Sliced Optimal Transport

Kimia Nadjahi

Optimal Transport for Machine Learning (IASD/MVA)

October 29, 2025

1 Sliced Optimal Transport: Definition and Properties

Definition 1 (Slicing operator). *Let $u \in \mathbb{R}^d$. The slicing operator based on u, $p_u : \mathbb{R}^d \to \mathbb{R}$, is defined for any $x \in \mathbb{R}^d$ as,*

$$p_u(x) = \langle x, u \rangle = \sum_{i=1}^{d} x^{(i)} u^{(i)} \,,$$

where $y^{(i)}$ denotes the i-th component of a vector $y \in \mathbb{R}^d$.

Consider $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ with each $x_i \in \mathbb{R}^d$. Then, slicing μ yields a discrete distribution on $\mathbb{R}$, defined as

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{p_u(x_i)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\langle x_i, u \rangle}$$

We can generalize this to any continuous measure μ on $\mathbb{R}^d$ using the push-forward operator. Formally, for any $x \sim \mu$, then $\langle u, x \rangle \sim (p_u)_\sharp \mu$.

In words: the pushforward operator allows us to lift operations on points in $\mathbb{R}^d$ (like the projection p_u) to operations on measures.

Useful property: apply the change of variable formula for push-forward measures on $(p_u)_\sharp\mu$: for any measurable function g on $\mathbb{R}$ s.t. g is integrable with respect to $(p_u)_\sharp\mu$ ($\Leftrightarrow g \circ p_u$ is integrable w.r.t. μ), one has

$$\int_{\mathbb{R}} g(s)\mathrm{d}((p_u)_\sharp)\mu)(s) = \int_{\mathbb{R}^d} (g \circ p_u)(x)\mathrm{d}\mu(x) = \int_{\mathbb{R}^d} g(\langle u, x \rangle)\mathrm{d}\mu(x) \,. \tag{1}$$

Definition 2 (Sliced-Wasserstein distance, Rabin et al. [2012]). *Let μ, ν be two distributions on $\mathbb{R}^d$. Denote by $\mathbf{W}_c$ the* one-dimensional *Wasserstein distance based on the cost function $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. The* Sliced-Wasserstein distance *(based on c) between μ and ν is defined as,*

$$\mathbf{SW}_c(\mu, \nu) = \mathbb{E}_{\theta \sim \mathcal{U}_{\mathbb{S}^{d-1}}}\left[\mathbf{W}_c\big((p_\theta)_\sharp\mu, (p_\theta)_\sharp\nu\big)\right] \tag{2}$$

$$= \int_{\mathbb{S}^{d-1}} \mathbf{W}_c\big((p_\theta)_\sharp\mu, (p_\theta)_\sharp\nu\big)\mathrm{d}\mathcal{U}_{\mathbb{S}^{d-1}}(\theta) \tag{3}$$

where $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$, and $\mathcal{U}_{\mathbb{S}^{d-1}}$ is the uniform distribution on $\mathbb{S}^{d-1}$.

Remark 1 (Uniform distribution on $\mathbb{S}^{d-1}$). *There is a unique Borel measure σ on $\mathbb{S}^{d-1}$ such that for every non-negative Borel measurable function f on $\mathbb{R}^d$ (more explanations here)*

$$\int_{\mathbb{R}^d} f(x)\mathrm{d}x = \int_0^{+\infty} \int_{\mathbb{S}^{d-1}} f(r\theta) r^{d-1} \mathrm{d}\sigma(\theta)\mathrm{d}r \,. \tag{4}$$

This is the change of variable formula from Cartesian coordinates to polar coordinates in $(0,+\infty)\times\mathbb{S}^{d-1}$*) The uniform distribution on* $\mathbb{S}^{d-1}$ *is then defined as, for any Borel set* $B\subset\mathbb{S}^{d-1}$*,*

$$\mathcal{U}_{\mathbb{S}^{d-1}}(B)=\frac{\sigma(B)}{\mathcal{A}(\mathbb{S}^{d-1})}\,,$$

where $\mathcal{A}(\mathbb{S}^{d-1})=\frac{2\pi^{d/2}}{\Gamma(d/2)}$ *is the surface area of the sphere, with* $\Gamma(d/2)=\int_0^{+\infty}e^{-t}t^{d/2-1}\mathrm{d}t$*.*

Remark 2 (Why uniformly sample on $\mathbb{S}^{d-1}$ instead of $\mathbb{R}^d$?)**.** *In polar coordinates, every vector* $x\in\mathbb{R}^d$ *is described as* $x=r\theta$ *where* $r\geq 0$ *is a radius and* $\theta\in\mathbb{S}^{d-1}$ *the direction. Projection along* $r\theta$ *rescales the 1-D Wasserstein by* $|r|^p$*, so integrating over* $\mathbb{R}^d$ *repeats directional information and gives a divergent radial factor with the Lebesgue measure. Hence, we average over* $\mathbb{S}^{d-1}$ *instead.* (Rigorous justification left as an exercise!)

