Appendix A Proof of the injectivity of the spatial Radon transform

We prove that the spatial Radon transform defined with a mapping $g(\cdot):\mathbb{R}^d\to\mathbb{R}^{d_\theta}$ is injective if and only if $g(\cdot)$ is injective. In the following contents, we use $P_k(\mathbb{R}^d)$ to denote a set of Borel probability measures with finite k-th moment on $\mathbb{R}^d$, and $f_1\equiv f_2$ is used to denote functions $f_1(\cdot):X\to\mathbb{R}$ and $f_2(\cdot):X\to\mathbb{R}$ that satisfy $f_1(x)=f_2(x)$ for $\forall x\in X$, and $f_1\not\equiv f_2$ is used to denote functions $f_1(\cdot):X\to\mathbb{R}$ and $f_2(\cdot):X\to\mathbb{R}$ that satisfy $f_1(x)\neq f_2(x)$ for certain $x\in X$. With a slight abuse of notation, we interchangeably use $f_1(x)\equiv f_2(x)$ for $\forall x\in X$ and $f_1\equiv f_2$.

Proof. By using proof by contradiction, we first prove that if $g(\cdot)$ is injective, the corresponding spatial Radon transform is injective. If the spatial Radon transform defined with an injective mapping $g(\cdot):\mathbb{R}^d\to\mathbb{R}^{d_\theta}$ is not injective, there exist $\mu,\nu\in P_k(\mathbb{R}^d)$, $\mu\not\equiv\nu$, such that $\mathcal{H}p_\mu(t,\theta;g)\equiv\mathcal{H}p_\nu(t,\theta;g)$ for $\forall t\in\mathbb{R}$ and $\forall\theta\in\mathbb{S}^{d_\theta-1}$, where p_μ and p_ν are probability density functions defined on $\mathbb{R}^d$ and $p_\mu\not\equiv p_\nu$.

From Equation (12), for $\forall t\in\mathbb{R}$ and $\forall\theta\in\mathbb{S}^{d_\theta-1}$, the spatial Radon transform can be written as:

$$\mathcal{H}p_\mu(t,\theta;g)=\mathcal{R}p_{\hat{\mu}_g}(t,\theta), \tag{20}$$

$$\mathcal{H}p_\nu(t,\theta;g)=\mathcal{R}p_{\hat{\nu}_g}(t,\theta), \tag{21}$$

where $p_{\hat{\mu}_g}$ and $p_{\hat{\nu}_g}$ refer to the probability density functions of $\hat{x}=g(x)$ and $\hat{y}=g(y)$ respectively, where $x\sim\mu$ and $y\sim\nu$. From Equations (20) and (21), we know $\mathcal{R}p_{\hat{\mu}_g}(t,\theta)\equiv\mathcal{R}p_{\hat{\nu}_g}(t,\theta)$ for $\forall t\in\mathbb{R}$ and $\forall\theta\in\mathbb{S}^{d_\theta-1}$, which implies $p_{\hat{\mu}_g}\equiv p_{\hat{\nu}_g}$ as the Radon transform is injective.

Since $g(\cdot)$ is injective, for $\forall\mathcal{X}\subseteq\mathbb{R}^d$, $x\in\mathcal{X}$ if and only if $\hat{x}=g(x)\in g(\mathcal{X})$, which implies $P(x\in\mathcal{X})=P(\hat{x}\in g(\mathcal{X}))$, $P(y\in\mathcal{X})=P(\hat{y}\in g(\mathcal{X}))$. Therefore,

$$\int_{g(\mathcal{X})}p_{\hat{\mu}_g}(\hat{x})d\hat{x}=\int_{\mathcal{X}}p_\mu(x)dx, \tag{22}$$

$$\int_{g(\mathcal{X})}p_{\hat{\nu}_g}(\hat{y})d\hat{y}=\int_{\mathcal{X}}p_\nu(y)dy. \tag{23}$$

Since $p_{\hat{\mu}_g}\equiv p_{\hat{\nu}_g}$, from Equations (22) and (23): $\int_{\mathcal{X}}p_\mu(x)dx=\int_{\mathcal{X}}p_\nu(y)dy$ for $\forall\mathcal{X}\subseteq\mathbb{R}^d$. Hence, for $\forall\mathcal{X}\subseteq\mathbb{R}^d$:

$$\int_{\mathcal{X}}\big(p_\mu(x)-p_\nu(x)\big)dx=0, \tag{24}$$

which implies $p_\mu\equiv p_\nu$, contradicting with the assumption $p_\mu\not\equiv p_\nu$. Therefore, if $\mathcal{H}_{p_\mu}\equiv\mathcal{H}_{p_\nu}$, $p_\mu\equiv p_\nu$. In addition, from the definition of the spatial Radon transform in Equation (11), it is trivial to show that if $p_\mu\equiv p_\nu$, $\mathcal{H}_{p_\mu}(t,\theta;g)\equiv\mathcal{H}_{p_\nu}(t,\theta;g)$. Therefore, $\mathcal{H}_{p_\mu}\equiv\mathcal{H}_{p_\nu}$ if and only if $p_\mu\equiv p_\nu$, i.e. the spatial Radon transform $\mathcal{H}$ defined with an injective mapping $g(\cdot):\mathbb{R}^d\to\mathbb{R}^{d_\theta}$ is injective.

We now prove that if the spatial Radon transform defined with a mapping $g(\cdot):\mathbb{R}^d\to\mathbb{R}^{d_\theta}$ is injective, $g(\cdot)$ must be injective. Again, we use proof by contradiction. If $g(\cdot)$ is not injective, there exist $x_0,y_0\in\mathbb{R}^d$ such that $x_0\neq y_0$ and $g(x_0)=g(y_0)$. For two Dirac measures μ_1 and ν_1 which probability density functions are $p_{\mu_1}(x)=\delta(x-x_0)$ and $p_{\nu_1}(y)=\delta(y-y_0)$, respectively, we know $\mu_1\not\equiv\nu_1$ as $x_0\neq y_0$.

We define variables $x\sim\mu_1$ and $y\sim\nu_1$. Then for variables $\hat{x}=g(x)$ and $\hat{y}=g(y)$, we denote their probability density functions by p_{μ_2} and p_{ν_2}, respectively. It is trivial to derive

$$p_{\mu_2}(\hat{x})=\delta(\hat{x}-g(x_0)), \tag{25}$$

$$p_{\nu_2}(\hat{y})=\delta(\hat{y}-g(y_0)), \tag{26}$$

which implies $p_{\mu_2}\equiv p_{\nu_2}$ as $g(x_0)=g(y_0)$.

From Equations (20), (21), (25) and (26), for $\forall t\in\mathbb{R}$ and $\forall\theta\in\mathbb{S}^{d_\theta-1}$:

$$\begin{aligned}\mathcal{H}p_{\mu_1}(t,\theta;g)&=\mathcal{R}p_{\mu_2}(t,\theta),\\&=\mathcal{R}p_{\nu_2}(t,\theta),\\&=\mathcal{H}p_{\nu_1}(t,\theta;g),\end{aligned} \tag{27}$$

which implies $\mathcal{H}_{p_{\mu_1}}\equiv\mathcal{H}_{p_{\nu_1}}$, contradicting with the assumption that the spatial Radon transform is injective. Therefore, if the spatial Radon transform is injective, $g(\cdot)$ must be injective. We conclude that the spatial Radon transform is injective if and only if the mapping $g(\cdot)$ is an injection. □

Appendix B Proof of Remark 2

We provide a proof for the claim in Remark 2 that the spatial Radon transform includes the vanilla Radon transform and the polynomial GRT as special cases.

