A Finite-Particle Convergence Rate for Stein Variational Gradient Descent

author names withheld

Under Review for OPT 2022

Abstract

We provide a first finite-particle convergence rate for Stein variational gradient descent (SVGD). Specifically, with certain choices of step size sequences, SVGD with n particles drive the kernel Stein discrepancy to zero at the rate $O\left(\frac{1}{\sqrt{\log\log n}}\right)$. 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, Liu and Wang [7]) is an algorithm for drawing samples from an (unnormalized) target distribution. The algorithm iteratively evolves a group of particles to approach the target distribution through optimization in the space of probability measures.

Despite the popularity of SVGD, theoretical understanding of the method is quite limited. Prior work [1, 4–6, 9, 10] has only studied the convergence of SVGD when there is infinite number of particles. They typically rely on a descent lemma showing that the KL-divergence between the approximation and the target distribution decreases at each step. Since the KL-divergence is only defined when the support of two distributions match, the lemma cannot be directly applied to finite-particle discrete measures.

In this work, we present the first finite-particle convergence rate of SVGD. In order to derive this, we bound the discretization error of SVGD at all steps using Wasserstein-1 distances between finite and infinite-particle version of the algorithm. We then combine this with the established infinite-particle analysis of SVGD to arrive at an error bound for finite-particle approximations.

1.1. Related work

Liu [6] made the first attempt to the theoretical analysis of SVGD and established the first version of the descent lemma in KL divergence. The descent lemma was then improved by Korba et al. [5], Salim et al. [9], Sun et al. [10] with cleaner proofs, path-independent constants, and weaker smoothness conditions. Duncan et al. [1] studied the continuous-time limit of the algorithm and provide conditions for exponential convergence. Gorham et al. [4] improved the discretization error bound of Liu [6] by using Wasserstein distances. Only Gorham et al. [4], Liu [6] bound the distance between finite and infinite-particle versions of SVGD. However, they did not use them to further develop a finite-particle error bound.

© .

1.2. Notation

Let $\mathcal{H}$ denote the Reproducing Kernel Hilbert Space (RKHS) of kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. $\mathcal{H}^d \triangleq \bigotimes_{i=1}^d \mathcal{H}$ denotes the product RKHS. For any two probability measures μ and ν, $W_1(\mu, \nu) \triangleq \inf_{X\sim\mu, Z\sim\nu} \mathbb{E}[\|Z - X\|]$ denotes the Wasserstein-1 distance between them. For each measure μ, if μ has integrable first and second moments, we define $m_\mu \triangleq \mu(\|\cdot\|_2)$, $m_{\mu,P} \triangleq \mathbb{E}_{X\sim\mu \perp\!\!\!\perp Z\sim P}[\|X - Z\|_2]$, and $M_{\mu,P} \triangleq \mathbb{E}_{X\sim\mu \perp\!\!\!\perp Z\sim P}[\|X - Z\|_2^2]$.

Definition 1 *The Langevin Stein operator [2] for a target measure P is defined by*

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

Definition 2 *The vector-valued Langevin Stein operator [8] for a target measure P is defined by*

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

with components

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

Definition 3 *The SVGD transport map [7] for a target measure P, an approximating measure μ, and a step size ϵ is given by*

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

Definition 4 *We let $\Phi_\epsilon(\mu)$ represent the distribution of $T_{\mu,\epsilon}(X)$ when $X \sim \mu$. In other words, $\Phi_\epsilon(\mu)$ is the pushforward of μ under the action of $T_{\mu,\epsilon}$.*

2. Convergence Analysis for Finite Number of Particles

Assumption 1 $\sup_{x,y} |D_x^I D_y^I k(x, y)| \le \kappa^2$ *for any multi-index $I = (I_1, I_2, \ldots, I_d)$ with $|I| \le 2$, where $|I| = \sum_{i=1}^d I_i$ and $D^I \triangleq \frac{d^{|I|}}{dx_1^{I_1} dx_2^{I_2} \ldots dx_d^{I_d}}$.*