Remark 3 (The Sliced-Wasserstein distance *of order p*)**.** *If* $c(x,y)=|x-y|^p$ *with* $p\in[1,+\infty)$*, the resulting SW distance is denoted by* $\mathbf{SW}_p$ *and given by,*

$$\mathbf{SW}_p(\mu,\nu)=\big(\mathbb{E}_{\theta\sim\mathcal{U}_{\mathbb{S}^{d-1}}}\big[\mathbf{W}_p\big((p_\theta)_\sharp\mu,(p_\theta)_\sharp\nu\big)^p\big]\big)^{1/p}\tag{5}$$

From now on, we will focus on $\mathbf{SW}_p$. The following theoretical properties and proofs are based on results from Bonnotte [2013].

Theorem 1. *The Sliced-Wasserstein distance of order p satisfies all metric axioms (hence the name distance!)*

Proof. Let μ,ν,ξ be three arbitrary probability distributions on $\mathbb{R}^d$.

(a) Symmetry. $\mathbf{SW}_p(\mu,\nu)=\mathbf{SW}_p(\nu,\mu)$. Indeed,

$$\begin{aligned}\mathbf{SW}_p(\mu,\nu)&=\big(\mathbb{E}_{\theta\sim\mathcal{U}_{\mathbb{S}^{d-1}}}\big[\mathbf{W}_p\big((p_\theta)_\sharp\mu,(p_\theta)_\sharp\nu\big)^p\big]\big)^{1/p}\\&=\big(\mathbb{E}_{\theta\sim\mathcal{U}_{\mathbb{S}^{d-1}}}\big[\mathbf{W}_p\big((p_\theta)_\sharp\nu,(p_\theta)_\sharp\mu\big)^p\big]\big)^{1/p}\quad\text{(by the symmetry of }\mathbf{W}_p)\\&=\mathbf{SW}_p(\nu,\mu)\,.\end{aligned}$$

(b) Triangle inequality. $\mathbf{SW}_p(\mu,\nu)\leq\mathbf{SW}_p(\mu,\xi)+\mathbf{SW}_p(\xi,\nu)$. Indeed,

$$\begin{aligned}\mathbf{SW}_p(\mu,\nu)&=\big(\mathbb{E}_{\theta\sim\mathcal{U}_{\mathbb{S}^{d-1}}}\big[\mathbf{W}_p\big((p_\theta)_\sharp\mu,(p_\theta)_\sharp\nu\big)^p\big]\big)^{1/p}\\&=\big(\mathbb{E}_{\theta\sim\mathcal{U}_{\mathbb{S}^{d-1}}}\big[\big\{\mathbf{W}_p\big((p_\theta)_\sharp\mu,(p_\theta)_\sharp\xi\big)+\mathbf{W}_p\big((p_\theta)_\sharp\xi,(p_\theta)_\sharp\nu\big)\big\}^p\big]\big)^{1/p}\end{aligned}$$

since $\mathbf{W}_p$ satisfies the triangle inequality. We conclude by using Minkowski's inequality: for every real-valued random variables X and Y,

$$\mathbb{E}[|X+Y|^p]^{1/p}\leq\mathbb{E}[|X|^p]^{1/p}+\mathbb{E}[|Y|^p]^{1/p}\,.$$

(c) Identity of indiscernibles (*Séparation*) $\mathbf{SW}_p(\mu,\nu)=0\Leftrightarrow\mu=\nu$.

($\Leftarrow$) Suppose $\mu=\nu$. Then,

$$\mathbf{SW}_p(\mu,\mu)=\big(\mathbb{E}_{\theta\sim\mathcal{U}_{\mathbb{S}^{d-1}}}\big[\mathbf{W}_p\big((p_\theta)_\sharp\mu,(p_\theta)_\sharp\mu\big)^p\big]\big)^{1/p}=0\tag{6}$$

since $\mathbf{W}_p(\mu',\mu')=0$ for any probability distribution μ'.

