb{E}[g_t(\mathbf{x})|\mathcal{F}_t] = h_{\mu_t|\mathcal{F}_t}(\mathbf{x})$. Thus *lower triangular dynamics* (5) is a stochastic approximation in $\mathcal{P}_2(\mathbb{R}^d)$ to the population limit of SVGD (2). Setting the batch size to general K and tracking the evolution of the first $KT+n$ particles, we obtain VP-SVGD (Algorithm 1).

Virtual Particles In Algorithm 1, $(\mathbf{x}_t^{(l)})_{KT \leq l \leq KT+n-1}$ are the *real particles* which constitute the output. $(\mathbf{x}_t^{(l)})_{l < KT}$ are *virtual particles* which propagate information about the probability measure $\mu_t|\mathcal{F}_t$ to enable computation of g_t, an unbiased estimate of the projected Wasserstein gradient $h_{\mu_t|\mathcal{F}_t}$.

Intuition Behind GB-SVGD We note that VP-SVGD (Algorithm 1)is a without-replacement random-batch approximation of SVGD (1), where a different batch is used across timesteps, but the same batch is used across particles given a fixed timestep. With i.i.d. initialization, picking the 'virtual particles' in a fixed order or from a random permutation does not change the evolution of the real particles. With this insight, we design GB-SVGD (Algorithm 2) where we consider n particles *and* output n particles (instead of wasting KT particles as 'virtual particles') via a random-batch approximation of SVGD. In GB-SVGD, with replacement sampling means selecting a batch of K particles i.i.d. from $\mathsf{Uniform}((n))$. Without replacement sampling means fixing a random permutation $\sigma \sim \mathsf{Uniform}(S_{(n)})$ and selecting the batches in the order specified by σ (essentially ensuring that no data point is repeated during an iteration).

Algorithm 2 Global Batch SVGD (`GB-SVGD`)

Input: # of time steps T, # of particles n, $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(n-1)} \overset{i.i.d.}{\sim} \mu_0$, Kernel k, step size γ, Batch size K, Sampling method $\in$ {with replacement, without replacement}

for $t \in \{0, \dots, T-1\}$ **do**
  $\mathcal{K}_t \leftarrow$ random subset of $[n]$ of size K (via. sampling method)
  for $s \in \{0, \dots, n-1\}$ **do**
    $\mathbf{x}_{t+1}^{(s)} = \mathbf{x}_t^{(s)} - \frac{\gamma}{K} \sum_{r \in \mathcal{K}_t} [k(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(r)}) \nabla F(\mathbf{x}_t^{(r)}) - \nabla_2 k(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(r)})]$
  end for
end for
Draw S uniformly at random from $\{0, 1, \dots, T-1\}$
Output $(\bar{\mathbf{y}}^{(0)}, \dots, \bar{\mathbf{y}}^{(n-1)}) = (\mathbf{x}_S^{(0)}, \dots, \mathbf{x}_S^{(n-1)})$

5 Assumptions

We now discuss the key assumptions required for our analysis of VP-SVGD and GB-SVGD.

Assumption 1 (L-Smoothness)**.** *∇F exists and is L Lipschitz. Moreover $\|\nabla F(0)\| \leq \sqrt{L}$.*

Lipschitzness of ∇F is standard in optimization and sampling. It is also easy find a point $\mathbf{x}^*$ such that $\|\nabla F(\mathbf{x}^*)\| \leq \sqrt{L}$ (e.g., using $\Theta(1)$ steps of GD [32]) and center the initialization at $\mathbf{x}^*$. We take $\mathbf{x}^* = 0$ without loss of generality. We now impose the following growth condition on F.

Assumption 2 (Growth Condition)**.** *There exist $\alpha, d_1, d_2 > 0$ such that $F(\mathbf{x}) \geq d_1 \|\mathbf{x}\|^\alpha - d_2$*

Note that Assumption 1 ensures $\alpha \leq 2$. Assumption 2 is a tail decay assumption on $\pi^\star(\mathbf{x}) \propto e^{-F(\mathbf{x})}$, ensuring that its tails decay as $\propto e^{-\|\mathbf{x}\|^\alpha}$. Thus, it holds with $\alpha = 2$ when $\pi^\star$ is subgaussian and with $\alpha = 1$ when $\pi^\star$ is subexponential (See Appendix B for proofs). Subgaussianity is equivalent to $\pi^\star$ satisfying the T_1 inequality [5], commonly assumed in prior works on SVGD [36, 37]. Moreover, subexponentiality holds whenever $\pi^\star$ satisfies the Poincare Inequality [4], which is a mild condition in the sampling literature [41, 8, 2, 12, 7]. This makes Assumption 1 much weaker than the isoperimetric or information-transport assumptions considered in prior works. We also make the following mild assumptions on the k that appear in prior work [23, 17] and are satisfied by several standard kernels (e.g. RBF Kernels, Matérn kernels of order $\geq {}^3\!/_2$)

Assumption 3 (Kernel Regularity)**.** *For any $\mathbf{y} \in \mathbb{R}^d$, $k(\cdot, \mathbf{y})$ satisfies $\|k(\cdot, \mathbf{y})\|_{\mathcal{H}_0} \leq B$ and $\nabla_2 k(\cdot, \mathbf{y}) \in \mathcal{H}$ with $\|\nabla_2 k(\cdot, \mathbf{y})\|_{\mathcal{H}} \leq B$. Moreover, there exist $A_1, A_2, A_3 > 0$ such that $0 \leq k(\mathbf{x}, \mathbf{y}) \leq \frac{A_1}{1 + \|\mathbf{x} - \mathbf{y}\|^2}$, $\|\nabla_2 k(\mathbf{x}, \mathbf{y})\| \leq A_2$, and $\|\nabla_2 k(\mathbf{x}, \mathbf{y})\|^2 \leq A_3 k(\mathbf{x}, \mathbf{y})$.*

For ease of exposition, we make the following mild assumption on the initialization.

Assumption 4 (Initialization)**.** *The initial density is $\mu_0 = \mathsf{Uniform}(\mathcal{B}(R))$ with $\mathsf{KL}\left(\mu_0 || \pi^\star\right) < \infty$.*

Since $\mathcal{N}(0, \mathbf{I})$ and $\mathsf{Uniform}(\mathcal{B}(R))$ are nearly indistinguishable with high probability when $R = \tilde{\Theta}(\sqrt{d})$, Assumption 4 can be easily replaced by the Gaussian initialization assumed in prior works. Furthermore, we show in Appendix B that $R = \sqrt{{}^d\!/_L}$ ensures $\mathsf{KL}\left(\mu_0 || \pi^\star\right) = O(d)$

6 Results

6.1 VP-SVGD

Our first result, proved in Appendix C, shows that the law of the *real particles of* VP-SVGD , when conditioned on the virtual particles, is close to $\pi^\star$ in KSD. Consequently, it shows that the particles output by VP-SVGD are i.i.d. samples from a random probability measure $\bar{\mu}(\cdot; \mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(KT-1)}, S)$ which is close to $\pi^\star$ in KSD. Appendix C also presents a high-probability version of Theorem 1.

Theorem 1 (**Convergence of VP-SVGD**). *Let μ_t be as defined in Section 4. Let Assumptions 1 2, 3, and 4 be satisfied and let $\gamma \leq \min\{1/2A_1 L, 1/(4+L)B\}$. There exist $(\zeta_i)_{0\leq i\leq 3}$ depending polynomially on $A_1, A_2, A_3, B, L, d_1, d_2$ for any fixed $\alpha \in (0, 2]$, such that whenever $\gamma\xi \leq \frac{1}{2B}$, with $\xi = \zeta_0 + \zeta_1(\gamma T)^{1/\alpha} + \zeta_2(\gamma^2 T)^{1/\alpha} + \zeta_3 R^{2/\alpha}$, the following holds:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathsf{KSD}^2_{\pi^\star}(\mu_t|\mathcal{F}_t||\pi^\star)\right] \leq \frac{2\mathsf{KL}\left(\mu_0\|\pi^\star\right)}{\gamma T} + \frac{\gamma B(4+L)\xi^2}{K}$$

Define the probability kernel $\bar{\mu}(\cdot\,;\cdot)$ as follows: For any $x_\tau \in \mathbb{R}^d$, $\tau \in (KT)$ and $s \in (T)$, $\bar{\mu}(\cdot\,; x_0, \dots, x_{KT-1}, s) := \mu_s(\cdot\,; x_0, \dots, x_{Ks-1})$ and $\bar{\mu}(\cdot\,; x_0, \dots, x_{KT-1}, s = 0) := \mu_0(\cdot)$. Conditioned on $\mathbf{x}^{(0)}_\tau = x_\tau,\ S = s$ for every $\tau \in (KT)$, the outputs $\mathbf{y}^{(0)}, \dots, \mathbf{y}^{(n-1)}$ of VP-SVGD are i.i.d samples from $\bar{\mu}(\cdot\,; x_0, \dots, x_{KT-1}, s)$. Furthermore,

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\bar{\mu}(\cdot\,; \mathbf{x}^{(0)}_0, \dots, \mathbf{x}^{(KT-1)}_0, S)||\pi^\star)] \leq \frac{2\mathsf{KL}\left(\mu_0\|\pi^\star\right)}{\gamma T} + \frac{\gamma B(4+L)\xi^2}{K}$$

Convergence Rates Setting $R = \sqrt{d/L}$ ensures $\mathsf{KL}\left(\mu_0\|\pi^\star\right) = O(d)$ (see Appendix B). Hence, choosing $\gamma = O(\frac{(Kd)^\eta}{T^{1-\eta}})$ ensures that $\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\bar{\mu}||\pi^\star)] = O(\frac{d^{1-\eta}}{(KT)^\eta})$ where $\eta = \frac{\alpha}{2(1+\alpha)}$. Thus, for $\alpha = 2$, (i.e, sub-Gaussian $\pi^\star$), $\mathsf{KSD}^2 = O(\frac{d^{2/3}}{(KT)^{1/3}})$. For $\alpha = 1$ (i.e, sub-Exponential $\pi^\star$), the rate (in squared KSD) becomes $O(\frac{d^{3/4}}{(KT)^{1/4}})$. To the best of our knowledge, our convergence guarantee for sub-exponential $\pi^\star$ is the first of its kind.

Comparison with Prior Works Salim et al. [36] analyzes population-limit SVGD for subgaussian $\pi^\star$, obtaining $\mathsf{KSD}^2 = O(d^{3/2}/T)$ rate. We note that population-limit SVGD is not implementable whereas VP-SVGD is an implementable algorithm whose outputs are samples from a distribution with guaranteed convergence to $\pi^\star$.

6.2 GB-SVGD

We now use VP-SVGD as the basis to analyze GB-SVGD. Assume $n > KT$. Then, with probability $\geq 1 - K^2T^2/n$ (for with-replacement sampling) and 1 (for without-replacement sampling), the random batches $\mathcal{K}_t$ in GB-SVGD (Algorithm 2) are disjoint and contain distinct elements. Conditioned on this event $\mathcal{E}$, we note that the $n - KT$ particles that were not included in any random batch $\mathcal{K}_t$ evolve exactly like the n real particles of VP-SVGD. With this insight, we show that, conditioned on $\mathcal{E}$, the outputs of VP-SVGD and GB-SVGD can be coupled such that the first $n - KT$ particles output by both the algorithms are exactly equal. This can be used to derive the following squared KSD bound between their empirical measures. We prove this result in Appendix D

Theorem 2 (**KSD Bounds for GB-SVGD**). *Let $n > KT$ and let $\mathbf{Y} = (\mathbf{y}^{(0)}, \dots, \mathbf{y}^{(n-1)})$ and $\bar{\mathbf{Y}} = (\bar{\mathbf{y}}^{(0)}, \dots, \bar{\mathbf{y}}^{(n-1)})$ denote the outputs of VP-SVGD and GB-SVGD respectively. Moreover, let $\hat{\mu}^{(n)} = \frac{1}{n}\sum_{i=0}^{n-1}\delta_{\mathbf{y}^{(i)}}$ and $\hat{\nu}^{(n)} = \frac{1}{n}\sum_{i=0}^{n-1}\delta_{\bar{\mathbf{y}}^{(i)}}$ denote their respective empirical measures. Under the assumptions and parameter settings of Theorem 1, there exists a coupling of $\mathbf{Y}$ and $\bar{\mathbf{Y}}$ such that:*

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\hat{\mu}^{(n)})] \leq \begin{cases} \frac{2K^2T^2\xi^2}{n^2} & \textit{(without replacement sampling)} \\ \frac{2K^2T^2\xi^2}{n^2}\left(1 - \frac{K^2T^2}{n}\right) + \frac{2K^2T^2\xi^2}{n} & \textit{(with replacement sampling)} \end{cases} \quad (6)$$

6.3 Convergence of the Empirical Measure to the Target

As a corollary of Theorem 1 and Theorem 2, we show that the empirical measure of the output of VP-SVGD and GB-SVGD rapidly converges to $\pi^\star$ in KSD. We refer to Appendix E for the full statement and proof.

Corollary 1 (**VP-SVGD and GB-SVGD: Fast Finite Particle Rates**). *Let the assumptions and parameter settings of Theorem 1 be satisfied. Let $\hat{\mu}^{(n)}$ be the empirical measures of the n particles output by VP-SVGD, run with run with $KT = d^{\frac{\alpha}{2+\alpha}}$, $R = \sqrt{d/L}$ and appropriately chosen γ. Then:*

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\mu}^{(n)}||\pi^\star)] \leq O(\frac{d^{\frac{2}{2+\alpha}}}{n^{\frac{\alpha}{2+\alpha}}} + \frac{d^{2/\alpha}}{n})$$

Let $\hat{\nu}^{(n)}$ be the empirical measure of the output of GB-SVGD under without-replacement sampling, run with $KT = \sqrt{n}$, $R = \sqrt{d/L}$ and appropriately chosen γ. Then, the following holds:

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] \leq O(\frac{d^{2/\alpha}}{n} + \frac{d^{\frac{1}{1+\alpha}}}{n^{\frac{1+2\alpha}{2(1+\alpha)}}} + \frac{d^{\frac{2+\alpha}{2(1+\alpha)}}}{n^{\frac{\alpha}{4(1+\alpha)}}})$$

Comparison to Prior Work For subgaussian $\pi^\star$ (i.e. $\alpha = 2$), VP-SVGD has a finite-particle rate of $\mathbb{E}[\mathsf{KSD}_{\pi^\star}(\hat{\mu}^{(n)}||\pi^\star)] = O((d/n)^{1/4} + (d/n)^{1/2})$ while that of GB-SVGD is $\mathbb{E}[\mathsf{KSD}_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] = O(d^{1/3}/n^{1/12} + (d/n)^{1/2})$. Both these rates are a *double exponential improvement* over the $\tilde{O}(\frac{\mathsf{poly}(d)}{\sqrt{\log\log n^{\Theta(1/d)}}})$ KSD rate obtained by Shi and Mackey [37] for SVGD with subgaussian $\pi^\star$. For subexponential $\pi^\star$ (i.e. $\alpha = 1$) the KSD rate of VP-SVGD is $O(\frac{d^{1/3}}{n^{1/6}} + \frac{d}{n^{1/2}})$ while that of GB-SVGD is $O(\frac{d^{3/8}}{n^{1/16}} + \frac{d}{n^{1/2}})$. To our knowledge, both these results are the first of their kind.

Oracle Complexity As illustrated in Section E.3, for subgaussian $\pi^\star$, the oracle complexity of VP-SVGD to achieve ϵ-convergence in KSD is $O(d^4/\epsilon^{12})$ and that of GB-SVGD is $O(d^6/\epsilon^{18})$. To our knowledge, these results are the *first known oracle complexities for this problem with polynomial dimension dependence*, and significantly improve upon the $O(\frac{\mathsf{poly}(d)}{\epsilon^2}e^{\Theta(de^{\mathsf{poly}(d)/\epsilon^2})})$ oracle complexity of SVGD as implied by Shi and Mackey [37]. For subexponential $\pi^\star$, the oracle complexity of VP-SVGD is $O(d^6/\epsilon^{16})$ and that of GB-SVGD is $O(d^9/\epsilon^{24})$.

7 Proof Sketch

We now present a sketch of our analysis. As shown in Section 4, the particles $(\mathbf{x}_t^{(l)})_{l \geq Kt}$ are i.i.d conditioned on the filtration $\mathcal{F}_t$, and the random measure $\mu_t|\mathcal{F}_t$ is the law of $(\mathbf{x}_t^{(Kt)})$ conditioned on $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(Kt-1)}$. Moreover, from equation (5), we know that $\mu_t|\mathcal{F}_t$ is a stochastic approximation of population limit SVGD dynamics, i.e., $\mu_{t+1}|\mathcal{F}_{t+1} = (I - \gamma g_t)_{\#}\mu_t|\mathcal{F}_t$. Lemma 1 (similar to Salim et al. [36, Proposition 3.1] and Korba et al. [23, Proposition 5]) shows that under appropriate conditions, the KL between $\mu_t|\mathcal{F}_t$ and $\pi^\star$ satisfies a (stochastic) descent lemma . Hence $\mu_t|\mathcal{F}_t$ admits a density and $\mathsf{KL}\left(\mu_t|\mathcal{F}_t||\pi^\star\right)$ is almost surely finite.

Lemma 1 (Descent Lemma for $\mu_t|\mathcal{F}_t$)**.** *Let Assumptions 1, 3 and 4 be satisfied and let $\beta > 1$ be an arbitrary constant. On the event $\gamma\|g_t\|_{\mathcal{H}} \leq \frac{\beta-1}{\beta B}$, the following holds almost surely*

$$\mathsf{KL}\left(\mu_{t+1}|\mathcal{F}_{t+1}||\pi^\star\right) \leq \mathsf{KL}\left(\mu_t|\mathcal{F}_t||\pi^\star\right) - \gamma\left\langle h_{\mu_t|\mathcal{F}_t}, g_t\right\rangle_{\mathcal{H}} + \frac{\gamma^2(\beta^2+L)B}{2}\|g_t\|^2_{\mathcal{H}}$$

Lemma 1 is analogous to the noisy descent lemma which is used in the analysis of SGD for smooth functions. Notice that $\mathbb{E}[g_t|\mathcal{F}_t] = h_{\mu_t|\mathcal{F}_t}$ (when interpreted as a Gelfand-Pettis integral [40], as discussed in Appendix B and Appendix C) and hence in expectation, the KL divergence decreases in time. In order to apply Lemma 1, we establish an almost-sure bound on $\|g_t\|_{\mathcal{H}}$ below.

