

# IMPROVED FINITE-PARTICLE CONVERGENCE RATES FOR STEIN VARIATIONAL GRADIENT DESCENT

KRISHNAKUMAR BALASUBRAMANIAN<sup>1</sup>, SAYAN BANERJEE<sup>2</sup>, AND PROMIT GHOSAL<sup>3</sup>

**ABSTRACT.** We provide finite-particle convergence rates for the Stein Variational Gradient Descent (SVGD) algorithm in the Kernel Stein Discrepancy (KSD) and Wasserstein-2 metrics. Our key insight is the observation that the time derivative of the relative entropy between the joint density of  $N$  particle locations and the  $N$ -fold product target measure, starting from a regular initial distribution, splits into a dominant ‘negative part’ proportional to  $N$  times the expected KSD<sup>2</sup> and a smaller ‘positive part’. This observation leads to KSD rates of order  $1/\sqrt{N}$ , providing a near optimal double exponential improvement over the recent result by [Shi and Mackey \(2024\)](#). Under mild assumptions on the kernel and potential, these bounds also grow linearly in the dimension  $d$ . By adding a bilinear component to the kernel, the above approach is used to further obtain Wasserstein-2 convergence. For the case of ‘bilinear + Matérn’ kernels, we derive Wasserstein-2 rates that exhibit a curse-of-dimensionality similar to the i.i.d. setting. We also obtain marginal convergence and long-time propagation of chaos results for the time-averaged particle laws.

## 1. INTRODUCTION

Stein Variational Gradient Descent (SVGD) ([Liu and Wang, 2016](#)) is a widely-used algorithm for sampling from a target density  $\pi \propto \exp(-V)$ , where  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  is the potential function. For a given symmetric, positive-definite kernel  $\mathbf{k} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ , discrete time-step  $k \geq 0$ , step-size  $\eta > 0$ , and for  $1 \leq i \leq N$ , the SVGD algorithm is given by

$$x_i^N(k+1) = x_i^N(k) - \frac{\eta}{N} \sum_j [\mathbf{k}(x_i^N(k), x_j^N(k)) \nabla V(x_j^N(k)) - \nabla_2 \mathbf{k}(x_i^N(k), x_j^N(k))]. \quad (1)$$

SVGD provides a compelling alternative to more classical Markov Chain Monte Carlo techniques that are known to be harder to scale ([Blei et al., 2017](#)), and has attracted considerable attention in the machine learning and applied mathematics communities because of its fascinating theoretical properties and broad range of applications ([Feng et al., 2017](#); [Haarnoja et al., 2017](#); [Liu et al., 2021](#); [Xu et al., 2022](#)). Our focus in this work is on deriving rates of convergence of SVGD by considering the continuous-time,  $N$ -particle SVGD dynamics on  $\mathbb{R}^d$ , obtained by letting  $\eta \rightarrow 0_+$ , given by

$$\dot{x}_i^N(t) = -\frac{1}{N} \sum_j \mathbf{k}(x_i^N(t), x_j^N(t)) \nabla V(x_j^N(t)) + \frac{1}{N} \sum_j \nabla_2 \mathbf{k}(x_i^N(t), x_j^N(t)), \quad (2)$$

with  $\dot{x}$  denoting the time derivative and  $\nabla_2$  represents gradient with respect to the second argument.

**Background and past work:** The motivation for SVGD originates from the *gradient flow for the relative entropy* on the Wasserstein-2 space of probability measures on  $\mathbb{R}^d$ . More precisely, for a probability measure  $\mu$  on  $\mathbb{R}^d$  possessing a regular enough positive density, the Wasserstein gradient flow is given by the measure-valued trajectory  $\mu_t$  satisfying the continuity equation

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0, \quad \mu_0 = \mu, \quad (3)$$

<sup>1</sup>DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, DAVIS, EMAIL: KBALA@UCDAVIS.EDU. SUPPORTED IN PART BY NATIONAL SCIENCE FOUNDATION (NSF) GRANT DMS-2413426.

<sup>2</sup>DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH, UNIVERSITY OF NORTH CAROLINA, CHAPEL HILL. EMAIL: SAYAN@EMAIL.UNC.EDU. SUPPORTED IN PART BY NSF CAREER AWARD DMS-2141621.

<sup>3</sup>DEPARTMENT OF STATISTICS, UNIVERSITY OF CHICAGO. EMAIL: PROMIT@UCHICAGO.EDU

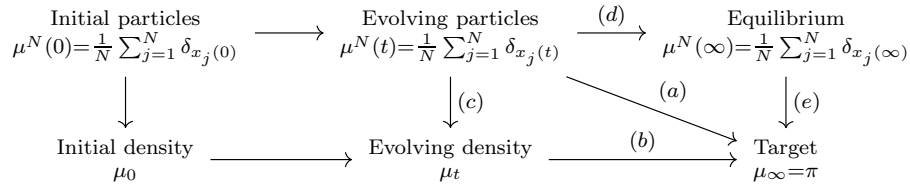
where  $v_t = -\nabla \log(p_t/\pi)$  and  $p_t$  is the density of  $\mu_t$ . Under suitable conditions,  $\mu_t$  can be shown to converge (often with quantifiable fast rates) to  $\pi$ . Unfortunately, this approach is not practically implementable via particle discretization as the associated empirical measure approximating  $\mu_t$  does not possess a density.

In a very influential paper, [Liu and Wang \(2016\)](#) devised a *projected* gradient descent algorithm by projecting the velocity vector  $v_t$  along a reproducing kernel Hilbert space (RKHS) associated with a symmetric positive definite kernel  $k$ . This leads to a flow analogous to (3) but with  $v_t = -P_{\mu_t} \nabla \log(p_t/\pi)$ , where the projection  $P_{\mu_t}$  is given by  $P_{\mu_t} f(x) := \int k(x, y) f(y) \nu(dy)$  for probability measure  $\nu$  and function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  for which the integral is well-defined. The key observation of [Liu and Wang \(2016\)](#) was that, by applying integration by parts, one obtains

$$-P_{\mu_t} \nabla \log(p_t/\pi)(x) = \int (-k(x, y) \nabla V(y) + \nabla_2 k(x, y)) \mu_t(dy).$$

The right hand side is well-defined even when  $\mu_t$  lacks a density and is hence amenable to particle discretization, which leads to the SVGD equations (1) and (2). See [Korba et al. \(2020\)](#) for a more detailed description of this approach.

With this motivation, the following diagram from [Liu et al. \(2024\)](#) highlights the major goals involved in rigorously analyzing the SVGD dynamics.



- (a) Unified convergence of the empirical measure for  $N < \infty$  particles to the continuous target as time  $t$  and  $N$  jointly grow to infinity;
- (b) Convergence of mean-field SVGD to the target distribution over time;
- (c) Convergence of the empirical measure for finite particles to the mean-field distribution at any finite given time  $t \in [0, \infty)$ ;
- (d) Convergence of finite-particle SVGD to equilibrium over time;
- (e) Convergence of the empirical measure for finite particles to the continuous target at time  $t = \infty$ .

Practically speaking, (a) is the ideal outcome that completely defines the algorithmic behavior of SVGD. One approach towards this is to combine either (b) and (c) or (d) and (e) in a quantitative way to yield (a). Regarding (b), [Liu \(2017\)](#) showed the convergence of mean-field SVGD (solution to (3) with  $v_t = -P_{\mu_t} \nabla \log(p_t/\pi)$ ) in Kernelized Stein Discrepancy (KSD, [Chwialkowski et al. \(2016\)](#); [Liu et al. \(2016\)](#); [Gorham and Mackey \(2017\)](#)), which is known to imply weak convergence under appropriate assumptions. [Korba et al. \(2020\)](#); [Chewi et al. \(2020\)](#); [Salim et al. \(2022\)](#); [Sun et al. \(2023\)](#); [Duncan et al. \(2023\)](#) sharpened the results with weaker conditions or explicit rates. [He et al. \(2024\)](#) extended the above result to the stronger Fisher information metric and Kullback–Leibler divergence based on a regularization technique. [Lu et al. \(2019\)](#); [Gorham et al. \(2020\)](#); [Korba et al. \(2020\)](#) obtained time-dependent mean-field convergence (c) under various assumptions using techniques from partial differential equations and from the literature of ‘propagation of chaos’. In particular, [Lu et al. \(2019\)](#) derived the mean-field PDE (3) for the evolving density that emerges as the mean-field limit of the finite-particle SVGD systems, and showed the well-posedness of the PDE solutions. [Carrillo and Skrzeczkowski \(2023\)](#) established refined stability estimates in comparison to [Lu et al. \(2019\)](#) for the mean-field system when the initial distribution is close to the target distribution in a suitable sense. In particular, they increase the length of the time interval

in which mean-field approximation is meaningful from  $\approx \log \log N$  to  $\approx \sqrt{N}$  for such initial data close to the target.

