Marginal Fairness Sliced Wasserstein Barycenter

Khai Nguyen [1] **Hai Nguyen** [2] **Nhat Ho** [1]

Abstract

The Sliced Wasserstein barycenter (SWB) is widely acknowledged for generalizing the averaging operation within probability measure spaces but achieving marginal fairness SWB, ensuring approximately equal distances from the barycenter to marginals, remains unexplored as the uniform weighted SWB might not be the optimal choice due to structure of marginals and the non-optimality of the optimizations. We introduce the marginal fairness sliced Wasserstein barycenter (MFSWB) problem as a constrained SWB problem, and proposes three surrogate MFSWB problems that implicitly minimize distances to marginals and encourage marginal fairness, then discusses their relationship to the sliced multi-marginal Wasserstein distance. Finally, we conduct experiments on finding 3D point-clouds averaging, color harmonization, and training of sliced Wasserstein autoencoder with class-fairness representation to show the favorable performance of the proposed surrogate MFSWB problems.

1. Introduction

Wasserstein barycenter (Agueh & Carlier, 2011) generalizes "averaging" to the space of probability measures. In particular, a Wasserstein barycenter is a probability measure that minimizes a weighted sum of Wasserstein distances between it and some given marginal probability measures. Due to the rich geometry, Wasserstein barycenter has been applied widely to various applications in machine learning such as Bayesian inference (Srivastava et al., 2018; Staib et al., 2017), domain adaptation (Montesuma & Mboula, 2021), clustering (Ho et al., 2017), sensor fusion (Elvander et al., 2018), text classification (Kusner et al., 2015), and so on. Moreover, Wasserstein barycenter is also a powerful tool for computer graphics since it can be used for texture mixing (Rabin et al., 2012), style transfer (Mroueh, 2020), shape interpolation (Solomon et al., 2015), and many other tasks on many other domains.

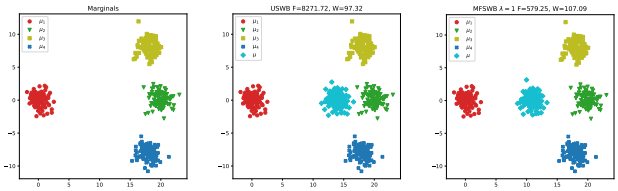


Figure 1. The uniform SWB and the MFSWB of 4 Gaussian distributions.

Computing Wasserstein barycenters is computationally expensive, ranging from $\mathcal{O}(n^3 \log n)$ using linear programming (Anderes et al., 2016) to $\mathcal{O}(n^2)$ with entropic regularization (Cuturi, 2013). To address this, sliced Wasserstein barycenter (SWB) was introduced by Bonneel et al. (Bonneel et al., 2015), offering a time complexity of $\mathcal{O}(n \log n)$. SWB benefits from the equivalence of sliced Wasserstein, making it a scalable alternative choice for Wasserstein barycenters.

In some applications, we might want to find a barycenter that minimizes the distances to marginals and has equal distances to marginals at the same time e.g., constructing shape template for a group of shapes (Bongratz et al., 2022; Sun et al., 2023) that can be further used in downstream tasks, exact balance style mixing between images (Bonneel et al., 2015), fair generative modeling (Choi et al., 2020), and so on. However, obtaining such a barycenter is challenging. Even though uniform barycenter weights are commonly used, they do not guarantee a marginal fairness barycenter as shown in Figure 1. To the best of our knowledge, there is no prior work that investigates finding a marginal fairness barycenter.

In this work, we make the first attempt to tackle the marginal fairness barycenter problem i.e., we focus on finding marginal fairness SW barycenter (MFSWB) to utilize the scalability of SW distance.

Contribution: Our main contributions are four-fold:

1. We define the marginal fairness SW barycenter (MFSWB) problem which is a constrained barycenter problem, where the constraint tries to limit the average pair-wise absolute difference between distances from the barycenter to

[1]Department of Statistics and Data Sciences, University of Texas at Austin, USA [2]VinAI Research, Vietnam. Correspondence to: Khai Nguyen <khainb@utexas.edu>.

ICML 2024 Workshop on the Next Generation of AI Safety, Vienna, Austria.

marginals. We derive the dual form of MFSWB, discuss its computation, and address its computational challenges.

2. We propose hyperparameter-free and computationally tractable surrogate definitions of MFSWB. Inspired by Fair PCA (Samadi et al., 2018), our first surrogate minimizes the largest SW distance from the barycenter to the marginals. Then, we introduce a second surrogate which is the improvement of the first one, aiming to provide an unbiased gradient estimator. We extend this to a third surrogate using slicing distribution selection, proving it to be an upper bound of the previous two.

3. We connect the proposed surrogate MFSWB problems with sliced multi-marginal Wasserstein (SMW) distance using the maximal ground metric. Solving our MFSWB problems equals minimizing a lower bound of the SMW.

4. We conduct simulations with Gaussian and experiments on various applications including 3D point-cloud averaging, color harmonization, and sliced Wasserstein autoencoder with class-fairness representation to demonstrate the favorable performance of the proposed surrogate definitions.

2. Preliminaries

Sliced Wasserstein distance. The sliced Wasserstein (SW) distance (Bonneel et al., 2015) between two probability measures $\mu_1 \in \mathcal{P}_p(\mathbb{R}^d)$ and $\mu_2 \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as:

$$\text{SW}_p^p(\mu_1, \mu_2) = \mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}[\text{W}_p^p(\theta\sharp\mu_1, \theta\sharp\mu_2)],$$

where the Wasserstein distance has a closed form in one-dimension which is $\text{W}_p^p(\theta\sharp\mu_1, \theta\sharp\mu_2) = \int_0^1 |F^{-1}_{\theta\sharp\mu_1}(z) - F^{-1}_{\theta\sharp\mu_2}(z)|^p dz$ where $F_{\theta\sharp\mu_1}$ and $F_{\theta\sharp\mu_2}$ are the cumulative distribution function (CDF) of $\theta\sharp\mu_1$ and $\theta\sharp\mu_2$ respectively.

Sliced Wasserstein Barycenter. The definition of the sliced Wasserstein barycenter (SWB) problem (Bonneel et al., 2015) of $K \geq 2$ marginals $\mu_{1:K} \in \mathcal{P}_p(\mathbb{R}^d)$ with marginal weights $\omega_{1:K} > 0$ ($\sum_{i=k}^K \omega_k = 1$) is defined as: $\min_\mu \mathcal{F}(\mu; \mu_{1:K}, \omega_{1:K}) \coloneqq \min_\mu \sum_{k=1}^K \omega_k \text{SW}_p^p(\mu, \mu_k)$.

Computation of parametric SWB. Let μ_ϕ be parameterized by $\phi \in \Phi$, SWB can be solved by gradient-based optimization. In that case, the interested quantity is the gradient $\nabla_\phi \mathcal{F}(\mu_\phi; \mu_{1:K}, \omega_{1:K}) = \sum_{k=1}^K \omega_k \nabla_\phi \text{SW}_p^p(\mu_\phi, \mu_k)$. However, the gradient of SW term is intractable due to the intractability of SW with the expectation with respect to the uniform distribution over the unit-hypersphere. Therefore, Monte Carlo estimation is used. With the stochastic gradient, the SWB can be solved by using a stochastic gradient descent algorithm. We refer the reader to Algorithm 1 in Appendix B for more detail. We discuss the discrete SWB i.e., marginals and the barycenter are discrete measures in Appendix B

Parametric SWB computation: The parameterized barycenter μ_ϕ with $\phi \in \Phi$ enables solving SWB through gradient-based optimization. In that case, the interested quantity is the gradient $\nabla_\phi \mathcal{F}(\mu_\phi; \mu_{1:K}, \omega_{1:K}) = \sum_{k=1}^K \omega_k \nabla_\phi \text{SW}_p^p(\mu_\phi, \mu_k)$. The intractable gradient of the Sliced Wasserstein term requires Monte Carlo estimation due to the expectation over the unit hypersphere. Utilizing stochastic gradient descent, SWB can be solved. Further details are provided in Algorithm 1 in Appendix B. The barycenter μ_ϕ can be expressed in two parameterizations: either as $\mu_\phi = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ with $\phi = (x_{1:n})$ (when ϕ denotes supports) or as $\mu_\phi = \sum_{i=1}^n \phi_i \delta_{x_i}$ (when ϕ represents weights). Both parameterizations allow for the computation of the gradient of SWB with respect to ϕ. When the supports or weights of the barycenter are outcomes of a parametric function, the gradient of the function's parameters can be estimated using the chain rule. The detailed computation of the parametric SWB gradient in both settings is provided in Appendix B.

Sliced Multi-marginal Wasserstein. Given $K \geq 1$ marginals $\mu_{1:K} \in \mathcal{P}_p(\mathbb{R}^d)$, sliced Multi-marginal Wasserstein (Cohen et al., 2021) (SMW) is defined as:

$$SMW_p^p(\mu_{1:K}; c) = \mathbb{E}\Big[\inf_{\pi \in \Pi(\mu_{1:K})} \int c(\theta^\top x_1, \ldots, \theta^\top x_K)^p d\pi(x_{1:K})\Big] \tag{1}$$

where the expectation is under $\theta \sim \mathcal{U}(\mathbb{S}^{d-1})$. When using the barycentric cost i.e., $c(\theta^\top x_1, \ldots, \theta^\top x_K)^p = \sum_{k=1}^K \beta_k \left|\theta^\top x_k - \sum_{k'=1}^K \beta_{k'} \theta^\top x_{k'}\right|^p$ for $\beta_k > 0 \forall k$ and $\sum_k \beta_k = 1$. Minimizing $SMW_p^p(\mu_{1:K}, \mu; c)$ with respect to μ is equivalent to a barycenter problem. We refer the reader to Proposition 7 in (Cohen et al., 2021) for detail.

3. Marginal Fairness Sliced Wasserstein Barycenter

3.1. Formal Definition

Now, we define the Marginal Fairness Sliced Wasserstein barycenter (MFSWB) problem by adding marginal fairness constraints to the SWB problem.

Definition 3.1. Given $K \geq 2$ marginals $\mu_{1:K} \in \mathcal{P}_p(\mathbb{R}^d)$, admissible $\epsilon \geq 0$ for $i = 1:K$ and $j = i+1:K$, the Marginal Fairness Sliced Wasserstein barycenter (MFSWB) is defined as:

$$\min_\mu \frac{1}{K}\sum_{k=1}^K \text{SW}_p^p(\mu, \mu_k)$$
$$\text{s.t. } \frac{2}{(K-1)K}\sum_{i=1}^K \sum_{j=i+1}^K |\text{SW}_p^p(\mu, \mu_i) - \text{SW}_p^p(\mu, \mu_j)| \leq \epsilon. \tag{2}$$

Duality objective. For admissible $\epsilon > 0$, there exist a Lagrange multiplier λ such that we have the dual form

Given $K \geq 2$ marginals $\mu_1, \ldots, \mu_K \in \mathcal{P}_p(\mathbb{R}^d)$, admissible $\epsilon \geq 0$ for $i = 1, \ldots, K$ and $j = i+1, \ldots, K$, the Marginal Fairness Sliced Wasserstein barycenter (MFSWB) is defined as:

$$\mathcal{L}(\mu, \lambda) = \frac{1}{K}\sum_{k=1}^{K} \text{SW}_p^p(\mu, \mu_k) + \frac{2\lambda}{(K-1)K}\sum_{i=1}^{K}\sum_{j=i+1}^{K} |\text{SW}_p^p(\mu, \mu_i) - \text{SW}_p^p(\mu, \mu_j)| - \lambda\epsilon. \tag{3}$$

Computational challenges. Firstly, MFSWB in Definition 3.1 requires an admissible $\epsilon > 0$ for the barycenter μ to exist, which is difficult to determine. Secondly, obtaining the optimal Lagrange multiplier $\lambda^\star$ in Equation (3) to minimize the duality gap is challenging and can result in weak duality. Thirdly, using Equation (3) requires hyperparameter tuning for λ and might not provide a good optimization landscape. Additionally, unbiased gradient estimates of ϕ for the parametric barycenter μ_ϕ are not possible due to the biased Monte Carlo estimation of the distance between SW distances. Finally, Equation (3) has quadratic time and space complexity, $\mathcal{O}(K^2)$, relative to the number of marginals.