Lemma 2. *Let Assumptions 1, 2, 3 and 4 hold. Then, for $\gamma \leq 1/2A_1L$, $\|g_t\|_{\mathcal{H}} \leq \xi$ holds almost surely, where ξ is as defined in Theorem 1*

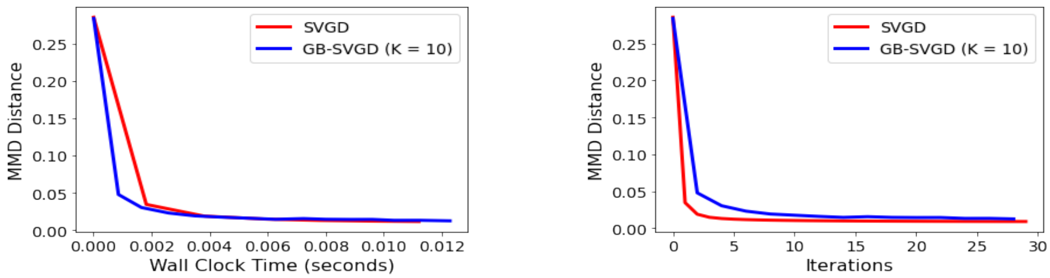
Let $K = 1$ for clarity. To prove Lemma 2, we first note via smoothness of $F(\cdot)$ and Assumption 3 that $\|g_t\|_{\mathcal{H}} \leq C_0\|\mathbf{x}_t^{(t)}\| + C_1$, and then bound $\|\mathbf{x}_t^{(t)}\|$. Now, $g_s(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}_s^{(s)})\nabla F(\mathbf{x}_s^{(s)}) - \nabla_2 k(\mathbf{x}, \mathbf{x}_s^{(s)})$. When $\|\mathbf{x}_s^{(s)} - \mathbf{x}\|$ is large, $\|g_s(\mathbf{x})\|$ is small due to decay assumptions on the kernel (Assumption 3) implying that the particle does not move much. When $\mathbf{x}_s^{(s)} \approx \mathbf{x}$, we have $g_s(\mathbf{x}) \approx k(\mathbf{x}, \mathbf{x}_s^{(s)})\nabla F(\mathbf{x}) - \nabla_2 k(\mathbf{x}, \mathbf{x}_s^{(s)})$ and $k(\mathbf{x}, \mathbf{x}_s^{(s)}) \geq 0$. This is approximately a gradient descent update on $F(\cdot)$ along with a bounded term $\nabla_2 k(\mathbf{x}, \mathbf{x}_s^{(s)})$. Thus, the value of $F(\mathbf{x}_t^{(l)})$ cannot grow too large after T iterations. By Assumption 2, $F(\mathbf{x}_t^{(l)})$ being small implies that $\|\mathbf{x}_t^{(l)}\|$ is small.

Equipped with Lemma 2, we set the step-size γ to ensure that the descent lemma (Lemma 1) always holds. The remainder of the proof involves unrolling through Lemma 1 by taking iterated expectations on both sides. To this end we control $\left\langle h_{\mu_t|\mathcal{F}_t}, g_t\right\rangle_{\mathcal{H}}$ and $\|g_t\|^2_{\mathcal{H}}$ in expectation, in Lemma 3.

Lemma 3. *Let Assumptions 1,2,3 and4 hold and ξ be as defined in Theorem 1. Then, for $\gamma \leq 1/2A_1L$, $\mathbb{E}\left[\left\langle h_{\mu_t|\mathcal{F}_t}, g_t\right\rangle_{\mathcal{H}}|\mathcal{F}_t\right] = \|h_{\mu_t|\mathcal{F}_t}\|^2_{\mathcal{H}}$ and $\mathbb{E}[\|g_t\|^2_{\mathcal{H}}] \leq \xi^2/K + \|h_{\mu_t|\mathcal{F}_t}\|^2_{\mathcal{H}}$*

8 Experiments

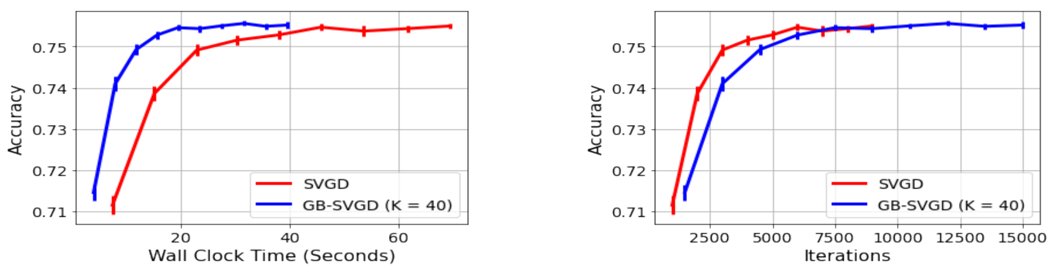
We compare the performance of GB-SVGD and SVGD on the standard baselines used by prior work [27]. We take $n = 100$ and use the Laplace kernel with $h = 1$ for both the algorithms. We pick the stepsize γ by a grid search independently for each algorithm. For both our experimental setups, we observe that while SVGD takes fewer iterations to converge, the compute time for GB-SVGD is considerably lower. This is similar to the typical behavior of stochastic optimization algorithms like SGD.



(a) MMD vs Compute Time

(b) MMD vs Iterations

Figure 1: Gaussian Experiment Comparing SVGD and GB-SVGD averaged over 10 experiments.



(a) Accuracy vs Compute Time

(b) Accuracy vs Iterations

Figure 2: Covertype Experiment, averaged over 50 runs. The error bars represent 95% CI.

Sampling from Isotropic Gaussian (Figure 1): As a sanity check, we set $\pi^\star = \mathcal{N}(0, \mathbf{I})$ with $d = 5$. We pick $K = 10$ for GB-SVGD. The metric of convergence is MMD with respect to the empirical measure of 1000 i.i.d. sampled Gaussians.

Bayesian Logistic Regression (Figure 2) We consider the Covertype dataset which contains $\sim 580,000$ data points with $d = 54$. We consider the same priors suggested in Gershman et al. [16] and implemented in Liu and Wang [27]. We take $K = 40$ for GB-SVGD. For both VP-SVGD and GB-SVGD, we use AdaGrad with momentum to set the step-sizes as per Liu and Wang [27]

We ran our experiments using Python 3 on a 2.20 GHz Intel Xeon CPU with 13 GB of memory.

9 Conclusion

We develop two computationally efficient variants of SVGD with provably fast convergence guarantees in the finite-particle regime, and present a wide range of improvements over prior work. A promising avenue of future work could be to establish convergence guarantees for SVGD with general non-logconcave targets, as was considered in recent works on LMC and SGLD [2, 12]. Other important avenues include establishing minimax lower bounds for SVGD and related particle-based variational inference algorithms. Beyond this, we also conjecture that the rates of GB-SVGD can be improved even in the regime $n \ll KT$. However, we believe this requires new analytic tools.

Acknowledgements

We thank Jiaxin Shi, Lester Mackey and the anonymous reviewers for their helpful feedback. We are particularly grateful to Lester Mackey for providing insightful pointers on the properties of Kernel Stein Discrepancy, which greatly helped us in removing the curse of dimensionality from our finite-particle convergence guarantees.

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[2] K. Balasubramanian, S. Chewi, M. A. Erdogdu, A. Salim, and S. Zhang. Towards a theory of non-log-concave sampling:first-order stationarity guarantees for langevin monte carlo. In P.-L. Loh and M. Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2896–2923. PMLR, 02–05 Jul 2022. URL `https://proceedings.mlr.press/v178/balasubramanian22a.html`.

[3] E. Bernton. Langevin monte carlo and jko splitting. In *Conference On Learning Theory*, pages 1777–1798. PMLR, 2018.

[4] S. Bobkov and M. Ledoux. Poincaré's inequalities and talagrand's concentration phenomenon for the exponential distribution. *Probability Theory and Related Fields*, 107:383–400, 1997.

[5] S. G. Bobkov and F. Götze. Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999.

[6] C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

[7] S. Chewi, T. Le Gouic, C. Lu, T. Maunu, and P. Rigollet. Svgd as a kernelized wasserstein gradient flow of the chi-squared divergence. *Advances in Neural Information Processing Systems*, 33:2098–2109, 2020.

[8] S. Chewi, M. A. Erdogdu, M. Li, R. Shen, and S. Zhang. Analysis of langevin monte carlo from poincare to log-sobolev. In P.-L. Loh and M. Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1–2. PMLR, 02–05 Jul 2022. URL `https://proceedings.mlr.press/v178/chewi22a.html`.

[9] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International conference on machine learning*, pages 2606–2615. PMLR, 2016.

[10] J. B. Conway. *A course in functional analysis*, volume 96. Springer, 2019.

[11] A. Das, B. Schölkopf, and M. Muehlebach. Sampling without replacement leads to faster rates in finite-sum minimax optimization. *Advances in Neural Information Processing Systems*, 2022.

[12] A. Das, D. Nagaraj, and A. Raj. Utilising the clt structure in stochastic gradient based sampling: Improved analysis and faster algorithms. In *Conference on Learning Theory*, 2023.

[13] R. M. Dudley. *Real analysis and probability*. CRC Press, 2018.

[14] A. Duncan, N. Nüsken, and L. Szpruch. On the geometry of stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

[15] A. El Alaoui, A. Montanari, and M. Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 323–334. IEEE, 2022.

[16] S. Gershman, M. D. Hoffman, and D. M. Blei. Nonparametric variational inference. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, 2012.

[17] J. Gorham and L. Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.

[18] J. Gorham, A. Raj, and L. Mackey. Stochastic stein discrepancies. *Advances in Neural Information Processing Systems*, 33:17931–17942, 2020.

[19] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[20] P. Jain, D. M. Nagaraj, and P. Netrapalli. Making the last iterate of sgd information theoretically optimal. *SIAM Journal on Optimization*, 31(2):1108–1130, 2021.

[21] P. Jaini, L. Holdijk, and M. Welling. Learning equivariant energy based models with equivariant stein variational gradient descent. *Advances in Neural Information Processing Systems*, 34: 16727–16737, 2021.

[22] Y. Kinoshita and T. Suzuki. Improved convergence rate of stochastic gradient langevin dynamics with variance reduction and its application to optimization. *arXiv preprint arXiv:2203.16217*, 2022.

[23] A. Korba, A. Salim, M. Arbel, G. Luise, and A. Gretton. A non-asymptotic analysis for stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33: 4672–4682, 2020.

[24] H. J. Kushner and D. S. Clark. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media, 2012.

[25] Y. T. Lee and S. S. Vempala. The manifold joys of sampling (invited talk). In *49th International Colloquium on Automata, Languages, and Programming (ICALP 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

[26] Q. Liu. Stein variational gradient descent as gradient flow. *Advances in neural information processing systems*, 30, 2017.

[27] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

[28] Q. Liu, J. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

[29] Y. Liu, P. Ramachandran, Q. Liu, and J. Peng. Stein variational policy gradient. *arXiv preprint arXiv:1704.02399*, 2017.

[30] J. Lu, Y. Lu, and J. Nolen. Scaling limit of the stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671, 2019.

[31] R. M. Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

[32] Y. Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998.

[33] N. Nüsken and D. Renger. Stein variational gradient descent: many-particle and long-time asymptotics. *arXiv preprint arXiv:2102.12956*, 2021.

[34] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

[35] G. O. Roberts and R. L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

[36] A. Salim, L. Sun, and P. Richtarik. A convergence theory for svgd in the population limit under talagrand's inequality t1. In *International Conference on Machine Learning*, pages 19139–19152. PMLR, 2022.

[37] J. Shi and L. Mackey. A finite-particle convergence rate for stein variational gradient descent. *arXiv preprint arXiv:2211.09721*, 2022.

[38] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[39] L. Sun, A. Karagulyan, and P. Richtarik. Convergence of stein variational gradient descent under a weaker smoothness condition. In *International Conference on Artificial Intelligence and Statistics*, pages 3693–3717. PMLR, 2023.

[40] M. Talagrand. *Pettis integral and measure theory*. American Mathematical Soc., 1984.

[41] S. Vempala and A. Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

[42] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[43] D. Wang, Z. Zeng, and Q. Liu. Stein variational message passing for continuous graphical models. In *International Conference on Machine Learning*, pages 5219–5227. PMLR, 2018.

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

[45] A. Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.

Contents

A Additional Notation and Organization

We use Γ to denote the Gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathsf{d}t$, and recall that for any $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$. For any Lebesgue measurable $A \subseteq \mathbb{R}^d$, we use $\mathsf{vol}(A)$ to denote it's Lebesgue Measure and $\mathsf{Uniform}(A)$ to denote the uniform distribution supported on A. We use $\mathcal{B}(R)$ to denote the ball of radius R centered at the origin, and recall that $\mathsf{vol}(\mathcal{B}(R)) = \frac{\pi^{d/2}}{\Gamma(d/2+1)}R^d$. For ease of exposition, we assume $d \geq 2$. We further assume $\pi^\star(\mathbf{x}) = e^{-F(\mathbf{x})}$. We note that this can be easily ensured by absorbing the normalizing constant into $F(0)$, and does not affect the dynamics of SVGD, VP-SVGD or GB-SVGD (since they only use the gradient information of F). We highlight that both these assumptions are made purely for the sake of clarity and are very easily removable with negligible changes to our analysis.

We empirically benchmark SVGD and GB-SVGD in Appendix 8. In Appendix B, we discuss the technical lemmas used in our analysis, and present a short exposition to the Gelfand-Pettis integral in Appendix B.1, which we use to analyze VP-SVGD. We analyze VP-SVGD in Appendix C and GB-SVGD in Appendix D. Convergence guarantees for the empirical measure of VP-SVGD and GB-SVGD are presented in Appendix E. We give a brief review of the related work in Section F.

B Preliminaries

The following lemma shows that setting the initial distribution $\mu_0 = \mathsf{Uniform}(\mathcal{B}(R))$ with $R = \sqrt{d/L}$ suffices to ensure $\mathsf{KL}\left(\mu_0 || \pi^\star\right) = O(d)$.The proof of this result is similar to that of Vempala and Wibisono [41, Lemma 1] with the Gaussian initialization replaced by $\mathsf{Uniform}(\mathcal{B}(R))$ initialization.

Lemma 4 (KL Upper Bound for Uniform Initialization). *Let Assumption 1 be satisfied and let $\mu_0 = \mathsf{Uniform}(\mathcal{B}(R))$ with $R = \sqrt{d/L}$. Then, the following holds:*

$$\mathsf{KL}\left(\mu_0 || \pi^\star\right) \leq \frac{d}{2}\log(L/2\pi) + d + F(0) + 1/2 \leq O(d)$$

Proof. For any $\mathbf{x} \in \mathbb{R}^d$, the following holds by Assumption 1

$$\begin{aligned} F(\mathbf{x}) &\leq F(0) + \langle \nabla F(0), \mathbf{x} \rangle + \tfrac{L}{2}\|\mathbf{x}\|^2 \\ &\leq F(0) + \sqrt{L}\|\mathbf{x}\| + \tfrac{L}{2}\|\mathbf{x}\|^2 \\ &\leq F(0) + 1/2 + L\|\mathbf{x}\|^2 \end{aligned}$$

where the second inequality uses $\|\nabla F(0)\| \leq \sqrt{L}$ and the Cauchy Schwarz inequality, and the last inequality uses the identity $ab \leq a^2 + b^2/4$. It follows that,

$$\mathbb{E}_{\mathbf{x}\sim\mu_0}[F(\mathbf{x})] \leq F(0) + 1/2 + LR^2$$

By a slight abuse of notation, let μ_0 denote the density of $\mathsf{Uniform}(\mathcal{B}(R))$. Clearly. $\mu_0(\mathbf{x}) = \frac{1}{\mathsf{vol}(\mathcal{B}(R))}\mathbb{I}_{\mathbf{x}\in\mathcal{B}(R)}$. It follows that,

$$\int_{\mathbb{R}^d} \mu_0(\mathbf{x})\ln(\mu_0(\mathbf{x}))\mathsf{d}\mathbf{x} = \int_{\mathcal{B}(R)} \tfrac{1}{\mathsf{vol}(\mathcal{B}(R))}\log(1/\mathsf{vol}(\mathcal{B}(R)))\mathsf{d}\mathbf{x} = -\log(\mathsf{vol}(\mathcal{B}(R)))$$

Now, $\mathsf{vol}(\mathcal{B}(R)) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}R^d$. Furthermore, by Stirling's Approximation, $(x/e)^{x-1} \leq \Gamma(x) \leq (x/2)^{x-1}$. Hence,

$$\frac{d}{2}\log\left(\frac{2\pi R^2}{d/2+1}\right) \leq \log(\mathsf{vol}(\mathcal{B}(R))) \leq \frac{d}{2}\log\left(\frac{e\pi R^2}{d/2+1}\right)$$

Without loss of generality, assume $\pi^\star(\mathbf{x}) = e^{-F(\mathbf{x})}$ (this can be easily ensured by appropriately adjusting $F(0)$ upto constant factors). It follows that,

$$\begin{aligned} \mathsf{KL}\left(\mu_0 || \pi^\star\right) &= \int_{\mathbb{R}^d} \mu_0(\mathbf{x})\log(\tfrac{\mu_0(\mathbf{x})}{\pi^\star(\mathbf{x})})\mathsf{d}\mathbf{x} = \int_{\mathbb{R}^d} \mu_0(\mathbf{x})\ln(\mu_0(\mathbf{x}))\mathsf{d}\mathbf{x} + \mathbb{E}_{\mathbf{x}\sim\mu_0}[F(\mathbf{x})] \\ &\leq -\log(\mathsf{vol}(\mathcal{B}(R))) + F(0) + 1/2 + LR^2 \\ &\leq \frac{d}{2}\log\left(\frac{d/2+1}{2\pi R^2}\right) + F(0) + 1/2 + LR^2 \end{aligned}$$

Setting $R = \sqrt{d/L}$, we conclude that,

$$\mathsf{KL}\left(\mu_0 \| \pi^\star\right) \leq \frac{d}{2}\log(L/2\pi) + d + F(0) + 1/2 \leq O(d)$$

$\square$

We now show that the growth condition on F, i.e. Assumption 2 is more general than specific concentration assumptions on $\pi^\star$ (e.g. subgaussianity, subexponentiality etc.). To this end, we define the notion of α-tail decay as follows:

Definition 2 (α-Tail Decay)**.** *A probability distribution ν on $\mathbb{R}^d$ is said to satisfy α-tail decay for some $\alpha > 0$ if there exists some $C > 0$ such that $\mathbb{E}_{\mathbf{x}\sim\nu}\left[\exp\left(\|\frac{\mathbf{x}}{C}\|^\alpha\right)\right] < \infty$*

The α-tail decay condition essentially implies that the tails of $\pi^\star$ decay as $\propto e^{-\|\mathbf{x}\|^\alpha}$. In particular, Vershynin [42, Proposition 2.5.2 and Proposition 2.7.1] shows that $\pi^\star$ satisfying the tail decay condition with $\alpha = 2$ is equivalent to $\pi^\star$ being subgaussian, whereas tail deay with $\alpha = 1$ is equivalent to $\pi^\star$ being subexponential.

In the following lemma, we establish that, under smoothness of F, the α-tail decay condition is equivalent to the growth condition on F with the same exponent α. Consequently, Assumption 2 is much weaker than the standard isoperimetric and information transport assumptions generally used in the literature.

Lemma 5 (Growth Condition and Tail Decay)**.** *Let Assumption 2 be satisfied for some $\alpha > 0$. Then, $\pi^\star$ satisfies the α-tail decay condition. Conversely, let Assumption 1 be satisfied and suppose $\pi^\star$ satisfies the α-tail decay condition. Then, F satisfies Assumption 2 with the same exponent α.*

Proof. **Growth Condition Implies Tail Decay** Since Assumption 2 is satisfied, $F(\mathbf{x}) \geq d_1\|\mathbf{x}\|^\alpha - d_2$ for some $d_1, d_2, \alpha > 0$. Let $C = (2/d_1)^{1/\alpha}$. It follows that,

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim\pi^\star}\left[e^{\|\mathbf{x}/C\|^\alpha}\right] &= \int_{\mathbb{R}^d} e^{\frac{d_1}{2}\|\mathbf{x}\|^\alpha}\pi^\star(\mathbf{x})\mathsf{d}\mathbf{x} \\
&\leq \int_{\mathbb{R}^d} e^{\frac{d_1}{2}\|\mathbf{x}\|^\alpha - d_1\|\mathbf{x}\|^\alpha + d_2}\mathsf{d}\mathbf{x} \\
&= e^{d_2}\int_{\mathbb{R}^d} e^{-\frac{d_1}{2}\|\mathbf{x}\|^\alpha}\mathsf{d}\mathbf{x} < \infty
\end{aligned}$$

From Definition 2, we conclude that $\pi^\star$ satisfies α-tail decay.

