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
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.

Index Terms—Optimal Transport, Sliced-Wasserstein distance

1. INTRODUCTION
Sliced-Wasserstein (SW) distance [1] has become a core member in the family of probability metrics that are based on optimal transport [2]. Compared to Wasserstein distance, SW provides a lower computational cost thanks to the closed-form solution of optimal transport in one-dimensional setting.

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 [3, 4, 5]; a SW variant that augments original measures to higher dimensions for better linear separation is introduced in [6]; a SW variant that uses convolution slicer for projecting images is proposed in [7]. The current SW research trend focuses on its application, including its use in generative modeling [8, 9, 10, 11], domain adaptation [12], and Bayesian inference [14, 15].

To enhance SW’s capabilities, [16] proposes using non-linear projection defining functions, introducing 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 [17].

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 the use of Monte Carlo random projecting directions. We follow the approach of deterministic approximation of the SW in [18], which utilizes the Gaussian concentration of the distribution of low-dimensional projections of high-dimensional random variables [19, 20]. 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 [16], while the result for circular function setting is left for future work. It should be emphasized that despite sharing the same framework as [18], 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.

Notation. 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, \( \gamma_d \) represents the Gaussian distribution in \( \mathbb{R}^d \), \( \mathcal{N}(0, d^{-1}I_d) \) in which \( I_d \) 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 \( \mathbb{R}^d \). For any set \( A \), we denote by \( |A| \) its cardinality. Next, for any two sequences \( (a_n) \) and \( (b_n) \), the notation \( a_n = O(b_n) \) indicates that \( a_n \leq Cb_n \) for all \( n \in \mathbb{N} \) where \( C \) is some universal constant. Lastly, \( W_p(\mu, \nu) \) and \( SW_p(\mu, \nu) \) are the order \( p \) Wasserstein distance and sliced-Wasserstein distance between the measures \( \mu \) and \( \nu \), respectively.
2. BACKGROUND

In this section, we first 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. We then recall the definition of generalized sliced-Wasserstein distance, which is mainly focused on in this paper.

Theorem 1 ([21]). For any \( d \geq 1 \), let \( \mu \) 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:

\[
\int_{\mathbb{R}^d} W_2^2(\theta^* \mu, N(0, d^{-1} m_2(\mu))) \, d\gamma_d(\theta) \leq C \Xi_d(\mu),
\]

where \( \theta^* : \mathbb{R}^d \rightarrow \mathbb{R} \) denotes the linear form \( x \mapsto (\theta, x) \), \( \theta^* \mu \) indicates the push-forward measure of \( \mu \) by \( \theta^* \) and

\[
\Xi_d(\mu) = d^{-1} \left\{ A(\mu) + \left[ m_2(\mu) B_1(\mu) \right]^{1/2} 
+ m_2(\mu)^{1/2} B_2(\mu)^{1/2} \right\},
\]

(1)

\[
m_2(\mu) = \mathbb{E} \left[ \|X_{1:d}\|^2 \right], \quad A(\mu) = \mathbb{E} \left[ \|X_{1:d}\|^2 - m_2(\mu) \right],
\]

\[
B_k(\mu) = \mathbb{E}^{1/k} \left[ \|X_{1:d}\|^k \right],
\]

with \( k \in \{1, 2\} \) and \( X_{1:d} \) is an independent copy of \( X_{1:d} \).

Generalized Sliced-Wasserstein Distance: Denote the generalized Radon transform (GRT) of an integrable function \( I \in L^1(\mathbb{R}^d) \) as \( GI \), the GSW distance is given by:

\[
GSW_p(\mu, \nu) := \left( \int_{\mathbb{R}^d} W_p^p \left( GI(\cdot, \theta), GI(\cdot, \theta) \right) \, d\gamma_d(\theta) \right)^{\frac{1}{p}},
\]

where \( I_\mu, I_\nu \in L^1(\mathbb{R}^d) \) are probability density functions of measures \( \mu \) and \( \nu \), respectively. Here, with a slight abuse of notation, we use \( W_p(\mu, \nu) \) and \( W_p(I_\mu, I_\nu) \) interchangeably.