3.2. Surrogate Definitions

First Surrogate Definition. Motivated by Fair PCA (Samadi et al., 2018), we propose a practical surrogate MFSWB problem that is hyperparameter-free.

Definition 3.2. Given $K \geq 2$ marginals $\mu_{1:K} \in \mathcal{P}_p(\mathbb{R}^d)$, the surrogate marginal fairness sliced Wasserstein barycenter (s-MFSWB) problem is defined as:

$$\min_{\mu} \mathcal{SF}(\mu; \mu_{1:K}); \quad \text{s.t. } \mathcal{SF}(\mu; \mu_{1:K}) = \max_{k \in \{1:K\}} SW_p^p(\mu, \mu_k). \tag{4}$$

The s-MFSWB problem tries to minimize the maximal distance from the barycenter to the marginals. Therefore, it can minimize indirectly the overall distances between the barycenter to the marginals and implicitly make the distances to marginals approximately the same. The downside is that the gradient estimator is biased.

Second Surrogate Definition. To address the biased gradient issue of the first surrogate problem, we propose the second surrogate MFSWB problem.

Definition 3.3. Given $K \geq 2$ marginals $\mu_{1:K} \in \mathcal{P}_p(\mathbb{R}^d)$, the unbiased surrogate marginal fairness sliced Wasserstein barycenter (us-MFSWB) problem is defined as:

$$\min_{\mu} \mathcal{USF}(\mu; \mu_{1:K}); \quad \text{s.t. } \mathcal{USF}(\mu; \mu_{1:K}) = \mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}\left[\max_{k \in \{1:K\}} W_p^p(\theta\sharp\mu, \theta\sharp\mu_k)\right]. \tag{5}$$

In contrast to s-MFSWB which minimizes the maximal SW distance among marginals, us-MFSWB minimizes the expected value of the maximal one-dimensional Wasserstein distance among marginals. By considering fairness on one-dimensional projections, us-MFSWB can yield an unbiased gradient estimate which is the reason why it is named as unbiased s-MFSWB. Fortunately, the gradient estimator of s-MFSWB is unbiased.

Proposition 3.4. *Given $K \geq 2$ marginals $\mu_{1:K} \in \mathcal{P}_p(\mathbb{R}^d)$, we have $\mathcal{SF}(\mu; \mu_{1:K}) \leq \mathcal{USF}(\mu; \mu_{1:K})$.*

Proof of Proposition 3.4 is given in Appendix A.1. From the proposition, we see that minimizing the objective of us-MFSWB also reduces the objective of s-MFSWB implicitly.

Proposition 3.5. *Given $K \geq 2$ marginals $\mu_{1:K} \in \mathcal{P}_p(\mathbb{R}^d)$, $\theta_{1:L} \overset{i.i.d}{\sim} \mathcal{U}(\mathbb{S}^{d-1})$, we have:*

$$\mathbb{E}\left|\nabla_\phi \frac{1}{L}\sum_{l=1}^{L} W_p^p(\theta_l\sharp\mu_\phi, \theta_l\sharp\mu_{k_\theta^\star}) - \nabla_\phi \mathcal{USF}(\mu_\phi; \mu_{1:K})\right| \leq \frac{1}{\sqrt{L}} Var\left[\nabla_\phi W_p^p(\theta\sharp\mu_\phi, \theta\sharp\mu_{k_\theta^\star})\right]^{\frac{1}{2}},$$

where $k_\theta^\star = \arg\max_{k \in \{1:K\}} W_p^p(\theta\sharp\mu_\phi, \theta\sharp\mu_k)$; and the expectation and variance are under the random projecting direction $\theta \sim \mathcal{U}(\mathbb{S}^{d-1})$

Proof of Proposition 3.5 is given in Appendix A.2. From the proposition, we know that the approximation error of the gradient estimator of us-MFSWB reduces at the order of $\mathcal{O}(L^{-1/2})$. Therefore, increasing L leads to a better gradient approximation. The approximation could be further improved via Quasi-Monte Carlo methods (Nguyen et al., 2024a).

Third Surrogate Definition. The us-MFSWB in Definition 3.3 utilizes the uniform distribution as the slicing distribution, which is empirically shown to be non-optimal in statistical estimation (Nguyen et al., 2021). Following the slicing distribution selection approach in (Nguyen & Ho, 2023), we propose the third surrogate with a new slicing distribution that focuses on unfair projecting directions.

Definition 3.6. Given $K \geq 2$ marginals $\mu_{1:K} \in \mathcal{P}_p(\mathbb{R}^d)$, the marginal fairness energy-based slicing distribution

$\sigma(\theta; \mu, \mu_{1:K}) \in \mathcal{P}(\mathbb{S}^{d-1})$ is defined with the density function as follow:

$$f_\sigma(\theta; \mu, \mu_{1:K}) \propto \exp\left(\max_{k\in\{1:K\}} W_p^p(\theta\sharp\mu, \theta\sharp\mu_k)\right), \quad (6)$$

Marginal fairness energy-based slicing distribution in Definition 3.6 put more mass to a projecting direction θ that has the larger maximal one-dimensional Wasserstein distance to marginals. Therefore, it will penalize more marginally unfair projecting directions. From the new proposed slicing distribution, we can define a new surrogate MFSWB problem, named energy-based surrogate MFSWB.

Definition 3.7. Given $K \geq 2$ marginals $\mu_1, \ldots, \mu_K \in \mathcal{P}_p(\mathbb{R}^d)$, the energy-based surrogate marginal fairness sliced Wasserstein barycenter (us-MFSWB) problem is defined as:

$$\min_{\mu} \mathcal{ESF}(\mu; \mu_{1:K});$$
$$\mathcal{ESF}(\mu; \mu_{1:K}) = \mathbb{E}_{\theta\sim\sigma(\theta;\mu,\mu_{1:K})}\left[\max_{k\in\{1:K\}} W_p^p(\theta\sharp\mu, \theta\sharp\mu_k)\right] \quad (7)$$

Similar to the us-MFSWB, es-MFSWB utilizes the implicit one-dimensional marginal fairness. Nevertheless, es-MFSWB utilizes the marginal fairness energy-based slicing distribution to reweight the importance of each projecting direction instead of considering them equally.

Proposition 3.8. *Given $K \geq 2$ marginals $\mu_{1:K} \in \mathcal{P}_p(\mathbb{R}^d)$, we have $\mathcal{USF}(\mu; \mu_{1:K}) \leq \mathcal{ESF}(\mu; \mu_{1:K})$.*

Proof of Proposition 3.8 is given in Appendix A.3. From the proposition, we see that minimizing the objective of es-MFSWB reduces the objective of us-MFSWB implicitly which also decreases the objective of s-MFSWB (Proposition 3.4). Moreover, the gradient estimator of es-MFSW is asymptotically unbiased.

Computational complexities of proposed surrogates. The three proposed surrogates have linear time and space complexity, $\mathcal{O}(K)$, for the number of marginals K, matching the conventional SWB and outperforming the formal MFSWB's $\mathcal{O}(K^2)$. For the number of projections L, supports n, and dimensions d, the surrogates have a time complexity of $\mathcal{O}(Ln(\log n + d))$ and a space complexity of $\mathcal{O}(L(n + d))$, similar to both the formal MFSWB and SWB.

Gradient estimators of proposed surrogates A detailed discussion on the gradient estimators of all proposed surrogates is provided in Appendix B. In summary, the s-MFSWB surrogate has a biased gradient estimator, whereas the us-MFSWB and es-MFSWB surrogates have unbiased gradient estimators.

3.3. Sliced multi-marginal Wasserstein distance with the maximal ground metric

To shed light on the proposed substrates, we connect them to a special variant of Sliced multi-marginal Wasserstein (SMW) (see Equation 1) i.e., SMW with the maximal ground metric $c(\theta^\top x_1, \ldots, \theta^\top x_K) = \max_{i\in\{1:K\}, j\in\{1:K\}} |\theta^\top x_i - \theta^\top x_j|$. We first show that SMW with the maximal ground metric is a generalized metric on the space of probability measures.

Proposition 3.9. *Sliced multi-marginal Wasserstein distance with the maximal ground metric is a generalized metric i.e., it satisfies non-negativity, marginal exchangeability, generalized triangle inequality, and identity of indiscernibles.*

Proof of Proposition 3.9 is given in Appendix A.4. It is worth noting that SMW with the maximal ground metric has never been defined before. Since our work focuses on the MFSWB problem, we will leave the careful investigation of this variant of SMW to future work.

Proposition 3.10. *Given $K \geq 2$ marginals $\mu_{1:K} \in \mathcal{P}_p(\mathbb{R}^d)$, the maximal ground metric $c(\theta^\top x_1, \ldots, \theta^\top x_K) = \max_{i\in\{1:K\}, j\in\{1:K\}} |\theta^\top x_i - \theta^\top x_j|$, we have:*

$$\min_{\mu_1} \mathcal{USF}(\mu_1; \mu_{2:K}) \leq \min_{\mu_1} SMW_p^p(\mu_1; \mu_{2:K}, c). \quad (8)$$

The proof of Proposition 3.10 is in Appendix A.5. The inequality holds for any μ_i with $i = 2, \ldots, K$. Combining this with Proposition 3.4, we get the corollary $\min_{\mu_1} \mathcal{SF}(\mu_1; \mu_{2:K}) \leq \min_{\mu_1} SMW_p^p(\mu_1; \mu_{2:K}, c)$. This shows that minimizing us-MFSWB is equivalent to minimizing a lower bound of SMW with the maximal ground metric, which implies that us-MFSWB aims to minimize the multi-marginal distance. Extending this, minimizing es-MFSWB corresponds to minimizing a lower bound of energy-based SMW with the maximal ground metric. See Proposition B.1 in Appendix B for more details.

4. Experiments

We use two metrics i.e., the F-metric (F) and the W-metric (W) which are defined as follows: $F = \frac{2}{K(K-1)} \sum_{i=1}^K \sum_{j=i+1}^K |W_p^p(\mu, \mu_i) - W_p^p(\mu, \mu_j)|$, $W = \frac{1}{K} \sum_{i=1}^K W_p^p(\mu, \mu_i)$, where μ is the barycenter, $\mu_1, \ldots, \mu_K$ are the given marginals, and W_p^p is the Wasserstein distance (Flamary et al., 2021) of the order p. Here, the F-metric represents the marginal fairness degree of the barycenter and the W-metric represents the centerness of the barycenter. For all following experiments, we use $p = 2$ for the Wasserstein distance and barycenter problems.

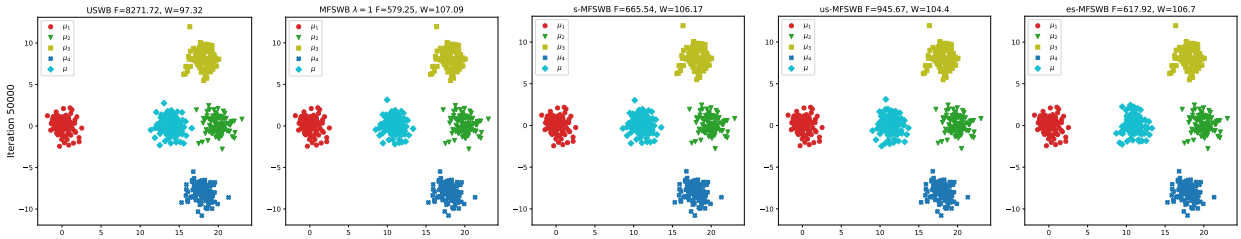


Figure 2. Barycenters from USWB, MFSWB with $\lambda = 1$, s-MFSWB, us-MFSWB, and es-MFSWB along gradient iterations with the corresponding F-metric and W-metric.