Smoothness and Tail Decay Imply the Growth Condition Since F is smooth, it suffices to consider $\alpha \in (0, 2]$. By Assumption 1, the following inequalities hold,

$$F(\mathbf{y}) - F(\mathbf{x}) \leq \|\nabla F(\mathbf{x})\|\|\mathbf{y} - \mathbf{x}\| + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq (L\|\mathbf{x}\| + \sqrt{L})\|\mathbf{y} - \mathbf{x}\| + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \quad (7)$$

Ww now prove this result by contradiction. Since $\pi^\star$ satisfies α-tail decay, there exists a constant $C > 0$ such that $\mathbb{E}_{\mathbf{x}\sim\pi^\star}[e^{\|\mathbf{x}/C\|^\alpha}] < \infty$. Now, suppose F does not satisfy the growth condition with exponent α, i.e., *assume there does not exist* any $d_1, d_2 > 0$ such that $F(\mathbf{x}) \geq d_1\|\mathbf{x}\|^\alpha - d_2 \; \forall\, \mathbf{x} \in \mathbb{R}^d$. This implies that, $\liminf_{\|\mathbf{x}\|\to\infty} \frac{F(\mathbf{x})}{\|\mathbf{x}\|^\alpha} = 0$. Thus, without loss of generality, we can assume there exists a diverging sequence $a_n \in \mathbb{R}$ and a diverging sequence $\mathbf{x}_n \in \mathbb{R}^d$ that satisfy the following for every $n \in \mathbb{N}$:

$$\frac{F(\mathbf{x}_n)}{\|\mathbf{x}_n\|^\alpha} \leq \frac{1}{a_n}, \quad \|\mathbf{x}_n\| \geq 2n, \quad \|\mathbf{x}_{n+1} - \mathbf{x}_n\| \geq 1 \quad (8)$$

where, without loss of generality, we assume $a_n, \|\mathbf{x}_n\| > 0$. Now, let $r_n = \frac{1}{\|\mathbf{x}_n\|^2}$ and $B_n \subseteq \mathbb{R}^d$ denote the ball of radius r_n centered at $\mathbf{x}_n$. Since $r_n \leq 1/4n^2$ and $\|\mathbf{x}_{n+1} - \mathbf{x}_n\| \geq 1$, B_n is a family of disjoint subsets of $\mathbb{R}^d$. We shall now prove that there exists some diverging sequence $b_n \in \mathbb{R}$ such that $\frac{F(\mathbf{y})}{\|\mathbf{y}\|^\alpha} \leq \frac{1}{b_n}$ for every $\mathbf{y} \in B_n$.

Consider any arbitrary $n \in \mathbb{N}$ and let $\mathbf{y} \in B_n$. Applying (7) to $\mathbf{y}$ and $\mathbf{x}_n$, we obtain,

$$\begin{aligned}\frac{F(\mathbf{y})}{\|\mathbf{x}_n\|^\alpha} &\leq \frac{F(\mathbf{x}_n)}{\|\mathbf{x}_n\|^\alpha} + \frac{L\|\mathbf{x}_n\|r_n}{\|\mathbf{x}_n\|^\alpha} + \frac{r_n\sqrt{L}}{\|\mathbf{x}_n\|^\alpha} + \frac{Lr_n^2}{2\|\mathbf{x}_n\|^\alpha} \\ &\leq \frac{1}{a_n} + \frac{L}{\|\mathbf{x}\|^{\alpha+1}} + \frac{\sqrt{L}}{\|\mathbf{x}_n\|^{\alpha+2}} + \frac{L}{2\|\mathbf{x}_n\|^{\alpha+4}} \end{aligned} \tag{9}$$

where we use (8) and $r_n = 1/\|\mathbf{x}_n\|^2$. Moreover, we note that

$$\|\mathbf{y}\| \geq \|\mathbf{x}_n\| - \|\mathbf{y} - \mathbf{x}_n\| \geq \|\mathbf{x}_n\| - r_n = \|\mathbf{x}_n\| - \frac{1}{\|\mathbf{x}_n\|^2} \geq \frac{\|\mathbf{x}_n\|}{2} \tag{10}$$

where we use the fact that $\|\mathbf{x}_n\| \geq 2n > 2^{1/3}$. It follows that,

$$\begin{aligned}\frac{F(\mathbf{y})}{\|\mathbf{y}\|^\alpha} &\leq \frac{2^\alpha F(\mathbf{y})}{\|\mathbf{x}_n\|^\alpha} \\ &\leq \frac{4}{a_n} + \frac{4L}{\|\mathbf{x}\|^{\alpha+1}} + \frac{4\sqrt{L}}{\|\mathbf{x}_n\|^{\alpha+2}} + \frac{2L}{\|\mathbf{x}_n\|^{\alpha+4}} \end{aligned} \tag{11}$$

where we use (9) and the fact that $\alpha \in (0, 2]$. We now define the sequence $b_n \in \mathbb{R}$ as follows:

$$b_n = \left(\frac{4}{a_n} + \frac{4L}{\|\mathbf{x}\|^{\alpha+1}} + \frac{4\sqrt{L}}{\|\mathbf{x}_n\|^{\alpha+2}} + \frac{2L}{\|\mathbf{x}_n\|^{\alpha+4}}\right)^{-1}$$

Since $\alpha > 0$, and $a_n, \|\mathbf{x}_n\| \to \infty$, it is clear that b_n is a diverging sequence. Furthermore, from (11), we conclude that $\frac{F(\mathbf{y})}{\|\mathbf{y}\|^\alpha} \leq \frac{1}{b_n} \; \forall \, \mathbf{y} \in B_n$. Equipped with this construction, we note that

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\pi^\star}\left[\exp\left(\frac{\|\mathbf{x}\|^\alpha}{C^\alpha}\right)\right] &= \int_{\mathbb{R}^d} \exp\left(\frac{\|\mathbf{y}\|^\alpha}{C^\alpha}\right)\exp(-F(\mathbf{y}))\mathsf{d}\mathbf{y} \\ &\geq \sum_{n=1}^{\infty}\int_{B_n} \exp\left(\frac{\|\mathbf{y}\|^\alpha}{C^\alpha}\right)\exp(-F(\mathbf{y}))\mathsf{d}\mathbf{y} \\ &\geq \sum_{n=1}^{\infty}\int_{B_n} \exp\left(\frac{\|\mathbf{y}\|^\alpha}{C^\alpha} - \frac{\|\mathbf{y}\|^\alpha}{b_n}\right)\mathsf{d}\mathbf{y}\end{aligned}$$

where the second inequality use the fact that B_n is a disjoint family of subsets of $\mathbb{R}^d$ and the third inequality uses the fact that $\frac{F(\mathbf{y})}{\|\mathbf{y}\|^\alpha} \leq \frac{1}{b_n} \; \forall \, \mathbf{y} \in B_n$. Since b_n is a diverging sequence, there exists some $N_0 \in \mathbb{N}$ such that $b_n \geq 2C^\alpha \; \forall \, n \geq N_0$. It follows that,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\pi^\star}\left[\exp\left(\frac{\|\mathbf{x}\|^\alpha}{C^\alpha}\right)\right] &\geq \sum_{n=1}^{\infty}\int_{B_n} \exp\left(\frac{\|\mathbf{y}\|^\alpha}{C^\alpha} - \frac{\|\mathbf{y}\|^\alpha}{b_n}\right)\mathsf{d}\mathbf{y} \\ &\geq \sum_{n=N_0}^{\infty}\int_{B_n} \exp\left(\frac{\|\mathbf{y}\|^\alpha}{2C^\alpha}\right)\mathsf{d}\mathbf{y} \\ &= \sum_{n=N_0}^{\infty}\mathsf{vol}(B_n)\mathbb{E}_{\mathbf{y}\sim\mathsf{Uniform}(B_n)}\left[\exp\left(\frac{\|\mathbf{y}\|^\alpha}{2C^\alpha}\right)\right]\end{aligned}$$

Consider the function $g : [0, \infty) \to [0, \infty)$ defined as $g(t) = e^{t^\alpha}$. We note that for $\alpha \geq 1$, g is a convex function for every $t \geq 0$, and for $\alpha \in (0, 1)$, g is convex for every $t \geq (1/\alpha - 1)^{1/\alpha}$. From (10), we note that $\|\mathbf{y}\| \geq \|\mathbf{x}\|/2 \geq n$ for every $\mathbf{y} \in B_n$. Hence, there exists an $N_1 \in \mathbb{N}$ such that e^{t^α}

is a convex function for all $t \geq \|\mathbf{y}\|/2,\ \forall\, \mathbf{y} \in B_n,\ n \geq N_1$. Let $N = \max\{N_0, N_1\} + 1$. Then,

$$
\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim\pi^\star}\left[\exp\left(\frac{\|\mathbf{x}\|^\alpha}{C^\alpha}\right)\right] &\geq \sum_{n=N_0}^{\infty} \mathsf{vol}(B_n)\mathbb{E}_{\mathbf{y}\sim\mathsf{Uniform}(B_n)}\left[\exp\left(\frac{\|\mathbf{y}\|^\alpha}{2C^\alpha}\right)\right] \\
&\geq \sum_{n=N}^{\infty} \mathsf{vol}(B_n)\exp\left(\tfrac{1}{2C^\alpha}\mathbb{E}_{\mathbf{y}\sim\mathsf{Uniform}(B_n)}[\|\mathbf{y}\|]^\alpha\right) \\
&\geq \sum_{n=N}^{\infty} \mathsf{vol}(B_n)\exp\left(\tfrac{1}{2C^\alpha}\|\mathbb{E}_{\mathbf{y}\sim\mathsf{Uniform}(B_n)}[\mathbf{y}]\|^\alpha\right) \\
&\geq \sum_{n=N}^{\infty} \mathsf{vol}(B_n)\exp\left(\tfrac{1}{2C^\alpha}\|\mathbf{x}_n\|^\alpha\right) \\
&= \sum_{n=N}^{\infty} C_d(r_n)^d\exp\left(\tfrac{1}{2C^\alpha}\|\mathbf{x}_n\|^\alpha\right) \\
&= \sum_{n=N}^{\infty} \frac{C_d\exp\left(\tfrac{1}{2C^\alpha}\|\mathbf{x}_n\|^\alpha\right)}{\|\mathbf{x}_n\|^{2d}}
\end{aligned}
$$

where $C_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$. Let k be any positive integer such that $\alpha k \geq 2d + 1$. It follows that,

$$
\frac{\exp\left(\tfrac{1}{2C^\alpha}\|\mathbf{x}_n\|^\alpha\right)}{\|\mathbf{x}_n\|^{2d}} \geq \frac{C_d}{2^k k! C^{\alpha k}}\|\mathbf{x}_n\|^{\alpha k - 2d} \geq \frac{C_d n}{2^{k-1}k!C^{\alpha k}}
$$

Thus, we infer that,

$$
\mathbb{E}_{\mathbf{x}\sim\pi^\star}\left[\exp\left(\frac{\|\mathbf{x}\|^\alpha}{C^\alpha}\right)\right] \geq \frac{C_d}{2^{k-1}k!C^{\alpha k}}\sum_{n=N_0}^{\infty} n = \infty
$$

which is a contradiction. Thus, there exists some $d_1, d_2 > 0$ such that $F(\mathbf{x}) \geq d_1\|\mathbf{x}\|^\alpha - d_2$, i.e., F satisfies the growth condition with exponent α. □

The following lemma establishes boundedness and contractivity properties of the function $h(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{y})\nabla F(\mathbf{y}) - \nabla_2 k(\cdot, \mathbf{y})$, that are vital for proving almost-sure bounds such as Lemma 2.

Lemma 6 (Properties of h)**.** *Let Assumptions 1 and 3 be satisfied. Then, the following holds,*

$$
\begin{aligned}
\|h(., \mathbf{y})\|_{\mathcal{H}} &\leq BL\|\mathbf{y}\| + B\|\nabla F(0)\| + B \\
\|h(\mathbf{x}, \mathbf{y})\| &\leq \frac{A_1 L}{2} + A_2 + k(\mathbf{x}, \mathbf{y})\|\nabla F(\mathbf{x})\| \\
-\langle \nabla F(\mathbf{x}), h(\mathbf{x}, \mathbf{y})\rangle &\leq -\tfrac{1}{2}k(\mathbf{x}, \mathbf{y})\|\nabla F(\mathbf{x})\|^2 + L^2 A_1 + A_3
\end{aligned}
$$

Proof. Recalling the definition of h from Section 2, we observe that,

$$
h(\cdot, \mathbf{y}) = k(\cdot, \mathbf{y})\nabla F(\mathbf{y}) - \nabla_2 k(\cdot, \mathbf{y})
$$

Thus, by triangle inequality of $\|\cdot\|_{\mathcal{H}}$, Assumptions 1 and 3, we obtain

$$
\begin{aligned}
\|h(\cdot, \mathbf{y})\|_{\mathcal{H}} &\leq \|\nabla F(\mathbf{y})\|\|k(\cdot, \mathbf{y})\|_{\mathcal{H}_0} + \|\nabla_2 k(\cdot, \mathbf{y})\|_{\mathcal{H}} \\
&\leq BL\|\mathbf{y}\| + B\|\nabla F(0)\| + B
\end{aligned}
$$

To prove the remaining inequalities, we first note that,

$$
\begin{aligned}
h(\mathbf{x}, \mathbf{y}) &= k(\mathbf{x}, \mathbf{y})\nabla F(\mathbf{y}) - \nabla_2 k(\mathbf{x}, \mathbf{y}) \\
&= k(\mathbf{x}, \mathbf{y})\nabla F(\mathbf{x}) + k(\mathbf{x}, \mathbf{y})\left[\nabla F(\mathbf{y}) - \nabla F(\mathbf{x})\right] - \nabla_2 k(\mathbf{x}, \mathbf{y})
\end{aligned} \tag{12}
$$

Using Assumptions 1 and 3, we note that,

$$
\begin{aligned}
\|h(\mathbf{x}, \mathbf{y})\| &\leq k(\mathbf{x}, \mathbf{y})\|\nabla F(\mathbf{x})\| + \frac{LA_1\|\mathbf{x} - \mathbf{y}\|}{1 + \|\mathbf{x} - \mathbf{y}\|^2} + A_2 \\
&\leq \frac{A_1 L}{2} + A_2 + k(\mathbf{x}, \mathbf{y})\|\nabla F(\mathbf{x})\|
\end{aligned}
$$

where the second inequality uses the fact $\frac{t}{1+t^2} \leq 1/2$

To prove the last inequality, we infer the following from (12)

$$\begin{aligned}
-\langle \nabla F(\mathbf{x}), h(\mathbf{x},\mathbf{y})\rangle &\leq -k(\mathbf{x},\mathbf{y})\|\nabla F(\mathbf{x})\|^2 + k(\mathbf{x},\mathbf{y})\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\|\|\nabla F(\mathbf{x})\| \\
&\quad + \|\nabla_2 k(\mathbf{x},\mathbf{y})\|\|\nabla F(\mathbf{x})\| \\
&\leq -k(\mathbf{x},\mathbf{y})\|\nabla F(\mathbf{x})\|^2 + L\sqrt{k(\mathbf{x},\mathbf{y})}\sqrt{\frac{A_1\|\mathbf{x}-\mathbf{y}\|^2}{1+\|\mathbf{x}-\mathbf{y}\|^2}}\|\nabla F(\mathbf{x})\| \\
&\quad + \sqrt{A_3 k(\mathbf{x},\mathbf{y})}\|\nabla F(\mathbf{x})\| \\
&\leq -\frac{1}{2}k(\mathbf{x},\mathbf{y})\|\nabla F(\mathbf{x})\|^2 + L^2 A_1 + A_3
\end{aligned}$$

where the second inequality uses Assumptions 1 and 3, and the last inequality uses the identity $ab \leq a^2 + b^2/4$ □

To analyze the dynamics of VP-SVGD in the Wasserstein space, we use the following lemma presented in Salim et al. [36]

Lemma 7 (Salim et al. [36], Proposition 3.1)**.** *Let Assumptions 1 and 3 be satisfied. Consider any $\nu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mathsf{KL}(\nu_0 \| \pi^\star) < \infty$, $f \in \mathcal{H}$ and let $\nu_1 = (I - \eta f)_{\#}\nu_0$ with $\eta\|f\|_{\mathcal{H}} \leq \frac{\beta-1}{\beta B}$ for some $\beta > 1$. Then, the following holds,*

$$\mathsf{KL}(\mu_1 \| \pi^\star) \leq \mathsf{KL}(\mu_0 \| \pi^\star) - \eta \langle h_{\mu_0}, f\rangle + \frac{\eta^2(\beta^2 + L)B}{2}\|f\|^2_{\mathcal{H}}$$

B.1 Gelfand-Pettis Integrals for Reproducing Kernel Hilbert Spaces

The Gelfand-Pettis integral is a generalization of the Lebesgue integral to functions that take values in an arbitrary topological vector space. In this section, we describe the Gelfand-Pettis integral for an arbitrary Hilbert space $(V, \langle \cdot, \cdot \rangle_V)$ and refer the readers to Talagrand [40] for a more general treatment.

Let (X, Σ, λ) be a measure space and $(V, \langle \cdot, \cdot \rangle_V)$ be a Hilbert Space. A function $g : X \to V$ is said to be Gelfand-Pettis integrable if there exists a vector $w_g \in V$ such that $\langle u, w_g \rangle_V = \int_X \langle u, g(x) \rangle_V \, \mathsf{d}\lambda(x) \; \forall \, u \in V$. The vector w_g is called the Gelfand-Pettis integral of g

We now establish the following lemma for Gelfand-Pettis integrals with respect to the RKHS $\mathcal{H}$, which is a key component of our analysis of VP-SVGD.

Lemma 8. *Let μ be a probability measure on $\mathbb{R}^d$. Let $G : \mathbb{R}^d \times \mathbb{R}^d$ be a function such that for every $\mathbf{y} \in \mathbb{R}^d$, $G(.,\mathbf{y}) \in \mathcal{H}$ with $\|G(.,\mathbf{y})\|_{\mathcal{H}} \leq C$ holding μ-almost surely. Let $G_\mu(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim\mu}[G(\mathbf{x},\mathbf{y})]$. Then, the map $\psi : \mathbb{R}^d \to \mathcal{H}$ defined as $\psi(\mathbf{y}) = G(\cdot,\mathbf{y})$ is Gelfand-Pettis integrable and G_μ is the Gelfand-Pettis integral of ψ with respect to μ, i.e. $G_\mu \in \mathcal{H}$ and for any $f \in \mathcal{H}$, $\mathbb{E}_{\mathbf{y}\sim\mu}[\langle f, G(.,\mathbf{y})\rangle_{\mathcal{H}}] = \langle f, G_\mu \rangle_{\mathcal{H}}$*

Proof. Let $\Phi : \mathcal{H} \to \mathbb{R}$ denote the map $\Phi(f) = \mathbb{E}_{\mathbf{y}\sim\mu}[\langle f, G(.,\mathbf{y})\rangle_{\mathcal{H}}] \; \forall \; f \in \mathcal{H}$. By linearity of expectations and inner products, we note that Φ is a linear functional on $\mathcal{H}$. Furthermore, since $\|G(.,\mathbf{y})\|_{\mathcal{H}} \leq C$ holds μ-almost surely, we note that for any $f \in \mathcal{H}$, $|\Phi(f)| \leq \mathbb{E}_{\mathbf{y}\sim\mu}[|\langle f, G(.,\mathbf{y})\rangle_{\mathcal{H}}|] \leq C\|f\|_{\mathcal{H}}$ by Jensen's inequality and Cauchy Schwarz inequality for $\mathcal{H}$. We conclude that Φ is a bounded linear functional of $\mathcal{H}$. Thus, by Reisz Representation Theorem [10], there exists $g \in \mathcal{H}$ such that for any $f \in \mathcal{H}$, the following holds

$$\mathbb{E}_{\mathbf{y}\in\mu}[\langle f, G(.,\mathbf{y})\rangle_{\mathcal{H}}] = \langle f, g \rangle_{\mathcal{H}}$$

Hence, we conclude that the map ψ is Gelfand-Pettis integrable. We now use the reproducing property of $\mathcal{H}$ to show that $g = G_\mu$, i.e., G_μ is the Gelfand-Pettis integral of ψ. To this end, let $\mathbf{x} \in \mathbb{R}^d$ be arbitrary. Setting $f = k(\mathbf{x}, .)$ and using the fact that $g \in \mathcal{H}$, $G(.,\mathbf{y}) \in \mathcal{H}$ for any $\mathbf{y} \in \mathbb{R}^d$,

$$g(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\in\mu}[G(\mathbf{x},\mathbf{y})] = G_\mu(\mathbf{x})$$

Hence, $g = G_\mu$, i.e., $\mathbb{E}_{\mathbf{y}\sim\mu}[\langle f, G(.,\mathbf{y})\rangle_{\mathcal{H}}] = \langle f, G_\mu \rangle_{\mathcal{H}}$ □

C Analysis of VP-SVGD

In this section, we present our analysis of VP-SVGD. Throughout this section, we define the random function $g_t : \mathbb{R}^d \times \mathbb{R}^d$ as $g_t(\mathbf{x}) = \frac{1}{K}\sum_{l=0}^{K-1} h(\mathbf{x}, \mathbf{x}_t^{(Kt+l)})$ where $t \in \mathbb{N} \cup \{0\}$, K is the batch-size of VP-SVGD, and $h : \mathbb{R}^d \times \mathbb{R}^d$ is as defined in Section 2, i.e., $h(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{y})\nabla F(\mathbf{y}) - \nabla_2 k(\mathbf{x}, \mathbf{y})$.

