Supplement to “Energy-Based Sliced Wasserstein Distance"

We first provide skipped proofs in the main text in Appendix A. We then provide additional materials including additional background, detailed algorithms, and discussion in Appendix B. Additional experimental results in point-cloud gradient flows, color transfer, and deep point-cloud reconstruction in Appendix C. Finally, we report the computational infrastructure in Appendix D.

A Proofs

A.1 Proof of Theorem 1

Non-negativity and Symmetry. the non-negativity and symmetry properties of the EBSW follow directly by the non-negativity and symmetry of the Wasserstein distance since it is an expectation of the one-dimensional Wasserstein distance.

Identity. We need to show that $\text{EBSW}_p(\mu,\nu;f)=0$ if and only if $\mu=\nu$. First, from the definition of EBSW, we obtain directly $\mu=\nu$ implies $\text{EBSW}_p(\mu,\nu;f)=0$. For the reverse direction, we use the same proof technique in [4]. If $\text{EBSW}_p(\mu,\nu;f)=0$, we have $\int_{\mathbb{S}^{d-1}} \text{W}_p\left(\theta\sharp\mu,\theta\sharp\nu\right) \mathrm{d}\sigma_{\mu,\nu}(\theta;f)=0$. Hence, we have $W_p(\theta\sharp\mu,\theta\sharp\nu)=0$ for $\sigma_{\mu,\nu}(\theta;f)$-almost surely $\theta\in\mathbb{S}^{d-1}$. Since $\sigma_{\mu,\nu}(\theta;f)$ is continuous, we have $W_p(\theta\sharp\mu,\theta\sharp\nu)=0$ for all $\theta\in\mathbb{S}^{d-1}$. From the identity property of the Wasserstein distance, we obtain $\theta\sharp\mu=\theta\sharp\nu$ for $\sigma_{\mu,\nu}(\theta;f)$-a.e $\theta\in\mathbb{S}^{d-1}$. Therefore, for any $t\in\mathbb{R}$ and $\theta\in\mathbb{S}^{d-1}$, we have:

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

where $\mathcal{F}[\gamma](w)=\int_{\mathbb{R}^{d'}} e^{-i\langle w,x\rangle}d\gamma(x)$ denotes the Fourier transform of $\gamma\in\mathcal{P}(\mathbb{R}^{d'})$. By the injectivity of the Fourier transform, we obtain $\mu=\nu$ which concludes the proof.

A.2 Proof of Proposition 1

(a) We first provide the proof for the inequality $\text{SW}_p(\mu,\nu)\leq\text{EBSW}_p(\mu,\nu;f)$. It is equivalent to prove that

$$\mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[\text{W}_p^p(\theta\sharp\mu,\theta\sharp\nu)\right]\leq\mathbb{E}_{\theta\sim\sigma_{\mu,\nu}(\theta;f)}\left[\text{W}_p^p(\theta\sharp\mu,\theta\sharp\nu)\right].$$

From the law of large number, it is sufficient to demonstrate that

$$\frac{1}{L}\sum_{i=1}^L W_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu)\leq\sum_{i=1}^L\frac{W_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu)f(W_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu))}{\sum_{i=1}^L f(W_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu))}, \tag{4}$$

for any $L\geq1$ and $\theta_1,\ldots,\theta_L\overset{\text{i.i.d.}}{\sim}\mathcal{U}(\mathbb{S}^{d-1})$. To ease the presentation, we denote $a_i=W_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu)$ and $b_i=f(W_p^p(\theta_i\sharp\mu,\theta_i\sharp\nu))$ for all $1\leq i\leq L$. The inequality (4) becomes:

$$\frac{1}{L}(\sum_{i=1}^L a_i)(\sum_{i=1}^L b_i)\leq\sum_{i=1}^L a_ib_i. \tag{5}$$

We prove the inequality (5) via an induction argument. It is clear that this inequality holds when $L=1$. We assume that this inequality holds for any L. We now verify that the inequality (5) also holds for $L+1$. Without loss of generality, we assume that $a_1\leq a_2\leq\ldots\leq a_L\leq a_{L+1}$. Since the function f is an increasing function, it indicates that $b_1\leq b_2\leq\ldots\leq b_L\leq b_{L+1}$. Applying the induction hypothesis for $a_1,\ldots,a_L$ and $b_1,\ldots,b_L$, we find that

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

This inequality leads to

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

Therefore, to obtain the conclusion of the hypothesis for $L+1$, it is sufficient to demonstrate that

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

which is equivalent to show that

$$(\sum_{i=1}^{L} a_i) b_{L+1} + (\sum_{i=1}^{L} b_i) a_{L+1} \leq \sum_{i=1}^{L} a_i b_i + L a_{L+1} b_{L+1}. \tag{6}$$

Since $a_{L+1} \geq a_i$ and $b_{L+1} \geq b_i$ for all $1 \leq i \leq L$, we have $(a_{L+1} - a_i)(b_{L+1} - b_i) \geq 0$, which is equivalent to $a_{L+1} b_{L+1} + a_i b_i \geq a_{L+1} b_i + b_{L+1} a_i$ for all $1 \leq i \leq L$. By taking the sum of these inequalities over i from 1 to L, we obtain the conclusion of inequality (6). Therefore, we obtain the conclusion of the induction argument for $L+1$, which indicates that inequality (5) holds for all L. As a consequence, we obtain the inequality $\text{SW}_p(\mu, \nu) \leq \text{EBSW}_p(\mu, \nu; f)$.

(b) We recall the definition of the Max-SW:

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

Since $\mathbb{S}^{d-1}$ is compact and the function $\theta \to W_p(\theta \sharp \mu, \theta \sharp \nu)$ is continuous, we have $\theta^\star = \text{argmax}_{\theta \in \mathbb{S}^{d-1}} W_p(\theta \sharp \mu, \theta \sharp \nu)$. From Definition 2, for any $p \geq 1$, dimension $d \geq 1$, energy-function f, and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ we have:

$$\begin{aligned}
\text{EBSW}_p(\mu, \nu) &= \left(\mathbb{E}_{\theta \sim \sigma_{\mu,\nu}(\theta; f))}\left[W_p^p\left(\theta \sharp \mu, \theta \sharp \nu\right)\right]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}_{\theta \sim \sigma_{\mu,\nu}(\theta; f))}\left[W_p^p\left(\theta^\star \sharp \mu, \theta^\star \sharp \nu\right)\right]\right)^{\frac{1}{p}} = W_p^p\left(\theta^* \sharp \mu, \theta^* \sharp \nu\right) = \text{Max-SW}_p(\mu, \nu).
\end{aligned}$$

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

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

after taking the p-rooth, we completes the proof.

A.3 Proof of Theorem 2

We aim to show that for any sequence of probability measures $(\mu_k)_{k \in \mathbb{N}}$ and μ in $\mathcal{P}_p(\mathbb{R}^d)$, $\lim_{k \to +\infty} \text{EBSW}_p(\mu_k, \mu; f) = 0$ if and only if for any continuous and bounded function $f : \mathbb{R}^d \to \mathbb{R}$, $\lim_{k \to +\infty} \int f \, \mathrm{d}\mu_k = \int f \, \mathrm{d}\mu$. We follow the proof techniques in [26]. We first state the following lemma.

Lemma 1. *For any $p \geq 1$, energy function f, and dimension $d \geq 1$, a sequence of probability measures $(\mu_k)_{k \in \mathbb{N}}$ satisfies $\lim_{k \to +\infty}$ EBSW$_p(\mu_k, \mu; f) = 0$ with μ in $\mathcal{P}_p(\mathbb{R}^d)$, there exists an increasing function $\phi : \mathbb{N} \to \mathbb{N}$ such that the subsequence $\left(\mu_{\phi(k)}\right)_{k \in \mathbb{N}}$ converges weakly to μ.*

Proof. Since $\lim_{k\to+\infty}\text{EBSW}_p(\mu_k,\mu;f)=0$, we have $\lim_{k\to\infty}\int_{\mathbb{S}^{d-1}}\text{W}_p\left(\theta\sharp\mu_k,\theta\sharp\mu\right)\mathrm{d}\sigma_{\mu,\nu}(\theta;f)=0$. From Theorem 2.2.5 in [1], there exists an increasing function $\phi:\mathbb{N}\to\mathbb{N}$ such that $\lim_{k\to\infty}\text{W}_p(\theta\sharp\mu_{\phi(k)},\theta\sharp\nu)=0$ for $\sigma_{\mu,\nu}(\theta;f)$-a.e $\theta\in\mathbb{S}^{d-1}$. From [39], the Wasserstein distance of order p implies weak convergence in $\mathcal{P}_p(\mathbb{R}^d)$, hence,$\left(\theta\sharp\mu_{\phi(k)}\right)_{k\in\mathbb{N}}$ converges weakly to $\theta\sharp\mu$ for $\sigma_{\mu,\nu}(\theta;f)$-a.e $\theta\in\mathbb{S}^{d-1}$.

Let $\Phi_\mu=\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\langle v,w\rangle}\mathrm{d}\mu(w)$ be the characteristic function of $\mu\in\mathcal{P}_p(\mathbb{R}^d)$, the weak convergence implies the convergence of characteristic function (Theorem 4.3 [17]): $\lim_{k\to\infty}\Phi_{\theta\sharp\mu_{\phi(k)}}(s)=\Phi_{\theta\sharp\mu}(s),\quad\forall s\in\mathbb{R}$, for $\sigma_{\mu,\nu}(\theta;f)$-a.e $\theta\in\mathbb{S}^{d-1}$. Therefore, $\lim_{k\to\infty}\Phi_{\mu_{\phi(k)}}(z)=\Phi_\mu(z)$, for almost most every $z\in\mathbb{R}^d$.

