Fast Approximation of the Generalized Sliced-Wasserstein Distance

Dung Le [* 1]   **Huy Nguyen** [* 2]   **Khai Nguyen** [* 2]   **Nhat Ho** [2]

Abstract

Generalized sliced-Wasserstein distance is a variant of sliced-Wasserstein distance that exploits the power of non-linear projection through a given defining function to better capture the complex structures of probability distributions. Similar to the sliced-Wasserstein distance, generalized sliced-Wasserstein is defined as an expectation over random projections which can be approximated by the Monte Carlo method. However, the complexity of that approximation can be expensive in high-dimensional settings. To that end, we propose to form deterministic and fast approximations of the generalized sliced-Wasserstein distance by using the concentration of random projections when the defining functions are polynomial function and neural network type function. Our approximations hinge upon an important result that one-dimensional projections of a high-dimensional random vector are approximately Gaussian.

1. Introduction

Sliced-Wasserstein (SW) distance (Bonneel et al., 2015) has become a core member in the family of probability metrics that are based on optimal transport (Villani, 2008). Compared to Wasserstein distance, SW provides a lower computational cost thanks to the closed-form solution of optimal transport in one-dimensional setting. In particular, when dealing with probability measures with at most n support points, the computational complexity of SW is $\mathcal{O}(n \log n)$ while that of Wasserstein distance is $\mathcal{O}(n^3 \log n)$ (Pele & Werman, 2009). Furthermore, the memory complexity of SW is only $\mathcal{O}(n)$ in comparison with $\mathcal{O}(n^2)$ of Wasserstein distance (due to the storage of a cost matrix). Additionally, the statistical estimation rate (or the sample complexity) of

SW does not depend on the dimension (denoted as d) like Wasserstein distance. In particular, the sample complexity of the former is $\mathcal{O}(n^{-1/2})$ (Bobkov & Ledoux, 2019), whereas it is of $\mathcal{O}(n^{-1/d})$ (Fournier & Guillin, 2015) for the latter. Therefore, the SW does not suffer from the curse of dimensionality.

Due to the practicality of SW, several improvements and variants of that distance have been explored recently. For instance, selective discriminative projecting direction techniques are proposed in (Deshpande et al., 2019; Nguyen et al., 2021a;b); a SW variant that augments original measures to higher dimensions for better linear separation is introduced in (Chen et al., 2022); a SW variant on the sphere is defined in (Bonet et al., 2022); a SW variant that uses convolution slicer for projecting images is proposed in (Nguyen & Ho, 2022b). However, the prevailing trend of the current work on SW is focused on its application. Indeed, SW is used in generative modeling (Deshpande et al., 2018; Nguyen & Ho, 2022a; Dai & Seljak, 2021; Nguyen et al., 2023), domain adaptation (Wu et al., 2019), clustering (Kolouri et al., 2018), and Bayesian inference (Nadjahi et al., 2020a; Yi & Liu, 2021).

To further enhance the ability of SW, (Kolouri et al., 2019) propose using non-linear projecting defining functions for SW instead of conventional linear projecting. This extension leads to a generalized class of sliced probability distances, named the generalized sliced-Wasserstein (GSW) distance. Despite being more expressive, GSW also needs to be approximated by the Monte Carlo method as SW. In greater detail, the definition of GSW is an expectation over random projections via certain defining functions of Wasserstein distance between corresponding one-dimensional projected probability measures. In general, the expectation is intractable to compute; hence, Monte Carlo samples are used to approximate the expectation as mentioned. It is shown in both theory and practice that the number of Monte Carlo samples (the number of projections) should be large for good performance and approximation of sliced probability metrics (Nadjahi et al., 2020b).

Contribution. In this work, we aim to overcome the projection complexity of the GSW by deriving fast approximations of that distance that do not require using Monte Carlo random projecting directions. We follow the approach of deter-

---

*Equal contribution   [1]École Polytechnique   [2]Department of Statistics and Data Sciences, The University of Texas at Austin. Correspondence to: Huy Nguyen <huynm@utexas.edu>.

ministic approximation of the SW in (Nadjahi et al., 2021), which utilizes the Gaussian concentration of the distribution of low-dimensional projections of high-dimensional random variables (Sudakov, 1978; Diaconis & Freedman, 1984). Our paper covers the settings when the (non-linear) defining functions are polynomial function with odd degree and neural network type, which had been discussed in (Kolouri et al., 2019), while the result for circular function setting is left for future work. It should be emphasized that despite sharing the same framework as (Nadjahi et al., 2021), the error analysis in our paper is much more challenging due to the structures of considered defining functions, which requires us to derive non-trivial combinatorial results to handle.

Organization. The paper is organized as follows. We provide background on Wasserstein distance, sliced-Wasserstein distance and its fast approximation, as well as revisit the generalized sliced-Wasserstein distance in Section 2. We then study the fast approximation of the generalized sliced-Wasserstein distance when the defining function is polynomial with odd degree in Section 3 and when the defining function is neural network type in Section 4. Meanwhile, the approximation when the defining function is circular is briefly discussed in Appendix B. Finally, we give empirical results for the approximation errors of the proposed approximated generalized sliced-Wasserstein distance in Section 5 before concluding the paper in Section 6. The remaining proofs of key results in the paper are deferred to Appendix A.

Notation. We use the following notations throughout our paper. Firstly, we denote by $\mathbb{N}$ the set of all positive integers. For any $d \in \mathbb{N}$ and $p \in \mathbb{N}$, $\mathcal{P}_p(\mathbb{R}^d)$ stands for the set of all probability measures in $\mathbb{R}^d$ with finite moments of order p whereas $\mathbb{S}^{d-1} := \{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$ denotes the d-dimensional unit sphere where $\|\cdot\|$ is the Euclidean norm. Additionally, γ_d represents the Gaussian distribution in $\mathbb{R}^d$, $\mathcal{N}(0, d^{-1}\mathbf{I}_d)$ in which $\mathbf{I}_d$ is an identity matrix of size $d \times d$. Meanwhile, we denote $L^1(\mathbb{R}^d) := \{f : \mathbb{R}^d \to \mathbb{R} : \int_{\mathbb{R}^d} |f(x)|dx < \infty\}$ as the set of all absolutely integrable functions on $\mathbb{R}^d$. Next, for any set A, we denote by $|A|$ its cardinality. Finally, for any two sequences (a_n) and (b_n), the notation $a_n = \mathcal{O}(b_n)$ indicates that $a_n \leq Cb_n$ for all $n \in \mathbb{N}$ where C is some universal constant.

2. Background

In this section, we first revisit Wasserstein distance and the conditional central limit theorem for Gaussian projections. We then present background on sliced-Wasserstein distance and its fast approximation. Finally, we recall the definition of generalized sliced-Wasserstein distance, which is mainly focused on in this paper.

2.1. Wasserstein Distance

Let $p \geq 1$ and μ, ν be two probability measures on $\mathbb{R}^d$, $d \geq 1$, with finite moments of order p. Then, the p-Wasserstein distance between μ and ν is defined as follows:

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi(x, y) \right)^{\frac{1}{p}},$$

where $\|\cdot\|$ denotes the Euclidean norms, and $\Pi(\mu, \nu)$ is the set of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ which admit μ and ν as their marginals.

Next, we review an important result about the concentration of measure phenomenon, which states that under mild assumptions, one-dimensional projections of a high-dimensional random vector are approximately Gaussian. Specifically, we have the following theorem.

Theorem 2.1 ((Reeves, 2017)). *For any $d \geq 1$, let μ be the distribution of $X_{1:d} = (X_1, \ldots, X_d)$. Assume that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, then there exists a universal constant $C \geq 0$ such that:*

$$\int_{\mathbb{R}^d} W_2^2 \left(\theta_\sharp^* \mu, N(0, d^{-1}\mathfrak{m}_2(\mu)) \right) d\gamma_d(\theta) \leq C \, \Xi_d(\mu),$$

where $\theta^* : \mathbb{R}^d \to \mathbb{R}$ denotes the linear form $x \mapsto \langle \theta, x \rangle$, $\theta_\sharp^* \mu$ indicates the push-forward measure of μ by θ^* and

$$\Xi_d(\mu) = d^{-1}\Big\{ A(\mu) + [\mathfrak{m}_2(\mu)B_1(\mu)]^{1/2}$$
$$+ \mathfrak{m}_2(\mu)^{1/5}B_2(\mu)^{4/5} \Big\}, \tag{1}$$
$$\mathfrak{m}_2(\mu) = \mathbb{E}\left[\|X_{1:d}\|^2 \right], A(\mu) = \mathbb{E}\left[\left| \|X_{1:d}\|^2 - \mathfrak{m}_2(\mu) \right| \right],$$
$$B_k(\mu) = \mathbb{E}^{1/k}\left[|\langle X_{1:d}, X'_{1:d}\rangle|^k \right],$$

with $k \in \{1, 2\}$ and $X'_{1:d}$ is an independent copy of $X_{1:d}$.

It is worth noting that the above result only holds for the 2-Wasserstein distance.

2.2. Sliced-Wasserstein Distance And Its Fast Approximation

To adapt the result of Theorem 2.1 to the sliced-Wasserstein setting, (Nadjahi et al., 2021) introduce a new version of SW distance in which projections are sampled from the Gaussian distribution rather than uniformly on the unit sphere as usual. In particular,

Sliced-Wasserstein Distance: For any $p \geq 1$, the sliced-Wasserstein distance of order p with Gaussian projections between two probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as follows:

$$SW_p(\mu, \nu) := \left(\int_{\mathbb{R}^d} W_p^p(\theta_\sharp^* \mu, \theta_\sharp^* \nu) d\gamma_d(\theta) \right)^{\frac{1}{p}}. \tag{2}$$

The notation $\theta^* \sharp \mu$ is equivalent to the Radon Transform of μ given the projecting direction θ^* (Kolouri et al., 2019). By leveraging Theorem 2.1, (Nadjahi et al., 2021) provide the following bound for the sliced-Wasserstein distance between any two probability measures with finite second moments.

Proposition 2.2 ((Nadjahi et al., 2021)). *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be two probability measures in $\mathbb{R}^d$. Consider two Gaussian distributions $\eta_\mu = \mathcal{N}(0, d^{-1}\mathfrak{m}_2(\mu))$ and $\eta_\nu = \mathcal{N}(0, d^{-1}\mathfrak{m}_2(\nu))$, where $\mathfrak{m}_2(\mu), \mathfrak{m}_2(\nu)$ are given in equation* (1). *Then, there exists a universal constant $C > 0$ such that:*

$$|SW_2(\mu, \nu) - W_2(\eta_\mu, \eta_\nu)| \leq C(\Xi_d(\mu) + \Xi_d(\nu))^{\frac{1}{2}}, \quad (3)$$

where $\Xi_d(\mu)$ and $\Xi_d(\nu)$ are defined in equation (1).

Note that equation (3) can be simplified by using the closed-form expression of Wasserstein distance between two Gaussians distributions η_μ and η_ν, which is given by

$$W_2(\eta_\mu, \eta_\nu) = d^{-\frac{1}{2}} \left| \sqrt{\mathfrak{m}_2(\mu)} - \sqrt{\mathfrak{m}_2(\nu)} \right|.$$

According to (Nadjahi et al., 2021), $\Xi_d(\mu)$ and $\Xi_d(\nu)$ cannot be shown to converge to 0 if the data are not centered. Fortunately, they demonstrate that there is a relation between $SW_2(\mu, \nu)$ and $SW_2(\bar{\mu}, \bar{\nu})$, where $\bar{\mu}$ and $\bar{\nu}$ are centered versions of μ and ν, respectively.

