Sliced Wasserstein with Random-Path Projecting Directions

Khai Nguyen [1] **Shujian Zhang** [1] **Tam Le** [2 3] **Nhat Ho** [1]

Abstract

Slicing distribution selection has been used as an effective technique to improve the performance of parameter estimators based on minimizing sliced Wasserstein distance in applications. Previous works either utilize expensive optimization to select the slicing distribution or use slicing distributions that require expensive sampling methods. In this work, we propose an optimization-free slicing distribution that provides a fast sampling for the Monte Carlo estimation of expectation. In particular, we introduce the *random-path* projecting direction (RPD) which is constructed by leveraging the normalized difference between two random vectors following the two input measures. From the RPD, we derive the random-path slicing distribution (RPSD) and two variants of sliced Wasserstein, i.e., the Random-Path Projection Sliced Wasserstein (RPSW) and the Importance Weighted Random-Path Projection Sliced Wasserstein (IWRPSW). We then discuss the topological, statistical, and computational properties of RPSW and IWRPSW. Finally, we showcase the favorable performance of RPSW and IWRPSW in gradient flow and the training of denoising diffusion generative models on images.

1. Introduction

Utilizing the closed-form solution of optimal transport in one dimension (Peyré & Cuturi, 2020), the sliced Wasserstein (Bonneel et al., 2015) (SW) distance is computationally and statistically scalable. In particular, SW has the time complexity of $\mathcal{O}(n \log n)$ and the space complexity of $\mathcal{O}(n)$ when dealing with two probability measures that have at most n supports. Moreover, the SW does not suffer from the curse of dimensionality since it has sample complexity of $\mathcal{O}(n^{-1/2})$ (Nadjahi et al., 2020b; Nietert et al., 2022). As evidence for the effectiveness of the SW distance, many applications have leveraged SW as the key to improving performance e.g., generative models (Deshpande et al., 2018), domain adaptation (Lee et al., 2019), point-cloud upsampling (Savkin et al., 2022), clustering (Kolouri et al., 2018), gradient flows (Bonet et al., 2022), and so on.

A key component of the SW is the slicing distribution, i.e., the distribution of the projecting direction. The slicing distribution controls the contribution of each projecting direction to the final value of the SW. A *statistically* desired slicing distribution should satisfy two properties. The first property is to assign a high probability to "informative" projecting directions. In some applications that involve weak convergence of measures, such as generative modeling (Deshpande et al., 2018) and gradient flow (Bonet et al., 2022), an informative direction can be interpreted as a discriminative direction that highlights the difference between two measures after projection. The second property is to have the maximal projecting direction in the support set, i.e., the projecting direction that can lead to the maximum projected distance. This property ensures the identity of indiscernible of the expected value. To ensure the second property, a more relaxed criterion can be considered, which involves having a continuous density on the unit-hypersphere.

Beyond statistical properties, two *computational* properties are desired for the slicing distribution. Firstly, the slicing distribution should be stable to obtain, meaning that parameters of the slicing distribution can be computationally identifiable. Secondly, it should be fast and efficient to sample projecting directions from the slicing distribution. This property is essential, as Monte Carlo samples are used to approximate the intractable expectation with respect to the slicing distribution.

The conventional SW (Bonneel et al., 2015) utilizes the uniform distribution over the unit-hypersphere as the slicing distribution, which is a fixed, continuous, and easy-to-sample distribution. However, the uniform slicing distribution is not discriminative. Distributional sliced Wasserstein (Nguyen et al., 2021) (DSW) performs optimization to find the best slicing distribution belonging to a chosen family of distributions over the unit-hypersphere, aiming to maximize the expected projected distance. When the

[1]Department of Statistics and Data Sciences, University of Texas at Austin, USA [2]Department of Advanced Data Science, The Institute of Statistical Mathematics (ISM), Japan [3]RIKEN AIP, Japan. Correspondence to: Khai Nguyen <khainb@utexas.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

family is a collection of Dirac measures, DSW reverts to max sliced Wasserstein (Deshpande et al., 2019) (Max-SW). DSW and Max-SW are discriminative, but they are not stable, computationally expensive, and require a differentiable ground metric. This is because they involve iterative gradient-based optimization, which may not guarantee a global optimum. Energy-based sliced Wasserstein (Nguyen & Ho, 2023) (EBSW) utilizes the energy-based slicing distribution, which has a density proportional to an increasing function of the projected distance. The energy-based slicing distribution is continuous, optimization-free, and discriminative. However, it requires more sophisticated sampling techniques to obtain projecting directions, such as importance sampling and Markov chain Monte Carlo (MCMC), due to the unnormalized density. Moreover, these sampling techniques lead to biased weighted estimates.

To design an effective slicing distribution, the key is to relate the *continuous* slicing distribution with the two input measures. This connection can be achieved through optimization, as demonstrated in DSW (Nguyen et al., 2021), or by directly defining the slicing distribution based on the two input measures, as seen in EBSW (Nguyen & Ho, 2023). Designing an efficient slicing distribution poses a more challenging task. Interactive procedures, such as optimization (as in DSW) and sequential sampling (as in EBSW with MCMC), should be avoided. The ideal scenario is to have a slicing distribution that allows for i.i.d. sampling, making it parallelizable. Using a proposal distribution, importance sampling in EBSW could yield a fast estimate. However, importance sampling requires a careful choice of the proposal distribution for accurate estimation, especially in high dimensions. Additionally, importance sampling leads to biased estimates in the case of unnormalized densities.

To address the challenge, we propose a novel, effective, and efficient slicing distribution obtained based on a new notation of projecting direction, named *random-path* projecting direction (RPD). The idea is to first create a 'path' between two input measures, then project the two measures along that path and conduct the one-dimensional Wasserstein comparison. Here, a path is defined as the normalized difference between two random variables that are distributed with two input measures, respectively, with an additional random perturbation for continuity guarantee. Since the path is defined as the random difference, it captures the directions in which the two measures differ. Therefore, using the distribution of the random-paths can lead to a discriminative slicing distribution. In addition, sampling a random path is efficient since it only requires the mild assumption that two input measures are easy to sample from, a condition usually held in machine learning applications.

Contribution. In summary, our contributions are three-fold:

1. We propose a new type of projecting direction, named *random-path* projecting direction (RPD), which involves a random perturbation of the normalized difference between two random variables distributed under two input measures. From the RPD, we derive the random-path slicing distribution (RPSD), which is guaranteed to be continuous. Despite having intractable density, the RPSD is fast and simple to sample random projecting directions.

2. From the RPSD, we introduce two novel variants of sliced Wasserstein. The first variant is called *random-path projection sliced Wasserstein* (RPSW), which replaces the uniform distribution in SW with the RPSD. The second variant is *importance-weighted random-path projection sliced Wasserstein* (IWRPSW). It is defined as an expectation of a weighted average of multiple randomly projected distances with RPDs. We then establish theoretical properties of RPSW and IWRPSW, encompassing topological, statistical, and computational aspects. In greater detail, we delve into the metricity of RPSW and IWRPSW, their connections to other SW variants, their sample complexity, and their computational complexities using Monte Carlo methods.

3. We conduct a comparative analysis of the proposed RPSW and IWRPSW against existing SW variants, including SW, Max-SW, DSW, and EBSW. Our first application involves gradient flow, where distances are utilized to guide a source distribution towards a target distribution. Furthermore, we introduce a new framework for training denoising diffusion models using SW metrics through the augmented generalized mini-batch energy (AGME) distance. In this framework, SW variants serve as the kernel to assess the difference between two random sets. Here, AGME functions as the loss function, aiming to minimize the difference between forward transition distributions and their corresponding reverse transition distributions in denoising diffusion models. In addition to serving as a benchmark for comparing SW variants, our proposed training approach can improve generative quality and sampling speed.

Organization. We begin by reviewing some preliminaries in Section 2. Following that, in Section 3, we define the random-path projection direction, the random-path slicing distribution, RPSW, and IWRPSW. We then derive their theoretical and computational properties. Section 4 presents experiments comparing RPSW and IWRPSW with other SW variants. In Section 5, we provide concluding remarks. Finally, we defer the proofs of key results and additional materials to the Appendices.

Notations. For any $d \geq 2$, we denote $\mathbb{S}^{d-1} := \{\theta \in \mathbb{R}^d \mid ||\theta||_2^2 = 1\}$ and $\mathcal{U}(\mathbb{S}^{d-1})$ as the unit hyper-sphere and its corresponding uniform distribution . We denote $\mathcal{P}(\mathcal{X})$ as the set of all probability measures on the set $\mathcal{X}$. For $p \geq 1$, $\mathcal{P}_p(\mathcal{X})$ is the set of all probability measures on the set $\mathcal{X}$ that have finite p-moments. For any two sequences a_n and b_n, the notation $a_n = \mathcal{O}(b_n)$ means that $a_n \leq Cb_n$ for all

$n \geq 1$, where C is some universal constant.

2. Preliminaries

We first review some essential preliminaries.

One-dimensional Wasserstein. Let μ and ν be two one-dimensional measures belongs to $\mathcal{P}_p(\mathbb{R})$, the Wasserstein distance has a closed form which is

$$\begin{aligned}\mathrm{W}_p^p(\mu,\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} |x-y|^p d\pi(x,y) \\ &= \int_0^1 |F_\mu^{-1}(z) - F_\nu^{-1}(z)|^p dz\end{aligned}$$

where F_μ and F_ν are the cumulative distribution function (CDF) of μ and ν respectively.

Sliced Wasserstein. To utilize the closed-form, the sliced Wasserstein distance averages all possible projected Wasserstein distances i.e., the definition of sliced Wasserstein (SW) distance (Bonneel et al., 2015) between two probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ is:

$$\mathrm{SW}_p^p(\mu,\nu) = \mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^{d-1})}[\mathrm{W}_p^p(\theta\sharp\mu,\theta\sharp\nu)], \quad (1)$$

where $\theta\sharp\mu$ is the push-forward measures of μ through the function $f:\mathbb{R}^d \to \mathbb{R}$ that is $f(x) = \theta^\top x$. However, the expectation in the definition of the SW distance is intractable to compute. Therefore, the Monte Carlo scheme is employed to approximate the value:

$$\widehat{\mathrm{SW}}_p^p(\mu,\nu;L) = \frac{1}{L}\sum_{l=1}^{L} \mathrm{W}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu),$$

where $\theta_1,\ldots,\theta_L \overset{i.i.d}{\sim} \mathcal{U}(\mathbb{S}^{d-1})$ and are referred to as projecting directions. The pushfoward measures $\theta_1\sharp\mu,\ldots,\theta_L\sharp\mu$ are called projections of μ (similarly for ν). The number of Monte Carlo samples L is often referred to as the number of projections. When μ and ν are discrete measures that have at most n supports, the time complexity and memory complexity of the SW are $\mathcal{O}(Ln\log n + Ldn)$ and $\mathcal{O}(Ld+Ln)$ respectively.

Distributional sliced Wasserstein. To select a better-slicing distribution, Nguyen et al. (2021) introduce the distributional sliced Wasserstein (DSW) distance which is defined between two probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ as:

$$\mathrm{DSW}_p^p(\mu,\nu) = \max_{\psi\in\Psi} \mathbb{E}_{\theta\sim\sigma_\psi(\theta))}[\mathrm{W}_p^p(\theta\sharp\mu,\theta\sharp\nu)], \quad (2)$$

where $\sigma_\psi(\theta) \in \mathcal{P}(\mathbb{S}^{d-1})$, e.g., an implicit distribution (Nguyen et al., 2021), von Mises-Fisher (Jupp & Mardia, 1979) (vMF) distribution and Power Spherical (PS) (Cao et al., 2018) with unknown location parameter $\sigma_\psi(\theta) := (\mathrm{PS})\mathrm{vMF}(\theta|\epsilon,\kappa)$, $\psi = \epsilon$. After using $T \geq 1$ (projected) stochastic (sub)-gradient ascent iterations to obtain an estimation of the parameter $\hat{\psi}_T$, Monte Carlo samples $\theta_1,\ldots,\theta_L \overset{i.i.d}{\sim} \sigma_{\hat{\psi}_T}(\theta)$ are used to approximate the value of the DSW. The time complexity and space complexity of the DSW are $\mathcal{O}(LTn\log n + LTdn)$ and $\mathcal{O}(Ld+Ln)$ without counting the complexities of sampling from $\sigma_{\hat{\psi}_T}(\theta)$.

Max sliced Wasserstein. By letting the concentration parameter $\kappa \to \infty$, the vMF and PS distributions degenerate to the Dirac distribution, we obtain the max sliced Wasserstein distance (Deshpande et al., 2019). The definition of max sliced Wasserstein (Max-SW) distance between two probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ is:

$$\text{Max-SW}_p(\mu,\nu) = \max_{\theta\in\mathbb{S}^{d-1}} \mathrm{W}_p(\theta\sharp\mu,\theta\sharp\nu). \quad (3)$$

The Max-SW is computed by using $T \geq 1$ iterations of (projected) (sub)-gradient ascent to obtain the approximation of the "max" projecting direction $\hat{\theta}_T$. The approximated value of the Max-SW is then set to $\mathrm{W}_p(\hat{\theta}_T\sharp\mu,\hat{\theta}_T\sharp\nu)$. It is worth noting that the optimization problem is non-convex (Nietert et al., 2022) which leads to the fact that we cannot obtain the global optimum $\theta^\star$. As a result, the approximation of Max-SW is not a metric even when we let the number of iterations $T \to \infty$. The time complexity and space complexity of the Max-SW are $\mathcal{O}(Tn\log n + Tdn)$ and $\mathcal{O}(d+n)$.