Proof. Given a probability measure $\mu \in P(\mathbb{R}^d)$ which probability density function is p_μ, the spatial Radon transform of p_μ is defined as:

$$\mathcal{H}p_\mu(t,\theta;g) = \int_{\mathbb{R}^d} p_\mu(x)\delta(t - \langle g(x),\theta\rangle)dx, \tag{28}$$

where $t \in \mathbb{R}$ and $\theta \in \mathbb{S}^{d_\theta - 1}$ are the parameters of hypersurfaces in $\mathbb{R}^d$. When the mapping $g(\cdot)$ is an identity mapping, i.e. $g(x) = x$ for $\forall x \in \mathbb{R}^d$, the spatial Radon transform degenerates to the vanilla Radon transform:

$$\begin{aligned}\mathcal{H}p_\mu(t,\theta;g) &= \int_{\mathbb{R}^d} p_\mu(x)\delta(t - \langle x,\theta\rangle)dx \\ &= \mathcal{R}p_\mu(t,\theta).\end{aligned} \tag{29}$$

[Ehrenpreis, 2003] provides a class of injective GRTs named polynomial GRTs by adopting homogeneous polynomial functions with an odd degree m as the defining function:

$$\begin{aligned}\mathcal{G}p_\mu(t,\theta) = \int_{\mathbb{R}^d} p_\mu(x)\delta(t - \sum_{i=1}^{d_\alpha}\theta_i x^{\alpha_i})dx, \\ \text{s.t.} |\alpha_i| = m,\end{aligned} \tag{30}$$

where $\alpha_i = (\eta_{i,1},...,\eta_{i,d}) \in \mathbb{N}^d$, $|\alpha_i| = \sum_{j=1}^d \eta_{i,j}$, $x^{\alpha_i} = \prod_{j=1}^d x_j^{\eta_{i,j}}$ for $x = (x_1,...,x_d) \in \mathbb{R}^d$, d_α is the number of all possible multi-indices α_i that satisfies $|\alpha_i| = m$, and $\theta = (\theta_1,...,\theta_{d_\alpha}) \in \mathbb{S}^{d_\alpha - 1}$.

In spatial Radon transform, for $\forall x \in \mathbb{R}^d$, when the mapping $g(\cdot)$ is defined as:

$$g(x) = (x^{\alpha_1},...,x^{\alpha_{d_\alpha}}), \tag{31}$$

the spatial Radon transform is equivalent to the polynomial GRT defined in Equation (30):

$$\begin{aligned}\mathcal{H}p_\mu(t,\theta;g) &= \int_{\mathbb{R}^d} p_\mu(x)\delta(t - \langle g(x),\theta\rangle)dx \\ &= \int_{\mathbb{R}^d} p_\mu(x)\delta(t - \sum_{i=1}^{d_\alpha}\theta_i x^{\alpha_i})dx.\end{aligned} \tag{32}$$

□

Appendix C Proof of Theorem 1

We provide a proof that the ASWD defined with a mapping $g(\cdot):\mathbb{R}^d\to\mathbb{R}^{d_\theta}$ is a metric on $P_k(\mathbb{R}^d)$, if and only if $g(\cdot)$ is injective. In what follows, we denote a set of Borel probability measures with finite k-th moment on $\mathbb{R}^d$ by $P_k(\mathbb{R}^d)$, and use $\mu,\nu\in P_k(\mathbb{R}^d)$ to refer to two probability measures whose probability density functions are p_μ and p_ν.

Proof. **Symmetry:** Since the k-Wasserstein distance is a metric thus symmetric [Villani, 2008]:

$$W_k\big(\mathcal{H}p_\mu(\cdot,\theta;g),\mathcal{H}p_\nu(\cdot,\theta;g)\big)=W_k\big(\mathcal{H}p_\nu(\cdot,\theta;g),\mathcal{H}p_\mu(\cdot,\theta;g)\big). \tag{33}$$

Therefore,

$$\begin{aligned}\text{ASWD}_k(\mu,\nu;g)&=\left(\int_{\mathbb{S}^{d_\theta-1}}W_k^k\big(\mathcal{H}p_\mu(\cdot,\theta;g),\mathcal{H}p_\nu(\cdot,\theta;g)\big)d\theta\right)^{\frac{1}{k}}\\&=\left(\int_{\mathbb{S}^{d_\theta-1}}W_k^k\big(\mathcal{H}p_\nu(\cdot,\theta;g),\mathcal{H}p_\mu(\cdot,\theta;g)\big)d\theta\right)^{\frac{1}{k}}=\text{ASWD}_k(\nu,\mu;g).\end{aligned}$$

Triangle inequality: Given an injective mapping $g(\cdot):\mathbb{R}^d\to\mathbb{R}^{d_\theta}$ and probability measures $\mu_1,\mu_2,\mu_3\in P_k(\mathbb{R}^d)$, since the k-Wasserstein distance satisfies the triangle inequality [Villani, 2008], the following inequality holds:

$$\begin{aligned}\text{ASWD}_k(\mu_1,\mu_3;g)&=\left(\int_{\mathbb{S}^{d_\theta-1}}W_k^k\big(\mathcal{H}p_{\mu_1}(\cdot,\theta;g),\mathcal{H}p_{\mu_3}(\cdot,\theta;g)\big)d\theta\right)^{\frac{1}{k}}\\&\le\left(\int_{\mathbb{S}^{d_\theta-1}}\big(W_k(\mathcal{H}p_{\mu_1}(\cdot,\theta;g),\mathcal{H}p_{\mu_2}(\cdot,\theta;g))\right.\\&\quad\left.+W_k(\mathcal{H}p_{\mu_2}(\cdot,\theta;g),\mathcal{H}p_{\mu_3}(\cdot,\theta;g))\big)^k d\theta\right)^{\frac{1}{k}}\\&\le\left(\int_{\mathbb{S}^{d_\theta-1}}W_k^k\big(\mathcal{H}p_{\mu_1}(\cdot,\theta;g),\mathcal{H}p_{\mu_2}(\cdot,\theta;g)\big)d\theta\right)^{\frac{1}{k}}\\&\quad+\left(\int_{\mathbb{S}^{d_\theta-1}}W_k^k\big(\mathcal{H}p_{\mu_2}(\cdot,\theta;g),\mathcal{H}p_{\mu_3}(\cdot,\theta;g)\big)d\theta\right)^{\frac{1}{k}}\\&=\text{ASWD}_k(\mu_1,\mu_2;g)+\text{ASWD}_k(\mu_2,\mu_3;g),\end{aligned}$$

where the second inequality is due to the Minkowski inequality in $L^k(\mathbb{S}^{d_\theta-1})$.

Identity of indiscernibles: Since $W_k(\mu,\mu)=0$ for $\forall\mu\in P_k(\mathbb{R}^d)$,we have

$$\text{ASWD}_k(\mu,\mu;g)=\left(\int_{\mathbb{S}^{d_\theta-1}}W_k^k\big(\mathcal{H}p_\mu(\cdot,\theta;g),\mathcal{H}p_\mu(\cdot,\theta;g)\big)d\theta\right)^{\frac{1}{k}}=0, \tag{34}$$

for $\forall\mu\in P_k(\mathbb{R}^d)$. Conversely, for $\forall\mu,\nu\in P_k(\mathbb{R}^d)$, if $\text{ASWD}_k(\mu,\nu;g)=0$, from the definition of the ASWD:

$$\text{ASWD}_k(\mu,\nu;g)=\left(\int_{\mathbb{S}^{d_\theta-1}}W_k^k\big(\mathcal{H}p_\mu(\cdot,\theta;g),\mathcal{H}p_\nu(\cdot,\theta;g)\big)d\theta\right)^{\frac{1}{k}}=0, \tag{35}$$

Due to the non-negativity of k-th Wasserstein distance as it is a metric on $P_k(\mathbb{R}^d)$ and the continuity of $W_k(\cdot,\cdot)$ on $P_k(\mathbb{R}^d)$ [Villani, 2008], $W_k\big(\mathcal{H}p_\mu(\cdot,\theta;g),\mathcal{H}p_\nu(\cdot,\theta;g)\big)=0$ holds for $\forall\theta\in\mathbb{S}^{d_\theta-1}$ if and only if $\mathcal{H}p_\mu(\cdot,\theta;g)\equiv\mathcal{H}p_\nu(\cdot,\theta;g)$. Again, given the spatial Radon transform is injective when $g(\cdot)$ is injective (see the proof in Appendix A), $\mathcal{H}p_\mu(\cdot,\theta;g)\equiv\mathcal{H}p_\nu(\cdot,\theta;g)$ implies $p_\mu\equiv p_\nu$ and $\mu\equiv\nu$ if $g(\cdot)$ is injective.