We denote $f_\gamma(x)=f*g_\gamma(x)=\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}f(x-z)\exp\left(-\|z\|^2/\left(2\gamma^2\right)\right)\mathrm{d}z$ for any $\gamma>0$ and a continuous function $f:\mathbb{R}^d\to\mathbb{R}$ with compact support, and g_γ is the density function of $\mathcal{N}(0,\gamma I_d)$. Now, we have:

$$\begin{aligned}
\int_{\mathbb{R}^d}f_\gamma(z)\mathrm{d}\mu_{\phi(k)}(z)&=\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}f(w)g_\gamma(z-w)\mathrm{d}w\;\mathrm{d}\mu_{\phi(k)}(z)\\
&=\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}f(w)\left(2\pi\gamma^2\right)^{-d/2}\exp(-||z-w||^2/(2\gamma^2))\mathrm{d}w\;\mathrm{d}\mu_{\phi(k)}(z)\\
&=\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}f(w)\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\langle z-w,x\rangle}g_{1/\gamma}(x)\mathrm{d}x\;\mathrm{d}w\;\mathrm{d}\mu_{\phi(k)}(z)\\
&=\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}f(w)\int_{\mathbb{R}^d}e^{-\mathrm{i}\langle w,x\rangle}e^{\mathrm{i}\langle z,x\rangle}g_{1/\gamma}(x)\mathrm{d}x\;\mathrm{d}w\;\mathrm{d}\mu_{\phi(k)}(z)\\
&=\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}f(w)e^{-\mathrm{i}\langle w,x\rangle}g_{1/\gamma}(x)\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\langle z,x\rangle}\;\mathrm{d}\mu_{\phi(k)}(z)\mathrm{d}x\;\mathrm{d}w\\
&=\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}f(w)\mathrm{e}^{-\mathrm{i}\langle w,x\rangle}g_{1/\gamma}(x)\Phi_{\mu_{\phi(k)}}(x)\mathrm{d}x\;\mathrm{d}w\\
&=\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\mathcal{F}[f](x)g_{1/\gamma}(x)\Phi_{\mu_{\phi(k)}}(x)\mathrm{d}x,
\end{aligned}$$

where the third equality is because $\int_{\mathbb{R}^d}e^{\mathrm{i}\langle z-w,x\rangle}g_{1/\gamma}(x)\mathrm{d}x=\exp(-||z-w||^2/(2\gamma^2))$, and $\mathcal{F}[f](w)=\int_{\mathbb{R}^{d'}}f(x)e^{-i\langle w,x\rangle}dx$ denotes the Fourier transform of the bounded function f. Similarly, we have:

$$\begin{aligned}
\int_{\mathbb{R}^d}f_\gamma(z)\mathrm{d}\mu(z)&=\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}f(w)g_\gamma(z-w)\mathrm{d}w\;\mathrm{d}\mu(z)\\
&=\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}f(w)\left(2\pi\gamma^2\right)^{-d/2}\exp(-||z-w||^2/(2\gamma^2))\mathrm{d}w\;\mathrm{d}\mu(z)\\
&=\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}f(w)\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\langle z-w,x\rangle}g_{1/\gamma}(x)\mathrm{d}x\;\mathrm{d}w\;\mathrm{d}\mu(z)\\
&=\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}f(w)\int_{\mathbb{R}^d}e^{-\mathrm{i}\langle w,x\rangle}e^{\mathrm{i}\langle z,x\rangle}g_{1/\gamma}(x)\mathrm{d}x\;\mathrm{d}w\;\mathrm{d}\mu(z)\\
&=\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}f(w)e^{-\mathrm{i}\langle w,x\rangle}g_{1/\gamma}(x)\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\langle z,x\rangle}\;\mathrm{d}\mu(z)\mathrm{d}x\;\mathrm{d}w\\
&=\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}f(w)\mathrm{e}^{-\mathrm{i}\langle w,x\rangle}g_{1/\gamma}(x)\Phi_\mu(x)\mathrm{d}x\;\mathrm{d}w\\
&=\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\mathcal{F}[f](x)g_{1/\gamma}(x)\Phi_\mu(x)\mathrm{d}x.
\end{aligned}$$

We know that $\mathcal{F}[f]$ exists and is bounded by $\int_{\mathbb{R}^d}|f(w)|\mathrm{d}w<+\infty$ since f has compact support. Hence, for any $x\in\mathbb{R}^d$ and $k\in\mathbb{R}$, we have $\left|\mathcal{F}[f](x)g_{1/\gamma}(x)\Phi_{\mu_{\phi(k)}}(x)\right|\leq g_{1/\gamma}(x)\int_{\mathbb{R}^d}|f(w)|\mathrm{d}w$ and $\left|\mathcal{F}[f](x)g_{1/\gamma}(x)\Phi_\mu(x)\right|\leq g_{1/\gamma}(x)\int_{\mathbb{R}^d}|f(w)|\mathrm{d}w$. Using the proved result of $\lim_{k\to\infty}\Phi_{\mu_{\phi(k)}}(z)=\Phi_\mu(z)$ and Lebesgue's Dominated Convergence Therefore, we obtain

$$\begin{aligned}\lim_{k\to\infty}\int_{\mathbb{R}^d} f_\gamma(z)\mathrm{d}\mu_{\phi(k)}(z) &= \lim_{k\to\infty}\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\mathcal{F}[f](x)g_{1/\gamma}(x)\Phi_{\mu_{\phi(k)}}(x)\mathrm{d}x\\ &= \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\mathcal{F}[f](x)g_{1/\gamma}(x)\Phi_{\mu_{\phi(k)}}(x)\mathrm{d}x\\ &= \int_{\mathbb{R}^d} f_\gamma(z)\mathrm{d}\mu(z).\end{aligned}$$

Moreover, we have:

$$\begin{aligned}&\lim_{\gamma\to 0}\limsup_{k\to+\infty}\left|\int_{\mathbb{R}^d} f(z)\mathrm{d}\mu_{\phi(k)}(z)-\int_{\mathbb{R}^d} f(z)\mathrm{d}\mu(z)\right|\\ &\leq \lim_{\gamma\to 0}\limsup_{k\to+\infty}\left[2\sup_{z\in\mathbb{R}^d}|f(z)-f_\gamma(z)|+\left|\int_{\mathbb{R}^d} f_\gamma(z)\mathrm{d}\mu_{\phi(k)}(z)-\int_{\mathbb{R}^d} f_\gamma(z)\mathrm{d}\mu(z)\right|\right]\\ &= \lim_{\gamma\to 0}2\sup_{z\in\mathbb{R}^d}|f(z)-f_\gamma(z)|=0,\end{aligned}$$

which implies $\left(\mu_{\phi(k)}\right)_{k\in\mathbb{N}}$ converges weakly to μ. □

Continuing the proof of Theorem 2, we show that $\lim_{k\to\infty}\text{EBSW}_p(\mu_k,\mu;f)=0$ implies $(\mu_k)_{k\in\mathbb{N}}$ converges weakly to μ. Let $\left(\mu_{\phi(k)}\right)_{k\in\mathbb{N}}$ be a sequence such that $\lim_{k\to\infty}\text{EBSW}_p(\mu_k,\mu;f)=0$, we suppose $\left(\mu_{\phi(k)}\right)_{k\in\mathbb{N}}$ does not converge weakly to μ. So, let $\mathcal{D}_{\mathcal{P}}$ be the Lévy-Prokhorov metric, $\lim_{k\to\infty}\mathcal{D}_{\mathcal{P}(\mu_k,\mu)}\neq 0$ that implies there exists $\varepsilon>0$ and a subsequence $\left(\mu_{\psi(k)}\right)_{k\in\mathbb{N}}$ with an increasing function $\psi:\mathbb{N}\to\mathbb{N}$ such that for any $k\in\mathbb{N}$: $\mathcal{D}_{\mathcal{P}}(\mu_{\psi(k)},\mu)\geq\varepsilon$. Using the Holder inequality with $\mu,\nu\in\mathbb{P}_p(\mathbb{R}^d)$, we have:

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

Therefore, $\lim_{k\to\infty}\text{EBSW}_1(\mu_{\psi(k)},\mu;f)=0$ which implies that there exists s a subsequence $\left(\mu_{\phi(\psi(k))}\right)_{k\in\mathbb{N}}$ with an increasing function $\phi:\mathbb{N}\to\mathbb{N}$ such that $\left(\mu_{\phi(\psi(k))}\right)_{k\in\mathbb{N}}$ converges weakly to μ by Lemma 1. Therefore a contradiction appears, namely, $\lim_{k\to\infty}d_{\mathcal{P}}\left(\mu_{\phi(\psi(k))},\mu\right)=0$. Therefore, $\lim_{k\to\infty}\text{EBSW}_p(\mu_k,\mu;f)=0$, $(\mu_k)_{k\in\mathbb{N}}$ converges weakly to μ.

We have $(\theta\sharp\mu_k)_{k\in\mathbb{N}}$ converges weakly to $\theta\sharp\mu$ for any $\theta\in\mathbb{S}^{d-1}$ by the continuous mapping theorem. From [39], the weak convergence implies the convergence under the Wasserstein distance. So, we have $\lim_{k\to\infty}W_p(\theta\sharp\mu_k,\mu)=0$. Moreover, using the fact that the Wasserstein distance is also bounded, hence, the bounded convergence theorem implies:

$$\begin{aligned}\lim_{k\to\infty}\text{EBSW}_p^p(\mu_k,\mu;f) &= \mathbb{E}_{\theta\sim\sigma_{\mu,\nu}(\theta;f)}\left[W_p^p\left(\theta\sharp\mu_k,\theta\sharp\mu\right)\right]\\ &= \mathbb{E}_{\theta\sim\sigma_{\mu,\nu}(\theta;f)}\left[0\right]=0.\end{aligned}$$

Again, usingthe continuous mapping theorem with function $x\to x^{1/p}$, we have $\lim_{k\to\infty}\text{EBSW}_p(\mu_k,\mu;f)\to 0$. We conclude the proof.

A.4 Proof of Proposition 2

We first show that the following inequality holds

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

where $C>0$ is some universal constant and the outer expectation is taken with respect to the random variables $X_1,\ldots,X_n$. We now follow the proof technique from in [28]. Let $F_{n,\theta}^{-1}$ and F_θ^{-1} be the

Algorithm 1 Computational algorithm of the SW distance

```
Input: Probability measures $\mu$ and $\nu$, $p \geq 1$, and the number of projections $L$.
for $l = 1$ to $L$ do
    Sample $\theta_l \sim \mathcal{U}(\mathbb{S}^{d-1})$
    Compute $v_l = \text{W}_p(\theta_l \sharp \mu, \theta_l \sharp \nu)$
end for
Compute $\widehat{SW}_p(\mu, \nu; L) = \left(\frac{1}{L}\sum_{l=1}^{L} v_l\right)^{\frac{1}{p}}$
Return: $\widehat{SW}_p(\mu, \nu; L)$
```

inverse cumulative distributions of two push-forward measures $\theta \sharp \mu_n$ and $\theta \sharp \mu$. Using the closed-form expression of the Wasserstein distance in one dimension, we obtain to the following equations and inequalities:

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

where $\mathcal{X} \subset \mathbb{R}^d$ is the compact set of the probability measure μ. We can check that

$$\max_{\theta \in \mathbb{R}^d : \|\theta\| \leq 1} |F_{n,\theta}(x) - F_\theta(x)| = \sup_{A \in \mathcal{B}} |\mu_n(A) - \mu(A)|,$$

where $\mathcal{B}$ is the set of half-spaces $\{z \in \mathbb{R}^d : \theta^\top z \leq x\}$ for all $\theta \in \mathbb{R}^d$ such that $\|\theta\| \leq 1$. We know that the Vapnik-Chervonenkis (VC) dimension of $\mathcal{B}$ is at most $d + 1$ [40]. Therefore, using the VC inequality, we obtain:

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{32}{n}[(d+1)\log(n+1) + \log(8/\delta)]},$$

with probability at least $1 - \delta$. Therefore, we obtain that

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

where $C > 0$ is some universal constant. Moreover, we have $\mathbb{E}[\text{EBSW}_p(\mu_n, \mu; f)] \leq \mathbb{E}\left[\text{Max-SW}_p(\mu_n, \mu)\right]$ from Proposition 1. Therefore, As a consequence, we obtain:

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

which completes the proof.