Proposition 2.3 ((Nadjahi et al., 2021)). *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be two probability measures in $\mathbb{R}^d$ with respective means m_μ and m_ν. Then, the sliced-Wasserstein distance of order 2 between μ and ν can be decomposed as:*

$$SW_2^2(\mu, \nu) = SW_2^2(\bar{\mu}, \bar{\nu}) + d^{-1}\|m_\mu - m_\nu\|^2. \quad (4)$$

As a consequence, (Nadjahi et al., 2021) successfully derive a deterministic approximation for $SW_2(\mu, \nu)$ as follows:

$$\widehat{SW}_2^2(\mu, \nu) = W_2^2(\eta_{\bar{\mu}}, \eta_{\bar{\nu}}) + d^{-1}\|m_\mu - m_\nu\|^2. \quad (5)$$

2.3. Generalized Sliced-Wasserstein Distance

Inspired by the approximation of SW distance in equation (5), we manage to extend that result to the setting of generalized sliced-Wasserstein (GSW) distance in this work. Before exploring the aforementioned extension, it is necessary to recall the definition of the GSW distance.

Generalized Sliced-Wasserstein Distance: Let g be a defining function (Kolouri et al., 2019) and δ be the Dirac delta function, then the generalized Radon transform (GRT) of an integrable function $I \in L^1(\mathbb{R}^d)$, denoted by $\mathcal{G}I$, is defined as follows:

$$\mathcal{G}I(t, \theta) := \int_{\mathbb{R}^d} I(x)\delta(t - g(x, \theta))dx. \quad (6)$$

When $g(x, \theta) = \langle x, \theta \rangle$, GRT reverts into the conventional Radon Transform which is used in SW distance. By using the GRT, the GSW distance is given by:

$$GSW_p(\mu, \nu) := \left(\int_{\mathbb{R}^d} W_p^p \left(\mathcal{G}I_\mu(\cdot, \theta), \mathcal{G}I_\nu(\cdot, \theta) \right) d\gamma_d(\theta) \right)^{\frac{1}{p}},$$

where $I_\mu, I_\nu \in L^1(\mathbb{R}^d)$ are probability density functions of measures μ and ν, respectively. Here, with a slight abuse of notation, we use $W_p(\mu, \nu)$ and $W_p(I_\mu, I_\nu)$ interchangeably. In this paper, we will also use the pushforward measures notation to define GSW e.g., $g^\theta \sharp \mu$ denotes the GRT of μ given the defining function g and its parameter θ.

In order for the GSW distance to become a proper metric, the GRT must be essentially an injective function. There is a line of work (Homan & Zhou, 2017; Beylkin, 1984) studying the sufficient and necessary conditions for the injectivity of GRT, which finds that the GRT is injective when g is either a polynomial defining function or a circular defining function. By contrast, it is non-trivial to show that GRT is injective when g is a neural network type function; therefore, GSW, in this case, is a pseudo-metric.

As mentioned in Section 2.1, the result in Theorem 2.1 only applies to the 2-Wasserstein distance. Thus, we only consider the GSW of the same order throughout this paper.

3. Polynomial Defining Function

In this section, we consider the problem of finding a deterministic approximation for the generalized sliced-Wasserstein distance under the setting when the defining function g is a polynomial function with an odd degree, which is defined as follows:

Definition 3.1 (**Polynomial defining function**). For a multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ and a vector $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we denote $|\boldsymbol{\alpha}| = \alpha_1 + \ldots + \alpha_d$ and $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \ldots x_d^{\alpha_d}$. Then, a polynomial defining function with an odd degree m is given by: $g_{\mathsf{poly}}(\boldsymbol{x}, \theta) = \sum_{|\boldsymbol{\alpha}|=m} \theta_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}}$, where $\theta := (\theta_{\boldsymbol{\alpha}})_{|\boldsymbol{\alpha}|=m} \in \mathbb{S}^{q-1}$ with $q = \binom{m+d-1}{d-1}$ be the number of non-negative solutions to the equation $\alpha_1 + \ldots + \alpha_d = m$.

Accordingly, the generalized sliced-Wasserstein distance in this case is denoted as $\mathrm{poly}-GSW$.

Subsequently, we introduce some necessary notations for our analysis. Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^\top$ be random vectors following probability distributions $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, respectively. For an odd positive integer $m \in \mathbb{N}$, by denoting μ_q and ν_q as the probability distributions in $\mathbb{R}^q$ of random vectors $\boldsymbol{U} := (\boldsymbol{X}^{\boldsymbol{\alpha}})_{|\boldsymbol{\alpha}|=m} \in \mathbb{R}^q$ and $\boldsymbol{V} := (\boldsymbol{Y}^{\boldsymbol{\alpha}})_{|\boldsymbol{\alpha}|=m} \in \mathbb{R}^q$, we find that there is a connection between the GSW distance and the SW distance as follows:

Proposition 3.2. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be two probability measures in $\mathbb{R}^d$ with finite second moments and $\mu_q, \nu_q \in \mathcal{P}_2(\mathbb{R}^q)$ be defined as above where $q = \binom{m+d-1}{d-1}$ with $m \in \mathbb{N}$ is an odd positive integer. Then, we have:* poly$-GSW_2(\mu, \nu) = SW_2(\mu_q, \nu_q)$.

Proof of Proposition 3.2. For $\theta \in \mathbb{R}^q$, we denote $g_{\text{poly}}^\theta : \mathbb{R}^d \to \mathbb{R}$ as a function $\boldsymbol{x} \mapsto g_{\text{poly}}(\boldsymbol{x}, \theta)$. It follows from the definition of poly$-GSW$ distance that

$$
\begin{aligned}
&\text{poly} - GSW_2^2(\mu, \nu) \\
&= \int_{\mathbb{R}^q} W_2^2\Big((g_{\text{poly}}^\theta)_\sharp \mu, (g_{\text{poly}}^\theta)_\sharp \nu\Big) d\gamma_q(\theta) \\
&= \int_{\mathbb{R}^q} W_2^2(\theta_\sharp^* \mu_q, \theta_\sharp^* \nu_q) d\gamma_q(\theta) = SW_2^2(\mu_q, \nu_q).
\end{aligned}
$$

Hence, we obtain the conclusion of this proposition. □

As a consequence, the original problem of approximating the poly$-GSW$ distance between μ and ν boils down to estimating the SW distance between μ_q and ν_q. Combining this result with Proposition 2.2, we obtain the following bound for the poly$-GSW$ distance between μ and ν.

Theorem 3.3. *For any probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with finite second moments, there exists a universal constant $C > 0$ such that*

$$
\begin{aligned}
&\left| \text{poly} - GSW_2(\mu, \nu) - q^{-\frac{1}{2}} \left| \sqrt{\mathfrak{m}_2(\mu_q)} - \sqrt{\mathfrak{m}_2(\nu_q)} \right| \right| \\
&\qquad\qquad \leq C(\Xi_q(\mu_q) + \Xi_q(\nu_q))^{\frac{1}{2}},
\end{aligned}
$$

where $\mathfrak{m}_2(\zeta)$ and $\Xi_q(\zeta)$ are defined as in equation (1) for $\zeta \in \{\mu_q, \nu_q\}$.

It is observed that components in the above approximation error, which are $\Xi_q(\mu_q)$ and $\Xi_q(\nu_q)$, cannot be shown to converge to 0 as $d \to \infty$ unless μ_q and ν_q are centered. We overcome this issue by using the following equality:

$$
\begin{aligned}
\text{poly} - GSW_2^2(\mu, \nu) = &\text{poly} - GSW_2^2(\bar{\mu}, \bar{\nu}) \\
&+ q^{-1} \|m_{\mu_q} - m_{\nu_q}\|^2,
\end{aligned}
$$

where $m_\zeta, \bar{\zeta}$ are mean and centered versions of ζ for $\zeta \in \{\mu_q, \nu_q\}$. This equality is achieved by putting Proposition 2.3 and Proposition 3.2 together. Thus, we firstly try to approximate poly$-GSW_2^2(\bar{\mu}, \bar{\nu})$. It is sufficient to estimate $m_{\mu_q}, m_{\nu_q}, \mathfrak{m}_2(\bar{\mu}_q) = \mathbb{E}[\|\boldsymbol{U}\|^2] - \|\mathbb{E}[\boldsymbol{U}]\|^2$, and $\mathfrak{m}_2(\bar{\nu}_q) = \mathbb{E}[\|\boldsymbol{V}\|^2] - \|\mathbb{E}[\boldsymbol{V}]\|^2$. Let $\{\boldsymbol{x}^{(j)}\}_{j=1}^N$ and $\{\boldsymbol{y}^{(j)}\}_{j=1}^N$ be samples drawn from probability distributions μ and ν, respectively. Denote $\boldsymbol{U}(\boldsymbol{x}) = (\boldsymbol{U}_i(\boldsymbol{x}))_{i=1}^q := (\boldsymbol{x}^{\boldsymbol{\alpha}})_{|\boldsymbol{\alpha}|=m}$, and $\boldsymbol{V}(\boldsymbol{y}) = (\boldsymbol{V}_i(\boldsymbol{y}))_{i=1}^q := (\boldsymbol{y}^{\boldsymbol{\alpha}})_{|\boldsymbol{\alpha}|=m}$, be two q-dimensional vectors. Then, the estimators of $\mathfrak{m}_2(\bar{\mu}_q), \mathfrak{m}_2(\bar{\nu}_q), m_{\mu_q}$ and

m_{ν_q} can be calculated as:

$$
\widehat{\mathfrak{m}}_2(\bar{\mu}_q) = \sum_{i=1}^q \left(\frac{\sum_{j=1}^N \boldsymbol{U}_i^2(\boldsymbol{x}^{(j)})}{N} - \left(\frac{\sum_{j=1}^N \boldsymbol{U}_i(\boldsymbol{x}^{(j)})}{N} \right)^2 \right),
$$

$$
\widehat{\mathfrak{m}}_2(\bar{\nu}_q) = \sum_{i=1}^q \left(\frac{\sum_{j=1}^N \boldsymbol{V}_i^2(\boldsymbol{y}^{(j)})}{N} - \left(\frac{\sum_{j=1}^N \boldsymbol{V}_i(\boldsymbol{y}^{(j)})}{N} \right)^2 \right),
$$

$$
\widehat{m}_{\mu_q} = \frac{1}{N} \sum_{j=1}^N \boldsymbol{U}(\boldsymbol{x}^{(j)}), \quad \widehat{m}_{\nu_q} = \frac{1}{N} \sum_{j=1}^N \boldsymbol{V}(\boldsymbol{y}^{(j)}).
$$

Corollary 3.4. *Consequently, an approximation of the poly$-GSW_2(\mu, \nu)$ can be written as*

$$
\begin{aligned}
\widehat{\text{poly} - GSW}_2^2(\mu, \nu) = q^{-1} &\left(\sqrt{\widehat{\mathfrak{m}}_2(\bar{\mu}_q)} - \sqrt{\widehat{\mathfrak{m}}_2(\bar{\nu}_q)} \right)^2 \\
&+ q^{-1} \|\widehat{m}_{\mu_q} - \widehat{m}_{\nu_q}\|^2.
\end{aligned}
$$

To validate our approximation of poly$-GSW_2(\mu, \nu)$, we provide in the following theorem an upper bound for the approximation error $(\Xi_q(\mu_q) + \Xi_q(\nu_q))^{\frac{1}{2}}$.

