A Finite-Particle Convergence Rate for Stein Variational Gradient Descent

Jiaxin Shi[*] JIAXINS@STANFORD.EDU
Stanford University

Lester Mackey LMACKEY@MICROSOFT.COM
Microsoft Research New England

Abstract

We provide a first finite-particle convergence rate for Stein variational gradient descent (SVGD). Specifically, whenever the target distribution is sub-Gaussian with a Lipschitz score, SVGD with n particles and an appropriate step size sequence drives the kernel Stein discrepancy to zero at an order $1/\sqrt{\log\log n}$ rate. We suspect that the dependence on n can be improved, and we hope that our explicit, non-asymptotic proof strategy will serve as a template for future refinements.

1. Introduction

Stein variational gradient descent [SVGD, 14] is an algorithm for approximating a target probability distribution P on $\mathbb{R}^d$ with a collection of n particles. Given an initial particle approximation $\mu_0^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ with locations $x_i \in \mathbb{R}^d$, SVGD (Algorithm 1) iteratively evolves the particle locations to provide a more faithful approximation of the target P by performing optimization in the space of probability measures. SVGD has demonstrated encouraging results for a wide variety of inferential tasks, including approximate inference [14, 22, 23], generative modeling [21], and reinforcement learning [9, 16].

Algorithm 1 Stein Variational Gradient Descent [14]: SVGD(μ_0, r)

Input: Target P with density p, kernel k, step sizes $(\epsilon_s)_{s\geq 0}$, approximating measure μ_0, rounds r
for $s = 0, \cdots, r-1$ **do**
  Let μ_{s+1} be the distribution of $X^s + \epsilon_s \int k(x, X^s)\nabla \log p(x) + \nabla_x k(x, X^s) d\mu_s(x)$ for $X^s \sim \mu_s$.
Output: Updated approximation μ_r of the target P

Despite the popularity of SVGD, relatively little is known about its approximation quality. A first analysis by Liu [13, Thm. 3.3] showed that *continuous SVGD*—that is, Algorithm 1 initialized with a continuous distribution μ_0^∞ in place of the discrete particle approximation μ_0^n—converges to P in kernel Stein discrepancy [KSD, 4, 7, 15]. KSD convergence is also known to imply weak convergence under various conditions on the target P and the SVGD kernel k [1, 3, 7, 10]. Follow-up work by Korba et al. [11], Salim et al. [17], Sun et al. [19] sharpened the result of Liu with path-independent constants, weaker smoothness conditions, and explicit rates of convergence. In addition, Duncan et al. [5] analyzed the continuous-time limit of continuous SVGD to provide

[*] Part of this work was done at Microsoft Research New England.

conditions for exponential convergence. However, each of these analyses applies only to continuous SVGD and not to the finite-particle algorithm used in practice.

To bridge this gap, Liu [13, Thm. 3.2] showed that n-particle SVGD converges to continuous SVGD in bounded-Lipschitz distance but only under boundedness assumptions violated by most applications of SVGD. To provide a more broadly applicable proof of convergence, Gorham et al. [8, Thm. 7] showed that n-particle SVGD converges to continuous SVGD in 1-Wasserstein distance under assumptions commonly satisfied in SVGD applications. However, both convergence results are asymptotic, providing neither explicit error bounds nor rates of convergence. Korba et al. [11, Prop. 7] explicitly bounded the expected squared Wasserstein distance between n-particle and continuous SVGD but only under the assumption of bounded $\nabla \log p$, an assumption that rules out all strongly log concave or dissipative distributions and all distributions for which the KSD is currently known to control weak convergence [1, 3, 7, 10]. In addition, Korba et al. [11] do not provide a unified bound for the convergence of n-particle SVGD to P.

In this work, we derive a first unified convergence bound for finite-particle SVGD to its target. To achieve this, we first bound the 1-Wasserstein discretization error between finite-particle and continuous SVGD under assumptions commonly satisfied in SVGD applications and compatible with KSD weak convergence control (see Theorem 1). We next bound KSD in terms of 1-Wasserstein distance and SVGD moment growth to explicitly control KSD discretization error in Theorem 2. Finally, Theorem 3 combines our results with the established KSD analysis of continuous SVGD to arrive at an explicit KSD error bound for n-particle SVGD.

2. Notation and Assumptions

Throughout, we fix a target distribution P in the set $\mathcal{P}$ of probability measures on $\mathbb{R}^d$, a nonnegative step size sequence $(\epsilon_s)_{s\geq 0}$, and a reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ and product RKHS $\mathcal{H}^d \triangleq \bigotimes_{i=1}^d \mathcal{H}$ [2]. For all $\mu, \nu \in \mathcal{P}$, we let $W_1(\mu,\nu) \triangleq \inf_{X\sim\mu, Z\sim\nu} \mathbb{E}[\|Z - X\|_2]$ denote the 1-Wasserstein distance between them and introduce the shorthand $m_\mu \triangleq \mathbb{E}_\mu[\|\cdot\|]_2$, $m_{\mu,P} \triangleq \mathbb{E}_{X\sim\mu \perp\!\!\!\perp Z\sim P}[\|X - Z\|_2]$, and $M_{\mu,P} \triangleq \mathbb{E}_{X\sim\mu \perp\!\!\!\perp Z\sim P}[\|X - Z\|_2^2]$. We allow each of these quantities to take on the value ∞ when the random variables are not suitably integrable. We further define the Kullback-Leibler divergence as $\mathrm{KL}(\mu\|\nu) \triangleq \mathbb{E}_\mu[\log(\frac{d\mu}{d\nu})]$ when μ is absolutely continuous with respect to ν (denoted by $\mu \ll \nu$) and as ∞ otherwise.

Our analysis will make use of the following standard assumptions on the SVGD kernel and target distribution.

Assumption 1 (Lipschitz, mean-zero score function) *The target distribution $P \in \mathcal{P}$ has a twice differentiable density p with an L-Lipschitz* score function $s_p \triangleq \nabla \log p$, *i.e.,* $\|s_p(x) - s_p(y)\|_2 \leq L\|x - y\|_2$ *for all $x, y \in \mathbb{R}^d$. Moreover, $\mathbb{E}_P[s_p] = 0$ and $s_p(x^\star) = 0$ for some $x^\star \in \mathbb{R}^d$.*

Assumption 2 (Bounded kernel derivatives) *For any multi-index $I = (I_1, I_2, \ldots, I_d)$ with $|I| \triangleq \sum_{i=1}^d I_i \leq 2$, we have $\sup_{x\in\mathbb{R}^d}\left(D_x^I D_y^I k(x,y)|_{y=x}\right) \leq \kappa^2$. Here, D^I is the differential operator defined as $D_x^I = \frac{d^{|I|}}{dx_1^{I_1} dx_2^{I_2} \ldots dx_d^{I_d}}$.*

Assumption 3 (Decaying kernel derivatives) *There exists $\gamma > 0$ such that*

$$\sup_{x,y\in\mathbb{R}^d, \|x-y\|_2\geq r} \|\nabla_x k(x,y)\|_2 \leq \gamma/r.$$

To leverage the continuous SVGD convergence rates of Salim et al. [17], we additionally assume that the target P satisfies Talagrand's T_1 inequality [20, Def. 22.1]. Remarkably, Villani [20, Thm. 22.10] showed that Assumption 4 is *equivalent* to P being a sub-Gaussian distribution. Hence, this mild assumption holds for all strongly log concave P [18, Def. 2.9], all P satisfying the log Sobolev inequality [20, Thm. 22.17], and all *distantly dissipative* P for which KSD is known to control weak convergence [7, Def. 4].