After proving the key lemmas required for our analysis of VP-SVGD, we present the proof of Theorem 1 in Appendix C.4. We also present a high-probability version of Theorem 1 in Appendix C.5

C.1 Population Level Dynamics : Proof of Lemma 1

Proof. We now derive the population-limit dynamics of VP-SVGD for arbitrary batch-size K, and subsequently prove the descent lemma (i.e. Lemma 1) for VP-SVGD. The arguments of this section are a straightforward generalization of that used in Section 4.

To this end, we recall from Section 4 that the countably infinite number of particles $\mathbf{x}_0^{(l)}$, $l \in \mathbb{N} \cup \{0\}$ are i.i.d samples from the measure μ_0, which has a density w.r.t the Lebesgue measure. Thus, by the strong law of large numbers (Dudley [13, Theorem 11.4.1]), the empirical measure of $(\mathbf{x}_0^{(l)})_{l \geq 0}$ is almost surely equal to μ_0. Furthermore, we recall the filtration $\mathcal{F}_t$ defined in Section 4 as $\mathcal{F}_t = \sigma(\mathbf{x}_0^{(l)} \mid l \leq Kt - 1)$, $t \in \mathbb{N}$ with $\mathcal{F}_0$ being the trivial σ algebra. We now consider the following dynamics in $\mathbb{R}^d$:

$$\mathbf{x}_{t+1}^{(s)} = \mathbf{x}_t^{(s)} - \frac{\gamma}{K}\sum_{l=0}^{K-1} h(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(tK+l)}), \quad s \in \mathbb{N} \cup \{0\} \tag{13}$$

We note that the above updates are the same as that of VP-SVGD for $s \in \{0, \dots, KT + n - 1\}$. Now, for each time-step t, we focus on the lower triangular evolution, i.e., the time evolution of the particles $(\mathbf{x}_t^{(l)})_{l \geq Kt}$. From (13), we infer that for any $t \in \mathbb{N}$ and $s \geq Kt$, $\mathbf{x}_t^{(s)}$ depends only on $(\mathbf{x}_0^{(l)})_{l \leq Kt-1}$ and $\mathbf{x}_0^{(s)}$. Hence, there exists a measurable function H_t for every $t \in \mathbb{N}$ such that the following holds almost surely:

$$\mathbf{x}_t^{(s)} = H_t(\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(Kt-1)}, \mathbf{x}_0^{(s)}); \quad \forall\, s \geq Kt \tag{14}$$

Since $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(Kt-1)}, \mathbf{x}_0^{(s)} \overset{i.i.d.}{\sim} \mu_0$, we conclude from (14) that $(\mathbf{x}_t^{(s)})_{s \geq Kt}$ are i.i.d when conditioned on $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(Kt-1)}$. To this end, we define the random measure $\mu_t|\mathcal{F}_t$ as the law of $\mathbf{x}_t^{(Kt)}$ conditioned on $\mathcal{F}_t$, i.e. $\mu_t|\mathcal{F}_t$ is a probability kernel $\mu_t(\cdot; \mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(Kt-1)})$ with $\mu_0|\mathcal{F}_0 \coloneqq \mu_0$. By the strong law of large numbers, $\mu_t|\mathcal{F}_t$ is equal to the empirical measure of $(\mathbf{x}_t^{(l)})_{l \geq Kt}$ conditioned on $\mathcal{F}_t$. Furthermore, we infer from (13) that the particles satisfy the following:

$$\mathbf{x}_{t+1}^{(s)}(I - \gamma g_t)(\mathbf{x}_t^{(s)}), \quad s \geq K(t+1)$$

Recall that $\mathbf{x}_{t+1}^{(s)}|\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(K(t+1)-1)} \sim \mu_{t+1}|\mathcal{F}_{t+1}$ for any $s \geq K(t+1)$. Furthermore, from Equation (14), we note that for $s \geq K(t+1)$, $\mathbf{x}_t^{(s)}$ depends only on $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(Kt-1)}$ and $\mathbf{x}_0^{(s)}$, which implies that $\mathrm{Law}(\mathbf{x}_t^{(s)}|\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(K(t+1)-1)}) = \mathrm{Law}(\mathbf{x}_t^{(s)}|\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(Kt-1)}) = \mu_t|\mathcal{F}_t$. Finally, we note that g_t is an $\mathcal{F}_{t+1}$-measurable random function. With these insights, we conclude that the population-level dynamics of the lower triangular evolution in $\mathcal{P}_2(\mathbb{R}^d)$ is almost surely described by the following update:

$$\mu_{t+1}|\mathcal{F}_{t+1} = (I - \gamma g_t)_{\#}\mu_t|\mathcal{F}_t \tag{15}$$

Setting $\gamma\|g_t\|_{\mathcal{H}} \leq \frac{\beta-1}{\beta B}$ for some arbitrary $\beta > 1$ and applying Lemma 7 to the population-level update (15), we conclude that the following holds almost surely:

$$\mathsf{KL}\left(\mu_{t+1}|\mathcal{F}_{t+1} \| \pi^\star\right) \leq \mathsf{KL}\left(\mu_t|\mathcal{F}_t \| \pi^\star\right) - \gamma\left\langle h_{\mu_t|\mathcal{F}_t}, g_t\right\rangle_{\mathcal{H}} + \frac{\gamma^2(\beta^2 + L)B}{2}\|g_t\|_{\mathcal{H}}^2$$

□

C.2 Iterate Bounds : Proof of Lemma 2

To establish almost sure bounds on $\|g_t\|_{\mathcal{H}}$, we prove the following result which is stronger than Lemma 2.

Lemma 9 (Almost-Sure Iterate Bounds for VP-SVGD)**.** *Let Assumptions 1, 2, 3 and 4 be satisfied. Then, the following holds almost surely for any $s \in \mathbb{N} \cup \{0\}$ and $t \in (T+1)$ whenever $\gamma \leq 1/2A_1L$*

$$\|\mathbf{x}_t^{(s)}\| \leq \zeta_0 + \zeta_1(\gamma T)^{1/\alpha} + \zeta_2(\gamma^2 T)^{1/\alpha} + \zeta_3 R^{2/\alpha}$$
$$\|h(\cdot, \mathbf{x}_t^{(s)})\|_{\mathcal{H}} \leq \zeta_0 + \zeta_1(\gamma T)^{1/\alpha} + \zeta_2(\gamma^2 T)^{1/\alpha} + \zeta_3 R^{2/\alpha}$$
$$\|g_t\|_{\mathcal{H}} \leq \zeta_0 + \zeta_1(\gamma T)^{1/\alpha} + \zeta_2(\gamma^2 T)^{1/\alpha} + \zeta_3 R^{2/\alpha}$$

where $\zeta_0, \ldots, \zeta_3$ are problem-dependent constants that depend polynomially on $A_1, A_2, A_3, B, d_1, d_2, L$ for any fixed α.

Proof. Let $c_t^{(s)} = \frac{1}{K}\sum_{l=0}^{K-1} k(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(Kt+l)})$. Note that by Assumption 3, $c_t^{(s)} \geq 0$ Since $\mathbf{x}_{t+1}^{(s)} = \mathbf{x}_t^{(s)} - \gamma g_t(\mathbf{x}_t^{(s)})$, it follows from the smoothness of F that,

$$F(\mathbf{x}_{t+1}^{(s)}) - F(\mathbf{x}^{(s)}) \leq -\gamma \left\langle \nabla F(\mathbf{x}_t^{(s)}), g_t(\mathbf{x}_t^{(s)}) \right\rangle + \frac{\gamma^2 L}{2}\|g_t(\mathbf{x}_t^{(s)})\|^2 \tag{16}$$

By Lemma 6, we note that,

$$\begin{aligned}
-\gamma \left\langle \nabla F(\mathbf{x}_t^{(s)}), g_t(\mathbf{x}_t^{(s)}) \right\rangle &= -\frac{\gamma}{K}\sum_{l=0}^{K-1} \left\langle \nabla F(\mathbf{x}_t^{(s)}), h(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(tK+l)}) \right\rangle \\
&\leq \frac{\gamma}{K}\sum_{l=0}^{L-1} \left[-\tfrac{1}{2}k(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(tK+l)})\|\nabla F(\mathbf{x}_t^{(s)})\|^2 + L^2 A_1 + A_3\right] \\
&\leq -\frac{\gamma c_t^{(s)}}{2}\|\nabla F(\mathbf{x}_t^{(s)})\|^2 + \gamma L^2 A_1 + \gamma A_3
\end{aligned} \tag{17}$$

Moreover, by Jensen's Inequality and Lemma 6

$$\begin{aligned}
\|g_t(\mathbf{x}_t^{(s)})\|^2 &\leq \frac{1}{K}\sum_{l=0}^{K-1} \|h(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(Kt+l)})\|^2 \\
&\leq \frac{1}{K}\sum_{l=0}^{K-1} 2(A_1L/2 + A_2)^2 + 2k(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(tK+l)})^2\|F(\mathbf{x}_t^{(s)})\|^2 \\
&\leq \frac{1}{K}\sum_{l=0}^{K-1} 2(A_1L/2 + A_2)^2 + 2A_1 k(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(tK+l)})\|F(\mathbf{x}_t^{(s)})\|^2 \\
&\leq 2(A_1L/2 + A_2)^2 + 2A_1 c_t^{(s)}\|F(\mathbf{x}_t^{(s)})\|^2
\end{aligned} \tag{18}$$

Substituting (17) and (18) into (16), we obtain,

$$\begin{aligned}
F(\mathbf{x}_{t+1}^{(s)}) - F(\mathbf{x}_t^{(s)}) &\leq -\frac{\gamma c_t^{(s)}}{2}\|\nabla F(\mathbf{x}_t^{(s)})\|^2 + \gamma L^2 A_1 + \gamma A_3 \\
&\quad + \gamma^2 L(A_1L/2 + A_2)^2 + \gamma^2 L A_1 c_t^{(s)}\|F(\mathbf{x}_t^{(s)})\|^2 \\
&\leq -\frac{\gamma c_t^{(s)}}{2}(1 - 2A_1L\gamma)\|\nabla F(\mathbf{x}_t^{(s)})\|^2 + \gamma A_3 + \gamma L^2 A_1 + \gamma^2 L(A_1L/2 + A_2)^2 \\
&\leq \gamma A_3 + \gamma L^2 A_1 + \gamma^2 L(A_1L/2 + A_2)^2
\end{aligned}$$

where the last inequality uses the fact that $c_t^{(s)} \geq 0$ and $\gamma \leq 1/2A_1L$. Now, iterating through the above inequality, we obtain the following for any $t \in [T],\ s \in \mathbb{N} \cup \{0\}$

$$F(\mathbf{x}_t^{(s)}) \leq F(\mathbf{x}_0^{(s)}) + \gamma T L^2 A_1 + \gamma T A_3 + \gamma^2 T L(A_1L/2 + A_2)^2 \tag{19}$$

Furthermore, by Assumption 1

$$F(\mathbf{x}_0^{(s)}) \leq F(0) + \|\nabla F(0)\|\|\mathbf{x}_0^{(s)}\| + \frac{L}{2}\|\mathbf{x}_0^{(s)}\|^2$$
$$\leq F(0) + {}^1\!/_2 + L\|\mathbf{x}_0^{(s)}\|^2$$

Substituting the above inequality into (19), and using Assumption 2, we obtain the following for any $t \in [T],\ s \in \mathbb{N} \cup \{0\}$

$$d_1\|\mathbf{x}_t^{(s)}\|^\alpha - d_2 \leq F(\mathbf{x}_t^s) \leq F(0) + {}^1\!/_2 + L\|\mathbf{x}_0^{(s)}\|^2 + \gamma T L^2 A_1 + \gamma T A_3$$
$$+ \gamma^2 T L({}^{A_1 L}\!/_2 + A_2)^2$$

Rearranging and applying Assumption 4, we obtain

$$\|\mathbf{x}_t^{(s)}\| \leq d_1^{-1/\alpha}\left[F(0) + {}^1\!/_2 + LR^2 + \gamma T L^2 A_1 + \gamma T A_3 + \gamma^2 T L(A_1 L + A_2)^2\right]^{1/\alpha}$$
$$\leq \tilde{\zeta}_0 + \tilde{\zeta}_1(\gamma T)^{1/\alpha} + \tilde{\zeta}_2(\gamma^2 T)^{1/\alpha} + \tilde{\zeta}_3 R^{2/\alpha}$$

where $\tilde{\zeta}_0, \dots, \tilde{\zeta}_3$ are constants that depend polynomially on L, A_1, A_2, A_3, R. We note that, since $0 < \alpha \leq 2$, the above inequality also holds for $t = 0$.

Using the above inequality Lemma 6 and Assumption 1, we conclude that the following holds almost surely for any $t \in (T+1), s \in \mathbb{N} \cup \{0\}$

$$\|h(\cdot, \mathbf{x}_t^{(s)})\|_{\mathcal{H}} \leq BL\|\mathbf{x}_t^{(s)}\| + B\sqrt{L} + B$$
$$\leq \tilde{\eta}_0 + \tilde{\eta}_1(\gamma T)^{1/\alpha} + \tilde{\eta}_2(\gamma^2 T)^{1/\alpha} + \tilde{\eta}_3 R^{2/\alpha}$$

where $\tilde{\eta}_0, \dots, \tilde{\eta}_3$ are constants that depend polynomially on L, B, A_1, A_2, A_3, R. Using the above inequality, we conclude that the following also holds for any $t \in (T+1)$.

$$\|g_t\|_{\mathcal{H}} \leq \tilde{\eta}_0 + \tilde{\eta}_1(\gamma T)^{1/\alpha} + \tilde{\eta}_2(\gamma^2 T)^{1/\alpha} + \tilde{\eta}_3 R^{2/\alpha}$$

Taking $\zeta_i = \max\{\tilde{\zeta}_i, \tilde{\eta}_i\}$, the proof is complete. □

C.3 Controlling g_t in Expectation : Proof of Lemma 3

Proof. Let $\xi = \zeta_0 + \zeta_1(\gamma T)^{1/\alpha} + \zeta_2(\gamma^2 T)^{1/\alpha} + \zeta_3 R^{2/\alpha}$ where $\zeta_0, \dots, \zeta_3$ are as defined in Lemma 9. Recall that $g_t = \frac{1}{K}\sum_{l=0}^{K-1} h(., \mathbf{x}_t^{(Kt+l)})$. Since $\gamma \leq {}^1\!/_{2A_1 L}$, $\|h(\cdot, \mathbf{x}_t^{(Kt+l)})\|_{\mathcal{H}} \leq \xi$ holds almost surely y Lemma 9.

Consider any $l \in (K)$. Conditioned on the filtration $\mathcal{F}_t$, $\text{Law}(\mathbf{x}_t^{(Kt+l)}|\mathcal{F}_t) = \mu_t|\mathcal{F}_t$. Moreover, for any $\mathbf{x} \in \mathbb{R}^d$, $\mathbb{E}_{\mathbf{x}_t^{(Kt+l)}}[h(\mathbf{x}, \mathbf{x}_t^{(Kt+l)})|\mathcal{F}_t] = h_{\mu_t|\mathcal{F}_t}(\mathbf{x})$. Thus, from Lemma 8, we conclude that $h_{\mu_t|\mathcal{F}_t}$ is the Gelfand-Pettis Integral of the map $\mathbf{x} \to h(\mathbf{x}, \mathbf{x}_t^{(Kt+l)})$ with respect to $\mu_t|\mathcal{F}_t$. Hence, the following holds

$$\mathbb{E}_{\mathbf{x}_t^{(Kt+l)}}\left[\left\langle h(\cdot, \mathbf{x}_t^{(Kt+l)}), f\right\rangle_{\mathcal{H}}\middle|\mathcal{F}_t\right] = \langle h_{\mu_t|\mathcal{F}_t}, f\rangle_{\mathcal{H}}$$

In particular, setting $f = h_{\mu_t|\mathcal{F}_t}$ and using linearity of expectation, we conclude,

$$\mathbb{E}\left[\langle g_t, h_{\mu_t|\mathcal{F}_t}\rangle_{\mathcal{H}}|\mathcal{F}_t\right] = \frac{1}{K}\sum_{l=0}^{K-1}\mathbb{E}_{\mathbf{x}_t^{(Kt+l)}}\left[\left\langle h(\cdot, \mathbf{x}_t^{(Kt+l)}), h_{\mu_t|\mathcal{F}_t}\right\rangle_{\mathcal{H}}\middle|\mathcal{F}_t\right]$$
$$= \|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}}^2$$

To control $\mathbb{E}[\|g_t\|_{\mathcal{H}}^2|\mathcal{F}_t]$, we note that,

$$\|g_t\|_{\mathcal{H}}^2 = \frac{1}{K^2}\sum_{l_1, l_2=0}^{K-1}\left\langle h(\cdot, \mathbf{x}_t^{(Kt+l_1)}), h(\cdot, \mathbf{x}_t^{(Kt+l_2)})\right\rangle_{\mathcal{H}}$$
$$= \frac{1}{K^2}\sum_{l=0}^{K-1}\|h(\cdot, \mathbf{x}_t^{(Kt+l)})\|_{\mathcal{H}}^2 + \sum_{0 \leq l_1 \neq l_2 \leq K-1}\left\langle h(\cdot, \mathbf{x}_t^{(Kt+l_1)}), h(\cdot, \mathbf{x}_t^{(Kt+l_2)})\right\rangle_{\mathcal{H}}$$
$$\leq \frac{\xi^2}{K} + \sum_{0 \leq l_1 \neq l_2 \leq K-1}\left\langle h(\cdot, \mathbf{x}_t^{(Kt+l_1)}), h(\cdot, \mathbf{x}_t^{(Kt+l_2)})\right\rangle_{\mathcal{H}}$$

where the last inequality uses the fact that $\|h(\cdot, \mathbf{x}_t^{(Kt+l)})\|_{\mathcal{H}} \leq \xi$ almost surely as per Lemma 9.