B Additional Materials

B.1 Additional Background

Sliced Wasserstein. When two probability measures are empirical probability measures on n supports: $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, the SW distance can be computed by sorting projected supports. In particular, we have $\theta \sharp \mu = \frac{1}{n}\sum_{i=1}^n \delta_{\theta^\top x_i}$, $\theta \sharp \nu = \frac{1}{n}\sum_{i=1}^n \delta_{\theta^\top y_i}$, and $\text{W}_p^p(\theta \sharp \mu, \theta \sharp \nu) = \frac{1}{n}\sum_{i=1}^n (\theta^\top x_{(i)} - \theta^\top y_{(i)})^p$ where $\theta^\top x_{(i)}$ is the ordered projected supports. We provide the pseudo-code for computing the SW in Algorithm 1.

Max sliced Wasserstein. The Max-SW is often computed by the projected gradient ascent. The sub-gradient is used when the one-dimensional optimal matching is not unique e.g., in discrete cases. We provide the projected (sub)-gradient ascent algorithm for computing the Max-SW in Algorithm 2. Compared to the SW, the Max-SW needs two hyperparameters which are the number of iterations T and the step size η. Moreover, the empirical estimation of the Max-SW might not converge to the Max-SW when $T \to \infty$.

Algorithm 2 Computational algorithm of the Max-SW distance

```
Input: Probability measures $\mu$ and $\nu$, $p \geq 1$, the number of iterations $T$, and the step size $\eta$.
Sample $\hat{\theta}_0 \sim \mathcal{U}(\mathbb{S}^{d-1})$
for $t = 1$ to $T$ do
    Compute $\hat{\theta}_t = \hat{\theta}_{t-1} + \eta \nabla_{\hat{\theta}_{t-1}} \mathrm{W}_p(\hat{\theta}_{t-1}\sharp\mu, \hat{\theta}_{t-1}\sharp\nu)$
    Compute $\hat{\theta}_t = \frac{\hat{\theta}_t}{||\hat{\theta}_t||_2}$
end for
Compute $\widehat{\text{Max-SW}}_p(\mu, \nu; T) = \mathrm{W}_p(\hat{\theta}_T\sharp\mu, \hat{\theta}_T\sharp\nu)$
Return: $\widehat{\text{Max-SW}}_p(\mu, \nu; T)$
```

Algorithm 3 Computational algorithm of the DSW distance

```
Input: Probability measures $\mu$ and $\nu$, $p \geq 1$, the number of projections $L$, the number of iterations $T$, and the step size $\eta$.
Initialize $\hat{\psi}_0$
for $t = 1$ to $T$ do
    $\nabla_\psi = 0$
    for $l = 1$ to $L$ do
        Sample $\theta_{l,\psi} \sim \sigma_{\hat{\psi}_{t-1}(\theta)}$ via reparameterization.
        Compute $\hat{\theta}_t = \frac{\hat{\theta}_t}{||\hat{\theta}_t||_2}$
    end for
    Compute $\hat{\psi}_t = \hat{\psi}_{t-1} + \eta \frac{1}{p} \left( \frac{1}{L} \sum_{l=1}^{L} \mathrm{W}_p^p(\theta_{l,\psi}\sharp\mu, \theta_{l,\psi}\sharp\nu) \right)^{\frac{1-p}{p}} \frac{1}{L} \sum_{l=1}^{l} \nabla_\psi \mathrm{W}_p^p(\theta_{l,\psi}\sharp\mu, \theta_{l,\psi}\sharp\nu))$
end for
for $l = 1$ to $L$ do
    Sample $\theta_l \sim \sigma_{\hat{\psi}_T(\theta)}$ via reparameterization.
end for
Compute $\widehat{\mathrm{DSW}}_p(\mu, \nu; T, L) = \left( \frac{1}{L} \sum_{l=1}^{L} \mathrm{W}_p^p(\theta_l\sharp\mu, \theta_l\sharp\nu) \right)^{\frac{1}{p}}$
Return: $\widehat{\mathrm{DSW}}_p(\mu, \nu; T, L)$
```

Distributional sliced Wasserstein. To solve the optimization of the DSW, we need to use the stochastic (sub)-gradient ascent algorithm. In particular, we first need to estimate the gradient $\nabla_\psi \left(\mathbb{E}_{\theta \sim \sigma_\psi(\theta)} \mathrm{W}_p^p(\theta\sharp\mu, \theta\sharp\nu) \right)^{\frac{1}{p}}$:

$$\nabla_\psi \left(\mathbb{E}_{\theta \sim \sigma_\psi(\theta)} \mathrm{W}_p^p(\theta\sharp\mu, \theta\sharp\nu) \right)^{\frac{1}{p}} = \frac{1}{p} \left(\mathbb{E}_{\theta \sim \sigma_\psi(\theta)} \mathrm{W}_p^p(\theta\sharp\mu, \theta\sharp\nu) \right)^{\frac{1-p}{p}} \nabla_\psi \mathbb{E}_{\theta \sim \sigma_\psi(\theta)} \mathrm{W}_p^p(\theta\sharp\mu, \theta\sharp\nu).$$

To estimate the gradient $\nabla_\psi \mathbb{E}_{\theta \sim \sigma_\psi(\theta)} \mathrm{W}_p^p(\theta\sharp\mu, \theta\sharp\nu)$, we need to use reparameterization trick for $\sigma_\psi(\theta)$ e.g., the vMF distribution. After using the reparameterization trick, we can approximate the gradient $\nabla_\psi \mathbb{E}_{\theta \sim \sigma_\psi(\theta)} \mathrm{W}_p^p(\theta\sharp\mu, \theta\sharp\nu) = \frac{1}{L} \sum_{l=1}^{l} \nabla_\psi \mathrm{W}_p^p(\theta_{l,\psi}\sharp\mu, \theta_{l,\psi}\sharp\nu)$ where $\theta_{1,\psi}, \ldots, \theta_{L,\psi}$ are i.i.d reparameterized samples from $\sigma_\psi(\theta)$. Similarly, we approximiate $\mathbb{E}_{\theta \sim \sigma_\psi(\theta)} \mathrm{W}_p^p(\theta\sharp\mu, \theta\sharp\nu) = \frac{1}{L} \sum_{l=1}^{L} \mathrm{W}_p^p(\theta_l\sharp\mu, \theta_l\sharp\nu)$. We refer to the details in the following papers [7, 30]. We review the algorithm for computing the DSW in Algorithm 3. Compared to the SW, the DSW needs three hyperparameters i.e., the number of projections L, the number of iterations T, and the step size η.

Minimum Distance Estimator and Gradient Estimation. In statistical inference, we are given the empirical samples $X_1, \ldots, X_n$ from the interested distribution ν. Since we do not know the form of ν, we might want to find an alternative representation. In particular, we want to find the best member μ_ϕ in a family of distribution parameterized by $\phi \in \Phi$. To do that, we want to minimize the distance between μ_ϕ and the empirical distribution $\nu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$. This framework is named the minimum distance estimator [41]:

$$\min_{\phi \in \Phi} \mathcal{D}(\mu_\phi, \nu_n),$$

where $\mathcal{D}$ is a discrepancy between two distributions. The gradient ascent algorithm is often used to solve the problem. To do so, we need to compute the gradient $\nabla_\phi \mathcal{D}(\mu_\phi, \nu_n)$. When using sliced

Algorithm 4 Computational algorithm of the IS-EBSW distance

```
Input: Probability measures $\mu$ and $\nu$, $p \geq 1$, the number of projections $L$, the energy function $f$.
for $l = 1$ to $L$ do
    Sample $\theta_l \sim \mathcal{U}(\mathbb{S}^{d-1})$
    Compute $v_l = \text{W}_p(\theta_l \sharp \mu, \theta_l \sharp \nu)$
    Compute $w_l = f(\text{W}_p(\theta_l \sharp \mu, \theta_l \sharp \nu))$
end for
Compute $\text{IS-}\widehat{\text{EBSW}}_p(\mu, \nu; L, f) = \left(\frac{1}{L}\sum_{l=1}^{L} v_l \frac{w_l}{\sum_{i=1}^{L} w_i}\right)^{\frac{1}{p}}$
Return: $\text{IS-}\widehat{\text{EBSW}}_p(\mu, \nu; L, f)$
```

Wasserstein distances, the gradient $\nabla_\phi \mathcal{D}(\mu_\phi, \nu_n)$ is often approximated by a stochastic gradient since the SW distances involve an intractable expectation. In previous SW variants, the expectation does not depend on ϕ, hence, we can use directly the Leibniz rule to exchange the gradient and the expectation, then perform the Monte Carlo approximation. In particular, we have $\nabla_\phi \mathbb{E}_{\theta \sim \sigma(\theta)}[\text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu)] = \mathbb{E}_{\theta \sim \sigma(\theta)}[\nabla_\phi \text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu)] \approx \frac{1}{L}\sum_{l=1}^{L} \nabla_\phi \text{W}_p^p(\theta_l\sharp\mu, \theta_l\sharp\nu)$ for $\theta_1, \ldots, \theta_L \overset{i.i.d}{\sim} \sigma(\theta)$.

B.2 Importance Sampling

Derivation. We first provide the derivation of the importance sampling estimation of EBSW. From the definition of the EBSW, we have:

$$\begin{aligned}
\text{EBSW}_p(\mu, \nu; f) &= \left(\mathbb{E}_{\theta \sim \sigma_{\mu,\nu}(\theta; f)}\left[\text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu)\right]\right)^{\frac{1}{p}} \\
&= \left(\frac{\int_{\mathbb{S}^{d-1}} \text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu) f(\text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu)) d\theta}{\int_{\mathbb{S}^{d-1}} f(\text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu)) d\theta}\right)^{\frac{1}{p}} \\
&= \left(\frac{\int_{\mathbb{S}^{d-1}} \text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu) \frac{f(\text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu))}{\sigma_0(\theta)} \sigma_0(\theta) d\theta}{\int_{\mathbb{S}^{d-1}} \frac{f(\text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu))}{\sigma_0(\theta)} \sigma_0(\theta) d\theta}\right)^{\frac{1}{p}} \\
&= \left(\frac{\mathbb{E}_{\theta \sim \sigma_0(\theta)}\left[\text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu) \frac{f(\text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu))}{\sigma_0(\theta)}\right]}{\mathbb{E}_{\theta \sim \sigma_0(\theta)}\left[\frac{f(\text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu))}{\sigma_0(\theta)}\right]}\right)^{\frac{1}{p}} \\
&= \left(\frac{\mathbb{E}_{\theta \sim \sigma_0(\theta)}\left[\text{W}_p^p(\theta\sharp\mu, \theta\sharp\nu) w_{\mu,\nu,\sigma_0}(\theta)\right]}{\mathbb{E}_{\theta \sim \sigma_0(\theta)}\left[w_{\mu,\nu,\sigma_0}(\theta)\right]}\right)^{\frac{1}{p}}.
\end{aligned}$$

Algorithms. We provide the algorithm for the IS estimation of the EBSW in Algorithm 4. Compared to the algorithm of the SW in Algorithm 1, the IS-EBSW can be obtained by only adding one or two lines of code in practice. Therefore, the computation of the IS-EBSW is as fast as the SW while being more meaningful.