Assumption 4 (Talagrand's T_1 inequality [20, Def. 22.1]) $\mathbb{E}_P[\|\cdot\|_2] < \infty$, *and there exists* $\lambda > 0$ *such that for all* μ *with* $\mathbb{E}_\mu[\|\cdot\|_2] < \infty$

$$W_1(\mu, P) \leq \sqrt{2\mathrm{KL}(\mu\|P)/\lambda}.$$

Finally we make use of the following notation specific to the SVGD algorithm.

Definition 1 (Stein operator) *For any vector-valued function* $g : \mathbb{R}^d \to \mathbb{R}^d$, *the* Langevin Stein operator [6] *for* P *satisfying Assumption 1 is defined by*

$$(\mathcal{T}_P g)(x) \triangleq \langle s_p(x), g(x)\rangle + \nabla \cdot g(x).$$

Definition 2 (Vector-valued Stein operator) *For any function* $h : \mathbb{R}^d \to \mathbb{R}$, *the* vector-valued Langevin Stein operator [15] *for* P *satisfying Assumption 1 is defined by*

$$(\mathcal{A}_P h)(x) \triangleq s_p(x)h(x) + \nabla h(x)$$

with components

$$(\mathcal{A}_P^j h)(x) \triangleq (s_p(x))_j\, h(x) + \nabla_j h(x).$$

Definition 3 (SVGD transport map and pushforward) *The* SVGD transport map [14] *for a kernel* k, *target* P *satisfying Assumption 1, approximating* $\mu \in \mathcal{P}$, *and step size* ϵ *is given by*

$$T_{\mu,\epsilon}(x) \triangleq x + \epsilon \mathbb{E}_{X\sim\mu}[(\mathcal{A}_P k(\cdot, x))(X)].$$

Moreover, the SVGD pushforward $\Phi_\epsilon(\mu)$ *represents the distribution of* $T_{\mu,\epsilon}(X)$ *when* $X \sim \mu$.

Definition 4 (Kernel Stein discrepancy) *For a target* P *satisfying Assumption 1 and measures* $\mu, \nu \in \mathcal{P}$, *we define the* Langevin kernel Stein discrepancy [KSD, 4, 7, 15] *with base kernel* k *by*[1]

$$\mathrm{KSD}_P(\mu, \nu) \triangleq \sup_{\|g\|_{\mathcal{H}^d}\leq 1} \mathbb{E}_\mu[\mathcal{T}_P g] - \mathbb{E}_\nu[\mathcal{T}_P g].$$

3. Wasserstein Discretization Error of SVGD

Our first main result concerns the discretization error of SVGD and shows that n-particle SVGD does not stray too far from its continuous SVGD limit when its step sizes are chosen appropriately.

1. Prior work [4, 7, 15] only defined $\mathrm{KSD}_P(\mu, \nu)$ for the case $\nu = P$. Definition 4 extends this definition to general ν.

Theorem 1 (Wasserstein discretization error of SVGD) *Suppose Assumptions 1, 2, and 3 hold. For any $\mu_0^n, \mu_0^\infty \in \mathcal{P}$, the Algorithm 1 outputs $\mu_r^n = \mathrm{SVGD}(\mu_0^n, r)$ and $\mu_r^\infty = \mathrm{SVGD}(\mu_0^\infty, r)$ satisfy*

$$W_1(\mu_r^n, \mu_r^\infty) \leq W_1(\mu_0^n, \mu_0^\infty) \exp(b_{r-1}(A + B\exp(Cb_{r-1})))$$

for $b_{r-1} \triangleq \sum_{s=0}^{r-1} \epsilon_s$, $A = (c_1 + c_2)(1 + m_P)$, $B = c_1 m_{\mu_0^n, P} + c_2 m_{\mu_0^\infty, P}$, and $C = \kappa^2(3L + d)$. Here $c_1 = \max(\sqrt{d}\kappa^2 L, \sqrt{d}\kappa^2 L\|x^\star\|_2 + d\kappa^2)$ and $c_2 = \kappa^2 L + d\kappa^2 + L(\gamma + \sqrt{d}\kappa^2)(1 + \|x^\star\|_2)$.

The proof of Theorem 1 in Appendix A relies on two lemmas. The first, due to Gorham et al. [8], shows that the one-step SVGD pushforward Φ_ϵ (Definition 3) is pseudo-Lipschitz with respect to the 1-Wasserstein distance whenever the score function $\nabla \log p$ and kernel k fulfill a commonly-satisfied pseudo-Lipschitz condition.

Lemma 1 (Wasserstein pseudo-Lipschitzness of SVGD [8, Lem. 12]) *For $P \in \mathcal{P}$ with differentiable density p, suppose that the following pseudo-Lipschitz bounds hold*

$$\sup_{z \in \mathbb{R}^d} \|\nabla_z (s_p(x)k(x,z) + \nabla_x k(x,z))\|_{\mathrm{op}} \leq c_1(1 + \|x\|_2),$$
$$\sup_{x \in \mathbb{R}^d} \|\nabla_x (s_p(x)k(x,z) + \nabla_x k(x,z))\|_{\mathrm{op}} \leq c_2(1 + \|z\|_2).$$

for some constants $c_1, c_2 > 0$. Then, for any $\mu, \nu \in \mathcal{P}$,

$$W_1(\Phi_\epsilon(\mu), \Phi_\epsilon(\nu)) \leq W(\mu, \nu)(1 + \epsilon c_{\mu,\nu}),$$

where Φ_ϵ is the one-step SVGD pushforward (Definition 3) and $c_{\mu,\nu} = c_1(1 + m_\mu) + c_2(1 + m_\nu)$.

The second lemma, proved in Appendix B, controls the growth of the first and second absolute moments under SVGD.

Lemma 2 (SVGD moment growth) *Suppose Assumptions 1 and 2 hold, and let $C = \kappa^2(3L + d)$. Then the SVGD output μ_r of Algorithm 1 with $b_{r-1} \triangleq \sum_{s=0}^{r-1} \epsilon_s$ satisfies*

$$m_{\mu_r} \leq m_{\mu_0, P} \prod_{s=0}^{r-1}(1 + \epsilon_s C) + m_P \leq m_{\mu_0, P} \exp(Cb_{r-1}) + m_P,$$
$$M_{\mu_r, P} \leq M_{\mu_0, P} \prod_{s=0}^{r-1}(1 + \epsilon_s C)^2 \leq M_{\mu_0, P} \exp(2Cb_{r-1}).$$

4. KSD Discretization Error of SVGD

Our next result translates the Wasserstein error bounds of Theorem 1 into KSD error bounds.