Under Assumption 1, we could show that for any $f \in \mathcal{H}$,

$$|f(x)| \le \|f\|_{\mathcal{H}} \sqrt{k(x, x)} \le \kappa \|f\|_{\mathcal{H}};$$
$$|D^I f(x)| \le \|f\|_{\mathcal{H}} \sqrt{D_x^I D_y^I k(x, x)} \le \kappa \|f\|_{\mathcal{H}}.$$

Assumption 2 *The target distribution P has a differentiable density p on $\mathbb{R}^d$ and $s_p \triangleq \nabla \log p$ is L-Lipschitz, i.e., $\|s_p(x) - s_p(y)\|_2 \le L\|x - y\|_2$ for any $x \ne y$.*

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

Theorem 1 (Discretization error of SVGD) *Suppose Assumptions 1 and 2 hold. Let μ_r^n, μ_r^∞ denote the approximation given by n-particle and infinite-number-of-particle SVGD with step size ϵ_r at step r. And let $b_r \triangleq \sum_{s=0}^r \epsilon_s$. Then*

$$W_1(\mu_{r+1}^n, \mu_{r+1}^\infty) \le W_1(\mu_0^n, \mu_0^\infty) \exp(b_r(A + B \exp(Db_r))),$$

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 $D = \kappa^2(3L + d)$. Here c_1, c_2 are the constants defined in Lemma 1.

Proof By repeated application of Lemma 1 and the inequality $(1+x) \leq e^x$, we have

$$\begin{aligned} W_1(\mu_{r+1}^n, \mu_{r+1}^\infty) &= W_1(\Phi_{\epsilon_r}(\mu_r^n), \Phi_{\epsilon_r}(\mu_r^\infty)) \leq (1+\epsilon_r C_r) W(\mu_r^n, \mu_r^\infty) \\ &\leq W_1(\mu_0^n, \mu_0^\infty) \prod\nolimits_{s=0}^r (1+\epsilon_s C_s) \leq W_1(\mu_0^n, \mu_0^\infty) \exp(\sum\nolimits_{s=0}^r \epsilon_s C_s) \end{aligned} \tag{1}$$

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

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 $B = c_1 m_{\mu_0^n,P} + c_2 m_{\mu_0^\infty,P}$. 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 (1) proves the result.

■

The above proof of Theorem 1 relies on two lemmas. The first, due to Gorham et al. [4], shows that the one-step SVGD mapping Φ_ϵ 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 [4, Lem. 12]) *Suppose that, for some constants $c_1, c_2 > 0$, the following pseudo-Lipschitz bounds hold:*

$$\begin{aligned} \sup_{z\in\mathbb{R}^d} \|\nabla_z(\nabla \log p(x) k(x,z) + \nabla_x k(x,z))\|_{\mathrm{op}} &\leq c_1(1+\|x\|_2), \\ \sup_{x\in\mathbb{R}^d} \|\nabla_x(\nabla \log p(x) k(x,z) + \nabla_x k(x,z))\|_{\mathrm{op}} &\leq c_2(1+\|z\|_2). \end{aligned}$$

Then,

$$W_1(\Phi_\epsilon(\mu), \Phi_\epsilon(\nu)) \leq W(\mu,\nu)(1+\epsilon C(\mu,\nu)),$$

where $C(\mu,\nu) = c_1(1+m_\mu) + c_2(1+m_\nu)$.

The second lemma, proved in Appendix A, controls first absolute moment growth under SVGD.

Lemma 2 (SVGD moment growth) *Suppose Assumptions 1 and 2 hold. Let μ_r denote the approximation given by SVGD at step r. Then,*

$$\begin{aligned} m_{\mu_r} &\leq m_{\mu_0,P} \textstyle\prod_{s=0}^{r-1}(1+\epsilon_s D) + m_P \leq m_{\mu_0,P}\exp(Db_{r-1}) + m_P, \\ M_{\mu_r,P} &\leq M_{\mu_0,P} \textstyle\prod_{s=0}^{r-1}(1+2\epsilon_s D + \epsilon_s^2 D^2) \leq M_{\mu_0,P}\exp((2D+\epsilon_0 D^2)b_{r-1}) \end{aligned} \tag{2}$$

for $D = \kappa^2(3L+d)$.