Gradient Estimators. Let μ_ϕ be parameterized by ϕ, we derive now the gradient estimator $\nabla_\phi \text{EBSW}_p(\mu, \nu; f)$ through importance sampling. We have:

$$\begin{aligned}
\nabla_\phi \text{EBSW}_p(\mu_\phi, \nu; f) = \frac{1}{p}&\left(\frac{\mathbb{E}_{\theta \sim \sigma_0(\theta)}\left[\text{W}_p^p(\theta\sharp\mu_\phi, \theta\sharp\nu) w_{\mu_\phi,\nu,\sigma_0,f}(\theta)\right]}{\mathbb{E}_{\theta \sim \sigma_0(\theta)}\left[w_{\mu_\phi,\nu,\sigma_0,f}(\theta)\right]}\right)^{\frac{1-p}{p}} \\
&\nabla_\phi \frac{\mathbb{E}_{\theta \sim \sigma_0(\theta)}\left[\text{W}_p^p(\theta\sharp\mu_\phi, \theta\sharp\nu) w_{\mu_\phi,\nu,\sigma_0,f}(\theta)\right]}{\mathbb{E}_{\theta \sim \sigma_0(\theta)}\left[w_{\mu_\phi,\nu,\sigma_0,f}(\theta)\right]}.
\end{aligned}$$

We denote $A(\phi) = \mathbb{E}_{\theta \sim \sigma_0(\theta)}\left[\text{W}_p^p(\theta\sharp\mu_\phi, \theta\sharp\nu) w_{\mu_\phi,\nu,\sigma_0,f}(\theta)\right]$, $B(\phi) = \mathbb{E}_{\theta \sim \sigma_0(\theta)}\left[w_{\mu_\phi,\nu,\sigma_0,f}(\theta)\right]$, we have

$$\nabla_\phi \frac{A(\phi)}{B(\phi)} = \frac{B(\phi)\nabla_\phi A(\phi) - A(\phi)\nabla\phi B(\phi)}{B^2(\phi)}.$$

Using Monte Carlo samples $\theta_1,\ldots,\theta_L \sim \sigma_0(\theta)$ after using the Lebnitz rule to exchange the differentiation and the expectation, we obtain:

$$\left(\frac{\mathbb{E}_{\theta\sim\sigma_0(\theta)}\left[\mathrm{W}_p^p(\theta\sharp\mu_\phi,\theta\sharp\nu)w_{\mu_\phi,\nu,\sigma_0,f}(\theta)\right]}{\mathbb{E}_{\theta\sim\sigma_0(\theta)}\left[w_{\mu_\phi,\nu,\sigma_0,f}(\theta)\right]}\right)^{\frac{1-p}{p}} \approx \left(\frac{\frac{1}{L}\sum_{l=1}^L\left[\mathrm{W}_p^p(\theta_l\sharp\mu_\phi,\theta_l\sharp\nu)w_{\mu_\phi,\nu,\sigma_0,f}(\theta_l)\right]}{\frac{1}{L}\sum_{l=1}^L\left[w_{\mu_\phi,\nu,\sigma_0,f}(\theta_l)\right]}\right)^{\frac{1-p}{p}},$$

$$\nabla_\phi A(\phi) \approx \frac{1}{L}\sum_{l=1}^L \nabla_\phi\left(\mathrm{W}_p^p(\theta_l\sharp\mu_\phi,\theta_l\sharp\nu)w_{\mu_\phi,\nu,\sigma_0,f}(\theta)\right),$$

$$\nabla_\phi B(\phi) \approx \frac{1}{L}\sum_{l=1}^L \nabla_\phi w_{\mu_\phi,\nu,\sigma_0,f}(\theta),$$

which yields the gradient estimation. If we construct the slicing distribution by using a copy of μ_ϕ i.e., $\mu_{\phi'}$ with $\phi'=\phi$ in terms of value, the gradient estimator can be derived by:

$$\nabla_\phi \mathrm{EBSW}_p(\mu_\phi,\nu;f) = \frac{1}{p}\left(\frac{\mathbb{E}_{\theta\sim\sigma_0(\theta)}\left[\mathrm{W}_p^p(\theta\sharp\mu_\phi,\theta\sharp\nu)w_{\mu_{\phi'},\nu,\sigma_0,f}(\theta)\right]}{\mathbb{E}_{\theta\sim\sigma_0(\theta)}\left[w_{\mu_{\phi'},\nu,\sigma_0,f}(\theta)\right]}\right)^{\frac{1-p}{p}} \frac{\nabla_\phi\mathbb{E}_{\theta\sim\sigma_0(\theta)}\left[\mathrm{W}_p^p(\theta\sharp\mu_\phi,\theta\sharp\nu)w_{\mu_{\phi'},\nu,\sigma_0,f}(\theta)\right]}{\mathbb{E}_{\theta\sim\sigma_0(\theta)}\left[w_{\mu_{\phi'},\nu,\sigma_0,f}(\theta)\right]},$$

Using Monte Carlo samples $\theta_1,\ldots,\theta_L \sim \sigma_0(\theta)$ after using the Lebnitz rule to exchange the differentiation and the expectation, we obtain:

$$\left(\frac{\mathbb{E}_{\theta\sim\sigma_0(\theta)}\left[\mathrm{W}_p^p(\theta\sharp\mu_\phi,\theta\sharp\nu)w_{\mu_{\phi'},\nu,\sigma_0,f}(\theta)\right]}{\mathbb{E}_{\theta\sim\sigma_0(\theta)}\left[w_{\mu_{\phi'},\nu,\sigma_0,f}(\theta)\right]}\right)^{\frac{1-p}{p}} \approx \left(\frac{\frac{1}{L}\sum_{l=1}^L\left[\mathrm{W}_p^p(\theta_l\sharp\mu_\phi,\theta_l\sharp\nu)w_{\mu_{\phi'},\nu,\sigma_0,f}(\theta_l)\right]}{\frac{1}{L}\sum_{l=1}^L\left[w_{\mu_{\phi'},\nu,\sigma_0,f}(\theta_l)\right]}\right)^{\frac{1-p}{p}},$$

$$\nabla_\phi\mathbb{E}_{\theta\sim\sigma_0(\theta)}\left[\mathrm{W}_p^p(\theta\sharp\mu_\phi,\theta\sharp\nu)w_{\mu_{\phi'},\nu,\sigma_0,f}(\theta)\right] \approx \frac{1}{L}\sum_{l=1}^L \left(\nabla_\phi \mathrm{W}_p^p(\theta_l\sharp\mu_\phi,\theta_l\sharp\nu)\right) w_{\mu_\phi,\nu',\sigma_0,f}(\theta),$$

$$\mathbb{E}_{\theta\sim\sigma_0(\theta)}\left[w_{\mu_{\phi'},\nu,\sigma_0,f}(\theta)\right] \approx \frac{1}{L}\sum_{l=1}^L w_{\mu_{\phi'},\nu,\sigma_0,f}(\theta).$$

It is worth noting that using a copy of μ_ϕ does not change the value of the distance. This trick will show its true benefit when dealing with the SIR, and the MCMC methods. However, we still discuss it in the IS case for completeness. We refer to the "copy" trick is the "parameter-copy" gradient estimator while the original one is the conventional estimator.

Importance Weighted sliced Wasserstein distance. Although the IS estimation of the EBSW is not an unbiased estimation for finite L, it is an unbiased estimation of a valid distance on the space of probability measures. We refer to the distance as the importance weighted sliced Wasserstein distance (IWSW) which has the following definition.

Definition 3. *For any $p \geq 1$, dimension $d \geq 1$, energy function f, a continuous proposal distribution $\sigma_0(\theta) \sim \mathcal{P}(\mathbb{S}^{d-1})$ and two probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathbb{R}^d$, the importance weighted sliced Wasserstein (IWSW) distance is defined as follows:*

$$\mathit{IWSW}_p(\mu,\nu;f) = \left(\mathbb{E}\left[\frac{\frac{1}{L}\sum_{l=1}^L\left[\mathrm{W}_p^p(\theta_l\sharp\mu,\theta_l\sharp\nu)w_{\mu,\nu,\sigma_0,f}(\theta_l)\right]}{\frac{1}{L}\sum_{l=1}^L\left[w_{\mu,\nu,\sigma_0,f}(\theta_l)\right]}\right]\right)^{\frac{1}{p}}, \tag{7}$$

where the expectation is with respect to $\theta_1,\ldots,\theta_L \overset{i.i.d}{\sim} \sigma_0(\theta)$, and $w_{\mu,\nu,\sigma_0,f}(\theta) = \frac{f(\mathrm{W}_p^p(\theta\sharp\mu,\theta\sharp\nu))}{\sigma_0(\theta)}$.

The IWSW is semi-metric, it also does not suffer from the curse of dimensionality, and it induces weak convergence. The proofs can be derived by following directly the proofs of the EBSW in Appendix A.1, Apendix A.3, and Appendix A.4. Therefore, using the IS estimation of the EBSW is as safe as the SW.

Algorithm 5 Computational algorithm of the SIR-EBSW distance

```
Input: Probability measures $\mu$ and $\nu$, $p \geq 1$, the number of projections $L$, the energy function $f$.
for $l = 1$ to $L$ do
    Sample $\theta_l \sim \mathcal{U}(\mathbb{S}^{d-1})$
    Compute $w_l = f(\text{W}_p(\theta_l \sharp \mu, \theta_l \sharp \nu))$
end for
for $l = 1$ to $L$ do
    Compute $\hat{w}_l = \frac{f(\text{W}_p(\theta_l \sharp \mu, \theta_l \sharp \nu))}{\sum_{i=1}^L f(\text{W}_p(\theta_i \sharp \mu, \theta_i \sharp \nu))}$
end for
for $l = 1$ to $L$ do
    Sample $\theta_l \sim \text{Cat}(\hat{w}_1, \ldots, \hat{w}_L)$
    Compute $v_l = \text{W}_p(\theta_l \sharp \mu, \theta_l \sharp \nu)$
end for
Compute $\widehat{\text{SIR-SW}}_p(\mu, \nu; L, f) = \left(\frac{1}{L} \sum_{l=1}^L v_l\right)^{\frac{1}{p}}$
Return: $\widehat{\text{SIR-SW}}_p(\mu, \nu; L, f)$
```

Algorithm 6 Computational algorithm of the SW distance and the IMH-EBSW distance

```
Input: Probability measures $\mu$ and $\nu$, $p \geq 1$, the number of projections $L$, the energy function $f$.
Sample $\theta_1 \sim \mathcal{U}(\mathbb{S}^{d-1})$
Compute $v_1 = \text{W}_p(\theta_1 \sharp \mu, \theta_1 \sharp \nu)$
for $l = 2$ to $L$ do
    Sample $\theta'_l \sim \mathcal{U}(\mathbb{S}^{d-1})$
    Compute $\alpha = \min\left(1, \frac{f(\text{W}_p^p(\theta'_l \sharp \mu, \theta'_l \sharp \nu)))}{f(\text{W}_p^p(\theta_{l-1} \sharp \mu, \theta_{l-1} \sharp \nu)))}\right)$
    Sample $u \sim \mathcal{U}([0,1])$
    if $\alpha \geq u$ then
        Set $\theta_l = \theta'_l$
    else if $\alpha < u$ then
        Set $\theta_l = \theta_{l-1}$
    end if
    $v_l = \text{W}_p(\theta_l \sharp \mu, \theta_l \sharp \nu)$
end for
Compute $\widehat{\text{IMH-EBSW}}_p(\mu, \nu; L, f) = \left(\frac{1}{L} \sum_{l=1}^L v_l\right)^{\frac{1}{p}}$
Return: $\widehat{\text{IMH-EBSW}}_p(\mu, \nu; L)$
```

B.3 Sampling Importance Resampling and Markov Chain Monte Carlo

Algorithms. We first provide the algorithm for computing the EBSW via the SIR, the IMH, and the RMH in Algorithm 5-7.