Theorem 2 (KSD discretization error of SVGD) *Suppose Assumptions 1 and 2 hold. For any $\mu_0^n, \mu_0^\infty \in \mathcal{P}$, the Algorithm 1 outputs $\mu_r^n = \mathrm{SVGD}(\mu_0^n, r)$ and $\mu_r^\infty = \mathrm{SVGD}(\mu_0^\infty, r)$ satisfy*

$$\mathrm{KSD}_P(\mu_r^n, \mu_r^\infty) \leq \kappa(d + L) w_{0,n} \exp(b_{r-1}(A + B\exp(Cb_{r-1})))$$
$$+ d^{1/4}\kappa L \sqrt{2M_{\mu_0^\infty, P} w_{0,n}} \exp(b_{r-1}(2C + A + B\exp(Cb_{r-1}))/2)$$

for $w_{0,n} \triangleq W_1(\mu_0^n, \mu_0^\infty)$ and A, B, C defined as in Theorem 1.

Our proof of Theorem 2 relies on the following lemma, proved in Appendix C, that shows that the KSD is controlled by the 1-Wasserstein distance.

Lemma 3 (KSD-Wasserstein bound) *Suppose Assumptions 1 and 2 hold. For any $\mu, \nu \in \mathcal{P}$,*

$$\begin{aligned}\mathrm{KSD}_P(\mu,\nu) &\leq \kappa(d+L)W_1(\mu,\nu)+L\mathbb{E}_{(X\sim\mu,Y\sim\nu)\perp\!\!\!\perp Z\sim P}[\|Y-Z\|_2 \min(2\kappa, \sqrt{d}\kappa\|X-Y\|_2)] \\ &\leq \kappa(d+L)W_1(\mu,\nu)+d^{1/4}\kappa L\sqrt{2M_{\nu,P}W_1(\mu,\nu)}.\end{aligned}$$

Proof of Theorem 2 The result follows directly from Lemma 3, Lemma 2, and Theorem 1. ■

5. A Finite-particle Convergence Rate for SVGD

To establish our main SVGD convergence result, we combine Theorems 1 and 2 with the following descent lemma for continuous SVGD error due to Salim et al. [17] which shows that continuous SVGD decreases the KL divergence to P and drives the KSD to P to zero.

Lemma 4 (Continuous SVGD descent lemma [17, Thm. 3.2]) *Suppose Assumptions 1, 2, and 4 hold, and consider the outputs $\mu_r^\infty = \mathrm{SVGD}(\mu_0^\infty, r)$ and $\mu_{r+1}^\infty = \mathrm{SVGD}(\mu_0^\infty, r+1)$ of Algorithm 1 with $\mu_0^\infty \ll P$. If $\max_{0\leq s\leq r}\epsilon_s \leq R_{\alpha,2}$ for some $\alpha > 1$ and*

$$R_{\alpha,p} \triangleq \min\Big(\tfrac{p}{\kappa^2(L+\alpha^2)}, (\alpha-1)(1+L\mathbb{E}_{\mu_0^\infty}[\|\cdot - x^\star\|_2] + 2L\sqrt{2\mathrm{KL}(\mu_0^\infty\|P)/\lambda})\Big) \text{ for } p\in\{1,2\},$$

then

$$\mathrm{KL}(\mu_{r+1}^\infty\|P) - \mathrm{KL}(\mu_r^\infty\|P) \leq -\epsilon_r\Big(1 - \tfrac{\kappa^2(L+\alpha^2)}{2}\epsilon_r\Big)\mathrm{KSD}_P(\mu_r^\infty, P)^2. \tag{1}$$

By summing the result (1) over $r = 0, \ldots, t$, we obtain the following corollary.

Corollary 1 *Under the assumptions and notation of Lemma 4, suppose $\max_{0\leq r\leq t}\epsilon_r \leq R_{\alpha,1}$ for some $\alpha > 1$, and let $\pi_r \triangleq \frac{c(\epsilon_r)}{\sum_{r=0}^t c(\epsilon_r)}$ for $c(\epsilon) \triangleq \epsilon\Big(1 - \frac{\kappa^2(L+\alpha^2)}{2}\epsilon\Big)$. Since $\frac{\epsilon}{2} \leq c(\epsilon) < \epsilon$, we have*

$$\textstyle\sum_{r=0}^t \pi_r \mathrm{KSD}_P(\mu_r^\infty, P)^2 \leq \frac{1}{\sum_{r=0}^t c(\epsilon_r)}\mathrm{KL}(\mu_0^\infty\|P) \leq \frac{2}{\sum_{r=0}^t \epsilon_r}\mathrm{KL}(\mu_0^\infty\|P).$$

Finally, we arrive at our main result that bounds the approximation error of n-particle SVGD in terms of the chosen step size sequence and the initial discretization error $W_1(\mu_0^n, \mu_0^\infty)$.

Theorem 3 (KSD error of finite-particle SVGD) *Suppose Assumptions 1, 2, 3, and 4 hold, fix any $\mu_0^\infty \ll P$ and $\mu_0^n \in \mathcal{P}$, and let $w_{0,n} \triangleq W_1(\mu_0^n, \mu_0^\infty)$. If $\max_{0\leq r<t}\epsilon_r \leq \epsilon_t \triangleq R_{\alpha,1}$ for some $\alpha > 1$ and $R_{\alpha,1}$ defined in Lemma 4, then the Algorithm 1 outputs $\mu_r^n = \mathrm{SVGD}(\mu_0^n, r)$ satisfy*

$$\min_{0\leq r\leq t}\mathrm{KSD}_P(\mu_r^n, P) \leq \textstyle\sum_{r=0}^t \pi_r \mathrm{KSD}_P(\mu_r^n, P) \leq a_{t-1} + \sqrt{\frac{2}{R_{\alpha,1}+b_{t-1}}\mathrm{KL}(\mu_0^\infty\|P)}, \tag{2}$$

for π_r as defined in Lemma 4, (A, B, C) as defined in Theorem 1, $b_{t-1} \triangleq \sum_{r=0}^{t-1}\epsilon_r$, and

$$\begin{aligned}a_{t-1} \triangleq\ & \kappa(d+L)w_{0,n}\exp(b_{t-1}(A + B\exp(Cb_t))) \\ &+ d^{1/4}\kappa L\sqrt{2M_{\mu_0^\infty,P}w_{0,n}}\exp(b_{t-1}(2C + A + B\exp(Cb_{t-1}))/2).\end{aligned} \tag{3}$$