Table 1. F-metric and W-metric along iterations in point-cloud averaging application.

Methods	Iteration 1000		Iteration 5000		Iteration 10000	
	F	W	F	W	F	W
USWB	4.89	85.72	3.79	45.37	1.55	39.81
MFSWB $\lambda = 0.1$	4.76	84.86	3.78	45.2	1.32	39.73
MFSWB $\lambda = 1$	0.49	79.08	3.64	44.71	1.03	39.45
MFSWB $\lambda = 10$	4.03	**71.24**	7.32	45.21	4.13	42.56
s-MFSWB	2.52	81.84	4.01	44.9	1.15	39.58
us-MFSWB	0.3	78.69	3.74	44.38	0.87	39.26
es-MFSWB	**0.2**	78.1	**3.5**	**44.37**	**0.84**	**39.18**

Table 2. Results of training SWAE with different regularization losses.

Methods	RL	$W^2_{2,\text{latent}} \times 10^2$	$W^2_{2,\text{image}}$	$F_{\text{latent}} \times 10^2$	$W_{\text{latent}} \times 10^2$
SWAE	2.95	9.41	26.91	7.05	23.86
USWB	3.15	10.46	27.41	7.02	12.73
MFSWB $\lambda = 0.1$	3.11	8.52	**26.62**	8.66	20.01
MFSWB $\lambda = 1.0$	3.12	9.71	26.92	10.15	21.16
MFSWB $\lambda = 10.0$	**2.79**	10.32	26.95	8.27	23.39
s-MFSWB	3.22	10.50	28.69	1.30	13.71
us-MFSWB	3.05	**7.79**	27.81	2.29	**9.11**
es-MFSWB	3.35	9.60	28.27	**0.98**	9.92

4.1. Barycenter of Gaussians

We perform a Gaussian barycenter simulation with four marginals and visualize the results over last iteration in Figure 2. Settings and details are provided in the in Appendix D.

Result. USWB does not yield a fair barycenter, while the three proposed surrogates achieve better and faster convergence in both metrics. At iteration 50,000, USWB fails to produce a fair barycenter, whereas the proposed surrogates do. Among them, es-MFSWB achieves the highest marginal fairness with competitive centerness. The formal MFSWB (dual form with $\lambda = 1$) provides the fairest barycenter but is sensitive to λ.

4.2. 3D Point-cloud Averaging

We aim to find the mean shape of point-cloud shapes. We report F-metric and W-metric at iterations 1000, 5000, and 10000 in Table 1 and take average from three independent runs. We refer the reader to Appendix D for a detailed setting.

Result. As in the Gaussian simulation, proposed surrogates help to reduce the two metrics faster than the USWB. With the slicing distribution selection, es-MFSWB performs the best at every iteration, even better than the formal MFSWB with three choices of λ i.e., $0.1, 1, 10$. We also observe a similar phenomenon for two plane shapes in Figure 5 and Table 3 in Appendix D.

4.3. Color Harmonization

We conducted an experiment to transform the color palette of an image into a hybrid of two target images. Details of the experiment and results are provided in Appendix D.

Result. As in previous experiments, we see that the three proposed surrogates yield a better barycenter faster than USWB. The proposed es-MFSWB is the best variant among all surrogates since it has the lowest F-metric and W-metric at all iterations. We refer the reader to Figure 9-Figure 11 in Appendix D for additional flowers-images example, where a similar relative comparison happens.

4.4. Sliced Wasserstein Autoencoder with Class-Fair Representation

We consider training the sliced Wasserstein autoencoder (SWAE)(Kolouri et al., 2018) with a class-fairness regularization. Details of experiment is in Appendix D.

Results. From the Table 2, the proposed surrogate MFSWB generally yield better scores than USWB, except for $W^2_{2,\text{image}}$. The formal MFSWB performs well in reconstruction loss and $W^2_{2,\text{image}}$, though its F and W scores are high. The $W^2_{2,\text{latent}}$ varies slightly across runs, with minor differences in performance order, indicating relatively similar results. Overall, es-MFSWB is the best variant among the surrogates. Compared to SWAE, using a barycenter loss results in a more class-fair latent representation but sacrifices image reconstruction and generative quality.

5. Conclusion

We introduced marginal fairness sliced Wasserstein barycenter (MFSWB), a special case of sliced Wasserstein barycenter (SWB) which has approximately the same distance to

marginals. We first defined the MFSWB as a constrainted uniform SWB problem. After that, to overcome the computational drawbacks of the original problem, we propose three surrogate definitions of MFSWB which are hyperparameter-free and easy to compute. We discussed the relationship of the proposed surrogate problems and their connection to the sliced Multi-marginal Wasserstein distance with the maximal ground metric. Finally, we conduct simulations with Gaussian and experiments on 3D point-cloud averaging, color harmonization, and sliced Wasserstein autoencoder with class-fairness representation to show the benefits of the proposed surrogate MFSWB definitions.

References

Agueh, M. and Carlier, G. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924, 2011.

Anderes, E., Borgwardt, S., and Miller, J. Discrete Wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84:389–409, 2016.

Bongratz, F., Rickmann, A.-M., Pölsterl, S., and Wachinger, C. Vox2cortex: Fast explicit reconstruction of cortical surfaces from 3d mri scans with geometric deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20773–20783, 2022.

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

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Choi, K., Grover, A., Singh, T., Shu, R., and Ermon, S. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pp. 1887–1898. PMLR, 2020.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Fair regression with Wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33: 7321–7331, 2020.

Claici, S., Chien, E., and Solomon, J. Stochastic Wasserstein barycenters. In *International Conference on Machine Learning*, pp. 999–1008. PMLR, 2018.

Cohen, S., Terenin, A., Pitcan, Y., Amos, B., Deisenroth, M. P., and Kumar, K. Sliced multi-marginal optimal transport. *arXiv preprint arXiv:2102.07115*, 2021.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Cuturi, M. and Doucet, A. Fast computation of Wasserstein barycenters. In *International conference on machine learning*, pp. 685–693. PMLR, 2014.

Danskin, J. M. *The theory of max-min and its application to weapons allocation problems*, volume 5. Springer Science & Business Media, 2012.

Elvander, F., Haasler, I., Jakobsson, A., and Karlsson, J. Tracking and sensor fusion in direction of arrival estimation using optimal mass transport. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1617–1621. IEEE, 2018.

Fan, J., Taghvaei, A., and Chen, Y. Scalable computations of Wasserstein barycenter via input convex neural networks. In *International Conference on Machine Learning*, pp. 1571–1581. PMLR, 2021.

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.

Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J.-M. Obtaining fairness using optimal transport theory. In *International conference on machine learning*, pp. 2357–2365. PMLR, 2019.

Ho, N., Nguyen, X., Yurochkin, M., Bui, H. H., Huynh, V., and Phung, D. Multilevel clustering via Wasserstein means. In *International Conference on Machine Learning*, pp. 1501–1509, 2017.

Hu, F., Ratz, P., and Charpentier, A. Fairness in multi-task learning via W asserstein barycenters. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 295–312. Springer, 2023.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. Wasserstein fair classification. In *Uncertainty in artificial intelligence*, pp. 862–872. PMLR, 2020.

Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Korotin, A., Egiazarian, V., Li, L., and Burnaev, E. Wasserstein iterative networks for barycenter estimation. *Advances in Neural Information Processing Systems*, 35: 15672–15686, 2022.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. From word embeddings to document distances. In *International conference on machine learning*, pp. 957–966. PMLR, 2015.

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.

Manole, T., Balakrishnan, S., and Wasserman, L. Minimax confidence intervals for the sliced Wasserstein distance. *Electronic Journal of Statistics*, 16(1):2252–2345, 2022.

Montesuma, E. F. and Mboula, F. M. N. Wasserstein barycenter for multi-source domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16785–16793, 2021.

Mroueh, Y. Wasserstein style transfer. In *International Conference on Artificial Intelligence and Statistics*, pp. 842–852. PMLR, 2020.

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.

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

Nguyen, K. and Ho, N. Hierarchical hybrid sliced Wasserstein: A scalable metric for heterogeneous joint distributions. *arXiv preprint arXiv:2404.15378*, 2024.

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., Bariletto, N., and Ho, N. Quasi-monte carlo for 3d sliced Wasserstein. In *The Twelfth International Conference on Learning Representations*, 2024a.

Nguyen, K., Zhang, S., Le, T., and Ho, N. Sliced Wasserstein with random-path projecting directions. *International Conference on Machine Learning*, 2024b.

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.

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

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pp. 435–446. Springer, 2012.

Samadi, S., Tantipongpipat, U., Morgenstern, J. H., Singh, M., and Vempala, S. The price of fair pca: One extra dimension. *Advances in neural information processing systems*, 31, 2018.

Séjourné, T., Vialard, F.-X., and Peyré, G. Faster unbalanced optimal transport: Translation invariant sinkhorn and 1-d frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pp. 4995–5021. PMLR, 2022.

Silvia, C., Ray, J., Tom, S., Aldo, P., Heinrich, J., and John, A. A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3633–3640, 2020.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

Srivastava, S., Li, C., and Dunson, D. B. Scalable bayes via barycenter in Wasserstein space. *Journal of Machine Learning Research*, 19(8):1–35, 2018.

Staib, M., Claici, S., Solomon, J. M., and Jegelka, S. Parallel streaming Wasserstein barycenters. *Advances in Neural Information Processing Systems*, 30, 2017.

Sun, S., Le, T.-T., You, C., Tang, H., Han, K., Ma, H., Kong, D., Yan, X., and Xie, X. Hybrid-csr: Coupling explicit and implicit shape representation for cortical surface reconstruction. *arXiv preprint arXiv:2307.12299*, 2023.

Zhuang, Y., Chen, X., and Yang, Y. Wasserstein k-means for clustering probability distributions. *Advances in Neural Information Processing Systems*, 35:11382–11395, 2022.

We present skipped proofs in Appendix A. We then provide some additional materials which are mentioned in the main paper in Appendix B. After that, related works are discussed in Appendix C. We then provide additional experimental results in Appendix D. Finally, we report the used computational devices in Appendix E.

A. Proofs

A.1. Proof of Proposition 3.4

Proof. From Definition 3.2, we have

$$\begin{aligned}\mathcal{SF}(\mu, \mu_{1:K}) &= \max_{k\in\{1,\dots,K\}} SW_p^p(\mu, \mu_k) \\ &= \max_{k\in\{1,\dots,K\}} \mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^{d-1})}[W_p^p(\theta\sharp\mu, \theta\sharp\mu_k)]\end{aligned}$$

Let $k^\star = \arg\max_{k\in\{1,\dots,K\}} \mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^{d-1})}[W_p^p(\theta\sharp\mu, \theta\sharp\mu_k)]$, we have

$$\begin{aligned}\mathcal{SF}(\mu, \mu_{1:K}) &= \mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^{d-1})}[W_p^p(\theta\sharp\mu, \theta\sharp\mu_{k^\star})] \\ &\leq \mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[\max_{k\in\{1,\dots,K\}} W_p^p(\theta\sharp\mu, \theta\sharp\mu_k)\right] \\ &= \mathcal{USF}(\mu, \mu_{1:K}),\end{aligned}$$

as from Definition 3.3, which completes the proof.