Energy-based sliced Wasserstein. In practice, optimization often costs more computation than sampling from a fixed slicing distribution. To avoid expensive optimization, the recent work (Nguyen & Ho, 2023) proposes an optimization-free approach that is based on energy-based slicing distribution. The energy-based sliced Wasserstein (EBSW) distance between two probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as follow:

$$\mathrm{EBSW}_p^p(\mu,\nu;f) = \mathbb{E}_{\theta\sim\sigma_{\mu,\nu}(\theta;f,p)}\left[\mathrm{W}_p^p(\theta\sharp\mu,\theta\sharp\nu)\right], \quad (4)$$

where $f:[0,\infty) \to (0,\infty)$ is an increasing energy function e.g., $f(x) = e^x$, and $\sigma_{\mu,\nu}(\theta;f,p) \propto f(\mathrm{W}_p^p(\theta\sharp\mu,\theta\sharp\nu))$. EBSW can be approximated by importance sampling with the uniform proposal distribution $\sigma_0 = \mathcal{U}(\mathbb{S}^{d-1})$. For $\theta_1,\ldots,\theta_L \overset{i.i.d}{\sim} \sigma_0(\theta)$, we have: $\widehat{\mathrm{EBSW}}_p^p(\mu,\nu;f,L) =$

$$\sum_{l=1}^{L} \mathrm{W}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu)\hat{w}_{\mu,\nu,\sigma_0,f,p}(\theta_l), \quad (5)$$

for $w_{\mu,\nu,\sigma_0,f,p}(\theta) = \frac{f(\mathrm{W}_p^p(\theta\sharp\mu,\theta\sharp\nu))}{\sigma_0(\theta)}$ is the importance weighted function and $\hat{w}_{\mu,\nu,\sigma_0,f,p}(\theta_l) = \frac{w_{\mu,\nu,\sigma_0,f,p}(\theta_l)}{\sum_{l'=1}^{L} w_{\mu,\nu,\sigma_0,f,p}(\theta_{l'})}$ is the normalized importance weights. The time complexity and space complexity of the EBSW are also $\mathcal{O}(Ln\log n + Ldn)$ and $\mathcal{O}(Ld+Ln)$ in turn. However, the estimation is biased, i.e.,

$\mathbb{E}[\widehat{\text{EBSW}}_p^p(\mu,\nu;f,L)] \neq \text{EBSW}_p^p(\mu,\nu;f)$. In addition to importance sampling, Markov Chain Monte Carlo (MCMC) can be used to approximate EBSW, however, they are very computationally expensive (Nguyen & Ho, 2023).

3. Random-Path Projection Sliced Wasserstein

We first discuss the random-path projecting direction and the random-path slicing distribution in Section 3.1. We then discuss the Random-Path Projection Sliced Wasserstein variants in Section 3.2.

3.1. Random-Path Projecting Direction

As discussed, discriminative projecting directions, i.e., directions that can highlight the difference between two distributions after the projection, are preferred. Although the energy-based slicing distribution in EBSW can do the job, it is not possible to sample directly from the energy-based slicing distribution. To address the issue, we propose a novel slicing distribution that is based on the new notion of projecting direction, referred to as "random-path".

Random paths between two measures. We now define the "random path" between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$.

Definition 1 (Random-path)**.** For any $p \geq 1$, dimension $d \geq 1$, $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$, let $X \sim \mu$ and $Y \sim \nu$, the random-path (RP) is defined as:

$$Z_{X,Y} = X - Y. \tag{6}$$

The measure of $Z_{X,Y}$ can be written as $\sigma_{\mu-\nu} := \mu * (-)\sharp\nu$ where $*$ denotes the convolution operator, and $(-)\sharp\nu$ denotes the pushforward measures of ν through the function $f(x) = -x$. The density of $\sigma_{\mu-\nu}$ is intractable, however, it is simple to sample from it, i.e., sampling $X \sim \mu$, $Y \sim \nu$, set $Z_{X,Y} = X - Y$, then $Z_{X,Y} \sim \sigma_{\mu-\nu}$. It is worth noting that $\sigma_{\mu-\nu} \neq \sigma_{\nu-\mu}$.

Random-path projecting direction. From the random path, we can create a random projecting direction by normalizing it i.e., $\frac{Z_{X,Y}}{\|Z_{X,Y}\|_2}$. However, the density of the RP, i.e., $\sigma_{\mu-\nu}$ does not always have full support on $\mathbb{R}^d$ e.g., in discrete cases (μ and ν are discrete). Therefore, $\frac{Z_{X,Y}}{\|Z_{X,Y}\|_2}$ does not fully support $\mathbb{S}^{d-1}$ which cannot guarantee theoretical property and harm practical performance. As a solution, we can add a random perturbation around the normalized random path. As long as the random perturbation has continuous density, the marginal density of the final projecting direction is continuous. Now, we define the random-path projecting direction as follows:

Definition 2 (Random-path projecting direction)**.** For any $\kappa > 0$, dimension $d \geq 1$, $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$, let $X \sim \mu$ and $Y \sim \nu$, the random-path projecting direction (RPD) is:

$$\theta | X, Y, \kappa \sim \sigma_\kappa(\cdot; P_{\mathbb{S}^{d-1}}(X - Y)), \tag{7}$$

where $P_{\mathbb{S}^{d-1}}(x) = \frac{x}{\|x\|_2}$, and σ_κ is a location scale distribution on $\mathbb{S}^{d-1}$ such as vMF and PS.

In Definition 2, the reason for choosing the location-scale family is to guarantee the projecting direction concentrates around the normalized random path. We recall the density of vMF distribution $\text{vMF}(\theta;\epsilon,\kappa) \propto \exp(\kappa\epsilon^\top\theta)$ (Jupp & Mardia, 1979) and the PS distribution $\text{PS}(\theta;\epsilon,\kappa) \propto (1+\epsilon^\top\theta)^\kappa$ (De Cao & Aziz, 2020).

Remark 1. When μ and ν have shared supports, $Z_{X,Y}$ has support at 0 which makes $P_{\mathbb{S}^{d-1}}(x)$ undefined. For continuous cases, it is not the problem since the probability of $Z_{X,Y} = 0$ is 0. For the discrete cases, we can solve the issue by simply adding a small constant to $Z_{X,Y}$ i.e., $Z_{X,Y} + c$, and treat it as the random path. We find that this is not the problem in practice since two interested distributions are often continuous or rarely have shared supports.

Random-path slicing distribution. From Definition 2, we can obtain the slicing distribution of the RPD by marginalizing out X, Y. In particular, we have the random-path slicing distribution (RPSD) defined as follows:

Definition 3. For any $0 < \kappa < \infty$, dimension $d \geq 1$, $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$, let $X \sim \mu$ and $Y \sim \nu$, the random-path slicing distribution (RPSD) is:

$$\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa) = \int \sigma_\kappa(\theta; P_{\mathbb{S}^{d-1}}(x-y))d\mu(x)d\nu(y).$$

Although $\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)$ is intractable, we can sample from it easily by sampling $X \sim \mu, Y \sim \nu$, then $\theta \sim \sigma_\kappa(P_{\mathbb{S}^{d-1}}(X-Y))$. The sampling process can also be parallelized for multiple samples $\theta_1, \ldots, \theta_L$ for $L > 1$.

Remark 2. We can rewrite $\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa) = \int \sigma_\kappa(\theta; P_{\mathbb{S}^{d-1}}(x-y))d\pi(x,y)$ where $\pi = \mu \otimes \nu$ is the independent coupling between μ and ν (efficiently for computation). As a natural extension, we can use other coupling between μ and ν. However, a more complicated coupling could cost more computation for doing sampling while the benefit of using such coupling is not trivial.

3.2. Random-Path Projection Sliced Wasserstein

We now discuss the two sliced Wasserstein variants that are based on the random-path slicing distribution.

Definitions. We first define the random-path projection sliced Wasserstein (RPSW) distance.

Definition 4. For any $p \geq 1$, $d \geq 1$, $0 < \kappa < \infty$, two probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^d)$, the random-path projection sliced Wasserstein (RPSW) distance

is defined as follows:

$$\text{RPSW}_p^p(\mu,\nu;\sigma_\kappa) = \mathbb{E}_{\theta\sim\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)}[W_p^p(\theta\sharp\mu,\theta\sharp\nu)],$$

where σ_{RP} is defined as in Definition 3.

Remark 3. The equivalent definition of RPSW to Definition 4 is: $\text{RPSW}_p^p(\mu,\nu;\sigma_\kappa) = \mathbb{E}_{X\sim\mu,Y\sim\nu}\mathbb{E}_{\theta\sim\sigma_\kappa(\theta;P_{\mathbb{S}^{d-1}}(X-Y))}[W_p^p(\theta\sharp\mu,\theta\sharp\nu)]$.

Motivated by the importance of sampling estimation of EBSW in Equation (5), we can further adjust the weight of RPDs i.e., $\theta_1,\ldots,\theta_L \sim \sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)$ by their corresponding projected distance $W_p^p(\theta_1\sharp\mu,\theta_1\sharp\nu)),\ldots,W_p^p(\theta_L\sharp\mu,\theta_L\sharp\nu))$. We now define the importance weighted random-path projection sliced Wasserstein (IWRPSW) distance.

Definition 5. For any $p\geq 1$, $d\geq 1$, $L\geq 1$, $0<\kappa<\infty$, an increasing function $f:[0,\infty)\to(0,\infty)$, two probability measures $\mu\in\mathcal{P}_p(\mathbb{R}^d)$ and $\nu\in\mathcal{P}_p(\mathbb{R}^d)$, the importance weighted random-path projection sliced Wasserstein (IWRPSW) is defined as:

$$\text{IWRPSW}_p^p(\mu,\nu;\sigma_\kappa,L,f) = \mathbb{E}_{\theta_1,\ldots,\theta_L\sim\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)} \left[\sum_{l=1}^{L} W_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu)\frac{f(W_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu))}{\sum_{j=1}^{L} f(W_p^p(\theta_j\sharp\mu,\theta_j\sharp\nu))}\right],$$

where σ_{RP} is defined as in Definition 3.

Compared to only one random projecting direction in RPSW, IWRPSW utilizes L random projecting directions in the *population* form.

Topological Properties. We first investigate the metricity of RPSW and IWRPSW.

Theorem 1. *For any $p\geq 1$, $L\geq 1$, $f:[0,\infty)\to(0,\infty)$ and $0<\kappa<\infty$, the random-path projection sliced Wasserstein $RPSW_p(\cdot,\cdot;\sigma_\kappa)$ and the importance weighted random-path projection sliced Wasserstein $IWRPSW_p(\cdot,\cdot;\sigma_\kappa,L)$ are semi-metrics in the probability space on $\mathbb{R}^d$, namely RPSW and IWRPSW satisfy non-negativity, symmetry, and identity of indiscernible. The RPSW satisfies the "quasi"-triangle inequality i.e.,*

$$RPSW_p(\mu_1,\mu_2;\sigma_\kappa) \leq RPSW_p(\mu_1,\mu_3;\sigma_\kappa,\mu_1,\mu_2) + RPSW_p(\mu_3,\mu_2;\sigma_\kappa,\mu_1,\mu_2)$$

where we have $RPSW_p^p(\mu_1,\mu_3;\sigma_\kappa,\mu_1,\mu_2) = \mathbb{E}_{\theta\sim\sigma_{RP}(\theta;\mu_1,\mu_2,\sigma_\kappa)}[W_p^p(\theta\sharp\mu_1,\theta\sharp\mu_3)]$ and a similar definition of $RPSW_p^p(\mu_3,\mu_2;\sigma_\kappa,\mu_1,\mu_2)$.

The proof of Theorem 1 is given in Appendix A.1. It is worth noting that the triangle inequality for RPSW and IWRPSW is challenging to prove due to the dependency of the two input measures with the slicing distribution. Therefore, it is unknown if they satisfy the triangle inequality.

Remark 4. Although the $\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)$ is not symmetric with respective to μ and ν, RPSW and IWRPSW are still symmetric since $W_p^p(\theta\sharp\mu,\theta\sharp\nu)$ is symmetric with respect to θ i.e., $W_p^p(\theta\sharp\mu,\theta\sharp\nu) = W_p^p(-\theta\sharp\mu,-\theta\sharp\nu)$ and we have $\sigma_\kappa(\theta;P_{\mathbb{S}^{d-1}}(x-y)) = \sigma_\kappa(-\theta;P_{\mathbb{S}^{d-1}}(y-x))$. We refer to the proof of Theorem 1 for a more detail.

We now discuss the connection between RPSW and IWRPSW, and their connection to other SW variants and Wasserstein distance.

Proposition 1. *For any $p\geq 1$, $L\geq 1$, increasing function $f:[0,\infty)\to(0,\infty)$, and $0<\kappa<\infty$, we have the following relationship:*
(i) $RPSW_p(\mu,\nu;\sigma_\kappa) \leq IWRPSW_p(\mu,\nu;\sigma_\kappa,L,f) \leq$ Max-$SW_p(\mu,\nu) \leq W_p(\mu,\nu)$,
(ii) $\lim_{\kappa\to 0} RPSW_p(\mu,\nu;\sigma_\kappa) \to SW_p(\mu,\nu)$,
(iii) $\lim_{L\to\infty} IWRPSW_p(\mu,\nu;\sigma_\kappa,L,f) \to EBSW_p(\mu,\nu;f)$,
(iv) $RPSW_p(\mu,\nu;\sigma_\kappa) \geq RPSW_1(\mu,\nu;\sigma_\kappa)$.