Shi and Mackey (2024) obtained refined results for (c) and combined them with (b) to get the *first unified convergence* (a) in terms of KSD. However, they have a rather slow rate of order  $1/\sqrt{\log \log N}$ , resulting from the fact that their bounds for (c) still depend on the time  $t$  (sum of step sizes) double-exponentially. Note that studying the convergence (d) and (e), provides another way to characterize the unified convergence (a) for SVGD. Liu et al. (2024) analyzed this strategy for the *Gaussian SVGD* case where the target distribution  $\pi$  and initial distribution  $\mu$  are both Gaussian and the kernel  $k$  is bilinear. In this case, the flow of measures for the mean-field SVGD remains Gaussian for all time and this fact was exploited to obtain detailed rates and ‘uniform-in-time’ propagation of chaos results. Das and Nagaraj (2023) obtained a polynomial convergence rate ( $O(N^{-\alpha})$  for some  $\alpha > 0$ ) for a related but different algorithm, which they called *SVGD with virtual particles*, by adding more randomness to the dynamics and using stochastic approximation techniques. The recent work of Priser et al. (2024) also studies finite-particle asymptotics, albeit for not the original SVGD iterates (as in (1)) but for a modified one where a Langevin-type regularization including a Gaussian noise is added at each step. Hence, they leverage existing techniques for Langevin Monte Carlo to establish their results. However, their techniques are not applicable to the deterministic SVGD system in (1).

**Challenges for finite-particle SVGD:** In comparison to the numerous works quantifying convergence rates for the mean-field SVGD equation, only Shi and Mackey (2024) and Liu et al. (2024) discussed above have made attempts towards obtaining rates for the finite-particle version of (deterministic) SVGD. This has been perceived as a challenging open problem till date. The tractability of the mean-field SVGD equation comes from the observation that it has a (projected) gradient structure which leads to the following monotonicity property of the KL-divergence:

$$\partial_t \text{KL}(\mu_t || \pi) = -\text{KSD}^2(\mu_t || \pi), \quad t \geq 0.$$

The non-negativity of KL then leads to bounds on the KSD. In the finite-particle versions (1) and (2), there is no gradient structure to the dynamics, which renders the above approach inapplicable. Moreover, the vector-field driving the finite-particle dynamics is not globally Lipschitz and lacks suitable convexity properties. This results in double-exponentially growing bounds in time between the particle empirical distribution and the mean-field limit (see Lu et al. (2019, Prop. 2.6) and Shi and Mackey (2024, Thm. 1)). As a consequence, trying to obtain (a) using (b) and (c) in the general (non-Gaussian) setting yields the slow convergence rate  $1/\sqrt{\log \log N}$ . In Das and Nagaraj (2023), the authors tried to deliberately bypass this approach in the finite-particle setting, which leads to better rates, but their approach needs an altogether different, albeit related, algorithm with additional randomness driving the dynamics.

**Our contribution:** A key insight in this paper is to work with the *joint density* of the particle locations, when started from a suitably regular initial distribution, and track the evolution of its relative entropy with respect to the  $N$ -fold product measure  $\pi^{\otimes N}$ . It turns out that the time derivative of this relative entropy has a ‘negative part’ that is exactly  $N$  times the expected  $\text{KSD}^2$  of the empirical measure at time  $t$  with respect to  $\pi$ , and a ‘positive part’ that can be separately handled and shown to be small in comparison to the negative part (see (9)). This gives a novel connection between the joint particle dynamics and the empirical measure evolution. While analyzing the deterministic SVGD, in comparison to Shi and Mackey (2024) which uses (b) and (c) or Liu et al. (2024) which uses (d) and (e), we directly provide convergence rates for (a) using the above strategy.

Our first main result, Theorem 1, exploits this observation to obtain  $O(1/\sqrt{N})$  bounds for the expected KSD between  $\mu_{av}^N := \frac{1}{N} \int_0^N \mu^N(t) dt$  and  $\pi$ . This is a *double exponential improvement over Shi and Mackey (2024) for the true SVGD algorithm*. As discussed in Remark 2, these bounds are *essentially optimal* when compared with the KSD in the i.i.d. setting, and *grow linearly in  $d$*

(that is KSD is  $O(d/\sqrt{N})$ ) under mild assumptions on the kernel and the potential. Moreover, unlike previous works even for the mean-field SVGD, we do not require any assumptions on the tail behavior (such as sub-Gaussianity) of the target  $\pi$ . Further, it follows from [Gorham and Mackey \(2017\)](#) that the KSD bound alone does not even guarantee weak convergence of the particle marginal laws as  $N \rightarrow \infty$ , unless one establishes tightness of these laws. Our approach gives control on the relative entropy of the joint law in time which, in turn, gives the desired tightness and weak convergence for the time-averaged particle marginal laws  $\bar{\mu}^N(dx) := \frac{1}{N} \int_0^N \mathbb{P}(x_1(t) \in dx) dt$  for exchangeable initial conditions, see [Theorem 2](#).

In [Section 3](#), we obtain *Wasserstein-2 convergence and associated rates*. For this purpose, we heavily rely on the treatise of [Kanagawa et al. \(2022\)](#) which connects KSD convergence to Wasserstein convergence when the kernel has a bilinear component and a translation invariant component of the form  $(x, y) \mapsto \Psi(x - y)$  (see [\(10\)](#)). Such kernels are typically unbounded, in contrast with standard boundedness assumptions in most papers on SVGD (a notable exception is [Liu et al. \(2024\)](#)). In [Theorem 3](#), under dissipativity and growth assumptions on the potential  $V$ , we obtain polynomial KSD convergence rates for SVGD finite-particle dynamics with such kernels, which by [Kanagawa et al. \(2022\)](#) imply Wasserstein convergence. When the translation invariant part of the kernel is of Matérn type, we obtain Wasserstein convergence rates in [Theorem 4](#) of the form  $O(1/N^{\alpha/d})$  (where  $\alpha > 0$  does not depend on  $d$ ) for the particle SVGD using [Theorem 3](#) in conjunction with results in [Kanagawa et al. \(2022\)](#). This is the first work on Wasserstein convergence for non-Gaussian SVGD finite-particle dynamics. Unlike the KSD bound, the  $d$  dependence leads to *curse-of-dimensionality* in the Wasserstein bound, but this is to be expected when compared to Wasserstein bounds for empirical distribution of i.i.d. random variables ([Dudley, 1969](#); [Weed and Bach, 2019](#)). Finally, we obtain a *long time propagation of chaos* result in [Proposition 1](#), namely, we show that the time-averaged marginals of the particle locations over the time interval  $[0, N]$ , started from an exchangeable initial configuration, become asymptotically independent as  $N \rightarrow \infty$  and essentially produce i.i.d samples from  $\pi$ .

We note here that, although our analysis explicitly uses the continuous time description [\(2\)](#) of SVGD, one can extend the entire analysis to the discrete time setting given by [\(1\)](#). In the later case, under additional smoothness assumptions on the potential function  $V$ , the evolution of the joint relative entropy can be similarly tracked at the discrete time steps to yield an equation analogous to [\(9\)](#) below, with an added  $O(\eta^2)$  term that is negligible in comparison to the negative part. This can be leveraged to prove versions of all the stated results, as long as the step-size  $\eta$  from [\(1\)](#) is chosen to be sufficiently small.

*Notation:* Throughout the article, we will often suppress the superscript  $N$  for various objects when it is clear from context.