To establish our main SVGD convergence result, we combine Theorem 1 with the following infinite-particle SVGD convergence rate due to Salim et al. [9] (see also Korba et al. [5], Liu [6]).

Lemma 3 (Infinite-particle SVGD descent lemma, Salim et al. [9]) *For any constant $\alpha > 1$ and $\epsilon_r < \min\left(\frac{2}{\kappa^2(L+\alpha^2)}, R(\alpha)\right)$ ($R(\alpha)$ is the right hand side of Salim et al. [9, (12)]). SVGD with infinite number of particles decreases KL and drives KSD to zero:*

$$\mathrm{KL}(\mu_{r+1}^{\infty}\|P) - \mathrm{KL}(\mu_r^{\infty}\|P) \leq -\epsilon_r\left(1 - \frac{\kappa^2(L+\alpha^2)}{2}\epsilon_r\right)\mathrm{KSD}(\mu_r^{\infty}, P)^2. \tag{3}$$

.

By summing (3) 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 $\epsilon_r \leq \min\left(\frac{1}{\kappa^2(L+\alpha^2)}, R(\alpha)\right)$, then we have $\frac{\epsilon_r}{2} \leq c(\epsilon_r) < \epsilon_r$ and thereby*

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

The above infinite-particle SVGD convergence rate is computed with respect to KSD. To relate it to the Wasserstein discretization error bound in Theorem 1, we establish the following KSD-Wasserstein bound.

Theorem 2 (KSD-Wasserstein bound) *Suppose Assumptions 1 and 2 holds. Let μ_r^n, μ_r^{∞} denote the approximation by n particle and infinitely many particle SVGD at iteration r. Then,*

$$\begin{aligned}\mathrm{KSD}(\mu_r^n, P) \leq\ & \kappa(d+L)W_1(\mu_r^n, \mu_r^{\infty}) + d^{1/4}\kappa L\exp((2D+\epsilon_0 D^2)b_{r-1})\sqrt{2M_{\mu_0^{\infty},P}W_1(\mu_r^n, \mu_r^{\infty})} \\ & + \mathrm{KSD}(\mu_r^{\infty}, P).\end{aligned}$$

Proof Using the result of Lemma 4, Cauchy-Schwarz, Eq. (2) of Lemma 2, and $\min(a,b)^2 \leq ab$, we have

$$\begin{aligned}\mathrm{KSD}(\mu_r^n, P) &= \sup_{\|g\|_{\mathcal{H}^d}\leq 1}\mathbb{E}_{\mu_r^n}[(\mathcal{T}_P g)(\cdot)] - \mathbb{E}_{\mu_r^{\infty}}[(\mathcal{T}_P g)(\cdot)] + \mathbb{E}_{\mu_r^{\infty}}[(\mathcal{T}_P g)(\cdot)] \\ &\leq \sup_{\|g\|_{\mathcal{H}^d}\leq 1}\mathbb{E}_{\mu_r^n}[(\mathcal{T}_P g)(\cdot)] - \mathbb{E}_{\mu_r^{\infty}}[(\mathcal{T}_P g)(\cdot)] + \sup_{\|g\|_{\mathcal{H}^d}\leq 1}\mathbb{E}_{\mu_r^{\infty}}[(\mathcal{T}_P g)(\cdot)] \\ &\leq \kappa(d+L)W_1(\mu_r^n, \mu_r^{\infty}) + L\mathbb{E}_{X\sim\mu_r^n, Y\sim\mu_r^{\infty}, Z\sim P}[\|Y-Z\|_2\min(2\kappa, \sqrt{d}\kappa\|X-Y\|_2)] \\ &\quad + \mathrm{KSD}(\mu_r^{\infty}, P) \\ &\leq \kappa(d+L)W_1(\mu_r^n, \mu_r^{\infty}) + LM_{\mu_r^{\infty},P}^{1/2}\mathbb{E}_{X\sim\mu_r^n, Y\sim\mu_r^{\infty}}[\min(2\kappa, \sqrt{d}\kappa\|X-Y\|_2)^2]^{1/2} \\ &\quad + \mathrm{KSD}(\mu_r^{\infty}, P) \\ &\leq \kappa(d+L)W_1(\mu_r^n, \mu_r^{\infty}) + \sqrt{2M_{\mu_r^{\infty},P}}d^{1/4}\kappa L\mathbb{E}_{X\sim\mu_r^n, Y\sim\mu_r^{\infty}}[\|X-Y\|_2]^{1/2} \\ &\quad + \mathrm{KSD}(\mu_r^{\infty}, P).\end{aligned}$$