The proof of Proposition 1 is given in Appendix A.2.

Statistical Properties. We now discuss the one-sided sample complexity of RPSW and IWRPSW.

Proposition 2. *Let $X_1,X_2,\ldots,X_n$ be i.i.d. samples from the probability measures μ being supported on compact set of $\mathbb{R}^d$. We denote the empirical measures $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. Then, for any $p>1$, $L\geq 1$ and $0<\kappa<\infty$, there exists a universal constant $C>0$ such that*

$$\mathbb{E}[RPSW_p(\mu_n,\mu;\sigma_\kappa)] \leq \mathbb{E}[IWRPSW_p(\mu_n,\mu;\sigma_\kappa,L,f)] \leq C\sqrt{\frac{(d+1)\log(n+1)}{n}},$$

where the outer expectation is taken with respect to $X_1,X_2,\ldots,X_n$.

The proof of Proposition 2 is given in Appendix A.3. From the proposition, we can see that the approximation rate of using an empirical probability measure to a population measure of the proposed RPSW and IWRPSW is at the order of $n^{-1/2}$. Therefore, it is possible to say that the proposed RPSW and IWRPSW do not suffer from the curse of dimensionality as other SW variants. Due to the missing proof of triangle inequality of RPSW and IWRPSW, it is non-trivial to extend from the one-sided sample complexity to the two-sided sample complexity as in the conventional SW (Nadjahi et al., 2019).

Computational Properties. We now present how to compute RPSW and IWRPSW in practice.

Monte Carlo estimation. The expectations in RPSW is intractable (Definition 4), hence, Monte Carlo samples are used to estimate RPSW. In particular, we sample $\theta_1,\ldots,\theta_L \sim \sigma_{RP}(\theta;\mu,\nu,\kappa)$ as described in Section 3.1,

then we form the Monte Carlo estimate of RPSW:

$$\widehat{\text{RPSW}}_p^p(\mu,\nu;\sigma_\kappa,L)=\frac{1}{L}\sum_{l=1}^{L}W_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu). \quad (8)$$

We refer to the reader to Algorithm 1- 2 in Appendix C for the detailed computation of RPSW and IWRPSW. We then discuss the approximation error of the estimator.

Proposition 3. *For any $p\geq 1$, $0<\kappa<\infty$, dimension $d\geq 1$, and $\mu,\nu\in\mathcal{P}_p(\mathbb{R}^d)$, we have:*

$$\mathbb{E}|\widehat{RPSW}_p^p(\mu,\nu;\sigma_\kappa)-RPSW_p^p(\mu,\nu;\sigma_\kappa)| \leq \frac{1}{\sqrt{L}}Var_{\theta\sim\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)}\left[W_p^p\left(\theta\sharp\mu,\theta\sharp\nu\right)\right]^{\frac{1}{2}}.$$

The proof of Proposition 3 is given in Appendix 3. From the proposition, we observe that RPSW has the same error rate as SW in terms of L i.e., $L^{-1/2}$.

Similarly for IWRPSW, let $H\geq 1$, we sample $\theta_{11},\ldots,\theta_{HL}\sim\sigma_{RP}(\theta;\mu,\nu,\kappa)$, then we form the Monte Carlo estimate of IWRPSW as follows:

$$\widehat{\text{IWRPSW}}_p^p(\mu,\nu;\sigma_\kappa,L,H)=\frac{1}{H}\sum_{h=1}^{H}\left[\sum_{l=1}^{L}W_p^p(\theta_{hl}\sharp\mu,\theta_{hl}\sharp\nu)\frac{f(W_p^p(\theta_{hl}\sharp\mu,\theta_{hl}\sharp\nu))}{\sum_{j=1}^{L}f(W_p^p(\theta_{hj}\sharp\mu,\theta_{hj}\sharp\nu))}\right]. \quad (9)$$

Since IWRPSW has L random projecting directions, we approximate the expectation by H sets of Monte Carlo samples. To simplify and make L as the only parameter for the number of projections, we set $H=1$ in this paper. In contrast to RPSW, the error rate of IWRPSW with respect to L is non-trivial due to the importance weights.

Unbiasedness. Since $\text{RPSW}_p^p(\mu,\nu;\sigma_\kappa)$ and $\text{IWRPSW}_p^p(\mu,\nu;\sigma_\kappa,L,f)$ can be approximated directly by Monte Carlo samples, their corresponding estimators $\widehat{\text{RPSW}}_p^p(\mu,\nu;\sigma_\kappa)$ and $\widehat{\text{IWRPSW}}_p^p(\mu,\nu;\sigma_\kappa,L,H)$ are unbiased estimates.

Computational complexities. When μ and ν are discrete measures that have a most n supports, sampling from them only cost $\mathcal{O}(n)$ in terms of time and space. Hence, sampling L random paths between μ and ν cost $\mathcal{O}(Ldn)$ in time and memory. After that, sampling from the vMF distribution and the PS distribution costs $\mathcal{O}(Ld)$ (Algorithm 1 in (De Cao & Aziz, 2020)) in time and memory. Adding the complexities for computing one-dimensional Wasserstein distance, the time complexity and space complexity for RPSW is $\mathcal{O}(Ln\log n+Ldn)$ and $\mathcal{O}(Ld+Ln)$ respectively. For IWRPSW, the complexities are multiplied by H, however, they can be kept the same if setting $H=1$.

Gradient estimation. In applications where RPSW and IWRPSW are used as a risk to estimate some parameters of interest i.e., μ_ϕ with $\phi\in\Phi$, we might want to estimate the gradient of RPSW and IWRPSW with respect to ϕ. For RPSW, we have:

$$\nabla_\phi\text{RPSW}_p^p(\mu_\phi,\nu;\sigma_\kappa) = \nabla_\phi\mathbb{E}_{X\sim\mu_\phi,Y\sim\nu}\mathbb{E}_{\theta\sim\sigma_\kappa(\theta;P_{\mathbb{S}^{d-1}}(X-Y))}[W_p^p(\theta\sharp\mu_\phi,\theta\sharp\nu)].$$

Since vMF and PS are reparameterizable (De Cao & Aziz, 2020), we can reparameterize $\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)$ if μ_ϕ is reparameterizable (e.g., $\mu_\phi:=f_\phi\sharp\varepsilon$ for ε is a fixed distribution, or discussed other distributions in (Kingma & Welling, 2013)). With parameterized sampling, we can sample $\theta_{1,\phi},\ldots,\theta_{L,\phi}\sim\sigma_{RP}(\theta;\mu_\phi,\nu,\sigma_\kappa)$, then form an unbiased gradient estimator as follow:

$$\nabla_\phi\text{RPSW}_p^p(\mu_\phi,\nu;\sigma_\kappa)\approx\frac{1}{L}\sum_{l=1}^{L}\nabla_\phi W_p^p(\theta_{l,\phi}\sharp\mu_\phi,\theta_{l,\phi}\sharp\nu).$$

For IWRPSW, we sample $\theta_{11,\phi},\ldots,\theta_{HL,\phi}\sim\sigma_{RP}(\theta;\mu_\phi,\nu,\kappa)$, then form the following estimator:

$$\nabla_\phi\text{IWRPSW}_p^p(\mu_\phi,\nu;\sigma_\kappa,L,H)=\frac{1}{H}\sum_{h=1}^{H}\left[\nabla_\phi\sum_{l=1}^{L}W_p^p(\theta_{hl,\phi}\sharp\mu_\phi,\theta_{hl,\phi}\sharp\nu)\frac{f(W_p^p(\theta_{hl,\phi}\sharp\mu_\phi,\theta_{hl,\phi}\sharp\nu))}{\sum_{j=1}^{L}f(W_p^p(\theta_{hj,\phi}\sharp\mu_\phi,\theta_{hj,\phi}\sharp\nu))}\right].$$

Remark 5. We found in later experiments that removing the dependent of ϕ in the slicing distribution i.e., using a dependent copy $\sigma_{RP}(\theta;\mu_{\phi'},\nu,\kappa)$ with $\phi'=\phi$ in value, leads to a more stable estimator in practice. It is worth noting that using a copy of both μ and ν (i.e., μ' and ν') in RPSW can even lead to a variant that satisfies the triangle inequality since the slicing distribution becomes a fixed distribution in this situation. We refer to this gradient estimator as the simplified gradient estimator.

4. Experiments

In this section, we compare the proposed RPSW and IWRPSW to the existing SW variants such as SW, Max-SW, DSW, and EBSW in gradient flow in Section 4.1 and training denoising diffusion models in 4.3.

4.1. Toy Gradient Flows

The gradient flow models a dynamic distribution $\mu(t)$ flowing with time t along the gradient flow of a loss functional $\mu(t)\rightarrow f(\mu(t))$. In this experiment, we set $f(\mu(t))=\mathcal{D}(\mu(t),\nu)$ with ν as the target distribution and $\mathcal{D}$ is a given SW variant between probability measures. The flow will drive the source distribution towards ν (Feydy et al., 2019;

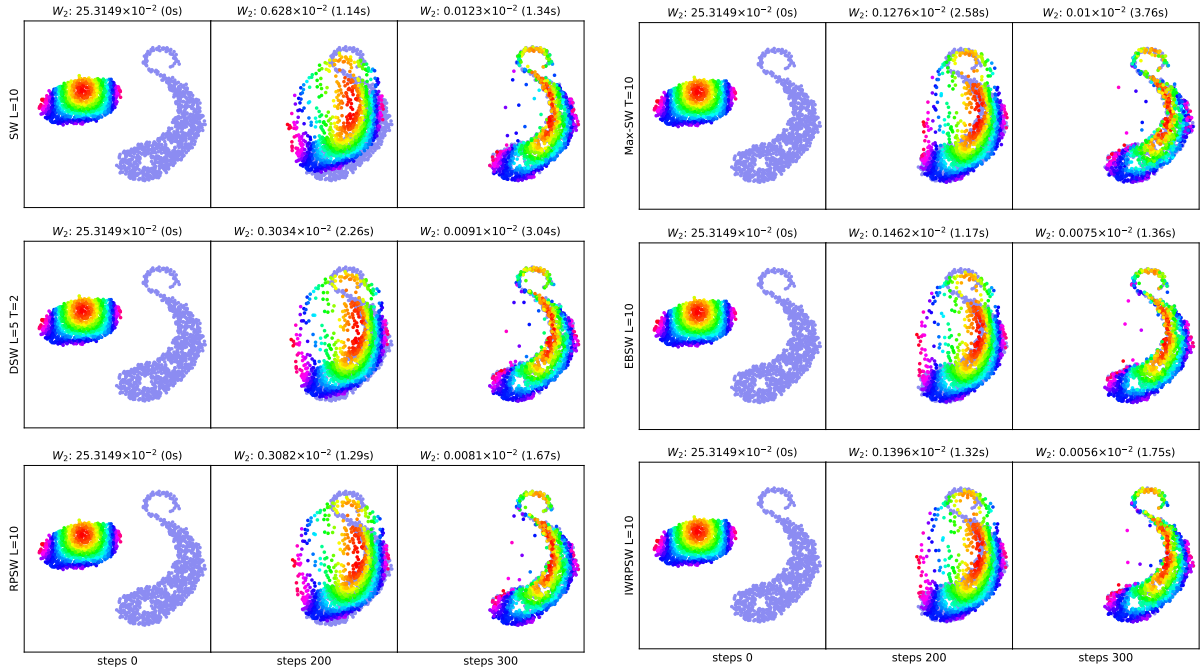


Figure 1. Results for gradient flows that are from the empirical distribution over the color points to the empirical distribution over S-shape points produced by different SW variants. The corresponding Wasserstein-2 distance between the empirical distribution at the current step and the S-shape distribution and the computational time (in second) to reach the step is reported at the top of the figure.

Santambrogio, 2015). In this setup, we consider the discrete setting i.e., $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ and $\mu(t) = \frac{1}{n}\sum_{i=1}^n \delta_{X_i(t)}$.

Setting. To solve the flow, we integrate the ordinary differential equation $\dot{X}(t) = -n\nabla_{X(t)}\left[\mathcal{D}\left(\frac{1}{n}\sum_{i=1}^n \delta_{X_i(t)}, \nu\right)\right]$ with the Euler scheme with 300 timesteps and the step size is 10^{-4}. We set the total number of projections for SW variants to 10. For DSW, RPSW, and IWRPSW, we set the concentration parameter κ of the PS distribution as a dynamic quantity i.e., $\kappa(t) = (\kappa_0 - 1)\left(\frac{N-t-1}{N-1}\right)^{10} + 1$ with $N = 300$ and $\kappa_0 \in \{100, 50\}$. The reason for the dynamic choice of κ is to stabilize the convergence of the flow since when two distributions are relatively closed, the uniform distribution is nearly optimal.

Results. We show both the qualitative visualization and quantitative comparison (in Wasserstein-2 distance (Flamary et al., 2021)) in Figure 1. From the figure, we observe that RPSW leads to a faster and better convergence than SW and IWRPSW leads to a faster and better convergence than EBSW. We also observe that optimization-based variants such as Max-SW and DSW are slower than sampling-based variants like SW, EBSW, RPSW, and IWRPSW. Moreover, the convergence of Max-SW and DSW is not as good as EBSW, RPSW, and IWRPSW. Overall, the experiment has shown the benefit of the random-path slicing distribution which improves the performance while having a fast computation. In the experiment, we use the simplified gradient estimator for RPSW and IWRPSW. We refer the reader to Figure 4 in Appendix D for the result of the original gradient estimator. Despite the fact that such estimators can lead to faster convergence, they seem to be worse in remaining the original topology of the source distribution.