The result in Theorem 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 finding a deterministic approximation for the generalized sliced-Wasserstein distance under the setting where the defining function \( q \) is a polynomial function with an odd degree, which is defined as follows:

Definition 1 (Polynomial defining 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 \( |\alpha| = \alpha_1 + \ldots + \alpha_d \) and \( x^\alpha = x_1^{\alpha_1} \ldots x_d^{\alpha_d} \).

Then, a polynomial defining function with an odd degree \( m \) is given by:

\[
g_{poly}(x, \theta) = \sum_{|\alpha| = m} \theta_\alpha x^\alpha,
\]

where \( \theta := (\theta_\alpha)_{|\alpha| = m} \in \mathbb{R}^{q-1} \) 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 \( \text{poly}-GSW \).

Let \( X = (X_1, \ldots, X_d)^T \) and \( Y = (Y_1, \ldots, Y_d)^T \) 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 \( \mu_q \) and \( \nu_q \) as the probability distributions in \( \mathbb{R}^q \) of random vectors \( U := (X^\alpha)_{|\alpha| = m} \in \mathbb{R}^q \) and \( V := (Y^\alpha)_{|\alpha| = m} \in \mathbb{R}^q \), we find that there is a connection between the GSW distance and the SW distance as follows:

Proposition 1. 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: \( \text{poly}-GSW_2(\mu, \nu) = SW_2(\mu_q, \nu_q) \).

As a consequence, the original problem of approximating the \( \text{poly}-GSW \) distance between \( \mu \) and \( \nu \) boils down to estimating the SW distance between \( \mu_q \) and \( \nu_q \). Combining this result with [[18], Theorem 1], we obtain the following bound for the \( \text{poly}-GSW \) distance between \( \mu \) and \( \nu \).

Theorem 2. 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

\[
\left| \text{poly}-GSW_2(\mu, \nu) - q^{-\frac{1}{2}} \left( \sqrt{m_2(\mu_q)} - \sqrt{m_2(\nu_q)} \right) \right| 
\leq C(\Xi_q(\mu_q) + \Xi_q(\nu_q))^2,
\]

where \( m_2(\zeta) \) and \( \Xi_q(\zeta) \) are defined as in equation (1) for \( \zeta \in \{\mu_q, \nu_q\} \).

Let \( m_\zeta, \tilde{\zeta} \) be the mean and centered versions of \( \zeta \) for \( \zeta \in \{\mu_q, \nu_q\} \). Putting Proposition 1 together with [[18], Proposition 2], we first try to approximate \( \text{poly}-GSW_2(\mu, \nu) \).

It is sufficient to estimate \( \hat{m}_\mu_q, \hat{m}_\nu_q, \hat{m}_2(\mu_q), \hat{m}_2(\nu_q) = \mathbb{E}[\|U\|^2] - \mathbb{E}[\|U\|^2] \) and \( \hat{m}_2(\nu_q) = \mathbb{E}[\|V\|^2] - \mathbb{E}[\|V\|^2] \). Let \( \{x_j\}_{j=1}^N \) and \( \{y_j\}_{j=1}^N \) be samples drawn from probability distributions \( \mu \) and \( \nu \), respectively. Denote \( U(x) = U_j(x)_{|\alpha| = m} := (x^\alpha)_{|\alpha| = m} \) and \( V(y) = V_j(y)_{|\alpha| = m} := (y^\alpha)_{|\alpha| = m} \) be two \( q \)-dimensional vectors. Then, the estimators of \( m_2(\hat{\mu}_q), m_2(\hat{\nu}_q), m_{\mu_q} \) and \( m_{\nu_q} \) can be calculated as:

\[
\hat{m}_2(\hat{\mu}_q) = \frac{q}{N} \sum_{j=1}^N U_j(x_j)^2 \left( \frac{\sum_{j=1}^N U_j(x_j)^2}{N} \right)^2,
\]

\[
\hat{m}_2(\hat{\nu}_q) = \frac{q}{N} \sum_{j=1}^N V_j(y_j)^2 \left( \frac{\sum_{j=1}^N V_j(y_j)^2}{N} \right)^2,
\]

\[
\hat{m}_{\mu_q} = \frac{1}{N} \sum_{j=1}^N U_j(x_j), \quad \hat{m}_{\nu_q} = \frac{1}{N} \sum_{j=1}^N V_j(y_j).
\]