To control the off-diagonal terms, let $i = Kt + l_1$ and $j = Kt + l_2$ for any arbitrary l_1, l_2 with $0 \leq l_1 \neq l_2 \leq K-1$. Conditioned on $\mathcal{F}_t$, $\mathbf{x}_t^{(i)}$ and $\mathbf{x}_t^{(j)}$ are i.i.d samples from $\mu_t|\mathcal{F}_t$. Thus, by Lemma 8 and Fubini's Theorem,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}_t^{(i)}, \mathbf{x}_t^{(j)}}\left[\left\langle h(\cdot, \mathbf{x}_t^{(i)}), h(\cdot, \mathbf{x}_t^{(j)})\right\rangle_{\mathcal{H}} | \mathcal{F}_t\right] &= \mathbb{E}_{\mathbf{x}_t^{(i)}}\left[\mathbb{E}_{\mathbf{x}_t^{(j)}}\left[\left\langle h(\cdot, \mathbf{x}_t^{(i)}), h(\cdot, \mathbf{x}_t^{(j)})\right\rangle_{\mathcal{H}} |\right]\right] \\ &= \mathbb{E}_{\mathbf{x}_t^{(i)}}\left[\left\langle h_{\mu_t|\mathcal{F}_t}, h(\cdot, \mathbf{x}_t^{(i)})\right\rangle_{\mathcal{H}} | \mathcal{F}_t\right] \\ &= \|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}}^2\end{aligned}$$

Thus, we conclude that,

$$\mathbb{E}\left[\|g_t\|_{\mathcal{H}}^2 | \mathcal{F}_t\right] \leq \|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}}^2 + \frac{\xi^2}{K}$$

□

C.4 Proof of Theorem 1

Proof. Let $\xi = \zeta_0 + \zeta_1(\gamma T)^{1/\alpha} + \zeta_2(\gamma^2 T)^{1/\alpha} + \zeta_3 R^{2/\alpha}$ where $\zeta_0, \ldots, \zeta_3$ are as defined in Lemma 9. Since $\gamma \leq 1/2A_1L$, $\|g_t\|_{\mathcal{H}} \leq \xi$ holds almost surely as per Lemma 9.

Since $\gamma\xi \leq 1/2B$, Lemma 1 ensures that the following holds almost surely

$$\mathsf{KL}\left(\mu_{t+1}|\mathcal{F}_{t+1}\|\pi^\star\right) \leq \mathsf{KL}\left(\mu_t|\mathcal{F}_t\|\pi^\star\right) - \gamma\left\langle h_{\mu_t|\mathcal{F}_t}, g_t\right\rangle_{\mathcal{H}} + \frac{\gamma^2(4+L)B}{2}\|g_t\|_{\mathcal{H}}^2$$

Taking conditional expectations w.r.t $\mathcal{F}_t$ on both sides and applying Lemma 3, we obtain,

$$\begin{aligned}\mathbb{E}\left[\mathsf{KL}\left(\mu_{t+1}|\mathcal{F}_{t+1}\|\pi^\star\right)|\mathcal{F}_t\right] &\leq \mathsf{KL}\left(\mu_t|\mathcal{F}_t\|\pi^\star\right) - \gamma\left(1 - \frac{\gamma(4+L)B}{2}\right)\|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}}^2 + \frac{\gamma^2(4+L)B\xi^2}{2K} \\ &\leq \mathsf{KL}\left(\mu_t|\mathcal{F}_t\|\pi^\star\right) - \frac{\gamma}{2}\|h_{\mu_t|\mathcal{F}_t}\|^2 + \frac{\gamma^2(4+L)B\xi^2}{2K} \\ &= \mathsf{KL}\left(\mu_t|\mathcal{F}_t\|\pi^\star\right) - \frac{\gamma}{2}\mathsf{KSD}_{\pi^\star}(\mu_t|\mathcal{F}_t||\pi^\star)^2 + \frac{\gamma^2(4+L)B\xi^2}{2K}\end{aligned}$$

where the second inequality uses the fact that $\gamma \leq 1/(4+L)B$. Taking expectations on both sides and rearranging,

$$\frac{\gamma}{2}\mathbb{E}\left[\mathsf{KSD}_{\pi^\star}(\mu_t|\mathcal{F}_t||\pi^\star)^2\right] \leq \mathbb{E}\left[\mathsf{KL}\left(\mu_t|\mathcal{F}_t\|\pi^\star\right) - \mathsf{KL}\left(\mu_{t+1}|\mathcal{F}_{t+1}\|\pi^\star\right)\right] + \frac{\gamma^2(4+L)B\xi^2}{2K}$$

Telescoping and averaging, we conclude,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathsf{KSD}_{\pi^\star}(\mu_t|\mathcal{F}_t||\pi^\star)^2\right] \leq \frac{2\mathsf{KL}\left(\mu_0|\mathcal{F}_0\|\pi^\star\right)}{\gamma T} + \frac{\gamma(4+L)B\xi^2}{K} \tag{20}$$

Now, recall from the proof of Lemma 1 in Section C.1 that for any $t \in [T]$ and $l \geq Kt$, $\mathbf{x}_t^{(l)}$ depends only on $\mathbf{x}_0^{(0)}, \ldots, \mathbf{x}_0^{(Kt-1)}, \mathbf{x}_0^{(l)}$, i.e., there exists a deterministic measurable function H_t such that $\mathbf{x}_t^{(l)} = H_t(\mathbf{x}_0^{(0)}, \ldots, \mathbf{x}_0^{(Kt-1)}, \mathbf{x}_0^{(l)})$ holds almost surely. We note that the output $\mathbf{Y} = (\mathbf{y}^{(0)}, \ldots, \mathbf{y}^{(n-1)})$ satisfies $\mathbf{y}^{(l)} = \mathbf{x}_S^{(KT+l)}\ \forall\ l \in (n)$, where $S \sim \mathsf{Uniform}((T))$ is sampled independently of everything else.

Thus, we infer that $\mathbf{y}^{(l)}$ depends only on $\mathbf{x}_0^{(0)}, \ldots, \mathbf{x}_0^{(KT-1)}, S, \mathbf{x}_0^{(KT+l)}$, i.e., there exists a deterministic measurable function G such that $\mathbf{y}^{(l)} = G(\mathbf{x}_0^{(0)}, \ldots, \mathbf{x}_0^{(KT-1)}, S, \mathbf{x}_0^{(KT+l)})$ for every $l \in (n)$. Since $\mathbf{x}_0^{(KT)}, \ldots, \mathbf{x}_0^{(KT+n-1)} \overset{i.i.d.}{\sim} \mu_0$, we infer that $\mathbf{y}^{(0)}, \ldots, \mathbf{y}^{(n-1)}$ are i.i.d when conditioned on $\mathbf{x}_0^{(0)}, \ldots, \mathbf{x}_0^{(KT-1)}, S$.

We now show that, when conditioned on $\mathbf{x}_0^{(0)}, \ldots, \mathbf{x}_0^{(KT-1)}, S$, $\mathbf{y}^{(l)}$ is distributed as $\bar{\mu}$, where $\bar{\mu}$ is the probability kernel defined as $\bar{\mu}(\cdot; \mathbf{x}_0^{(0)} = x_0, \ldots, \mathbf{x}_0^{(KT-1)} = \mathbf{x}_{KT-1}, S = s) := \mu_s(\cdot, \mathbf{x}_0^{(0)} =$

$\mathbf{x}_0, \dots, \mathbf{x}_0^{(Ks-1)} = \mathbf{x}_{Ks-1})$. For any arbitrary fixed $s \in (T)$, note that, under the event $S = s$, $\mathbf{y}^{(l)} = \mathbf{x}_s^{(KT+l)}$ for every $l \in (n)$. Thus, for any Borel measurable set $A \subseteq \mathbb{R}^d$, $\{\mathbf{y}^{(l)} \in A\} \cap \{S = s\} = \{\mathbf{x}_s^{(KT+l)} \in A\} \cap \{S = s\}$. For the sake of clarity, we denote the conditioning $\mathbf{x}_0^{(0)} = \mathbf{x}_0, \mathbf{x}_0^{(KT-1)} = \mathbf{x}_{KT-1}$ as $\mathcal{C}$, only in this proof. Since S is independent of $\mathbf{x}_t^{(l)}$ for every $t \in (T+1), l \in (KT+n)$, we infer the following:

$$
\begin{aligned}
\mathbb{P}\left(\{\mathbf{y}^{(l)} \in A\}\middle|\mathcal{C}, S = s\right) &= \frac{\mathbb{P}\left(\{\mathbf{y}^{(l)} \in A\} \cap \{S = s\}\middle|\mathcal{C}\right)}{\mathbb{P}\left(S = s\right)} \\
&= T\mathbb{P}\left(\{\mathbf{x}_s^{(KT+l)} \in A\} \cap \{S = s\}\middle|\mathcal{C}\right) \\
&= T\mathbb{P}\left(\{S = s\}\right)\mathbb{P}\left(\{\mathbf{x}^{(KT+l)} \in A\}\middle|\mathcal{C}\right) \\
&= \mathbb{P}\left(\{\mathbf{x}_s^{(KT+l)} \in A\}\middle|\mathcal{C}\right)
\end{aligned}
$$

As discussed above, $\mathbf{x}_s^{(KT+l)}$ depends only on $\mathbf{x}_0^{(0)}, \mathbf{x}_0^{(Ks-1)}, \mathbf{x}_0^{(KT+l)}$. It follows that $\mathbb{P}\left(\{\mathbf{x}_s^{(KT+l)} \in A\}|\mathcal{C}\right) = \mu_s(A; \mathbf{x}_0^{(0)} = x_0, \dots, \mathbf{x}_0^{(Ks-1)} = x_{Ks-1})$ and,

$$
\begin{aligned}
\mathbb{P}\left(\{\mathbf{y}^{(l)} \in A\}\middle|\mathcal{C}, S = s\right) &= \mu_s(A; \mathbf{x}_0^{(0)} = x_0, \dots, \mathbf{x}_0^{(Ks-1)} = x_{Ks-1}) \\
&= \bar{\mu}(A; \mathbf{x}_0^{(0)} = x_0, \dots, \mathbf{x}_0^{(KT-1)} = x_{Kt-1}, S = s)
\end{aligned}
$$

Thus, $\mathbf{y}^{(0)}, \dots, \mathbf{y}^{(n-1)}$ are i.i.d samples from $\bar{\mu}$ when conditioned on $\mathbf{x}_0^{(0)}, \dots \mathbf{x}_0^{(KT-1)}, S$.

We now obtain an upper bound on the expected squared KSD between $\bar{\mu}$ and $\pi^\star$. We recall from the proof of Lemma 1 in Section C.1 that, for any $t \in (T+1)$, conditioned on $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(Kt-1)}$, $(\mathbf{x}_t^{(l)})_{l \geq t}$ are i.i.d samples from $\mu_t|\mathcal{F}_t$ where $\mu_t|\mathcal{F}_t \coloneqq \mu_t(\cdot; \mathbf{x}_0^{(0)}, \mathbf{x}_0^{(Kt-1)})$. Hence, from (20), we conclude that,

$$
\begin{aligned}
\mathbb{E}[\mathsf{KSD}_{\pi^\star}(\bar{\mu}(\cdot; (\mathbf{x}_0^{(l)})_{l \in (KT)}, S) || \pi^\star)^2] &= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathbb{E}\left[\mathsf{KSD}_{\pi^\star}(\bar{\mu}(\cdot; (\mathbf{x}_0^{(l)})_{l \in (KT)}, S = t) || \pi^\star)^2 \middle| (\mathbf{x}_0^{(l)})_{l \in (KT)}\right]\right] \\
&= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathsf{KSD}_{\pi^\star}(\mu_t(\cdot; \mathbf{x}_0^{(0)}, \cdot, \mathbf{x}_0^{(Kt-1)}) || \pi^\star)^2\right] \\
&= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathsf{KSD}_{\pi^\star}(\mu_t|\mathcal{F}_t || \pi^\star)^2\right] \\
&\leq \frac{2\mathsf{KL}\left(\mu_0|\mathcal{F}_0 \| \pi^\star\right)}{\gamma T} + \frac{\gamma(4+L)B\xi^2}{K}
\end{aligned}
$$

where we use the fact that $S \sim \mathsf{Uniform}((T))$ is sampled independent of everything else. ☐

C.5 High-Probability Guarantees

We establish the convergence guarantee for VP-SVGD which holds with high probability, when conditioned on the virtual particles $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(KT-1)}$

Theorem 3 (VP-SVGD: High-Probability Rates). *Let the assumptions and parameter settings of Theorem 1 apply and let $\delta \in (0,1)$. Then, the following holds with probability at least $1-\delta$:*

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \mathsf{KSD}_{\pi^\star}(\mu_t|\mathcal{F}_t || \pi^\star)^2 &\leq \frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0 \| \pi^\star\right)}{\gamma T} + \frac{2\gamma(4+L)B\xi^2}{K} \\
&\quad + \frac{32\xi^2\log(2/\delta)}{KT} + 12\gamma(4+L)B\xi^2\sqrt{\frac{\log(2/\delta)}{T}}
\end{aligned}
$$

Let $\bar{\mu}(\cdot; \mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(KT-1)}, S)$ be the probability kernel defined in the statement of Theorem 1. Then, conditioned on $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(KT-1)}, S$, the n particles output by VP-SVGD are i.i.d samples from

$\bar{\mu}(\cdot;\mathbf{x}_0^{(0)},\dots,\mathbf{x}_0^{(KT-1)},S)$. *Furthermore, with probability at least* $1-\delta$

$$\mathbb{E}_S[\mathsf{KSD}_{\pi^\star}(\bar{\mu}(\cdot;\mathbf{x}_0^{(0)},\dots,\mathbf{x}_0^{(KT-1)},S)||\pi^\star)^2] \leq \frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0\|\pi^\star\right)}{\gamma T} + \frac{2\gamma(4+L)B\xi^2}{K} + \frac{32\xi^2\log(2/\delta)}{KT} + 12\gamma(4+L)B\xi^2\sqrt{\frac{\log(2/\delta)}{T}}$$

where $\mathbb{E}_S$ *denotes that the expectation is being taken only with respect to* $S\sim\mathsf{Uniform}((T))$

Proof. Following the same steps as Theorem 1, we note that the following holds almost surely.

$$\begin{aligned}\mathsf{KL}\left(\mu_{t+1}|\mathcal{F}_{t+1}\|\pi^\star\right) &\leq \mathsf{KL}\left(\mu_t|\mathcal{F}_t\|\pi^\star\right) - \gamma\left\langle h_{\mu_t|\mathcal{F}_t}, g_t\right\rangle_{\mathcal{H}} + \frac{\gamma^2(4+L)B}{2}\|g_t\|^2_{\mathcal{H}} \\ &\leq \mathsf{KL}\left(\mu_t|\mathcal{F}_t\|\pi^\star\right) - \frac{\gamma}{2}\|h_{\mu_t|\mathcal{F}_t}\|^2_{\mathcal{H}} + \gamma\left\langle h_{\mu_t|\mathcal{F}_t}, h_{\mu_t|\mathcal{F}_t} - g_t\right\rangle_{\mathcal{H}} \\ &+ \frac{\gamma^2(4+L)B\xi^2}{2K} + \frac{\gamma^2(4+L)B}{2}\left[\|g_t\|^2_{\mathcal{H}} - \|h_{\mu_t|\mathcal{F}_t}\|^2_{\mathcal{H}} - \frac{\xi^2}{K}\right]\end{aligned} \tag{21}$$

where the last inequality uses the fact that $\gamma \leq 1/(4+L)B$. We now define $\Delta_t^{(l)}$, Δ_t and r_t for $l\in(K),\ t\in(T)$ as follows:

$$\begin{aligned}\Delta_t^{(l)} &= \left\langle h_{\mu_t|\mathcal{F}_t}, h_{\mu_t|\mathcal{F}_t} - h(\cdot,\mathbf{x}_t^{(Kt+l)})\right\rangle_{\mathcal{H}} \\ \Delta_t &= \frac{1}{K}\sum_{l=0}^{K-1}\Delta_t^{(l)} = \left\langle h_{\mu_t|\mathcal{F}_t}, h_{\mu_t|\mathcal{F}_t} - g_t\right\rangle_{\mathcal{H}} \\ r_t &= \|g_t\|^2_{\mathcal{H}} - \|h_{\mu_t|\mathcal{F}_t}\|^2_{\mathcal{H}} - \frac{\xi^2}{K}\end{aligned}$$

Substituting the above into (21), we obtain the following:

$$\mathsf{KL}\left(\mu_{t+1}|\mathcal{F}_{t+1}\|\pi^\star\right) \leq \mathsf{KL}\left(\mu_t|\mathcal{F}_t\|\pi^\star\right) - \frac{\gamma}{2}\|h_{\mu_t|\mathcal{F}_t}\|^2_{\mathcal{H}} + \gamma\Delta_t + \frac{\gamma^2(4+L)B\xi^2}{2K} + \frac{\gamma^2(4+L)Br_t}{2}$$

Telescoping and averaging both sides, and using $\|h_{\mu_t|\mathcal{F}_t}\|^2_{\mathcal{H}} = \mathsf{KSD}_{\pi^\star}(\mu_t|\mathcal{F}_t||\pi^\star)^2$, we obtain the following:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathsf{KSD}_{\pi^\star}(\mu_t|\mathcal{F}_t||\pi^\star)^2 \leq \frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0\|\pi^\star\right)}{\gamma T} + \frac{2\gamma(4+L)B\xi^2}{K} + \frac{4}{T}\sum_{t=0}^{T-1}\left(\Delta_t - \frac{\|h_{\mu_t|\mathcal{F}_t}\|^2_{\mathcal{H}}}{4}\right) + \frac{2\gamma(4+L)B}{T}\sum_{t=0}^{T-1}r_t \tag{22}$$

We note that the first two terms are the same as that of the in-expectation guarantee for VP-SVGD in Theorem 1. The third and fourth term are random quantities that vanish in expectation. The remainder of our analysis upper bounds them with high probability.