In addition, if $g(\cdot)$ is not injective, the spatial Radon transform is not injective (see the proof in Appendix A), then $\exists\mu,\nu\in P_k(\mathbb{R}^d)$, $\mu\not\equiv\nu$ such that $\mathcal{H}_{p_\mu}(\cdot,\theta;g)\equiv\mathcal{H}_{p_\nu}(\cdot,\theta;g)$, which implies $\text{ASWD}_k(\mu,\nu;g)=0$ for $\mu\not\equiv\nu$. Therefore, the ASWD satisfies the identity of indiscernibles if and only if $g(\cdot)$ is injective.

Non-negativity: The three axioms of a distance metric, i.e. symmetry, triangle inequality, and identity of indiscernibles imply the non-negativity of the ASWD. Since the Wasserstein distance is non-negative, for $\forall\mu,\nu\in P_k(\mathbb{R}^d)$, it can also be straightforwardly proved the ASWD between μ and ν is non-negative:

$$\begin{aligned}\text{ASWD}_k(\mu,\nu;g)&=\left(\int_{\mathbb{S}^{d_\theta-1}}W_k^k\big(\mathcal{H}p_\mu(\cdot,\theta;g),\mathcal{H}p_\nu(\cdot,\theta;g)\big)d\theta\right)^{\frac{1}{k}}\\&\ge\left(\int_{\mathbb{S}^{d_\theta-1}}0^k d\theta\right)^{\frac{1}{k}}=0.\end{aligned} \tag{36}$$

Therefore, the ASWD is a metric on $P_k(\mathbb{R}^d)$ if and only if $g(\cdot)$ is injective. □

Appendix D Proof of Corollary 1.1

We prove that the ASWD defined with optimal mappings $g^*(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d_\theta}$ is also a metric on $P_k(\mathbb{R}^d)$ when the optimization is confined to the set of bounded and injective functions $\{g(x) : \mathbb{R}^d \to \mathbb{R}^{d_\theta} | \exists M \in \mathbb{R}, \forall x \in \mathbb{R}^d, ||g(x)||_2 \leq M\}$. Recall that given two measures $\mu, \nu \in P_k(\mathbb{R}^d)$, the ASWD defined with the optimal mapping $g^*(\cdot) = \underset{g}{\text{argmax}}(\text{ASWD}_k(\mu,\nu;g))$ is defined as:

$$\text{ASWD}_k(\mu,\nu;g^*) = \sup_g \{\text{ASWD}_k(\mu,\nu;g)\}. \tag{37}$$

Proof. **Symmetry:** Since the k-Wasserstein distance is a metric thus symmetric [Villani, 2008]:

$$W_k\big(\mathcal{H}p_\mu(\cdot,\theta;g^*), \mathcal{H}p_\nu(\cdot,\theta;g^*)\big) = W_k\big(\mathcal{H}p_\nu(\cdot,\theta;g^*), \mathcal{H}p_\mu(\cdot,\theta;g^*)\big). \tag{38}$$

Therefore,

$$\begin{aligned}\text{ASWD}_k(\mu,\nu;g^*) &= \left(\int_{\mathbb{S}^{d_\theta-1}} W_k^k\big(\mathcal{H}p_\mu(\cdot,\theta;g^*), \mathcal{H}p_\nu(\cdot,\theta;g^*)\big) d\theta\right)^{\frac{1}{k}} \\ &= \left(\int_{\mathbb{S}^{d_\theta-1}} W_k^k\big(\mathcal{H}p_\nu(\cdot,\theta;g^*), \mathcal{H}p_\mu(\cdot,\theta;g^*)\big) d\theta\right)^{\frac{1}{k}} = \text{ASWD}_k(\nu,\mu;g^*).\end{aligned}$$

Triangle inequality: It is trivial to prove that the ASWD defined in Eq. (37) is finite when the mapping $g(\cdot)$ is confined to the set of bounded functions $\{g(x) : \mathbb{R}^d \to \mathbb{R}^{d_\theta} | \exists M \in \mathbb{R}, \forall x \in \mathbb{R}^d, ||g(x)||_2 \leq M\}$. We then prove that the the ASWD defined in Eq. (37) satisfies the triangle inequality.

Denote by μ_1, μ_2, $\mu_3 \in P_k(\mathbb{R}^d)$ three measures, then the following equations hold for the ASWD defined with optimal mappings:

$$\begin{aligned}\text{ASWD}_k(\mu_1,\mu_2;g_1^*) &\leq \big(\text{ASWD}_k(\mu_1,\mu_3;g_1^*)^k + \text{ASWD}_k(\mu_2,\mu_3;g_1^*)^k\big)^{\frac{1}{k}} &(39)\\ &\leq \text{ASWD}_k(\mu_1,\mu_3;g_1^*) + \text{ASWD}_k(\mu_2,\mu_3;g_1^*) &(40)\\ &\leq \sup_g\{\text{ASWD}_k(\mu_1,\mu_3;g)\} + \sup_g\{\text{ASWD}_k(\mu_2,\mu_3;g)\} &(41)\\ &= \text{ASWD}_k(\mu_1,\mu_3;g_2^*) + \text{ASWD}_k(\mu_2,\mu_3;g_3^*), &(42)\end{aligned}$$

where the first two inequalities are from the metric property of the ASWD, and g_1^*, g_2^*, and g_3^* correspond to optimal mappings that result in the supremum of ASWDs between μ_1 and μ_2, μ_1 and μ_3, μ_2 and μ_3, respectively.

Identity of indiscernibles: Since $W_k(\mu,\mu) = 0$ for $\forall \mu \in P_k(\mathbb{R}^d)$,we have

$$\text{ASWD}_k(\mu,\mu;g^*) = \left(\int_{\mathbb{S}^{d_\theta-1}} W_k^k\big(\mathcal{H}p_\mu(\cdot,\theta;g^*), \mathcal{H}p_\mu(\cdot,\theta;g^*)\big) d\theta\right)^{\frac{1}{k}} = 0, \tag{43}$$

for $\forall \mu \in P_k(\mathbb{R}^d)$. Conversely, for $\forall \mu,\nu \in P_k(\mathbb{R}^d)$, if $\text{ASWD}_k(\mu,\nu;g^*) = 0$, from Eq. (37):

$$\text{ASWD}_k(\mu,\nu;g^*) = \left(\int_{\mathbb{S}^{d_\theta-1}} W_k^k\big(\mathcal{H}p_\mu(\cdot,\theta;g^*), \mathcal{H}p_\nu(\cdot,\theta;g^*)\big) d\theta\right)^{\frac{1}{k}} = 0. \tag{44}$$

Due to the non-negativity of k-th Wasserstein distance as it is a metric on $P_k(\mathbb{R}^d)$ and the continuity of $W_k(\cdot,\cdot)$ on $P_k(\mathbb{R}^d)$ [Villani, 2008], $W_k\big(\mathcal{H}p_\mu(\cdot,\theta;g^*), \mathcal{H}p_\nu(\cdot,\theta;g^*)\big) = 0$ holds for $\forall \theta \in \mathbb{S}^{d_\theta-1}$, which implies $\mathcal{H}p_\mu(\cdot,\theta;g^*) \equiv \mathcal{H}p_\nu(\cdot,\theta;g^*)$ for $\forall \theta \in \mathbb{S}^{d_\theta-1}$. Therefore, given the spatial Radon transform is injective when $g^*(\cdot)$ is injective, $\mathcal{H}p_\mu(\cdot,\theta;g^*) \equiv \mathcal{H}p_\nu(\cdot,\theta;g^*)$ implies $p_\mu \equiv p_\nu$ and $\mu \equiv \nu$.

Non-negativity: Since the Wasserstein distance is non-negative, for $\forall \mu, \nu \in P_k(\mathbb{R}^d)$, the ASWD defined with optimal mappings $g(\cdot)$ between μ and ν is also non-negative:

$$\begin{aligned}\text{ASWD}_k(\mu,\nu;g^*) &= \left(\int_{\mathbb{S}^{d_\theta-1}} W_k^k\big(\mathcal{H}p_\mu(\cdot,\theta;g^*), \mathcal{H}p_\nu(\cdot,\theta;g^*)\big) d\theta\right)^{\frac{1}{k}} \\ &\geq \left(\int_{\mathbb{S}^{d_\theta-1}} 0^k d\theta\right)^{\frac{1}{k}} = 0. &(45)\end{aligned}$$

Therefore, the ASWD defined in Eq. (37) is non-negative, symmetric, and satisfies the triangle inequality and the identity of indiscernibles, i.e. the ASWD defined with optimal mappings $g^*(\cdot)$ is also a metric.