Proof By the triangle inequality and Theorem 2 we have

$$|\mathrm{KSD}_P(\mu_r^n, P) - \mathrm{KSD}_P(\mu_r^\infty, P)| \le \mathrm{KSD}_P(\mu_r^n, \mu_r^\infty) \le a_{r-1}$$

for each r. Therefore

$$\textstyle\sum_{r=0}^t \pi_r(\mathrm{KSD}_P(\mu_r^n, P) - a_{r-1})^2 \le \sum_{r=0}^t \pi_r \mathrm{KSD}_P(\mu_r^\infty, P)^2 \le \frac{2}{R_{\alpha,1}+b_{t-1}}\mathrm{KL}(Q_0^\infty \| P), \quad (4)$$

where the last inequality follows from Corollary 1. Moreover, by Jensen's inequality,

$$\textstyle\sum_{r=0}^t \pi_r(\mathrm{KSD}_P(\mu_r^n, P) - a_{r-1})^2 \ge \left(\sum_{r=0}^t \pi_r \mathrm{KSD}_P(\mu_r^n, P) - \sum_{r=0}^t \pi_r a_{r-1}\right)^2. \quad (5)$$

Combining (4) and (5), we have

$$\textstyle\sum_{r=0}^t \pi_r \mathrm{KSD}_P(\mu_r^n, P) \le \sum_{r=0}^t \pi_r a_{r-1} + \sqrt{\frac{2}{R_{\alpha,1}+b_{t-1}}\mathrm{KL}(Q_0^\infty \| P)}.$$

We finish the proof by noticing that $\sum_{r=0}^t \pi_r a_{r-1} \le \max_{0\le r\le t} a_{r-1} = a_{t-1}$. ■

The following corollary, proved in Appendix D, provides an explicit SVGD convergence bound and rate by choosing the step size sum to balance the terms on the right-hand side of (2).

Corollary 2 (A finite-particle convergence rate for SVGD) *Instantiate the notation and assumptions of Theorem 3, let $(\bar{w}_{0,n}, \bar{A}, \bar{B}, \bar{C})$ be any upper bounds on $(w_{0,n}, A, B, C)$ respectively, and define the growth functions*

$$\phi(w) \triangleq \log\log(e^e + \tfrac{1}{w}) \quad \textit{and} \quad \psi_{\bar{B},\bar{C}}(x, y, \beta) \triangleq \tfrac{1}{\bar{C}}\log(\tfrac{1}{\bar{B}}\max(\bar{B}, \tfrac{1}{\beta}\log\tfrac{1}{x} - y)).$$

If the step size sum

$$b_{t-1} = \textstyle\sum_{r=0}^{t-1}\epsilon_r = \min\left(\psi_{\bar{B},\bar{C}}\big(\bar{w}_{0,n}\sqrt{\phi(\bar{w}_{0,n})}, \bar{A}, \beta_1\big), \psi_{\bar{B},\bar{C}}\big(\bar{w}_{0,n}\,\phi(\bar{w}_{0,n}), \bar{A}+2\bar{C}, \beta_2\big)\right),$$
$$\textit{for} \quad \beta_1 \triangleq \max\big(1, \psi_{\bar{B},\bar{C}}\big(\bar{w}_{0,n}\sqrt{\phi(\bar{w}_{0,n})}, \bar{A}, 1\big)\big) \quad \textit{and}$$
$$\beta_2 \triangleq \max\big(1, \psi_{\bar{B},\bar{C}}\big(\bar{w}_{0,n}\,\phi(\bar{w}_{0,n}), \bar{A}+2\bar{C}, 1\big)\big)$$

then

$$\min_{0\le r\le t} \mathrm{KSD}_P(\mu_r^n, P)$$
$$\le \begin{cases} \kappa(d+L)\bar{w}_{0,n} + d^{1/4}\kappa L\sqrt{2M_{\mu_0^\infty,P}\bar{w}_{0,n}} + \sqrt{\frac{2}{R_{\alpha,1}}\mathrm{KL}(\mu_0^\infty\|P)} & \textit{if } b_{t-1}=0 \\ \frac{\kappa(d+L)+d^{1/4}\kappa L\sqrt{2M_{\mu_0^\infty,P}}}{\sqrt{\phi(\bar{w}_{0,n})}} + \sqrt{\frac{2\mathrm{KL}(\mu_0^\infty\|P)}{R_{\alpha,1}+\frac{1}{\bar{C}}\log(\frac{1}{\bar{B}}(\frac{\log(1/(\bar{w}_{0,n}\,\phi(\bar{w}_{0,n})))}{\max(1,\psi_{\bar{B},\bar{C}}(\bar{w}_{0,n},0,1))}-\bar{A}-2\bar{C}))}} & \textit{otherwise.} \end{cases} \quad (6)$$

$$= O\left(\frac{1}{\sqrt{\log\log(e^e+\frac{1}{\bar{w}_{0,n}})}}\right). \quad (7)$$

If, in addition, $\mu_0^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ for $x_i \overset{i.i.d.}{\sim} \mu_0^\infty$ with $M_{\mu_0^\infty} \triangleq \mathbb{E}_{\mu_0^\infty}[\|\cdot\|_2^2] < \infty$, then

$$\bar{w}_{0,n} \triangleq \frac{M_{\mu_0^\infty}\log(n)^{\mathbb{I}[d=2]}}{\delta\, n^{1/(2\vee d)}} \ge w_{0,n} \quad (8)$$

with probability at least $1 - c\delta$ for a universal constant $c > 0$. Hence, with this choice of $\bar{w}_{0,n}$,

$$\min_{0\le r\le t} \mathrm{KSD}_P(\mu_r^n, P) = O\left(\frac{1}{\sqrt{\log\log(n\delta)}}\right)$$

with probability at least $1 - c\delta$.

References

[1] Alessandro Barp, Carl-Johann Simon-Gabriel, Mark Girolami, and Lester Mackey. Targeted separation and convergence with kernel discrepancies. *arXiv preprint arXiv:2209.12835*, 2022.

[2] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[3] Wilson Ye Chen, Lester Mackey, Jackson Gorham, François-Xavier Briol, and Chris Oates. Stein points. In *International Conference on Machine Learning*, pages 844–853, 2018.

[4] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pages 2606–2615, 2016.

[5] Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of Stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

[6] Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015.

[7] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301, 2017.

[8] Jackson Gorham, Anant Raj, and Lester Mackey. Stochastic Stein discrepancies. *Advances in Neural Information Processing Systems*, 33:17931–17942, 2020.

[9] 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, 2017.

[10] Jonathan Huggins and Lester Mackey. Random feature Stein discrepancies. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 1903–1913. 2018.

[11] 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.

[12] Jing Lei. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020.

[13] Qiang Liu. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems*, pages 3115–3123, 2017.

[14] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29:2378–2386, 2016.

[15] 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, 2016.

[16] Yang Liu, Prajit Ramachandran, Qiang Liu, and Jian Peng. Stein variational policy gradient. In *33rd Conference on Uncertainty in Artificial Intelligence*, 2017.

[17] Adil Salim, Lukang Sun, and Peter 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.

[18] Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45, 2014.

[19] Lukang Sun, Avetik Karagulyan, and Peter Richtarik. Convergence of Stein variational gradient descent under a weaker smoothness condition. *arXiv preprint arXiv:2206.00508*, 2022.

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

[21] Dilin Wang and Qiang Liu. Learning to draw samples: With application to amortized MLE for generative adversarial learning. *arXiv preprint arXiv:1611.01722*, 2016.