## 2. FINITE-PARTICLE CONVERGENCE RATES IN KSD METRIC

We first provide rates in the KSD metric. Let  $\mathcal{H}_0$  denote the reproducing kernel Hilbert space (RKHS) of real-valued functions associated with the positive definite kernel  $k$  ([Aronszajn, 1950](#)). Then  $\mathcal{H} := \mathcal{H}_0 \times \cdots \times \mathcal{H}_0$  inherits a natural RKHS structure comprising  $\mathbb{R}^d$ -valued functions. The *Stein operator*  $\mathcal{T}_\pi$ , associated with  $\pi \propto e^{-V}$ , acts on differentiable functions  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^d$  by

$$\mathcal{T}_\pi \phi(x) := -\nabla V(x) \cdot \phi(x) + \nabla \cdot \phi(x), \quad x \in \mathbb{R}^d.$$

The *Kernelized Stein Discrepancy* (KSD) ([Liu et al., 2016](#); [Liu and Wang, 2016](#)), associated with the kernel  $k$ , of a probability measure  $P$  on  $\mathbb{R}^d$  with respect to  $\pi$  is defined as

$$\text{KSD}(P||\pi) := \sup\{\mathbb{E}[\mathcal{T}_\pi \phi(X)] : X \sim P, \phi \in \mathcal{H}, \|\phi\|_{\mathcal{H}} \leq 1\}. \quad (4)$$

The definition of KSD is motivated by *Stein's identity* which says that, for any sufficiently regular  $\phi$ ,  $\mathbb{E}_{X \sim \pi}[\mathcal{T}_\pi \phi(X)] = 0$  and thus, the above measures the ‘distance’ of  $P$  from  $\pi$  via the maximum discrepancy of this expectation from 0 when  $X \sim P$ , as  $\phi$  varies over  $\mathcal{H}$ .

The appeal of KSD lies in the fact that, unlike most distances on the space of probability measures, KSD has an explicit tractable expression. The function  $\phi^* \in \mathcal{H}$  for which the above supremum is attained has a closed form expression  $\phi^*(x) \propto \mathbb{E}_{Y \sim P} [-\mathbf{k}(Y, x) \nabla V(Y) + \nabla_1 \mathbf{k}(Y, x)]$ . Using this, we get the following expression for KSD:

$$\text{KSD}^2(P||\pi) = \mathbb{E}_{(X,Y) \sim P \otimes P} [\nabla V(X) \cdot (\mathbf{k}(X, Y) \nabla V(Y)) - \nabla V(X) \cdot \nabla_2 \mathbf{k}(X, Y) - \nabla V(Y) \cdot \nabla_1 \mathbf{k}(X, Y) + \nabla_1 \cdot \nabla_2 \mathbf{k}(X, Y)]. \quad (5)$$

Before proceeding, we introduce the following regularity conditions, and state an existence and regularity result (proved in Appendix A) for the joint particle density.

**Assumption 1.** *We make the following regularity assumptions.*

- The maps  $(x, y) \mapsto \mathbf{k}(x, y)$  and  $x \mapsto V(x)$  are  $C^3$ .
- $\underline{x}^N(0) = (x_1^N(0), \dots, x_N^N(0))$  has a  $C^1$  density  $p_0^N$ .

**Lemma 1.** *Consider the SVGD dynamics (2) under Assumption 1. Then the particle locations  $(x_1(t), \dots, x_N(t))$  have a joint density  $p(t, \cdot)$  for every  $t \geq 0$ , and the map  $(t, \underline{z}) \mapsto p(t, \underline{z})$  is  $\mathcal{C}^2$ .*

Denote the KL-divergence as

$$\text{KL}(p^N(t)||\pi^{\otimes N}) := \int \log \left( \frac{p^N(t, \underline{z})}{\pi^{\otimes N}(\underline{z})} \right) p^N(t, \underline{z}) d\underline{z}. \quad (6)$$

The following theorem furnishes the key bound on the KSD between the empirical law

$$\mu^N(t) := \frac{1}{N} \sum_{i=1}^N \delta_{x_i(t)}, \quad t \geq 0,$$

and the target distribution  $\pi$ . Define

$$C^*(z) := \nabla_2 \mathbf{k}(z, z) \cdot \nabla V(z) + \mathbf{k}(z, z) \Delta V(z) - \Delta_2 \mathbf{k}(z, z), \quad z \in \mathbb{R}^d. \quad (7)$$

In the above,  $\nabla_2 \mathbf{k}(z, z) := \nabla_2 \mathbf{k}(z, \cdot)(z)$  and  $\Delta_2 \mathbf{k}(z, z) := \Delta_2 \mathbf{k}(z, \cdot)(z)$ .

**Theorem 1.** *Let Assumption 1 hold. Then, we have for every  $T > 0$ ,*

$$\frac{1}{T} \int_0^T \mathbb{E}[\text{KSD}^2(\mu^N(t)||\pi)] dt \leq \frac{\text{KL}(p^N(0)||\pi^{\otimes N})}{NT} + \frac{1}{N^2 T} \int_0^T \mathbb{E} \left[ \sum_{k=1}^N C^*(x_k(t)) \right] dt,$$

where the expectation is with respect to  $p(t)$ . In addition, we have that

$$\text{KL}(p^N(T)||\pi^{\otimes N}) \leq \text{KL}(p^N(0)||\pi^{\otimes N}) + \frac{1}{N} \int_0^T \mathbb{E} \left[ \sum_{k=1}^N C^*(x_k(t)) \right] dt.$$

Moreover, if  $C^* := \sup_{z \in \mathbb{R}^d} C^*(z) < \infty$  and  $\limsup_{N \rightarrow \infty} \text{KL}(p^N(0)||\pi^{\otimes N})/N < \infty$ , then

$$(\mathbb{E}[\text{KSD}(\mu_{av}^N||\pi)])^2 \leq \frac{1}{N} \int_0^T \mathbb{E}[\text{KSD}^2(\mu^N(t)||\pi)] dt \leq \frac{\sup_L \frac{\text{KL}(p^L(0)||\pi^{\otimes L})}{L} + C^*}{N}, \quad (8)$$

where  $\mu_{av}^N(dx) := \frac{1}{N} \int_0^T \mu^N(t, dx) dt$ .

**Remark 1.** *The condition  $C^* < \infty$  holds, for example, when  $\mathbf{k}(u, v) = \Psi(u - v)$  for a positive-definite  $\mathcal{C}^2$  function  $\Psi : \mathbb{R}^d \rightarrow \mathbb{R}$  (Bochner, 1933), and  $\sup_{x \in \mathbb{R}^d} \Delta V(x) < \infty$ . Examples of such kernels include the radial basis kernel (e.g., Gaussian), the Laplacian and the Matérn kernel. The condition on the potential allows for a large class of non-log-concave densities as well.*

**Remark 2** (Optimality and dimension dependence). According to [Sriperumbudur \(2016\)](#); [Hagrass et al. \(2024\)](#), we have under mild regularity conditions, that the empirical measure  $P_N := \frac{1}{N} \sum_{i=1}^N \delta_{X_i}$ , where  $X_i \sim P$ , i.i.d., satisfies  $\mathbb{E}[\text{KSD}(P_N||P)] = O(1/\sqrt{N})$ . This points to the fact that our rates in Theorem 1 are presumably optimal with respect to  $N$ . While there is no curse-of-dimensionality in the KSD rates, the dimension factor appears in the numerator of the bound (8). When  $k(u, v) = \Psi(u - v)$  as in Remark 1 with  $\sup_{x \in \mathbb{R}^d} \Delta V(x) \leq Cd$  for some dimension independent constant  $C$ , it can be checked that  $C^* \leq \Psi(0)Cd - \Delta\Psi(0)$ , which gives a linear in  $d$  upper bound on  $C^*$  for a wide range of kernels (including the Gaussian kernel) and potentials. Moreover, as long as mild regularity conditions are assumed about the kernel and the potential function, then according to [Vempala and Wibisono \(2019, Lemma 1\)](#), the initialization dependent term could be taken to be linear in  $d$  when  $V$  has Lipschitz gradients. These combine to give an  $O(d/\sqrt{N})$  bound on the KSD.