□

Appendix E Pseudocode for the empirical version of the ASWD

Algorithm 1 The augmented sliced Wasserstein distance. All of the for loops can be parallelized.

Require: Sets of samples $\{x_n \in \mathbb{R}^d\}_{n=1}^N$, $\{y_n \in \mathbb{R}^d\}_{n=1}^N$;
Require: Randomly initialized injective neural network $g_\omega(\cdot): \mathbb{R}^d \to \mathbb{R}^{d_\theta}$;
Require: Number of projections L, hyperparameter λ, learning rate ϵ, number of iterations M;
Initialize $D=0, L_\lambda=0, m=1$;
while ω has not converged and $m \leq M$ **do**
    Draw a set of samples $\{\theta_l\}_{l=1}^L$ from $\in \mathbb{S}^{d_\theta - 1}$;
    for $n=1$ to N **do**
        Compute $g_\omega(x_n)$ and $g_\omega(y_n)$;
        Calculate the regularization term $L_\lambda \leftarrow L_\lambda + \frac{\lambda}{N}(||g_\omega(x_n)||_2 + ||g_\omega(y_n)||_2)$;
    end for
    for $l=1$ to L **do**
        Compute $\beta(x_n,\theta_l) = \langle g_\omega(x_n), \theta_l \rangle$, $\beta(y_n,\theta_l) = \langle g_\omega(y_n), \theta_l \rangle$ for each n;
        Sort $\beta(x_n,\theta_l)$ and $\beta(y_n,\theta_l)$ in ascending order s.t. $\beta(x_{I_x^l[n]},\theta_l) \leq \beta(x_{I_x^l[n+1]},\theta_l)$ and $\beta(y_{I_y^l[n]},\theta_l) \leq \beta(y_{I_y^l[n+1]},\theta_l)$;
        Calculate the ASWD: $D \leftarrow D + (\frac{1}{L}\sum_{n=1}^N |\beta(x_{I_x^l[n]},\theta_l) - \beta(y_{I_y^l[n]},\theta_l)|^k)^{\frac{1}{k}}$;
    end for
    $\mathcal{L} \leftarrow D - L_\lambda$;
    Update ω by gradient ascent $\omega \leftarrow \omega + \epsilon \cdot \nabla_\omega \mathcal{L}$;
    Reset $D=0, L_\lambda=0$, update $m \leftarrow m+1$;
end while
Draw a set of samples $\{\theta_l\}_{l=1}^L$ from $\in \mathbb{S}^{d_\theta - 1}$;
for $n=1$ to N **do**
    Compute $g_\omega(x_n)$ and $g_\omega(y_n)$;
end for
for $l=1$ to L **do**
    Compute $\beta(x_n,\theta_l) = \langle g_\omega(x_n), \theta_l \rangle$, $\beta(y_n,\theta_l) = \langle g_\omega(y_n), \theta_l \rangle$ for each n;
    Sort $\beta(x_n,\theta_l)$ and $\beta(y_n,\theta_l)$ in ascending order s.t. $\beta(x_{I_x^l[n]},\theta_l) \leq \beta(x_{I_x^l[n+1]},\theta_l)$ and $\beta(y_{I_y^l[n]},\theta_l) \leq \beta(y_{I_y^l[n+1]},\theta_l)$;
    Calculate the ASWD: $D \leftarrow D + (\frac{1}{L}\sum_{n=1}^N |\beta(x_{I_x^l[n]},\theta_l) - \beta(y_{I_y^l[n]},\theta_l)|^k)^{\frac{1}{k}}$;
end for
Output: Augmented sliced Wasserstein distance D.

Appendix F Experimental setups

F.1 Hyperparameters in the sliced Wasserstein flow experiment

We randomly generate 500 samples both for target distributions and source distributions. We initialize the source distributions μ_0 as standard normal distributions $\mathcal{N}(0,I)$, where I is a 2-dimensional identity matrix. We update source distributions using Adam optimizer, and set the learning rate=0.002. For all methods, we set the order $k=2$. When testing the ASWD, the number of iterations M in Algorithm 1 is set to 10. Empirical errors in the experiment are found to be not sensitive to the choice of λ in a candidate set of $\{0.01, 0.05, 0.1, 0.5\}$. The reported results are produced with $\lambda=0.1$.

F.2 Network architecture in the generative modeling experiment

Denote a convolutional layer whose kernel size is s with C kernels by $\text{Conv}_C(s\times s)$, and a fully-connected layer whose input and output layer have s_1 and s_2 neurons by $\text{FC}(s_1\times s_2)$. The network structure used in the generative modeling experiment is configured to be the same as described in [Nguyen et al., 2021]:

$$
\begin{aligned}
&h_\psi : (64\times 64\times 3) \to \text{Conv}_{64}(4\times 4) \to \text{LeakyReLU}(0.2) \to \\
&\text{Conv}_{128}(4\times 4) \to \text{BatchNormalization} \to \text{LeakyReLU}(0.2) \to \\
&\text{Conv}_{256}(4\times 4) \to \text{BatchNormalization} \to \text{LeakyReLU}(0.2) \to \\
&\text{Conv}_{512}(4\times 4) \to \text{BatchNormalization} \to \text{Tanh} \xrightarrow{\text{Output}} (512\times 4\times 4) \\
&D_\Psi : \text{Conv}_1(4\times 4) \to \text{Sigmoid} \xrightarrow{\text{Output}} (1\times 1\times 1) \\
&G_\Phi : z\in\mathbb{R}^{32} \to \text{ConvTranspose}_{512}(4\times 4) \to \\
&\text{BatchNormalization} \to \text{ReLU} \to \text{ConvTranspose}_{256}(4\times 4) \to \\
&\text{BatchNormalization} \to \text{ReLU} \to \text{ConvTranspose}_{128}(4\times 4) \to \\
&\text{BatchNormalization} \to \text{ReLU} \to \text{ConvTranspose}_{64}(4\times 4) \to \\
&\text{BatchNormalization} \to \text{ConvTranspose}_{3}(4\times 4) \to \text{Tanh} \\
&\xrightarrow{\text{Ouput}} (64\times 64\times 3) \\
&\phi : \text{FC}(8192\times 8192) \xrightarrow{\text{Output}} (8192)\text{-dimensional vector}
\end{aligned}
$$

We train the models with the Adam optimizer, and set the batch size to 512. Following the setup in [Nguyen et al., 2021], the learning rate is set to 0.0005 and beta=(0.5, 0.999) for both CIFAR10 dataset and CelebA dataset. For all methods, we set the order k to 2. For the ASWD, the number of iterations M in Algorithm 1 is set to 5. The hyperparameter λ is set to 0.5 to introduce slightly larger regularization of the optimization objective due to the small output values from the feature layer h_ψ.

Appendix G Additional results in the sliced Wasserstein flow experiment

G.1 Full experimental results on the sliced Wasserstein experiment

Figure 4 shows the full experimental results on the sliced Wasserstein flow experiment.