We begin by deriving a high probability upper bound for the fourth term in (22). To this end, we note that, since $\gamma \leq 1/2A_1L$, $\|h(\cdot,\mathbf{x}_t^{(Kt+l)})\|_{\mathcal{H}} \leq \xi$ for any $t\in(T),\ l\in(K)$ as per Lemma 9. Furthermore, since $\mathbb{E}[h(\cdot,\mathbf{x}_t^{(Kt+l)}|\mathcal{F}_t] = h_{\mu_t|\mathcal{F}_t}$ (both pointwise and in the sense of the Gelfand-Pettis integral, see proof of Lemma 3 in Appendix C.3), it follows by Jensen's inequality that $\|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}} \leq \xi$. This further implies that $|r_t| \leq 3\xi^2$. Moreover, r_t is $\mathcal{F}_{t+1}$ measurable (as g_t is an $\mathcal{F}_{t+1}$ measurable random function) with $\mathbb{E}[r_t|\mathcal{F}_t] \leq 0$ (as per Lemma 3)

Thus, $S_t = \sum_{s=0}^{t-1} r_t$ is an $\mathcal{F}$-adapted supermartingale difference sequence with bounded increments. Thus, by the Hoeffding-Azuma inequality, we conclude that the following holds with probability at least $1-\delta/2$

$$\frac{1}{T}\sum_{t=0}^{T-1}r_t \leq 6\xi^2\sqrt{\frac{\log(2/\delta)}{T}} \tag{23}$$

We now proceed to control the third term in (22). Recall from the proof of Theorem 1 in Appendix C.4, that, for any fixed $t \in (T)$, $(\mathbf{x}_t^{(l)})_{l \in (KT)}$ are i.i.d when conditioned on $\mathcal{F}_t$. As discussed above, $\mathbb{E}[h(\cdot, \mathbf{x}_t^{(Kt+l)})] = h_{\mu_t|\mathcal{F}_t}$ in the sense of the Gelfand-Pettis integral, implying $\mathbb{E}[\Delta_t^{(l)}] = 0$. Moreover, $|\Delta^{(l)}\|_t \leq 2\xi \|h_{\mu_t|\mathcal{F}_t}\|$. Thus, when conditioned on $\mathcal{F}_t$, $\Delta_t^{(l)}$ are independent zero-mean bounded random variables. Hence, we conclude the following by Hoeffding's Lemma

$$\mathbb{E}\left[e^{\theta \Delta_t}|\mathcal{F}_t\right] \leq \prod_{l=0}^{K-1} \mathbb{E}[e^{\frac{\theta \Delta_t^{(l)}}{K}}|\mathcal{F}_t] \leq e^{\frac{2\theta^2 \xi^2}{K}\|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}}^2}, \quad \forall\, \theta \in \mathbb{R} \tag{24}$$

We now define the sequence M_t as follows, where $\lambda = {}^{K}/{8\xi^2}$

$$M_t = \exp(\sum_{s=0}^{t-1} \lambda \Delta_s - \tfrac{\lambda}{4}\|h_{\mu_s|\mathcal{F}_s}\|_{\mathcal{H}}^2)$$

Since g_t is $\mathcal{F}_{t+1}$ measurable, so is Δ_t, which implies M_t is $\mathcal{F}_{t+1}$ measurable. Furthermore,

$$\begin{aligned}\mathbb{E}[M_t|\mathcal{F}_t] &= M_{t-1} e^{-\frac{\lambda}{4}\|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}}^2}\mathbb{E}[e^{\lambda \Delta_t}|\mathcal{F}_t] \\ &\leq M_{t-1} e^{(-\frac{\lambda}{4} + \frac{2\lambda^2\xi^2}{K})\|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}}^2} \leq M_{t-1}\end{aligned}$$

Thus, M_t is an $\mathcal{F}$-adapted supermartingale sequence. Following the same steps, we conclude $E[M_1] \leq 1$, which implies $\mathbb{E}[M_T] \leq \mathbb{E}[M_1] \leq 1$. Thus, from Markov's Inequality

$$\mathbb{P}\left[\sum_{t=0}^{T-1} \Delta_t - \tfrac{1}{4}\|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}}^2 > x\right] \leq e^{-\lambda x}\mathbb{E}[M_T] \leq e^{-\lambda x}$$

Hence, the following holds with probability at least $1 - {}^{\delta}/{2}$.

$$\sum_{t=0}^{T-1} \Delta_t - \tfrac{1}{4}\|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}}^2 \leq \frac{8\xi^2}{K}\log({}^{2}/{\delta}) \tag{25}$$

Substituting (24) and (25) into (20) and taking a union bound, we conclude that the following holds with probability at least $1 - \delta$:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mathsf{KSD}_{\pi^\star}(\mu_t|\mathcal{F}_t||\pi^\star)^2 &\leq \frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0\|\pi^\star\right)}{\gamma T} + \frac{2\gamma(4+L)B\xi^2}{K} \\ &+ \frac{32\xi^2\log({}^{2}/{\delta})}{KT} + 12\gamma(4+L)B\xi^2\sqrt{\frac{\log({}^{2}/{\delta})}{T}}\end{aligned} \tag{26}$$

Recall from the proof of Theorem 1 in Appendix C.4 that the outputs $(\mathbf{y}^{(l)})_{l \in (n)}$ of VP-SVGD are i.i.d samples from the random measure $\bar{\mu}(\cdot; \mathbf{x}_0^{(0)}, \ldots, \mathbf{x}_{KT-1}^{(0)}, S)$ when conditioned on $\mathbf{x}_0^{(0)}, \ldots, \mathbf{x}_{KT-1}^{(0)}, S$. Furthermore, when conditioned on $S = t$, $\bar{\mu}(\cdot; \mathbf{x}_0^{(0)}, \ldots, \mathbf{x}_{KT-1}^{(0)}, S = t) = \mu_t|\mathcal{F}_t$. Thus, from (26), we conclude that, upon taking an expectation over $S \sim \mathsf{Uniform}((T))$ while conditioning on the virtual particles $\mathbf{x}_0^{(0)}, \ldots, \mathbf{x}_0^{(KT-1)}$, the following holds with probability at least $1 - \delta$:

$$\begin{aligned}\mathbb{E}_S[\mathsf{KSD}_{\pi^\star}(\bar{\mu}(\cdot; \mathbf{x}_0^{(0)}, \ldots, \mathbf{x}_0^{(KT-1)}, S)||\pi^\star)^2] &\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathsf{KSD}_{\pi^\star}(\mu_t|\mathcal{F}_t||\pi^\star)^2 \\ &\leq \frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0\|\pi^\star\right)}{\gamma T} + \frac{2\gamma(4+L)B\xi^2}{K} \\ &+ \frac{32\xi^2\log({}^{2}/{\delta})}{KT} + 12\gamma(4+L)B\xi^2\sqrt{\frac{\log({}^{2}/{\delta})}{T}}\end{aligned}$$

□

D Analysis of GB-SVGD

In this section, we present our analysis of GB-SVGD. For any $t \in (T)$, we use $\tilde{g}_t$ to denote the random function $\tilde{g}_t(\mathbf{x}) = \frac{1}{K}\sum_{r\in\mathcal{K}_t} h(\mathbf{x}, \mathbf{x}_t^{(r)})$ where $\mathcal{K}_t$ is the random batch of size K sampled at time-step t of GB-SVGD.

In order to prove Theorem 2, we first establish an almost-sure iterate bound for GB-SVGD which is similar to that of Lemma 9 for VP-SVGD.

Lemma 10 (Almost-Sure Iterate Bounds)**.** *Let Assumptions 1, 2, 3 and 4 be satisfied. Then, the following holds almost surely for any $s \in \mathbb{N}\cup\{0\}$ and $t \in (T+1)$ whenever $\gamma \leq 1/2A_1L$*

$$\|\mathbf{x}_t^{(s)}\| \leq \zeta_0 + \zeta_1(\gamma T)^{1/\alpha} + \zeta_2(\gamma^2 T)^{1/\alpha} + \zeta_3 R^{2/\alpha}$$
$$\|h(\cdot, \mathbf{x}_t^{(s)})\|_{\mathcal{H}} \leq \zeta_0 + \zeta_1(\gamma T)^{1/\alpha} + \zeta_2(\gamma^2 T)^{1/\alpha} + \zeta_3 R^{2/\alpha}$$
$$\|\tilde{g}_t\|_{\mathcal{H}} \leq \zeta_0 + \zeta_1(\gamma T)^{1/\alpha} + \zeta_2(\gamma^2 T)^{1/\alpha} + \zeta_3 R^{2/\alpha}$$

where $\zeta_0, \dots, \zeta_3$ are problem-dependent constants that depend polynomially on $A_1, A_2, A_3, B, d_1, d_2, L$ for any fixed α.

Proof. The proof of this Lemma is identical to that of Lemma 9. To this end, let $c_t^{(s)} = \frac{1}{K}\sum_{r\in\mathcal{K}_t} k(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(r)})$. Note that by Assumption 3, $c_t^{(s)} \geq 0$ Since $\mathbf{x}_{t+1}^{(s)} = \mathbf{x}_t^{(s)} - \gamma\tilde{g}_t(\mathbf{x}_t^{(s)})$, it follows from the smoothness of F that,

$$F(\mathbf{x}_{t+1}^{(s)}) - F(\mathbf{x}^{(s)}) \leq -\gamma\left\langle \nabla F(\mathbf{x}_t^{(s)}), \tilde{g}_t(\mathbf{x}_t^{(s)})\right\rangle + \frac{\gamma^2 L}{2}\|\tilde{g}_t(\mathbf{x}_t^{(s)})\|^2 \tag{27}$$

By Lemma 6, we note that,

$$\begin{aligned}
-\gamma\left\langle \nabla F(\mathbf{x}_t^{(s)}), \tilde{g}_t(\mathbf{x}_t^{(s)})\right\rangle &= -\frac{\gamma}{K}\sum_{r\in\mathcal{K}_t}\left\langle \nabla F(\mathbf{x}_t^{(s)}), h(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(r)})\right\rangle \\
&\leq \frac{\gamma}{K}\sum_{r\in\mathcal{K}_t}\left[-\tfrac{1}{2}k(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(r)})\|\nabla F(\mathbf{x}_t^{(s)})\|^2 + L^2A_1 + A_3\right] \\
&\leq -\frac{\gamma c_t^{(s)}}{2}\|\nabla F(\mathbf{x}_t^{(s)})\|^2 + \gamma L^2 A_1 + \gamma A_3
\end{aligned} \tag{28}$$

Moreover, by Jensen's Inequality and Lemma 6

$$\begin{aligned}
\|\tilde{g}_t(\mathbf{x}_t^{(s)})\|^2 &\leq \frac{1}{K}\sum_{r\in\mathcal{K}_t}\|h(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(r)})\|^2 \\
&\leq \frac{1}{K}\sum_{r\in\mathcal{K}_t} 2(A_1L/2 + A_2)^2 + 2k(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(r)})^2\|F(\mathbf{x}_t^{(s)})\|^2 \\
&\leq \frac{1}{K}\sum_{r\in\mathcal{K}_t} 2(A_1L/2 + A_2)^2 + 2A_1k(\mathbf{x}_t^{(s)}, \mathbf{x}_t^{(r)})\|F(\mathbf{x}_t^{(s)})\|^2 \\
&\leq 2(A_1L/2 + A_2)^2 + 2A_1c_t^{(s)}\|F(\mathbf{x}_t^{(s)})\|^2
\end{aligned} \tag{29}$$

Substituting (28) and (29) into (27), we obtain,

$$\begin{aligned}
F(\mathbf{x}_{t+1}^{(s)}) - F(\mathbf{x}_t^{(s)}) &\leq -\frac{\gamma c_t^{(s)}}{2}\|\nabla F(\mathbf{x}_t^{(s)})\|^2 + \gamma L^2A_1 + \gamma A_3 \\
&\quad + \gamma^2 L(A_1L/2 + A_2)^2 + \gamma^2 LA_1c_t^{(s)}\|F(\mathbf{x}_t^{(s)})\|^2 \\
&\leq -\frac{\gamma c_t^{(s)}}{2}(1 - 2A_1L\gamma)\|\nabla F(\mathbf{x}_t^{(s)})\|^2 + \gamma A_3 + \gamma L^2A_1 + \gamma^2 L(A_1L/2 + A_2)^2 \\
&\leq \gamma A_3 + \gamma L^2 A_1 + \gamma^2 L(A_1L/2 + A_2)^2
\end{aligned}$$

where the last inequality uses the fact that $c_t^{(s)} \geq 0$ and $\gamma \leq 1/2A_1L$. Now, iterating through the above inequality, we obtain the following for any $t \in [T],\ s \in \mathbb{N}\cup\{0\}$

$$F(\mathbf{x}_t^{(s)}) \leq F(\mathbf{x}_0^{(s)}) + \gamma TL^2A_1 + \gamma TA_3 + \gamma^2 TL(A_1L/2 + A_2)^2 \tag{30}$$

Furthermore, by Assumption 1

$$F(\mathbf{x}_0^{(s)}) \leq F(0) + \|\nabla F(0)\| \|\mathbf{x}_0^{(s)}\| + \frac{L}{2}\|\mathbf{x}_0^{(s)}\|^2$$
$$\leq F(0) + 1/2 + L\|\mathbf{x}_0^{(s)}\|^2$$

Substituting the above inequality into (30), and using Assumption 2, we obtain the following for any $t \in [T],\ s \in \mathbb{N} \cup \{0\}$

$$d_1\|\mathbf{x}_t^{(s)}\|^\alpha - d_2 \leq F(\mathbf{x}_t^s) \leq F(0) + 1/2 + L\|\mathbf{x}_0^{(s)}\|^2 + \gamma T L^2 A_1 + \gamma T A_3 + \gamma^2 T L(A_1 L/2 + A_2)^2$$

Rearranging and applying Assumption 4, we obtain

$$\|\mathbf{x}_t^{(s)}\| \leq d_1^{-1/\alpha}\left[F(0) + 1/2 + LR^2 + \gamma T L^2 A_1 + \gamma T A_3 + \gamma^2 T L(A_1 L + A_2)^2\right]^{1/\alpha}$$
$$\leq \tilde{\zeta}_0 + \tilde{\zeta}_1(\gamma T)^{1/\alpha} + \tilde{\zeta}_2(\gamma^2 T)^{1/\alpha} + \tilde{\zeta}_3 R^{2/\alpha}$$

where $\tilde{\zeta}_0, \dots, \tilde{\zeta}_3$ are constants that depend polynomially on L, A_1, A_2, A_3, R. We note that, since $0 < \alpha \leq 2$, the above inequality also holds for $t = 0$.

Using the above inequality Lemma 6 and Assumption 1, we conclude that the following holds almost surely for any $t \in (T+1), s \in \mathbb{N} \cup \{0\}$

$$\|h(\cdot, \mathbf{x}_t^{(s)})\|_{\mathcal{H}} \leq BL\|\mathbf{x}_t^{(s)}\| + B\sqrt{L} + B$$
$$\leq \tilde{\eta}_0 + \tilde{\eta}_1(\gamma T)^{1/\alpha} + \tilde{\eta}_2(\gamma^2 T)^{1/\alpha} + \tilde{\eta}_3 R^{2/\alpha}$$

where $\tilde{\eta}_0, \dots, \tilde{\eta}_3$ are constants that depend polynomially on L, B, A_1, A_2, A_3, R. Using the above inequality, we conclude that the following also holds for any $t \in (T+1)$.

$$\|\tilde{g}_t\|_{\mathcal{H}} \leq \tilde{\eta}_0 + \tilde{\eta}_1(\gamma T)^{1/\alpha} + \tilde{\eta}_2(\gamma^2 T)^{1/\alpha} + \tilde{\eta}_3 R^{2/\alpha}$$

Taking $\zeta_i = \max\{\tilde{\zeta}_i, \tilde{\eta}_i\}$, the proof is complete. □

D.1 Proof of Theorem 2

Proof. Let $\xi = \zeta_0 + \zeta_1(\gamma T)^{1/\alpha} + \zeta_2(\gamma^2 T)^{1/\alpha} + \zeta_3 R^{2/\alpha}$ where $\zeta_0, \dots, \zeta_3$ are constants as described in Lemma 9 and Lemma 10. Since the assumptions and parameter settings of Theorem 1 holds, $\gamma \leq 1/2A_1L$ and thus, by Lemma 9 and Lemma 10, the particles output by VP-SVGD and GB-SVGD are bounded as $\|\mathbf{y}^{(l)}\| \leq \xi$ and $\|\bar{\mathbf{y}}^{(l)}\| \leq \xi$.

Let $\mathbf{Y} = (\mathbf{y}^{(0)}, \dots, \mathbf{y}^{(n-1)})$ and $\bar{\mathbf{Y}} = (\bar{\mathbf{y}}^{(0)}, \dots, \bar{\mathbf{y}}^{(n-1)})$ denote the outputs of VP-SVGD and GB-SVGD. Let $\hat{\mu}^{(n)} = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{\mathbf{y}^{(i)}}$ and $\hat{\nu}^{(n)} = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{\bar{\mathbf{y}}^{(i)}}$ be their respective empirical distributions. We shall now explicitly construct a coupling between the inputs of VP-SVGD and GB-SVGD such that the first $n - KT$ particles of their respective outputs are equal. This in turn will allow us to control the expected squared KSD between $\hat{\mu}^{(n)}$ and $\hat{\nu}^{(n)}$.

To this end, let $\mathcal{E}$ denote the event that each random batch $\mathcal{K}_t$ of GB-SVGD is disjoint and contains unique elements for every $t \in (T)$. Subsequently, let $\mathcal{K}$ denote the set of all indices that were chosen to be part of some random batch $\mathcal{K}_t$. Let Λ be a uniformly random permutation over $\{0, \dots, n-1\}$. We note that, conditioned on $\mathcal{E}$, the distribution of the random set $\mathcal{K}$ is the same as the distribution of $\{\Lambda(0), \dots, \Lambda(KT-1)\}$. We can couple a uniformly random permutation Λ and $\mathcal{K}_t$ for $0 \leq t \leq T$ such that under the event $\mathcal{E}$, $\mathcal{K} = \{\Lambda(0), \dots, \Lambda(KT-1)\}$ and $\{\Lambda(tK), \dots, \Lambda((t+1)K-1)\}$ is the random batch $\mathcal{K}_t$. Thus, under the event $\mathcal{E}$, one can couple a uniformly random permutation Λ and $\mathcal{K}_t$ for $t \in (T)$ such that $\mathcal{K} = \{\Lambda(0), \dots, \Lambda(KT-1)\}$ and $\mathcal{K}_t = \{\Lambda(tK), \dots, \Lambda((t+1)K-1)\}$

With this insight, we couple VP-SVGD and GB-SVGDas follows. We note that, the random batch $\mathcal{K}_t$ in GB-SVGD is sampled independently of the initial particles. To this end, let $\bar{\mathbf{x}}_0^{(0)}, \dots, \bar{\mathbf{x}}_0^{(n-1)} \overset{i.i.d.}{\sim} \mu_0$, and let the random batches $\mathcal{K}_t$ and permutation Λ be jointly distributed as described above, independently of $\bar{\mathbf{x}}_0^{(0)}, \dots, \bar{\mathbf{x}}_0^{(n-1)}$, i.e.

$$\Lambda \sim \mathsf{Uniform}(\mathbb{S}_{(n)}), \quad \mathcal{K}_t = \{\Lambda(tK), \dots, \Lambda((t+1)K-1)\}, \quad t \in (T)$$

We now define $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(KT+n-1)}$ as:

$$\mathbf{x}_0^{(l)} := \begin{cases} = \bar{\mathbf{x}}_0^{(\Lambda(l))} & \text{for } 0 \leq l \leq n-1 \\ \sim \mu_0 \text{ independent of everything else} & \text{for } n \leq l \leq KT+n-1 \end{cases} \tag{31}$$

Let $\bar{\mathbf{x}}_0^{(0)}, \dots, \bar{\mathbf{x}}_0^{(n-1)}$ and $\mathcal{K}_t$ as the initialization and random batches for GB-SVGD, and let $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(KT+n-1)}$ be the initialization for GB-SVGD. We first show that this construction is indeed a valid coupling between VP-SVGD and GB-SVGD.

Claim 1. *Conditioned on $\mathcal{E}$, the inputs to VP-SVGD and GB-SVGD, as constructed above is a valid coupling, i.e., the marginal distribution of $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(KT+n-1)}$ is equal to the distribution of initial particles in VP-SVGD, and the marginal distribution of $\bar{\mathbf{x}}_0^{(0)}, \dots, \bar{\mathbf{x}}_0^{(n-1)}, (\mathcal{K}_t)_{t \in (T)}$ is the same as the distribution of initial particles and random batches in $\mathcal{K}_t$*

Proof. By construction $\bar{\mathbf{x}}_0^{(0)}, \dots, \bar{\mathbf{x}}_0^{(n-1)} \overset{i.i.d.}{\sim} \mu_0$. Moreover, conditioned on $\mathcal{E}$, the distribution of $\mathcal{K}_t = \{\Lambda(tK), \dots, \Lambda((t+1)K-1)\}$, has the distribution of a uniform random batch of size K since $\Lambda \sim \mathsf{Uniform}(\mathbb{S}_n)$. Furthermore, since Λ is sampled independently of $\bar{\mathbf{x}}_0^{(0)}, \dots, \bar{\mathbf{x}}_0^{(n-1)}$, $\mathcal{K}_t$ is independent of $\bar{\mathbf{x}}_0^{(0)}, \dots, \bar{\mathbf{x}}_0^{(n-1)}$ for any $t \in (T)$. Thus, the coupling constructed above has the correct marginal with respect to GB-SVGD.