*Proof of Theorem 1.* We will abbreviate  $H(t) := \text{KL}(p^N(t)||\pi^{\otimes N})$ . Using the particle dynamics equation (2) and integration by parts, it is easy to verify that  $p(t, \underline{z})$  is a weak solution of the following  $N$ -body Liouville equation (see, for example, [Golse et al. \(2013, Pg. 7\)](#) and [Ambrosio et al. \(2005, Chapter 8\)](#)) given by

$$\partial_t p(t, \underline{z}) + \frac{1}{N} \sum_{k, \ell=1}^N \text{div}_{z_k} (p(t, \underline{z}) \Phi(z_k, z_\ell)) = 0,$$

where  $\Phi(z, w) := -k(z, w)\nabla V(w) + \nabla_2 k(z, w)$ . Recalling (6), and using the density regularity obtained in Lemma 1, we have that

$$\begin{aligned} H'(t) &= \int \partial_t p(t, \underline{z}) d\underline{z} + \int \log \left( \frac{p(t, \underline{z})}{\pi^{\otimes N}(\underline{z})} \right) \partial_t p(t, \underline{z}) d\underline{z} \\ &= - \int \frac{1}{N} \sum_{k, \ell} \log \left( \frac{p(t, \underline{z})}{\pi^{\otimes N}(\underline{z})} \right) \text{div}_{z_k} (p(t, \underline{z}) \Phi(z_k, z_\ell)) d\underline{z} \\ &= \frac{1}{N} \sum_{k, \ell} \int \nabla_{z_k} \log \left( \frac{p(t, \underline{z})}{\pi^{\otimes N}(\underline{z})} \right) \cdot (p(t, \underline{z}) \Phi(z_k, z_\ell)) d\underline{z} \\ &= \frac{1}{N} \sum_{k, \ell} \int \nabla_{z_k} p(t, \underline{z}) \cdot \Phi(z_k, z_\ell) d\underline{z} + \frac{1}{N} \sum_{k, \ell} \int \nabla V(z_k) \cdot \Phi(z_k, z_\ell) p(t, \underline{z}) d\underline{z} \\ &= \frac{1}{N} \sum_{k, \ell} \int (-\text{div}_{z_k} \Phi(z_k, z_\ell) + \nabla V(z_k) \cdot \Phi(z_k, z_\ell)) p(t, \underline{z}) d\underline{z}. \end{aligned}$$

Now, observe that

$$\begin{aligned} -\text{div}_{z_k} \Phi(z_k, z_\ell) &= \text{div}_{z_k} (k(z_k, z_\ell) \nabla V(z_\ell)) - \text{div}_{z_k} (\nabla_2 k(z_k, z_\ell)) \\ &= \nabla_1 k(z_k, z_\ell) \cdot \nabla V(z_\ell) - \nabla_1 \cdot \nabla_2 k(z_k, z_\ell) \\ &\quad + (\nabla_2 k(z_k, z_k) \cdot \nabla V(z_k) + k(z_k, z_\ell) \Delta V(z_k) - \Delta_2 k(z_k, z_\ell)) \mathbb{1}_{\{k=\ell\}}. \end{aligned}$$

Similarly,

$$\nabla V(z_k) \cdot \Phi(z_k, z_\ell) = -\nabla V(z_k) \cdot (k(z_k, z_\ell) \nabla V(z_\ell)) + \nabla V(z_k) \cdot \nabla_2 k(z_k, z_\ell).$$

Therefore, using the explicit form of KSD in (5), we have

$$\sum_{k, \ell} (-\text{div}_{z_k} \Phi(z_k, z_\ell) + \nabla V(z_k) \cdot \Phi(z_k, z_\ell)) = -N^2 \text{KSD}^2(\mu(\underline{z})||\pi) + \sum_k C^*(z_k),$$



where  $\mu(\underline{z}) := \frac{1}{N} \sum_{i=1}^N \delta_{z_i}$ . Hence, we have

$$\begin{aligned} H'(t) &= -N \int \text{KSD}^2(\mu(\underline{z})||\pi) p(t, \underline{z}) d\underline{z} + \frac{1}{N} \mathbb{E} \left[ \sum_k C^*(x_k(t)) \right] \\ &= -N \mathbb{E}[\text{KSD}^2(\mu^N(t)||\pi)] + \frac{1}{N} \mathbb{E} \left[ \sum_k C^*(x_k(t)) \right], \end{aligned} \quad (9)$$

where we recall that  $\mu^N(t) = \frac{1}{N} \sum_{i=1}^N \delta_{x_i(t)}$  is the empirical measure. Hence, we have

$$\frac{1}{T} \int_0^T \mathbb{E}[\text{KSD}^2(\mu^N(t)||\pi)] dt \leq \frac{H(0)}{NT} + \frac{1}{N^2 T} \int_0^T \mathbb{E} \left[ \sum_k C^*(x_k(t)) \right] dt,$$

which completes the first claim. The entropy bound follows from (9).

To prove the final claim, recall  $\mu_{av}^N(dx) := \frac{1}{N} \int_0^N \mu^N(t, dx) dt$  and note that the map  $Q \mapsto \text{KSD}(Q||\pi)$  is convex, which follows immediately from the representation of KSD given in (4). From this and repeated applications of Jensen's inequality, we obtain

$$\begin{aligned} \mathbb{E}[\text{KSD}(\mu_{av}^N||\pi)] &\leq \frac{1}{N} \int_0^N \mathbb{E}[\text{KSD}(\mu^N(t)||\pi)] dt \\ &\leq \frac{1}{N} \int_0^N \sqrt{\mathbb{E}[\text{KSD}^2(\mu^N(t)||\pi)]} dt \\ &\leq \left( \frac{1}{N} \int_0^N \mathbb{E}[\text{KSD}^2(\mu^N(t)||\pi)] dt \right)^{1/2} \\ &\leq \frac{\left( \sup_L \frac{H(0)}{L} + C^* \right)^{\frac{1}{2}}}{\sqrt{N}}. \end{aligned}$$

This completes the proof of the theorem.  $\square$

**2.1. Marginal convergence.** We now address the convergence in law of the time-averaged marginals when the initial particle locations are drawn from an exchangeable law.

**Theorem 2.** *Suppose Assumption 1 holds,  $C^* < \infty$  and let  $k(u, v) = \Psi(x - y)$ , where  $\Psi$  is a  $\mathcal{C}^2$  function with non-vanishing generalized Fourier transform. Suppose also that the law  $p_0$  of the initial particle locations  $(x_1(0), \dots, x_N(0))$  is exchangeable for each  $N \in \mathbb{N}$  and  $\limsup_{N \rightarrow \infty} \frac{1}{N} \text{KL}(p^N(0)||\pi^{\otimes N}) < \infty$ . Define*

$$\bar{\mu}^N(dx) := \frac{1}{N} \int_0^N \mathbb{P}(x_1(t) \in dx) dt.$$

*Then,  $\bar{\mu}^N \rightarrow \pi$ , weakly.*

*Proof of Theorem 2.* We will first show tightness of  $\{\bar{\mu}^N\}_N$ . By subadditivity of relative entropy (which follows from Budhiraja and Dupuis (2019, Lem. 2.4(b) and Thm. 2.6)), we have

$$\text{KL}(\mathcal{L}(x_1(t))||\pi) \leq \frac{1}{N} \text{KL}(p(t)||\pi^{\otimes N}) \leq \frac{\text{KL}(p(0)||\pi^{\otimes N})}{N} + \frac{C^* t}{N}.$$

Hence, there exists  $C > 0$  such that  $\text{KL}(\mathcal{L}(x_1(t))||\pi) \leq C$  for all  $t \in [0, N]$ ,  $\forall N \geq 1$ . Fix any  $\epsilon > 0$ . Let  $\delta > 0$  such that  $\delta C < \epsilon/2$ . Let  $K$  be a compact subset of  $\mathbb{R}^d$  such that  $\delta \log(1 + \pi(K^c)(e^{1/\delta} - 1)) \leq \epsilon/2$ . By the variational representation of relative entropy (Budhiraja and Dupuis, 2019, Prop. 2.3) and Theorem 1, we have

$$\mathbb{P}(x_1(t) \notin K) \leq \delta \left[ \log \left( \int e^{\frac{1}{\delta} 1_{K^c}(z)} \pi(dz) \right) + \text{KL}(\mathcal{L}(x_1(t))||\pi) \right]$$

$$\leq \delta \log(1 + \pi(K^c)(e^{1/\delta} - 1)) + \delta C < \epsilon$$

for all  $t \in [0, N]$ ,  $\forall N \geq 1$ . In particular,  $\{\bar{\mu}^N\}_N$  is tight. Moreover, we have

$$\begin{aligned} \text{KSD}(\bar{\mu}^N || \pi) &= \text{KSD}\left(\frac{1}{N} \int_0^N \mathcal{L}(x_1(t)) dt || \pi\right) \\ &\leq \frac{1}{N} \int_0^N \text{KSD}(\mathcal{L}(x_1(t)) || \pi) dt \quad \text{by the convexity of KSD} \\ &= \frac{1}{N} \int_0^N \text{KSD}(\mathbb{E}[\mu^N(t)] || \pi) dt \quad \text{using exchangeability, where } \mathbb{E}[\mu^N(t)](dx) := \mathbb{E}[\mu^N(t, dx)] \\ &\leq \frac{1}{N} \int_0^N \mathbb{E}[\text{KSD}(\mu^N(t)) || \pi] dt \rightarrow 0 \quad \text{by KSD convexity and Theorem 1.} \end{aligned}$$