Gradient estimators. We derive the reinforce gradient estimator of the EBSW for the SIR, the IMH, and the RHM sampling.

$$\nabla_\phi \text{EBSW}_p(\mu_\phi, \nu; f) = \frac{1}{p} \left(\mathbb{E}_{\theta \sim \sigma_{\mu_\phi, \nu}(\theta; f)} \left[\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu)\right]\right)^{\frac{1-p}{p}} \nabla_\phi \mathbb{E}_{\theta \sim \sigma_{\mu_\phi, \nu}(\theta; f)} \left[\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu)\right].$$

We have:

$$\nabla_\phi \mathbb{E}_{\theta \sim \sigma_{\mu_\phi, \nu}(\theta; f)} \left[\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu)\right] = \mathbb{E}_{\theta \sim \sigma_{\mu_\phi, \nu; f}(\theta)} \left[\text{W}_p^p(\theta_\phi \sharp \mu, \theta \sharp \nu) \nabla_\phi \log \left(\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu) \sigma_{\mu_\phi, \nu}(\theta; f)\right)\right]$$

Algorithm 7 Computational algorithm of the SW distance and the RMH-EBSW distance

```
Input: Probability measures $\mu$ and $\nu$, $p \geq 1$, the number of projections $L$, the energy function $f$,
the concentration parameter $\kappa$.
Sample $\theta_1 \sim \mathcal{U}(\mathbb{S}^{d-1})$
Compute $v_1 = \text{W}_p(\theta_1 \sharp \mu, \theta_1 \sharp \nu)$
for $l = 2$ to $L$ do
    Sample $\theta'_l \sim \text{vMF}(\theta_{l-1}, \kappa)$
    Compute $\alpha = \min\left(1, \frac{f(\text{W}_p^p(\theta'_l \sharp \mu, \theta'_l \sharp \nu)))}{f(\text{W}_p^p(\theta_{l-1} \sharp \mu, \theta_{l-1} \sharp \nu)))}\right)$
    Sample $u \sim \mathcal{U}([0,1])$
    if $\alpha \geq u$ then
        Set $\theta_l = \theta'_l$
    else if $\alpha < u$ then
        Set $\theta_l = \theta_{l-1}$
    end if
    $v_l = \text{W}_p(\theta_l \sharp \mu, \theta_l \sharp \nu)$
end for
Compute $\widehat{\text{RMH-EBSW}}_p(\mu, \nu; L, f) = \left(\frac{1}{L}\sum_{l=1}^{L} v_l\right)^{\frac{1}{p}}$
Return: $\widehat{\text{RMH-EBSW}}_p(\mu, \nu; L)$
```

and

$$\begin{aligned}
\nabla_\phi \log\left(\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu)\sigma_{\mu_\phi,\nu}(\theta; f)\right) &= \nabla_\phi \log(\text{W}_p^p \theta \sharp \mu_\phi, \theta \sharp \nu)) + \nabla_\phi \log(f(\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu))) \\
&\quad - \nabla_\phi \log\left(\int_{\mathbb{S}^{d-1}} f(\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu))d\theta\right) \\
&= \frac{1}{\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu))} \nabla_\phi \text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu) \\
&\quad + \frac{1}{f(\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu))} \nabla_\phi f(\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu)) \\
&\quad - \nabla_\phi \log\left(\int_{\mathbb{S}^{d-1}} f(\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu))d\theta\right),
\end{aligned}$$

and

$$\begin{aligned}
\nabla_\phi \log\left(\int_{\mathbb{S}^{d-1}} f(\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu))d\theta\right) &= \nabla_\phi \log\left(\mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}\left[f(\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu))\frac{2\pi^{d/2}}{\Gamma(d/2)}\right]\right) \\
&= \nabla_\phi \log\left(\mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}\left[f(\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu))\right]\right) \\
&= \frac{1}{\mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}\left[f(\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu))\right]} \nabla_\phi \mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}\left[f(\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu)\right].
\end{aligned}$$

Using L Monte Carlo samples from the SIR (or the IMH or the RMH) to approximate the expectation $\mathbb{E}_{\theta \sim \sigma_{\mu_\phi,\nu}(\theta; f)}$, and L samples from $\mathcal{U}(\mathbb{S}^{d-1})$ to approximate the expectation $\mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}$, we obtain the gradient estimator of the EBSW. However, the reinforce gradient estimator is unstable in practice, especially with the energy function $f_e(x) = e^x$. Therefore, we propose a more simple gradient estimator which is

$$\nabla_\phi \text{EBSW}_p(\mu_\phi, \nu; f) \approx \frac{1}{p}\left(\mathbb{E}_{\theta \sim \sigma_{\mu_{\phi'},\nu}(\theta; f)}\left[\text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu)\right]\right)^{\frac{1-p}{p}} \mathbb{E}_{\theta \sim \sigma_{\mu_{\phi'},\nu}(\theta; f)}\left[\nabla_\phi \text{W}_p^p(\theta \sharp \mu_\phi, \theta \sharp \nu)\right].$$

The key is to use a copy of the parameter ϕ' for constructing the slicing distribution $\sigma_{\mu_{\phi'},\nu}(\theta; f)$, hence, we can exchange directly the differentiation and the expectation. It is worth noting that using the copy also affects the gradient estimation, it does not change the value of the distance. We refer to the "copy" trick is the "parameter-copy" gradient estimator while the original one is the conventional estimator.

Population distance. The approximated values of p-power EBSW from using the SIR, the IMH, and the RMH can be all written as $\frac{1}{L}\sum_{l=1}^{L} \text{W}_p^p(\theta_l \sharp \mu, \theta_l \sharp \nu)$. Here, the distributions of $\theta_1, \ldots, \theta_L$

are different. Therefore, they are not an unbiased estimation of the $\text{EBSW}_p^p(\mu,\nu;f)$. However, the population distance of the estimation can be defined as in Definition 4.

Definition 4. *For any $p \geq 1$, dimension $d \geq 1$, energy function f, and two probability measures $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathbb{R}^d$, the projected sliced Wasserstein (PSW) distance is defined as follows:*

$$PSW_p(\mu,\nu;f) = \left(\mathbb{E}\left[\frac{1}{L}\sum_{l=1}^{L} W_p^p(\theta_l\sharp\mu,\theta\sharp\nu)\right]\right)^{\frac{1}{p}}, \tag{8}$$

where the expectation is with respect to $(\theta_1,\ldots,\theta_L) \sim \sigma(\theta_1,\ldots,\theta_L)$ which is a distribution defined by the SIR (the IMH or the RHM).

The PSW is a valid metric since it satisfies the triangle inequality in addition to the symmetry, the non-negativity, and the identity. In particular, given three probability measures $\mu_1,\mu_2,\mu_3 \in \mathcal{P}_p(\mathbb{R}^d)$ we have:

$$\begin{aligned}
\text{PSW}_p(\mu_1,\mu_3) &= \left(\mathbb{E}_{(\theta_{1:L})\sim\sigma(\theta_{1:L})}\left[\frac{1}{L}\sum_{l=1}^{L} W_p^p\left(\theta_l\sharp\mu_1,\theta_l\sharp\mu_3\right)\right]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}_{(\theta_{1:L})\sim\sigma(\theta_{1:L})}\left[\frac{1}{L}\sum_{t=1}^{L}\left(W_p\left(\theta_l\sharp\mu_1,\theta_l\sharp\mu_2\right)+W_p\left(\theta_l\sharp\mu_2,\theta_l\sharp\mu_3\right)\right)^p\right]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}_{(\theta_{1:L})\sim\sigma(\theta_{1:L})}\left[\frac{1}{L}\sum_{t=1}^{L} W_p^p\left(\theta_l\sharp\mu_1,\theta_l\sharp\mu_2\right)\right]\right)^{\frac{1}{p}} \\
&\quad + \left(\mathbb{E}_{(\theta_{1:L})\sim\sigma(\theta_{1:L})}\left[\frac{1}{L}\sum_{l=1}^{T} W_p^p\left(\theta_l\sharp\mu_2,\theta_l\sharp\mu_3\right)\right]\right)^{\frac{1}{p}} \\
&= \text{PSW}_p(\mu_1,\mu_2) + \text{PSW}_p(\mu_2,\mu_3),
\end{aligned}$$

where the first inequality is due to the triangle inequality of Wasserstein distance and the second inequality is due to the Minkowski inequality. The PSW also does not suffer from the curse of dimensionality, and it induces weak convergence. The proofs can be derived by following directly the proofs of the EBSW in Appendix A.1, Apendix A.3, and Appendix A.4. Therefore, using the SIR, the IMH, and the RMH estimation of the EBSWs are as safe as the SW.

C Additional Experiments

In this section, we provide additional results for point-cloud gradient flows in Appendix C.1, color transfer in Appendix C.2, and deep point-cloud reconstruction in Appendix C.3.

C.1 Point-Cloud Gradient Flows

We provide the full experimental results including the IS-EBSW, the SIR-EBSW, the IMH-EBSW, and the RMH-EBSW with both the exponential energy function and the identity energy function in Table 3. In the table, we also include the results for the number of projections $L = 10$. In Table 3, we use the conventional gradient estimator for the IS-EBSW while the "parameter-copy" estimator is used for other variants of the EBSW. Therefore, we also provide the ablation studies comparing the gradient estimators in Table 4 by adding the results for the "parameter-copy" estimator for the IS-EBSW and the conventional estimator for other variants. Experimental settings are the same as in the main text.

Quantitative Results. From the two tables, we observe that the IS-EBSW is the best variant of the EBSW in both performance and computational time. Also, we observe that the exponential energy function is better than the identity energy function in this application. It is worth noting that the EBSW variants of all computational methods and energy functions are better than the baselines in terms of Wasserstein-2 distances at the last epoch. For all sliced Wasserstein variants, we see that reducing the number of projections leads to worsening performance which is consistent with previous

Table 3: Summary of Wasserstein-2 scores (multiplied by 10^4) from three different runs, computational time in second (s) to reach step 500 of different sliced Wasserstein variants in gradient flows.