Theorem 3.5. *Let $(X_j)_{j \in \mathbb{N}}$ and $(Y_j)_{j \in \mathbb{N}}$ be sequences of independent random variables in $\mathbb{R}$ with zero means such that $\mathbb{E}[X_j^{4m}] < \infty$ and $\mathbb{E}[Y_j^{4m}] < \infty$ for all $j \in \mathbb{N}$, where m is an odd positive integer. For $d \in \mathbb{N}$, let $\boldsymbol{X} = \{X_j\}_{j=1}^d$ and $\boldsymbol{Y} = \{Y_j\}_{j=1}^d$ and denote by μ, ν the distributions of $\boldsymbol{X}, \boldsymbol{Y}$, respectively, while μ_q, ν_q are defined as above. Then,*

$$
(\Xi_q(\mu_q) + \Xi_q(\nu_q))^{\frac{1}{2}} \leq \mathcal{O}(d^{-\frac{m}{8}}), \tag{7}
$$

Remark 3.6. When $m = 1$, the polynomial defining function reduces to the linear case $g_{\text{poly}}(\boldsymbol{x}, \theta) = \langle \boldsymbol{x}, \theta \rangle$, leading to the fact that poly$-GSW_p(\mu, \nu) = SW_p(\mu, \nu)$ for any $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. Under that setting, the approximation error in Theorem 3.5 approaches 0 at a rate of $\left(d^{-\frac{1}{8}} \right)$, which matches the result provided in [(Nadjahi et al., 2021), Corollary 1].

We now provide the proof of Theorem 3.5.

Proof of Theorem 3.5. Recall that

$$
\begin{aligned}
\Xi_q(\mu_q) = q^{-1} \Big\{ &A(\mu_q) + (\mathfrak{m}_2(\mu_q) B_1(\mu_q))^{1/2} \\
&+ \mathfrak{m}_2(\mu_q)^{1/5} B_2(\mu_q)^{4/5} \Big\}. \tag{8}
\end{aligned}
$$

Thus, it is sufficient to bound $\mathfrak{m}_2(\mu_q), A(\mu_q)$ and $B_k(\mu_q)$ for $k \in \{1, 2\}$. For any $\boldsymbol{\alpha} \in \mathbb{R}^d$ such that $|\boldsymbol{\alpha}| = m$, by using the Hölder's inequality, we have

$$
\begin{aligned}
\mathbb{E}\left[(\boldsymbol{X}^{\boldsymbol{\alpha}})^2\right] &= \mathbb{E}\left[X_1^{2\alpha_1} \ldots X_d^{2\alpha_d}\right] \\
&\leq \mathbb{E}\left[X_1^{2m}\right]^{\alpha_1/m} \ldots \mathbb{E}\left[X_d^{2m}\right]^{\alpha_d/m} \\
&\leq \max_{1 \leq j \leq d} \mathbb{E}[X_j^{2m}],
\end{aligned}
$$

which leads to the following bound for $\mathfrak{m}_2(\mu_q)$:

$$\mathfrak{m}_2(\mu_q) = \sum_{|\boldsymbol{\alpha}|=m} \mathbb{E}[(\boldsymbol{X^\alpha})^2] \leq q \max_{1\leq j\leq d} \mathbb{E}[X_j^{2m}].$$

Denote $\boldsymbol{U} := (\boldsymbol{X^\alpha})_{|\boldsymbol{\alpha}|=m}$ as a q-dimensional vector, then $[A(\mu_q)]^2 = \mathbb{E}^2 \left| \|\boldsymbol{U}\|^2 - \mathbb{E}\|\boldsymbol{U}\|^2 \right|$ is upper bounded by

$$\mathrm{Var}[\|\boldsymbol{U}\|^2] = \mathrm{Var}\left[\sum_{|\boldsymbol{\alpha}|=m} (\boldsymbol{X^\alpha})^2 \right]$$
$$= \sum_{|\boldsymbol{\alpha}|=|\boldsymbol{\beta}|=m} \mathrm{Cov}\left[(\boldsymbol{X^\alpha})^2, (\boldsymbol{X^\beta})^2 \right].$$

Note that if two terms $\boldsymbol{X^\alpha}$ and $\boldsymbol{X^\beta}$ do not share any variables, then they are independent which implies that $\mathrm{Cov}\left[(\boldsymbol{X^\alpha})^2, (\boldsymbol{X^\beta})^2 \right] = 0$. Otherwise, by denoting $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ and using the Hölder's inequality, $\left| \mathrm{Cov}\left[(\boldsymbol{X^\alpha})^2, (\boldsymbol{X^\beta})^2 \right] \right|^2$ is bounded by

$$\mathbb{E}\left[(\boldsymbol{X^\alpha})^4 \right] \mathbb{E}\left[\left(\boldsymbol{X^\beta}\right)^4 \right]$$
$$= \mathbb{E}\left[X_1^{4\alpha_1} \ldots X_d^{4\alpha_d} \right] \mathbb{E}\left[X_1^{4\beta_1} \ldots X_d^{4\beta_d} \right]$$
$$\leq \left(\mathbb{E}\left[X_1^{4m} \right] \right)^{\alpha_1/m} \ldots \left(\mathbb{E}\left[X_d^{4m} \right] \right)^{\alpha_d/m}$$
$$\times \left(\mathbb{E}\left[X_1^{4m} \right] \right)^{\beta_1/m} \ldots \left(\mathbb{E}\left[X_d^{4m} \right] \right)^{\beta_d/m}$$
$$\leq \max_{1\leq i\leq d} \left(\mathbb{E}\left[X_i^{4m} \right] \right)^2.$$

Next, applying part (ii) of the Lemma 3.7 (cf. the end of this proof), we obtain that $[A(\mu_q)]^2$ is upper bounded by

$$\left(\frac{1}{[(m-1)!]^2} d^{2m-1} + \mathcal{O}(d^{2m-2}) \right) \max_{1\leq j\leq d} \mathbb{E}[X_j^{4m}].$$

Finally, we need to bound $B_1(\mu_q)$ and $B_2(\mu_q)$. By utilizing the Cauchy-Schwarz inequality, we have $B_1(\mu_q) \leq B_2(\mu_q)$. Thus, it is sufficient to bound $B_2(\mu_q)$. Let us denote by $\boldsymbol{X}'$ an independent copy of $\boldsymbol{X}$ and $\boldsymbol{U}' := (\boldsymbol{X'^\alpha})_{|\boldsymbol{\alpha}|=m}$, then

$$\langle \boldsymbol{U}, \boldsymbol{U}' \rangle^2 = \left(\sum_{|\boldsymbol{\alpha}|=m} \boldsymbol{X^\alpha X'^\alpha} \right)^2$$
$$= \sum_{|\boldsymbol{\alpha}|=m} (\boldsymbol{X^\alpha X'^\alpha})^2 + \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \boldsymbol{X^\alpha X^\beta X'^\alpha X'^\beta}. \quad (9)$$

Noting that $\mathbb{E}[\boldsymbol{X^\alpha X^\beta}] = 0$ if this product contains a variable of degree 1. Otherwise, using the Hölder inequality as in the above calculation, we have $\mathbb{E}[\boldsymbol{X^\alpha X^\beta}] \leq \max_{1\leq j\leq d} \mathbb{E}[X_j^{2m}]$ by the Hölder inequality. Thus, taking the expectation of equation (9), and using part (iii) of

Lemma 3.7 (cf. the end of this proof), we have

$$\mathbb{E}\left[\langle \boldsymbol{U}, \boldsymbol{U}' \rangle^2 \right] \leq \mathcal{O}(d^m) \left(\max_{1\leq j\leq d} \mathbb{E}[X_j^{2m}] \right)^2.$$

Therefore, $B_2(\mu_q) \leq \mathcal{O}(d^{m/2}) \max_{1\leq j\leq d} \mathbb{E}[X_j^{2m}]$. In summary, we get

$$\mathfrak{m}_2(\mu_q) \leq \mathcal{O}(d^m) \max_{1\leq j\leq d} \mathbb{E}[X_j^{2m}],$$
$$A(\mu_q) \leq \mathcal{O}(d^{m-1/2})\left\{ \max_{1\leq j\leq d} \mathbb{E}[X_j^{4m}] \right\}^{1/2},$$
$$B_1(\mu_q) \leq B_2(\mu_q) \leq \mathcal{O}(d^{m/2}) \max_{1\leq j\leq d} \mathbb{E}[X_j^{2m}]. \quad (10)$$

Combining above results with equation (8) with a note that $q^{-1} = \mathcal{O}(d^{-m})$, we obtain that $\Xi_q(\mu_q) \leq \mathcal{O}(d^{-\frac{m}{4}})$. Similarly, we have $\Xi_q(\nu_q) \leq \mathcal{O}(d^{-\frac{m}{4}})$. Hence, we reach the conclusion of the theorem. $\qquad\square$

Lemma 3.7. *For two positive integer numbers $d \in \mathbb{N}$ and $m \in \mathbb{N}$, let P_d^m be the set of all multivariate polynomials of $x_1, \ldots, x_d$ of degree m. Then, the followings hold:*

(i) $|P_d^m| = \binom{d+m-1}{m} = \dfrac{d(d+1)\ldots(d+m-1)}{m!}.$

(ii) Let A_d^m be the set of all pairs of polynomials $P, Q \in P_d^m$ such that there is at least one variable appearing in the two polynomials. Then, there exists two polynomials $P_{A,m}^u$ and $P_{A,m}^\ell$ of variable d with degree $2m-1$ and the leading coefficient $1/((m-1)!)^2$ such that $P_{A,m}^\ell(d) \leq |A_d^m| \leq P_{A,m}^u(d)$ for any $d \in \mathbb{N}$.

(iii) Let B_d^m be the set of all pairs of polynomials $P, Q \in P_d^m$ such that the product of two polynomials in each pair does not contain a variable with degree 1. Then, there exist two polynomials $P_{B,m}^u$ and $P_{B,m}^\ell$ of variable d with degree m such that $P_{B,m}^\ell(d) \leq |B_d^m| \leq P_{B,m}^u(d)$.

The proof of Lemma 3.7 is deferred to Appendix A.2.

4. Neural Network Type Function

Using the polynomial defining function in Section 3 yields an interesting approximation for the generalized sliced-Wasserstein distance, which generalizes the result in [(Nadjahi et al., 2021), Theorem 1] for the linear case. However, the memory complexity of polynomial projections grows exponentially with the dimension of data d and the degree of the polynomial, therefore, restricting their usage in machine learning and deep learning applications. To address that issue, we consider in this section the problem of approximating the generalized sliced-Wasserstein distance equipped with a neural network type defining function.

Prior to introducing the definition of a neural network type defining function, let us present some necessary notations for this section. Firstly, recall that $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^\top$ are random vectors having probability distributions $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, respectively. Then, let n be the number of layers of a neural network, we denote by $\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(n)}$ n random matrices of size $d \times d$ such that they are independent of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and their entries are i.i.d random variables following a zero-mean Gaussian distribution $\mathcal{N}(0, d^{-1})$. Now, we are ready to define a neural network type defining function.

Definition 4.1 (Neural network type defining function). Let $\boldsymbol{x}, \theta$ be two vectors in $\mathbb{R}^d$ and $\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(n)}$ be random matrices defined as above, then a neural network type defining function is given by $g_{\mathsf{neural}}(\boldsymbol{x}, \theta) = \langle \theta, \boldsymbol{\Theta}^{(1)} \ldots \boldsymbol{\Theta}^{(n)} \boldsymbol{x} \rangle$. Accordingly, the generalized sliced-Wasserstein distance with this neural network type defining function is denoted as $\mathsf{neural}-GSW$.