A.2. Proof of Proposition 3.5

Using the Holder's inequality, we have:

$$\begin{aligned}&\mathbb{E}\left|\nabla_\phi \frac{1}{L}\sum_{l=1}^{L} \mathrm{W}_p^p(\theta_l\sharp\mu_\phi, \theta_l\sharp\mu_{k^\star_{\theta_l}}) - \nabla_\phi \mathcal{USF}(\mu_\phi; \mu_{1:K})\right| \\ &\leq \left(\mathbb{E}\left|\nabla_\phi \frac{1}{L}\sum_{l=1}^{L} \mathrm{W}_p^p(\theta_l\sharp\mu_\phi, \theta_l\sharp\mu_{k^\star_{\theta_l}}) - \nabla_\phi \mathcal{USF}(\mu_\phi; \mu_{1:K})\right|^2\right)^{\frac{1}{2}} \\ &= \left(\mathbb{E}\left(\nabla_\phi \frac{1}{L}\sum_{l=1}^{L} \mathrm{W}_p^p(\theta_l\sharp\mu_\phi, \theta_l\sharp\mu_{k^\star_{\theta_l}}) - \nabla_\phi \mathbb{E}\left[\mathrm{W}_p^p(\theta\sharp\mu_\phi, \theta\sharp\mu_{k^\star_\theta})\right]\right)^2\right)^{\frac{1}{2}} \\ &= \left(\mathbb{E}\left(\frac{1}{L}\sum_{l=1}^{L} \nabla_\phi \mathrm{W}_p^p(\theta_l\sharp\mu_\phi, \theta_l\sharp\mu_{k^\star_{\theta_l}}) - \mathbb{E}\left[\nabla_\phi \mathrm{W}_p^p(\theta\sharp\mu_\phi, \theta\sharp\mu_{k^\star_\theta})\right]\right)^2\right)^{\frac{1}{2}} \\ &= \left(\mathrm{Var}\left[\frac{1}{L}\sum_{l=1}^{L} \nabla_\phi \mathrm{W}_p^p(\theta_l\sharp\mu_\phi, \theta_l\sharp\mu_{k^\star_{\theta_l}})\right]\right)^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{L}}\mathrm{Var}\left[\nabla_\phi \mathrm{W}_p^p(\theta\sharp\mu_\phi, \theta\sharp\mu_{k^\star_\theta})\right]^{\frac{1}{2}},\end{aligned}$$

which completes the proof.

A.3. Proof of Proposition 3.8

We first restate the following Lemma from (Nguyen et al., 2024b) and provide the proof for completeness.

Lemma A.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:*

$$\frac{1}{L}(\sum_{i=1}^{L} a_i)(\sum_{i=1}^{L} b_i) \leq \sum_{i=1}^{L} a_i b_i. \tag{9}$$

Proof. For $L = 1$, we directly have $a_i b_i = a_i b_i$. Assuming that for L the inequality holds i.e., $\frac{1}{L}(\sum_{i=1}^{L} a_i)(\sum_{i=1}^{L} b_i) \leq \sum_{i=1}^{L} a_i b_i$ which is equivalent to $(\sum_{i=1}^{L} a_i)(\sum_{i=1}^{L} b_i) \leq L \sum_{i=1}^{L} a_i b_i$. Now, we show that $\frac{1}{L}(\sum_{i=1}^{L} a_i)(\sum_{i=1}^{L} b_i) \leq \sum_{i=1}^{L} a_i b_i$ i.e., the inequality holds for $L+1$. We have

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

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

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

Then, 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. □

Now, we go back to the main inequality which is $\mathcal{USF}(\mu; \mu_{1:K}) \leq \mathcal{ESF}(\mu; \mu_{1:K})$. From Definition 3.7, we have:

$$\begin{aligned}\mathcal{ESF}(\mu; \mu_{1:K}) &= \mathbb{E}_{\theta \sim \sigma(\theta; \mu, \mu_{1:K})} \left[\max_{k \in \{1,\dots,K\}} W_p^p(\theta \sharp \mu, \theta \sharp \mu_k) \right] \\ &= \mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})} \left[\max_{k \in \{1,\dots,K\}} W_p^p(\theta \sharp \mu, \theta \sharp \mu_k) \frac{f_\sigma(\theta; \mu, \mu_{1:K})}{\frac{\Gamma(d/2)}{2\pi^{d/2}}} \right],\end{aligned}$$

where $f_\sigma(\theta; \mu, \mu_{1:K}) \propto \exp\left(\max_{k \in \{1,\dots,K\}} W_p^p(\theta \sharp \mu, \theta \sharp \mu_k)\right)$. Now, we consider a Monte Carlo estimation of $\mathcal{ESF}(\mu; \mu_{1:K})$ by importance sampling:

$$\widehat{\mathcal{ESF}}(\mu; \mu_{1:K}, L) = \frac{1}{L} \sum_{l=1}^{L} \left[\max_{k \in \{1,\dots,K\}} W_p^p(\theta_l \sharp \mu, \theta_l \sharp \mu_k) \frac{\exp\left(\max_{k \in \{1,\dots,K\}} W_p^p(\theta_l \sharp \mu, \theta_l \sharp \mu_k)\right)}{\sum_{i=1}^{L} \exp\left(\max_{k \in \{1,\dots,K\}} W_p^p(\theta_i \sharp \mu, \theta_i \sharp \mu_k)\right)} \right],$$

where $\theta_1, \dots, \theta_L \overset{i.i.d}{\sim} \mathcal{U}(\mathbb{S}^{d-1})$. Similarly, we consider a Monte Carlo estimation of $\mathcal{USF}(\mu; \mu_{1:K})$:

$$\widehat{\mathcal{USF}}(\mu; \mu_{1:K}, L) = \frac{1}{L} \sum_{l=1}^{L} \left[\max_{k \in \{1,\dots,K\}} W_p^p(\theta_l \sharp \mu, \theta_l \sharp \mu_k) \right],$$

for the same set of $\theta_1, \dots, \theta_L$. Without losing generality, we assume that $\max_{k \in \{1,\dots,K\}} W_p^p(\theta_1 \sharp \mu, \theta_1 \sharp \mu_k) \leq \dots \leq \max_{k \in \{1,\dots,K\}} W_p^p(\theta_L \sharp \mu, \theta_L \sharp \mu_k)$. Let $\max_{k \in \{1,\dots,K\}} W_p^p(\theta_i \sharp \mu, \theta_i \sharp \mu_k) = a_i$ and $\exp\left(\max_{k \in \{1,\dots,K\}} W_p^p(\theta_i \sharp \mu, \theta_i \sharp \mu_k)\right) = b_i$, applying Lemma A.1, we have:

$$\widehat{\mathcal{USF}}(\mu; \mu_{1:K}, L) \leq \widehat{\mathcal{ESF}}(\mu; \mu_{1:K}, L) \quad \forall L \geq 1.$$

By letting $L \to \infty$ and applying the law of large numbers, we obtain:

$$\mathcal{USF}(\mu; \mu_{1:K}) \leq \mathcal{ESF}(\mu; \mu_{1:K}),$$

which completes the proof.

A.4. Proof of Proposition 3.9

We first recall the definition of the SMW with the maximal ground metric:

$$SMW_p^p(\mu_1,\dots,\mu_K;c) = \mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\dots,\mu_K)}\int \max_{i\in\{1,\dots,K\},j\in\{1,\dots,K\}} |\theta^\top x_i - \theta^\top x_j|^p d\pi(x_1,\dots,x_K)\right].$$

Non-negativity. Since $\max_{i\in\{1,\dots,K\},j\in\{1,\dots,K\}} |\theta^\top x_i - \theta^\top x_j|^p \geq 0$ for any $x_1,\dots,x_K$ and for any θ, we can obtain the desired property $SMW_p^p(\mu_1,\dots,\mu_K;c) \geq 0$ which implies $SMW_p(\mu_1,\dots,\mu_K;c) \geq 0$.

Marginal Exchangeability. For any permutation $\sigma : [[K]] \to [[K]]$, we have:

$$\begin{aligned}
SMW_p^p(\mu_1,\dots,\mu_K;c) &= \mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\dots,\mu_K)}\int \max_{i\in\{1,\dots,K\},j\in\{1,\dots,K\}} |\theta^\top x_i - \theta^\top x_j|^p d\pi(x_1,\dots,x_K)\right] \\
&= \mathbb{E}\left[\inf_{\pi\in\Pi(\mu_{\sigma(1)},\dots,\mu_{\sigma(K)})}\int \max_{i\in\{1,\dots,K\},j\in\{1,\dots,K\}} |\theta^\top x_i - \theta^\top x_j|^p d\pi(x_1,\dots,x_K)\right] \\
&= SMW_p^p(\mu_{\sigma(1)},\dots,\mu_{\sigma(K)};c).
\end{aligned}$$

Generalized Triangle Inequality. For $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, we have :

$$\begin{aligned}
&SMW_p^p(\mu_1,\dots,\mu_K;c) \\
&= \mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\dots,\mu_K)}\int \max_{i\in\{1,\dots,K\},j\in\{1,\dots,K\}} |\theta^\top x_i - \theta^\top x_j|^p d\pi(x_1,\dots,x_K)\right] \\
&\leq \mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\dots,\mu_K)}\int \sum_{k=1}^K \max_{i\in\{1,\dots,K\}\setminus\{k\},j\in\{1,\dots,K\}\setminus\{k\}} |\theta^\top x_i - \theta^\top x_j|^p d\pi(x_1,\dots,x_K)\right] \\
&= \mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\dots,\mu_K)}\sum_{k=1}^K \int \max_{i\in\{1,\dots,K\}\setminus\{k\},j\in\{1,\dots,K\}\setminus\{k\}} |\theta^\top x_i - \theta^\top x_j|^p d\pi(x_1,\dots,x_K)\right] \\
&= \mathbb{E}\left[\sum_{k=1}^K \int \max_{i\in\{1,\dots,K\}\setminus\{k\},j\in\{1,\dots,K\}\setminus\{k\}} |\theta^\top x_i - \theta^\top x_j|^p d\pi^\star(x_1,\dots,x_{k-1},x_{k+1},\dots x_K)\right]
\end{aligned}$$

for $\pi^\star$ is the optimal multi-marginal transportation plan and $\pi^\star(x_1,\dots,x_{k-1},x_{k+1},x_K)$ is the marginal joint distribution by integrating out x_k. By the gluing lemma (Peyré & Cuturi, 2020), there exists optimal plans $\pi^\star(x_1,\dots,x_{k-1},y,x_{k+1},x_K)$ for any $k \in [[K]]$ and y follows μ. We further have:

$$\begin{aligned}
&SMW_p^p(\mu_1,\dots,\mu_K;c) \\
&\leq \mathbb{E}\left[\sum_{k=1}^K \int \max\Big(\max_{i\in\{1,\dots,K\}\setminus\{k\},j\in\{1,\dots,K\}\setminus\{k\}} |\theta^\top x_i - \theta^\top x_j|^p, \right. \\
&\qquad \left. \max_{i\in\{1,\dots,K\}\setminus\{k\}} |\theta^\top x_i - \theta^\top y|^p \Big) d\pi^\star(x_1,\dots,x_{k-1},y,x_{k+1},\dots x_K)\right] \\
&= \sum_{k=1}^K \mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\dots,\mu_{k-1},\mu,\mu_{k+1},\dots,\mu_K)}\int \max_{i\in\{1,\dots,K\},j\in\{1,\dots,K\}} |\theta^\top x_i - \theta^\top x_j|^p d\pi(x_1,\dots,x_K)\right] \\
&= \sum_{k=1}^K SMW_p^p(\mu_1,\dots,\mu_{k-1},\mu,\mu_{k+1},\dots,\mu_K;c).
\end{aligned}$$

Applying the Minkowski's inequality, we obtain the desired property:

$$SMW_p(\mu_1,\dots,\mu_K;c) \leq \sum_{k=1}^K SMW_p(\mu_1,\dots,\mu_{k-1},\mu,\mu_{k+1},\dots,\mu_K;c).$$

Identity of Indiscernibles. From the proof in Appendix A.5, we have:

$$\begin{aligned} SMW_p^p(\mu_1,\ldots,\mu_K;c) &\geq \mathbb{E}\left[\max_{i\in\{1,\ldots,K\},j\in\{1,\ldots,K\}} W_p^p(\theta\sharp\mu_i,\theta\sharp\mu_j)\right] \\ &\geq \max_{i\in\{1,\ldots,K\},j\in\{1,\ldots,K\}} \mathbb{E}\left[W_p^p(\theta\sharp\mu_i,\theta\sharp\mu_j)\right] \\ &= \max_{i\in\{1,\ldots,K\},j\in\{1,\ldots,K\}} SW_p^p(\mu_i,\mu_j). \end{aligned}$$

Therefore, when $SMW_p(\mu_1,\ldots,\mu_K;c)=0$, we have $SW_p^p(\mu_i,\mu_j)=0$ which implies $\mu_i=\mu_j$ for any $i,j\in[[K]]$. As a result, $\mu_1=\ldots=\mu_K$ from the metricity of the SW distance. For the other direction, it is easy to see that if $\mu_1=\ldots\mu_K$, we have $SMW_p(\mu_1,\ldots,\mu_K;c)=0$ based on the definition and the metricity of the Wasserstein distance.