Distances	Step 0 ($W_2 \downarrow$)	Step 100 ($W_2 \downarrow$)	Step 200 ($W_2 \downarrow$)	Step 300 ($W_2 \downarrow$)	Step 400($W_2 \downarrow$)	Step 500 ($W_2 \downarrow$)	Time (s $\downarrow$)
SW L=100	2048.29 ± 0.0	986.93 ± 9.55	350.66 ± 5.32	99.69 ± 1.85	27.03 ± 0.65	9.41 ± 0.27	$\mathbf{17.06 \pm 0.45}$
Max-SW T=100	2048.29 ± 0.0	506.56 ± 9.28	93.54 ± 3.39	22.2 ± 0.79	9.62 ± 0.22	6.83 ± 0.22	28.38 ± 0.05
v-DSW L*T=100	2048.29 ± 0.0	649.33 ± 8.77	127.4 ± 5.06	29.44 ± 1.25	10.95 ± 1.0	5.68 ± 0.56	21.2 ± 0.02
IS-EBSW-e L=100	2048.29 ± 0.0	$\mathbf{419.09 \pm 2.64}$	$\mathbf{71.02 \pm 0.46}$	$\mathbf{18.2 \pm 0.05}$	$\mathbf{6.9 \pm 0.08}$	$\mathbf{3.3 \pm 0.08}$	17.63 ± 0.02
SIR-EBSW-e L=100	2048.29 ± 0.0	435.02 ± 1.1	85.26 ± 0.11	21.96 ± 0.12	7.9 ± 0.22	3.79 ± 0.17	29.8 ± 0.04
IMH-EBSW-e L=100	2048.29 ± 0.0	460.19 ± 3.46	91.28 ± 1.19	23.35 ± 0.52	8.26 ± 0.26	3.93 ± 0.14	49.3 ± 0.54
RMH-EBSW-e L=100	2048.29 ± 0.0	454.92 ± 3.25	87.92 ± 0.69	22.66 ± 0.46	8.14 ± 0.31	3.82 ± 0.24	62.5 ± 0.09
IS-EBSW-1 L=100	2048.29 ± 0.0	692.63 ± 7.21	167.75 ± 3.12	41.8 ± 0.93	12.31 ± 0.27	5.35 ± 0.1	17.91 ± 0.28
SIR-EBSW-1 L=100	2048.29 ± 0.0	704.08 ± 2.75	169.88 ± 0.47	41.85 ± 0.28	12.58 ± 0.24	5.64 ± 0.18	30.56 ± 0.05
IMH-EBSW-1 L=100	2048.29 ± 0.0	715.97 ± 4.49	171.42 ± 1.25	42.05 ± 0.42	12.6 ± 0.1	5.63 ± 0.06	50.01 ± 0.01
RMH-EBSW-1 L=100	2048.29 ± 0.0	712.11 ± 1.64	173.47 ± 1.49	42.94 ± 0.4	12.68 ± 0.15	5.54 ± 0.09	64.01 ± 0.08
SW L=10	2048.29 ± 0.0	988.57 ± 14.01	351.63 ± 2.63	101.54 ± 2.45	28.19 ± 1.04	10.11 ± 0.34	$\mathbf{3.84 \pm 0.04}$
Max-SW T=10	2048.29 ± 0.0	525.72 ± 7.35	134.8 ± 4.6	34.07 ± 0.34	10.77 ± 0.15	7.36 ± 0.31	6.55 ± 0.06
IS-EBSW-e L=10	2048.29 ± 0.0	519.73 ± 8.63	$\mathbf{92.14 \pm 1.29}$	$\mathbf{23.94 \pm 0.07}$	$\mathbf{9.03 \pm 0.33}$	$\mathbf{4.59 \pm 0.22}$	5.57 ± 0.03
SIR-EBSW-e L=10	2048.29 ± 0.0	$\mathbf{508.86 \pm 8.49}$	104.47 ± 1.93	28.27 ± 0.68	10.56 ± 0.08	5.61 ± 0.16	6.84 ± 0.06
IMH-EBSW-e L=10	2048.29 ± 0.0	621.51 ± 22.49	131.75 ± 7.09	34.42 ± 1.89	11.55 ± 0.38	5.56 ± 0.09	8.41 ± 0.04
RMH-EBSW-e L=10	2048.29 ± 0.0	642.87 ± 5.25	135.91 ± 8.39	36.11 ± 2.13	12.57 ± 0.75	5.94 ± 0.31	9.69 ± 0.04
IS-EBSW-1 L=10	2048.29 ± 0.0	713.65 ± 5.68	177.16 ± 1.19	45.07 ± 0.17	13.6 ± 0.26	6.16 ± 0.22	5.69 ± 0.0
SIR-EBSW-1 L=10	2048.29 ± 0.0	731.4 ± 9.37	181.28 ± 5.05	44.99 ± 1.07	13.59 ± 0.51	6.68 ± 0.27	6.9 ± 0.03
IMH-EBSW-1 L=10	2048.29 ± 0.0	772.86 ± 28.09	199.29 ± 7.02	48.73 ± 1.69	14.1 ± 0.49	6.25 ± 0.35	8.61 ± 0.02
RMH-EBSW-1 L=10	2048.29 ± 0.0	810.1 ± 10.2	212.11 ± 9.53	54.62 ± 2.63	15.44 ± 0.93	6.74 ± 0.32	9.86 ± 0.06

Table 4: Summary of Wasserstein-2 scores (multiplied by 10^4) from three different runs, computational time in second (s) to reach step 500 of different sliced Wasserstein variants in gradient flows.

Distances	Step 0 ($W_2 \downarrow$)	Step 100 ($W_2 \downarrow$)	Step 200 ($W_2 \downarrow$)	Step 300 ($W_2 \downarrow$)	Step 400($W_2 \downarrow$)	Step 500 ($W_2 \downarrow$)	Time (s $\downarrow$)
IS-EBSW-e L=100 (c)	2048.29 ± 0.0	435.39 ± 1.82	85.31 ± 0.44	21.9 ± 0.09	7.81 ± 0.06	3.68 ± 0.07	17.51 ± 0.01
IS-EBSW-1 L=100 (c)	2048.29 ± 0.0	711.33 ± 7.2	170.69 ± 2.91	42.2 ± 0.79	12.62 ± 0.2	5.7 ± 0.11	17.72 ± 0.02
SIR-EBSW-1 L=100	2048.29 ± 0.0	685.87 ± 8.35	166.39 ± 2.65	41.52 ± 0.56	12.29 ± 0.32	5.56 ± 0.1	44.51 ± 0.16
IMH-EBSW-1 L=100	2048.29 ± 0.0	700.47 ± 9.13	173.25 ± 1.26	44.08 ± 0.52	13.03 ± 0.18	5.93 ± 0.2	63.83 ± 0.02
RMH-EBSW-1 L=100	2048.29 ± 0.0	711.0 ± 10.98	175.76 ± 1.45	44.5 ± 0.56	13.39 ± 0.13	6.06 ± 0.05	77.32 ± 0.2
IS-EBSW-e L=10 (c)	2048.29 ± 0.0	524.69 ± 7.38	107.37 ± 2.18	28.46 ± 0.35	10.13 ± 0.38	4.93 ± 0.37	5.54 ± 0.04
IS-EBSW-1 L=10 (c)	2048.29 ± 0.0	729.53 ± 6.74	179.35 ± 1.7	45.03 ± 0.79	13.32 ± 0.82	6.15 ± 0.46	5.7 ± 0.03
SIR-EBSW-1 L=10	2048.29 ± 0.0	762.23 ± 9.66	202.2 ± 5.23	56.48 ± 1.55	19.05 ± 0.83	10.42 ± 0.53	8.45 ± 0.02
IMH-EBSW-1 L=10	2048.29 ± 0.0	762.67 ± 14.63	200.3 ± 6.48	54.28 ± 1.17	18.11 ± 0.36	9.29 ± 0.26	10.02 ± 0.02
RMH-EBSW-1 L=10	2048.29 ± 0.0	817.92 ± 23.86	220.66 ± 2.55	60.15 ± 1.53	20.0 ± 0.7	9.8 ± 0.36	11.35 ± 0.03

studies in previous works [27, 19]. In Table 3, the IS-EBSW uses the conventional gradient estimator while the SIR-EBSW, the IMH-EBSW, and the RMH-EBSW use the "parameter-copy" estimator. Therefore, we report the IS-EBSW with the "parameter-copy" estimator and the SIR-EBSW, the IMH-EBSW, and the RMH-EBSW with the Reinforce estimator (conventional estimator) in Table 4. From the table, we observe the "parameter-copy" estimator is worse than the conventional estimator in the case of IS-EBSW. For the SIR-EBSW, the IMH-EBSW, and the RMH-EBSW, we cannot use the exponential energy function due to the numerically unstable Reinforce estimator. In the case of the identity energy function, the exponential energy function is also worse than the "parameter-copy" estimator. Therefore, we recommend to use the IS-EBSW-e with the conventional gradient estimator.

Qualitative Results. We provide the visualization of the gradient flows from SW (L=100), Max-SW (T=100), v-DSW (L=10,T=10), and all the EBSW-e variants in Figure 4. Overall, we see that EBSW-e variants give smoother flows than other baselines. Despite having slightly different quantitative scores due to the approximation methods, the visualization from the EBSW-e variants is consistent. Therefore, the energy-based slicing function helps to improve the convergence of the source point-cloud to the target point-cloud.

C.2 Color Transfer

Similar to the point-cloud gradient flow, we follow the same experimental settings of color transfer in the main text. We provide the full experimental results including the IS-EBSW, the SIR-EBSW, the IMH-EBSW, and the RMH-EBSW with both the exponential energy function and the identity energy function, with both $L = 10$ and $L = 100$, and with both gradient estimators in Figure 5.