Next, we consider a random vector $\boldsymbol{X}^* = (X_1^*, \ldots, X_d^*)^\top \in \mathbb{R}^d$ (resp. $\boldsymbol{Y}^* = (Y_1^*, \ldots, Y_d^*)^\top$) which is achieved by multiplying n random matrices $\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(n)}$ (each corresponding to a layer of the neural network) and $\boldsymbol{X}$ (resp. $\boldsymbol{Y}$):

$$\boldsymbol{X}^* = \boldsymbol{\Theta}^{(1)} \ldots \boldsymbol{\Theta}^{(n)} \boldsymbol{X}, \quad (11)$$

$$\boldsymbol{Y}^* = \boldsymbol{\Theta}^{(1)} \ldots \boldsymbol{\Theta}^{(n)} \boldsymbol{Y}. \quad (12)$$

Based on the above definitions, we also attain in this section a relation between the neural-GSW distance and SW distance. In particular,

Proposition 4.2. *Let μ^* and ν^* be two probability measures corresponding to the random vectors $\boldsymbol{X}^*$ and $\boldsymbol{Y}^*$ given in equations* (11) *and* (12)*, we obtain*

$$\mathsf{neural}-GSW_2(\mu, \nu) = SW_2(\mu^*, \nu^*).$$

Proof of Proposition 4.2. For $\theta \in \mathbb{R}^d$, we denote $g_{\mathsf{neural}}^\theta : \mathbb{R}^d \to \mathbb{R}$ as a function $x \mapsto g_{\mathsf{neural}}(\boldsymbol{x}, \theta)$. It follows from the definition of $\mathsf{poly}-GSW$ distance that

$$\mathsf{neural} - GSW_2^2(\mu, \nu)$$
$$= \int_{\mathbb{R}^d} W_2^2 \left((g_{\mathsf{neural}}^\theta)_\sharp \mu, (g_{\mathsf{neural}}^\theta)_\sharp \nu \right) d\gamma_d(\theta)$$
$$= \int_{\mathbb{R}^d} W_2^2 (\theta_\sharp^* \mu^*, \theta_\sharp^* \nu^*) d\gamma_d(\theta) := SW_2^2(\mu^*, \nu^*).$$

Thus, we reach the conclusion of this proposition. $\square$

Consequently, deriving an approximation of the $\mathsf{neural}-GSW$ between μ and ν is equivalent to estimating the SW distance between μ^* and ν^*. Putting this result and Proposition 2.2 together, we achieve the following bound for the $\mathsf{neural}-GSW(\mu, \nu)$.

Theorem 4.3. *For any probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with finite second moments, there exists a universal constant $C > 0$ such that*

$$\left| \mathsf{neural}-GSW_2(\mu, \nu) - d^{-\frac{1}{2}} \left| \sqrt{\mathfrak{m}_2(\mu^*)} - \sqrt{\mathfrak{m}_2(\nu^*)} \right| \right|$$
$$\leq C(\Xi_d(\mu^*) + \Xi_d(\nu^*))^{\frac{1}{2}},$$

where $\mathfrak{m}_2(\zeta)$ and $\Xi_d(\zeta)$ are defined as in equation (1) *for $\zeta \in \{\mu^*, \nu^*\}$.*

Subsequently, we estimate the values of $\mathfrak{m}_2(\mu^*)$ and $\mathfrak{m}_2(\nu^*)$. Since $\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(n)}$ are independent random matrices with zero means and they are independent of $\boldsymbol{X}$ and $\boldsymbol{Y}$, it follows from equations (11) and (12) that $\mathbb{E}[\boldsymbol{X}^*] = 0$ and $\mathbb{E}[\boldsymbol{Y}^*] = 0$. In other words, μ^* and ν^* are zero-mean distributions. Therefore, let $\{\boldsymbol{x}^{(j)}\}_{j=1}^N$ and $\{\boldsymbol{y}^{(j)}\}_{j=1}^N$ be samples drawn from μ and ν, respectively, we compute estimations of $\mathfrak{m}_2(\mu^*)$ and $\mathfrak{m}_2(\nu^*)$ as follows:

$$\widehat{\mathfrak{m}}_2(\mu^*) = \frac{1}{N} \sum_{j=1}^N \|\boldsymbol{x}^{(j)}\|^2, \quad \widehat{\mathfrak{m}}_2(\nu^*) = \frac{1}{N} \sum_{j=1}^N \|\boldsymbol{y}^{(j)}\|^2.$$

Corollary 4.4. *As a consequence, an approximation of $\mathsf{neural}-GSW_2(\mu, \nu)$ can be written as*

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

Finally, we provide in the following theorem an upper bound of the approximation error $(\Xi_d(\mu^*) + \Xi_d(\nu^*))^{\frac{1}{2}}$.

Theorem 4.5. *Let $(X_j)_{j\in\mathbb{N}}$ and $(Y_j)_{j\in\mathbb{N}}$ be sequences of independent random variables in $\mathbb{R}$ with zero means such that $\mathbb{E}[X_j^4] < \infty$ and $\mathbb{E}[Y_j^4] < \infty$ for all $j \in \mathbb{N}$. For $d \in \mathbb{N}$, let $\boldsymbol{X} = \{X_j\}_{j=1}^d$ and $\boldsymbol{Y} = \{Y_j\}_{j=1}^d$ and denote by μ, ν the distributions of $\boldsymbol{X}, \boldsymbol{Y}$, respectively, while μ^*, ν^* are defined as above. Then, we have*

$$(\Xi_d(\mu^*) + \Xi_d(\nu^*))^{\frac{1}{2}} \leq \mathcal{O}(3^{\frac{n}{4}} d^{-\frac{1}{4}} + d^{-\frac{1}{8}}), \quad (13)$$

where $n \in \mathbb{N}$ is the number of neural network layers.

When there are no layers, i.e. $n = 0$, the neural defining function reduces to the classic case $g_{\mathsf{neural}}(\boldsymbol{x}, \theta) = \langle \boldsymbol{x}, \theta \rangle$, implying that $\mathsf{neural}-GSW_p(\mu, \nu) = SW_p(\mu, \nu)$ for any $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. Additionally, the approximation error in Theorem 4.5 goes to 0 at the same rate as in [(Nadjahi et al., 2021), Corollary 1], which is of $\mathcal{O}(d^{-\frac{1}{8}})$.

Proof sketch of Theorem 4.5. The full proof of Theorem 4.5 is in Appendix A.1. From the definition of $\Xi_d(\mu^*)$ in equation (1), it is sufficient to bound $\mathfrak{m}_2(\mu^*)$, $A(\mu^*)$ and $B_k(\mu^*)$ for $k \in \{1, 2\}$. Thus, this proof is divided into three parts :

Bounding $\mathfrak{m}_2(\mu^*)$: As $\boldsymbol{X}$ and $(\boldsymbol{\Theta}^{(i)})_{i=1}^n$ are independent,

$$\mathfrak{m}_2(\mu^*) = \sum_{j_0=1}^d \mathbb{E}\left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta^{(1)}_{j_0 j_1} \cdots \Theta^{(n)}_{j_{n-1} j_n} X_{j_n}\right)^2$$

$$= \sum_{j_0=1}^d \sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \prod_{i=1}^n \mathbb{E}\left[(\Theta^{(i)}_{j_{i-1} j_i})^2\right] \mathbb{E}[X_{j_n}^2]$$

$$= \sum_{j_n=1}^d \mathbb{E}[X_{j_n}^2] \le d \max_{1 \le j \le d} \mathbb{E}[X_j^2] = \mathcal{O}(d). \quad (14)$$

Bounding $A(\mu^*)$: It follows from the definition of $A(\mu^*)$ in equation (1) that $[A(\mu^*)]^2 \le \mathrm{Var}(\|\boldsymbol{X}^*\|^2)$. Thus, $[A(\mu^*)]^2$ is upper bounded by

$$\mathrm{Var}\left[\sum_{j_0=1}^d \left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta^{(1)}_{j_0 j_1} \cdots \Theta^{(n)}_{j_{n-1} j_n} X_{j_n}\right)^2\right]$$

$$= d \mathrm{Var}\left[\left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta^{(1)}_{1 j_1} \cdots \Theta^{(n)}_{j_{n-1} j_n} X_{j_n}\right)^2\right]$$

$$+ d(d-1) \mathrm{Cov}\left[\left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta^{(1)}_{1 j_1} \cdots \Theta^{(n)}_{j_{n-1} j_n} X_{j_n}\right)^2,\right.$$

$$\left.\left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta^{(1)}_{2 j_1} \cdots \Theta^{(n)}_{j_{n-1} j_n} X_{j_n}\right)^2\right]. \quad (15)$$

Regarding the variance term in equation (15): Since $\boldsymbol{X}$ and $(\boldsymbol{\Theta}^{(i)})_{i=1}^n$ are independent, this term is equal to

$$\sum \mathrm{Cov}\left[\Theta^{(1)}_{j_0^1 j_1^1} \cdots \Theta^{(n)}_{j_{n-1}^1 j_n^1} X_{j_n^1} \Theta^{(1)}_{j_0^2 j_1^2} \cdots \Theta^{(n)}_{j_{n-1}^2 j_n^2} X_{j_n^2},\right.$$

$$\left.\Theta^{(1)}_{j_0^3 j_1^3} \cdots \Theta^{(n)}_{j_{n-1}^3 j_n^3} X_{j_n^3} \Theta^{(1)}_{j_0^4 j_1^4} \cdots \Theta^{(n)}_{j_{n-1}^4 j_n^4} X_{j_n^4}\right], \quad (16)$$

where the sum is subject to tuples $(\boldsymbol{j}^1, \boldsymbol{j}^2, \boldsymbol{j}^3, \boldsymbol{j}^4)$ with $\boldsymbol{j}^i = (j_0^i, \ldots, j_n^i)$ such that $j_0^i = 1$. It is easy to see that if $(\boldsymbol{j}^1, \boldsymbol{j}^2, \boldsymbol{j}^3, \boldsymbol{j}^4)$ is not a reducible tuple (defined in Lemma 4.6), the respected covariance value in equation (16) equals 0. Thus, by utilizing part (i) of Lemma 4.6 (cf. the end of this proof), the variance term in equation (15) is bounded by $\mathcal{O}(3^n)$.

Regarding the covariance term in equation (15): This term is equal to

$$\sum \mathrm{Cov}\left[\Theta^{(1)}_{j_0^1 j_1^1} \cdots \Theta^{(n)}_{j_{n-1}^1 j_n^1} X_{j_n^1} \Theta^{(1)}_{j_0^2 j_1^2} \cdots \Theta^{(n)}_{j_{n-1}^2 j_n^2} X_{j_n^2},\right.$$

$$\left.\Theta^{(1)}_{j_0^3 j_1^3} \cdots \Theta^{(n)}_{j_{n-1}^3 j_n^3} X_{j_n^3} \Theta^{(1)}_{j_0^4 j_1^4} \cdots \Theta^{(n)}_{j_{n-1}^4 j_n^4} X_{j_n^4}\right], \quad (17)$$

where the sum is subject to tuples $(\boldsymbol{j}^1, \boldsymbol{j}^2, \boldsymbol{j}^3, \boldsymbol{j}^4)$ with $\boldsymbol{j}^i = (j_0^i, \ldots, j_n^i)$ such that $j_0^1 = j_0^2 = 1$, $j_0^3 = j_0^4 = 2$. It is

obvious that if $(\boldsymbol{j}^1, \boldsymbol{j}^2, \boldsymbol{j}^3, \boldsymbol{j}^4)$ is not a reducible tuple as in the definition of Lemma 4.6 (cf. the end of this proof), the respected value of the covariance term in equation (17) equals 0. Therefore, by using part (ii) of Lemma 4.6, the covariance term in equation (15) is bounded by $\mathcal{O}(3^n)/d$.