4.2. Gradient Flows on Images

Setting. Utilizing MNIST dataset (LeCun et al., 1998), we select images of digit 1 to construct the source distribution and images of digit 0 to construct the target distribution. Following the previous section, we use the Euler discretization scheme with step size 1 and $N = 5000$ iterations. We compare SW variants i..e, SW, EBSW, RPSW, and IWRPSW with the number of projections $L = 1000$, Max-SW with $T = 1000$, DSW with $L = 100, T = 10$. For DSW, RPSW, and IWRPSW, we set $\kappa(t) = (\kappa_0 - 1)\left(\frac{N-t-0.001}{N-1}\right)^{10} + 0.001$ with $N = 5000$ and $\kappa_0 \in 1000$ For evaluation, we use the Wasserstein-2 distance.

Results. We report the Wasserstein-2 distance across timesteps and the computational time in Table 1. Moreover, we visualize the flows in Figure 2 and Figure 5 in Appendix D. We see that RPSW and IWRPSW help to reduce the Wasserstein-2 very fast compared to other optimization-free variants i.e., SW and EBSW. Moreover, RPSW and IWRPSW have significantly lower computation than DSW and Max-SW. It is worth noting that the quality of the gradient flow can be improved using convolution slicing operator as in (Nguyen & Ho, 2022; Du et al., 2023).

Table 1. Wasserstein-2 distance and computational times across timesteps in gradient flow on MNIST.

Distance	Step 100		Step 500		Step 1000		Step 3000		Step 5000	
	W_2(↓)	Time (s)(↓)	W_2(↓)	Time (s)(↓)	W_2(↓)	Time (s)(↓)	W_2(↓)	Time (s)(↓)	W_2(↓)	Time (s)(↓)
SW	83.86	**7.11**	67.74	**16.76**	49.34	**28.09**	29.14	**67.65**	25.17	**107.83**
Max-SW	**34.74**	128.88	33.79	753.47	33.42	1435.14	33.00	38145.11	32.37	6436.89
DSW	50.26	136.14	48.22	462.57	40.71	764.98	29.96	202.79	27.49	291.20
EBSW	80.29	7.25	67.13	16.99	48.69	28.51	29.03	68.53	25.11	108.65
RPSW	46.54	7.58	28.97	18.13	27.57	30.61	23.83	74.71	21.86	118.83
IWRPSW	44.28	7.59	**28.75**	18.20	**27.49**	30.66	**23.82**	77.71	**21.81**	120.05

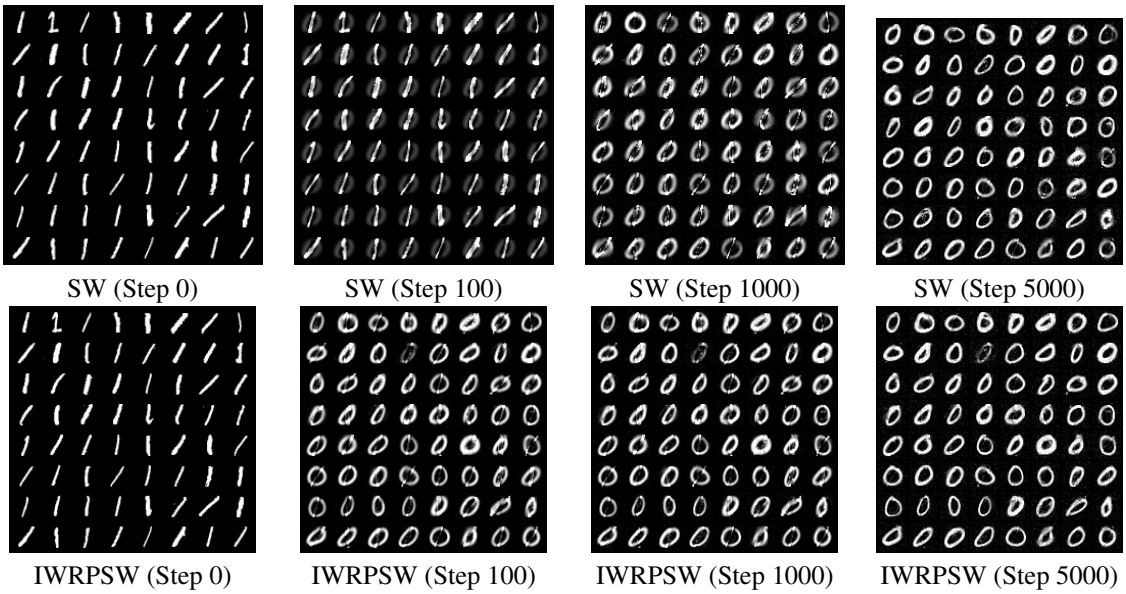


Figure 2. Gradient flows from MNIST digit 1 to MNIST digit 0. .

4.3. Denoising Diffusion Models

Denoising diffusion model (Ho et al., 2020; Sohl-Dickstein et al., 2015) defines a forward process that gradually adds noise to the data $x_0 \sim q(x_0)$. The process has $T > 0$ steps and the noise is Gaussian.

$$q(x_{1:T}|x_0) = \prod_{t\geq 1} q(x_t|x_{t-1}),$$

where $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ with the pre-defined variance schedule β_t. Training denoising diffusion model is to estimate some parameters ϕ of a reverse denoising process defined by:

$$p_\phi(x_{0:T}) = p(x_T)\prod_{t\geq 1} p_\phi(x_{t-1}|x_t),$$

where $p_\phi(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\phi(x_t, t), \sigma_t^2 I)$. The conventional training uses the maximum likelihood approach by maximizing the evidence lower bound $\mathcal{L} \leq p_\phi(x_0) = \int p_\phi(x_{0:T})dx_{1:T}$, which can be rewritten as:

$$\mathcal{L} = -\sum_{t\geq 1} \mathbb{E}_{q(x_t)}[KL(q(x_{t-1}|x_t)||p_\phi(x_{t-1}|x_t))] + C,$$

where KL denotes the Kullback-Leibler divergence, and C is a known constant.

Implicit denoising model. To reduce the number of steps T for faster generation, denoising diffusion GANs (Xiao et al., 2021) proposes to use the implicit denoising model $p_\phi(x_{t-1}|x_t) = \int p(\epsilon)q(x_{t-1}|x_t, x_0 = G_\phi(x_t, \epsilon, t))$ where $\epsilon \sim \mathcal{N}(0, I)$. After that, Xiao et al. (2021) use adversarial training to estimate the parameters i.e,

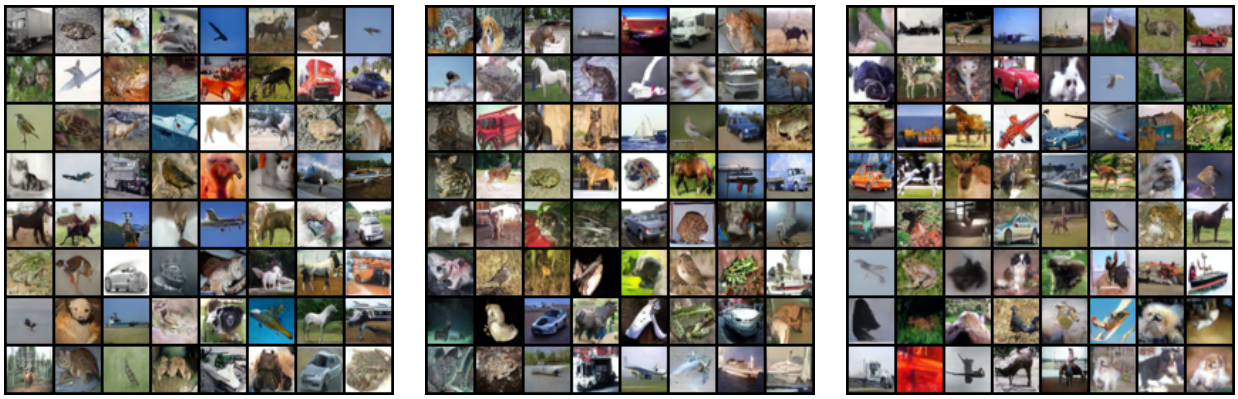
$$\min_\phi \sum_{t\geq 1} \mathbb{E}_{q(x_t)}[D_{adv}(q(x_{t-1}|x_t)||p_\phi(x_{t-1}|x_t))],$$

where D_{adv} is the GAN objective or the Jensen Shannon divergence (Goodfellow et al., 2014).

Augmented Generalized Mini-batch Energy distances. In this paper, we replace D_{adv} by the generalized mini-batch Energy distance (Salimans et al., 2018). In particular, for two measures μ and ν, and mini-batch size $m \geq 1$, we have the generalized Mini-batch Energy distance with SW kernel:

$$\begin{aligned} GME^2_{\mathcal{D}}(\mu,\nu) &= 2\mathbb{E}[\mathcal{D}(P_X, P_Y)] \\ &- \mathbb{E}[\mathcal{D}(P_X, P_{X'}) - \mathbb{E}[\mathcal{D}(P_Y, P_{Y'})], \end{aligned}$$

where $X, X' \overset{i.i.d}{\sim} \mu^{\otimes m}$ and $X, X' \overset{i.i.d}{\sim} \nu^{\otimes m}$, $P_X =$



DDGAN RPSW-DD IWRPSW-DD

Figure 3. Random generated images on CIFAR10 from DDGAN, RPSW-DD, and IWRPSW-DD.

Table 2. Results for unconditional generation on CIFAR-10.

Model	FID↓	NFE↓	G-Time (s ↓)
DDPM (Ho et al., 2020)	3.17	1000	80.5
Score SDE (VE) (Song et al., 2020b)	2.20	2000	423.2
Score SDE (VP) (Song et al., 2020b)	2.41	2000	421.5
DDIM (Song et al., 2020a)	4.67	50	4.01
Prob-Flow (VP) (Song et al., 2020b)	3.08	140	50.9
LSGM (Vahdat et al., 2021)	2.10	147	44.5
DDGAN (Xiao et al., 2021)	3.64	4	0.21
SW-DD (ours)	2.90	4	0.21
Max-SW-DD (ours)	2.99	4	0.21
DSW-DD (ours)	2.88	4	0.21
EBSW-DD (ours)	2.87	4	0.21
RPSW-DD (ours)	2.82	4	0.21
IWRPSW-DD (ours)	**2.70**	4	0.21

$\frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ with $X = (x_1, \ldots, x_m)$, and $\mathcal{D}$ is a SW variant. In the denoising diffusion case, μ is $q(x_{t-1}|x_t)$ and ν is $p_\phi(x_{t-1}|x_t))$ for $t = 1, \ldots, T$. Although the generalized Mini-batch Energy distance can guarantee statistical convergence, it is known to be not good in practical training due to the weak training signal (lack of non-linearity which is essential for images) (Salimans et al., 2018). To address the issue, we propose the augmented generalized Mini-batch Energy distance: $AGME^2_{\mathcal{D}}(\mu, \nu; g) = GME^2_{\mathcal{D}}(\bar{\mu}, \bar{\nu})$, where $\bar{\mu} = f\sharp\mu$ and $\bar{\nu} = f\sharp\nu$ with $f(x) = (x, g(x))$ for $g : \mathbb{R}^d \to \mathbb{R}$ is a non-linear function. For any g, $AGME^2_{\mathcal{D}}(\mu, \nu; g)$ is still a valid distance since f is injective. In the diffusion model setting, we train g_γ (parameterized by γ) as a time-conditional discriminator as in (Xiao et al., 2021, Equation (5)). We refer the reader to Algorithm 3 in Appendix C for a detailed algorithm, and to Appendix D for more detailed experimental settings.

Results. We follow the setting in (Xiao et al., 2021) for diffusion models on CIFAR10 (Krizhevsky et al., 2009) with $N = 1800$ epochs. We set $L = 10^4$ for SW, DSW, EBSW, RPSW, and IWRPSW. We set $T \in \{2, 5, 10\}$ for DSW and $T \in \{100, 1000\}$ for Max-SW. It is worthing noting that the training time of Max-SW and DSW is more than two times the time for SW, EBSW, RPSW, and IWRPSW, hence, we cannot increase the value of T further. We adjust concentration parameter κ of the PS distribution for each epoch t as $\kappa(t) = (\kappa_0 - 1)\left(\frac{N-t-1}{N-1}\right)^{10} + 1$ with $N = 1800$ and $\kappa_0 \in \{100, 50\}$. We report the FID scores (Heusel et al., 2017), the number of function evaluations, and the generation time in Table 2. Using an implicit denoising model with only 4 time steps, our SW diffusion models have a much faster generation time compared to existing diffusion models such as DDPM, Score SDE, DDIM, Probability Flow, and LSGM. Compared to DDGAN with the same 4 time steps, our SW diffusion models lead to a lower FID score and the same generation time. Among SW variants, the proposed RPSW and IWRPSW achieve the lowest scores of 2.82 and 2.70 in turn. We show some random generated images from DDGAN, RPSW-DD, and IWRPSW-DD in Figure 3 as qualitative results. Overall, AGME distance with RPSW and IWRPSW kernels could be a potential effective loss for improving training implicit diffusion models[1].

