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 satisfies Talagrand's T_1 inequality, 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, 11] 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 [11, 18, 19], generative modeling [17], and reinforcement learning [8, 13].

Algorithm 1 Stein Variational Gradient Descent [11]: SVGD(μ_0)

Input: Target P with density p, approximating distribution μ_0, kernel k, step sizes $\epsilon_{0:r-1}$
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 [10, 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, 2, 6, 12]. Follow-up work by Korba et al. [9], Salim et al. [14], Sun et al. [15] sharpened this result with path-independent constants, weaker smoothness conditions, and explicit rates of convergence. In addition, Duncan et al. [4] analyzed the continuous-time limit of continuous SVGD to provide conditions for exponential convergence. However, each of these analyses applies only to continuous SVGD and not to the finite-particle algorithm used in practice.

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

© J. Shi & L. Mackey.

To bridge this gap, Liu [10, 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. [7, 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.

In this work, we derive the first explicit error bounds and rates of convergence for finite-particle SVGD. To achieve this, we first bound the 1-Wasserstein error between each round of finite-particle and continuous SVGD in Theorem 1. We next bound the KSD in terms of the 1-Wasserstein distance 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 $\mathbb{R}^d$ 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 $\mathcal{H}^d \triangleq \bigotimes_{i=1}^d \mathcal{H}$ [1]. 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 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 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 $\sup_{x\in\mathbb{R}^d}\big(D_x^I D_y^I k(x, y)|_{y=x}\big) \leq \kappa^2$. 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*

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

In line with Salim et al. [14], we also assume the distribution P satisfies Talagrand's T_1 inequality [16, Def. 22.1]. We note that this assumption is only necessary for leveraging the continuous SVGD convergence result of Salim et al. [14]. In case it does not hold, one can alternatively resort to the path-dependent assumptions in Korba et al. [9], Liu [10] .

Assumption 4 (Talagrand's T_1 inequality [16, Def. 22.1]) *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 [5] *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 [12] *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 [11] *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 measures $\mu, \nu \in \mathcal{P}$ and target P satisfying Assumption 1 we define the* Langevin kernel Stein discrepancy [KSD, 2, 6, 12] *with base kernel k by*[1]

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

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.

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 outputs $\mu_r^n = \mathrm{SVGD}(\mu_0^n)$ and $\mu_r^\infty = \mathrm{SVGD}(\mu_0^\infty)$ of Algorithm 1 satisfy*

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

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 $c_2 = \kappa^2 L + d\kappa^2 + (\gamma + \sqrt{d}\kappa^2)(1 + \|x^\star\|_2)$.

The proof of Theorem 1 in Appendix A relies on two lemmas. The first, due to Gorham et al. [7], 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.

1. Prior work [2, 6, 12] defined $\mathrm{KSD}(\mu, \nu)$ for the special case $\nu = P$. Definition 4 extends this definition to general ν.

Lemma 1 (Wasserstein pseudo-Lipschitzness of SVGD [7, Lem. 12]) *For $P \in \mathcal{P}$ with differentiable density p, suppose that the following pseudo-Lipschitz bounds hold*

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

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

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

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*

$$m_{\mu_r} \le m_{\mu_0,P}\prod_{s=0}^{r-1}(1+\epsilon_s C) + m_P \le m_{\mu_0,P}\exp(Cb_{r-1}) + m_P,$$
$$M_{\mu_r,P} \le M_{\mu_0,P}\prod_{s=0}^{r-1}(1+2\epsilon_s C + \epsilon_s^2 C^2) \le M_{\mu_0,P}\exp((2C+\epsilon_0 C^2)b_{r-1}).$$

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 outputs $\mu_r^n = \mathrm{SVGD}(\mu_0^n)$ and $\mu_r^\infty = \mathrm{SVGD}(\mu_0^\infty)$ of Algorithm 1 satisfy*

$$\begin{aligned}\mathrm{KSD}(\mu_r^n, \mu_r^\infty) &\le \kappa(d+L)w_0\exp(b_{r-1}(A + B\exp(Cb_{r-1}))) \\ &\quad + d^{1/4}\kappa L\sqrt{2M_{\mu_0^\infty,P}w_0}\exp((2C+\epsilon_0 C^2 + A + B\exp(Cb_{r-1}))/2)b_{r-1})\end{aligned}$$

for $w_0 \triangleq W_1(\mu_0^n, \mu_0^\infty)$ and A, B, C defined as in Theorem 1.