Putting the above results together, we obtain that

$$A(\mu^*) \le \mathcal{O}(3^{\frac{n}{2}} d^{\frac{1}{2}}). \quad (18)$$

Bounding $B_k(\mu^*)$: By the Cauchy-Schwartz inequality, we have $B_1(\mu^*) \le B_2(\mu^*)$. Therefore, it is sufficient to bound $B_2(\mu^*)$. Let $\boldsymbol{X}^{*\prime}$ be an independent copy of $\boldsymbol{X}^*$, then

$$\langle \boldsymbol{X}^*, \boldsymbol{X}^{*\prime}\rangle^2 = \sum_{j=1}^d (X_j^*)^2 (X_j^{*\prime})^2 + \sum_{i \ne j} X_i^* X_j^* X_i^{*\prime} X_j^{*\prime}.$$

As $\boldsymbol{X}^*$ and $\boldsymbol{X}^{*\prime}$ are independent, we have

$$\mathbb{E}\left[(X_j^*)^2 (X_j^{*\prime})^2\right] = (\mathbb{E}[(X_j^*)^2])^2$$

$$= \left(\mathbb{E}\left[\left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta^{(1)}_{j j_1} \cdots \Theta^{(n)}_{j_{n-1} j_n} X_{j_n}\right)^2\right]\right)^2$$

$$= \left(\mathbb{E}\left[\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \left(\Theta^{(1)}_{j j_1}\right)^2 \cdots \left(\Theta^{(n)}_{j_{n-1} j_n}\right)^2 X_{j_n}^2\right]\right)^2$$

$$\le \left(d^n \max_{1 \le j \le d} \frac{1}{d^n} \mathbb{E}[X_j^2]\right)^2 = (\max_{1 \le j \le d} \mathbb{E}[(X_j^*)^2])^2,$$

and $\mathbb{E}[X_i^* X_j^* X_i^{*\prime} X_j^{*\prime}] = (\mathbb{E}[X_i^* X_j^*])^2 = 0$ for $i \ne j$, Thus, we have

$$B_1(\mu^*) \le B_2(\mu^*) \le \mathcal{O}(d^{\frac{1}{2}}). \quad (19)$$

From equations (14), (18) and (19), we obtain $\Xi_d(\mu^*) \le \mathcal{O}(3^{\frac{n}{2}} d^{-\frac{1}{2}} + d^{-\frac{1}{4}})$. Similarly, we have $\Xi_d(\nu^*) \le \mathcal{O}(3^{\frac{n}{2}} d^{-\frac{1}{2}} + d^{-\frac{1}{4}})$. Hence, we conclude the theorem. $\square$

Lemma 4.6. *Let us consider tuples* $((j_0^1, j_1^1, \ldots, j_n^1), (j_0^2, j_1^2, \ldots, j_n^2), (j_0^3, j_1^3, \ldots, j_n^3), (j_0^4, j_1^4, \ldots, j_n^4))$, *where j_i^k is a positive integer number in $[1, d]$.*

*We call a tuple to be **reducible** if for each $i \in \{0, 1 \ldots, n-1\}$, there exists a way to divide the set $\{1, 2, 3, 4\}$ into two disjoint subsets $\{p, q\}$ and $\{r, s\}$ such that $j_i^p = j_i^q$, $j_{i+1}^p = j_{i+1}^q$, $j_i^r = j_i^s$, $j_{i+1}^r = j_{i+1}^s$. Then,*

(i) *The number of reducible tuples such that $j_0^1 = j_0^2 = j_0^3 = j_0^4 = 1$ is a 2n-degree polynomial of variable d.*

(ii) *The number of reducible tuples such that $j_0^1 = j_0^2 = 1, j_0^3 = j_0^4 = 2$ is a polynomial of variable d of degree $2n$ with the highest coefficient of 1.*

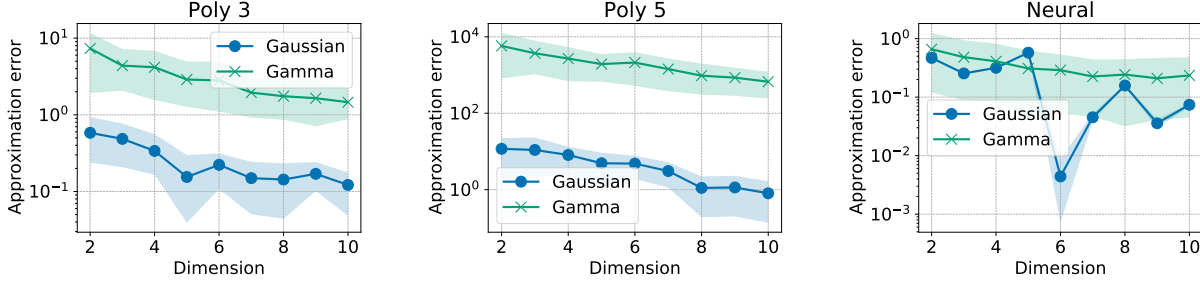The proof of Lemma 4.6 is deferred to Appendix A.3.

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.
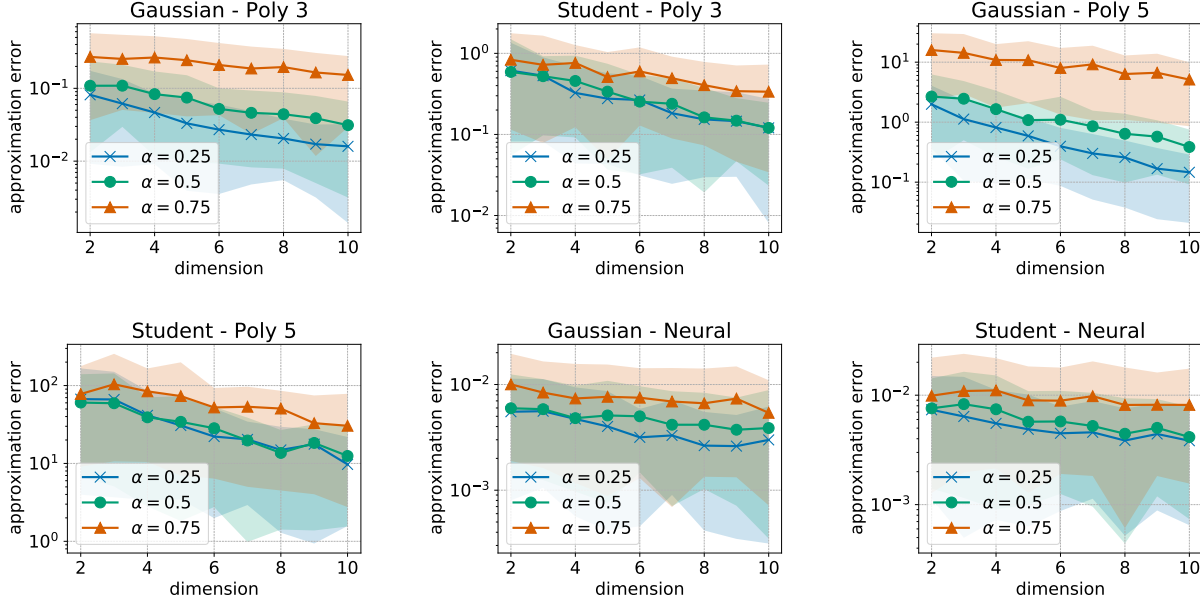

Figure 2. Approximation error between approximated GSW with the Monte Carlo GSW with a huge number of projections between empirical distributions on samples that are drawn from autoregressive processes of order one AR(1).

5. Experiments

In this section, we focus on testing the approximation error of our proposed approximated GSW. In particular, we try to increase the dimension of simulated data to see the change of the approximation error, namely, the $\mathbb{L}_1$ distance between the approximated GSW and the Monte Carlo GSW with a huge number of projections, e.g., 20000. For all experiments, we repeat the process 100 times and report the mean and the standard deviation each time.

Approximation error on Multivariate Gaussian and Gamma: In this setup, we first generate two sets of $n = 10^4$ d-dimensional samples from two Multivariate Gaussian distributions $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\mathcal{N}(\mathbf{1}, 2\mathbf{I}_d)$. We denote two empirical distributions as $\mu_d = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ and $\nu_d = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$. We then compute our approximated GSW and the Monte Carlo GSW with the polynomial defining function (degree 3 and 5) and the neural defining function. Finally, we plot the approximation error with respect to the number of dimensions in Figure 1. From the figure, we observe that the

approximation error has a decreasing trend when the number of dimensions increases for all defining functions. A similar phenomenon happens when we use empirical samples from multivariate random variables where each dimension follows $\mathrm{Gamma}(1, 2)$ and $\mathrm{Gamma}(1, 3)$. We also observe that the error in the Gamma case is larger than in the Gaussian case. The reason is that our approximation is based on the closed form of Wasserstein between two Gaussians.

Approximation error on autoregressive processes of order one (AR(1)): We would like to recall that in AR(1) process, $X_t = \alpha X_{t-1} + \varepsilon_t$ where $\alpha \in [0, 1]$ and $\{\varepsilon_i\}_{i=1}^{n}$ are i.i.d. real random variables with $\mathbb{E}[\varepsilon_i] = 0$ that have finite second-order moment. We use this process with $10^4 + d$ to generate samples. We only take the last d steps while the previous steps are for "burn in" that guarantees the stationary solution of the process. We generate empirical samples $\{x_i\}_{i=1}^{n}$ and $\{y_i\}_{i=1}^{n}$ using the same Gaussian noise (Student noise). We report the approximation errors for different values of α and defining functions in Figure 2. Similar to the previous experiment, the approximation error decreases

when the number of dimensions increases.

6. Conclusion

In this paper, we establish deterministic and fast approximations of the generalized sliced-Wasserstein distance without using random projections by leveraging the conditional central limit theorem for Gaussian projections. In both cases of polynomial defining function and neural network type function, we provide a rigorous guarantee that under some mild assumptions on two input probability measures, the approximation errors approach zero when the dimension increases. Finally, we carry out some simulation studies on different types of probability distributions to justify our theoretical results.

References

Beylkin, G. The inversion problem and applications of the generalized radon transform. *Communications on pure and applied mathematics*, 37(5):579–599, 1984.

Bobkov, S. and Ledoux, M. 'One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Memoirs of the American Mathematical Society*, 261, 2019.

Bonet, C., Berg, P., Courty, N., Septier, F., Drumetz, L., and Pham, M.-T. Spherical sliced-wasserstein. *arXiv preprint arXiv:2206.08780*, 2022.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 1(51):22–45, 2015.

Chen, X., Yang, Y., and Li, Y. Augmented sliced Wasserstein distances. *International Conference on Learning Representations*, 2022.

Dai, B. and Seljak, U. Sliced iterative normalizing flows. In *International Conference on Machine Learning*, pp. 2352–2364. PMLR, 2021.

Deshpande, I., Zhang, Z., and Schwing, A. G. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3483–3491, 2018.

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Diaconis, P. and Freedman, D. Asymptotics of graphical projection pursuit. *The annals of statistics*, pp. 793–815, 1984.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015.