To establish the same for VP-SVGD, we note that by (31), $\mathbf{x}_0^{(n)}, \dots, \mathbf{x}_0^{(KT+n-1)} \overset{i.i.d.}{\sim} \mu_0$, *sampled independently of everything else*. Moreover, since $\bar{\mathbf{x}}_0^{(0)}, \dots, \bar{\mathbf{x}}_0^{(n-1)} \overset{i.i.d.}{\sim} \mu_0$, we infer that $\bar{\mathbf{x}}_0^{(\Lambda(0))}, \dots, \bar{\mathbf{x}}_0^{(\Lambda(n-1))} \overset{i.i.d.}{\sim} \mu_0$ for any arbitrary permutation $\Lambda \in \mathbb{S}_n$. From this, and (31), we conclude that $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(KT+n-1)} \overset{i.i.d.}{\sim} \mu_0$. Hence, the coupling constructed above has the correct marginal with respect to VP-SVGD. $\square$

We now show that, under the constructed coupling, the time-evolution of the particles of VP-SVGD and GB-SVGD satisfy $\bar{\mathbf{x}}_t^{(\Lambda(l))} = \mathbf{x}_t^{(l)},\ KT \leq l \leq n-1, t \in (T+1)$, when conditioned on the event $\mathcal{E}$.

Claim 2. *Let the inputs to VP-SVGD and GB-SVGD be coupled as per the construction above. Then, conditioned on the event $\mathcal{E}$, the particles $\mathbf{x}_t^{(s)}$ and $\bar{\mathbf{x}}_t^{(s)}$ of VP-SVGD and GB-SVGD respectively, satisfy $\bar{\mathbf{x}}_t^{(\Lambda(l))} = \mathbf{x}_t^{(l)}$ for every $KT \leq l \leq n-1$ and $0 \leq t \leq T$*

Proof. We prove this by an inductive argument. Clearly, the claim holds for $t = 0$ by the construction of our coupling. Assume it holds for some arbitrary $t \in (T)$. Now, writing the update equation for GB-SVGD for $KT \leq l \leq n-1$,

$$\begin{aligned} \bar{\mathbf{x}}_{t+1}^{(\Lambda(l))} &= \bar{\mathbf{x}}_t^{(\Lambda(l))} - \frac{\gamma}{K} \sum_{r \in \mathcal{K}_t} h(\bar{\mathbf{x}}_t^{(\Lambda(l))}, \bar{\mathbf{x}}_t^{(r)}) \\ &= \bar{\mathbf{x}}_t^{(\Lambda(l))} - \frac{\gamma}{K} \sum_{l=0}^{K-1} h(\bar{\mathbf{x}}_t^{(\Lambda(l))}, \bar{\mathbf{x}}_t^{(\Lambda(Kt+l))}) \\ &= \mathbf{x}_t^{(l)} - \frac{\gamma}{K} \sum_{l=0}^{K-1} h(\mathbf{x}_t^{(l)}, \mathbf{x}_t^{(Kt+l)}) = \mathbf{x}_t^{(l+1)} \end{aligned}$$

where the second equality uses the fact that $\mathcal{K}_t = \{\Lambda(tK), \dots, \Lambda((t+1)K-1)\}$ when conditioned on $\mathcal{E}$ and the third equality uses the induction hypothesis $\bar{\mathbf{x}}_t^{(\Lambda(l))} = \mathbf{x}_t^{(l)}$ for $KT \leq l \leq n-1$. Hence, the claim is proven true by induction. $\square$

Equipped with the above coupling between the inputs of VP-SVGD and GB-SVGD, one can now couple their outputs by sampling an $S \sim \mathsf{Uniform}((n))$ and using this sampled S as the random timestep chosen by both VP-SVGD (Step 6 in Algorithm 1) and GB-SVGD (Step 7 in Algorithm 2) that are run with the coupled input constructed above. It is easy to see that this results in a coupling of the outputs $\mathbf{Y}$ and $\bar{\mathbf{Y}}$ of VP-SVGD and GB-SVGD respectively. Furthermore, by Claim 2, we note that, conditioned on the event $\mathcal{E}$, $\mathbf{y}^{(l-TK)} = \bar{\mathbf{y}}^{(\Lambda(l))}$ for every $KT \leq l \leq n-1$. We now define the permutation $\tau \in \mathbb{S}_{(n)}$ as follows:

$$\tau(\Lambda(l)) = \begin{cases} l + n - KT & \text{for } 0 \leq l \leq KT - 1 \\ l - KT & \text{for } KT \leq l \leq n-1 \end{cases} \tag{32}$$

It follows that $\bar{\mathbf{y}}^{\tau(l)} = \mathbf{y}^{(l)}$ for $KT \leq l \leq n-1$. Thus, by definition of Kernel Stein Discrepancy (Definition 1), we can infer that the following holds when conditioned on the event $\mathcal{E}$

$$\begin{aligned}
\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\hat{\mu}^{(n)})|\mathcal{E}] &= \mathbb{E}\left[\|h_{\hat{\nu}^{(n)}} - h_{\hat{\mu}^{(n)}}\|^2_{\mathcal{H}} \mid \mathcal{E}\right] \\
&= \mathbb{E}\left[\|\frac{1}{n}\sum_{l=0}^{n-1} h(\cdot, \bar{\mathbf{y}}^{(l)}) - \frac{1}{n}\sum_{l=0}^{n-1} h(\cdot, \mathbf{y}^{(l)})\|^2_{\mathcal{H}} \mid \mathcal{E}\right] \\
&= \frac{1}{n^2}\mathbb{E}[\|\sum_{l=0}^{n-1} h(\cdot, \bar{\mathbf{y}}^{(\tau(l))}) - h(\cdot, \mathbf{y}^{(l)})\|^2_{\mathcal{H}} \mid \mathcal{E}] \\
&= \frac{1}{n^2}\mathbb{E}[\|\sum_{l=0}^{KT-1} h(\cdot, \bar{\mathbf{y}}^{(\tau(l))}) - h(\cdot, \mathbf{y}^{(l)})\|^2_{\mathcal{H}} \mid \mathcal{E}] \\
&\leq \frac{KT}{n^2}\sum_{l=0}^{KT-1} \mathbb{E}[\|h(\cdot, \bar{\mathbf{y}}^{(\tau(l))}) - h(\cdot, \mathbf{y}^{(l)})\|^2_{\mathcal{H}} \mid \mathcal{E}] \\
&\leq \frac{2K^2T^2\xi^2}{n^2}
\end{aligned} \tag{33}$$

where the second step uses the permutation invariance of summation, the third step uses the fact that $\bar{\mathbf{y}}^{\tau(l)} = \bar{\mathbf{y}}^{(l)}$ for $KT \leq l \leq n-1$, the fourth step uses the convexity of $\|\cdot\|^2_{\mathcal{H}}$ and the last step uses the almost-sure iterate bounds of Lemma 9 and 10

Under the event $\mathcal{E}^c$, we directly apply the almost-sure iterate bounds of Lemma 9 and 10 to obtain the following:

$$\begin{aligned}
\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\hat{\mu}^{(n)})|\mathcal{E}^c] &= \mathbb{E}\left[\|h_{\hat{\mu}^{(n)}} - h_{\hat{\nu}^{(n)}}\|^2_{\mathcal{H}} \mid \mathcal{E}^c\right] \\
&= \frac{1}{n^2}\mathbb{E}[\|\sum_{l=0}^{n-1} h(\cdot, \bar{\mathbf{y}}^{(l)}) - h(\cdot, \mathbf{y}^{(l)})\|^2_{\mathcal{H}} \mid \mathcal{E}^c] \\
&\leq 2\xi^2
\end{aligned} \tag{34}$$

From Equations (33) and (34), it follows that:

$$\begin{aligned}
\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\hat{\mu}^{(n)})] &= \mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\hat{\mu}^{(n)})|\mathcal{E}]\mathbb{P}(\mathcal{E}) + \mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\hat{\mu}^{(n)})|\mathcal{E}^c]\mathbb{P}(\mathcal{E}^c) \\
&\leq \frac{2K^2T^2\xi^2}{n^2}\mathbb{P}(\mathcal{E}) + 2\xi^2\mathbb{P}(\mathcal{E}^c)
\end{aligned}$$

Recall that $P(\mathcal{E}) = 1$ under sampling without replacement and $P(\mathcal{E}) = 1 - \frac{K^2T^2}{n}$ under sampling with replacement. Thus, we conclude that the following holds under the constructed coupling of $\mathbf{Y}$ and $\bar{\mathbf{Y}}$

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\hat{\mu}^{(n)})] \leq \begin{cases} \frac{2K^2T^2\xi^2}{n^2} & \text{(without replacement sampling)} \\ \frac{2K^2T^2\xi^2}{n^2}\left(1 - \frac{K^2T^2}{n}\right) + \frac{2K^2T^2\xi^2}{n} & \text{(with replacement sampling)} \end{cases}$$

□

E Finite-Particle Convergence Guarantees for VP-SVGD and GB-SVGD

In this section, we show that the empirical measure of the particles output by VP-SVGD and GB-SVGD rapidly converge to the target distribution $\pi^\star$ in KSD. To this end, we prove the finite-particle convergence rates for VP-SVGD in Appendix E.1 and that of GB-SVGD in Appendix E.2. Finally, we compare the oracle complexity (i.e., the number of evaluations of ∇F) of VP-SVGD and GB-SVGD to that of SVGD in Appendix E.3

E.1 VP-SVGD

Corollary 2 (VP-SVGD : Fast Finite-Particle Convergence). *Let the assumptions and parameter settings of Theorem 1 be satisfied. Let $\hat{\mu}^{(n)}$ denote the empirical measure of the n particles output by VP-SVGD.*

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\mu}^{(n)}||\pi^\star)] \leq \frac{\xi^2}{n} + \frac{2\mathsf{KL}\left(\mu_0|\mathcal{F}_0\|\pi^\star\right)}{\gamma T} + \frac{\gamma B(4+L)\xi^2}{K}$$

where ξ is as defined in Theorem 1. Setting $R = \sqrt{d/L}, \gamma = O(\frac{(Kd)^\eta}{T^{1-\eta}})$ with $\eta = \frac{\alpha}{2(1+\alpha)}$ and $KT = d^{\frac{\alpha}{2+\alpha}} n^{\frac{2(1+\alpha)}{2+\alpha}}$ suffices to ensure,

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\mu}^{(n)}||\pi^\star)] \leq O\left(\frac{d^{\frac{2}{2+\alpha}}}{n^{\frac{\alpha}{2+\alpha}}} + \frac{d^{2/\alpha}}{n}\right)$$

Proof. Recall from Algorithm 1 that the outputs of VP-SVGD are $\mathbf{x}_S^{(KT)}, \dots, \mathbf{x}_S^{(KT+n-1)}$ where $S \sim \mathsf{Uniform}(\{0, \dots, T-1\})$. Hence, their empirical measure $\hat{\mu}^{(n)}$ is given by $\hat{\mu}^{(n)} = \frac{1}{n}\sum_{l=0}^{n-1}\delta_{\mathbf{x}_S^{(KT+l)}}$. From the definition of the Kernel Stein Discrepancy (Definition 1), it follows that,

$$\mathsf{KSD}^2_{\pi^\star}(\hat{\mu}^{(n)}||\pi^\star) = \|h_{\hat{\mu}^{(n)}}\|^2_{\mathcal{H}} = \|\tfrac{1}{n}\sum_{l=1}^{N} h(\cdot, \mathbf{x}_S^{(KT+l)})\|^2_{\mathcal{H}} \tag{35}$$

For the sake of clarity, only in this proof, we use $\mathcal{C}$ to denote the conditioning on the virtual particles $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(KT-1)}$. Now, consider any arbitrary $t \in \{0, \dots, T-1\}$. Taking conditional expectations on both sides of Equation (35) by conditioning on $\mathcal{C}$ and the event $\{S = t\}$, we obtain the following:

$$\begin{aligned}\mathbb{E}\left[\mathsf{KSD}^2_{\pi^\star}(\hat{\mu}^{(n)}||\pi^\star) \mid \mathcal{C}, S=t\right] &= \mathbb{E}\left[\|\tfrac{1}{n}\sum_{l=0}^{n-1} h(\cdot, \mathbf{x}_S^{(KT+l)})\|^2_{\mathcal{H}} \mid \mathcal{C}, S=t\right] \\ &= \mathbb{E}\left[\|\tfrac{1}{n}\sum_{l=0}^{n-1} h(\cdot, \mathbf{x}_t^{(KT+l)})\|^2_{\mathcal{H}} \mid (\mathbf{x}_0^{(s)})_{0\leq s\leq KT-1}\right] \end{aligned}\tag{36}$$

Recall from Equation (14) in Appendix C.1 that for any $l \in \{0, \dots, n-1\}$ $\mathbf{x}_t^{(KT+l)}$ depends only on $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(Kt-1)}$ and $\mathbf{x}_0^{(KT+l)}$. Furthermore, from Appendix C.1, we recall that the filtration $\mathcal{F}_t$ is defined as $\mathcal{F}_t = \sigma(\{\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(Kt-1)}\})$. It follows that,

$$\begin{aligned}\mathbb{E}\left[\|\tfrac{1}{n}\sum_{l=0}^{n-1} h(\cdot, \mathbf{x}_t^{(KT+l)})\|^2_{\mathcal{H}} \mid (\mathbf{x}_0^{(s)})_{0\leq s\leq KT-1}\right] &= \mathbb{E}\left[\|\tfrac{1}{n}\sum_{l=1}^{N} h(\cdot, \mathbf{x}_t^{(KT+l)})\|^2_{\mathcal{H}} \mid (\mathbf{x}_0^{(s)})_{0\leq s\leq Kt-1}\right] \\ &= \mathbb{E}\left[\|\tfrac{1}{n}\sum_{l=0}^{n-1} h(\cdot, \mathbf{x}_t^{(KT+l)})\|^2_{\mathcal{H}} \mid \mathcal{F}_t\right]\end{aligned}\tag{37}$$

To control $\mathbb{E}\left[\|\frac{1}{n}\sum_{l=0}^{n-1} h(\cdot, \mathbf{x}_t^{(KT+l)})\|^2_{\mathcal{H}} \mid \mathcal{F}_t\right]$, we apply the arguments used in the proof of Lemma 3. To this end, note that when conditioned on the virtual particles $\mathbf{x}_0^{(0)}, \dots, \mathbf{x}_0^{(Kt-1)}$, the particles $\mathbf{x}_t^{(KT)}, \dots, \mathbf{x}_t^{(KT+n-1)} \overset{i.i.d.}{\sim} \mu_t|\mathcal{F}_t$. Furthermore, since $\gamma \leq 1/2A_1 L$ (as per the parameter settings of Theorem 1), $\|h(\cdot, \mathbf{x}_t^{(KT+l)})\|_{\mathcal{H}} \leq \xi \ \forall\ l \in (n)$ by Lemma 6. Finally, $\mathbb{E}[h(\mathbf{x}, \mathbf{x}_t^{(KT+l)})|\mathcal{F}_t] = h_{\mu_t|\mathcal{F}_t}(\mathbf{x}) \ \forall\ l \in (n),\ \mathbf{x} \in \mathbb{R}^d$. Hence, from Lemma 8, we conclude that $h_{\mu_t|\mathcal{F}_t}$ is the Gelfand-Pettis integral of the map $\mathbf{x} \to h(\mathbf{x}, \mathbf{x}_t^{(KT+l)})$ with respect to the measure $\mu_t|\mathcal{F}_t$, i.e.,

$$\mathbb{E}[\left\langle h(\cdot, \mathbf{x}_t^{(KT+l)}), f\right\rangle_{\mathcal{H}} |\mathcal{F}_t] = \left\langle h_{\mu_t|\mathcal{F}_t}, f\right\rangle \ \forall\ f \in \mathcal{H} \tag{38}$$

To control $\mathbb{E}\left[\|\frac{1}{n}\sum_{l=0}^{n-1} h(\cdot, \mathbf{x}_t^{(KT+l)})\|_{\mathcal{H}}^2 \mid \mathcal{F}_t\right]$, we proceed as follows:

$$
\begin{aligned}
\|\tfrac{1}{n}\sum_{l=0}^{n-1} h(\cdot, \mathbf{x}_t^{(KT+l)})\|_{\mathcal{H}}^2 &= \frac{1}{n^2}\sum_{l_1, l_2=0}^{n-1} \left\langle h(\cdot, \mathbf{x}_t^{(Kt+l_1)}), h(\cdot, \mathbf{x}_t^{(Kt+l_2)})\right\rangle_{\mathcal{H}} \\
&= \frac{1}{n^2}\sum_{l=0}^{m-1} \|h(\cdot, \mathbf{x}_t^{(Kt+l)})\|_{\mathcal{H}}^2 + \frac{1}{n^2}\sum_{0 \leq l_1 \neq l_2 \leq n-1} \left\langle h(\cdot, \mathbf{x}_t^{(Kt+l_1)}), h(\cdot, \mathbf{x}_t^{(Kt+l_2)})\right\rangle_{\mathcal{H}} \\
&\leq \frac{\xi^2}{n} + \frac{1}{n^2}\sum_{0 \leq l_1 \neq l_2 \leq n-1} \left\langle h(\cdot, \mathbf{x}_t^{(Kt+l_1)}), h(\cdot, \mathbf{x}_t^{(Kt+l_2)})\right\rangle_{\mathcal{H}}
\end{aligned}
$$

where the last inequality uses the fact that $\|h(\cdot, \mathbf{x}_t^{(Kt+l)})\|_{\mathcal{H}} \leq \xi$ almost surely as per Lemma 9.