5. Conclusion

We have presented the new construction of discriminative projecting directions for SW i.e., the random-path projecting direction. From that, we derive the random-path slicing distribution and two optimization-free variants of SW, named random-path sliced Wasserstein (RPSW) and importance weighted random-path sliced Wasserstein (IWRPSW). We then discuss the topological properties, statistical properties, and computational properties of the proposed RPSW and IWRPSW. Finally, we show the favorable performance of RPSW and IWRPSW in gradient flow and training denoising diffusion models. In the future, we will extend the random-path directions to probability measures that have supported on the hyper-sphere (Bonet et al., 2023a), and hyperbolic manifolds (Bonet et al., 2023b). In such cases, the random path will not be straight lines but curves.

[1]Code for this paper is published at `https://github.com/khainb/RPSW`.

Impact Statement

The paper proposes a new variant of sliced Wasserstein to compare two probability measures. Given the numerous applications in machine learning, such as generative modeling, clustering, classification, domain adaptation, and more, the proposed random-path variants of sliced Wasserstein distance could enhance the performance of downstream applications in both quality and computation, as demonstrated in the paper.

Acknowledgements

We would like to thank Peter Pao-Huang for his insightful discussion during the course of this project. NH acknowledges support from the NSF IFML 2019844 and the NSF AI Institute for Foundations of Machine Learning. TL acknowledges the support of JSPS KAKENHI Grant number 23K11243, and Mitsui Knowledge Industry Co., Ltd. grant.

References

Bai, Y., Schmitzer, B., Thorpe, M., and Kolouri, S. Sliced optimal partial transport. *arXiv preprint arXiv:2212.08049*, 2022.

Bai, Y., Tran, H., Damelin, S. B., and Kolouri, S. Partial transport for point-cloud registration. *arXiv preprint arXiv:2309.15787*, 2023.

Bonet, C., Courty, N., Septier, F., and Drumetz, L. Efficient gradient flows in sliced-Wasserstein space. *Transactions on Machine Learning Research*, 2022.

Bonet, C., Berg, P., Courty, N., Septier, F., Drumetz, L., and Pham, M.-T. Spherical sliced-Wasserstein. *International Conference on Learning Representations*, 2023a.

Bonet, C., Chapel, L., Drumetz, L., and Courty, N. Hyperbolic sliced-wasserstein via geodesic and horospherical projections. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 334–370. PMLR, 2023b.

Bonet, C., Malézieux, B., Rakotomamonjy, A., Drumetz, L., Moreau, T., Kowalski, M., and Courty, N. Sliced-wasserstein on symmetric positive definite matrices for m/eeg signals. In *International Conference on Machine Learning*, pp. 2777–2805. PMLR, 2023c.

Bonneel, N. and Coeurjolly, D. Spot: sliced partial optimal transport. *ACM Transactions on Graphics (TOG)*, 38(4): 1–13, 2019.

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.

Bonnotte, N. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.

Cao, Z., Ma, L., Long, M., and Wang, J. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 135–150, 2018.

De Cao, N. and Aziz, W. The power spherical distribution. *arXiv preprint arXiv:2006.04437*, 2020.

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.

Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Du, C., Li, T., Pang, T., Yan, S., and Lin, M. Nonparametric generative modeling with conditional sliced-wasserstein flows. In *International Conference on Machine Learning*, pp. 8565–8584. PMLR, 2023.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouve, A., and Peyré, G. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, 2019.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jupp, P. E. and Mardia, K. V. Maximum likelihood estimators for the matrix von Mises-Fisher and bingham distributions. *The Annals of Statistics*, 7(3):599–606, 1979.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Le, T., Yamada, M., Fukumizu, K., and Cuturi, M. Tree-sliced variants of Wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

Le, T., Nguyen, T., Phung, D., and Nguyen, V. A. Sobolev transport: A scalable metric for probability measures with graph metrics. In *International Conference on Artificial Intelligence and Statistics*, pp. 9844–9868, 2022.

Le, T., Nguyen, K., Sun, S., Ho, N., and Xie, X. Integrating efficient optimal transport and functional maps for unsupervised shape correspondence learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024a.

Le, T. T., Nguyen, K., shanlin sun, Han, K., Ho, N., and Xie, X. Diffeomorphic mesh deformation via efficient optimal transport for cortical surface reconstruction. In *The Twelfth International Conference on Learning Representations*, 2024b.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, C.-Y., Batra, T., Baig, M. H., and Ulbricht, D. Sliced Wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10285–10295, 2019.

McCann, R. J. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

Nadjahi, K., Durmus, A., Simsekli, U., and Badeau, R. Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. In *Advances in Neural Information Processing Systems*, pp. 250–260, 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.

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

Nguyen, K. and Ho, N. Energy-based sliced Wasserstein distance. *Advances in Neural Information Processing Systems*, 2023.

Nguyen, K., Ho, N., Pham, T., and Bui, H. Distributional sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021.

Nguyen, K., Ren, T., and Ho, N. Markovian sliced Wasserstein distances: Beyond independent projections. *Advances in Neural Information Processing Systems*, 2023.

Nguyen, K., Bariletto, N., and Ho, N. Quasi-monte carlo for 3d sliced wasserstein. *International Conference on Learning Representations*, 2024.

Nietert, S., Sadhu, R., Goldfeld, Z., and Kato, K. Statistical, robustness, and computational guarantees for sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 2022.

Paulin, L., Bonneel, N., Coeurjolly, D., Iehl, J.-C., Webanck, A., Desbrun, M., and Ostromoukhov, V. Sliced optimal transport sampling. *ACM Transactions on Graphics (TOG)*, 39(4):99–1, 2020.

Peyré, G. and Cuturi, M. Computational optimal transport, 2020.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Savkin, A., Wang, Y., Wirkert, S., Navab, N., and Tombari, F. Lidar upsampling with sliced Wasserstein distance. *IEEE Robotics and Automation Letters*, 8(1):392–399, 2022.

Séjourné, T., Bonet, C., Fatras, K., Nadjahi, K., and Courty, N. Unbalanced optimal transport meets sliced-Wasserstein. *arXiv preprint arXiv:2306.07176*, 2023.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020b.

Tanguy, E. Convergence of sgd for training neural networks with sliced Wasserstein losses. *arXiv preprint arXiv:2307.11714*, 2023.

Vahdat, A., Kreis, K., and Kautz, J. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

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

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Xiao, Z., Kreis, K., and Vahdat, A. Tackling the generative learning trilemma with denoising diffusion gans. In *International Conference on Learning Representations*, 2021.

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

Supplement to "Sliced Wasserstein with Random-Path Projecting Projections"

First, we present skipped proofs in the main text in Appendix A. We then discuss some related works in B. After that, we present detailed algorithms for computing RPSW and IWRPSW with Monte Carlo estimation and training denoising diffusion models with augmented mini-batch energy distance in Appendix C. Then, we provide some additional experimental results on gradient estimators in Appendix D. Finally, we report the computational infrastructure in Appendix E.

A. Proofs

A.1. Proof of Theorem 1

Non-negativity. Since the Wasserstein distance is non-negative, we have $W_p(\theta\sharp\mu, \theta\sharp\nu) \geq 0$ for all $\theta \in \mathbb{S}^{d-1}$. Therefore, $\mathbb{E}_{\theta\sim\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)}[W_p(\theta\sharp\mu, \theta\sharp\nu)] \geq 0$ which leads to $\text{RPSW}_p(\mu, \nu; \sigma_\kappa) \geq 0$. Similarly, we have $\mathbb{E}_{\theta_1,\ldots,\theta_L\sim\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)}\left[\sum_{l=1}^L W_p^p(\theta_l\sharp\mu, \theta_l\sharp\nu)\frac{f(W_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu))}{\sum_{j=1}^L f(W_p^p(\theta_j\sharp\mu,\theta_j\sharp\nu))}\right] \geq 0$ which implies $\text{IWRPSW}_p(\mu, \nu; \sigma_\kappa, L) \geq 0$

Symmetry. From the definition of RPSW from Remark 3, and $f_{\sigma_\kappa(P_{\mathbb{S}^{d-1}}(x-y))}(\theta)$ as the density function of $\sigma_\kappa(P_{\mathbb{S}^{d-1}}(x-y))$ we have:

$$
\begin{aligned}
\text{RPSW}_p^p(\mu, \nu; \sigma_\kappa) &= \mathbb{E}_{X\sim\mu,Y\sim\nu}\mathbb{E}_{\theta\sim\sigma_\kappa(\theta;P_{\mathbb{S}^{d-1}}(X-Y))}[W_p^p(\theta\sharp\mu, \theta\sharp\nu)] \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} W_p^p(\theta\sharp\mu, \theta\sharp\nu) f_{\sigma_\kappa(P_{\mathbb{S}^{d-1}}(x-y))}(\theta) d\theta d\mu(x) d\nu(y) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} W_p^p(-\theta\sharp\mu, -\theta\sharp\nu) f_{\sigma_\kappa(P_{\mathbb{S}^{d-1}}(x-y))}(-\theta) d\theta d\mu(x) d\nu(y) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} W_p^p(\theta\sharp\mu, \theta\sharp\nu) f_{\sigma_\kappa(P_{\mathbb{S}^{d-1}}(y-x))}(\theta) d\theta d\mu(x) d\nu(y) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} W_p^p(\theta\sharp\nu, \theta\sharp\mu) f_{\sigma_\kappa(P_{\mathbb{S}^{d-1}}(y-x))}(\theta) d\theta d\nu(y) d\mu(x) \\
&= \mathbb{E}_{Y\sim\nu,X\sim\mu}\mathbb{E}_{\theta\sim\sigma_\kappa(\theta;P_{\mathbb{S}^{d-1}}(Y-X))}[W_p^p(\theta\sharp\nu, \theta\sharp\mu)] = \text{RPSW}_p^p(\nu, \mu; \sigma_\kappa),
\end{aligned}
$$

where we use the reflection property of $\mathbb{S}^{d-1}$, the property:

$$
\begin{aligned}
W_p^p(\theta\sharp\mu, \theta\sharp\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} \int |\theta^\top(x-y)|^p d\pi(x,y) \\
&= \inf_{\pi\in\Pi(\mu,\nu)} \int |-\theta^\top(x-y)|^p d\pi(x,y) = W_p^p(-\theta\sharp\mu, -\theta\sharp\nu),
\end{aligned}
$$

and

$$
f_{\sigma_\kappa(\theta;P_{\mathbb{S}^{d-1}}(x-y))}(\theta) = f_{\sigma_\kappa(\theta;P_{\mathbb{S}^{d-1}}(y-x))}(-\theta)
$$

which holds for both the vMF density $f_{\sigma_\kappa(\theta;P_{\mathbb{S}^{d-1}}(x-y)}(\theta) \propto \exp\left(\kappa\frac{(x-y)}{\|x-y\|_2}^\top\theta\right)$ and the PS density $f_{\sigma_\kappa(\theta;P_{\mathbb{S}^{d-1}}(x-y)}(\theta) \propto \left(1+\frac{(x-y)}{\|x-y\|_2}^\top\theta\right)^\kappa$. Similarly, we have $\text{IWRPSW}_p^p(\mu, \nu; \sigma_\kappa, L, f) = \text{IWRPSW}_p^p(\nu, \mu; \sigma_\kappa, L, f)$.

Identity. We need to show that $\text{RPSW}_p(\mu, \nu; \sigma_k) = 0$ if and only if $\mu = \nu$. For the forward direction, since $W_p(\theta\sharp\mu, \theta\sharp\nu) = 0$ when $\mu = \nu$, we obtain directly $\mu = \nu$ implies $\text{RPSW}_p(\mu, \nu; \sigma_k) = 0$. For the reverse direction, we use the same proof technique in (Bonnotte, 2013). If $\text{RPSW}_p(\mu, \nu; \sigma_k) = 0$, we have $\int_{\mathbb{S}^{d-1}} \text{W}_p(\theta\sharp\mu, \theta\sharp\nu)\, \text{d}\sigma_{RP}(\theta; \mu, \nu, \sigma_k) = 0$. Hence, we have $W_p(\theta\sharp\mu, \theta\sharp\nu) = 0$ for $\sigma_{RP}(\theta; \mu, \nu, \sigma_k)$-almost surely $\theta \in \mathbb{S}^{d-1}$. Since $\sigma_{RP}(\theta; \mu, \nu, \sigma_k)$ is continuous due to the continuity of σ_k, we have $W_p(\theta\sharp\mu, \theta\sharp\nu) = 0$ for all $\theta \in \mathbb{S}^{d-1}$. Therefore, we obtain $\theta\sharp\mu = \theta\sharp\nu$ for $\sigma_{\mu,\nu}(\theta; f, p)$-a.e $\theta \in \mathbb{S}^{d-1}$ due to the identity of Wasserstein distance. For any $t \in \mathbb{R}$ and $\theta \in \mathbb{S}^{d-1}$, we have:

$$
\begin{aligned}
\mathcal{F}[\mu](t\theta) &= \int_{\mathbb{R}^d} e^{-it\langle\theta,x\rangle} d\mu(x) = \int_{\mathbb{R}} e^{-itz} d\theta\sharp\mu(z) = \mathcal{F}[\theta\sharp\mu](t) \\
&= \mathcal{F}[\theta\sharp\nu](t) = \int_{\mathbb{R}} e^{-itz} d\theta\sharp\nu(z) = \int_{\mathbb{R}^d} e^{-it\langle\theta,x\rangle} d\nu(x) = \mathcal{F}[\nu](t\theta),
\end{aligned}
$$

where $\mathcal{F}[\gamma](w) = \int_{\mathbb{R}^{d'}} e^{-i\langle w,x\rangle} d\gamma(x)$ denotes the Fourier transform of $\gamma \in \mathcal{P}(\mathbb{R}^{d'})$. The above transformation is also known as the slice projection theorem. By the injectivity of the Fourier transform.