Our proof of Theorem 2 relies on the following lemma, proved 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 $\mu, \nu \in \mathcal{P}$,*

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

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 continuous SVGD error bound due to Salim et al. [14] (see also Korba et al. [9], Liu [10]).

Lemma 4 (Continuous SVGD descent lemma [14, Thm. 3.2]) *Suppose Assumptions 1, 2, and 4 hold. For any constant $\alpha > 1$ and $\epsilon_r < \min\left(\frac{2}{\kappa^2(L+\alpha^2)}, R\right)$, where $R \triangleq (\alpha-1)(1+L\mathbb{E}_{\mu_0}[\|\cdot - x^\star\|_2] + 2L\sqrt{2\mathrm{KL}(\mu_0^\infty\|P)/\lambda})$, continuous SVGD decreases KL and drives KSD to zero:*

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

By summing (1) over $r = 0, \ldots, t$, we have the following corollary:

Corollary 1 *Let $c(\epsilon) \triangleq \epsilon\left(1 - \frac{\kappa^2(L+\alpha^2)}{2}\epsilon\right)$. If $\epsilon_r \leq \min\left(\frac{1}{\kappa^2(L+\alpha^2)}, R\right)$, then we have $\frac{\epsilon_r}{2} \leq c(\epsilon_r) < \epsilon_r$ and thereby*

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

Finally, we arrive at our main result that shows finite-particle SVGD with appropriately chosen step size sequences drives KSD to zero as $W_1(\mu_0^n, \mu_0^\infty) \to 0$.

Theorem 3 (KSD error of finite-particle SVGD) *Suppose Assumptions 1, 2, 3, and 4 hold. Define $\pi_r \triangleq \frac{c(\epsilon_r)}{\sum_{r=0}^t c(\epsilon_r)}$ and $w_0 \triangleq W_1(\mu_0^n, \mu_0^\infty)$. Then,*

$$\min_{0\leq r\leq t}\mathrm{KSD}(\mu_r^n, P) \leq \sum_{r=0}^t \pi_r\mathrm{KSD}(\mu_r^n, P) \leq a_t + \left(\frac{2}{b_t}\mathrm{KL}(\mu_0^\infty\|P)\right)^{1/2}, \tag{2}$$

where $b_t \triangleq \sum_{r=0}^t \epsilon_r$,

$$\begin{aligned} a_t \triangleq\ & \kappa(d+L)w_0\exp(b_t(A + B\exp(Cb_t))) \\ & + d^{1/4}\kappa L\sqrt{2M_{\mu_0^\infty,P}w_0}\exp(((2C + \epsilon_0 C^2 + A + B\exp(Cb_t))/2)b_t), \end{aligned}$$

and A, B, C are defined as in Theorem 1.

Proof By the triangle inequality and Theorem 2 we have

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

for each r. Therefore

$$\sum_{r=0}^t \pi_r(\mathrm{KSD}(\mu_r^n, P) - a_{r-1})^2 \leq \sum_{r=0}^t \pi_r\mathrm{KSD}(\mu_r^\infty, P)^2 \leq \frac{2}{b_t}\mathrm{KL}(Q_0^\infty\|P), \tag{3}$$

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

$$\sum_{r=0}^t \pi_r(\mathrm{KSD}(\mu_r^n, P) - a_{r-1})^2 \geq \left(\sum_{r=0}^t \pi_r\mathrm{KSD}(\mu_r^n, P) - \sum_{r=0}^t \pi_r a_{r-1}\right)^2 \tag{4}$$