To control the conditional expectation of the off-diagonal terms, let $i = Kt + l_1$ and $j = Kt + l_2$ for any arbitrary l_1, l_2 with $0 \leq l_1 \neq l_2 \leq n-1$. Conditioned on $\mathcal{F}_t$, $\mathbf{x}_t^{(i)}$ and $\mathbf{x}_t^{(j)}$ are i.i.d samples from $\mu_t|\mathcal{F}_t$. Thus, by Equation (38) and Fubini's Theorem,

$$
\begin{aligned}
\mathbb{E}_{\mathbf{x}_t^{(i)}, \mathbf{x}_t^{(j)}}\left[\left\langle h(\cdot, \mathbf{x}_t^{(i)}), h(\cdot, \mathbf{x}_t^{(j)})\right\rangle_{\mathcal{H}} \middle| \mathcal{F}_t\right] &= \mathbb{E}_{\mathbf{x}_t^{(i)}}\left[\mathbb{E}_{\mathbf{x}_t^{(j)}}\left[\left\langle h(\cdot, \mathbf{x}_t^{(i)}), h(\cdot, \mathbf{x}_t^{(j)})\right\rangle_{\mathcal{H}} \middle| \mathcal{F}_t\right]\right] \\
&= \mathbb{E}_{\mathbf{x}_t^{(i)}}\left[\left\langle h_{\mu_t|\mathcal{F}_t}, h(\cdot, \mathbf{x}_t^{(i)})\right\rangle_{\mathcal{H}} \middle| \mathcal{F}_t\right] \\
&= \|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}}^2
\end{aligned}
$$

It follows that,

$$
\mathbb{E}\left[\|\tfrac{1}{n}\sum_{l=0}^{n-1} h(\cdot, \mathbf{x}_t^{(KT+l)})\|_{\mathcal{H}}^2 \mid \mathcal{F}_t\right] \leq \|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}}^2 + \frac{\xi^2}{n}
$$

Substituting the above into equation 36 and equation 37, we obtain the following:

$$
\mathbb{E}\left[\mathsf{KSD}_{\pi^\star}^2(\hat{\mu}^{(n)}||\pi^\star) \mid \mathcal{C}, S = t\right] \leq \frac{\xi^2}{n} + \|h_{\mu_t|\mathcal{F}_t}\|_{\mathcal{H}}^2 = \frac{\xi^2}{n} + \mathsf{KSD}_{\pi^\star}^2(\mu_t|\mathcal{F}_t||\pi^*)
$$

where the second step applies Definition 1. Finally, taking expectations with respect to $\mathcal{C}$ and $S \sim \mathsf{Uniform}(\{0, \dots, T-1\})$ on both sides of the above inequality, we get:

$$
\mathbb{E}\left[\mathsf{KSD}_{\pi^\star}^2(\hat{\mu}^{(n)}||\pi^\star)\right] \leq \frac{\xi^2}{n} + \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\mathsf{KSD}_{\pi^\star}^2(\mu_t|\mathcal{F}_t||\pi^\star)]
$$

Substituting the bound from Theorem 1 into the above inequality, we conclude that:

$$
\mathbb{E}[\mathsf{KSD}_{\pi^\star}^2(\hat{\mu}^{(n)}||\pi^\star)] \leq \frac{\xi^2}{n} + \frac{2\mathsf{KL}\left(\mu_0|\mathcal{F}_0\|\pi^\star\right)}{\gamma T} + \frac{\gamma B(4+L)\xi^2}{K}
$$

We note that for $\gamma = O(\frac{(Kd)^\eta}{T^{1-\eta}})$ and $R = \sqrt{d/L}$, $\mathsf{KL}\left(\mu_0|\mathcal{F}_0\|\pi^\star\right) = O(d)$ by Lemma 4 and

$$
\xi^2 \leq 4\zeta_0 + 4\zeta_1(\gamma T)^{2/\alpha} + 4\zeta_2(\gamma^2 T)^{2/\alpha} + 4\zeta_3 R^{4/\alpha} \leq O\left((KdT)^{\frac{1}{1+\alpha}} + d^{2/\alpha}\right)
$$

Furthermore,

$$
\begin{aligned}
\frac{2\mathsf{KL}\left(\mu_0|\mathcal{F}_0\|\pi^\star\right)}{\gamma T} + \frac{\gamma B(4+L)\xi^2}{K} &\leq O\left(\frac{d}{\gamma T} + \frac{\gamma B(4+L)\xi^2}{2K}\right) \leq O\left(\frac{d^{1-\eta}}{(KT)^\eta}\right) \\
&\leq O\left(\frac{d^{\frac{2+\alpha}{2(1+\alpha)}}}{(KT)^{\frac{\alpha}{2(1+\alpha)}}}\right)
\end{aligned}
$$

It follows that,

$$
\mathbb{E}[\mathsf{KSD}_{\pi^\star}^2(\hat{\mu}^{(n)}||\pi^\star)] \leq O\left(\frac{d^{2/\alpha}}{n} + \frac{(KTd)^{\frac{1}{1+\alpha}}}{n} + \frac{d^{\frac{2+\alpha}{2(1+\alpha)}}}{(KT)^{\frac{\alpha}{2(1+\alpha)}}}\right)
$$

$KT = d^{\frac{\alpha}{2+\alpha}} n^{\frac{2(1+\alpha)}{2+\alpha}}$, we conclude:

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\mu}^{(n)}||\pi^\star)] \leq O\left(\frac{d^{\frac{2}{2+\alpha}}}{n^{\frac{\alpha}{2+\alpha}}} + \frac{d^{2/\alpha}}{n}\right)$$

$\square$

E.2 GB-SVGD

Corollary 3 (GB-SVGD : Fast Finite-Particle Convergence). *Let the assumptions and parameter settings of Theorem 1 be satisfied. Let $\hat{\nu}^{(n)}$ denote the empirical measure of the n particles output by GB-SVGD. Then, under without-replacement sampling of the minibatches, the following holds:*

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] \leq \frac{4K^2T^2\xi^2}{n^2} + \frac{2\xi^2}{n} + \frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0||\pi^\star\right)}{\gamma T} + \frac{2\gamma B(4+L)\xi^2}{K}$$

and the following holds under with-replacement sampling of the minibatches

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] \leq \frac{4K^2T^2\xi^2}{n^2}(1 - \frac{K^2T^2}{n}) + \frac{4K^2T^2\xi^2}{n} + \frac{2\xi^2}{n} + \frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0||\pi^\star\right)}{\gamma T} + \frac{2\gamma B(4+L)\xi^2}{K}$$

where ξ is as defined in Theorem 1. In particular, for GB-SVGD under without-replacement sampling of the minibatches, setting $R = \sqrt{d/L}, \gamma = O(\frac{(Kd)^\eta}{T^{1-\eta}})$ with $\eta = \frac{\alpha}{2(1+\alpha)}$ and $KT = \sqrt{n}$ suffices to ensure the following

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] \leq O\left(\frac{d^{2/\alpha}}{n} + \frac{d^{\frac{1}{1+\alpha}}}{n^{\frac{1+2\alpha}{2(1+\alpha)}}} + \frac{d^{\frac{2+\alpha}{2(1+\alpha)}}}{n^{\frac{\alpha}{4(1+\alpha)}}}\right)$$

Proof. Let $\bar{\mathbf{Y}} = (\bar{\mathbf{y}}^{(0)}, \dots, \bar{\mathbf{y}}^{(n-1)})$ denote the n particles output by GB-SVGD and let $\hat{\nu}^{(n)} = \frac{1}{n}\sum_{l=0}^{n-1}\delta_{\bar{\mathbf{y}}^{(l)}}$ denote their empirical measure. Let $\mathcal{E}$ denote the event that each random batch $\mathcal{K}_t$ of GB-SVGD is disjoint and contains unique elements for every $t \in (T)$. Moreover, let $\mathbf{Y} = (\mathbf{y}^{(0)}, \dots, \mathbf{y}^{(n-1)})$ denote the n particles output by VP-SVGD, run with the parameter settings stated above, and coupled with $\bar{\mathbf{Y}}$ as per the coupling constructed in the proof of Theorem 2 in Appendix D.1. Let $\hat{\mu}^{(n)} = \frac{1}{n}\sum_{l=0}^{n-1}\delta_{\mathbf{y}^{(l)}}$ denote their empirical measure. By definition of Kernel Stein Discrepancy (Definition 1) and the convexity $\|\cdot\|^2_{\mathcal{H}}$, it follows that:

$$\begin{aligned}
\mathbb{E}[\mathsf{KSD}_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] &= \mathbb{E}[\|h_{\hat{\nu}^{(n)}}\|^2_{\mathcal{H}}] \\
&= \mathbb{E}[\|h_{\hat{\nu}^{(n)}} - h_{\hat{\mu}^{(n)}} + h_{\hat{\mu}^{(n)}}\|^2_{\mathcal{H}}] \\
&\leq 2\mathbb{E}[\|h_{\hat{\nu}^{(n)}} - h_{\hat{\mu}^{(n)}}\|^2_{\mathcal{H}}] + 2\mathbb{E}[\|h_{\hat{\mu}^{(n)}}\|^2_{\mathcal{H}}] \\
&= 2\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\hat{\mu}^{(n)})] + 2\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\mu}^{(n)}||\pi^\star)]
\end{aligned}$$

Substituting the bounds of Theorem 2 and Corollary 2 into the above inequality, we conclude the following:

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] \leq \frac{4K^2T^2\xi^2}{n^2}\mathbb{P}(\mathcal{E}) + 4\xi^2\mathbb{P}(\mathcal{E}^c) + \frac{2\xi^2}{n} + \frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0||\pi^\star\right)}{\gamma T} + \frac{2\gamma B(4+L)\xi^2}{K}$$

We recall that, $\mathbb{P}(\mathcal{E}) = 1$ under without-replacement sampling of the random batches $\mathcal{K}_t$ and $\mathbb{P}(\mathcal{E}) = 1 - {K^2T^2}/{n}$ under with-replacement sampling. Thus, under without-replacement sampling, the following holds:

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] \leq \frac{4K^2T^2\xi^2}{n^2} + \frac{2\xi^2}{n} + \frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0||\pi^\star\right)}{\gamma T} + \frac{2\gamma B(4+L)\xi^2}{K}$$

Moreover, the following holds under with-replacement sampling

$$\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] \leq \frac{4K^2T^2\xi^2}{n^2}\left(1 - \frac{K^2T^2}{n}\right) + \frac{4K^2T^2\xi^2}{n} + \frac{2\xi^2}{n} + \frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0||\pi^\star\right)}{\gamma T} + \frac{2\gamma B(4+L)\xi^2}{K}$$

Now, let us consider GB-SVGD without replacement with $R = \sqrt{d/L}$, $\gamma = O(\frac{(Kd)^{\eta}}{T^{1-\eta}})$ and $KT = n^{1/2}$ It follows that $\mathsf{KL}\left(\mu_0|\mathcal{F}_0||\pi^{\star}\right) = O(d)$ by Lemma 4 and

$$\begin{aligned}
\xi^2 &\leq 4\zeta_0 + 4\zeta_1(\gamma T)^{2/\alpha} + 4\zeta_2(\gamma^2 T)^{2/\alpha} + 4\zeta_3 R^{4/\alpha} \\
&\leq O\left((KdT)^{\frac{1}{1+\alpha}} + d^{2/\alpha}\right) \\
&\leq O\left(d^{2/\alpha} + d^{\frac{1}{1+\alpha}} n^{\frac{1}{2(1+\alpha)}}\right)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0||\pi^{\star}\right)}{\gamma T} + \frac{2\gamma B(4+L)\xi^2}{K} &\leq O\left(\frac{d}{\gamma T} + \frac{\gamma B(4+L)\xi^2}{2K}\right) \leq O\left(\frac{d^{1-\eta}}{(KT)^{\eta}}\right) \\
&\leq O\left(\frac{d^{\frac{2+\alpha}{2(1+\alpha)}}}{n^{\frac{\alpha}{4(1+\alpha)}}}\right)
\end{aligned}$$

Hence, we conclude that,

$$\begin{aligned}
\mathbb{E}[\mathsf{KSD}^2_{\pi^{\star}}(\hat{\nu}^{(n)}||\pi^{\star})] &\leq \frac{4K^2T^2\xi^2}{n^2} + \frac{2\xi^2}{n} + \frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0||\pi^{\star}\right)}{\gamma T} + \frac{2\gamma B(4+L)\xi^2}{K} \\
&\leq \frac{6\xi^2}{n} + \frac{4\mathsf{KL}\left(\mu_0|\mathcal{F}_0||\pi^{\star}\right)}{\gamma T} + \frac{2\gamma B(4+L)\xi^2}{K} \\
&\leq O\left(\frac{d^{2/\alpha}}{n} + \frac{d^{\frac{1}{1+\alpha}}}{n^{\frac{1+2\alpha}{2(1+\alpha)}}} + \frac{d^{\frac{2+\alpha}{2(1+\alpha)}}}{n^{\frac{\alpha}{4(1+\alpha)}}}\right)
\end{aligned}$$

□

E.3 Oracle Complexity of SVGD, VP-SVGD and GB-SVGD

We now compare the gradient oracle complexity, (i.e., the number of evaluations of ∇F) of VP-SVGD (as implied by Corollary 2) and GB-SVGD (as implied by Corollary 3) with that of SVGD as implied by the state-of-the-art finite particle guarantee of Shi and Mackey [37].

E.3.1 SVGD

From Equation (1), We note that T steps of SVGD run with n particles requires $n^2 T$ evaluations of ∇F.

Subgaussian $\pi^{\star}$ For subgaussian $\pi^{\star}$, the finite-particle convergence rate obtained by Shi and Mackey [37] is $\mathsf{KSD}_{\pi^{\star}}(\hat{\mu}^{(n)}_{\mathsf{SVGD}}||\pi^{\star}) = \tilde{O}(\frac{\mathsf{poly}(d)}{\sqrt{\log\log n^{\Theta(1/d)}}})$, where $\hat{\mu}^{(n)}_{\mathsf{SVGD}}$ denotes the empirical measure of the n particles output by SVGD. By carefully following the analysis of Shi and Mackey [37], we infer that, to achieve $\mathsf{KSD}_{\pi^{\star}}(\hat{\mu}^{(n)}_{\mathsf{SVGD}}||\pi^{\star}) \leq \epsilon$, SVGD requires $T = \tilde{O}(\frac{\mathsf{poly}(d)}{\epsilon^2})$ and $n = \tilde{O}(\exp(\Theta(de^{\frac{\mathsf{poly}(d)}{\epsilon^2}})))$. Thus the oracle complexity of SVGD (as implied by Shi and Mackey [37]) for achieving $\mathsf{KSD}_{\pi^{\star}}(\hat{\mu}^{(n)}_{\mathsf{SVGD}}||\pi^{\star})$ is $\tilde{O}(\frac{\mathsf{poly}(d)}{\epsilon^2} \cdot \exp(\Theta(de^{\frac{\mathsf{poly}(d)}{\epsilon^2}})))$

E.3.2 VP-SVGD

From Algorithm 1, we note that T steps of VP-SVGD run with n particles and a batch-size of K requires $K^2T^2 + KTn$ evaluations of ∇F.

Subgaussian $\pi^{\star}$ For subgaussian $\pi^{\star}$, Corollary 2 implies a finite-particle convergence rate of $\mathbb{E}[\mathsf{KSD}^2_{\pi^{\star}}(\hat{\mu}^{(n)}||\pi^{\star})] = O(\frac{d^{1/2}}{n^{1/2}} + \frac{d}{n})$ (where $\hat{\mu}^{(n)}$ denotes the empirical measure of the n particles output by VP-SVGD) assuming $KT = d^{1/2}n^{3/2}$. Hence, to achieve $\mathbb{E}[\mathsf{KSD}_{\pi^{\star}}(\hat{\mu}^{(n)}||\pi^{\star})] \leq \epsilon$, VP-SVGD requires $n = O(\frac{d}{\epsilon^4})$ and $KT = d^{1/2}n^{3/2} = \frac{d^2}{\epsilon^6}$. The resulting oracle complexity for achieving

$\mathbb{E}[\mathsf{KSD}_{\pi^\star}(\hat{\mu}^{(n)}||\pi^\star)] \leq \epsilon$ is $O(\frac{d^4}{\epsilon^{12}})$. Compared to the oracle complexity of SVGD obtained above, this is a *double exponential improvement in both* d *and* $1/\epsilon$. Notably, the obtained oracle complexity guarantee *completely eliminates the curse of dimensionality*.

Subexponential $\pi^\star$ For subexponential $\pi^\star$, Corollary 2 implies a finite-particle convergence rate of $\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\mu}^{(n)}||\pi^\star)] = O(\frac{d^{2/3}}{n^{1/3}} + \frac{d^2}{n})$ (where $\hat{\mu}^{(n)}$ denotes the empirical measure of the n particles output by VP-SVGD) assuming $KT = d^{1/3}n^{4/3}$. Hence, to achieve $\mathbb{E}[\mathsf{KSD}_{\pi^\star}(\hat{\mu}^{(n)}||\pi^\star)] \leq \epsilon$, VP-SVGD requires $n = O(\frac{d^2}{\epsilon^6})$ and $KT = d^{1/3}n^{4/3} = \frac{d^3}{\epsilon^8}$. The resulting oracle complexity for achieving $\mathbb{E}[\mathsf{KSD}_{\pi^\star}(\hat{\mu}^{(n)}||\pi^\star)] \leq \epsilon$ is $O(\frac{d^6}{\epsilon^{16}})$.

E.3.3 GB-SVGD

From Algorithm 2, we note that T steps of GB-SVGD run with n particles and a batch-size of K requires KTn evaluations of ∇F.

Subgaussian $\pi^\star$ For subgaussian $\pi^\star$, Corollary 3 implies a finite-particle convergence rate of $\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] = O(\frac{d^{2/3}}{n^{1/6}} + \frac{d}{n})$ (where $\hat{\nu}^{(n)}$ denotes the empirical measure of the n particles output by GB-SVGD) assuming $KT = n^{1/2}$. Hence, to achieve $\mathbb{E}[\mathsf{KSD}_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] \leq \epsilon$, GB-SVGD requires $n = \frac{d^4}{\epsilon^{12}}$ and $KT = \sqrt{n} = \frac{d^2}{\epsilon^6}$. Under this setting, the oracle complexity of GB-SVGD as implied by Corollary 3 is $O(\frac{d^6}{\epsilon^{18}})$. Compared to the oracle complexity of SVGD obtained above, this is *a double exponential improvement in both* d *and* $1/\epsilon$. Notably, the obtained oracle complexity guarantee *completely eliminates the curse of dimensionality*

Subexponential $\pi^\star$ For subexponential $\pi^\star$, Corollary 3 implies a finite-particle convergence rate of $\mathbb{E}[\mathsf{KSD}^2_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] = O(\frac{d^{3/4}}{n^{1/8}} + \frac{d^2}{n})$ (where $\hat{\nu}^{(n)}$ denotes the empirical measure of the n particles output by GB-SVGD) assuming $KT = n^{1/2}$. Hence, to achieve $\mathbb{E}[\mathsf{KSD}_{\pi^\star}(\hat{\nu}^{(n)}||\pi^\star)] \leq \epsilon$, GB-SVGD requires $n = \frac{d^6}{\epsilon^{16}}$ and $KT = \sqrt{n} = \frac{d^3}{\epsilon^8}$. Under this setting, the oracle complexity of GB-SVGD as implied by Corollary 3 is $O(\frac{d^9}{\epsilon^{24}})$.

F Literature Review

Initial works on the analysis of SVGD such as Liu [26], Lu et al. [30], Duncan et al. [14], Chewi et al. [7], Nüsken and Renger [33] consider the continuous-time population limit, i.e., the limit of infinite particles and vanishing step-sizes. In this regime, Liu [26], Lu et al. [30], Nüsken and Renger [33] show that the behavior of SVGD is characterized by a Partial Differential Equation (PDE), and established asymptotic convergence of this PDE to the target distribution. The work of Duncan et al. [14] proposes the Stein Logarithmic Sobolev Inequality which ensures exponential convergence of this PDE to the target distribution. However, characterizing the conditions under which this inequality holds is an open problem. The work of Chewi et al. [7] show that the PDE governing SVGD in the continuous-time population limit can be interpreted as an approximate Wasserstein gradient flow of the Chi-squared divergence. To this end, Chewi et al. [7] shows that the (exact) Wasserstein gradient flow of the Chi-squared divergence exhibits exponential convergence to the target distribution when $\pi^\star$ satisfies a Poincare Inequality. To the best of our knowledge, the first discrete-time non-asymptotic convergence result for population-limit SVGD was established in Korba et al. [23], where the authors interpreted population-limit SVGD as projected Wasserstein gradient descent. Their result relied on the assumption that the Kernel Stein Discrepancy to the target is uniformly bounded along the trajectory of SVGD, a condition which is hard to verify apriori. This result was significantly improved in Salim et al. [36], which established convergence of population-limit SVGD assuming the potential F is smooth the target $\pi^\star \propto e^{-F}$ satisfies Talagrand's inequality T_1, an assumption which is equivalent to subgaussianity of $\pi^\star$. This result was extended in Sun et al. [39] to accommodate for potentials F that satisfy a more general smoothness condition.

In comparison to prior works on population-limit SVGD, the literature on finite-particle SVGD is relatively sparse. The works of Liu [26] and Gorham et al. [18] establish that the dynamics of finite-particle SVGD asymptotically converge to that of population-limit SVGD in bounded Lipschitz distance and Wasserstein-1 distance respectively, as the number of particles approaches infinity. Under the stringent condition of bounded F (which is violated in various scenarios, e.g. log-strongly concave $\pi^\star$), Korba et al. [23] derived a non-asymptotic bound between the expected Wasserstein-2 distance between finite-particle SVGD and population-limit SVGD. To the best of our knowledge, Shi and Mackey [37] is the only prior work that explicitly establishes a non-asymptotic convergence guarantee of finite-particle SVGD to the target, which shows that the empirical measure of SVGD run with n particles converges to the target density in KSD at a rate of $O\left(\sqrt{\frac{\mathsf{poly}(d)}{\log\log n^{\Theta(1/d)}}}\right)$