($\Rightarrow$) Suppose $\mathbf{SW}_p(\mu,\nu) = 0$. Since $\mathbf{W}_p(\mu',\nu') \geq 0$ for any distributions μ',ν', then for almost every $\theta \in \mathbb{S}^{d-1}$,

$$\mathbf{W}_p\big((p_\theta)_\sharp\mu, (p_\theta)_\sharp\nu\big)^p = 0\,, \tag{7}$$

thus $(p_\theta)_\sharp\mu = (p_\theta)_\sharp\nu$ (since $\mathbf{W}_p$ is a metric). Taking the Fourier transform of μ,

$$\begin{aligned}
\mathcal{F}\mu(s\theta) &= \int_{\mathbb{R}^d} e^{-2i\pi s\langle\theta,x\rangle}\mathrm{d}\mu(x) && (8)\\
&= \int_{\mathbb{R}^d} e^{-2i\pi t}\mathrm{d}[(p_\theta)_\sharp\mu](t) && (9)\\
&= \mathcal{F}((p_\theta)_\sharp\mu)(s) && (10)\\
&= \mathcal{F}((p_\theta)_\sharp\nu)(s) && (11)\\
&= \mathcal{F}\nu(s\theta) && (12)
\end{aligned}$$

By injectivity of the Fourier transform, we conclude that $\mu = \nu$.

□

Proposition 1 ($\mathbf{SW}_p$ vs. $\mathbf{W}_p$)**.** *Let $p \in [1,+\infty)$. For any two distributions μ,ν on $\mathbb{R}^d$ with finite moments of order p,*

$$\mathbf{SW}_p(\mu,\nu)^p \leq c_{d,p}\mathbf{W}_p(\mu,\nu)^p$$

with

$$c_{d,p} = \frac{1}{d}\int_{\mathbb{S}^{d-1}} \|\theta\|_p^p\,\mathrm{d}\mathcal{U}_{\mathbb{S}^{d-1}}(\theta) \leq 1\,.$$

Proof. Let $\gamma \in \Gamma(\mu,\nu)$ be an optimal transport plan between μ and ν. Then, $\big((p_\theta)\otimes(p_\theta)\big)_\sharp\gamma$ is a transport plan between $(p_\theta)_\sharp\mu$ and $(p_\theta)_\sharp\nu$ *(rigorous justification is left as an exercise!)*.

Therefore,

$$\begin{aligned}
\mathbf{W}_p\big((p_\theta)_\sharp\mu, (p_\theta)_\sharp\nu\big)^p &\leq \int_{\mathbb{R}\times\mathbb{R}} |s-t|^p\mathrm{d}[\big((p_\theta)\otimes(p_\theta)\big)_\sharp\gamma](s,t) && (13)\\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\langle x-y,\theta\rangle\|^p\mathrm{d}\gamma(x,y) && (14)
\end{aligned}$$

Moreover, for any $z \in \mathbb{R}^d$,

$$\begin{aligned}
\int_{\mathbb{S}^{d-1}} |\langle z,\theta\rangle|^p\mathrm{d}\mathcal{U}_{\mathbb{S}^{d-1}}(\theta) &\leq \frac{1}{d}\|z\|^p\int_{\mathbb{S}^{d-1}} \|\theta\|_p^p\,\mathrm{d}\mathcal{U}_{\mathbb{S}^{d-1}} && (15)\\
&= c_{d,p}\|z\|^p\,. && (16)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{SW}_p(\mu,\nu)^p &\leq \int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}^d\times\mathbb{R}^d} |\langle x-y,\theta\rangle|^p\mathrm{d}\gamma(x,y)\mathrm{d}\mathcal{U}_{\mathbb{S}^{d-1}}(\theta) && (17)\\
&\leq \int_{\mathbb{R}^d\times\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} |\langle x-y,\theta\rangle|^p\mathrm{d}\mathcal{U}_{\mathbb{S}^{d-1}}(\theta)\mathrm{d}\gamma(x,y) && (18)\\
&\leq c_{d,p}\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p\mathrm{d}\gamma(x,y) && (19)\\
&\leq c_{d,p}\mathbf{W}_p(\mu,\nu)^p\,. && (20)
\end{aligned}$$

□

Remark 4 (Additional comments on Proposition 1)**.**

- *As $p \geq 2$, $c_{d,p} \leq 1/d$ (left as an exercise!)*

- $\mathbf{W}_p$ *and* $\mathbf{SW}_p$ *are equivalent for distributions supported on* $\mathcal{B}(\mathbf{0}, R) \subset \mathbb{R}^d$ *(closed Euclidean ball in* $\mathbb{R}^d$ *of center* $\mathbf{0}$ *and radius* $R > 0$*). More precisely, there exists a constant* $C_{d,p} > 0$ *such that, for any* μ, ν *supported on* $\mathcal{B}(\mathbf{0}, R)$,
$$\mathbf{SW}_p(\mu,\nu)^p \leq c_{d,p}\mathbf{W}_p(\mu,\nu)^p \leq C_{d,p}R^{p-1/(d+1)}\mathbf{SW}_p(\mu,\nu)^{1/(d+1)}\,. \tag{21}$$