Combining (3) and (4), we have

$$\sum_{r=0}^t \pi_r\mathrm{KSD}(\mu_r^n, P) \leq \sum_{r=0}^t \pi_r a_{r-1} + \left(\frac{2}{b_t}\mathrm{KL}(\mu_0^\infty\|P)\right)^{1/2}.$$

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

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

Corollary 2 (A finite-particle convergence rate for SVGD) *In the notation of Theorem 3, define the shorthand* $\psi_{B,C}(x,y,\beta) \triangleq \frac{1}{C}\log\Big(\frac{1}{B}(\frac{1}{\beta}\log\frac{1}{x}-y)\Big)$. *If* $\max_{0\le r\le t}\epsilon_r \le \min\Big(\frac{1}{\kappa^2(L+\alpha^2)}, R\Big)$ *and*

$$b_t = \textstyle\sum_{r=0}^t \epsilon_r = \min\Big(\psi_{B,C}\big(w_0\sqrt{\log\log\frac{1}{w_0}}, A, \beta_1\big), \psi_{B,C}\big(w_0\log\log\frac{1}{w_0}, A+2C+\epsilon_0 C^2, \beta_2\big)\Big),$$

$$\textit{for } \beta_1 \triangleq \max\Big(1, \psi_{B,C}\big(w_0\sqrt{\log\log\frac{1}{w_0}}, A, 1\big)\Big) \textit{ and}$$

$$\beta_2 \triangleq \max\Big(1, \psi_{B,C}\big(w_0\log\log\frac{1}{w_0}, A+2C+\epsilon_0 C^2, 1\big)\Big),$$

then,

$$\min_{0\le r\le t}\mathrm{KSD}(\mu_r^n, P) = O\Big(1/\sqrt{\log\log\frac{1}{w_0}}\Big).$$

If, in addition, $\mu_0^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ *for* $x_i \overset{i.i.d.}{\sim} \mu_0^\infty$, *then*

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

References

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

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

[3] Richard Mansfield Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

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

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

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

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

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] 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.

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

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

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

[18] 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.

[19] 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-Lipshitz conditions in Lemma 1 holds 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 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-Lipschitz condition. Similarly, we have

$$\begin{aligned}
&\|\nabla_x(s_p(x)k(x,z) + \nabla_x k(x,z))\|_{\mathrm{op}} \\
&= \|\nabla s_p(x)\|_{\mathrm{op}} k(x,z) + \left\|(s_p(x) - s_p(x^\star))\nabla_x k(x,z)^\top\right\|_{\mathrm{op}} + \left\|\nabla_x^2 k(x,z)\right\|_{\mathrm{op}} \\
&\le \kappa^2 L + L\|x - x^\star\|_2 \|\nabla_x k(x,z)\|_2 + d\kappa^2,
\end{aligned} \tag{5}$$

where we used

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

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 $\|x - z\|_2 \ge 1$,

$$\begin{aligned}
\|x - x^\star\|_2 \|\nabla_x k(x,z)\|_2 &\le \gamma\|x - x^\star\|_2 / \|x - z\|_2 \\
&\le \gamma(1 + \|z - x^\star\|_2 / \|x - z\|_2) \\
&\le \gamma(1 + \|z - x^\star\|_2).
\end{aligned} \tag{6}$$

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

$$\begin{aligned}\|x - x^\star\|_2 \|\nabla_x k(x, z)\|_2 &\leq \sqrt{d}\kappa^2(\|x - z\|_2 + \|z - x^\star\|_2) \\ &\leq \sqrt{d}\kappa^2(1 + \|z - x^\star\|_2). \end{aligned} \tag{7}$$

Combining (6) and (7),

$$\begin{aligned}\|x - x^\star\|_2 \|\nabla_x k(x, z)\|_2 &\leq (\gamma + \sqrt{d}\kappa^2)(1 + \|z - x^\star\|_2) \\ &\leq (\gamma + \sqrt{d}\kappa^2)(1 + \|x^\star\|_2 + \|z\|_2). \end{aligned} \tag{8}$$

Plugging (8) back into (5), we can show the second pseudo Lipschitzness condition holds for $c_2 = \kappa^2 L + d\kappa^2 + (\gamma + \sqrt{d}\kappa^2)(1 + \|x^\star\|_2)$.