Results. From the figure, we observe that IMH-EBSW-e gives the best Wasserstein-2 distance among all EBSW variants. Between the exponential energy function and the identity energy function, we see that the exponential energy function yields a better result for all EBSW variants. Similar to the gradient flow, reducing the number of projections to 10 also leads to worse results for all

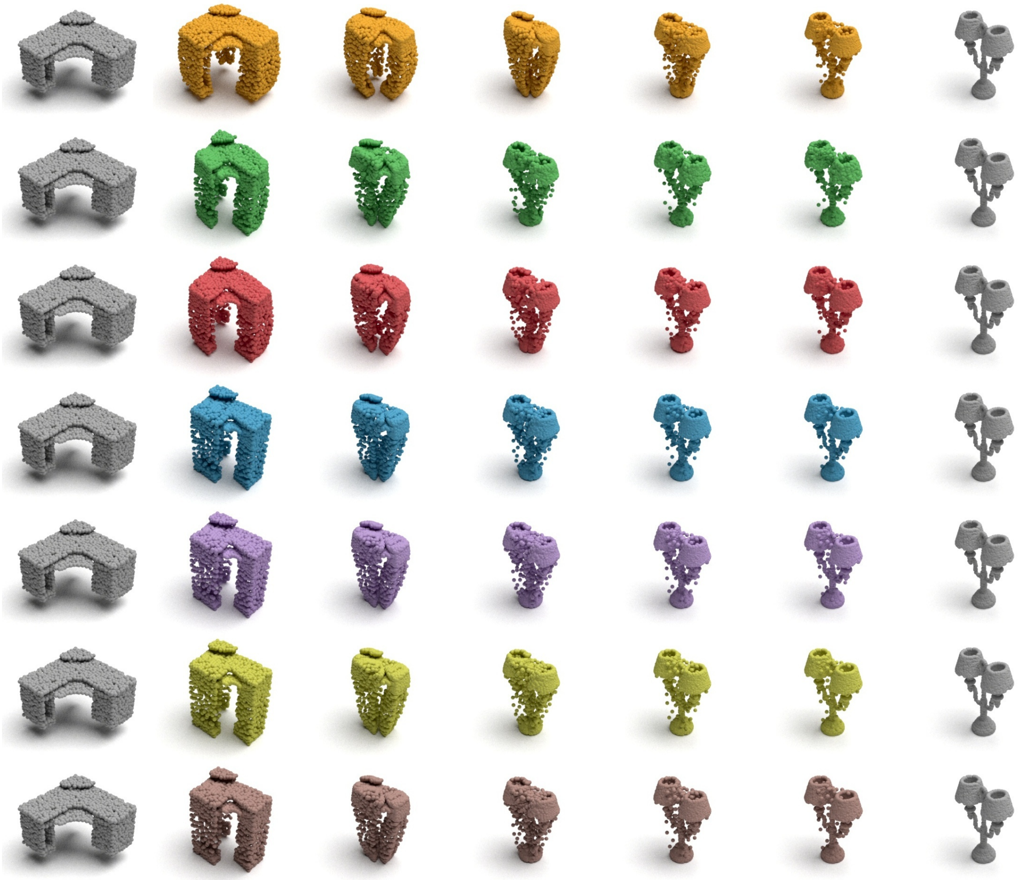

Figure 4: Gradient flows from the SW, the Max-SW, the v-DSW, the IS-EBSW-e, the SIR-EBSW-e, the IMH-EBSW-e, and the RMH-EBSW-e in turn.

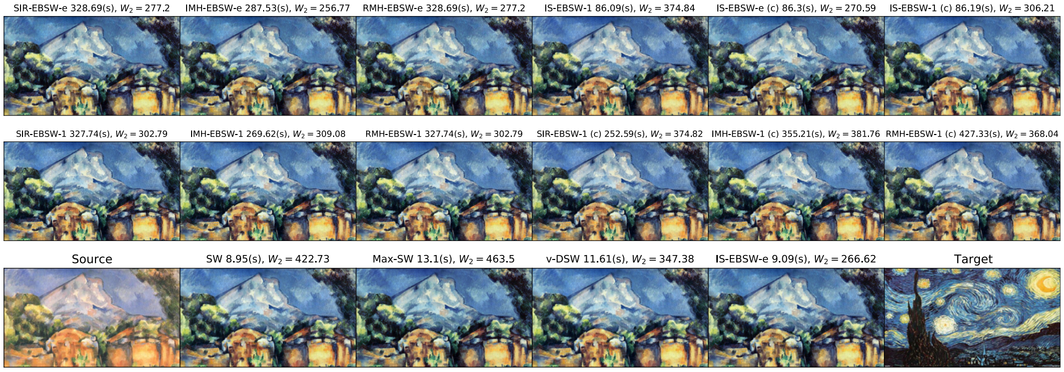
SIR-EBSW-e 328.69(s), $W_2 = 277.2$
IMH-EBSW-e 287.53(s), $W_2 = 256.77$
RMH-EBSW-e 328.69(s), $W_2 = 277.2$
IS-EBSW-1 86.09(s), $W_2 = 374.84$
IS-EBSW-e (c) 86.3(s), $W_2 = 270.59$
IS-EBSW-1 (c) 86.19(s), $W_2 = 306.21$
SIR-EBSW-1 327.74(s), $W_2 = 302.79$
IMH-EBSW-1 269.62(s), $W_2 = 309.08$
RMH-EBSW-1 327.74(s), $W_2 = 302.79$
SIR-EBSW-1 (c) 252.59(s), $W_2 = 374.82$
IMH-EBSW-1 (c) 355.21(s), $W_2 = 381.76$
RMH-EBSW-1 (c) 427.33(s), $W_2 = 368.04$
Source
SW 8.95(s), $W_2 = 422.73$
Max-SW 13.1(s), $W_2 = 463.5$
v-DSW 11.61(s), $W_2 = 347.38$
IS-EBSW-e 9.09(s), $W_2 = 266.62$
Target

Figure 5: The first two rows are with $L = 100$, (c) denotes the "parameter-copy" (the SIR-EBSW-e, the IMH-EBSW-e, the RMH-EBSW always use the "parameter-copy" estimator since the conventional estimator is not stable for them), and the last row is with $L = 10$.

sliced Wasserstein variants For the gradient estimators, the conventional estimator is preferred for the IS-EBSW while the "parameter-copy" estimator is preferred for other EBSW variants.

C.3 Deep Point-cloud Reconstruction

We follow the same experimental settings as in the main text. We provide the full experimental results including the IS-EBSW, the SIR-EBSW, the IMH-EBSW, and the RMH-EBSW with both the exponential energy function and the identity energy function, with both $L = 10$ and $L = 100$

Table 5: Reconstruction errors of different autoencoders measured by the (sliced) Wasserstein distance (×100). The results are from three different runs.

Distance	Epoch 20		Epoch 100		Epoch 200	
	$SW_2(\downarrow)$	$W_2(\downarrow)$	SW_2 $(\downarrow)$	$W_2(\downarrow)$	SW_2 $(\downarrow)$	$W_2(\downarrow)$
SW L=100	2.97 ± 0.14	12.67 ± 0.18	2.29 ± 0.04	10.63 ± 0.05	2.15 ± 0.04	9.97 ± 0.08
Max-SW T=100	2.91 ± 0.06	12.33 ± 0.05	2.24 ± 0.05	10.40 ± 0.06	2.14 ± 0.10	9.84 ± 0.12
v-DSW L*T=100	2.84 ± 0.02	12.64 ± 0.02	2.21 ± 0.01	10.52 ± 0.04	2.07 ± 0.09	9.81 ± 0.05
IS-EBSW-e L=100	$\mathbf{2.68 \pm 0.03}$	$\mathbf{11.90 \pm 0.04}$	$\mathbf{2.18 \pm 0.04}$	$\mathbf{10.27 \pm 0.01}$	2.04 ± 0.09	$\mathbf{9.69 \pm 0.14}$
SIR-EBSW-e L=100	2.77 ± 0.01	12.16 ± 0.04	2.24 ± 0.04	10.40 ± 0.01	2.00 ± 0.03	9.72 ± 0.04
IMH-EBSW-e L=100	2.75 ± 0.03	12.15 ± 0.04	2.19 ± 0.08	10.39 ± 0.09	$\mathbf{1.99 \pm 0.05}$	9.72 ± 0.10
RMH-EBSW-e L=100	2.83 ± 0.02	12.21 ± 0.03	2.20 ± 0.03	10.38 ± 0.07	2.02 ± 0.02	9.72 ± 0.03
IS-EBSW-1 L=100	2.83 ± 0.01	12.37 ± 0.01	2.27 ± 0.06	10.59 ± 0.07	2.11 ± 0.04	9.90 ± 0.02
SIR-EBSW-1 L=100	2.81 ± 0.02	12.32 ± 0.03	2.26 ± 0.08	10.56 ± 0.14	2.07 ± 0.01	9.81 ± 0.08
IMH-EBSW-1 L=100	2.82 ± 0.01	12.32 ± 0.02	2.28 ± 0.11	10.55 ± 0.13	2.03 ± 0.02	9.81 ± 0.02
RMH-EBSW-1 L=100	2.88 ± 0.04	12.42 ± 0.06	2.22 ± 0.07	10.37 ± 0.06	2.01 ± 0.02	9.73 ± 0.02
SW L=10	2.99 ± 0.12	12.70 ± 0.16	2.30 ± 0.01	10.64 ± 0.04	2.17 ± 0.06	10.01 ± 0.09
Max-SW T=10	3.00 ± 0.07	12.68 ± 0.05	2.31 ± 0.08	10.67 ± 0.06	2.14 ± 0.04	9.95 ± 0.05
IS-EBSW-e L=10	$\mathbf{2.76 \pm 0.04}$	$\mathbf{12.15 \pm 0.06}$	$\mathbf{2.20 \pm 0.08}$	$\mathbf{10.39 \pm 0.10}$	2.04 ± 0.07	$\mathbf{9.77 \pm 0.10}$
SIR-EBSW-e L=10	2.79 ± 0.03	12.26 ± 0.05	2.26 ± 0.08	10.53 ± 0.09	2.08 ± 0.11	9.87 ± 0.16
IMH-EBSW-e L=10	2.82 ± 0.02	12.33 ± 0.02	2.26 ± 0.12	10.53 ± 0.20	2.07 ± 0.02	9.86 ± 0.03
RMH-EBSW-e L=10	2.86 ± 0.04	12.37 ± 0.03	2.21 ± 0.01	10.45 ± 0.05	$\mathbf{2.02 \pm 0.0}2$	9.78 ± 0.01
IS-EBSW-1 L=10	2.84 ± 0.01	12.43 ± 0.01	2.28 ± 0.10	10.63 ± 0.11	2.10 ± 0.05	9.91 ± 0.05
SIR-EBSW-1 L=10	2.84 ± 0.01	12.38 ± 0.01	2.28 ± 0.07	10.59 ± 0.10	2.07 ± 0.07	9.88 ± 0.12
IMH-EBSW-1 L=10	2.82 ± 0.01	12.36 ± 0.03	2.28 ± 0.08	10.52 ± 0.05	2.08 ± 0.06	9.86 ± 0.09
RMH-EBSW-1 L=10	2.89 ± 0.04	12.47 ± 0.03	2.21 ± 0.03	10.45 ± 0.08	2.03 ± 0.03	9.80 ± 0.02

Table 6: Reconstruction errors of different autoencoders measured by the (sliced) Wasserstein distance (×100). We use (c) for the "parameter-copy" gradient estimator. The results are from three different runs.