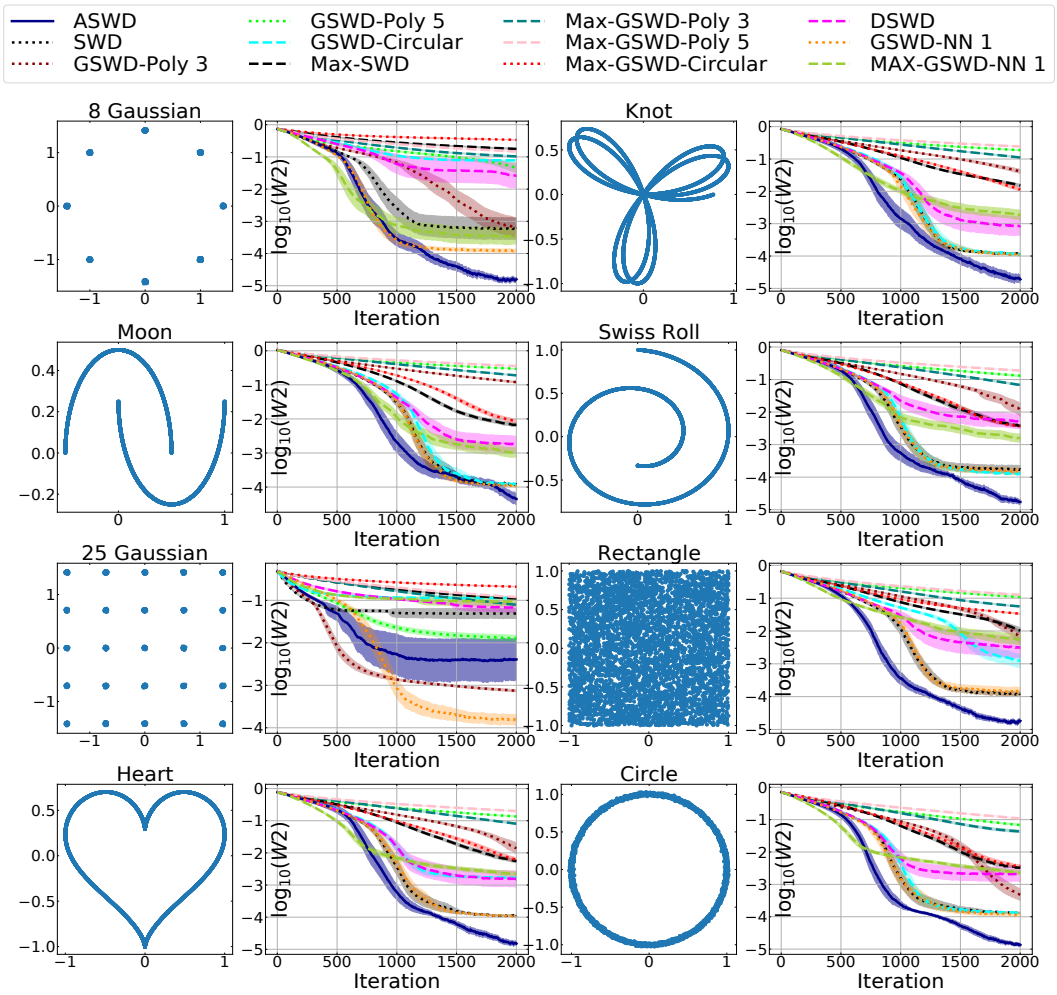


Figure 4: Full experimental results on the sliced Wasserstein flow example. The first and third columns are target distributions. The second and fourth columns are log 2-Wasserstein distances between the target distributions and the source distributions. The horizontal axis shows the number of training iterations. Solid lines and shaded areas represent the average values and 95% confidence intervals of log 2-Wasserstein distances over 50 runs.

G.2 Ablation study

In this ablation study, we compare ASWDs constructed by different mappings to GSWDs with different predefined defining functions, and investigate the effects of the optimization and injectivity of the adopted mapping $g_\omega(\cdot)$ used in the ASWDs. In what follows, "ASWD-vanilla" is used to denote ASWDs that employ randomly initialized neural network $\phi_\omega(\cdot)$ to parameterize the injective mapping $g_\omega(\cdot) = [\cdot, \phi_\omega(\cdot)]$, i.e. the mapping $g_\omega(\cdot)$ is not optimized in the ASWD-vanilla and the results of ASWD-vanilla reported in Figure 5 are obtained by slicing with random hypersurfaces. Furthermore, the "ASWD-non-injective" refers to ASWDs that do not use the injectivity trick, i.e. the mapping $g_\omega(\cdot) = \phi_\omega(\cdot)$ is not guaranteed to be injective. In addition, the "ASWD-vanilla-non-injective" adopts both setups in the "ASWD-vanilla" and "ASWD-non-injective", resulting in a random non-injective mapping $g_\omega(\cdot)$. The reported experiment results in this ablation study is calculated over 50 runs, and the neural network $\phi_\omega(\cdot)$ is reinitialized randomly in each run.

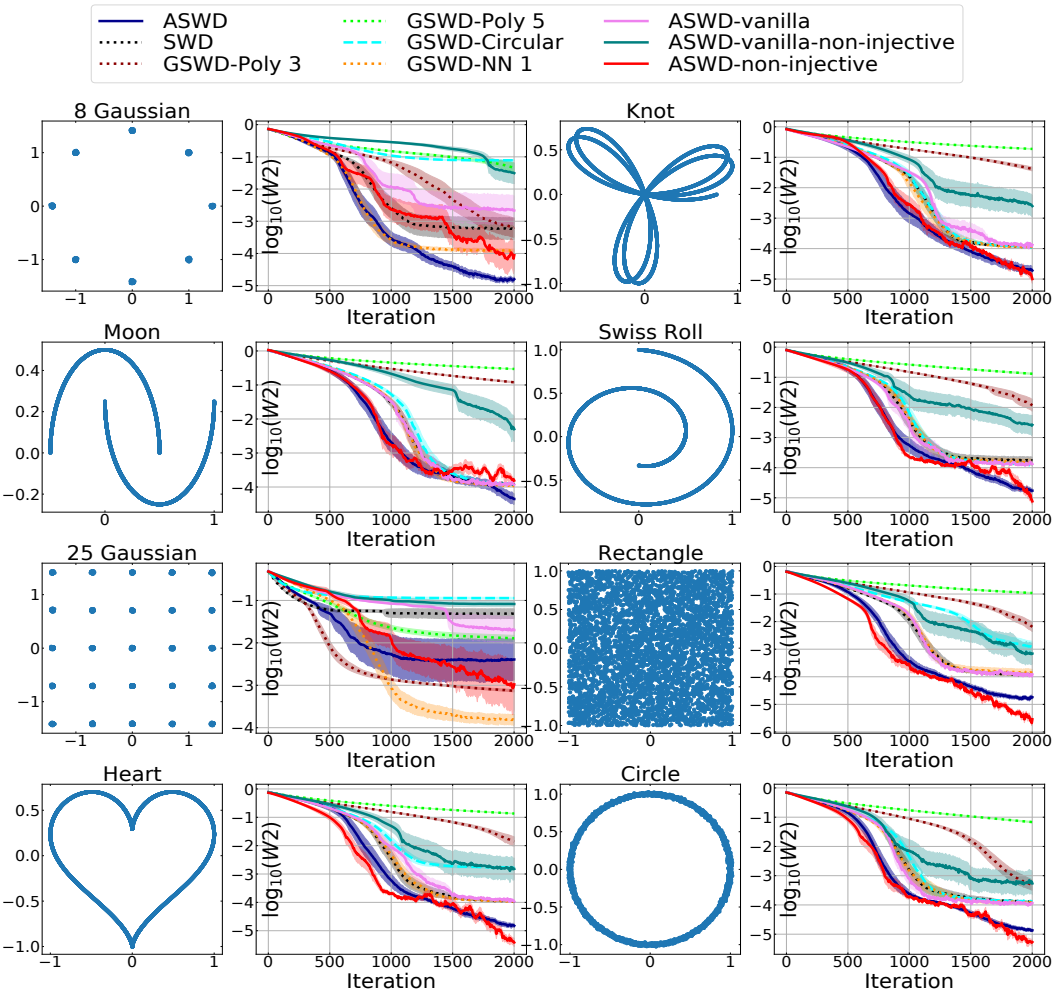
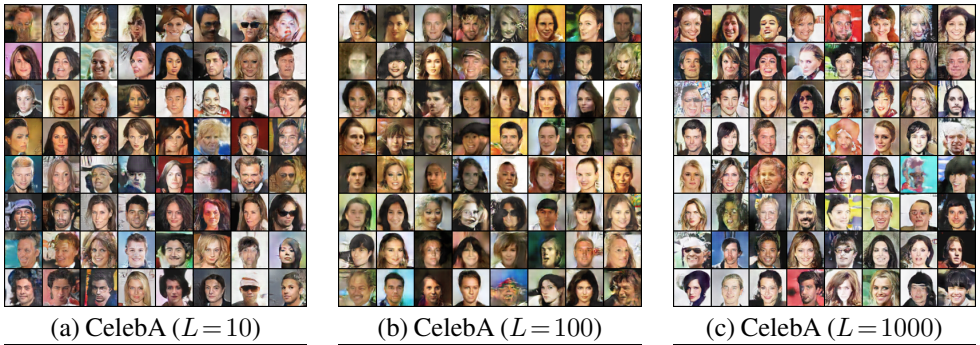


Figure 5: Ablation study on the impact from injective neural networks and the optimization of hypersurfaces on the ASWD. ASWDs with different mappings are compared to GSWDs with different defining functions. The first and third columns show target distributions. The second and fourth columns plot log 2-Wasserstein distances between the target distributions and the source distributions. In the second and fourth columns, the horizontal axis shows the number of training iterations. Solid lines and shaded areas represent the average values and 95% confidence intervals of log 2-Wasserstein distances over 50 runs.