The result now follows from [Gorham and Mackey \(2017, Theorem 7\)](#).  $\square$

### 3. PARTICLE RATES IN $W_2$ METRIC

We now explore convergence rates for SVGD in the  $L^2$ -Wasserstein metric. For  $s > 0$ , let  $\mathcal{P}_s$  be the set of all Borel measurable probability measures on  $\mathbb{R}^d$  with finite  $s$ -moment. For two measures  $\mu, \nu \in \mathcal{P}_s$ , the  $L^s$ -Wasserstein distance (based on the Euclidean distance) is defined as

$$W_s(\mu, \nu) := \left( \inf_{\pi \in \Pi(\mu, \nu)} \mathbb{E}_{(X, Y) \sim \pi} [\|X - Y\|_2^s] \right)^{1/s},$$

where  $\Pi(\mu, \nu)$  denotes the set of all possible couplings of the probability measures  $\mu$  and  $\nu$ .

To address  $W_2$  convergence, we consider the SVGD dynamics in (2) with kernels of the form

$$\tilde{k}(u, v) = 1 + \langle u, v \rangle + \Psi(u - v). \quad (10)$$

We further assume that the kernel obtained by  $(u, v) \mapsto \Psi(u - v)$  is positive-definite,  $\Psi(z) = \Psi(-z)$  for all  $z \in \mathbb{R}^d$ ,  $\sup_{z \in \mathbb{R}^d} \|\nabla \Psi(z)\| < \infty$ ,  $\Psi \in \mathcal{C}^2$  and has a non-vanishing and continuous generalized

Fourier transform. A classical example of a kernel that satisfies these assumptions is the Matérn kernel. We first state the following dissipativity and growth assumptions from [Kanagawa et al. \(2022\)](#) along with an additional Laplacian growth condition.

**Assumption 2.** *The following hold:*

(a) *Dissipativity: The potential  $V$  satisfies*

$$-\langle x, \nabla V(x) \rangle \leq -\alpha \|x\|^2 + \beta_1 \|x\| + (\beta_0 - d),$$

*for some  $\alpha > 0$  and  $\beta_1, \beta_0 \geq 0$ .*

(b) *Growth:  $\|\nabla V(x)\| \leq \lambda_b(1 + \|x\|)$ , for some  $\lambda_b > 0$ .*

(c)  $\sup_{z \in \mathbb{R}^d} \Delta V(z) < \infty$ .

The above assumptions are widely used in the MCMC literature and is satisfied in many cases including certain Gaussian mixture models, and allow for some degree of non-log-concavity in  $\pi$ .

**Theorem 3.** *Let Assumption 2 hold. Assume that the initialization is such that*

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \text{KL}(p(0) || \pi^{\otimes N}) < \infty.$$

*Fix any  $\eta > 0$  and let  $M = M(N) := \lceil N^{2+\eta} \rceil$ . Then, there exists a constant  $C_0 > 0$  such that*

$$\mathbb{E}[\text{KSD}(\mu_{av}^M || \pi)] \leq \frac{C_0}{N^{1+\eta/2}}, \quad \forall N \geq 1 \quad \text{and} \quad W_2(\mu_{av}^M, \pi) \xrightarrow{a.s.} 0, \quad \text{as } N \rightarrow \infty,$$



where recall  $\mu_{av}^M(dx) := \frac{1}{M} \int_0^M \mu^M(t, dx) dt$ .

The additional bilinear term in (10) is the key to tackling Wasserstein convergence. It gives uniform control in  $N, T$  over the second moment of the particle locations in the SVGD dynamics (2), given in the following lemma.

**Lemma 2.** *Under the same setting of Theorem 3, we have that*

$$\limsup_{N \rightarrow \infty} \sup_{T \geq 1} \mathbb{E} \left[ \frac{1}{T} \int_0^T \frac{1}{N} \sum_{i=1}^N \|x_i(t)\|^2 dt \right] < \infty.$$

This result is proved using Lyapunov function techniques and plays a key role in the proof of Theorem 3.

*Proof of Lemma 2.* For this proof, we will abbreviate  $x_i(t)$  as  $x_i$ . Note that, using the SVGD equations (2), we have

$$\begin{aligned} \frac{d}{dt} \left[ \frac{1}{N} \sum_{i=1}^N \nabla V(x_i) \right] &= - \left\| \frac{1}{N} \sum_{i=1}^N \nabla V(x_i) \right\|^2 - \frac{1}{N^2} \sum_{i,j} \langle x_i, x_j \rangle \langle \nabla V(x_i), \nabla V(x_j) \rangle \\ &\quad + \frac{1}{N} \sum_i \langle x_i, \nabla V(x_i) \rangle - \frac{1}{N^2} \sum_{i,j} \langle \nabla V(x_i), \nabla \Psi(x_i - x_j) \rangle \\ &\quad - \underbrace{\frac{1}{N^2} \sum_{i,j} \langle \nabla V(x_i), \Psi(x_i - x_j) \nabla V(x_j) \rangle}_{\geq 0}. \end{aligned} \quad (11)$$

The non-negativity claim above is a consequence of positive-definiteness of the kernel obtained by  $(u, v) \mapsto \Psi(u - v)$ . Note that

$$\begin{aligned} \frac{1}{N^2} \sum_{i,j} \langle x_i, x_j \rangle \langle \nabla V(x_i), \nabla V(x_j) \rangle &= \sum_{\ell, \ell'=1}^d \left( \frac{1}{N} \sum_{i=1}^N x_{i,\ell} (\nabla V(x_i))_{\ell'} \right)^2 \\ &\geq \sum_{\ell=1}^d \left( \frac{1}{N} \sum_{i=1}^N x_{i,\ell} (\nabla V(x_i))_{\ell} \right)^2 \\ &\geq \frac{1}{d} \left( \frac{1}{N} \sum_{i=1}^N \sum_{\ell=1}^d x_{i,\ell} (\nabla V(x_i))_{\ell} \right)^2 \\ &= \frac{1}{d} \left( \frac{1}{N} \sum_{i=1}^N \langle x_i, \nabla V(x_i) \rangle \right)^2, \end{aligned}$$

where the penultimate step follows by Cauchy-Schwartz inequality. Using the above inequality in (11), we obtain

$$\begin{aligned} \frac{d}{dt} \left[ \frac{1}{N} \sum_{i=1}^N \nabla V(x_i) \right] &\leq - \frac{1}{d} \left( \frac{1}{N} \sum_{i=1}^N \langle x_i, \nabla V(x_i) \rangle \right)^2 - \left\| \frac{1}{N} \sum_{i=1}^N \nabla V(x_i) \right\|^2 \\ &\quad + \frac{1}{N} \sum_i \langle x_i, \nabla V(x_i) \rangle - \frac{1}{N^2} \sum_{i,j} \langle \nabla V(x_i), \nabla \Psi(x_i - x_j) \rangle. \end{aligned} \quad (12)$$

By Assumption 2, there exists  $A, \alpha, \beta, \gamma > 0$  such that

$$\langle x, \nabla V(x) \rangle \geq \alpha \|x\|^2 \text{ for } \|x\| \geq A,$$

$$\begin{aligned}\|\nabla V(x)\| &\leq \beta\|x\| \text{ for } \|x\| \geq A, \\ \|\nabla \Psi\|_\infty &\leq \gamma.\end{aligned}$$