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

$$\begin{aligned}W_1(\mu^n_{r+1}, \mu^\infty_{r+1}) &= W_1(\Phi_{\epsilon_r}(\mu^n_r), \Phi_{\epsilon_r}(\mu^\infty_r)) \leq (1 + \epsilon_r C_r) W(\mu^n_r, \mu^\infty_r) \\ &\leq W_1(\mu^n_0, \mu^\infty_0) \prod\nolimits_{s=0}^{r} (1 + \epsilon_s C_s) \leq W_1(\mu^n_0, \mu^\infty_0) \exp(\sum\nolimits_{s=0}^{r} \epsilon_s C_s) \end{aligned} \tag{9}$$

for $C_s = c_1(1 + m_{\mu^n_s}) + c_2(1 + m_{\mu^\infty_s})$.

Using the result from Lemma 2, we have

$$C_{s+1} \leq A + B \exp(\kappa^2(3L + d)b_s)$$

for $A = (c_1 + c_2)(1 + m_P)$ and $B = c_1 m_{\mu^n_0, P} + c_2 m_{\mu^\infty_0, P}$. Therefore

$$\begin{aligned}\textstyle\sum_{s=0}^{r} \epsilon_s C_s &\leq \max_{0 \leq s \leq r} C_s \textstyle\sum_{s=0}^{r} \epsilon_s \\ &\leq b_r(A + B \exp(\kappa^2(3L + d)b_{r-1})) \\ &\leq b_r(A + B \exp(\kappa^2(3L + d)b_r)). \end{aligned}$$

Plugging this back into (9) proves the result.

Appendix B. Proof of Lemma 2: SVGD moment growth

From Assumption 2 we know

$$\begin{aligned}&|k(y, x)| \leq |\langle k(y, \cdot), k(x, \cdot)\rangle_{\mathcal{H}}| \leq \|k(y, \cdot)\|_{\mathcal{H}} \|k(x, \cdot)\|_{\mathcal{H}} = \sqrt{k(y, y)k(x, x)} \leq \kappa^2, \\ &|D^I_y k(y, x)| = |\langle D^I_y k(y, \cdot), k(x, \cdot)\rangle_{\mathcal{H}}| \leq \sqrt{k(x, x) D^I_y D^I_{y'} k(y, y')|_{y'=y}} \leq \kappa^2, \\ &\left\|\nabla^2_y k(y, x)\right\|_{op} \leq \left\|\nabla^2_y k(y, x)\right\|_F = (\textstyle\sum_{|I|=2} (D^I_y k(y, x))^2)^{1/2} \leq d\kappa^2. \end{aligned}$$

Let μ be any probability measure. Using 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

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

Let $m_{\mu_r,P} \triangleq \mathbb{E}_{X\sim\mu_r \perp\!\!\!\perp Z\sim P}[\|X - Z\|_2]$, $M_{\mu_r,P} \triangleq \mathbb{E}_{X\sim\mu_r \perp\!\!\!\perp Z\sim P}[\|X - Z\|_2^2]$, and $C = \kappa^2(3L+d)$. Then,