A.5. Proof of Proposition 3.10

Given the maximal ground metric $c(\theta^\top x_1,\ldots,\theta^\top x_K)=\max_{i\in\{1,\ldots,K\},j\in\{1,\ldots,K\}}|\theta^\top x_i-\theta^\top x_j|$, from Equation 1

$$\begin{aligned} SMW_p^p(\mu_1,\ldots,\mu_K;c) &= \mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\ldots,\mu_K)}\int c(\theta^\top x_1,\ldots,\theta^\top x_K)^p d\pi(x_1,\ldots,x_K)\right] \\ &= \mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\ldots,\mu_K)}\int \max_{i\in\{1,\ldots,K\},j\in\{1,\ldots,K\}}|\theta^\top x_i-\theta^\top x_j|^p d\pi(x_1,\ldots,x_K)\right] \end{aligned}$$

By Jensen inequality i.e., $(x_1,\ldots,x_K)\to\max_{i\in\{1,\ldots,K\},j\in\{1,\ldots,K\}}|\theta^\top x_i-\theta^\top x_j|^p$ is a convex function, we have:

$$SMW_p^p(\mu_1,\ldots,\mu_K;c)\geq\mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\ldots,\mu_K)}\max_{i\in\{1,\ldots,K\},j\in\{1,\ldots,K\}}\int|\theta^\top x_i-\theta^\top x_j|^p d\pi(x_1,\ldots,x_K)\right].$$

Using max-min inequality, we have:

$$\begin{aligned} SMW_p^p(\mu_1,\ldots,\mu_K;c) &\geq \mathbb{E}\left[\max_{i\in\{1,\ldots,K\},j\in\{1,\ldots,K\}}\inf_{\pi\in\Pi(\mu_1,\ldots,\mu_K)}\int|\theta^\top x_i-\theta^\top x_j|^p d\pi(x_1,\ldots,x_K)\right] \\ &\geq \mathbb{E}\left[\max_{i\in\{1,\ldots,K\},j\in\{1,\ldots,K\}}\inf_{\pi\in\Pi(\mu_i,\mu_j)}\int|\theta^\top x_i-\theta^\top x_j|^p d\pi(x_i,x_j)\right] \\ &= \mathbb{E}\left[\max_{i\in\{1,\ldots,K\},j\in\{1,\ldots,K\}} W_p^p(\theta\sharp\mu_i,\theta\sharp\mu_j)\right]. \end{aligned}$$

Therefore, minimizing two sides with respect to μ_1, we have:

$$\begin{aligned} \min_{\mu_1} SMW_p^p(\mu_1,\ldots,\mu_K;c) &\geq \min_{\mu_1}\mathbb{E}\left[\max_{i\in\{1,\ldots,K\},j\in\{1,\ldots,K\}} W_p^p(\theta\sharp\mu_i,\theta\sharp\mu_j)\right] \\ &\geq \min_{\mu_1}\mathbb{E}\left[\max_{i\in\{2,\ldots,K\}} W_p^p(\theta\sharp\mu_1,\theta\sharp\mu_i)\right] \\ &= \min_{\mu_1}\mathcal{USF}(\mu_1;\mu_{2:K}), \end{aligned}$$

which completes the proof.

B. Additional Materials

Gradient of Discrete Sliced Wasserstein Barycenter We discuss the discrete SWB i.e., marginals and the barycenter are discrete measures.

Free supports barycenter. In this setting, we have $\mu_\phi=\frac{1}{n}\sum_{i=1}^n\delta_{x_i}$, $\mu_k=\frac{1}{n}\sum_{i=1}^n\delta_{y_i}$, and $\phi=(x_{1:n})$, we can compute the (sub-)gradient with the time complexity $\mathcal{O}(n\log n)$:

$$\begin{aligned} \nabla_{x_i}\mathrm{W}_p^p(\theta\sharp\mu_\phi,\theta\sharp\mu_k) = p|\theta^\top x_i-\theta^\top y_{\sigma(i)}|^{p-1} \\ \mathrm{sign}(\theta^\top x_i-\theta^\top y_{\sigma(i)})\theta, \end{aligned} \tag{10}$$

where $\sigma = \sigma_1 \circ \sigma_2^{-1}$ with σ_1 and σ_2 are any sorted permutation of $\{x_{1:n}\}$ and $\{y_{1:n}\}$.

Fixed supports barycenter. In this setting, we have $\mu_\phi = \sum_{i=1}^n \phi_i \delta_{x_i}$, $\mu_k = \sum_{i=1}^n \beta_i \delta_{x_i}$, $\sum_{i=1}^n \phi_i = \sum_{i=1}^n \beta_i$ and $\phi = (\phi_{1:n})$. We can compute the gradient as follows:

$$\nabla_\phi \mathrm{W}_p^p(\theta\sharp\mu_\phi, \theta\sharp\mu_k) = \boldsymbol{f}^\star, \tag{11}$$

where $\boldsymbol{f}^\star$ is the first optimal Kantorovich dual potential of $\mathrm{W}_p^p(\theta\sharp\mu_\phi, \theta\sharp\mu_k)$ which can be obtained with the time complexity of $\mathcal{O}(n \log n)$. We refer the reader to Proposition 1 in (Cuturi & Doucet, 2014) for the detail and Algorithm 1 in (Séjourné et al., 2022) for the computational algorithm.

When the supports or weights of the barycenter are the output of a parametric function, we can use the chain rule to estimate the gradient of the parameters of the function. For the continuous case, we can approximate the barycenter and marginals by their empirical versions, and then perform the estimation in the discrete case. Since the sample complexity of SW is $\mathcal{O}(n^{-1/2})$ (Nadjahi et al., 2019; Nguyen et al., 2021; Manole et al., 2022; Nietert et al., 2022), the approximation error will reduce fast with the number of support n increases. Another option is to use continuous Wasserstein solvers (Fan et al., 2021; Korotin et al., 2022; Claici et al., 2018), however, this option is not as simple as the first one.

Gradient estimator of s-MFSWB. Let μ_ϕ be paramterized by $\phi \in \Phi$, and $\mathcal{F}(\phi, k) = SW_p^p(\mu_\phi, \mu_k)$, we would like to compute $\nabla_\phi \max_{k\in\{1,\ldots,K\}} \mathcal{F}(\phi, k)$. By Danskin's envelope theorem (Danskin, 2012), we have $\nabla_\phi \max_{k\in\{1,\ldots,K\}} \mathcal{F}(\phi, k) = \nabla_\phi \mathcal{F}(\phi, k^\star) = \nabla_\phi \mathrm{SW}_p^p(\mu_\phi, \mu_{k^\star})$ for $k^\star = \arg\max_{k\in\{1,\ldots,K\}} \mathcal{F}(\phi, k)$. Nevertheless, $k^\star$ is intractable due to the intractablity of $SW_p^p(\mu_\phi, \mu_k)$ for $k = 1, \ldots, K$. Hence, we can form the estimation $\hat{k}^\star = \arg\max_{k\in\{1,\ldots,K\}} \widehat{SW}_p^p(\mu_\phi, \mu_k; L)$ where $\widehat{SW}_p^p(\mu_\phi, \mu_k; L) = \frac{1}{L}\sum_{l=1}^L \mathrm{W}_p^p(\theta_l\sharp\mu_\phi, \theta_l\sharp\mu_k)$ with $\theta_1, \ldots, \theta_L \overset{i.i.d}{\sim} \mathcal{U}(\mathbb{S}^{d-1})$. The remaining work is to estimate $\nabla_\phi \mathrm{SW}_p^p(\mu_\phi, \mu_{\hat{k}^\star})$, which is easy. We refer the reader to Algorithm 2 for the gradient estimation and optimization procedure. The downside of this estimator is that it is biased.

Gradient estimator of us-MFSWB. Let μ_ϕ be paramterized by $\phi \in \Phi$, and $\mathcal{F}(\theta, \phi, k) = W_p^p(\theta\sharp\mu_\phi, \theta\sharp\mu_k)$, we would like to compute $\nabla_\phi \mathbb{E}_{\theta\sim\mathbb{S}^{d-1}}[\max_{k\in\{1,\ldots,K\}} \mathcal{F}(\theta, \phi, k)]$ which is equivalent to $\mathbb{E}_{\theta\sim\mathbb{S}^{d-1}}[\nabla_\phi \max_{k\in\{1,\ldots,K\}} \mathcal{F}(\theta, \phi, k)]$ due to the Leibniz's rule. By Danskin's envelope theorem, we have $\nabla_\phi \max_{k\in\{1,\ldots,K\}} \mathcal{F}(\theta, \phi, k) = \nabla_\phi \mathcal{F}(\theta, \phi, k^\star) = \nabla_\phi \mathrm{W}_p^p(\theta\sharp\mu_\phi, \theta\sharp\mu_{k^\star})$ for $k_\theta^\star = \arg\max_{k\in\{1,\ldots,K\}} \mathcal{F}(\theta, \phi, k)$ where we can estimate $\nabla_\phi \mathrm{W}_p^p(\theta\sharp\mu_\phi, \theta\sharp\mu_{k_\theta^\star})$ can be computed as in Equation 10- 11. Overall, with $\theta_1, \ldots, \theta_L \overset{i.i.d}{\sim} \mathcal{U}(\mathbb{S}^{d-1})$, we can form the final estimation $\frac{1}{L}\sum_{l=1}^L \nabla_\phi \mathrm{W}_p^p(\theta_l\sharp\mu_\phi, \theta_l\sharp\mu_{k_{\theta_l}^\star})$ which is an unbiased estimate. We refer the reader to Algorithm 3 for the gradient estimation and optimization procedure.

Gradient estimator of es-MFSWB. Let μ_ϕ be parameterized by $\phi \in \Phi$, we want to estimate $\nabla_\phi \mathcal{ESF}(\mu_\phi; \mu_{1:K})$. Since the slicing distribution is unnormalized, we use importance sampling to form an estimation. With $\theta_1, \ldots, \theta_L \overset{i.i.d}{\sim} \mathcal{U}(\mathbb{S}^{d-1})$, we can form the importance sampling stochastic gradient estimation:

$$\hat{\nabla}_\phi \mathcal{ESF}(\mu_\phi; \mu_{1:K}, L) = \frac{1}{L}\sum_{l=1}^L \left[\nabla_\phi \left(W_p^p(\theta_l\sharp\mu, \theta_l\sharp\mu_{k_{\theta_l}^\star}) \frac{\exp\left(W_p^p(\theta_l\sharp\mu, \theta_l\sharp\mu_{k_{\theta_l}^\star})\right)}{\frac{1}{L}\sum_{i=1}^L \left[\exp\left(W_p^p(\theta_i\sharp\mu, \theta_i\sharp\mu_{k_{\theta_i}^\star})\right)\right]}\right)\right],$$

which can be further derived by using the chain rule and previously discussed techniques. It is worth noting that the above estimation is only asymptotically unbiased. We refer the reader to Algorithm 4 for the gradient estimation and optimization procedure.

Algorithms. As mentioned in the main paper, we present the computational algorithm for SWB in Algorithm 1, for s-MFSWB in Algorithm 2, for us-MFSWB in Algorithm 3, and for es-MFSWB in Algorithm 4.