Homan, A. and Zhou, H. Injectivity and stability for a generic class of generalized radon transforms. *The Journal of Geometric Analysis*, 27(2):1515–1529, 2017.

Kolouri, S., Rohde, G. K., and Hoffmann, H. Sliced Wasserstein distance for learning Gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3427–3436, 2018.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, pp. 261–272, 2019.

Nadjahi, K., De Bortoli, V., Durmus, A., Badeau, R., and Şimşekli, U. Approximate Bayesian computation with the sliced-Wasserstein distance. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5470–5474. IEEE, 2020a.

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Simsekli, U. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020b.

Nadjahi, K., Durmus, A., Jacob, P. E., Badeau, R., and Simsekli, U. Fast approximation of the sliced-Wasserstein distance using concentration of random projections. *Advances in Neural Information Processing Systems*, 34, 2021.

Nguyen, K. and Ho, N. Amortized projection optimization for sliced Wasserstein generative models. *Advances in Neural Information Processing Systems*, 2022a.

Nguyen, K. and Ho, N. Revisiting sliced wasserstein on images: From vectorization to convolution. *Advances in Neural Information Processing Systems*, 2022b.

Nguyen, K., Ho, N., Pham, T., and Bui, H. Distributional sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=QYjO70ACDK.

Nguyen, K., Nguyen, S., Ho, N., Pham, T., and Bui, H. Improving relational regularized autoencoders with spherical sliced fused Gromov-Wasserstein. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=DiQD7FWL233.

Nguyen, K., Ren, T., Nguyen, H., Rout, L., Nguyen, T. M., and Ho, N. Hierarchical sliced wasserstein distance. In *The Eleventh International Conference on Learning Representations*, 2023.

Pele, O. and Werman, M. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 460–467. IEEE, September 2009.

Reeves, G. Conditional central limit theorems for gaussian projections. In *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017.

Sudakov, V. N. Typical distributions of linear functionals in finite-dimensional spaces of higher dimension. In *Doklady Akademii Nauk*, volume 243, pp. 1402–1405. Russian Academy of Sciences, 1978.

Villani, C. *Optimal transport: Old and New*. Springer, 2008.

Wu, J., Huang, Z., Acharya, D., Li, W., Thoma, J., Paudel, D. P., and Gool, L. V. Sliced Wasserstein generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3713–3722, 2019.

Yi, M. and Liu, S. Sliced Wasserstein variational inference. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2021.

Supplementary Materials for
"Fast Approximation of the Generalized Sliced-Wasserstein Distance"

In this supplementary material, we firstly provide proofs of remaining results in Appendix A. Then, we present the approximation of the generalized sliced-Wasserstein distance when the defining function is circular in Appendix B.

A. Proof of Remaining Results

This appendix is devoted to show the proofs of Theorem 4.5, Lemma 3.7 and Lemma 4.6.

A.1. Proof of Theorem 4.5

Given the definition of $\Xi_d(\mu^*)$ as follows:

$$\Xi_d(\mu^*) = d^{-1}\left\{A(\mu^*) + [\mathfrak{m}_2(\mu^*)B_1(\mu^*)]^{1/2} + \mathfrak{m}_2(\mu^*)^{1/5}B_2(\mu^*)^{4/5}\right\}, \tag{20}$$

this proof consists of bounding $\mathfrak{m}_2(\mu^*)$, $A(\mu^*)$, and $B_k(\mu^*)$ for $k \in \{1, 2\}$.

An upper bound of $\mathfrak{m}_2(\mu^*)$: Firstly, let us recall the definition of $\mathfrak{m}_2(\mu^*)$:

$$\mathfrak{m}_2(\mu^*) = \mathbb{E}[\|\boldsymbol{X}^*\|^2] = \mathbb{E}[\|\boldsymbol{\Theta}^{(1)} \ldots \boldsymbol{\Theta}^{(n)}\boldsymbol{X}\|^2] = \sum_{j_0=1}^{d} \mathbb{E}\left[\left(\sum_{j_1=1}^{d} \cdots \sum_{j_n=1}^{d} \Theta_{j_0 j_1}^{(1)} \cdots \Theta_{j_{n-1} j_n}^{(n)} X_{j_n}\right)^2\right].$$

Since $\boldsymbol{X}$ and $\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(n)}$ are independent, we have

$$\begin{aligned}
\mathfrak{m}_2(\mu^*) &= \sum_{j_0=1}^{d} \mathbb{E}\left[\sum_{j_1=1}^{d} \cdots \sum_{j_n=1}^{d} (\Theta_{j_0 j_1}^{(1)})^2 \cdots (\Theta_{j_{n-1} j_n}^{(n)})^2 X_{j_n}^2\right] \\
&= \sum_{j_0=1}^{d} \sum_{j_1=1}^{d} \cdots \sum_{j_n=1}^{d} \prod_{i=1}^{n} \mathbb{E}\left[(\Theta_{j_{i-1} j_i}^{(i)})^2\right] \mathbb{E}[X_{j_n}^2] \\
&= d^n \times \frac{1}{d^n} \sum_{j_n=1}^{d} \mathbb{E}[X_{j_n}^2] \\
&= \sum_{j_n=1}^{d} \mathbb{E}[X_{j_n}^2] \\
&\leq d \max_{1 \leq j \leq d} \mathbb{E}[X_j^2] = \mathcal{O}(d).
\end{aligned}$$

An upper bound of $A(\mu^*)$: It is worth noting that

$$[A(\mu^*)]^2 = \left[\mathbb{E}\left|\|\boldsymbol{X}^*\|^2 - \mathbb{E}\|\boldsymbol{X}^*\|^2\right|\right]^2 \leq \mathbb{E}\left[\|\boldsymbol{X}^*\|^2 - \mathbb{E}\|\boldsymbol{X}^*\|^2\right]^2 = \mathrm{Var}(\|\boldsymbol{X}^*\|^2).$$

As a result, $A(\mu^*)$ can be bounded as follows:

$$[A(\mu^*)]^2 \leq \text{Var}\left[\sum_{i=1}^d (X_i^*)^2\right] = \text{Var}\left[\sum_{j_0=1}^d \left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta_{j_0 j_1}^{(1)} \cdots \Theta_{j_{n-1} j_n}^{(n)} X_{j_n}\right)^2\right]$$

$$= \sum_{j_0=1}^d \text{Var}\left[\left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta_{j_0 j_1}^{(1)} \cdots \Theta_{j_{n-1} j_n}^{(n)} X_{j_n}\right)^2\right]$$

$$+ 2 \sum_{1 \leq j_0 < j_0' \leq d} \text{Cov}\left[\left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta_{j_0 j_1}^{(1)} \cdots \Theta_{j_{n-1} j_n}^{(n)} X_{j_n}\right)^2, \left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta_{j_0' j_1}^{(1)} \cdots \Theta_{j_{n-1} j_n}^{(n)} X_{j_n}\right)^2\right]$$

$$= d\,\text{Var}\left[\left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta_{1 j_1}^{(1)} \cdots \Theta_{j_{n-1} j_n}^{(n)} X_{j_n}\right)^2\right]$$

$$+ d(d-1)\,\text{Cov}\left[\left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta_{1 j_1}^{(1)} \cdots \Theta_{j_{n-1} j_n}^{(n)} X_{j_n}\right)^2, \left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta_{2 j_1}^{(1)} \cdots \Theta_{j_{n-1} j_n}^{(n)} X_{j_n}\right)^2\right], \quad (21)$$

where the last equality is due to the fact that $\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(n)}$ are i.i.d random matrices.

We now try to bound the variance term in equation (21). As $\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(n)}$ are independent of $\boldsymbol{X}^*$, we get

$$\text{Var}\left[\left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta_{1 j_1}^{(1)} \cdots \Theta_{j_{n-1} j_n}^{(n)} X_{j_n}\right)^2\right]$$

$$= \sum_{(\boldsymbol{j}^1, \boldsymbol{j}^2, \boldsymbol{j}^3, \boldsymbol{j}^4)} \text{Cov}\left[\Theta_{j_0^1 j_1^1}^{(1)} \cdots \Theta_{j_{n-1}^1 j_n^1}^{(n)} X_{j_n^1} \Theta_{j_0^2 j_1^2}^{(1)} \cdots \Theta_{j_{n-1}^2 j_n^2}^{(n)} X_{j_n^2}, \Theta_{j_0^3 j_1^3}^{(1)} \cdots \Theta_{j_{n-1}^3 j_n^3}^{(n)} X_{j_n^3} \Theta_{j_0^4 j_1^4}^{(1)} \cdots \Theta_{j_{n-1}^4 j_n^4}^{(n)} X_{j_n^4}\right], \quad (22)$$

where $(\boldsymbol{j}^1, \boldsymbol{j}^2, \boldsymbol{j}^3, \boldsymbol{j}^4)$ is a tuple of $\boldsymbol{j}^i = (j_0^i, \ldots, j_n^i)$ such that $j_0^i = 1$. It can be seen that if $(\boldsymbol{j}^1, \boldsymbol{j}^2, \boldsymbol{j}^3, \boldsymbol{j}^4)$ is not a reducible tuple as in the definition of Lemma 4.6, the respected covariance value in the sum in equation (22) is equal to 0. However, when $(\boldsymbol{j}^1, \boldsymbol{j}^2, \boldsymbol{j}^3, \boldsymbol{j}^4)$ is a reducible tuple, we have

$$\text{Cov}\left[\Theta_{j_0^1 j_1^1}^{(1)} \cdots \Theta_{j_{n-1}^1 j_n^1}^{(n)} X_{j_n^1} \Theta_{j_0^2 j_1^2}^{(1)} \cdots \Theta_{j_{n-1}^2 j_n^2}^{(n)} X_{j_n^2}, \Theta_{j_0^3 j_1^3}^{(1)} \cdots \Theta_{j_{n-1}^3 j_n^3}^{(n)} X_{j_n^3} \Theta_{j_0^4 j_1^4}^{(1)} \cdots \Theta_{j_{n-1}^4 j_n^4}^{(n)} X_{j_n^4}\right]$$

$$\leq \left(\mathbb{E}\left[\left(\Theta_{j_0^1 j_1^1}^{(1)} \cdots \Theta_{j_{n-1}^1 j_n^1}^{(n)} X_{j_n^1} \Theta_{j_0^2 j_1^2}^{(1)} \cdots \Theta_{j_{n-1}^2 j_n^2}^{(n)} X_{j_n^2}\right)^2\right]\right)^{1/2} \left(\mathbb{E}\left[\left(\Theta_{j_0^3 j_1^3}^{(1)} \cdots \Theta_{j_{n-1}^3 j_n^3}^{(n)} X_{j_n^3} \Theta_{j_0^4 j_1^4}^{(1)} \cdots \Theta_{j_{n-1}^4 j_n^4}^{(n)} X_{j_n^4}\right)^2\right]\right)^{1/2}$$

$$= \left(\prod_{i=1}^n \left(\mathbb{E}\left[\left(\Theta_{j_{i-1}^1 j_i^1}^{(i)} \Theta_{j_{i-1}^2 j_i^2}^{(i)}\right)^2\right] \mathbb{E}\left[\left(\Theta_{j_{i-1}^3 j_i^3}^{(i)} \Theta_{j_{i-1}^4 j_i^4}^{(i)}\right)^2\right]\right)^{1/2}\right) \left(\mathbb{E}\left[X_{j_n^1}^2 X_{j_n^2}^2\right] \mathbb{E}\left[X_{j_n^3}^2 X_{j_n^4}^2\right]\right)^{1/2}$$

$$\leq \left(\prod_{i=1}^n \left(\mathbb{E}\left[(\Theta_{j_{i-1}^1 j_i^1}^{(i)})^4\right] \mathbb{E}\left[(\Theta_{j_{i-1}^2 j_i^2}^{(i)})^4\right] \mathbb{E}\left[(\Theta_{j_{i-1}^3 j_i^3}^{(i)})^4\right] \mathbb{E}\left[(\Theta_{j_{i-1}^4 j_i^4}^{(i)})^4\right]\right)^{1/4}\right) \left(\mathbb{E}\left[X_{j_n^1}^4\right] \mathbb{E}\left[X_{j_n^2}^4\right] \mathbb{E}\left[X_{j_n^3}^4\right] \mathbb{E}\left[X_{j_n^4}^4\right]\right)^{1/4}$$

$$\leq \frac{3^n}{d^{2n}} \max_{1 \leq j \leq d} \mathbb{E}\left[X_j^4\right]. \quad (23)$$