[22] Dilin Wang, Zhe Zeng, and Qiang Liu. Stein variational message passing for continuous graphical models. In *International Conference on Machine Learning*, pages 5219–5227, 2018.

[23] Jingwei Zhuo, Chang Liu, Jiaxin Shi, Jun Zhu, Ning Chen, and Bo Zhang. Message passing Stein variational gradient descent. In *International Conference on Machine Learning*, pages 6018–6027, 2018.

Appendix A. Proof of Theorem 1: Wasserstein discretization error of SVGD

We first show that the pseudo-Lipschitzness conditions of Lemma 1 hold given our assumptions. According to Assumption 2, we have

$$\nabla_{z_i} k(x,z) = \langle \nabla_{z_i} k(z,\cdot), k(x,\cdot)\rangle_{\mathcal{H}} \le (\nabla_{z_i}\nabla_{z_i'} k(z,z')|_{z'=z} k(x,x))^{1/2} \le \kappa^2,$$
$$\|\nabla_z k(x,z)\|_2 = \sqrt{\textstyle\sum_{i=1}^d (\nabla_{z_i} k(x,z))^2} \le \sqrt{d}\kappa^2.$$
$$\nabla_{z_j}\nabla_{x_i} k(x,z) = \langle \nabla_{z_j} k(z,\cdot), \nabla_{x_i} k(x,\cdot)\rangle_{\mathcal{H}} \le (\nabla_{z_j}\nabla_{z_j'} k(z,z')|_{z'=z}\nabla_{x_i}\nabla_{x_i'} k(x,x')|_{x'=x})^{1/2} \le \kappa^2,$$
$$\|\nabla_z\nabla_x k(x,z)\|_{\mathrm{op}} \le \|\nabla_z\nabla_x k(x,z)\|_F = \sqrt{\textstyle\sum_{i=1}^d\sum_{j=1}^d (\nabla_{z_j}\nabla_{x_i} k(x,z))^2} \le d\kappa^2,$$

Recall that $x^\star$ satisfies $s_p(x^\star) = 0$ by Assumption 1. Then, by the triangle inequality, the definition of $\|\cdot\|_{\mathrm{op}}$, and Cauchy-Schwartz,

$$\begin{aligned}
&\|\nabla_z(s_p(x)k(x,z) + \nabla_x k(x,z))\|_{\mathrm{op}} \\
&= \left\|(s_p(x) - s_p(x^\star))\nabla_z k(x,z)^\top\right\|_{\mathrm{op}} + \|\nabla_z\nabla_x k(x,z)\|_{\mathrm{op}} \\
&\le \sup_{\|u\|_2\le 1}(\|s_p(x) - s_p(x^\star)\|_2 |\nabla_z k(x,z)^\top u|) + d\kappa^2 \\
&\le L\|x - x^\star\|_2\|\nabla_z k(x,z)\|_2 + d\kappa^2 \\
&\le \sqrt{d}\kappa^2 L(\|x\|_2 + \|x^\star\|_2) + d\kappa^2 \\
&\le \max(\sqrt{d}\kappa^2 L, \sqrt{d}\kappa^2 L\|x^\star\|_2 + d\kappa^2)(1 + \|x\|_2).
\end{aligned}$$

Letting $c_1 = \max(\sqrt{d}\kappa^2 L, \sqrt{d}\kappa^2 L\|x^\star\|_2 + d\kappa^2)$ and taking supremum over z proves the first pseudo-Lipschitzness condition. Similarly, we have

$$\begin{aligned}
&\|\nabla_x(s_p(x)k(x,z) + \nabla_x k(x,z))\|_{\mathrm{op}} \\
&= \|\nabla s_p(x)\|_{\mathrm{op}} k(x,z) + \left\|(s_p(x) - s_p(x^\star))\nabla_x k(x,z)^\top\right\|_{\mathrm{op}} + \left\|\nabla_x^2 k(x,z)\right\|_{\mathrm{op}} \\
&\le \kappa^2 L + L\|x - x^\star\|_2\|\nabla_x k(x,z)\|_2 + d\kappa^2,
\end{aligned} \tag{9}$$

where we used the Lipschitzness of s_p from Assumption 1 and

$$k(x,z) = \langle k(x,\cdot), k(z,\cdot)\rangle_{\mathcal{H}} \le \sqrt{k(x,x)k(z,z)} \le \kappa^2,$$
$$\left\|\nabla_x^2 k(x,z)\right\|_{\mathrm{op}} \le \left\|\nabla_x^2 k(x,z)\right\|_F = \sqrt{\textstyle\sum_{|I|=2}(D_x^I k(x,z))^2} \le d\kappa^2.$$

The last inequality is due to

$$|D_x^I k(x,z)| = |\langle D_x^I k(x,\cdot), k(z,\cdot)\rangle_{\mathcal{H}}| \le \sqrt{k(z,z) D_x^I D_{x'}^I k(x,x')|_{x'=x}} \le \kappa^2.$$

According to Assumption 3, there exists $\gamma > 0$ such that $\|\nabla_x k(x,z)\|_2 \le \gamma\|x - z\|_2^{-1}$. Therefore, for $\|x - z\|_2 \ge 1$,

$$\begin{aligned}
\|x - x^\star\|_2\|\nabla_x k(x,z)\|_2 &\le \gamma\|x - x^\star\|_2/\|x - z\|_2 \\
&\le \gamma(1 + \|z - x^\star\|_2/\|x - z\|_2) \\
&\le \gamma(1 + \|z - x^\star\|_2).
\end{aligned} \tag{10}$$

And for $\|x - z\|_2 < 1$,

$$\begin{aligned}\|x - x^\star\|_2 \|\nabla_x k(x, z)\|_2 &\leq \sqrt{d}\kappa^2(\|x - z\|_2 + \|z - x^\star\|_2) \\ &\leq \sqrt{d}\kappa^2(1 + \|z - x^\star\|_2). \end{aligned} \tag{11}$$

Combining (10) and (11),

$$\begin{aligned}\|x - x^\star\|_2 \|\nabla_x k(x, z)\|_2 &\leq (\gamma + \sqrt{d}\kappa^2)(1 + \|z - x^\star\|_2) \\ &\leq (\gamma + \sqrt{d}\kappa^2)(1 + \|x^\star\|_2 + \|z\|_2). \end{aligned} \tag{12}$$

Plugging (12) back into (9), we can show the second pseudo-Lipschitzness condition holds for $c_2 = \kappa^2 L + d\kappa^2 + L(\gamma + \sqrt{d}\kappa^2)(1 + \|x^\star\|_2)$.

Now that the pseudo-Lipschitzness conditions hold, by repeated application of Lemma 1 and the inequality $(1 + x) \leq e^x$, we have

$$\begin{aligned}W_1(\mu_{r+1}^n, \mu_{r+1}^\infty) = W_1(\Phi_{\epsilon_r}(\mu_r^n), \Phi_{\epsilon_r}(\mu_r^\infty)) &\leq (1 + \epsilon_r D_r) W(\mu_r^n, \mu_r^\infty) \\ &\leq W_1(\mu_0^n, \mu_0^\infty) \prod_{s=0}^r (1 + \epsilon_s D_s) \leq W_1(\mu_0^n, \mu_0^\infty) \exp(\sum_{s=0}^r \epsilon_s D_s) \end{aligned} \tag{13}$$

for $D_s = c_1(1 + m_{\mu_s^n}) + c_2(1 + m_{\mu_s^\infty})$.