From Figure 5, it can be observed that the ASWD-vanilla shows comparable performance to GSWDs defined by polynomial and circular defining functions, which implies GSWDs with predefined defining functions are as uninformative as slicing distributions with random hypersurfaces constructed by the ASWD. In GSWDs, the hypersurfaces are predefined and cannot be optimized since they are determined by the functional forms of the defining functions. On the contrary, we found that the optimization of hypersurfaces in the ASWD framework can help improve the performance of the slice-based Wasserstein distance. As in Figure 5, the ASWD and the ASWD-non-injective present significantly better performance than methods that do not optimize their hypersurfaces (ASWD-vanilla, ASWD-vanilla-non-injective, and GSWDs). In terms of the impact of the injectivity of the mapping g_ω, in this experiment, the ASWD-vanilla exhibits smaller 2-Wasserstein distances than the ASWD-vanilla-non-injective in all tested distributions, and the ASWD leads to more stable training than the ASWD-non-injective. Therefore, the injectivity of the mapping $g_\omega(\cdot)$ does not only guarantee the ASWD to be a valid distance metric as proved in Section 3, but also better empirical performance in this experiment setup.

Appendix H Additional results in the generative modeling experiment

H.1 Sampels of generated images of CIFAR10 and CelebA datasets



(a) CelebA ($L=10$) (b) CelebA ($L=100$) (c) CelebA ($L=1000$)

(d) CIFAR10 ($L=10$) (e) CIFAR10 ($L=100$) (f) CIFAR10 ($L=1000$)

Figure 6: Visualized experimental results of the ASWD on CelebA and CIFAR10 dataset with 10, 100, 1000 projections. The first row shows randomly selected samples of generated CelebA images, the second row shows randomly selected samples of generated CIFAR10 images.

H.2 Experiment results on MNIST dataset

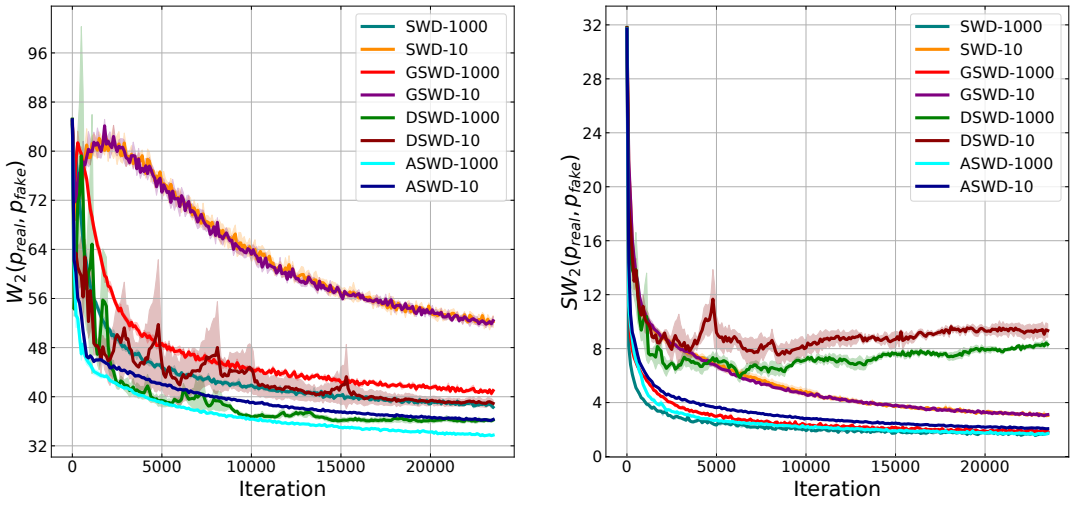
In the generative modelling experiment on the MNIST dataset, we train a generator by minimizing different slice-based Wasserstein metrics, including the ASWD, the DSWD, the GSWD (circular), and the SWD. Denote by G_Φ the generator, the training objective of the experiment can be formulated as [Bernton et al., 2019]:

$$\min_{\Phi} \mathbb{E}_{x\sim p_r, z\sim p_z}[\text{SWD}(x, G_\Phi(z))], \tag{46}$$

where p_z and p_r are the prior of latent variable z and the real data distribution, respectively. In other words, the SWD, or other slice-based Wasserstein metrics, can be considered as a discriminator in this framework. By replacing the SWD with the ASWD, the DSWD, and the GSWD, we compare the performance of learned generative models trained with different metrics. In this experiment, different methods are compared using different number of projections $L=\{10,1000\}$. The 2-Wasserstein distance and the SWD between generated images and real images are used as metrics for evaluating performances of different generative models. The experiment results are presented in Figure 7.

It can be observed from Figure 7 that the ASWD outperforms all the other methods regarding both the 2-Wasserstein distance and the SWD between generated and real images. In particular, the generative model trained with the ASWD produces smaller 2-Wasserstein distances within less iteration, which implies the generated images are of higher quality and the ASWD leads to higher convergence rates of generative models. In addition, the ASWD shows that it is able to generate higher quality images than the SWD and the GSWD with 1000 projections using only as less as 10 projections. In other words, the ASWD has higher projection efficiency than the other slice-based Wasserstein metrics. The ASWD also has the smallest SWD distance as shown in Figure 7. Although the SWD converges slightly faster than the ASWD in terms of the SWD between fake and real images, this is due to the training objective and the evaluated metric are the same for the SWD.

We also compare the execution time per mini-batch of different methods in Figure 8. We test the SWD by varying the number of projections in the set {10, 1000, 10000} and all the other methods in the set {10, 1000}. We found that although the SWD requires much fewer computational time than the DSWD and the ASWD, the quality of generated data is poor even when the number of projections L increases to 10000. The GSWD is also



(a) (b)

Figure 7: Visualized experimental results of different slice-based Wasserstein metrics on the MNIST dataset with 10, 1000 projections. (a) Comparison between the SWD, the GSWD, the DSWD, and the ASWD using the 2-Wasserstein distance between fake and real images as the evaluation metric. (b) Comparison between the SWD, the GSWD, the DSWD, and the ASWD using the SWD between fake and real images as the evaluation metric.

computationally efficient when using a 10 projections, but it requires the highest execution time and generates the highest 2-Wasserstein distance among all compared methods when the number of projections increases to 1000. The huge difference in the execution time of the GSWD with 10 and 1000 projections is due to the GSWD needs to calculate distance matrices of shape $N \times L$, where N and L are the number of samples and projections respectively, which is more computationally expensive than calculating inner products when the number of projections L increases. The DSWD requires a similar computational time as the ASWD in this example, while the ASWD generates higher quality images in terms of 2-Wasserstein distances. Randomly selected images generated by different slice-based Wasserstein metrics are presented in Figure 9.

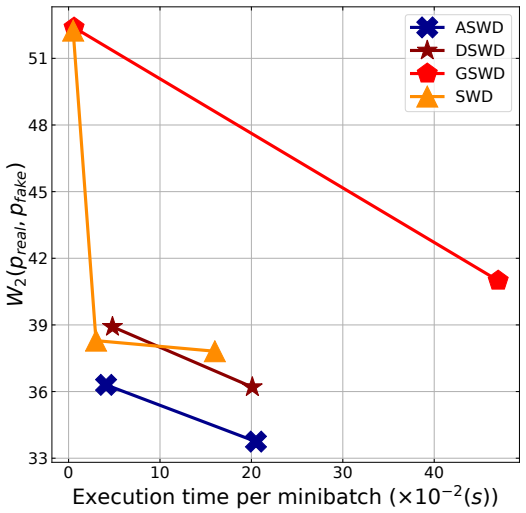


Figure 8: The execution time and the 2-Wasserstein distance between fake and real images of different methods. Each dot of the curve of SWD corresponds to the performance of the SWD with the number of projections $L = \{10, 1000, 10000\}$, in sequence. Each dot of the other curves correspond to the performance of the other methods with the number of projections $L = \{10, 1000\}$, in sequence.