Algorithm 1 Computational algorithm of the SWB problem

Input: Marginals $\mu_1, \ldots, \mu_K$, $p \geq 1$, weights $\omega_1, \ldots, \omega_K$, the number of projections L, step size η, the number of iterations T.
Initialize the barycenter μ_ϕ
for $t = 1$ to T **do**
  Set $\nabla_\phi = 0$
  Sample $\theta_l \sim \mathcal{U}(\mathbb{S}^{d-1})$
  for $l = 1$ to L **do**
    for $k = 1$ to K **do**
      Set $\nabla_\phi = \nabla_\phi + \nabla_\phi \frac{\omega_k}{L} \mathrm{W}_p^p(\theta_l \sharp \mu_\phi, \theta_l \sharp \mu_k)$
    end for
  end for
  $\phi = \phi - \eta \nabla_\phi$
end for
Return: μ_ϕ

Algorithm 2 Computational algorithm of the s-MFSWB problem

Input: Marginals $\mu_1, \ldots, \mu_K$, $p \geq 1$ the number of projections L, step size η, the number of iterations T.
Initialize the barycenter μ_ϕ
for $t = 1$ to T **do**
  Set $\nabla_\phi = 0$
  Sample $\theta_l \sim \mathcal{U}(\mathbb{S}^{d-1})$
  $k^\star = 1$
  for $k = 1$ to K **do**
    for $l = 1$ to L **do**
      if $\frac{1}{L} \sum_{l=1}^{L} \mathrm{W}_p^p(\theta_l \sharp \mu_\phi, \theta_l \sharp \mu_k) > \frac{1}{L} \sum_{l=1}^{L} \mathrm{W}_p^p(\theta_l \sharp \mu_\phi, \theta_l \sharp \mu_{k^\star})$ **then**
        $k^\star = k$
      end if
    end for
  end for
  $\nabla_\phi = \nabla_\phi + \frac{1}{L} \sum_{l=1}^{L} \nabla_\phi \mathrm{W}_p^p(\theta_l \sharp \mu_\phi, \theta_l \sharp \mu_{k^\star})$
  $\phi = \phi - \eta \nabla_\phi$
end for
Return: μ_ϕ

Algorithm 3 Computational algorithm of the us-MFSWB problem

Input: Marginals $\mu_1, \ldots, \mu_K$, $p \geq 1$ the number of projections L, step size η, the number of iterations T.
Initialize the barycenter μ_ϕ
for $t = 1$ to T **do**
  Set $\nabla_\phi = 0$
  Sample $\theta_l \sim \mathcal{U}(\mathbb{S}^{d-1})$
  for $l = 1$ to L **do**
    $k_l^\star = 1$
    for $k = 2$ to K **do**
      if $\mathrm{W}_p^p(\theta_l \sharp \mu_\phi, \theta_l \sharp \mu_k) > \mathrm{W}_p^p(\theta_l \sharp \mu_\phi, \theta_l \sharp \mu_{k_l^\star})$ **then**
        $k_l^\star = k$
      end if
    end for
    $\nabla_\phi = \nabla_\phi + \nabla_\phi \frac{1}{L} \mathrm{W}_p^p(\theta_l \sharp \mu_\phi, \theta_l \sharp \mu_{k_l^\star})$
  end for
  $\phi = \phi - \eta \nabla_\phi$
end for
Return: μ_ϕ

Algorithm 4 Computational algorithm of the es-MFSWB problem

Input: Marginals $\mu_1, \dots, \mu_K$, $p \geq 1$ the number of projections L, step size η, the number of iterations T.
Initialize the barycenter μ_ϕ
for $t = 1$ to T **do**
  Set $\nabla_\phi = 0$
  Sample $\theta_l \sim \mathcal{U}(\mathbb{S}^{d-1})$
  for $l = 1$ to L **do**
    $k_l^\star = 1$
    for $k = 2$ to K **do**
      if $\mathrm{W}_p^p(\theta_l \sharp \mu_\phi, \theta_l \sharp \mu_k) > \mathrm{W}_p^p(\theta_l \sharp \mu_\phi, \theta_l \sharp \mu_{k_l^\star})$ **then**
        $k_l^\star = k$
      end if
    end for
  end for
  for $l = 1$ to L **do**
    $w_{l,\phi} = \frac{\exp(\mathrm{W}_p^p(\theta_l \sharp \mu_\phi, \theta_l \sharp \mu_{k_l^\star}))}{\sum_{j=1}^L \exp(\mathrm{W}_p^p(\theta_j \sharp \mu_\phi, \theta_j \sharp \mu_{k_j^\star}))}$
  end for
  $\nabla_\phi = \nabla_\phi + \nabla_\phi \frac{w_{l,\phi}}{L} \mathrm{W}_p^p(\theta_l \sharp \mu_\phi, \theta_l \sharp \mu_{k_l^\star})$
  $\phi = \phi - \eta \nabla_\phi$
end for
Return: μ_ϕ

Energy-based sliced Multi-marginal Wasserstein. As shown in Proposition 3.10, us-MFSWB is equivalent to minimizing a lower bound of SMW with the maximal ground metric. We now show that es-MFSWB is also equivalent to minimizing a lower bound of a variant of SMW i.e., Energy-based sliced Multi-marginal Wasserstein with the maximal ground metric. We refer the reader to Proposition B.1 for a detailed definition. The proof of Proposition B.1 is similar to the proof of Proposition 3.10 in Appendix A.5.

Proposition B.1. *Given $K \geq 2$ marginals $\mu_1, \dots, \mu_K \in \mathcal{P}_p(\mathbb{R}^d)$, the maximal ground metric $c(\theta^\top x_1, \dots, \theta^\top x_K) = \max_{i \in \{1,\dots,K\}, j \in \{1,\dots,K\}} |\theta^\top x_i - \theta^\top x_j|$, we have:*

$$\min_{\mu_1} \mathcal{ESF}(\mu_1; \mu_{2:K}) \leq \min_{\mu_1} ESMW_p^p(\mu_1, \mu_2, \dots, \mu_K; c), \tag{12}$$

where

$$ESMW_p^p(\mu_1, \mu_2, \dots, \mu_K; c) = \mathbb{E}\left[\inf_{\pi \in \Pi(\mu_1, \dots, \mu_K)} \int c(\theta^\top x_1, \dots, \theta^\top x_K)^p d\pi(x_1, \dots, x_K)\right],$$

and the expectation is with respect to $\sigma(\theta)$ i.e.,

$$f_\sigma(\theta; \mu_1, \mu_{2:K}) \propto \exp\left(\max_{k \in \{2,\dots,K\}} W_p^p(\theta \sharp \mu_1, \theta \sharp \mu_k)\right).$$

C. Related Works

Fair Learning with Wasserstein Barycenter. A connection between fair regression and one-dimensional Wasserstein barycenter is established by deriving the expression for the optimal function minimizing squared risk under Demographic Parity constraints (Chzhen et al., 2020). Similarly, Demographic Parity fair classification is connected to one-dimensional Wasserstein-1 distance barycenter in (Jiang et al., 2020). The work (Hu et al., 2023) extends the Demographic Parity constraint to multi-task problems for regression and classification and connects them to the one-dimensional Wasserstein-2 distance barycenters. A method to augment the input so that predictability of the protected attribute is impossible, by using Wasserstein-2 distance Barycenters to repair the data is proposed in (Gordaliza et al., 2019). A general approach for using one-dimensional Wasserstein-1 distance barycenter to obtain Demographic Parity in classification and regression is proposed in (Silvia et al., 2020). Overall, all discussed works define fairness in terms of Demographic Parity constraints in

applications with a response variable (classification and regression) in one dimension. In contrast, we focus on marginal fairness barycenter i.e., using a set of measures only, in any dimensions.

Other possible applications. Wasserstein barycenter has been used to cluster measures in (Zhuang et al., 2022). In particular, a K-mean algorithm for measures is proposed with Wasserstein barycenter as the averaging operator. Therefore, our MFSWB can be directly used to enforce the fairness for averaging inside each cluster. The proposed MFSWB can be also used to average meshes by changing the SW to H2SW which is proposed in (Nguyen & Ho, 2024).

D. Additional Experiments

Gaussians barycenter with the formal MFSWB. We first start with a simple simulation with 4 marginals which are empirical distributions with 100 i.i.d samples from 4 Gaussian distributions i.e., $\mathcal{N}((0,0), I)$, $\mathcal{N}((20,0), I)$, $\mathcal{N}((18,8), I)$, and $\mathcal{N}((18,-8), I)$. We then find the barycenter which is represented as an empirical distribution with 100 supports initialized by sampling i.i.d from $\mathcal{N}((0,-5), I)$. We use stochastic gradient descent with 50000 iterations of learning rate 0.01, the number of projections 100. We show the visualization of the found barycenters with the corresponding F-metric and W-metric by using USWB, s-MFSWB, us-MFSWB, and es-MFSWB at iterations 0, 1000, 5000, and 50000 in Figure 3.

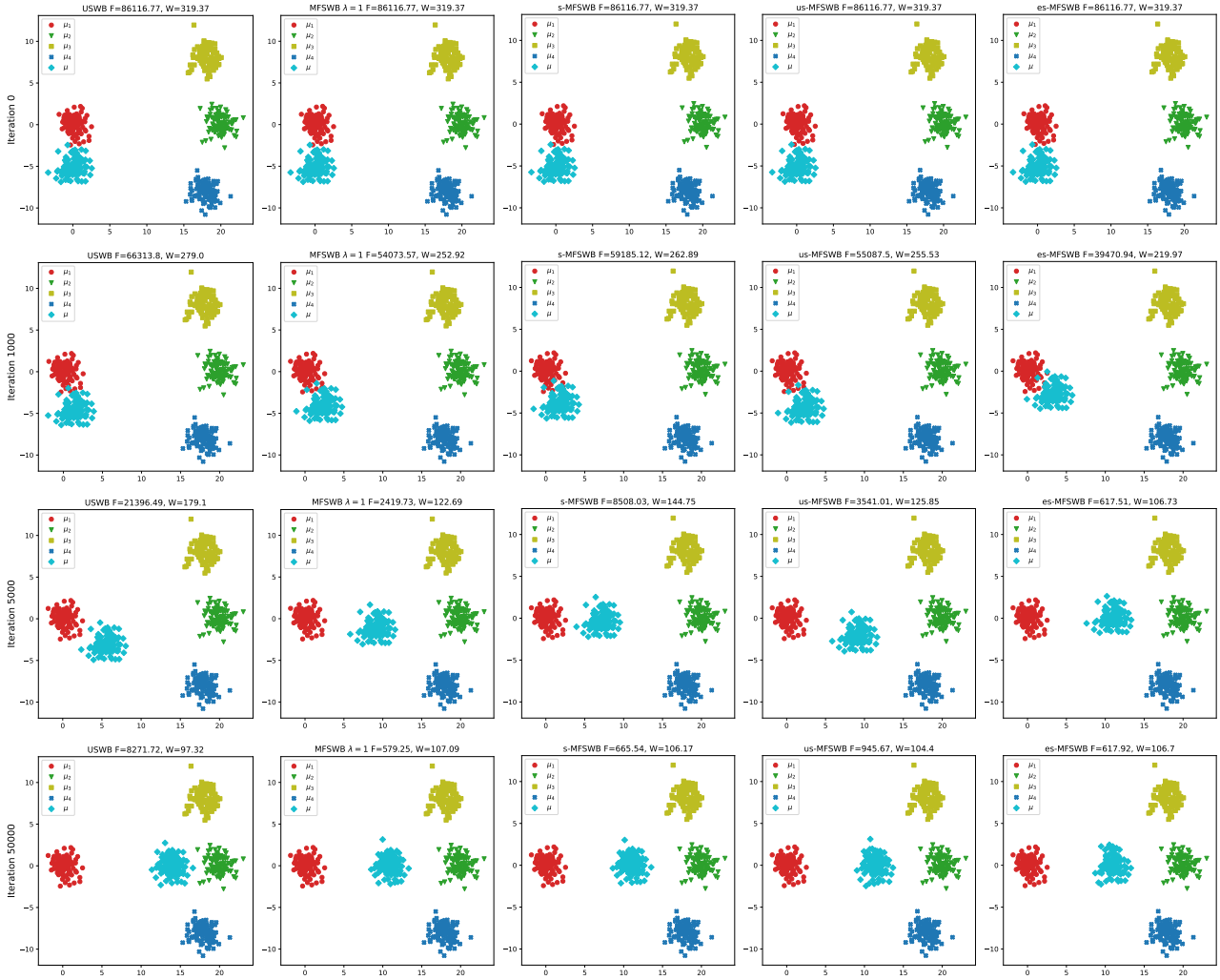


Figure 3. Barycenters from USWB, MFSWB with $\lambda = 1$, s-MFSWB, us-MFSWB, and es-MFSWB along gradient iterations with the corresponding F-metric and W-metric.

