Fast Approximation of the Generalized Sliced-Wasserstein Distance

Dung Le[†,*], Huy Nguyen[◇,*], Khai Nguyen[◇,*], Trang Nguyen[‡,*], Nhat Ho[◇]

[†]École Polytechnique, [◇]The University of Texas at Austin, [‡]VinAI Research

ABSTRACT

* A Monte Carlo approximation of the generalized Sliced-Wasserstein (GSW) distance has large complexity in high-dimensional settings.

* We propose deterministic and fast approximations of the GSW distance based on Gaussian projections when the defining functions are polynomial function and linear neural network function.

GENERALIZED SLICED-WASSERSTEIN DISTANCE (GSW)

Sliced-Wasserstein distance between two probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ based on Gaussian projections $\theta^* : x \mapsto \langle x, \theta \rangle$ is defined as follows:

$$SW_p(\mu,\nu) := \left(\int_{\mathbb{R}^d} W_p^p(\theta^*_\sharp \mu, \theta^*_\sharp \nu)\, \mathrm{d}\gamma_d(\theta) \right)^{\frac{1}{p}}, \qquad (1)$$

where $\gamma_d := \mathcal{N}(\mathbf{0}_d, d^{-1}\boldsymbol{I}_d)$ and $W_p(\theta^*_\sharp \mu, \theta^*_\sharp \nu)$ denotes the Wasserstein distance between one-dimensional probability measures, which admits the following closed-form:

$$W_p^p(\theta^*_\sharp \mu, \theta^*_\sharp \nu) = \int_0^1 \left| F^{-1}_{\theta^*_\sharp \mu}(z) - F^{-1}_{\theta^*_\sharp \nu}(z) \right| \mathrm{d}z, \qquad (2)$$

with $\theta^*_\sharp \mu(V) := \mu\big((\theta^*)^{-1}(V)\big)$ for any subset $V \subseteq \mathbb{R}$.

Generalized Sliced-Wasserstein Distance between probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ based on Gaussian projections is given by:

$$GSW_p(\mu,\nu) := \left(\int_{\mathbb{R}^d} W_p^p(g^\theta_\sharp \mu, g^\theta_\sharp \nu)\, \mathrm{d}\gamma_d(\theta) \right)^{\frac{1}{p}}, \qquad (3)$$

where g^θ is a possibly non-linear function.

CONDITIONAL CENTRAL LIMIT THEOREM FOR GAUSSIAN PROJECTIONS

Theorem 1 (Reeves et al., 2017)**.** *For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\int_{\mathbb{R}^d} W_2^2\Big(\theta^*_\sharp \mu, \mathcal{N}(\mathbf{0}_d, d^{-1}\mathfrak{m}_2(\mu))\Big) \mathrm{d}\gamma_d(\theta) \leq C d^{-1} \Big\{ A(\mu) + [\mathfrak{m}_2(\mu) B_1(\mu)]^{1/2} + \mathfrak{m}_2(\mu)^{1/5} B_2(\mu)^{4/5} \Big\}, \qquad (4)$$

where $C > 0$ is some universal constant and

$$\mathfrak{m}_2(\mu) := \mathbb{E}_\mu[\|X\|^2], \; A(\mu) := \mathbb{E}_\mu \big| \|X\|^2 - \mathfrak{m}_2(\mu) \big|,$$
$$B_k(\mu) := \Big[\mathbb{E}_{\mu \otimes \mu} \big| \langle X, X' \rangle^k \big| \Big]^{1/k}.$$

FAST APPROXIMATION OF GSW WITH POLYNOMIAL FUNCTION

Polynomial Function. For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ and a vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we denote $|a| = \alpha_1 + \ldots + \alpha_d$ and $x^a = x_1^{\alpha_1} \ldots x_d^{\alpha_d}$. Then, a polynomial function of degree m is defined as

$$g_{\text{poly}}(x, \theta) := \sum_{|\alpha| = m} \theta_\alpha x^\alpha, \qquad (5)$$

where $\theta := (\theta_\alpha)_{|\alpha|=m} \in \mathbb{R}^{q-1}$ such that $\|\theta\|_2 = 1$ with $q = \binom{m+d-1}{d-1}$ being the number of non-negative solutions to the equation $\alpha_1 + \ldots + \alpha_d = m$.