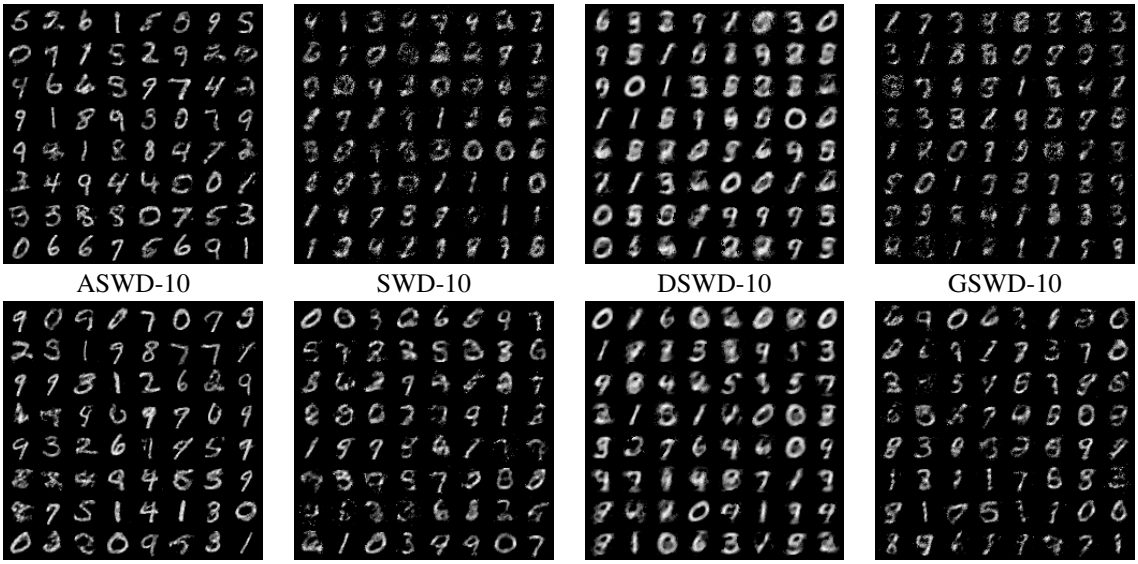


Figure 9: Randomly selected samples generated by different metrics, 10 and 1000 refer to the number of projections.

Appendix I Sliced Wasserstein autoencoders

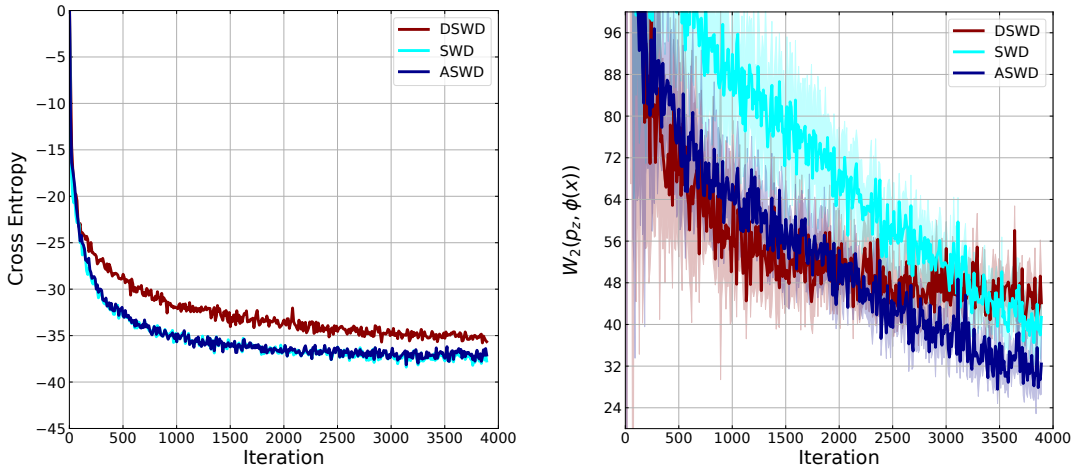
We train an autoencoder using the framework proposed in [Kolouri et al., 2019b], where an encoder and a decoder are jointly trained by minimizing the following objective:

$$\min_{\phi,\psi} \text{BCE}(\psi(\phi(x)),x)+\text{SWD}(p_z,\phi(x))+\text{SWD}(\psi(\phi(x)),x), \tag{47}$$

where ϕ is the encoder, ψ is the decoder, p_z is the prior distribution of latent variable, and $\text{BCE}(\cdot,\cdot)$ is the binary cross entropy loss between reconstructed images and real images. We train this model using different slice-based Wasserstein metrics, including the ASWD, the DSWD, and SWD. Here we use the ring distribution as the prior distribution as shown in Figure 11. We report the binary cross entropy loss during test time and the 2-Wasserstein distance between prior and the encoded latent variable $\phi(x)$ in Figure 10.

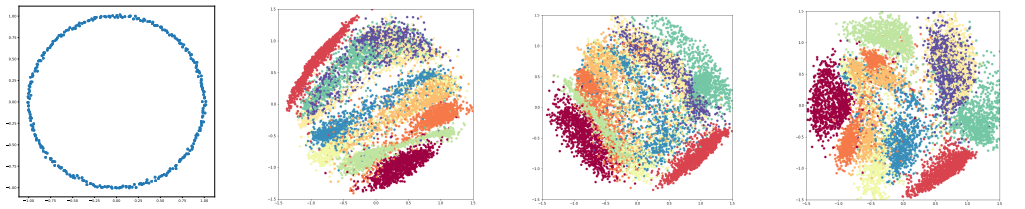
It can be observed from Figure 10 that the model trained with the ASWD converges faster to the smaller binary cross entropy loss than the model trained with the SWD, and has similar convergence behavior as the DSWD. In addition, the ASWD also leads to better coverage of the prior distribution as it can be observed from the second column of Figure 10 that the ASWD has the smallest 2-Wasserstein distance between the encoded latent variable distribution and the prior distribution. The latent spaces generated by different slice-based Wasserstein metrics are presented in Figure 11.

Some MNIST images randomly generated by SWAEs trained with different metrics are given in Figure 12.



(a) (b)

Figure 10: Convergence behavior of SWAEs trained with different slice-based Wasserstein metrics. (a) The binary cross entropy loss between the reconstruction and real data. (b) The 2-Wasserstein distance between the prior distribution p_z and the distribution of encoded feature $\phi(x)$.



(a) Prior distribution (b) ASWD latent space (c) SWD latent space (d) DSWD latent space

Figure 11: Comparisons between the encoded latent space generated by different slice-based Wasserstein metrics.

(a) ASWD samples (b) DSWD samples (c) SWD samples

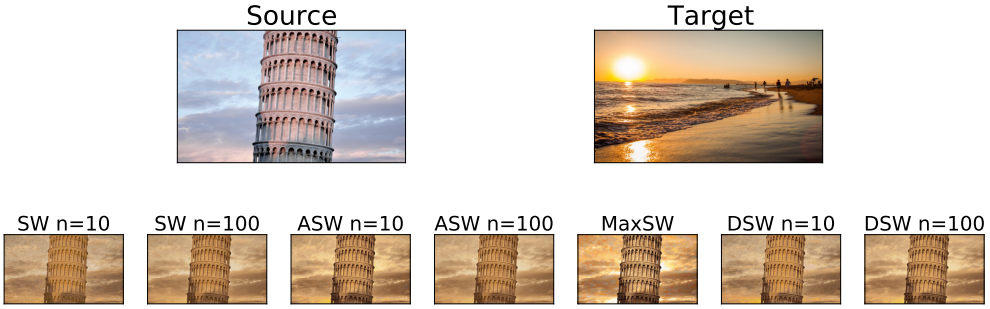
Figure 12: MNIST images randomly generated by SWAEs trained with different metrics.