Corollary 1. As a consequence, an approximation of the \( \text{poly}-GSW_2(\mu, \nu) \) can be written as

\[
\text{poly}-GSW_2^2(\mu, \nu) = q^{-1} \left( \sqrt{\hat{m}_2(\hat{\mu}_q)} - \sqrt{\hat{m}_2(\hat{\nu}_q)} \right)^2
+ q^{-1} \|\hat{m}_{\mu_q} - \hat{m}_{\nu_q}\|^2.
\]
To validate our approximation of poly−GSW$_2(\mu, \nu)$, we provide in the following theorem an upper bound for the approximation error $\left(\Xi_q(\mu_q) + \Xi_q(\nu_q)\right)^{\frac{1}{2}}$.

**Theorem 3.** 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 $j \in \mathbb{N}$, where $m$ is an odd positive integer. For $d \in \mathbb{N}$, let $X = \{X_j\}_{j=1}^d$ and $Y = \{Y_j\}_{j=1}^d$ and denote by $\mu, \nu$ the distributions of $X, Y$, respectively, while $\mu_q, \nu_q$ are defined as above. Then,

$$\left(\Xi_q(\mu_q) + \Xi_q(\nu_q)\right)^{\frac{1}{2}} \leq O(d^{-\frac{m}{2}}),$$

(2)

**Remark 1.** When $m = 1$, the polynomial defining function reduces to the linear case $g_{\text{poly}}(x, \theta) = (x, \theta)$, leading to the fact that poly−GSW$_2(\mu, \nu) = SW_2(\mu, \nu)$ for any $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. Under that setting, the approximation error in Theorem 3 approaches 0 at a rate of $(d^{-\frac{1}{2}})$, which matches the result provided in [[18], Corollary 1].

We now provide the proof of Theorem 3.

**Proof of Theorem 3.** From the definition of $\Xi_q(\mu_q)$ in equation (1), it is sufficient to bound $m_2(\mu_q), A(\mu_q)$ and $B_k(\mu_q)$ for $k \in \{1, 2\}$. For any $\alpha \in \mathbb{R}^d$ such that $|\alpha| = m$, by using the Hölder’s inequality, we have

$$m_2(\mu_q) = \sum_{|\alpha| = m} \mathbb{E}[(X^\alpha)^2] \leq q \max_{1 \leq j \leq d} \mathbb{E}[X_{2j}^m].$$

(3)

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

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

Next, applying part (ii) of the Lemma 1 (cf. the end of this proof), we obtain that $[A(\mu_q)]^2$ is upper bounded by

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

(4)

Finally, 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 $X'$ an independent copy of $X$ and $U' := (X'^\alpha)_{|\alpha|=m}$, then $\langle U, U' \rangle^2 = \sum_{|\alpha| = m} (X^\alpha X'^\alpha) + \sum_{\alpha \neq \beta} X^\alpha X'^\beta X^\alpha X'^\beta$. Taking the expectation and using part (iii) of Lemma 1 (cf. the end of this proof), we have

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

(5)

Combining the results in equations (3), (4) and (5) with a note that $q^{-1} = O(d^{-m})$, we reach the conclusion of the theorem. \qed

**Lemma 1.** 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:

- **Proposition 2.** Let $\mu^*$ and $\nu^*$ be two probability measures corresponding to the random vectors $X^*$ and $Y^*$ given as above, we obtain: $\text{neural−GSW}_2(\mu, \nu) = SW_2(\mu^*, \nu^*)$. 

4. NEURAL NETWORK TYPE FUNCTION

Polynomial projections’ memory complexity grows exponentially with data dimension $d$ and degree, constraining their ML and deep learning applications. To address this, we examine approximating the generalized sliced-Wasserstein distance equipped with a neural network type defining function.