Using part (i) of Lemma 4.6 about the number of reducible tuples, we obtain

$$\text{Var}\left[\left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta_{1 j_1}^{(1)} \cdots \Theta_{j_{n-1} j_n}^{(n)} X_{j_n}\right)^2\right] \leq \mathcal{O}(3^n) \max_{1 \leq j \leq d} \mathbb{E}\left[X_j^4\right]. \quad (24)$$

Subsequently, we will bound the covariance term in equation (21).

$$\mathrm{Cov}\left[\left(\sum_{j_1=1}^{d}\cdots\sum_{j_n=1}^{d}\Theta_{1j_1}^{(1)}\cdots\Theta_{j_{n-1}j_n}^{(n)}X_{j_n}\right)^2,\left(\sum_{j_1=1}^{d}\cdots\sum_{j_n=1}^{d}\Theta_{2j_1}^{(1)}\cdots\Theta_{j_{n-1}j_n}^{(n)}X_{j_n}\right)^2\right]$$

$$=\sum_{(\boldsymbol{j}^1,\boldsymbol{j}^2,\boldsymbol{j}^3,\boldsymbol{j}^4)}\mathrm{Cov}\left[\Theta_{j_0^1j_1^1}^{(1)}\cdots\Theta_{j_{n-1}^1j_n^1}^{(n)}X_{j_n^1}\Theta_{j_0^2j_1^2}^{(1)}\cdots\Theta_{j_{n-1}^2j_n^2}^{(n)}X_{j_n^2},\Theta_{j_0^3j_1^3}^{(1)}\cdots\Theta_{j_{n-1}^3j_n^3}^{(n)}X_{j_n^3}\Theta_{j_0^4j_1^4}^{(1)}\cdots\Theta_{j_{n-1}^4j_n^4}^{(n)}X_{j_n^4}\right], \quad (25)$$

where $(\boldsymbol{j}^1,\boldsymbol{j}^2,\boldsymbol{j}^3,\boldsymbol{j}^4)$ is a tuple of $\boldsymbol{j}^i=(j_0^i,\ldots,j_n^i)$ such that $j_0^1=j_0^2=1$, $j_0^3=j_0^4=2$. Note that if $(\boldsymbol{j}^1,\boldsymbol{j}^2,\boldsymbol{j}^3,\boldsymbol{j}^4)$ is not a reducible tuple as in the definition of Lemma 4.6, the respected covariance value in the sum in equation (25) is equal to 0. By contrast, when $(\boldsymbol{j}^1,\boldsymbol{j}^2,\boldsymbol{j}^3,\boldsymbol{j}^4)$ is a reducible tuple, note that if for each $\ell\in\{0,1,\ldots,n\}$, we have $j_\ell^1=j_\ell^2$, $j_\ell^3=j_\ell^4$ and $j_\ell^1\neq j_\ell^3$, we have

$$\mathrm{Cov}\left[\Theta_{j_0^1j_1^1}^{(1)}\cdots\Theta_{j_{n-1}^1j_n^1}^{(n)}X_{j_n^1}\Theta_{j_0^2j_1^2}^{(1)}\cdots\Theta_{j_{n-1}^2j_n^2}^{(n)}X_{j_n^2},\Theta_{j_0^3j_1^3}^{(1)}\cdots\Theta_{j_{n-1}^3j_n^3}^{(n)}X_{j_n^3}\Theta_{j_0^4j_1^4}^{(1)}\cdots\Theta_{j_{n-1}^4j_n^4}^{(n)}X_{j_n^4}\right]=0.$$

There are a total of $[d(d-1)]^n$ such tuples. Otherwise, by following the same arguments in the calculation in equation (23), we get

$$\mathrm{Cov}\left[\Theta_{j_0^1j_1^1}^{(1)}\cdots\Theta_{j_{n-1}^1j_n^1}^{(n)}X_{j_n^1}\Theta_{j_0^2j_1^2}^{(1)}\cdots\Theta_{j_{n-1}^2j_n^2}^{(n)}X_{j_n^2},\Theta_{j_0^3j_1^3}^{(1)}\cdots\Theta_{j_{n-1}^3j_n^3}^{(n)}X_{j_n^3}\Theta_{j_0^4j_1^4}^{(1)}\cdots\Theta_{j_{n-1}^4j_n^4}^{(n)}X_{j_n^4}\right]\leq\frac{3^n}{d^{2n}}\max_{1\leq i\leq n}\mathbb{E}\left[X_i^4\right].$$

Thus, using part (ii) of the Lemma 4.6 about the number of reducible tuples, we obtain an upper bound of the covariance term in equation (21)

$$\mathrm{Cov}\left[\left(\sum_{j_1=1}^{d}\cdots\sum_{j_n=1}^{d}\Theta_{1j_1}^{(1)}\cdots\Theta_{j_{n-1}j_n}^{(n)}X_{j_n}\right)^2,\left(\sum_{j_1=1}^{d}\cdots\sum_{j_n=1}^{d}\Theta_{2j_1}^{(1)}\cdots\Theta_{j_{n-1}j_n}^{(n)}X_{j_n}\right)^2\right]\leq\frac{\mathcal{O}(3^n)}{d}\max_{1\leq j\leq d}\mathbb{E}\left[X_j^4\right]. \quad (26)$$

Putting the equations (21), (24), and (26) together, we have $[A(\mu^*)]^2\leq\mathcal{O}(3^nd)\max_{1\leq j\leq d}\mathbb{E}\left[X_j^4\right]$, which implies that

$$A(\mu^*)\leq\mathcal{O}(3^{\frac{n}{2}}d^{\frac{1}{2}}).$$

An upper bound of $B_k(\mu^*)$: Finally, we need to bound $B_k(\mu^*)$ for $k\in\{1,2\}$. By applying the Cauchy-Schwartz inequality, we get $B_1(\mu^*)\leq B_2(\mu^*)$. Thus, it is sufficient to bound $B_2(\mu^*)$. Denote by $\boldsymbol{X}^{*\prime}=(X_1^{*\prime},\ldots,X_d^{*\prime})$ an independent copy of $\boldsymbol{X}^*=(X_1^*,\ldots,X_d^*)$, we consider

$$\langle\boldsymbol{X}^*,\boldsymbol{X}^{*\prime}\rangle^2=\left(\sum_{j=1}^{d}X_j^*X_j^{*\prime}\right)^2=\sum_{j=1}^{d}(X_j^*)^2(X_j^{*\prime})^2+2\sum_{1\leq i<j\leq d}X_i^*X_j^*X_i^{*\prime}X_j^{*\prime}.$$

Since $\boldsymbol{X}^*$ and $\boldsymbol{X}^{*\prime}$ are i.i.d random vectors, we have for any $1\leq j\leq d$ that

$$\mathbb{E}[(X_j^*)^2(X_j^{*\prime})^2]=(\mathbb{E}[(X_j^*)^2])^2=\left(\mathbb{E}\left[\left(\sum_{j_1=1}^{d}\cdots\sum_{j_n=1}^{d}\Theta_{jj_1}^{(1)}\cdots\Theta_{j_{n-1}j_n}^{(n)}X_{j_n}\right)^2\right]\right)^2$$

$$=\left(\mathbb{E}\left[\sum_{j_1=1}^{d}\cdots\sum_{j_n=1}^{d}\left(\Theta_{jj_1}^{(1)}\right)^2\cdots\left(\Theta_{j_{n-1}j_n}^{(n)}\right)^2X_{j_n}^2\right]\right)^2$$

$$\leq\left(d^n\max_{1\leq j\leq d}\frac{1}{d^n}\mathbb{E}[X_j^2]\right)^2=\left(\max_{1\leq j\leq d}\mathbb{E}[X_j^2]\right)^2.$$

Meanwhile, for $1 \leq i \neq j \leq d$, we get

$$\mathbb{E}[X_i^* X_j^* X_i^{*\prime} X_j^{*\prime}] = (\mathbb{E}[X_i^* X_j^*])^2$$

$$= \left(\mathbb{E}\left[\left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta_{ij_1}^{(1)} \cdots \Theta_{j_{n-1}j_n}^{(n)} X_{j_n} \right) \left(\sum_{j_1=1}^d \cdots \sum_{j_n=1}^d \Theta_{jj_1}^{(1)} \cdots \Theta_{j_{n-1}j_n}^{(n)} X_{j_n} \right) \right] \right)^2$$

$$= 0.$$

Consequently, we obtain $\mathbb{E}[\langle \boldsymbol{X}^*, \boldsymbol{X}^{*\prime} \rangle^2] \leq d \max_{1 \leq j \leq d}(\mathbb{E}[X_j^2])^2$, which leads to

$$B_2(\mu^*) \leq d^{\frac{1}{2}} \max_{1 \leq j \leq d} \mathbb{E}[X_j^2] = \mathcal{O}(d^{\frac{1}{2}}).$$

In summary, we have

$$\mathfrak{m}_2(\mu^*) \leq \mathcal{O}(d),$$
$$A(\mu^*) \leq \mathcal{O}(3^{\frac{n}{2}} d^{\frac{1}{2}}),$$
$$B_1(\mu^*) \leq B_2(\mu^*) \leq \mathcal{O}(d^{\frac{1}{2}}).$$

By plugging those results into equation (20), we obtain

$$\Xi_d(\mu^*) \leq \mathcal{O}(3^{\frac{n}{2}} d^{-\frac{1}{2}} + d^{-\frac{1}{4}}).$$

Likewise, we also have $\Xi_d(\nu^*) \leq \mathcal{O}(3^{\frac{n}{2}} d^{-\frac{1}{2}} + d^{-\frac{1}{4}})$. Hence, we reach the conclusion of the theorem, which is

$$(\Xi_d(\mu^*) + \Xi_d(\nu^*))^{\frac{1}{2}} \leq \mathcal{O}(3^{\frac{n}{4}} d^{-\frac{1}{4}} + d^{-\frac{1}{8}}).$$

A.2. Proof of Lemma 3.7

Part (i). Recall that an element of the set P_d^m is of the form $P(x_1, \ldots, x_d) := x_1^{a_1} \ldots x_d^{a_d}$, where $a_i \in \mathbb{N} \cup \{0\}$ for all $i \in \{1, \ldots, d\}$ and $\sum_{i=1}^d a_i = m$. Therefore, the cardinality of P_d^m is equal to the number of non-negative integer solutions of the following equation:

$$a_1 + \ldots + a_d = m.$$

Thus, we obtain that $|P_d^m| = \binom{d+m-1}{m}$.