Appendix J Color transferring

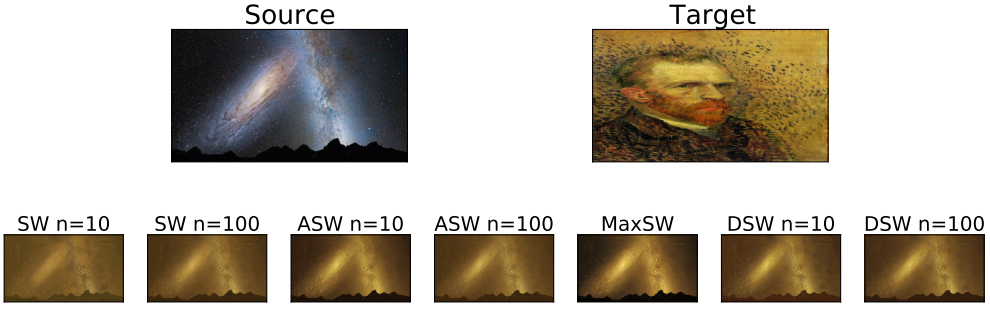
Color transferring can be formulated as an optimal transport problem [Bonneel et al., 2015; Radon, 1917]. In this task, the color palette of a source image is transferred to that of a target image, while keeping the content of source image unchanged. To achieve this, the optimal transport can be used to find the alignment between image pixels by calculating the optimal mapping of color palettes. In this experiment, instead of solving the optimal mapping in the original space, we first project the distribution onto one-dimensional spaces and average the alignment between one-dimensional samples as an approximation of the optimal mapping in the original space. After obtaining the approximation, we replace pixels of the source image with the averaged corresponding pixels in the target image. To reduce the computational cost, we utilize the approach proposed in [Muzellec and Cuturi, 2019], where the K-means algorithm is used to cluster the pixels of both source and target images, and then we implement color transfer for the quantized images whose pixels are consist of the centers of 3000 clusters rather than the original source and target images.

We present the results of color transferring in Figure 13. It can be observed that the ASWD and the DSWD produce sharper images than the SWD, we conjecture that is because the ASWD and the DSWD can generate better alignment of pixels. The Max-SWD has the highest contrast among all methods, but this is due to it only uses a single projection to obtain the transport mapping, thus there is no need to average different pixels from the target image. A disadvantage of the Max-SWD is that the transferred images generated by Max-SWD is not smooth enough and do not look realistic. The ASWD can generate smooth and realistic images than the SWD and the Max-SWD, even when the number of projections is as small as 10.

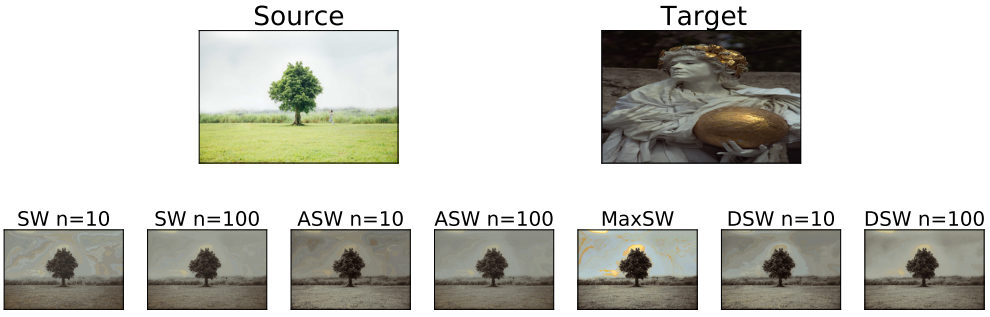
(a)



(b)



(c)



(d)

Figure 13: Top rows are source images and target images, lower rows show transferred images obtained by using different methods with different number of projections. Source and target images are from [Bonneel et al., 2015] and https://github.com/chia56028/Color-Transfer-between-Imageshttps://github.com/chia56028/Color-Transfer-between-Images.

Appendix K Sliced Wasserstein barycenter

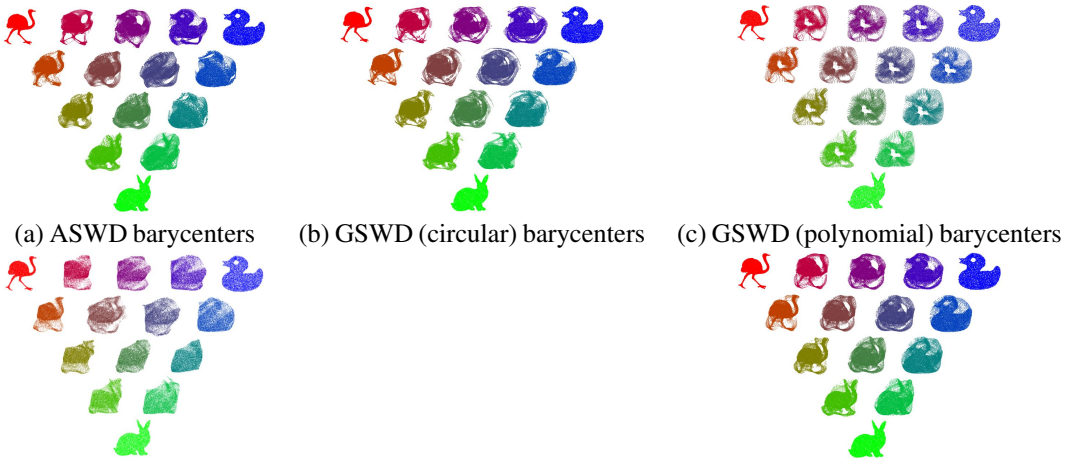
Sliced Wasserstein distances can also be applied in the barycenter calculation and shape interpolation [Bonneel et al., 2015]. Here we compare barycenters produced by different slice-based Wasserstein metrics, including the GSWD (circular and polynomial), the ASWD, the SWD, and the DSWD. Specifically, we compute barycenters of different shapes consisting of point clouds, as shown in Figure 14. Each object in Figure 14 corresponds to a specific barycenter with different weights.

Formally, a sliced-Wasserstein barycenter of objects $\mu = \{\mu_1, \mu_2, \cdots, \mu_N \in P_k(\mathbb{R}^d)\}$ assigned with weights $w = [w_1, w_2, \cdots, w_N \in \mathbb{R}]$ is defined as:

$$\text{Bar}(\mu, w) = \underset{\mu \in P_k(\mathbb{R}^d)}{\text{argmin}} \sum_{i=1}^{N} w_i \text{SWD}(\mu, \mu_i). \tag{48}$$

In this experiment, we set $N = 3$ and compute barycenters corresponding to different weights. The results are given in Figure 14.

From Figure 14, it can be observed that the ASWD produces similar barycenters as that of the SWD, which are sharper than the DSWD, and more meaningful than the GSWD (polynomial). The flexibility of the injective neural networks $g(\cdot)$ and its optimization in the ASWD can be potentially combined with specific requirements in particular tasks to generate calibrated barycenters - we leave this as a future research direction.



(a) ASWD barycenters (b) GSWD (circular) barycenters (c) GSWD (polynomial) barycenters

(d) DSWD barycenters (e) SWD barycenters

Figure 14: Sliced Wasserstein barycenters generated by the ASWD, the GSWD, the DSWD, and the SWD.

Appendix L Societal impacts

Research on comparing samples drawn from two probability distributions is inherently theoretical; it is also a fundamental topic in statistics and machine learning with a broad spectrum of downstream applications. In particular, our work has its root in the optimal transport theory and can be incorporated in a wide range of applications including computer vision (image retrieval and generation), natural language processing (alignment of word embedding for machine translation) and economics (resource allocation). Hence, this work developed a foundational tool with long-term societal and economic impact; the exact impact will be determined by the particular downstream applications. For example, the proposed distance metric can be used in machine translation to foster greater cross-cultural communication or employed in generative models to create new images or sound as a creativity tool; but it may also be used to produce 'fake' images for potentially harmful purposes. Those specific applications and their risk mitigation strategies are active research areas and out of the scope of this work, since this paper focuses on the theoretical and algorithmic development of a computational statistics tool.