Using the result from Lemma 2, we have

$$D_{s+1} \leq A + B \exp(C b_s)$$

for $A = (c_1 + c_2)(1 + m_P)$, $B = c_1 m_{\mu_0^n, P} + c_2 m_{\mu_0^\infty, P}$, and $C = \kappa^2(3L + d)$. Therefore

$$\begin{aligned}\textstyle\sum_{s=0}^r \epsilon_s D_s &\leq \max_{0 \leq s \leq r} D_s \textstyle\sum_{s=0}^r \epsilon_s \\ &\leq b_r(A + B \exp(C b_{r-1})) \\ &\leq b_r(A + B \exp(C b_r)). \end{aligned}$$

Plugging this back into (13) proves the result.

Appendix B. Proof of Lemma 2: SVGD moment growth

From Assumption 2 we know

$$\begin{aligned}&|k(y, x)| \leq |\langle k(y, \cdot), k(x, \cdot)\rangle_{\mathcal{H}}| \leq \|k(y, \cdot)\|_{\mathcal{H}} \|k(x, \cdot)\|_{\mathcal{H}} = \sqrt{k(y, y)k(x, x)} \leq \kappa^2, \\ &|D_y^I k(y, x)| = |\langle D_y^I k(y, \cdot), k(x, \cdot)\rangle_{\mathcal{H}}| \leq \sqrt{k(x, x) D_y^I D_{y'}^I k(y, y')|_{y'=y}} \leq \kappa^2, \\ &\left\|\nabla_y^2 k(y, x)\right\|_{op} \leq \left\|\nabla_y^2 k(y, x)\right\|_F = \sqrt{\textstyle\sum_{|I|=2} (D_y^I k(y, x))^2} \leq d\kappa^2. \end{aligned}$$

The last inequality implies

$$\|\nabla_y k(y, x) - \nabla_z k(z, x)\|_2 \leq d\kappa^2 \|y - z\|_2.$$

Let μ be any probability measure. Using the above results, Jensen's inequality and the facts that $\mathbb{E}_P[s_p(\cdot)] = 0$ and $\mathbb{E}_{Z\sim P}[(\mathcal{A}_P k(\cdot, x))(Z)] = 0$, we have

$$\begin{aligned}
&\|T_{\mu,\epsilon}(x) - x\|_2 \\
&\leq \epsilon \|\mathbb{E}_{X\sim\mu}[(\mathcal{A}_P k(\cdot, x))(X)]\|_2 \\
&= \epsilon \|\mathbb{E}_{X\sim\mu}[(\mathcal{A}_P k(\cdot, x))(X)] - \mathbb{E}_{Z\sim P}[(\mathcal{A}_P k(\cdot, x))(Z)]\|_2 \\
&= \epsilon \|\mathbb{E}_{X\sim\mu \perp\!\!\!\perp Z\sim P}[k(Z, x)(s_p(X) - s_p(Z)) + (k(X, x) - k(Z, x))(s_p(X) - \mathbb{E}_P[s_p(\cdot)]) \\
&\quad + (\nabla_X k(X, x) - \nabla_Z k(Z, x))]\|_2 \\
&\leq \epsilon \mathbb{E}_{X\sim\mu \perp\!\!\!\perp Z\sim P}[|k(Z, x)| \|s_p(X) - s_p(Z)\|_2 + (|k(X, x)| + |k(Z, x)|)\|s_p(X) - \mathbb{E}_P[s_p(\cdot)]\|_2 \\
&\quad + \|\nabla_X k(X, x) - \nabla_Z k(Z, x)\|_2] \\
&\leq \epsilon \mathbb{E}_{X\sim\mu \perp\!\!\!\perp Z\sim P}[\kappa^2 L \|X - Z\|_2 + 2\kappa^2 \|s_p(X) - \mathbb{E}_{Y\sim P}[s_p(Y)]\|_2 + d\kappa^2 \|X - Z\|_2] \\
&\leq \epsilon \mathbb{E}_{X\sim\mu \perp\!\!\!\perp Z\sim P}[\kappa^2 (L + d)\|X - Z\|_2] + \epsilon \cdot 2\kappa^2 L \mathbb{E}_{X\sim\mu \perp\!\!\!\perp Y\sim P}[\|X - Y\|_2] \\
&= \epsilon\kappa^2 (3L + d)\mathbb{E}_{X\sim\mu \perp\!\!\!\perp Z\sim P}[\|X - Z\|_2] \\
&= \epsilon C m_{\mu,P}.
\end{aligned} \tag{14}$$

The last step used the definitions $m_{\mu,P} \triangleq \mathbb{E}_{X\sim\mu \perp\!\!\!\perp Z\sim P}[\|X - Z\|_2]$ and $C = \kappa^2(3L + d)$. Then, applying triangle inequality and (14), we have

$$\begin{aligned}
m_{\mu_{r+1},P} &= \mathbb{E}_{X\sim\mu_{r+1} \perp\!\!\!\perp Z\sim P}[\|X - Z\|_2] \\
&= \mathbb{E}_{X\sim\mu_r \perp\!\!\!\perp Z\sim P}[\|T_{\mu_r,\epsilon_r}(X) - Z\|_2] \\
&\leq \mathbb{E}_{X\sim\mu_r \perp\!\!\!\perp Z\sim P}[\|T_{\mu_r,\epsilon_r}(X) - X\|_2 + \|X - Z\|_2] \\
&\leq \epsilon_r C \mathbb{E}_{X\sim\mu_r \perp\!\!\!\perp Z\sim P}[\|X - Z\|_2] + \mathbb{E}_{X\sim\mu_r \perp\!\!\!\perp Z\sim P}[\|X - Z\|_2] \\
&= (1 + \epsilon_r C) m_{\mu_r,P},
\end{aligned} \tag{15}$$

$$\begin{aligned}
M_{\mu_{r+1},P} &= \mathbb{E}_{X\sim\mu_{r+1} \perp\!\!\!\perp Z\sim P}[\|X - Z\|_2^2] \\
&= \mathbb{E}_{X\sim\mu_r \perp\!\!\!\perp Z\sim P}[\|T_{\mu_r,\epsilon_r}(X) - Z\|_2^2] \\
&\leq \mathbb{E}_{X\sim\mu_r \perp\!\!\!\perp Z\sim P}[\|T_{\mu_r,\epsilon_r}(X) - X\|_2^2 + 2\|T_{\mu_r,\epsilon_r}(X) - X\|_2 \|X - Z\|_2 + \|X - Z\|_2^2] \\
&\leq (\epsilon_r^2 C^2 + 2\epsilon_r C) m_{\mu_r,P}^2 + M_{\mu_r,P} \\
&\leq (1 + 2\epsilon_r C + \epsilon_r^2 C^2) M_{\mu_r,P} \\
&= (1 + \epsilon_r C)^2 M_{\mu_r,P},
\end{aligned} \tag{16}$$