Using the above in (12), and defining

$$\Gamma(t) := \sum_i \|x_i\|^2 \mathbb{1}[\|x_i\| \geq A],$$

we obtain

$$\frac{d}{dt} \left[ \frac{1}{N} \sum_{i=1}^N \nabla V(x_i) \right] \leq -\frac{\alpha^2}{d} (\Gamma(t))^2 + \left( \frac{2C\beta}{d} + \beta \right) \Gamma(t) + \beta\gamma (\Gamma(t))^{1/2} + C',$$

where the constants  $C, C' > 0$  are independent of  $N$  (but they depend on  $A$ ). Thus, choosing picking a sufficiently large constant  $B > 0$  (which is independent of  $N$ ), we obtain for a constant  $C_B > 0$  that, for all  $T > 0$ ,

$$\begin{aligned}\frac{\alpha^2}{2d} \int_0^T (\Gamma(t))^2 \mathbb{1}(\Gamma(t) \geq B) dt &\leq \frac{1}{N} \sum_i \nabla V(x_i(0)) + C_B, \\ \int_0^T (\Gamma(t))^2 \mathbb{1}(\Gamma(t) < B) dt &\leq B^2 T.\end{aligned}$$

Therefore, we obtain for all  $T > 0$ ,

$$\begin{aligned}\frac{1}{T} \int_0^T \frac{1}{N} \sum_i \|x_i(t)\|^2 dt &\leq \frac{1}{T} \int_0^T (\Gamma(t))^2 dt + A^2, \\ \text{with } \frac{1}{T} \int_0^T (\Gamma(t))^2 dt &\leq B^2 + \frac{C_B}{T} + \frac{1}{NT} \sum_i V(x_i(0)).\end{aligned}$$

Thus, for all  $T \geq 1$  and  $N \geq 1$  and for a constant  $D > 0$  (which is independent of  $N$ ),

$$\mathbb{E} \left[ \frac{1}{T} \int_0^T \frac{1}{N} \sum_{i=1}^N \|x_i(t)\|^2 \right] \leq D + \frac{1}{T} \mathbb{E} \left[ \frac{1}{N} \sum_i V(x_i(0)) \right].$$

By the variational representation of relative entropy, for  $\delta \in (0, 1)$ ,

$$\mathbb{E} \left[ \frac{1}{N} \sum_i V(x_i(0)) \right] \leq \frac{1}{\delta} \log \left( \int \exp^{\delta V(z)} \pi(dz) \right) + \frac{1}{N\delta} \text{KL}(p^N(0) || \pi^{\otimes N}).$$

By Assumption 2,  $\pi$  is sub-Gaussian. Hence, we have

$$\limsup_{N \rightarrow \infty} \mathbb{E} \left[ \frac{1}{N} \sum_i V(x_i(0)) \right] < \infty,$$

from which, the result follows.  $\square$

*Proof of Theorem 3.* Recall  $H(t) = \text{KL}(p^N(t) || \pi^{\otimes N})$ . From the general KSD bound obtained in Theorem 1, with  $\tilde{\mathbf{k}}$ , we obtain for every  $T > 0$ ,

$$\frac{1}{T} \int_0^T \mathbb{E}[\text{KSD}^2(\mu^N(t) || \pi)] dt \leq \frac{H(0)}{NT} + \frac{1}{N^2 T} \int_0^T \mathbb{E} \left[ \sum_{k=1}^N C^*(x_k(t)) \right] dt,$$

where  $C^*(x_i(t))$  is as defined in (7) with the kernel  $\tilde{\mathbf{k}}$ . Now note that we can obtain constant  $C > 0$  such that, for any  $z \in \mathbb{R}^d$ ,

$$\nabla_2 \tilde{\mathbf{k}}(z, z) \nabla V(z) = \langle z, \nabla V(z) \rangle \leq \|z\| \|\nabla V(z)\| \leq C(1 + \|z\|^2),$$

$$\begin{aligned}\tilde{\mathbf{k}}(z, z)\Delta V(z) &\leq C(\|z\|^2 + 1 + \Psi(0)), \\ -\Delta_2 \tilde{\mathbf{k}}(z, z) &= -\Delta \Psi(0).\end{aligned}$$

Hence, for all  $z \in \mathbb{R}^d$ , we have that  $C^*(z) \leq C_1\|z\|^2 + C_2$ , for some constants  $C_1, C_2 > 0$ . Therefore, we have for any  $t > 0$  and  $N \geq 1$ , that

$$\frac{1}{T} \int_0^T \mathbb{E}[\text{KSD}^2(\mu^N(t) \|\pi)] dt \leq \frac{H(0)}{NT} + \frac{C_1}{NT} \int_0^T \mathbb{E} \left[ \frac{1}{N} \sum_{k=1}^N \mathbb{E}[\|x_k(t)\|^2] \right] dt + \frac{C_2}{N}.$$

Now, using Lemma 2, there exists a constant  $C_4 > 0$  such that for any  $T \geq 1$  and  $N \geq 1$ ,

$$\frac{1}{T} \int_0^T \mathbb{E}[\text{KSD}^2(\mu^N(t) \|\pi)] dt \leq \frac{H(0)}{NT} + \frac{C_4}{N}.$$

By the convexity of KSD, we have for  $N \geq 1$ ,

$$\mathbb{E}[\text{KSD}(\mu_{av}^M \|\pi)] \leq \frac{C_0}{N^{1+\eta/2}}.$$

Hence, by Borel–Cantelli lemma, we have that  $\text{KSD}(\mu_{av}^M \|\pi) \xrightarrow{a.s.} 0$ , as  $N \rightarrow \infty$ . The stated Wasserstein convergence result now follows by Kanagawa et al. (2022, Thm. 3.1) taking  $m = \text{Id}$ ,  $q_m = 1$ ,  $q = 2$ ,  $L = L^{(2)}$  and  $\Phi = \Psi$ .  $\square$

By the results in Kanagawa et al. (2022, Section 3.2), we can translate the KSD bound in Theorem 3 into a bound in the  $W_2$  metric. For computational clarity, we restrict ourselves to the Matérn-family of kernels. Specifically we consider (10) with

$$\tilde{\mathbf{k}}_{\text{mk}}(u, v) := 1 + \langle u, v \rangle + \underbrace{\frac{2^{1-(d/2+\nu)}}{\Gamma(d/2+\nu)} \|\Sigma(u-v)\|_2^\nu K_{-\nu}(\|\Sigma(u-v)\|_2)}_{:= \Psi_{\text{mk}}(u-v)}, \quad (13)$$

where  $\Gamma$  is the Gamma function,  $\Sigma$  a strictly positive definite matrix, and  $K_{-\nu}$  the modified Bessel function of the second kind of order  $-\nu$ .

To proceed, we also require the following assumption from Kanagawa et al. (2022), which is motivated by the Langevin diffusion:

$$dZ_t = -\nabla V(Z_t)dt + \sqrt{2}dB_t, \quad (14)$$

where  $(B_t)_{t \geq 0}$  is a  $d$ -dimensional Brownian motion. Note that (14) is the stochastic differential equation equivalent of the Wasserstein gradient flow in (3); see, for example, Jordan et al. (1998); Bakry et al. (2014) for details. The connection between the two perspectives has in particular proved to be extremely useful for analyzing both Markov Chain Monte Carlo algorithms and particle-based methods.

**Assumption 3.** For  $s \in \{1, 2\}$ , let  $\rho_s : [0, \infty) \rightarrow [0, \infty)$  be an upper bounding function for the  $L^s$ -Wasserstein distance in the following sense:

$$W_s(\text{Law}(Z_t^x), \text{Law}(Z_t^y)) \leq \rho_s(t) \|x - y\|_2, \quad \forall x, y \in \mathbb{R}^d, t \geq 0, \quad (15)$$

where  $Z_t^x$  and  $Z_t^y$  are Langevin diffusion processes in (14) with initializations  $Z_0 = x$  and  $Z_0 = y$ .