Distance	Epoch 20		Epoch 100		Epoch 200	
	$SW_2(\downarrow)$	$W_2(\downarrow)$	SW_2 $(\downarrow)$	$W_2(\downarrow)$	SW_2 $(\downarrow)$	$W_2(\downarrow)$
IS-EBSW-e L=100 (c)	2.74 ± 0.04	12.14 ± 0.12	2.22 ± 0.07	10.42 ± 0.05	2.07 ± 0.01	9.77 ± 0.07
IS-EBSW-1 L=100 (c)	2.83 ± 0.01	12.34 ± 0.03	2.30 ± 0.05	10.60 ± 0.09	2.05 ± 0.07	9.83 ± 0.11
SIR-EBSW-1 L=100	2.80 ± 0.02	12.29 ± 0.01	2.21 ± 0.05	10.46 ± 0.08	2.04 ± 0.02	9.81 ± 0.07
IMH-EBSW-1 L=100	2.96 ± 0.05	12.67 ± 0.08	2.35 ± 0.05	10.82 ± 0.07	2.20 ± 0.11	10.20 ± 0.16
RMH-EBSW-1 L=100	3.00 ± 0.06	12.67 ± 0.10	2.27 ± 0.02	10.66 ± 0.06	2.15 ± 0.05	10.11 ± 0.11
IS-EBSW-e L=10 (c)	2.77 ± 0.01	12.22 ± 0.04	2.28 ± 0.09	10.63 ± 0.11	2.07 ± 0.07	9.80 ± 0.15
IS-EBSW-1 L=10 (c)	2.86 ± 0.02	12.42 ± 0.02	2.24 ± 0.08	10.52 ± 0.13	2.05 ± 0.04	9.84 ± 0.10
SIR-EBSW-1 L=10	2.87 ± 0.02	12.43 ± 0.08	2.36 ± 0.11	10.67 ± 0.19	2.08 ± 0.10	9.88 ± 0.14
IMH-EBSW-1 L=10	2.98 ± 0.02	12.65 ± 0.04	2.35 ± 0.05	10.84 ± 0.06	2.21 ± 0.11	10.22 ± 0.11
RMH-EBSW-1 L=10	3.01 ± 0.04	12.82 ± 0.05	2.37 ± 0.03	10.87 ± 0.03	2.11 ± 0.02	10.13 ± 0.06

in Table 5. In Table 5, we use the conventional gradient estimator for the IS-EBSW while other variants of EBSW use the "parameter-copy" gradient estimator. We also compare gradient estimators for the EBSW by adding the results for the "parameter-copy" gradient estimator for the IS-EBSW (denoted as (c)), and the conventional gradient estimator for the SIR-EBSW, the IMH-EBSW, and the RMH-EBSW in Table 6.

Quantitative Results. From the two tables, we observe that the IS-EBSW-e performs the best for both settings of the number of projections $L = 10$ and $L = 100$ in terms of the Wasserstein-2 reconstruction errors. For the SW reconstruction error, it is only slightly worse than the SIR-EBSW-e at epoch 200. Comparing the exponential energy function and the identity energy function, we observe that the exponential function is better in both settings of the number of projections. For the same number of projections, the EBSW variants with both types of energy function give lower errors than the baseline including the SW, the Max-SW, and the v-DSW. For all sliced Wasserstein variants, a higher value of the number of projections gives better results. For the gradient estimator of the EBSW, we see that the conventional gradient estimator is preferred for the IS-EBSW while the "parameter-copy" estimator is preferred for other EBSW variants.

Qualitative Results. We show some ground-truth point-clouds ModelNet40 and their corresponding reconstructed point-clouds from different models ($L = 100$) at epochs 200 and 20 in Figure 6- 7 respectively. From the top to the bottom is the ground truth, the SW, the Max-SW, the v-DSW, the IS-EBSW-e, the SIR-EBSW-e, the IMH-EBSW-e, and the RMH-EBSW-e.

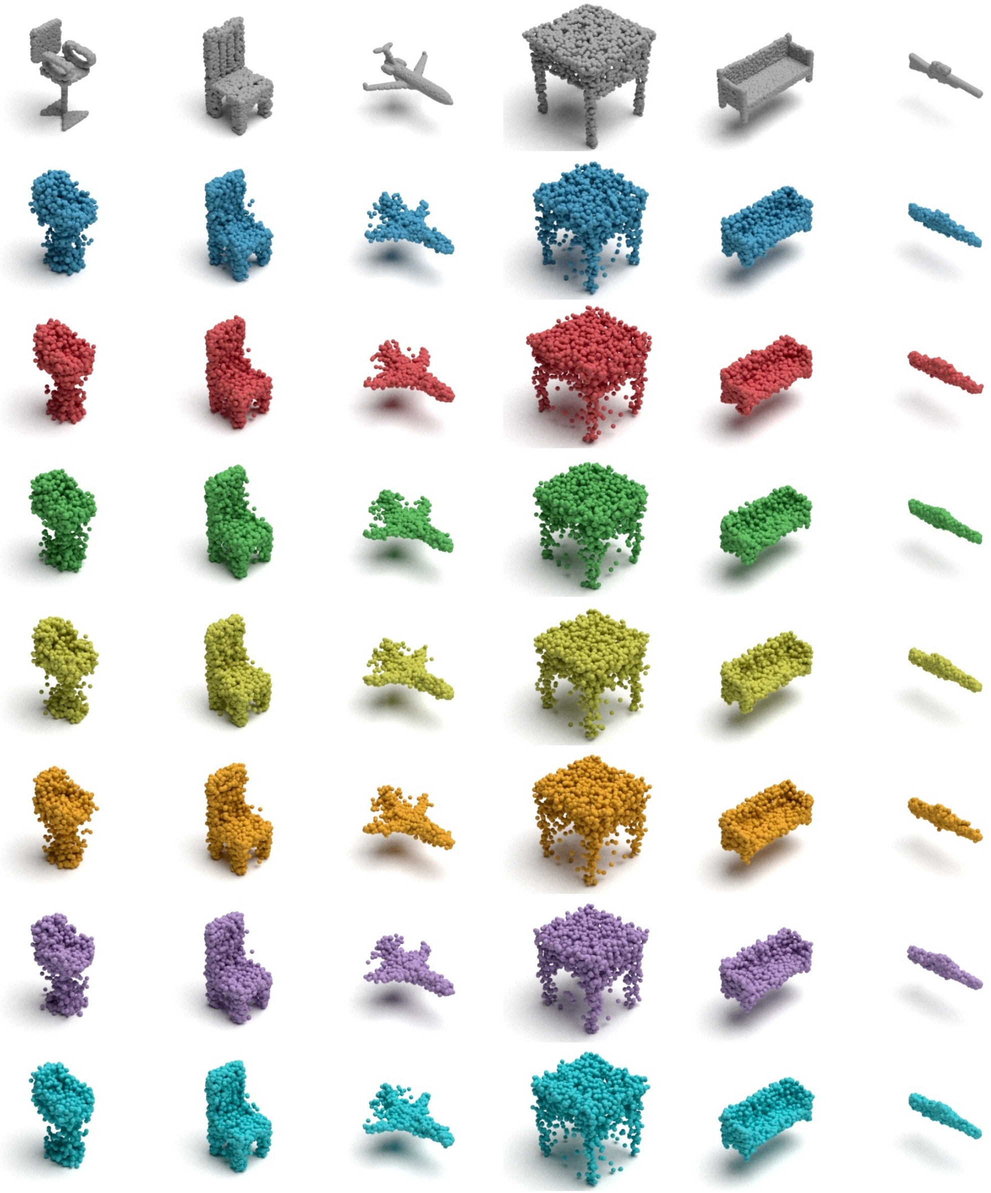

Figure 6: From the top to the bottom is the ground truth, the reconstructed point-clouds at epoch 200 of the SW, the Max-SW, the v-DSW, the IS-EBSW-e, the SIR-EBSW-e, the IMH-EBSW-e, and the RMH-EBSW-e respectively.

D Computational Infrastructure

For the point-cloud gradient flows and the color transfer, we use a Macbook Pro M1 for conducting experiments. For deep point-cloud reconstruction, experiments are run on a single NVIDIA V100 GPU.

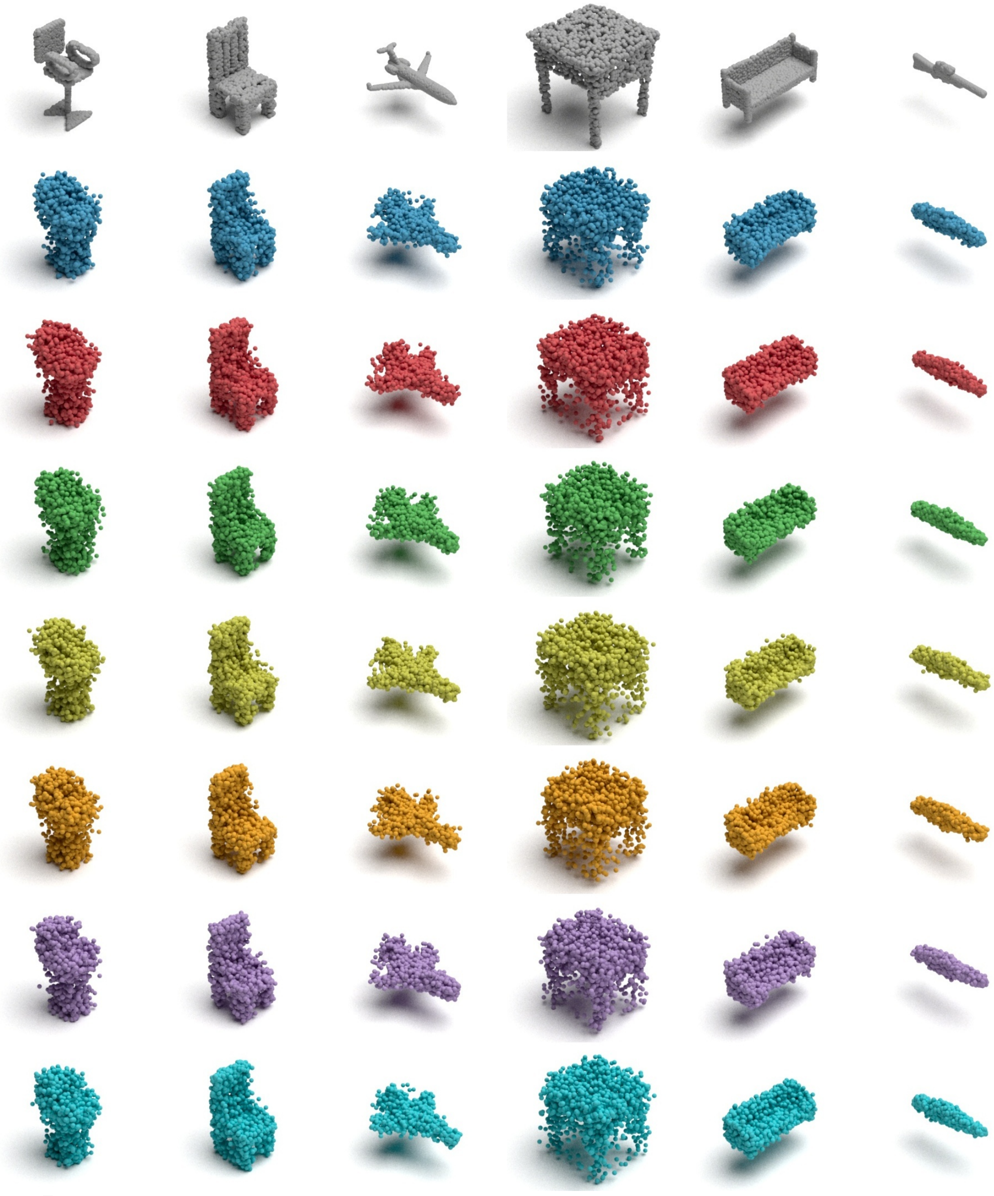

Figure 7: From the top to the bottom is the ground truth, the reconstructed point-clouds at epoch 20 of the SW, the Max-SW, the v-DSW, the IS-EBSW-e, the SIR-EBSW-e, the IMH-EBSW-e, and the RMH-EBSW-e respectively.