Theorem 2. *Let $X \sim \mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $Y \sim \nu \in \mathcal{P}_2(\mathbb{R}^d)$. Assume that $(X^\alpha)_{|\alpha|=m} \sim \mu_q$ and $(Y^\alpha)_{|\alpha|=m} \sim \nu_q$. Additionally, let $\eta_{\mu_q} = \mathcal{N}(\mathbf{0}_q, q^{-1}\mathfrak{m}_2(\mu_q))$ and $\eta_{\nu_q} = \mathcal{N}(\mathbf{0}_q, q^{-1}\mathfrak{m}_2(\nu_q))$. Then, under some mild assumptions on μ and ν, we have that*

$$\left| \text{poly} - GSW_2(\overline{\mu}, \overline{\nu}) - W_2(\eta_{\overline{\mu}_q}, \eta_{\overline{\nu}_q}) \right| \leq \mathcal{O}(d^{-\frac{m}{8}}), \qquad (6)$$

where $\overline{\mu}_q$ and $\overline{\nu}_q$ are centered versions of μ_q and ν_q.

Fast Approximation of $\text{poly} - GSW$**.**

$$\widehat{\text{poly} - GSW}_2^2(\mu,\nu) = \widehat{\text{poly} - GSW}_2^2(\overline{\mu}_q, \overline{\nu}_q) + q^{-1}\|\widehat{m}_{\mu_q} - \widehat{m}_{\nu_q}\|^2$$
$$= q^{-1} \left(\sqrt{\widehat{\mathfrak{m}}_2(\overline{\mu}_q)} - \sqrt{\widehat{\mathfrak{m}}_2(\overline{\nu}_q)} \right)^2 + q^{-1}\|\widehat{m}_{\mu_q} - \widehat{m}_{\nu_q}\|^2, \qquad (7)$$

where m_{μ_q} and m_{ν_q} are means of μ_q and ν_q, respectively.

FAST APPROXIMATION OF GSW WITH LINEAR NEURAL NETWORK TYPE FUNCTION

Linear Neural Network Function. Let $X \sim \mu$, $Y \sim \nu$ and $\Theta^{(1)}, \ldots, \Theta^{(n)}$ be random matrices of size $d \times d$ independent of X and Y such that their entries are i.i.d random variables following $\mathcal{N}(0, d^{-1})$. Then, a linear neural network function is:

$$g_{\text{neural}}(x, \theta) := \langle \theta, \Theta^{(1)} \ldots \Theta^{(n)} x \rangle \qquad (8)$$

Theorem 3. *Let $X^* := \Theta^{(1)} \ldots \Theta^{(n)} X$ and $Y^* := \Theta^{(1)} \ldots \Theta^{(n)} Y$ have probability distributions μ^* and ν^*, respectively. Then, under some mild assumptions on $\mu$and ν, we have that*

$$|\text{neural} - GSW_2(\mu,\nu) - W_2(\mu^*, \nu^*)| \leq \mathcal{O}(3^{\frac{n}{4}} d^{-\frac{1}{4}} + d^{-\frac{1}{8}}).$$

Fast Approximation of $\text{neural} - GSW$**.**

$$\widehat{\text{neural} - GSW}_2^2(\mu,\nu) = d^{-1} \left(\sqrt{\widehat{\mathfrak{m}}_2(\mu^*)} - \sqrt{\widehat{\mathfrak{m}}_2(\nu^*)} \right)^2. \qquad (9)$$

EXPERIMENTS

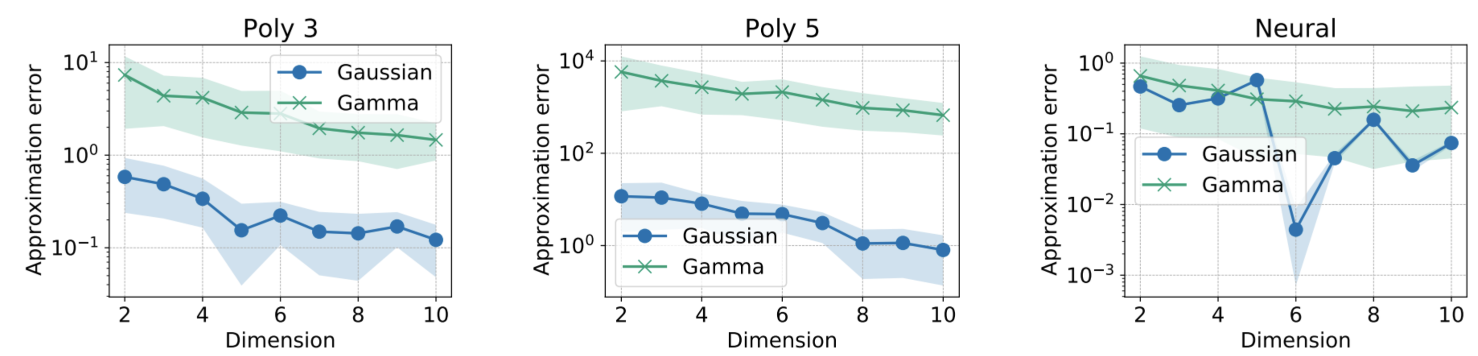

Figure 1: Approximation error between approximated GSW with the Monte Carlo GSW with a huge number of projections between empirical distributions on samples that are drawn from Multivariate Gaussian distributions and Gamma distributions.