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

where the last step used Jensen's inequality $m_{\mu_r,P}^2 \leq M_{\mu_r,P}$. Then

$$\begin{aligned}
M_{\mu_r,P} &\leq M_{\mu_0,P} \textstyle\prod_{s=0}^{r-1}(1 + 2\epsilon_s C + \epsilon_s^2 C^2) \leq M_{\mu_0,P} e^{\sum_{s=0}^{r-1}(2\epsilon_s C + \epsilon_s^2 C^2)} \leq M_{\mu_0,P} e^{(2C + \epsilon_0 C^2)\sum_{s=0}^{r-1}\epsilon_s}. \\
m_{\mu_r} &= \mathbb{E}_{\mu_r}[\|X\|_2] \leq m_{\mu_r,P} + \mathbb{E}_P[\|Z\|_2] \leq m_{\mu_0,P} \textstyle\prod_{s=0}^{r-1}(1 + \epsilon_s C) + m_P \\
&\leq m_{\mu_0,P} \exp(C b_{r-1}) + m_P.
\end{aligned}$$

The final inequality uses the relation $1 + x \leq e^x$.

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

Our proof generalizes that of Gorham and Mackey [6, 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

$$\|g(x)\|_2^2 \leq \textstyle\sum_{i=1}^d \kappa^2 \|g_i\|_{\mathcal{H}}^2 \leq \kappa^2, \tag{10}$$

$$\|\nabla g(x)\|_{op}^2 \leq \|\nabla g(x)\|_F^2 = \textstyle\sum_{i=1}^d \sum_{j=1}^d |\nabla_{x_i} g_j(x)|^2 \leq \sum_{i=1}^d \sum_{j=1}^d \kappa^2 \|g_j\|_{\mathcal{H}}^2 \leq d\kappa^2, \text{ and} \tag{11}$$

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

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 (10), and g and $\nabla \cdot g$ are Lipschitz (11), 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 first state and prove a useful lemma.

Lemma 5 *Suppose f is a monotonic decreasing function. Let $x = f(\beta)$ for $\beta = \max(1, f(1))$. Then $x \leq f(x)$.*

Proof If $\beta = 1 \geq f(1)$, then $x = f(\beta) = f(1) \leq 1$. Because f is monotonic decreasing, $f(x) \geq f(1) = x$. Otherwise, $\beta = f(1) > 1$ and hence $x = f(\beta) < f(1) = \beta$. Again, due to monotonicity, $f(x) > f(\beta) = x$. Combining the two cases finishes the proof. ■

Now we are ready to prove Corollary 2. Let $b_t^{(1)} = \psi_{B,C}(w_0\sqrt{\log\log\frac{1}{w_0}}, A, \beta_1)$ and $b_t^{(2)} = \psi_{B,C}(w_0 \log\log\frac{1}{w_0}, A + 2C + \epsilon_0 C^2, \beta_2)$. According to Lemma 5,

$$b_t^{(1)} \leq \psi_{B,C}(w_0\sqrt{\log\log\tfrac{1}{w_0}}, A, b_t^{(1)}).$$

Rearranging the terms, we have

$$w_0 \exp(b_t^{(1)}(A + B\exp(Cb_t^{(1)}))) \leq \frac{1}{\sqrt{\log\log\frac{1}{w_0}}}. \tag{12}$$

Similarly, we have $b_t^{(2)} \leq \psi_{B,C}(w_0 \log\log\frac{1}{w_0}, A + 2C + \epsilon_0 C^2, b_t^{(2)})$ and

$$\sqrt{w_0}\exp(b_t^{(2)}(2C + \epsilon_0 C^2 + (A + B\exp(Cb_t^{(2)}))/2)) \leq \frac{1}{\sqrt{\log\log\frac{1}{w_0}}}. \tag{13}$$

Since $b_t = \min(b_t^{(1)}, b_t^{(2)})$, we can substitute b_t for $b_t^{(1)}$ and $b_t^{(2)}$ in (12) and (13), respectively. Also, from the definition of b_t we know $\frac{1}{b_t} = O\left(\frac{1}{\log\log\frac{1}{w_0}}\right)$. Plugging them into (2) proves the first claim. The second claim follows from the $\Theta(n^{-1/d})$ convergence rate of an empirical measure to its population measure in 1-Wasserstein distance [3].