Let $n$ be the number of layers of a neural network, we denote by $\Theta^{(1)}, \ldots, \Theta^{(n)}$ $n$ random matrices of size $d \times d$ such that they are independent of $X$ and $Y$, and their entries are i.i.d random variables following a zero-mean Gaussian distribution $\mathcal{N}(0, d^{-1})$.

**Definition 2 (Neural network type defining function).** Let $x, \theta$ be two vectors in $\mathbb{R}^d$, then a neural network type defining function is given by $g_{\text{neural}}(x, \theta) = (\theta, \Theta^{(1)} \ldots \Theta^{(n)} x)$. Accordingly, the generalized sliced-Wasserstein distance with this neural network type defining function is denoted as neural−GSW.

Next, we consider a random vector $X^* = (X^*_1, \ldots, X^*_d)^\top \in \mathbb{R}^d$ (resp. $Y^* = (Y^*_1, \ldots, Y^*_d)^\top$) which is achieved by multiplying $n$ random matrices $\Theta^{(1)}, \ldots, \Theta^{(n)}$ (each corresponding to a layer of the neural network) and $X$ (resp. $Y$). In particular, $X^* = (\Theta^{(1)} \ldots \Theta^{(n)} X$ and $Y^* = (\Theta^{(1)} \ldots \Theta^{(n)} Y$.

...
Putting this result and [[18], Theorem 1] together, we achieve the following bound for the neural–GSW($\mu, \nu$).

**Theorem 4.** 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

$$\text{neural–GSW}_2(\mu, \nu) - d^{-\frac{1}{2}} \| \sqrt{m_2(\mu^*)} - \sqrt{m_2(\nu^*)} \| \leq C(\Xi_d(\mu^*) + \Xi_d(\nu^*))^\frac{1}{2},$$

where $m_2(\zeta)$ and $\Xi_d(\zeta)$ are defined as in equation (1) for $\zeta \in \{\mu^*, \nu^*\}$.

Subsequently, we estimate the values of $m_2(\mu^*)$ and $m_2(\nu^*)$. Since $\Theta^{(1)}, \ldots, \Theta^{(m)}$ are independent random matrices with zero means and they are independent of $X$ and $Y$, it follows that $E[X^*] = 0$ and $E[Y^*] = 0$. In other words, $\mu^*$ and $\nu^*$ are zero-mean distributions. Therefore, let $\{x^{(j)}\}_{j=1}^N$ and $\{y^{(j)}\}_{j=1}^N$ be samples drawn from $\mu$ and $\nu$, respectively, we compute estimations of $m_2(\mu^*)$ and $m_2(\nu^*)$ as follows:

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

**Corollary 2.** An approximation of neural–GSW$_2(\mu, \nu)$ can be written as

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

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

**Theorem 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 $E[X_j^4] < \infty$ and $E[Y_j^4] < \infty$ for all $j \in \mathbb{N}$. For $d \in \mathbb{N}$, let $X = \{X_j\}_{j=1}^d$ and $Y = \{Y_j\}_{j=1}^d$ and denote by $\mu, \nu$ the distributions of $X, Y$, respectively, while $\mu^*, \nu^*$ are defined as above. Then, we have

$$(\Xi_d(\mu^*) + \Xi_d(\nu^*))^\frac{1}{2} \leq O(3^\frac{3}{2}d^{-\frac{1}{2}} + d^{-\frac{1}{2}}),$$ (6)

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 $g_{\text{neural}}(x, \theta) = \langle x, \theta \rangle$, implying that $\text{neural–GSW}_p(\mu, \nu) = \text{SW}_p(\mu, \nu)$ for any $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. Additionally, the approximation error in Theorem 5 goes to 0 at the same rate as in [[18], Corollary 1], which is of $O(d^{-\frac{1}{2}})$.

**5. EXPERIMENTS**

In this section, we test the approximation error of our proposed GSW approximation by increasing the dimension of simulated data, observing the change in $L_1$ 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.