3D Point-cloud averaging. We aim to find the mean shape of point-cloud shapes by casting a point cloud $X = \{x_1, \ldots, x_n\}$ into an empirical probability measures $P_X = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$. We select two point-cloud shapes which consist of 2048 points in

ShapeNet Core-55 dataset (Chang et al., 2015). We initialize the barycenter with a spherical point-cloud. We use stochastic gradient descent with 10000 iterations of learning rate 0.01, the number of projections 10. We report the found barycenters for two car shapes in Figure 4 at the final iteration and the corresponding F-metric and W-metric at iterations 0, 1000, 5000, and 10000 in Table 1 from three independent runs. We also observe a similar phenomenon for two plane shapes in Figure 5 and Table 3.

Table 3. F-metric and W-metric along iterations in point-cloud averaging application.

Method	Iteration 0		Epoch 1000		Epoch 5000		Epoch 10000	
	F (↓)	W (↓)	F (↓)	W (↓)	F (↓)	W (↓)	F (↓)	W (↓)
USWB	746.67 ± 0.0	4814.71 ± 0.0	35.22 ± 1.04	161.11 ± 0.54	7.82 ± 0.26	109.82 ± 0.28	11.08 ± 0.06	108.52 ± 0.17
MFSWB $\lambda = 0.1$	746.67 ± 0.0	4814.71 ± 0.0	35.15 ± 0.36	159.84 ± 0.55	4.95 ± 0.23	109.14 ± 0.33	6.95 ± 0.8	107.83 ± 0.16
MFSWB $\lambda = 1$	746.67 ± 0.0	4814.71 ± 0.0	33.21 ± 2.72	151.24 ± 0.64	2.54 ± 1.5	109.66 ± 0.26	4.66 ± 2.1	$\mathbf{108.1 \pm 0.05}$
MFSWB $\lambda = 10$	746.67 ± 0.0	4814.71 ± 0.0	34.03 ± 22.6	158.66 ± 1.39	29.19 ± 14.29	122.66 ± 0.88	20.55 ± 13.57	123.65 ± 1.52
s-MFSWB	746.67 ± 0.0	4814.71 ± 0.0	36.23 ± 1.88	154.4 ± 0.67	$\mathbf{0.66 \pm 0.44}$	$\mathbf{109.17 \pm 0.34}$	2.54 ± 2.06	107.57 ± 0.19
us-MFSWB	746.67 ± 0.0	4814.71 ± 0.0	28.65 ± 1.37	144.27 ± 0.65	1.02 ± 0.8	109.67 ± 0.1	$\mathbf{1.35 \pm 0.77}$	108.2 ± 0.19
es-MFSWB	746.67 ± 0.0	4814.71 ± 0.0	$\mathbf{28.05 \pm 1.16}$	$\mathbf{143.24 \pm 0.76}$	0.99 ± 0.32	109.68 ± 0.14	1.36 ± 0.62	108.28 ± 0.07

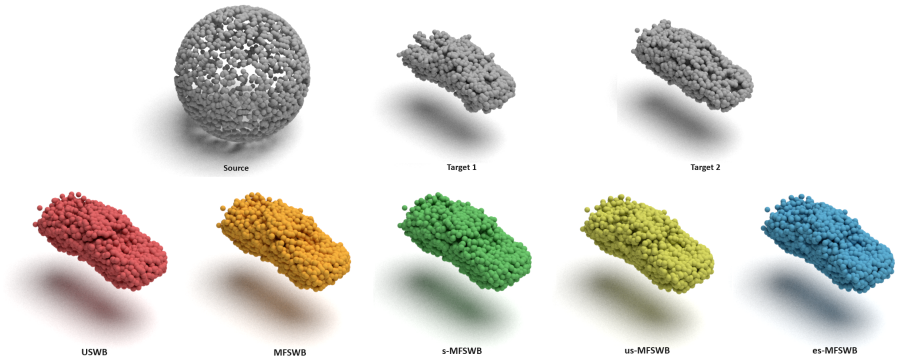


Figure 4. Averaging point-clouds with USWB, MFSWB ($\lambda = 1$), s-MFSWB, us-MFSWB, and es-MFSWB.

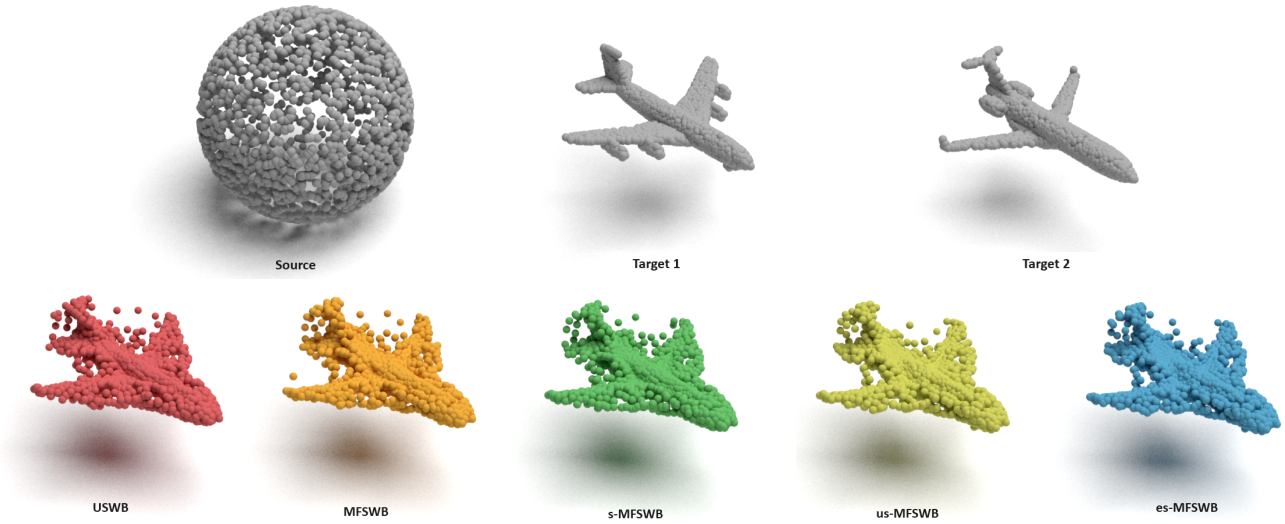


Figure 5. Averaging point-clouds with USWB, MFSWB ($\lambda = 1$), s-MFSWB, us-MFSWB, and es-MFSWB.

Color Harmonization. We first present the harmonized images of different methods including USWB, MFSWB ($\lambda = 1$), s-MFSWB, us-MFSWB, and es-MFSWB at iteration 5000 and 10000 for the demonstrated images in the main text in Figure 6-Figure 7. Moreover, we report the results of MFSWB ($\lambda = 0.1, 10$) at iteration 5000, 10000, and 20000 in Figure 8. Similarly, we repeat the same experiments with flower images in Figure 9- 11. Overall, we see that es-MFSWB helps to reduce both F-metric and W-metric faster than USWB and other surrogates. For the formal MFSWB, the performance depends significantly on the choice of λ.

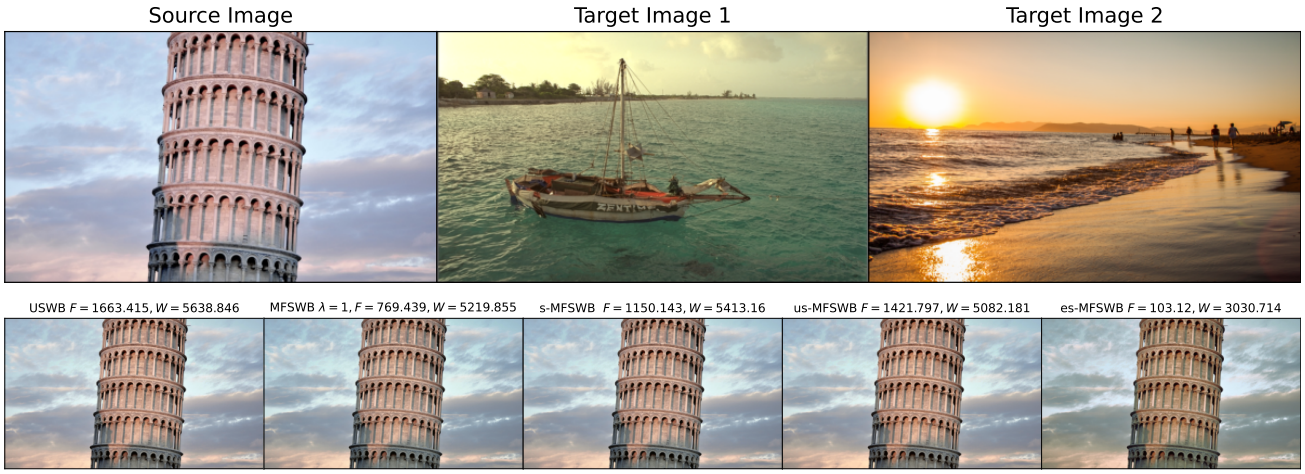


Figure 6. Harmonized images from USWB, MFSWB ($\lambda = 1$), s-MFSWB, us-MFSWB, and es-MFSWB at iteration 5000.

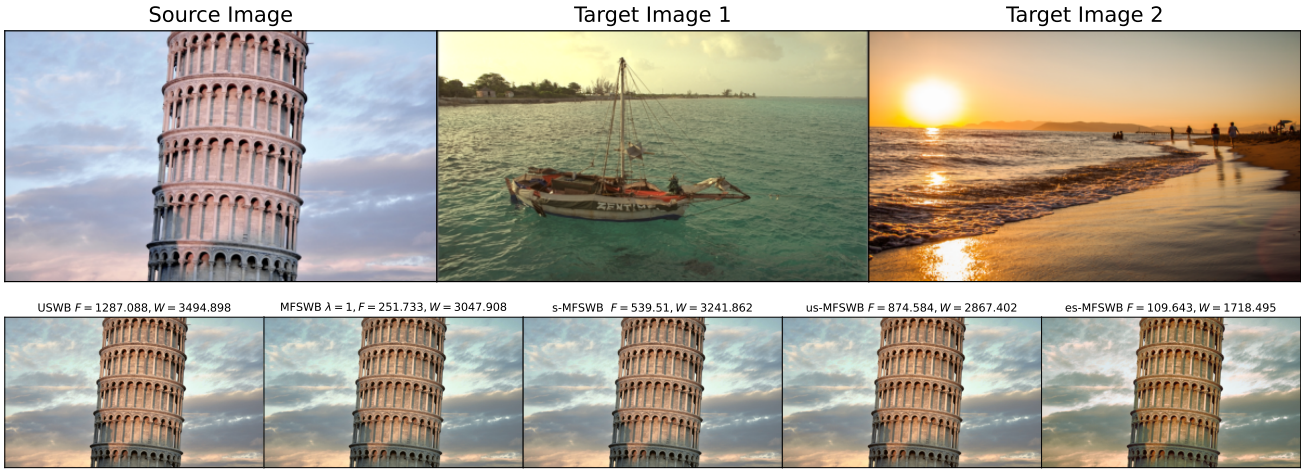


Figure 7. Harmonized images from USWB, MFSWB ($\lambda = 1$), s-MFSWB, us-MFSWB, and es-MFSWB at iteration 10000.