where the second last step used Jensen's inequality $m_{\mu_r,P}^2 \leq M_{\mu_r,P}$. Then, we repeatedly apply (15) and (16) together with the inequality $1 + x \leq e^x$ to get

$$\begin{aligned}
M_{\mu_r,P} &\leq M_{\mu_0,P} \textstyle\prod_{s=0}^{r-1}(1 + \epsilon_s C)^2 \leq M_{\mu_0,P} \exp(2C \textstyle\sum_{s=0}^{r-1} \epsilon_s) \leq M_{\mu_0,P} \exp(2C b_{r-1}), \\
m_{\mu_r} &= \mathbb{E}_{\mu_r}[\|X\|_2] \leq m_{\mu_r,P} + \mathbb{E}_P[\|Z\|_2] \leq m_{\mu_0,P} \textstyle\prod_{s=0}^{r-1}(1 + \epsilon_s C) + m_P \\
&\leq m_{\mu_0,P} \exp(C b_{r-1}) + m_P.
\end{aligned}$$

Appendix C. Proof of Lemma 3: KSD-Wasserstein bound

Our proof generalizes that of Gorham and Mackey [7, Lem. 18]. Consider any $g \in \mathcal{H}^d$ satisfying $\|g\|_{\mathcal{H}^d}^2 \triangleq \sum_{i=1}^d \|g_i\|_{\mathcal{H}}^2 \leq 1$. From Assumption 2 we know

$$\|g(x)\|_2^2 \leq \sum_{i=1}^d \kappa^2 \|g_i\|_{\mathcal{H}}^2 \leq \kappa^2, \tag{17}$$

$$\|\nabla g(x)\|_{op}^2 \leq \|\nabla g(x)\|_F^2 = \sum_{i=1}^d \sum_{j=1}^d |\nabla_{x_i} g_j(x)|^2 \leq \sum_{i=1}^d \sum_{j=1}^d \kappa^2 \|g_j\|_{\mathcal{H}}^2 \leq d\kappa^2, \text{ and} \tag{18}$$

$$\|\nabla(\nabla \cdot g(x))\|_2^2 = \sum_{i=1}^d \Big(\sum_{j=1}^d \nabla_{x_i} \nabla_{x_j} g_j(x)\Big)^2 \leq d \sum_{i=1}^d \sum_{j=1}^d |\nabla_{x_i} \nabla_{x_j} g_j(x)|^2 \leq d^2\kappa^2.$$

Suppose X, Y, Z are distributed so that (X, Y) is a 1-Wasserstein optimal coupling of (μ, ν) and Z is independent of (X, Y). Since s_p is L-Lipschitz with $\mathbb{E}_P[s_p] = 0$ (Assumption 1), g is bounded (17), and g and $\nabla \cdot g$ are Lipschitz (18), repeated use of Cauchy-Schwarz gives

$$\begin{aligned}
&\mathbb{E}_\mu[\mathcal{T}_P g] - \mathbb{E}_\nu[\mathcal{T}_P g] \\
&= \mathbb{E}[\nabla \cdot g(X) - \nabla \cdot g(Y)] + \mathbb{E}[\langle s_p(X) - s_p(Y), g(X)\rangle] + \mathbb{E}[\langle s_p(Y) - s_p(Z), g(X) - g(Y)\rangle] \\
&\leq \kappa(d + L) W_1(\mu, \nu) + L\mathbb{E}[\|Y - Z\|_2 \min(2\kappa, \sqrt{d}\kappa\|X - Y\|_2)].
\end{aligned}$$

Since our choice of g was arbitrary, the first advertised result now follows from the definition of KSD (Definition 4). The second claim then follows from Cauchy-Schwarz and the inequality $\min(a, b)^2 \leq ab$ for $a, b \geq 0$, since

$$\begin{aligned}
&\mathbb{E}[\|Y - Z\|_2 \min(2\kappa, \sqrt{d}\kappa\|X - Y\|_2)] \leq M_{\nu,P}^{1/2}\, \mathbb{E}[\min(2\kappa, \sqrt{d}\kappa\|X - Y\|_2)^2]^{1/2} \\
&\leq \sqrt{2M_{\nu,P}} d^{1/4} \kappa \mathbb{E}[\|X - Y\|_2]^{1/2} = \sqrt{2M_{\nu,P} W_1(\mu, \nu)} d^{1/4} \kappa.
\end{aligned}$$

Appendix D. Proof of Corollary 2: A finite-particle convergence rate for SVGD

We begin by establishing a lower bound on b_{t-1}. Let

$$b_{t-1}^{(1)} = \psi_{\bar{B},\bar{C}}(\bar{w}_{0,n}\sqrt{\phi(\bar{w}_{0,n})}, \bar{A}, \beta_1) \quad \text{and} \quad b_{t-1}^{(2)} = \psi_{\bar{B},\bar{C}}(\bar{w}_{0,n}\,\phi(\bar{w}_{0,n}), \bar{A} + 2\bar{C}, \beta_2)$$

so that $b_{t-1} = \min(b_{t-1}^{(1)}, b_{t-1}^{(2)})$. Since $\beta_1, \beta_2, \phi(\bar{w}_{0,n}) \geq 1$, we have

$$\begin{aligned}
\beta_1 &= \max(1, \tfrac{1}{\bar{C}} \log(\tfrac{1}{\bar{B}}(\log \tfrac{1}{\bar{w}_{0,n}\sqrt{\phi(\bar{w}_{0,n})}} - \bar{A}))) \\
&\leq \max(1, \tfrac{1}{\bar{C}} \log(\tfrac{1}{\bar{B}}(\log \tfrac{1}{\bar{w}_{0,n}\sqrt{\phi(\bar{w}_{0,n})}}))) \\
&\leq \max(1, \tfrac{1}{\bar{C}} \log(\tfrac{1}{\bar{B}}(\log \tfrac{1}{\bar{w}_{0,n}}))) \quad \text{and} \\
\beta_2 &= \max(1, \tfrac{1}{\bar{C}} \log(\tfrac{1}{\bar{B}}(\log \tfrac{1}{\bar{w}_{0,n}\phi(\bar{w}_{0,n})} - \bar{A} - 2\bar{C}))) \\
&\leq \max(1, \tfrac{1}{\bar{C}} \log(\tfrac{1}{\bar{B}}(\log \tfrac{1}{\bar{w}_{0,n}\phi(\bar{w}_{0,n})}))) \\
&\leq \max(1, \tfrac{1}{\bar{C}} \log(\tfrac{1}{\bar{B}}(\log \tfrac{1}{\bar{w}_{0,n}}))).
\end{aligned}$$