Taking the infinimum over the joint distribution of (X, Y) leads to

$$\begin{aligned}\mathrm{KSD}(\mu_r^n, P) \leq\ & \kappa(d+L)W_1(\mu_r^n, \mu_r^{\infty}) + d^{1/4}\kappa L\exp((2D+\epsilon_0 D^2)b_{r-1})\sqrt{2M_{\mu_0^{\infty},P}W_1(\mu_r^n, \mu_r^{\infty})} \\ & + \mathrm{KSD}(\mu_r^{\infty}, P).\end{aligned}$$

■

Theorem 2's proof uses the following lemma, a generalization of Gorham and Mackey [3, Lemma 18] that we prove in Appendix B, shows that the Stein MMD is controlled by the 1-Wasserstein distance.

Lemma 4 *Suppose Assumptions 1 and 2 hold. Let μ, ν be two probability measures on $\mathbb{R}^d$ and $g \in \mathcal{H}^d$ satisfying $\|g\|_{\mathcal{H}^d}^2 \triangleq \sum_{i=1}^d \|g_i\|_{\mathcal{H}}^2 \leq 1$. Then,*

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

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

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

where $a_t = \kappa(d+L)w_0 \exp(b_{t-1}(A + B\exp(Db_{t-1}))) + d^{1/4}\kappa L\sqrt{2M_{\mu_0^\infty,P} w_0}\exp((2D + \epsilon_0 D^2 + (A + B\exp(Db_{t-1}))/2)b_{t-1})$, and A, B, D are defined as in Theorem 1.

Proof Let $w_r \triangleq W_1(\mu_r^n, \mu_r^\infty)$. From Theorem 2, we know

$$\mathrm{KSD}(\mu_r^n, P) \leq \kappa(d+L)w_r + d^{1/4}\kappa L\exp((2D+\epsilon_0 D^2)b_{r-1})\sqrt{2M_{\mu_0^\infty,P} w_r} + \mathrm{KSD}(\mu_r^\infty, P).$$

Plugging in Theorem 1's result

$$w_r \leq w_0 \exp(b_{r-1}(A + B\exp(Db_{r-1}))),$$

we have $\mathrm{KSD}(\mu_r^n, P) - a_r \leq \mathrm{KSD}(\mu_r^\infty, P)$, where

$$\begin{aligned}a_r &= \kappa(d+L)w_0 \exp(b_{r-1}(A + B\exp(Db_{r-1}))) \\ &\quad + d^{1/4}\kappa L\sqrt{2M_{\mu_0^\infty,P} w_0}\exp((2D + \epsilon_0 D^2 + (A + B\exp(Db_{r-1}))/2)b_{r-1}).\end{aligned}$$