Assume that  $\tilde{\rho}(t) := \frac{\log \frac{\rho_1(t)}{\rho_1(0)} - \log \rho_2(t)}{\log \frac{\rho_1(t)}{\rho_1(0)}}$  is uniformly bounded in  $t$  and, moreover,

$$\int_0^\infty \rho_1(t) \left\{ 1 + \sqrt{\rho_1(t)} \tilde{\rho}(t) \right\} dt < \infty.$$

**Remark 3.** Suppose there exist  $U > 0$  and  $R, L \geq 0$  such that the potential  $V$  satisfies

$$\frac{\langle \nabla V(x) - \nabla V(y), x - y \rangle}{\|x - y\|_2^2} \leq \begin{cases} -U & \text{if } \|x - y\|_2 > R, \\ L & \text{if } \|x - y\|_2 \leq R. \end{cases} \quad (16)$$

Then, by [Eberle \(2011, 2016\)](#), there exists  $c, c_1 > 0$  such that we can set  $\rho_1(t) = ce^{-c_1 t}$ ,  $t \geq 0$ . Moreover, using (16) and Grönwall's lemma, we can set  $\rho_2(t) = e^{Lt}$ ,  $t \geq 0$ . Consequently,  $\tilde{\rho}(t) = \frac{L+c_1}{c_1}$  and hence Assumption 3 holds.

**Theorem 4.** Consider the SVGD updates in (2) with the kernel in (13). Suppose Assumption 3 and the assumptions made in Theorem 3 are satisfied. Then, there exists a constant  $C(d) > 0$  such that for any  $\epsilon \in (0, 1)$ , we have

$$\mathbb{P} \left[ W_2(\mu_{av}^M, \pi) \geq C(d) \left( \frac{C_0}{\epsilon N^{1+\eta/2}} \right)^{r(d)} \right] \leq \epsilon, \quad \forall N \geq \left( \frac{C_0}{\epsilon} \right)^{\frac{2}{2+\eta}},$$

where  $C_0$  is the same constant from Theorem 3,  $C(d)$  is the constant  $C_{P,d}(1)$  from [Kanagawa et al. \(2022, Theorem 3.5\)](#), and

$$r(d) := \frac{1}{3(\frac{4d+1}{d})} \cdot \frac{1}{\frac{3d}{2} + \frac{17}{6} + [\frac{d+1}{d} + \frac{5}{3}]} \nu. \quad (17)$$

**Remark 4.** Note that  $r(d) \approx \frac{1}{18d}$  for large  $d$ . Hence, unlike the KSD rates in Theorem 4, the  $W_2$  rates have a curse-of-dimensionality. However, this is expected, as even in the case of i.i.d. samples, we have a similar curse-of-dimensionality ([Dudley, 1969](#); [Weed and Bach, 2019](#)).

*Proof of Theorem 4.* By Theorem 3.5 in [Kanagawa et al. \(2022\)](#), we have that

$$W_2(\mu_{av}^M, \pi) \leq C(d)(1 \vee \text{KSD}(\mu_{av}^M || \pi)^{(1-r(d))}) \text{KSD}(\mu_{av}^M || \pi)^{r(d)},$$

where, from [Kanagawa et al. \(2022\)](#) we have

$$r(d) = \frac{1}{3(\frac{4d+1}{d})} \frac{1}{1+t_1} \quad \text{where } t_1 = \frac{3d+1}{2} + \frac{1}{3} + \left[ \frac{d+1}{d} + \frac{5}{3} \right] \nu,$$

resulting in (17). Define  $\mathcal{E} := \left\{ \text{KSD}(\mu_{av}^M || \pi) \leq \frac{C_0}{\epsilon N^{1+\eta/2}} \right\}$ . By Theorem 3 and Markov's inequality, we have that  $\mathbb{P}[\mathcal{E}^c] \leq \epsilon$ . On the event  $\mathcal{E}^c$ , for  $N \geq \left( \frac{C_0}{\epsilon} \right)^{\frac{2}{2+\eta}}$ , we have

$$W_2(\mu_{av}^M, \pi) \leq C(d) \left( \frac{C_0}{\epsilon N^{1+\eta/2}} \right)^{r(d)},$$

thereby proving the claim.  $\square$

#### 4. PROPAGATION OF CHAOS

We now exhibit a *long time propagation of chaos* (POC) for the particle system started from an exchangeable initial configuration and driven by the dynamics (2). More precisely, we show that, under the conditions of Theorem 3, the time-averaged marginals of particle locations over the time interval  $[0, N]$  become asymptotically independent, with distribution  $\pi$ , as  $N \rightarrow \infty$ .

**Proposition 1.** Suppose that the law  $p_0^N$  of the initial particle locations  $(x_1^N(0), \dots, x_N^N(0))$  is exchangeable for each  $N \in \mathbb{N}$ . For  $1 \leq k \leq N$ , define the  $k$ -dimensional marginal of the time-averaged occupancy measure of particle locations as follows:

$$\bar{\mu}_k^N(dx_1, \dots, dx_k) := \frac{1}{N} \int_0^N \mathbb{P}(x_1^N(t) \in dx_1, \dots, x_k^N(t) \in dx_k) dt.$$

Recall  $M = M(N) := \lceil N^{2+\eta} \rceil$ . Under the same setting as Theorem 3, we have that, for any fixed  $k \in \mathbb{N}$ ,  $W_1(\bar{\mu}_k^M, \pi^{\otimes k}) \rightarrow 0$ , as  $N \rightarrow \infty$ .

**Remark 5.** This result should be compared with [Shi and Mackey \(2024, Theorem 2\)](#) and [Lu et al. \(2019, Proposition 2.6\)](#) where finite-time POC results are shown: the particle marginal laws at a fixed time become asymptotically independent as  $N \rightarrow \infty$ . However, owing to the lack of Lipschitz property of the vector field driving (2), this POC can only be extended to growing times  $t = t_N = O(\log \log N)$  when the particle marginal laws are not necessarily close to  $\pi$ . In contrast, [Proposition 1](#) extends to the time interval  $[0, N]$  (hence, long-time POC) and the time-averaged particle trajectories essentially produce i.i.d. samples from  $\pi$ .

*Proof of Proposition 1.* Let  $\mathcal{P}(\mathbb{R}^d)$  denote the space of probability measures on  $\mathbb{R}^d$ , and denote by  $\mathcal{P}(\mathcal{P}(\mathbb{R}^d))$  the space of probability measures on  $\mathcal{P}(\mathbb{R}^d)$ . Let  $\mathcal{L}(\mu_{av}^M)$  denote the law of the random measure  $\mu_{av}^M$  and  $\delta_\pi$  denote the Dirac measure at  $\pi$  in  $\mathcal{P}(\mathcal{P}(\mathbb{R}^d))$ .

By [Lemma 2](#) and exchangeability,

$$\sup_{N \geq 1} \mathbb{E} \left[ \int_{\mathbb{R}^d} \|x\|^2 \mu_{av}^M(dx) \right] = \sup_{N \geq 1} \mathbb{E} \left[ \int_{\mathbb{R}^d} \|x\|^2 \bar{\mu}_1^M(dx) \right] < \infty. \quad (18)$$

Moreover, by [Assumption 2\(a\)](#),  $\int_{\mathbb{R}^d} \|x\|^2 \pi(dx) < \infty$ . Hence, by [Chaintron and Diez \(2022, Theorem 3.21\)](#), we conclude that  $W_1(\bar{\mu}_k^M, \pi^{\otimes k}) \rightarrow 0$  if and only if  $W_1(\mathcal{L}(\mu_{av}^M), \delta_\pi) \rightarrow 0$  as  $N \rightarrow \infty$ , where  $W_1$  is the Wasserstein distance on the space  $\mathcal{P}(\mathcal{P}(\mathbb{R}^d))$  equipped with the distance function  $W_1$  as defined in [Chaintron and Diez \(2022, Definition 3.5\)](#).