For IWRPSW, when $\mu = \nu$, we have $W_p(\theta_l \sharp \mu, \theta_l \sharp \nu) = 0$ for any $\theta_1, \ldots, \theta_L \in \mathbb{S}^{d-1}$. Therefore, we have $\sum_{l=1}^L W_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu) \frac{f(W_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu))}{\sum_{j=1}^L W_p^p(\theta_j \sharp \mu, \theta_j \sharp \nu)} = 0$ for any $\theta_1, \ldots, \theta_L \in \mathbb{S}^{d-1}$ which implies $\text{IWRPSW}_p(\mu, \nu; \sigma_\kappa, L) = \mathbb{E}\left[\sum_{l=1}^L W_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu) \frac{f(W_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu))}{\sum_{j=1}^L W_p^p(\theta_j \sharp \mu, \theta_j \sharp \nu)}\right] = 0$. In the reverse direction, when $\text{IWRPSW}_p(\mu, \nu; \sigma_\kappa, L) = 0$, it means that we have $\sum_{l=1}^L W_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu) \frac{f(W_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu))}{\sum_{j=1}^L W_p^p(\theta_j \sharp \mu, \theta_j \sharp \nu)} = 0$ for any $\theta_1, \ldots, \theta_L \in \mathbb{S}^{d-1}$. Since $f(W_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu)) > 0$ for any θ_j, it implies that $W_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu) = 0$ for all $\theta_l \in \mathbb{S}^{d-1}$. With similar arguments to the proof of RPSW, we obtain $\mu = \nu$ which completes the proof.

Quasi-Triangle Inequality. Given three probability measures $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_p(\mathbb{R}^d)$ we have:

$$\begin{aligned}
\text{RPSW}_p(\mu_1, \mu_2; \sigma_\kappa) &= \left(\mathbb{E}_{\theta \sim \sigma_{RP}(\theta; \mu_1, \mu_2, \sigma_\kappa)}[W_p^p(\theta \sharp \mu_1, \theta \sharp \mu_2)]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}_{\theta \sim \sigma_{RP}(\theta; \mu_1, \mu_2, \sigma_\kappa)}[(W_p(\theta \sharp \mu_1, \theta \sharp \mu_3) + W_p(\theta \sharp \mu_3, \theta \sharp \mu_2))^p]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}_{\theta \sim \sigma_{RP}(\theta; \mu_1, \mu_2, \sigma_\kappa)}[W_p^p(\theta \sharp \mu_1, \theta \sharp \mu_3)]\right)^{\frac{1}{p}} + \left(\mathbb{E}_{\theta \sim \sigma_{RP}(\theta; \mu_1, \mu_2, \sigma_\kappa)}[W_p^p(\theta \sharp \mu_3, \theta \sharp \mu_2)]\right)^{\frac{1}{p}} \\
&= \text{RPSW}_p(\mu_1, \mu_3; \sigma_\kappa, \mu_1, \mu_2) + \text{RPSW}_p(\mu_3, \mu_2; \sigma_\kappa, \mu_1, \mu_2),
\end{aligned}$$

where the first inequality is due to the triangle inequality of Wasserstein distance and the second inequality is due to the Minkowski inequality. We complete the proof here.

A.2. Proof of Proposition 1

(i) To prove that $\text{RPSW}_p(\mu, \nu; \sigma_\kappa) \leq \text{IWRPSW}_p(\mu, \nu; \sigma_\kappa, L)$, we introduce the following lemma which had been proved in (Nguyen & Ho, 2023). Here, we provide the proof for completeness.

Lemma 1. *For any $L \geq 1$, $0 \leq a_1 \leq a_2 \leq \ldots \leq a_L$ and $0 < b_1 \leq b_2 \leq \ldots \leq b_L$, we have:*

$$\frac{1}{L}(\sum_{i=1}^L a_i)(\sum_{i=1}^L b_i) \leq \sum_{i=1}^L a_i b_i. \tag{10}$$

Proof. We prove Lemma 1 via an induction argument. For $L = 1$, it is clear that $a_i b_i = a_i b_i$. Now, we assume that the inequality holds for L i.e., $\frac{1}{L}(\sum_{i=1}^L a_i)(\sum_{i=1}^L b_i) \leq \sum_{i=1}^L a_i b_i$ or $(\sum_{i=1}^L a_i)(\sum_{i=1}^L b_i) \leq L \sum_{i=1}^L a_i b_i$. Now, we want to show that the inequality holds for $L+1$ i.e., $(\sum_{i=1}^{L+1} a_i)(\sum_{i=1}^L b_i) \leq (L+1) \sum_{i=1}^{L+1} a_i b_i$. First, we have:

$$\begin{aligned}
(\sum_{i=1}^{L+1} a_i)(\sum_{i=1}^{L+1} b_i) &= (\sum_{i=1}^L a_i)(\sum_{i=1}^L b_i) + (\sum_{i=1}^L a_i) b_{L+1} + (\sum_{i=1}^L b_i) a_{L+1} + a_{L+1} b_{L+1} \\
&\leq L \sum_{i=1}^L a_i b_i + (\sum_{i=1}^L a_i) b_{L+1} + (\sum_{i=1}^L b_i) a_{L+1} + a_{L+1} b_{L+1}.
\end{aligned}$$

By rearrangement inequality, we have $a_{L+1} b_{L+1} + a_i b_i \geq a_{L+1} b_i + b_{L+1} a_i$ for all $1 \leq i \leq L$. By taking the sum of these inequalities over i from 1 to L, we obtain:

$$(\sum_{i=1}^L a_i) b_{L+1} + (\sum_{i=1}^L b_i) a_{L+1} \leq \sum_{i=1}^L a_i b_i + L a_{L+1} b_{L+1}.$$

Therefore, we have

$$
\begin{aligned}
(\sum_{i=1}^{L+1} a_i)(\sum_{i=1}^{L+1} b_i) &\leq L\sum_{i=1}^{L} a_i b_i + (\sum_{i=1}^{L} a_i) b_{L+1} + (\sum_{i=1}^{L} b_i) a_{L+1} + a_{L+1} b_{L+1} \\
&\leq L\sum_{i=1}^{L} a_i b_i + \sum_{i=1}^{L} a_i b_i + L a_{L+1} b_{L+1} + a_{L+1} b_{L+1} \\
&= (L+1)(\sum_{i=1}^{L+1} a_i b_i),
\end{aligned}
$$

which completes the proof. □

From Lemma 1, with $a_i = W_p^p(\theta_i \sharp \mu, \theta_j \sharp \nu)$ and $b_i = f(W_p^p(\theta_i \sharp \mu, \theta_j \sharp \nu))$, we have:

$$
\frac{1}{L}\sum_{i=1}^{l} W_p^p(\theta_i \sharp \mu, \theta_i \sharp \nu) \leq \sum_{i=1}^{L} W_p^p(\theta_i \sharp \mu, \theta_i \sharp \nu) \frac{f(W_p^p(\theta_i \sharp \mu, \theta_i \sharp \nu))}{\sum_{j=1}^{L} f(W_p^p(\theta_j \sharp \mu, \theta_j \sharp \nu))}.
$$

Taking the expectation with respect to $\theta_1, \ldots, \theta_L \overset{i.i.d}{\sim} \sigma_{RP}(\theta; \mu, \nu, \sigma_\kappa)$, we obtain $\text{RPSW}_p(\mu, \nu; \sigma_\kappa) \leq \text{IWRPSW}_p(\mu, \nu; \sigma_\kappa, L)$.

Now to show that $\text{IWRPSW}_p(\mu, \nu; \sigma_\kappa, L) \leq \text{Max-SW}(\mu, \nu)$, we have $\theta^\star = \text{argmax}_{\theta \in \mathbb{S}^{d-1}} W_p(\theta \sharp \mu, \theta \sharp \nu)$ since $\mathbb{S}^{d-1}$ is compact and the function $\theta \to W_p(\theta \sharp \mu, \theta \sharp \nu)$ is continuous. From the definition of the IWRPSW, for any $L \geq 1, \sigma_\kappa \in \mathcal{P}(\mathbb{S}^{d-1})$ we have:

$$
\begin{aligned}
\text{IWRPSW}_p(\mu, \nu; \sigma_\kappa, L) &= \left(\mathbb{E}_{\theta_1, \ldots, \theta_L \sim \sigma_{RP}(\theta; \mu, \nu, \sigma_\kappa)} \left[\sum_{l=1}^{L} W_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu) \frac{f(W_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu))}{\sum_{j=1}^{L} f(W_p^p(\theta_j \sharp \mu, \theta_j \sharp \nu))} \right] \right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}_{\theta \sim \sigma_{RP}(\theta; \mu, \nu, \sigma_\kappa)} \left[W_p^p\left(\theta^\star \sharp \mu, \theta^\star \sharp \nu\right) \right] \right)^{\frac{1}{p}} = W_p^p\left(\theta^\star \sharp \mu, \theta^\star \sharp \nu\right) = \text{Max-SW}_p(\mu, \nu).
\end{aligned}
$$

Furthermore, by applying the Cauchy-Schwartz inequality, we have:

$$
\begin{aligned}
\text{Max-SW}_p(\mu, \nu) &= \left(\max_{\theta \in \mathbb{S}^{d-1}} \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d} \left| \theta^\top x - \theta^\top y \right|^p d\pi(x, y) \right) \right)^{\frac{1}{p}} \\
&\leq \left(\max_{\theta \in \mathbb{S}^{d-1}} \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\theta\|^p \|x - y\|^p d\pi(x, y) \right) \right)^{\frac{1}{p}} \\
&= \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\theta\|^p \|x - y\|^p d\pi(x, y) \right)^{\frac{1}{p}} \\
&= \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi(x, y) \right)^{\frac{1}{p}} \\
&\leq \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^p d\pi(x, y) \right)^{\frac{1}{p}} \\
&= W_p(\mu, \nu),
\end{aligned}
$$

we complete the proof.

(ii) We first recall the density of The von-Mises Fisher distribution and the Power spherical distribution. In particular, we have $vMF(\theta; \epsilon, \kappa) := \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \exp(\kappa \epsilon^\top \theta)$ with I_s denotes the modified Bessel function of the first kind, and the

PS distribution $PS(\theta;\epsilon,\kappa) = \left(2^{d+\kappa-1}\pi^{(d-1)/2}\frac{\Gamma((d-1)/2+\kappa)}{\Gamma(d+\kappa-1)}\right)^{-1}(1+\epsilon^\top\theta)^\kappa$. We have:

$$\lim_{\kappa\to 0}\frac{\kappa^{d/2-1}}{(2\pi)^{d/2}I_{d/2-1}(\kappa)}\exp(\kappa\epsilon^\top\theta) \to C_1,$$
$$\lim_{\kappa\to 0}\left(2^{d+\kappa-1}\pi^{(d-1)/2}\frac{\Gamma((d-1)/2+\kappa)}{\Gamma(d+\kappa-1)}\right)^{-1}(1+\epsilon^\top\theta)^\kappa \to C_2$$

for some constant C_1 and C_2 which do not depend on θ, hence, the vMF distribution and the PS distribution converge to the uniform distribution when $\kappa \to 0$. Therefore, we have $vMF(\theta;\epsilon,\kappa) \to \mathcal{U}(\mathbb{S}^{d-1})$ and $PS(\theta;\epsilon,\kappa) \to \mathcal{U}(\mathbb{S}^{d-1})$. Therefore, we have $\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa) \to \mathcal{U}(\mathbb{S}^{d-1})$.

Now, we need to show that $W_p^p(\theta\sharp\mu,\theta\sharp\nu)$ is bounded and continuous with respective to θ. For the boundedness, we have:

$$\begin{aligned}
\mathrm{W}_p^p(\theta\sharp\mu,\theta\sharp\nu) &= \inf_{\pi\in\Pi(\nu,\mu)}\int_{\mathbb{R}^d}|\theta^\top x-\theta^\top y|^p d\pi(x,y)\\
&\leq \inf_{\pi\in\Pi(\nu,\mu)}\int_{\mathbb{R}^d}\|x-y\|^p d\pi(x,y)\\
&= \mathrm{W}_p^p(\mu,\nu) < \infty.
\end{aligned}$$

For the continuity, let $(\theta_t)_{t\geq 1}$ be a sequence on $\mathbb{S}^{d-1}$ which converges to $\theta \in \mathbb{S}^{d-1}$ i.e., $\|\theta_t-\theta\| \to 0$ as $t\to\infty$, and a arbitrary measure $\mu \in \mathcal{P}_p(\mathbb{R}^d)$. Then we have:

$$\begin{aligned}
\mathrm{W}_p(\theta\sharp\mu,\theta_t\sharp\mu) &= \left(\inf_{\pi\in\Pi(\mu,\mu)}\int_{\mathbb{R}^d}|\theta_t^\top x-\theta^\top y|^p d\pi(x,y)\right)^{1/p}\\
&\leq \left(\int_{\mathbb{R}^d}|\theta_t^\top x-\theta^\top x|^p d\mu(x)\right)^{1/p}\\
&\leq \left(\int_{\mathbb{R}^d}\|x\|^p\mu(dx)\right)^{1/p}\|\theta_t-\theta\| \to 0 \quad \text{as } t\to\infty,
\end{aligned}$$

where $\left(\int_{\mathbb{R}^d}\|x\|^p\mu(dx)\right)^{1/p} < \infty$ since $\mu\in\mathcal{P}_p(\mathbb{R}^d)$, and the second inequality is due to the Cauchy-Schwartz inequality.