Taking square on both sides and the weighted sum over $r = 0, \ldots, t-1$ with weights $\pi_r \triangleq \frac{c(\epsilon_r)}{\sum_{r=0}^{t-1} c(\epsilon_r)}$, we obtain

$$\sum_{r=0}^{t-1} \pi_r (\mathrm{KSD}(\mu_r^n, P) - a_r)^2 \leq \sum_{r=0}^{t-1} \pi_r \mathrm{KSD}(\mu_r^\infty, P)^2 \leq \tfrac{2}{b_{t-1}}\mathrm{KL}(Q_0^\infty \| P), \tag{4}$$

where the last inequality follows from Corollary 1. Moreover, by Cauchy-Schwarz,

$$\sum_{r=0}^{t-1} \pi_r (\mathrm{KSD}(\mu_r^n, P) - a_r)^2 \geq \left(\sum_{r=0}^{t-1} \pi_r \mathrm{KSD}(\mu_r^n, P) - \sum_{r=0}^{t-1} \pi_r a_r\right)^2 \tag{5}$$

Combining (4) and (5), we have

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

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

According to Theorem 3, we need to choose b_{t-1} by choosing the step size sequence and t to balance the two terms.

Corollary 2 (Convergence rate) *Choose* $\epsilon_r \leq \min\left(\frac{1}{\kappa^2(L+\alpha^2)}, R(\alpha)\right)$ *and* t *such that*

$$b_{t-1} = \frac{1}{D} \log\left(\frac{1}{B}\left(\frac{1}{\beta} \log \frac{1}{w_0 \log\log \frac{1}{w_0}} - A\right)\right),$$

where $\beta = \max\left(1, \frac{1}{D} \log\left(\frac{1}{B}\left(\frac{1}{\beta} \log \frac{1}{w_0 \log\log \frac{1}{w_0}} - A\right)\right)\right)$. *Then,*

$$\min_{0 \leq r < t} \mathrm{KSD}(\mu_r^n, P) = O\left(\frac{1}{\sqrt{\log\log \frac{1}{w_0}}}\right)$$

and if $\mu_0^n = \sum_{i=1}^n \delta_{x_i}$, $\{x_i\}_{i=1}^n$ *are i.i.d. samples of* μ_0^∞,

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

The last result is due to the well-known $\Theta(n^{-1/d})$ convergence rate of empirical measure in Wasserstein distance.

References

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

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

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

[4] Jackson Gorham, Anant Raj, and Lester Mackey. Stochastic stein discrepancies. *arXiv preprint arXiv:2007.02857*, 2020.

[5] 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.

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

[7] 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.

[8] 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.

[9] Adil Salim, Lukang Sun, and Peter Richtárik. Complexity analysis of Stein variational gradient descent under Talagrand's inequality T1. *arXiv preprint arXiv:2106.03076*, 2021.

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

Appendix A. Proof of Lemma 2: SVGD moment growth

From the assumptions we know

$$
\begin{aligned}
&|k(y,x)| \le \kappa^2, \\
&\left\|\nabla_y^2 k(y,x)\right\|_{op} \le \left\|\nabla_y^2 k(y,x)\right\|_F = (\textstyle\sum_{|I|=2}(D^I k(y,x))^2)^{1/2} \le d\kappa^2.
\end{aligned}
$$

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

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

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 $D = \kappa^2(3L+d)$. Then,

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

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

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

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

Appendix B. Proof of Lemma 4

From Assumption 1 we know

$$\|g(x)\|_2^2 \le \sum_{i=1}^d \kappa^2 \|g_i\|_{\mathcal{H}}^2 \le \kappa^2,$$
$$\|\nabla g(x)\|_{op}^2 \le \|\nabla g(x)\|_F^2 = \sum_{i=1}^d \sum_{j=1}^d |\nabla_{x_i} g_j(x)|^2 \le \sum_{i=1}^d \sum_{j=1}^d \kappa^2 \|g_j\|_{\mathcal{H}}^2 \le d\kappa^2,$$
$$\|\nabla(\nabla \cdot g(x))\|_2^2 = \sum_{i=1}^d \left(\sum_{j=1}^d \nabla_{x_i} \nabla_{x_j} g_j(x)\right)^2 \le d \sum_{i=1}^d \sum_{j=1}^d |\nabla_{x_i} \nabla_{x_j} g_j(x)|^2 \le d^2\kappa^2.$$

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 $\mathbb{E}_P[s_p(\cdot)] = 0$, repeated use of Cauchy-Schwarz gives

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