Part (ii). For any $i \in \{1, \ldots, d\}$, let A_i be the set of all pairs of elements in P_d^m such that both two elements in each pair contain x_i. Then, by using the principle of inclusion-exclusion, we have

$$\sum_{i=1}^d |A_i| - \sum_{1 \leq i < j \leq d} |A_i \cap A_j| \leq |A_d^m| \leq \sum_{i=1}^d |A_i|. \tag{27}$$

Compute $|A_i|$: It is worth noting that for any pair $(P, Q) \in A_i$, we have $P(x_1, \ldots, x_d)/x_i \in P_d^{m-1}$ and $Q(x_1, \ldots, x_d)/x_i \in P_d^{m-1}$. Since P is not necessarily different from Q, then we have

$$|A_i| = |P_d^{m-1}| \times |P_d^{m-1}| = |P_d^{m-1}|^2 = \binom{d+m-2}{m-1}^2. \tag{28}$$

Compute $|A_i \cap A_j|$, $i < j$: Similarly, note that for each $(P, Q) \in A_i \cap A_j$, we have $P(x_1, \ldots, x_d)/(x_i x_j) \in P_d^{m-2}$ and $Q(x_1, \ldots, x_d)/(x_i x_j) \in P_d^{m-2}$. Then, we obtain

$$|A_i \cap A_j| = |P_d^{m-2}| \times |P_d^{m-2}| = |P_d^{m-2}|^2 = \binom{d+m-3}{m-2}^2. \tag{29}$$

Plugging the results in equations (28) and (29) into equation (27), we get

$$d \binom{d+m-2}{m-1}^2 - \frac{d(d-1)}{2} \binom{d+m-3}{m-2}^2 \le |A_d^m| \le d \binom{d+m-2}{m-1}^2.$$

Hence, we achieve the conclusion of this part.

Part (iii). The product of two elements in each pair in B_d^m has the form

$$x_1^{a_1} \ldots x_d^{a_d},$$

where $a_i \neq 1$ is an integer number for any $i \in \{1, \ldots, d\}$, and $a_1 + \ldots + a_d = 2m$. We now count how many tuples $(a_1, \ldots, a_d)$ satisfy those constraints.

- Assume that exactly k elements of $\{a_1, \ldots, a_d\}$ are greater than 1 while others equal 0. There are a total of $\binom{d}{k}$ such tuples.

- Next, we consider a subset of k elements of $\{a_1, \ldots, a_d\}$ such that they are greater than 1 and they sum up to $2m$. Without loss of generality, we can assume that this subset is $\{a_1, \ldots, a_k\}$. Then, we have $a_i \geq 2$ for all $i \in \{1, \ldots, k\}$, and $a_1 + a_2 + \ldots + a_k = 2m$, or equivalently, $(a_1 - 2) + \ldots + (a_k - 2) = 2m - 2k$. Thus, there are $\binom{2m-k-1}{2m-2k} = \binom{2m-k+1}{k-1}$ ways to choose $(a_i)_{i=1}^k$ such that $a_i \geq 2$ for all $1 \leq i \leq k$ and $a_1 + \ldots + a_k = 2m$.

Thus, there are $\sum_{k=1}^d \binom{d}{k} \binom{2m-k-1}{k-1}$ ways to choose a non-negative integer tuple $(a_1, \ldots, a_d)$ such that $a_1 + \ldots + a_d = 2m$ and $a_i \neq 1$ for all $i \in \{1, \ldots, d\}$, which is a polynomial of d with degree m.

Subsequently, we consider a pair $(P, Q) \in B_d^m$ such that $P(x_1, \ldots, x_d)Q(x_1, \ldots, x_d) = x_1^{a_1} \ldots x_d^{a_d}$ where $(a_1, \ldots, a_d)$ is a tuple of non-negative integers satisfying that $a_i \geq 2$ for all $i \in \{1, \ldots, k\}$ ($1 \leq k \leq d$) while other a_i equal 0, and $a_1 + \ldots + a_d = 2m$. Suppose that P and Q can be written as

$$P(x_1, \ldots, x_d) = x_1^{u_1} \ldots x_d^{u_d},$$
$$Q(x_1, \ldots, x_d) = x_1^{v_1} \ldots x_d^{v_d},$$

where u_i and v_i are non-negative integers for all $i \in \{1, \ldots, d\}$. Then, we have

$$u_i + v_i = a_i,$$

for any $i \in \{1, \ldots, k\}$. Obviously, there are $(1 + a_1) \ldots (1 + a_d)$ pairs of $(u_i)_{i=1}^d$ and $(v_i)_{i=1}^d$ satisfying those constraints. As a result, we have

$$|B_d^m| = (1 + a_1) \ldots (1 + a_d) \sum_{k=1}^d \binom{d}{k} \binom{2m-k-1}{k-1}.$$

Note that

$$1 \leq (a_1 + 1) \ldots (a_k + 1) \leq \left(\frac{a_1 + \ldots + a_k + k}{k} \right)^k = \left(1 + \frac{2m}{k} \right)^k \leq e^{2m},$$

or equivalently,

$$\sum_{k=1}^d \binom{d}{k} \binom{2m-k-1}{k-1} \leq |B_d^m| \leq e^{2m} \sum_{k=1}^d \binom{d}{k} \binom{2m-k-1}{k-1}.$$

Hence, we reach the conclusion of this part.

A.3. Proof of Lemma 4.6

Part (i). For $n \in \mathbb{N}$, we denote by S_n the set of all reducible tuples of the form $((j_0^1, j_1^1, \ldots, j_n^1), (j_0^2, j_1^2, \ldots, j_n^2), (j_0^3, j_1^3, \ldots, j_n^3), (j_0^4, j_1^4, \ldots, j_n^4))$ such that $j_0^1 = j_0^2 = j_0^3 = j_0^4 = 1$. Among reducible tuples in S_n, let

- A_n be the set of tuples such that $j_n^1 = j_n^2 = j_n^3 = j_n^4$,

- B_n be the set of tuples such that there exist a way to divide the set $\{1, 2, 3, 4\}$ into two disjoint subsets $\{p, q\}$ and $\{r, s\}$ such that $j_n^p = j_n^q$, $j_n^r = j_n^s$, and $j_n^p \neq j_n^r$.

It is obvious that $A_n \cap B_n = \emptyset$ and $A_n \cup B_n = S_n$. Note that for any reducible tuple in A_{n+1}, by removing the last elements of its four sub-tuples, which are $j_{n+1}^1, j_{n+1}^2, j_{n+1}^3, j_{n+1}^4$, we receive a new reducible tuple belonging to either A_n or B_n. Since $j_{i+1}^k \in [1, d]$ have d possible values, we obtain the following equation:

$$|A_{n+1}| = d|A_n| + d|B_n|.$$

Subsequently, in order to form a reducible tuple in B_{n+1} from a reducible tuple in A_n, we can add $j_{n+1}^1, j_{n+1}^2, j_{n+1}^3, j_{n+1}^4$ such that there exist a way to divide the set $\{1, 2, 3, 4\}$ into two disjoint subsets $\{p, q\}$ and $\{r, s\}$ such that $j_n^p = j_n^q$, $j_i^r = j_i^s$, and $j_n^p \neq j_n^r$. There are $\frac{1}{2}\binom{4}{2} = 3$ ways to choose two pairs (j_n^p, j_n^q) and (j_i^r, j_i^s). As $j_n^p \in [1, d]$ have d possible values, there are d ways to choose the pair (j_n^p, j_n^q). Since $j_n^p \neq j_n^r$, we have $d - 1$ ways to choose the pair (j_i^r, j_i^s). Thus, there are a total of $3d(d-1)$ ways to form a reducible tuple in B_{n+1} from a reducible tuple in A_n.

Next, we consider how to build a reducible tuple in B_{n+1} from a reducible tuple in B_n. For an arbitrary reducible tuple in B_n, without loss of generality, we assume the last elements of its four sub-tuples are such that $j_n^1 = j_n^2$, $j_n^3 = j_n^4$ and $j_n^1 \neq j_n^3$. Then, by adding $j_{n+1}^1, j_{n+1}^2, j_{n+1}^3, j_{n+1}^4$ such that $j_{n+1}^1 = j_{n+1}^2$, $j_{n+1}^3 = j_{n+1}^4$ and $j_{n+1}^1 \neq j_{n+1}^3$, we obtain a reducible tuple in B_{n+1}. There are d ways to choose the pair (j_{n+1}^1, j_{n+1}^2) and $d - 1$ to choose the pair (j_{n+1}^3, j_{n+1}^4). Thus, we have a total of $d(d-1)$ ways to form a reducible tuple in B_{n+1} from a reducible tuple in B_n. Therefore, we have the following recursive formula:

$$|B_{n+1}| = 3(d^2 - d)|A_n| + (d^2 - d)|B_n|.$$

By some simple calculations, we get $|A_1| = d$ and $|B_1| = 3(d^2 - d)$. Putting the above results together, we have

$$|A_1| = d, \quad |B_1| = 3(d^2 - d),$$
$$|A_{n+1}| = d|A_n| + d|B_n|,$$
$$|B_{n+1}| = 3(d^2 - d)|A_n| + (d^2 - d)|B_n|.$$

Then, by using the induction method, we obtain that $|A_n|$ is a polynomial of variable d of degree $2n - 1$ while $|B_n|$ is a polynomial of variable d of degree $2n$. Hence, the number of reducible tuples such that $j_0^1 = j_0^2 = j_0^3 = j_0^4 = 1$ is equal to $|S_n| = |A_n| + |B_n|$, which is also a polynomial of variable d of degree $2n$.

Part (ii). We consider reducible tuples of the form $((j_0^1, j_1^1, \ldots, j_n^1), (j_0^2, j_1^2, \ldots, j_n^2), (j_0^3, j_1^3, \ldots, j_n^3), (j_0^4, j_1^4, \ldots, j_n^4))$ such that $j_0^1 = j_0^2 = 1, j_0^3 = j_0^4 = 2$. Let A_n and B_n be defined as in part (i), then we obtain the same recursive formula:

$$|A_{n+1}| = d|A_n| + d|B_n|, \quad |B_{n+1}| = 3(d^2 - d)|A_n| + (d^2 - d)|B_n|,$$

with the initial condition $|A_1| = d, |B_1| = d^2 - d$. Using the induction method, we receive the desired result.

B. Circular Defining Function

In this appendix, we present the challenge of approximating the generalized sliced-Wasserstein distance under the setting of a circular defining function, which is defined as follows:

Definition B.1 (Circular defining function). Let x and θ be vectors in $\mathbb{R}^d$ while $t \in \mathbb{R}^+$ be a positive real number. Then, the circular defining function is given by

$$g_{\text{circular}}(x, \theta) := \|x - t\theta\|,$$

where $\|\cdot\|$ is the Euclidean norm. In addition, the generalized sliced-Wasserstein in this case is denoted as $\text{circular} - GSW$.

It can be seen that the circular defining function cannot be written as the inner product of θ and some quantity dependent on $\boldsymbol{x}$. Therefore, under this setting, we are not able to utilize the conditional central limit theorem for Gaussian projections in Theorem 2.1, which requires to project the input measures μ and ν along the direction of θ. Additionally, we are also incapable of demonstrating the relation between the circular$-GSW$ distance and SW distance as in Proposition 3.2 (when the defining function is polynomial) and Proposition 4.2 (when the defining function is neural network type) due to the same reason. As a consequence, the problem of finding a deterministic approximation of the circular$-GSW$ distance remains open, and we believe that it is essential to develop a new technique to tackle the aforementioned issue, which is left for future work.