Using the triangle inequality, we have:

$$\begin{aligned}
|\mathrm{W}_p(\theta_t\sharp\mu,\theta_t\sharp\nu)-\mathrm{W}_p(\theta\sharp\mu,\theta\sharp\nu)| &\leq |\mathrm{W}_p(\theta_t\sharp\mu,\theta_t\sharp\nu)-\mathrm{W}_p(\theta\sharp\mu,\theta_t\sharp\nu)| + |\mathrm{W}_p(\theta\sharp\mu,\theta_t\sharp\nu)-\mathrm{W}_p(\theta\sharp\mu,\theta\sharp\nu)|\\
&\leq \mathrm{W}_p(\theta\sharp\mu,\theta_t\sharp\mu)+\mathrm{W}_p(\theta\sharp\nu,\theta_t\sharp\nu) \to 0 \quad \text{as } t\to\infty,
\end{aligned}$$

hence, $\mathrm{W}_p(\theta_t\sharp\mu,\theta_t\sharp\nu) \to \mathrm{W}_p(\theta\sharp\mu,\theta\sharp\nu)$ as $t\to\infty$ which complete the proof of continuity.

From the boundedness, continuity, and $\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa) \to \mathcal{U}(\mathbb{S}^{d-1})$, we have

$$\mathrm{RPSW}_p^p(\mu,\nu;\sigma_\kappa) = \mathbb{E}_{\theta\sim\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)}[W_p(\theta\sharp\mu,\theta\sharp\nu)] \to \mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^{d-1})}[W_p(\theta\sharp\mu,\theta\sharp\nu)] = SW_p^p(\mu,\nu).$$

Applying the continuous mapping theorem for $x \to x^{1/p}$, we obtain $\lim_{\kappa\to 0}\mathrm{RPSW}_p(\mu,\nu;\sigma_\kappa) \to SW_p(\mu,\nu)$.

(iii) Since we have proved that $W_p^p(\theta\sharp\mu,\theta\sharp\nu)$ is bounded and continuous with respective to θ, we can show that $\sum_{i=1}^L W_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu)\frac{f(W_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu))}{\sum_{j=1}^L f(W_p^p(\theta_j\sharp\mu,\theta_j\sharp\nu))}$ are bounded and continuous with respect to $\theta_1,\ldots,\theta_L$. As $L\to\infty$, we have $\sum_{i=1}^L W_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu)\frac{f(W_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu))}{\sum_{j=1}^L f(W_p^p(\theta_j\sharp\mu,\theta_j\sharp\nu))} \to \mathbb{E}_{\gamma\sim\sigma_{\mu,\nu,f}(\gamma)}[\gamma\sharp\mu,\gamma\sharp\nu] = \mathrm{EBSW}_p^p(\mu,\nu;f)$. Applying the continuous mapping theorem for $x\to x^{1/p}$, we obtain $\lim_{L\to\infty}\mathrm{IWRPSW}_p(\mu,\nu;\sigma_\kappa,L,f) \to \mathrm{EBSW}_p(\mu,\nu;f)$.

(iv) For any μ and ν, by the Holder inequality, we have:

$$\begin{aligned}
\mathrm{RPSW}_p(\mu,\nu;\sigma_\kappa) &= \left(\mathbb{E}_{\theta\sim\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)}[W_p^p(\theta\sharp\mu,\theta\sharp\nu)]\right)^{\frac{1}{p}}\\
&\geq \mathbb{E}_{\theta\sim\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)}[W_p(\theta\sharp\mu,\theta\sharp\nu)]\\
&\geq \mathbb{E}_{\theta\sim\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)}[W_1(\theta\sharp\mu,\theta\sharp\nu)]\\
&= \mathrm{RPSW}_1(\mu,\nu;\sigma_\kappa),
\end{aligned}$$

which completes the proof.

A.3. Proof of Proposition 2

The proof for this result follows from the proof of Proposition 3 in (Nguyen & Ho, 2023). We assume that μ has a compact set of support $\mathcal{X} \in \mathbb{R}^d$.

From Proposition 1, we have:

$$\mathbb{E}[\text{RPSW}_p(\mu_n, \mu; \sigma_\kappa)] \leq \mathbb{E}[\text{IWRPSW}_p(\mu_n, \mu; \sigma_\kappa, L)] \leq \mathbb{E}\left[\text{Max-SW}_p(\mu_n, \mu)\right],$$

for any $\sigma_\kappa \in \mathcal{P}(\mathbb{S}^{d-1})$ and $L \geq 1$. Therefore, the proposition follows as long as we can demonstrate that

$$\mathbb{E}[\text{Max-SW}_p(\mu_n, \mu)] \leq C\sqrt{(d+1)\log n/n}$$

where $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ with $X_1, \ldots, X_n \overset{i.i.d}{\sim} \mu$, and $C > 0$ is some universal constant and the outer expectation is taken with respect to $X_1, \ldots, X_n$.

Using the closed-form of one-dimensional Wasserstein distance, we have:

$$\text{Max-SW}_p(\mu_n, \mu) = \max_{\theta \in \mathbb{S}^{d-1}} \int_0^1 |F_{n,\theta}^{-1}(z) - F_\theta^{-1}(z)|^p dz,$$

where $F_{n,\theta}$ and F_θ as the cumulative distributions of $\theta\sharp\mu_n$ and $\theta\sharp\mu$. Since $W_p((t\theta)\sharp\mu, (t\theta)\sharp\nu) = tW_p(\theta\sharp\mu, \theta\sharp\nu)$ for $t > 0$. We can rewrite Max-SW as:

$$\begin{aligned}
\text{Max-SW}_p^p(\mu_n, \mu) &= \max_{\theta \in \mathbb{R}^d : \|\theta\| = 1} \int_0^1 |F_{n,\theta}^{-1}(z) - F_\theta^{-1}(z)|^p dz \\
&\leq \text{diam}(\mathcal{X}) \max_{x \in \mathbb{R}, \theta \in \mathbb{R}^d : \|\theta\| \leq 1} |F_{n,\theta}(x) - F_\theta(x)|^p \\
&= \text{diam}(\mathcal{X}) \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|,
\end{aligned}$$

where $\mathcal{A}$ is the set of half-spaces $\{z \in \mathbb{R}^d : \theta^\top z \leq x\}$ for all $\theta \in \mathbb{R}^d$ such that $\|\theta\| \leq 1$. From VC inequality (Theorem 12.5 in (Devroye et al., 2013)), we have

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |\mu_n(B) - \mu(A)| > t\right) \leq 8S(\mathcal{A}, n)e^{-nt^2/32}.$$

with $S(\mathcal{A}, n)$ is the growth function. From the Sauer Lemma (Proposition 4.18 in (Wainwright, 2019)), the growth function is upper bounded by $(n+1)^{VC(\mathcal{A})}$. Moreover, we can get $VC(\mathcal{A}) = d + 1$ from Example 4.21 in (Wainwright, 2019).

Let $8S(\mathcal{A}, n)e^{-nt^2/32} \leq \delta$, we have $t^2 \geq \frac{32}{n}\log\left(\frac{8S(\mathcal{A},n)}{\delta}\right)$. Therefore, we obtain

$$\mathbb{P}\left(\sup_{A \in \mathcal{B}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{32}{n}\log\left(\frac{8S(\mathcal{A}, n)}{\delta}\right)}\right) \geq 1 - \delta,$$

Using the Jensen inequality and the tail sum expectation for non-negative random variable, we have:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right] \\
&\leq \sqrt{\mathbb{E}\left[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right]^2} = \sqrt{\int_0^\infty \mathbb{P}\left(\left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right)^2 > t\right) dt} \\
&= \sqrt{\int_0^u \mathbb{P}\left(\left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right)^2 > t\right) dt + \int_u^\infty \mathbb{P}\left(\left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right)^2 > t\right) dt} \\
&\leq \sqrt{\int_0^u 1 dt + \int_u^\infty 8S(\mathcal{A}, n)e^{-nt/32} dt} = \sqrt{u + 256S(\mathcal{A}, n)\frac{e^{-nu/32}}{n}}.
\end{aligned}$$

Since the inequality holds for any u, we search for the best u that makes the inequality tight. Let $f(u) = u + 256S(\mathcal{A}, n)\frac{e^{-nu/32}}{n}$, we have $f'(u) = 1+8S(\mathcal{A}, n)e^{-nu/32}$. Setting $f'(u) = 0$, we obtain the minima $u^\star = \frac{32\log(8S(\mathcal{A},n))}{n}$. Plugging $u^\star$ in the inequality, we obtain:

$$\mathbb{E}\left[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right] \leq \sqrt{\frac{32\log(8S(\mathcal{A}, n))}{n} + 32} \leq C\sqrt{\frac{(d+1)\log(n+1)}{n}},$$

by using Sauer Lemma i.e., $S(\mathcal{A}, n) \leq (n+1)^{VC(\mathcal{A})} \leq (n+1)^{d+1}$. Putting the above results together leads to

$$\mathbb{E}[\text{Max-SW}_p(\mu_n, \mu)] \leq C\sqrt{(d+1)\log n/n},$$

where $C > 0$ is some universal constant. As a consequence, we obtain the conclusion of the proof.

A.4. Proof of Proposition 3

For any $p \geq 1$, $d \geq 1$, $\sigma_\kappa \in \mathcal{P}(\mathbb{S}^{d-1})$, and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, using the Holder's inequality, we have:

$$\begin{aligned}
&\mathbb{E}|\widehat{\text{RPSW}}_p^p(\mu, \nu; \sigma_\kappa) - \text{RPSW}_p^p(\mu, \nu; \sigma_\kappa)| \\
&\leq \left(\mathbb{E}|\widehat{\text{RPSW}}_p^p(\mu, \nu; \sigma_\kappa) - \text{RPSW}_p^p(\mu, \nu; \sigma_\kappa)|^2\right)^{\frac{1}{2}} \\
&= \left(\mathbb{E}\left(\frac{1}{L}\sum_{l=1}^{L} \text{W}_p^p(\theta_l\sharp\mu, \theta_l\sharp\nu) - \mathbb{E}_{\theta\sim\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)}\left[W_p^p\left(\theta\sharp\mu, \theta\sharp\nu\right)\right]\right)^2\right)^{\frac{1}{2}}
\end{aligned}$$

Since $\mathbb{E}[\frac{1}{L}\sum_{l=1}^{L} \text{W}_p^p(\theta_l\sharp\mu, \theta_l\sharp\nu)] = \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}[\text{W}_p^p(\theta_l\sharp\mu, \theta_l\sharp\nu)] = \frac{1}{L}\sum_{l=1}^{L} \widehat{\text{RPSW}}_p^p(\mu, \nu; \sigma_\kappa) = \widehat{\text{RPSW}}_p^p(\mu, \nu; \sigma_\kappa)$, we have:

$$\begin{aligned}
\mathbb{E}|\widehat{\text{RPSW}}_p^p(\mu, \nu; \sigma_\kappa) - \text{RPSW}_p^p(\mu, \nu; \sigma_\kappa)| &\leq \left(Var_{\theta\sim\sigma_{RP}(\theta;\mu,\nu,\sigma_\kappa)}\left[\frac{1}{L}\sum_{l=1}^{L} W_p^p\left(\theta_l\sharp\mu, \theta_l\sharp\nu\right)\right]\right)^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{L}} Var\left[W_p^p\left(\theta\sharp\mu, \theta\sharp\nu\right)\right]^{\frac{1}{2}},
\end{aligned}$$

since $\theta_1, \ldots, \theta_L \overset{i.i.d}{\sim} \sigma_{RP}(\theta; \mu, \nu, \sigma_\kappa)$, which completes the proof.

B. Related Works

Dynamic Optimal Transport. Wasserstein distance can also be seen as the shortest curve between two input measures (Villani, 2008) which creates an optimal interpolation path between two measures (McCann, 1997). In this work, we work with static sliced Wasserstein distance where the random-path is only used to construct discriminative projecting directions.

Sliced Wasserstein on Manifolds. On manifolds where a straight line is generalized into a curse, the random-path definition should be different. For example, on the sphere, the shortest path between two points is the great circle. Therefore, it is natural to define the random-path as a random great circle with two endpoints that are randomly drawn from a coupling between two measures. This definition can lead to a natural extension for spherical Sliced Wasserstein (Bonet et al., 2023a). Moreover, we can also create a similar definition on hyperbolic manifolds for hyperbolic sliced Wasserstein (Bonet et al., 2023b) and on the manifold of positive definite matrix (Bonet et al., 2023c).

Dependent Projecting Directions. Markovian Sliced Wasserstein distance is introduced in (Nguyen et al., 2023). It is a variant of SW that utilizes dependent projecting directions i.e., follows a Markovian structure joint distribution. The contribution of random-path slicing distribution is orthogonal to the mentioned approach since as can be used in either the prior distribution or the transition distribution in the Markovian process. In this work, we focus on slicing distribution selection variants of SW without dependency structure which can be computed very fast in parallel.