2 Sliced Optimal Transport in Practice

Consider $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, with each $x_i, y_i \in \mathbb{R}^d$ (practical setting: we only observe data points).

$$\mathbf{SW}_p(\mu,\nu)^p = \mathbb{E}_{\theta\sim\mathcal{U}_{\mathbb{S}^{d-1}}}\left[\mathbf{W}_p\big((p_\theta)_\sharp\mu,(p_\theta)_\sharp\nu\big)^p\right] \tag{22}$$
$$\approx \frac{1}{K}\sum_{j=1}^K \mathbf{W}_p\big((p_{\theta_j})_\sharp\mu,(p_{\theta_j})_\sharp\nu\big)^p \tag{23}$$
$$= \widehat{\mathbf{SW}}_p(\mu,\nu)^p \tag{24}$$

Main steps to compute SW in practice:

1. **Sample:** $(\theta_j)_{j=1}^K$, which are K independent and identically distributed samples from $\mathcal{U}_{\mathbb{S}^{d-1}}$
2. **Project:** For each $j \in \{1,\dots,K\}$, compute $(p_{\theta_j})_\sharp\mu$ and $(p_{\theta_j})_\sharp\nu$
3. **Sort and average:** Compute $\mathbf{W}_p\big((p_{\theta_j})_\sharp\mu,(p_{\theta_j})_\sharp\nu\big)$

Proposition 2 (How to uniformly sample on $\mathbb{S}^{d-1}$?)**.** *Let* $X_1, X_2, \dots, X_d \sim \mathcal{N}(0,1)$ *and be independent. Then, the vector*
$$\tilde{X} = \left(\frac{X_1}{Z}, \frac{X_2}{Z}, \dots, \frac{X_d}{Z}\right)$$
is a uniform random vector on $\mathbb{S}^{d-1}$*, where* $Z = \sqrt{\sum_{i=1}^d X_i^2}$.

Proof. Let $f : \mathbb{S}^{d-1} \to \mathbb{R}$ be a bounded, continuous function. We want to show that,
$$\mathbb{E}[f(\tilde{X})] = \frac{1}{\mathcal{A}(\mathbb{S}^{d-1})}\int_{\mathbb{S}^{d-1}} f(\theta)\mathrm{d}\sigma(\theta) \tag{25}$$
where σ is the surface measure on $\mathbb{S}^{d-1}$, and $\mathcal{A}(\mathbb{S}^{d-1})$ denotes the surface area of $\mathbb{S}^{d-1}$, *i.e.*, $\mathcal{A}(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ with $\Gamma(d/2) = \int_0^{+\infty} e^{-t}t^{\frac{d}{2}-1}\mathrm{d}t$.

Using the Gaussian density and switching to spherical coordinates $x = r\theta$ with $r \in (0,+\infty)$ and $\theta \in \mathbb{S}^{d-1}$, and using $\mathrm{d}x = r^{d-1}\mathrm{d}r\mathrm{d}\sigma(\theta)$ (see details here),
$$\mathbb{E}[f(\tilde{X})] = \int_{\mathbb{R}^d} f(x_1/z, x_2/z, \dots, x_d/z)(2\pi)^{-d/2}e^{-z^2/2}\mathrm{d}x_1\mathrm{d}x_2\dots\mathrm{d}x_d \tag{26}$$
$$= (2\pi)^{-d/2}\int_0^{+\infty}\left[\int_{\mathbb{S}^{d-1}} f(\theta)\mathrm{d}\sigma(\theta)\right]e^{-r^2/2}r^{d-1}\mathrm{d}r \tag{27}$$
$$= c_d\int_{\mathbb{S}^{d-1}} f(\theta)\mathrm{d}\sigma(\theta)\,. \tag{28}$$
where $c_d = (2\pi)^{-d/2}\int_0^{+\infty} e^{-r^2/2}r^{d-1}\mathrm{d}r = \frac{1}{2\pi^{d/2}}\Gamma(d/2)$. Thus, $c_d = \frac{1}{\mathcal{A}(\mathbb{S}^{d-1})}$, and this concludes the proof. □

Remark 5 (Computational complexity)**.**

- *Computing* $\widehat{\mathbf{SW}}_p(\mu,\nu)$ *costs* $O(Kdn + Kn\log n)$ *operations (projecting + sorting).*
- *Recall that computing* $\mathbf{W}_p(\mu,\nu)$ *scales in* $O(n^3)$ *operations when* $d > 1$.

$\Rightarrow$ *Computing* $\widehat{\mathbf{SW}}_p$ *is much cheaper than* $\mathbf{W}_p$ *for* $n \gg d$ *and moderate* K

References

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

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2012.