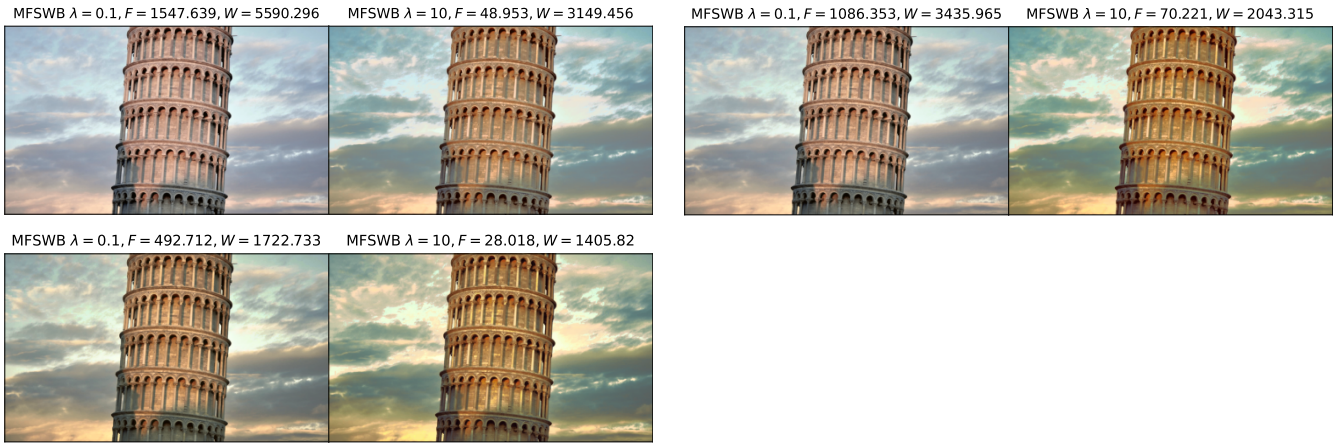


Figure 8. Harmonized images from MFSWB with $\lambda = 0.1$ and $\lambda = 10$ at iterations 5000, 10000, and 20000.

Figure 9. Harmonized images from USWB, MFSWB ($\lambda = 1$) s-MFSWB, us-MFSWB, and es-MFSWB at iteration 5000.

Figure 10. Harmonized images from USWB, MFSWB ($\lambda = 1$), s-MFSWB, us-MFSWB, and es-MFSWB at iteration 10000.

Figure 11. Color harmonized images from MFSWB with $\lambda = 0.1$ and $\lambda = 10$ at iterations 5000, 10000, and 20000.

Sliced Wasserstein autoencoder with class-fairness representation. We have the data distributions of $K \geq 1$ classes i.e., $\mu_k \in \mathcal{P}(\mathbb{R}^d)$ for $k = 1, \ldots, K$ and we would like to estimate an encoder network $f_\phi : \mathbb{R}^d \to \mathbb{R}^h$ ($\phi \in \Phi$) and a decoder network $g_\psi : \mathbb{R}^h \to \mathbb{R}^d$ ($\psi \in \Psi$ with $\mathbb{R}^h$ is a low-dimensional latent space. Given a prior distribution $\mu_0 \in \mathcal{P}(\mathbb{R}^h)$, $p \geq 1$, $\kappa_1 \in \mathbb{R}^+$, $\kappa_2 \in \mathbb{R}^+$, and a minibatch size $M \geq 1$, we perform the following optimization problem:

$$\min_{\phi,\psi} \mathbb{E}\left[\frac{1}{KM}\sum_{k=1}^{K}\sum_{i=1}^{M} c(X_{ki}, g_\psi(f_\phi(X_{ki})) + \kappa_1 SW_p^p(P_Z, P_{(f_\phi(X_k))_{k=1}^K}) + \kappa_2 \mathcal{B}(P_Z; P_{f(X_1)} : P_{f(X_K)})\right],$$

where $(X_1, \ldots, X_K) \sim \mu_1^{\otimes M} \otimes \ldots \otimes \mu_K^{\otimes M}$, $Z \sim \mu_0^{\otimes M}$, c is a reconstruction loss, $P_Z = \frac{1}{M}\sum_{i=1}^{M} \delta_{Z_i}$, $P_{(f_\phi(X_k))_{k=1}^K} = \frac{1}{KM}\sum_{k=1}^{K}\sum_{i=1}^{M} \delta_{f_\phi(X_{ki})}$, $P_{f(X_k)} = \frac{1}{M}\sum_{i=1}^{M} \delta_{X_{ki}}$ for $k = 1, \ldots, K$, and $\mathcal{B}$ denotes a barycenter loss i.e., USWB, MFSWB, s-MFSWB, us-MFSWB, and es-MFSWB. This setting can be seen as an inverse barycenter problem i.e., the barycenter is fixed and the marginals are learnt under some constraints (e.g., the reconstruction loss and the aggregated distribution loss). We train the autoencoder on MNIST dataset (LeCun et al., 1998) ($d = 28 \times 28$) with $\kappa_1 = 8.0$, $\kappa_2 = 0.5$, 250 epochs, and μ_0 is the uniform distribution on 2D ball ($h = 2$). Following the training phase, we evaluate the trained autoencoders on the MNIST test set. Similar to previous experiments, we use the metrics F (denoted as F_{latent}) and W (denoted as W_{latent}) in the latent space distributions $f_\phi \sharp \mu_1, \ldots, f_\phi \sharp \mu_K$ and the barycenter μ_0. We use the reconstruction loss (binary cross-entropy, denoted as RL), the Wasserstein-2 distance between the prior and aggregated posterior distribution in latent space $\text{W}^2_{2,\text{latent}} := W_2^2\left(\mu_0, \frac{1}{K}\sum_{k=1}^{K} f_\phi \sharp \mu_k\right)$, as well as in image space $\text{W}^2_{2,\text{image}} := W_2^2\left(g_\psi \sharp \mu_0, \frac{1}{K}\sum_{k=1}^{K} \mu_k\right)$. During evaluation, we approximate μ_0 by its empirical version of 10000 samples. We report the quantitative result in Table 2, and reconstructed images, generated images, and images of latent codes in Figure 12 in Appendix D.

We use the RMSprop optimizer with learning rate 0.01, alpha=0.99, eps=$1e-8$. As mentioned in the main text, we report the used neural network architectures:

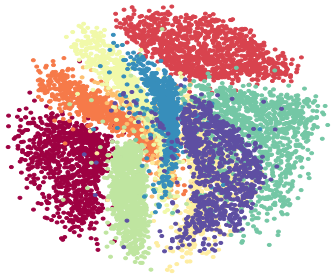
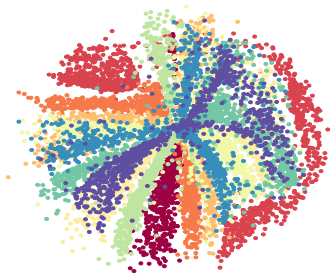
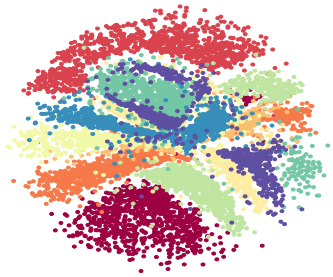
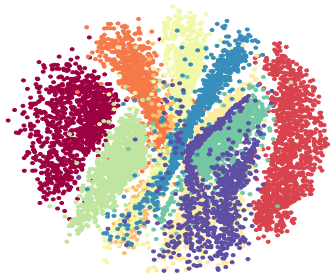
```
MNISTAutoencoder
    encoder:
        MNISTEncoder
            features:
                Conv2d(1, 16, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
```

```
            LeakyReLU(negative_slope=0.2, inplace=True)
            Conv2d(16, 16, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
            LeakyReLU(negative_slope=0.2, inplace=True)
            AvgPool2d(kernel_size=2, stride=2, padding=0)
            Conv2d(16, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
            LeakyReLU(negative_slope=0.2, inplace=True)
            Conv2d(32, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
            LeakyReLU(negative_slope=0.2, inplace=True)
            AvgPool2d(kernel_size=2, stride=2, padding=0)
            Conv2d(32, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
            LeakyReLU(negative_slope=0.2, inplace=True)
            Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
            LeakyReLU(negative_slope=0.2, inplace=True)
            AvgPool2d(kernel_size=2, stride=2, padding=1)
        fc:
            Linear(in_features=1024, out_features=128, bias=True)
            ReLU(inplace=True)
            Linear(in_features=128, out_features=2, bias=True)
decoder:
    MNISTDecoder
        fc:
            Linear(in_features=2, out_features=128, bias=True)
            Linear(in_features=128, out_features=1024, bias=True)
            ReLU(inplace=True)
        features:
            Upsample(scale_factor=2.0, mode='nearest')
            Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
            LeakyReLU(negative_slope=0.2, inplace=True)
            Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
            LeakyReLU(negative_slope=0.2, inplace=True)
            Upsample(scale_factor=2.0, mode='nearest')
            Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1))
            LeakyReLU(negative_slope=0.2, inplace=True)
            Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
            LeakyReLU(negative_slope=0.2, inplace=True)
            Upsample(scale_factor=2.0, mode='nearest')
            Conv2d(64, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
            LeakyReLU(negative_slope=0.2, inplace=True)
            Conv2d(32, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
            LeakyReLU(negative_slope=0.2, inplace=True)
            Conv2d(32, 1, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
```

We report some randomly selected reconstructed images, some randomly generated images, and the test latent codes of trained autoencoders in Figure 12. Overall, we observe that the qualitative results are consistent with the quantitive results in Table 2. From the latent spaces, we see that the proposed surrogates helps to make the codes of classes have approximately the same structure which do appear in the conventional SWAE's latent codes.

E. Computational Devices

For the Gaussian simulation, point-cloud averaging, and color harmonization, we use a HP Omen 25L desktop for conducting experiments. Additionally, for the Sliced Wasserstein Autoencoder with class-fair representation experiment, we employ the NVIDIA Tesla V100 GPU.

Method	Reconstructed Images	Generated Images	Latent Space
SWAE			
USWB			
MFSWB $\lambda = 0.1$			
MFSWB $\lambda = 1.0$			

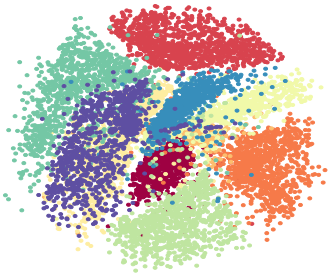
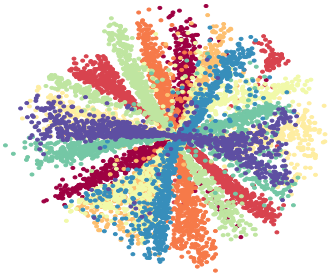
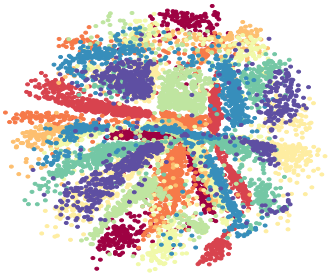
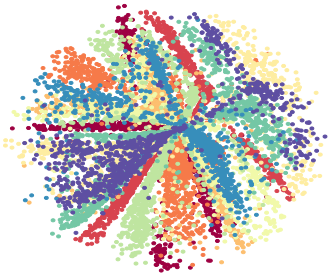
Method	Reconstructed Images	Generated Images	Latent Space
MFSWB $\lambda = 10.0$			
s-MFSWB			
us-MFSWB			
es-MFSWB			

Figure 12. Reconstructed images, generated images and latent space of all methods.