Properties of sliced Wasserstein losses. In this work, we use sliced Wasserstein distance as a loss for gradient flows and generative modeling. We refer the reader to (Tanguy, 2023) for a more discussion about its properties i.e., regularity, gradient definition, and so on.

Algorithm 1 Computational algorithm of RPSW

Input: Probability measures μ and ν, $p \geq 1$, the number of projections L, $0 < \kappa < \infty$
for $l = 1$ to L **do**
  Sample $X \sim \mu$, $Y \sim \nu$
  Sample $\theta_l \sim \sigma_\kappa \left(\theta; \frac{X-Y}{\|X-Y\|_2}\right)$
  Compute $v_l = \mathrm{W}_p(\theta_l \sharp \mu, \theta_l \sharp \nu)$
end for
Compute $\widehat{\mathrm{RPSW}}_p(\mu, \nu; L, \sigma_\kappa) = \left(\frac{1}{L}\sum_{l=1}^L v_l\right)^{\frac{1}{p}}$
Return: $\widehat{\mathrm{RPSW}}_p(\mu, \nu; L, \sigma_\kappa)$

Algorithm 2 Computational algorithm of the IWRPSW

Input: Probability measures μ and ν, $p \geq 1$, the number of projections L, $0 < \kappa < \infty$, and the energy function f.
for $l = 1$ to L **do**
  Sample $X \sim \mu$, $Y \sim \nu$
  Sample $\theta_l \sim \sigma_\kappa \left(\theta; \frac{X-Y}{\|X-Y\|_2}\right)$
  Compute $v_l = \mathrm{W}_p(\theta_l \sharp \mu, \theta_l \sharp \nu)$
  Compute $w_l = f(\mathrm{W}_p(\theta_l \sharp \mu, \theta_l \sharp \nu))$
end for
Compute $\widehat{\mathrm{IWRPSW}}_p(\mu, \nu; \sigma_\kappa, L, f) = \left(\sum_{l=1}^L v_l \frac{w_l}{\sum_{i=1}^L w_i}\right)^{\frac{1}{p}}$
Return: $\widehat{\mathrm{IWRPSW}}_p(\mu, \nu; \sigma_\kappa, L, f)$

Unbalanced Sliced Wasserstein. Random-path projecting direction can be applied directly to SW variants with unbalanced transport e.g., sliced partial optimal transport (Bonneel & Coeurjolly, 2019; Bai et al., 2022), sliced unbalanced optimal transport (Séjourné et al., 2023).

Other potential applications. Since RPSW and IWRPSW can be used as a replacement for SW, they can be applied in domain adaptation (Lee et al., 2019), point-cloud applications (Bai et al., 2023; Nguyen et al., 2024), 3D shapes matching and deformation (Le et al., 2024a;b), Bayesian inference (Yi & Liu, 2021; Nadjahi et al., 2020a), blue noise sampling (Paulin et al., 2020) and so on.

Other potential future directions. We consider extending the proposed ideas of random-path direction for a line in SW into OT problem with more general structures on supports, e.g., for a tree in tree-Wasserstein (Le et al., 2019), and for a graph in Sobolev transport (Le et al., 2022).

C. Algorithms

Computational algorithms of RPSW and IWRPSW. We present the pseudo-codes for computing RPSW and IWRPSW with Monte Carlo estimation in Algorithm 1 and Algorithm 2.

Algorithm for training denoising diffusion models with augmented mini-batch energy distance. The detailed algorithm is given in Algorithm 3.

D. Additional Experimental Details

Gradient flows with different gradient estimators. We run experiments on gradient flow with the original gradient estimators of RPSW and IWRPSW in Figure 4. We see that the original gradient estimators lead to faster convergence in terms of Wasserstein-2 distances. However, the inner topology between points, presented in colors, is destroyed especially for RPSW. The reason for this behavior is because of the independent coupling in constructing random paths. Therefore, we suggest using the simplified gradient estimator and treating the random-path slicing distribution as a separate distribution for projecting direction selection.

Algorithm 3 Training denoising diffusion models with augmented mini-batch energy distance

Input: $q(x_{1:T}|x_0) = \prod_{t\geq 1} q(x_t|x_{t-1})$, $p_\phi(x_{0:T}) = p(x_T)\prod_{t\geq 1} p_\phi(x_{t-1}|x_t)$, $p_\phi(x_{t-1}|x_t) = \int p(\epsilon) q(x_{t-1}|x_t, x_0 = G_\phi(x_t, \epsilon, t))$ with $p(\epsilon) \in \mathcal{P}(\mathbb{R}^z)$ and $G_\phi : \mathbb{R}^d \times \mathbb{R}^z \times \mathbb{R} \to \mathbb{R}^d$, $D_\gamma : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to [0, 1]$, metric $\mathcal{D}$.
while ϕ not converged or not reaching the maximum epochs **do**
  Sample a mini-batch $x_{0,1}, \ldots, x_{0,m} \sim q(x_0)$ with $m \geq 2$.
  Sample $t \sim \mathcal{U}(1, \ldots, T)$.
  Sample a mini-batch $x_{t-1,1} \sim q(x_{t-1}|x_{0,1}), \ldots, x_{t-1,m} \sim q(x_{t-1}|x_{0,m})$.
  Sample a mini-batch $x_{t,1} \sim q(x_t|x_{t-1,1}), \ldots, x_{t,m} \sim q(x_t|x_{t-1,m})$.
  Sample a mini-batch of noise $\epsilon_1, \ldots, \epsilon_m \sim p(\epsilon)$.
  Set $y_{0,1}, \ldots, y_{0,m} = G_\phi(x_{t,1}, \epsilon_1, t), \ldots, G_\phi(x_{t,m}, \epsilon_m, t)$.
  Sample a mini-batch $y_{t-1,1} \sim q(x_{t-1}|x_{t,1}, y_{0,1}), \ldots, y_{t-1,m} \sim q(x_{t-1}|x_{t,m}, y_{0,m})$.
  Update γ with $\nabla_\gamma \left(\frac{1}{m}\sum_{i=1}^m -\log(D_\gamma(x_{t-1,i}, x_{t,i}, t))\right)$.
  Update γ with $\nabla_\gamma \left(\frac{1}{m}\sum_{i=1}^m -\log(1 - D_\gamma(y_{t-1,i}, x_{t,i}, t))\right)$.
  Set $\bar{X} = ((x_{t-1,1}, D_\gamma(x_{t-1,1}, x_{t,1}, t)), \ldots, (x_{t-1,m}, D_\gamma(x_{t-1,m}, x_{t,m}, t)))$.
  Set $\bar{Y} = ((y_{t-1,1}, D_\gamma(y_{t-1,1}, x_{t,1}, t)), \ldots, (y_{t-1,m}, D_\gamma(y_{t-1,m}, x_{t,m}, t)))$.
  Set $\bar{X}_1, \bar{X}_2 = \bar{X}[: m/2], \bar{X}[m/2 :]$, and $\bar{Y}_1, \bar{Y}_2 = \bar{Y}[: m/2], \bar{Y}[m/2 :]$
  Update ϕ with $\nabla_\phi \left(2\mathcal{D}^2(\bar{X}, \bar{Y}) - \mathcal{D}^2(\bar{X}_1, \bar{X}_2) - \mathcal{D}^2(\bar{Y}_1, \bar{Y}_2)\right)$
end while
Return: ϕ

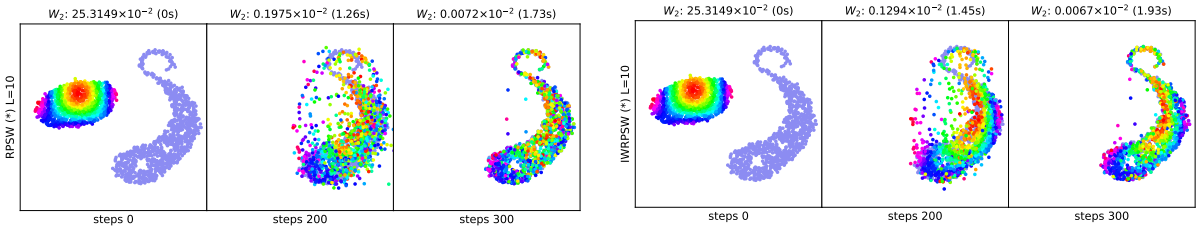


Figure 4. Results for gradient flows that are from the empirical distribution over the color points to the empirical distribution over S-shape points produced by different RPSW and IWRPSW with original gradient estimator.

Gradient flows on MNIST digits. We present the full visualization of the gradient flows on MNIST in Figure 5.

Neural network architecture for diffusion model. Our generator follows the U-net structure in (Xiao et al., 2021) which has 2 ResNet blocks per scale, 128 initial channels, channel multiplier for each scale: (1, 2, 2, 2), scale of attention block: 16, latent dimension 256, 3 latent mapping layers, latent embedding dimension: 512. For the discriminator, the order of the layers are 1 × 1 conv2d, 128 → ResBlock, 128 → ResBlock down, 256 → ResBlock down, 512 → ResBlock down, 512 → minibatch std layer → Global Sum Pooling → FC layer to scalar. For other settings, we refer the reader to (Xiao et al., 2021, Section C).

Hyper-parameters. We set initial learning rate for discriminator to 10^{-4}, initial learning rate for generator to 1.6×10^{-4}, Adam optimizer with parameters $(0.5, 0.9)$, EMA to 0.9999, batch-size to 256. For the learning rate scheduler, we use cosine learning rate decay.

E. Computational Infrastructure

For the gradient flow experiments, we use a HP Omen 25L desktop for conducting experiments. For diffusion model experiments, we use a single NVIDIA A100 GPU.

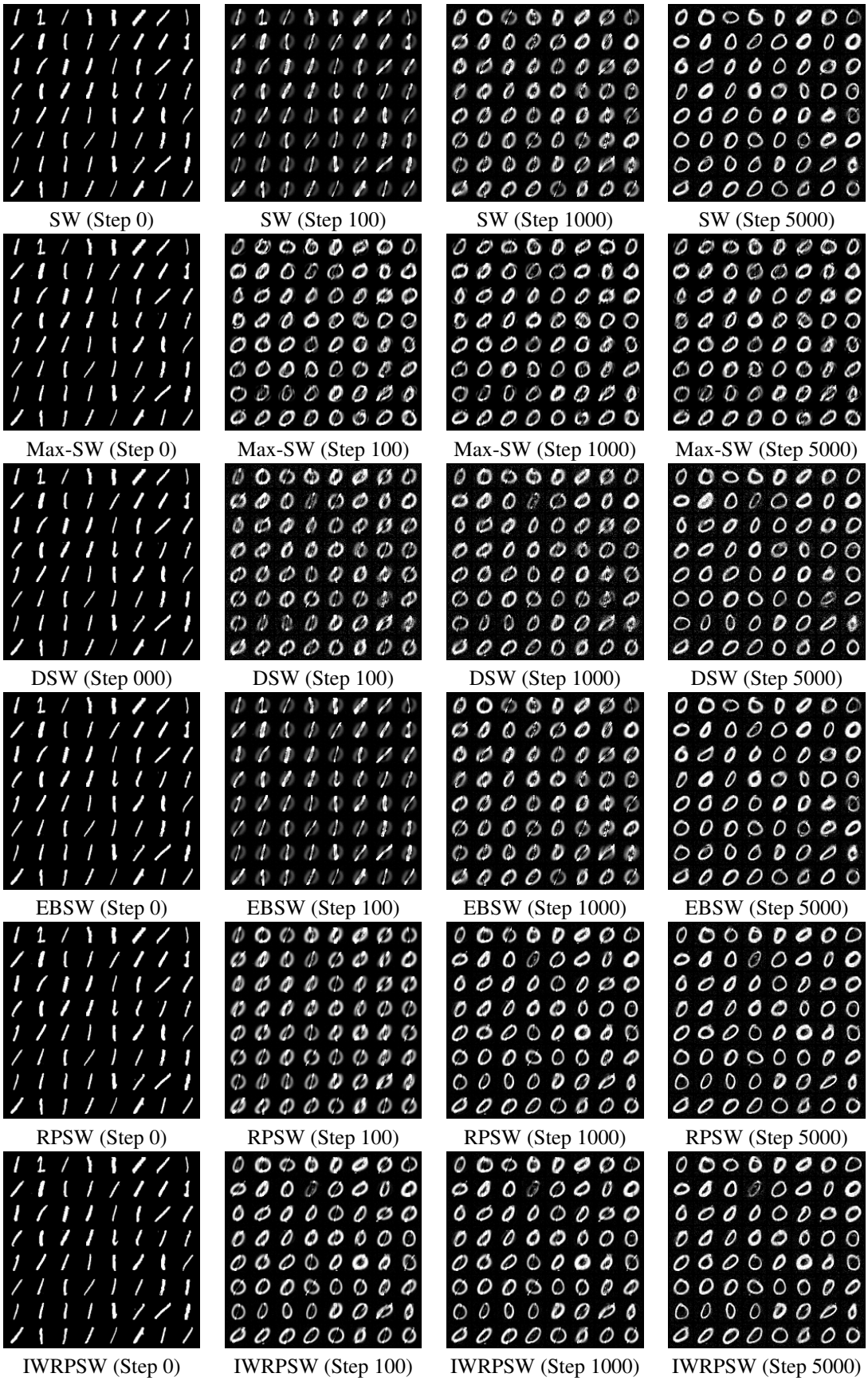


Figure 5. Gradient flows from MNIST digit 1 to MNIST digit 0 .