Note that  $W_1(\mathcal{L}(\mu_{av}^M), \delta_\pi) \leq \mathbb{E}[W_1(\mu_{av}^M, \pi)]$ . By [Theorem 3](#) and Jensen's inequality,

$$W_1(\mu_{av}^M, \pi) \xrightarrow{a.s.} 0 \quad \text{as} \quad N \rightarrow \infty.$$

Moreover, observe that

$$W_1^2(\mu_{av}^M, \pi) \leq W_2^2(\mu_{av}^M, \pi) \leq 2 \int_{\mathbb{R}^d} \|x\|^2 \mu_{av}^M(dx) + 2 \int_{\mathbb{R}^d} \|x\|^2 \pi(dx)$$

and hence, by (18) and [Assumption 2\(a\)](#),  $\sup_{N \geq 1} \mathbb{E}[W_1^2(\mu_{av}^M, \pi)] < \infty$ . In particular,  $\{W_1(\mu_{av}^M, \pi) : N \geq 1\}$  is uniformly integrable and thus  $\mathbb{E}[W_1(\mu_{av}^M, \pi)] \rightarrow 0$  as  $N \rightarrow \infty$ . The result follows.  $\square$

## REFERENCES

- Luigi Ambrosio, Nicola Gigli, and Giuseppe Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2005. (Cited on page 6.)
- Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950. (Cited on page 4.)
- Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 103. Springer, 2014. (Cited on page 11.)
- David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017. (Cited on page 1.)
- Salomon Bochner. Monotone funktionen, stieltjessche integrale und harmonische analyse. *Mathematische Annalen*, 108(1):378–410, 1933. (Cited on page 5.)
- Amarjit Budhiraja and Paul Dupuis. Analysis and approximation of rare events. *Representations and Weak Convergence Methods. Series Prob. Theory and Stoch. Modelling*, 94, 2019. (Cited on page 7.)
- José A Carrillo and Jakub Skrzeczkowski. Convergence and stability results for the particle system in the Stein gradient descent method. *arXiv preprint arXiv:2312.16344*, 2023. (Cited on page 2.)
- Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: A review of models, methods and applications. i. models and methods. *Kinetic and Related Models*, 15(6):895–1015, 2022. (Cited on page 13.)

- Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, and Philippe Rigollet. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. *Advances in Neural Information Processing Systems*, 33:2098–2109, 2020. (Cited on page 2.)
- Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pages 2606–2615. PMLR, 2016. (Cited on page 2.)
- Gianluca Crippa. *The flow associated to weakly differentiable vector fields*. PhD thesis, University of Zurich, 2008. (Cited on page 15.)
- Aniket Das and Dheeraj Nagaraj. Provably fast finite particle variants of SVGD via virtual particle stochastic approximation. *Advances in Neural Information Processing Systems*, 36:49748–49760, 2023. (Cited on page 3.)
- Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969. (Cited on pages 4 and 12.)
- Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 24:1–39, 2023. (Cited on page 2.)
- Andreas Eberle. Reflection coupling and wasserstein contractivity without convexity. *Comptes Rendus Mathématique*, 349(19-20):1101–1104, 2011. (Cited on page 12.)
- Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166:851–886, 2016. (Cited on page 12.)
- Yihao Feng, Dilin Wang, and Qiang Liu. Learning to draw samples with amortized Stein variational gradient descent. In *Conference on Uncertainty in Artificial Intelligence*, 2017. (Cited on page 1.)
- François Golse, Clément Mouhot, and Valeria Ricci. Empirical measures and Vlasov hierarchies. *arXiv preprint arXiv:1309.0222*, 2013. (Cited on page 6.)
- Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017. (Cited on pages 2, 4, and 8.)
- Jackson Gorham, Anant Raj, and Lester Mackey. Stochastic Stein discrepancies. *Advances in Neural Information Processing Systems*, 33:17931–17942, 2020. (Cited on page 2.)
- Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pages 1352–1361. PMLR, 2017. (Cited on page 1.)
- Omar Hagrass, Bharath Sriperumbudur, and Krishnakumar Balasubramanian. Minimax optimal goodness-of-fit testing with kernel Stein discrepancy. *arXiv preprint arXiv:2404.08278*, 2024. (Cited on page 6.)
- Philip Hartman. *Ordinary differential equations*. SIAM, 2002. (Cited on page 15.)
- Ye He, Krishnakumar Balasubramanian, Bharath K Sriperumbudur, and Jianfeng Lu. Regularized Stein variational gradient flow. *Foundations of Computational Mathematics*, pages 1–59, 2024. (Cited on page 2.)
- Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. (Cited on page 11.)
- Heishiro Kanagawa, Alessandro Barp, Arthur Gretton, and Lester Mackey. Controlling moments with kernel stein discrepancies. *arXiv preprint arXiv:2211.05408*, 2022. (Cited on pages 4, 8, 11, and 12.)
- Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for Stein Variational Gradient Descent. *Advances in Neural Information Processing Systems*, 33, 2020. (Cited on page 2.)
- Qiang Liu. Stein Variational Gradient Descent as gradient flow. *Advances in Neural Information Processing Systems*, 30, 2017. (Cited on page 2.)
- Qiang Liu and Dilin Wang. Stein Variational Gradient Descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29, 2016. (Cited on pages 1, 2, and 4.)



- Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284. PMLR, 2016. (Cited on pages 2 and 4.)
- Tianle Liu, Promit Ghosal, Krishnakumar Balasubramanian, and Natesh Pillai. Towards understanding the dynamics of Gaussian-Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 36, 2024. (Cited on pages 2, 3, and 4.)
- Xingchao Liu, Xin Tong, and Qiang Liu. Sampling with trustworthy constraints: A variational gradient framework. *Advances in Neural Information Processing Systems*, 34:23557–23568, 2021. (Cited on page 1.)
- Jianfeng Lu, Yulong Lu, and James Nolen. Scaling limit of the Stein Variational Gradient Descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671, 2019. (Cited on pages 2, 3, and 13.)
- Victor Priser, Pascal Bianchi, and Adil Salim. Long-time asymptotics of noisy SVGD outside the population limit. *arXiv preprint arXiv:2406.11929*, 2024. (Cited on page 3.)
- Adil Salim, Lukang Sun, and Peter Richtarik. A convergence theory for SVGD in the population limit under Talagrand’s inequality  $T_1$ . In *International Conference on Machine Learning*, pages 19139–19152. PMLR, 2022. (Cited on page 2.)
- Jiaxin Shi and Lester Mackey. A finite-particle convergence rate for Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 36, 2024. (Cited on pages 1, 3, and 13.)
- Bharath Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, pages 1839–1893, 2016. (Cited on page 6.)
- Lukang Sun, Avetik Karagulyan, and Peter 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. (Cited on page 2.)
- Santosh Vempala and Andre Wibisono. Rapid convergence of the Unadjusted Langevin Algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019. (Cited on page 6.)
- Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019. (Cited on pages 4 and 12.)
- Lantian Xu, Anna Korba, and Dejan Slepčev. Accurate quantization of measures via interacting particle-based optimization. In *International Conference on Machine Learning*, pages 24576–24595. PMLR, 2022. (Cited on page 1.)

## APPENDIX A. AUXILIARY RESULTS

*Proof of Lemma 1.* Let  $F : (\mathbb{R}^d)^N \rightarrow (\mathbb{R}^d)^N$  be given by

$$F_i(\underline{z}) := -\frac{1}{N} \sum_j \mathbf{k}(z_i, z_j) \nabla V(z_j) + \frac{1}{N} \sum_j \nabla_2 \mathbf{k}(z_i, z_j),$$

for  $1 \leq i \leq N$ . By Assumption 1,  $F$  is a  $\mathcal{C}^2$  map. Note that the SVGD particle trajectories can be written as  $\{\mathbf{x}(t, \underline{z}(0)) : t \geq 0\}$ , where the flow  $\mathbf{x} : [0, \infty) \times (\mathbb{R}^d)^N \rightarrow (\mathbb{R}^d)^N$  is given by

$$\dot{\mathbf{x}}(t, \underline{z}) = F(\mathbf{x}(t, \underline{z})), \quad \mathbf{x}(0, \underline{z}) = \underline{z},$$

with  $\dot{\mathbf{x}}$  denoting the time derivative. By Hartman (2002, Chapter 5, Cor. 4.1), the map  $(t, \underline{z}) \mapsto \mathbf{x}(t, \underline{z})$ , and consequently, the inverse map  $(t, \underline{z}) \mapsto \mathbf{x}(t, \cdot)^{-1}(\underline{z})$ , are  $\mathcal{C}^2$  maps on  $(0, \infty) \times (\mathbb{R}^d)^N$ .

A simple change of variable formula gives (see Crippa (2008, Pg. 21))

$$p(t, \underline{z}) = \frac{p_0}{\det(\nabla \mathbf{x}(t, \cdot))} \circ \mathbf{x}(t, \cdot)^{-1}(\underline{z}), \quad (t, \underline{z}) \in (0, \infty) \times (\mathbb{R}^d)^N.$$

The existence and regularity of  $p(\cdot, \cdot)$  then follows from the above observations.  $\square$