**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 $N(0, I_d)$ and $N(1, 2I_d)$. We denote two empirical distributions as $\mu_0 = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\nu_0 = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$. 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. 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.

**Approimation error on autoregressive processes of order one (AR(1)):** We would like to recall that in AR(1) process, $X_t = \alpha X_{t-1} + e_t$ where $\alpha \in [0, 1]$ and $\{e_t\}_{t=1}^\infty$ are i.i.d. real random variables with $E[e_t] = 0$ 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 $\alpha$ and defining functions in Figure 2. Similar to the previous experiment, the approximation error decreases when the number of dimensions increases.

**6. CONCLUSION**

We established deterministic and fast approximations of the generalized sliced-Wasserstein distance 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. Simulation studies on different types of probability distributions confirm our theoretical results.
7. REFERENCES

[1] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister, “Sliced and Radon Wasserstein barycenters of measures,” *Journal of Mathematical Imaging and Vision*, vol. 1, no. 51, pp. 22–45, 2015.

[2] Cédric Villani, *Optimal transport: Old and New*, Springer, 2008.

[3] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing, “Max-sliced Wasserstein distance and its use for GANs,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10648–10656.

[4] Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui, “Distributional sliced-Wasserstein and applications to generative modeling,” in *International Conference on Learning Representations*, 2021.

[5] Khai Nguyen, Son Nguyen, Nhat Ho, Tung Pham, and Hung Bui, “Improving relational regularized autoencoders with spherical sliced fused Gromov-Wasserstein,” in *International Conference on Learning Representations*, 2021.

[6] Xiongjie Chen, Yongxin Yang, and Yunpeng Li, “Augmented sliced Wasserstein distances,” *International Conference on Learning Representations*, 2022.

[7] Khai Nguyen and Nhat Ho, “Revisiting sliced wasserstein on images: From vectorization to convolution,” *Advances in Neural Information Processing Systems*, 2022.

[8] Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing, “Generative modeling using the sliced Wasserstein distance,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3483–3491.

[9] Khai Nguyen and Nhat Ho, “Amortized projection optimization for sliced Wasserstein generative models,” *Advances in Neural Information Processing Systems*, 2022.

[10] Biwei Dai and Uros Seljak, “Sliced iterative normalizing flows,” in *International Conference on Machine Learning*, PMLR, 2021, pp. 2352–2364.

[11] Khai Nguyen, Tongzheng Ren, Huy Nguyen, Litu Rout, Tan Minh Nguyen, and Nhat Ho, “Hierarchical sliced wasserstein distance,” in *The Eleventh International Conference on Learning Representations*, 2023.

[12] Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool, “Sliced Wasserstein generative models,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3713–3722.

[13] Soheil Kolouri, Gustavo K Rohde, and Heiko Hoffmann, “Sliced Wasserstein distance for learning Gaussian mixture models,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3427–3436.

[14] Kimia Nadjahi, Valentin De Bortoli, Alain Durmus, Roland Badeau, and Umut Şimşekli, “Approximate Bayesian computation with the sliced-Wasserstein distance,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5470–5474.

[15] Mingxuan Yi and Song Liu, “Sliced Wasserstein variational inference,” in *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2021.

[16] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde, “Generalized sliced Wasserstein distances,” in *Advances in Neural Information Processing Systems*, 2019, pp. 261–272.

[17] Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli, “Statistical and topological properties of sliced probability divergences,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 20802–20812, 2020.

[18] Kimia Nadjahi, Alain Durmus, Pierre E Jacob, Roland Badeau, and Umut Simsekli, “Fast approximation of the sliced-Wasserstein distance using concentration of random projections,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[19] Vladimir Nikolaevich Sudakov, “Typical distributions of linear functionals in finite-dimensional spaces of higher dimension,” in *Doklady Akademii Nauk*. Russian Academy of Sciences, 1978, vol. 243, pp. 1402–1405.

[20] Persi Diaconis and David Freedman, “Asymptotics of graphical projection pursuit,” *The annals of statistics*, pp. 793–815, 1984.

[21] Galen Reeves, “Conditional central limit theorems for gaussian projections,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017.