Hence, $\phi(\bar{w}_{0,n}) \geq 1$ implies that

$$\begin{aligned} b_{t-1}^{(1)} &\geq \tfrac{1}{\bar{C}} \log\Big(\tfrac{1}{\bar{B}}\Big(\frac{\log \frac{1}{\bar{w}_{0,n}\sqrt{\phi(\bar{w}_{0,n})}}}{\max(1, \frac{1}{\bar{C}}\log(\frac{1}{\bar{B}}(\log \frac{1}{\bar{w}_{0,n}})))} - \bar{A}\Big)\Big) \\ &\geq \tfrac{1}{\bar{C}} \log\Big(\tfrac{1}{\bar{B}}\Big(\frac{\log \frac{1}{\bar{w}_{0,n}\,\phi(\bar{w}_{0,n})}}{\max(1, \frac{1}{\bar{C}}\log(\frac{1}{\bar{B}}(\log \frac{1}{\bar{w}_{0,n}})))} - \bar{A} - 2\bar{C}\Big)\Big) \quad \text{and} \\ b_{t-1}^{(2)} &\geq \tfrac{1}{\bar{C}} \log\Big(\tfrac{1}{\bar{B}}\Big(\frac{\log \frac{1}{\bar{w}_{0,n}\,\phi(\bar{w}_{0,n})}}{\max(1, \frac{1}{\bar{C}}\log(\frac{1}{\bar{B}}(\log \frac{1}{\bar{w}_{0,n}})))} - \bar{A} - 2\bar{C}\Big)\Big). \end{aligned} \tag{19}$$

We divide the remainder of our proof into four parts. First we prove each of the two cases in the generic KSD bound (6) in Appendices D.1 and D.2. Next we show in Appendix D.3 that these two cases yield the generic convergence rate (7). Finally, we prove the high probability upper estimate (8) for $w_{0,n}$ under i.i.d. initialization in Appendix D.4.

D.1. Case $b_{t-1} = 0$

In this case, the error bound (6) follows directly from Theorem 3.

D.2. Case $b_{t-1} > 0$

We first state and prove a useful lemma.

Lemma 5 *Suppose $x = f(\beta)$ for a non-increasing function $f : \mathbb{R} \to \mathbb{R}$ and $\beta = \max(1, f(1))$. Then $x \leq f(x)$.*

Proof If $\beta = 1 \geq f(1)$, then $x = f(\beta) = f(1) \leq 1$. Because f is non-increasing, $f(x) \geq f(1) = x$. Otherwise, $\beta = f(1) > 1$, and hence $x = f(\beta) \leq f(1) = \beta$ since f is non-increasing. Since $x \leq \beta$ and f is non-increasing, we further have $f(x) \geq f(\beta) = x$ as advertised. ■

Since $\psi_{\bar{B},\bar{C}}$ is non-increasing in its third argument, Lemma 5 implies that

$$b_{t-1}^{(1)} \leq \psi_{\bar{B},\bar{C}}(\bar{w}_{0,n}\sqrt{\phi(\bar{w}_{0,n})}, \bar{A}, b_{t-1}^{(1)}) \leq \psi_{\bar{B},\bar{C}}(\bar{w}_{0,n}\sqrt{\phi(\bar{w}_{0,n})}, \bar{A}, 1) = \beta_1.$$

Rearranging the terms and noting that

$$\bar{B} < \tfrac{1}{\beta_1}\log\frac{1}{\bar{w}_{0,n}\sqrt{\phi(\bar{w}_{0,n})}} - \bar{A} \leq \tfrac{1}{b_{t-1}^{(1)}}\log\frac{1}{\bar{w}_{0,n}\sqrt{\phi(\bar{w}_{0,n})}} - \bar{A}$$

since $b_{t-1}^{(1)} \geq b_{t-1} > 0$, we have

$$\bar{w}_{0,n}\exp(b_{t-1}^{(1)}(\bar{A} + \bar{B}\exp(\bar{C}b_{t-1}^{(1)}))) \leq \frac{1}{\sqrt{\phi(\bar{w}_{0,n})}}. \tag{20}$$

Similarly, we have $b_{t-1}^{(2)} \leq \psi_{\bar{B},\bar{C}}(\bar{w}_{0,n}\log\log\frac{1}{\bar{w}_{0,n}}, \bar{A} + 2\bar{C}, b_{t-1}^{(2)})$ and

$$\sqrt{\bar{w}_{0,n}}\exp(b_{t-1}^{(2)}(2\bar{C} + \bar{A} + \bar{B}\exp(\bar{C}b_{t-1}^{(2)}))/2) \leq \frac{1}{\sqrt{\phi(\bar{w}_{0,n})}}. \tag{21}$$

Since $b_t = \min(b_{t-1}^{(1)}, b_{t-1}^{(2)})$, the inequalities (20) and (21) are also satisfied when b_t is substituted for $b_{t-1}^{(1)}$ and $b_{t-1}^{(2)}$. Since the error term a_t (3) is non-decreasing in each of $(w_{0,n}, A, B, C)$, we have

$$a_t \leq (\kappa(d+L) + d^{1/4}\kappa L\sqrt{2M_{\mu_0^\infty,P}})/\sqrt{\phi(\bar{w}_{0,n})}.$$

Since $b_{t-1} = \min(b_{t-1}^{(1)}, b_{t-1}^{(2)})$, the claim (6) follows from this estimate, the lower bounds (19), and Theorem 3.

D.3. Generic convergence rate

The generic convergence rate (7) holds as, by the lower bounds (19), $b_{t-1} = \min(b_{t-1}^{(1)}, b_{t-1}^{(2)}) > 0$ whenever

$$e^{-(\bar{B}e+\bar{A}+2\bar{C})} > \bar{w}_{0,n}\phi(\bar{w}_{0,n}) \quad \text{and} \quad \bar{B}^{(\bar{B}e+\bar{A}+2\bar{C})/\bar{C}} > \bar{w}_{0,n}\phi(\bar{w}_{0,n})\log(1/\bar{w}_{0,n})^{(\bar{B}e+\bar{A}+2\bar{C})/\bar{C}},$$

a condition which occurs whenever $\bar{w}_{0,n}$ is sufficiently small since the right-hand side of each inequality converges to zero as $\bar{w}_{0,n} \to 0$.

D.4. Initializing with i.i.d. particles

We begin by restating an expected Wasserstein bound due to Lei [12].

Lemma 6 (Lei [12, Thm. 3.1]) *Suppose $\mu_0^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ for $x_i \overset{i.i.d.}{\sim} \mu_0^\infty$ with $M_{\mu_0^\infty} \triangleq \mathbb{E}_{\mu_0^\infty}[\|\cdot\|_2^2] < \infty$. Then, for a universal constant $c > 0$,*

$$\mathbb{E}[W_1(\mu_0^n, \mu_0^\infty)] \leq cM_{\mu_0^\infty}\frac{\log(n)^{\mathbb{I}[d=2]}}{n^{1/(2\vee d)}}.$$

Together, Lemma 6 and Markov's inequality imply that

$$W_1(\mu_0^n, \mu_0^\infty) \leq \mathbb{E}[W_1(\mu_0^n, \mu_0^\infty)]/(c\delta) \leq M_{\mu_0^\infty}\frac{\log(n)^{\mathbb{I}[d=2]}}{n^{1/(2\vee d)}}/\delta$$

with probability at least $1 - c\delta$, proving the high probability upper estimate (8).