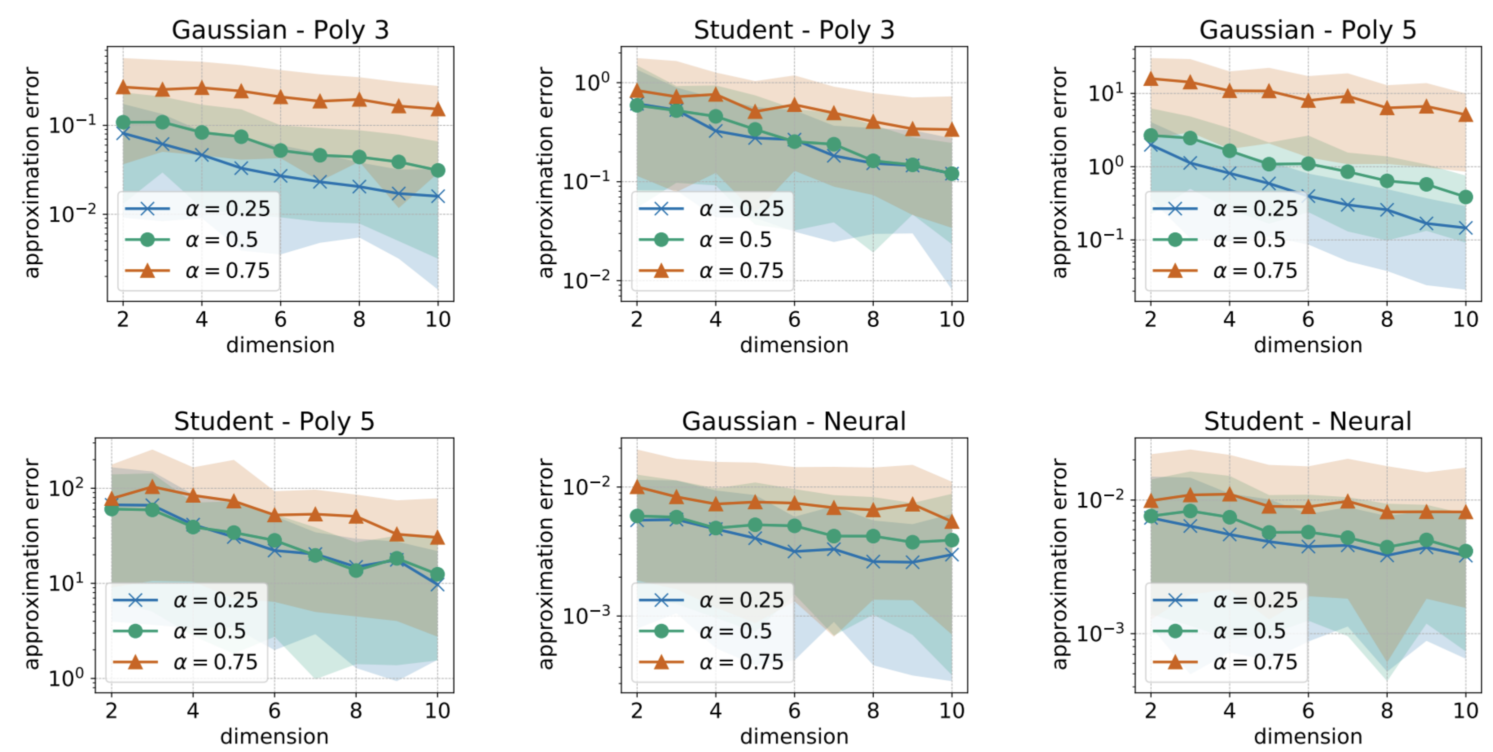

Figure 2: Approximation error between approximated GSW with the Monte Carlo GSW with a huge number of projections between empirical distributions on samples that are drawn from autoregressive processes of order one AR(1).