Spherical Sliced-Wasserstein

Clément Bonet[1], Paul Berg[1], Nicolas Courty[1], François Septier[1], Lucas Drumetz[2], Minh-Tan Pham[1]

[1]Université Bretagne Sud
[2]IMT Atlantique

Journée du laboratoire
07/06/2022

Motivation

Goal of Optimal Transport: Transport mass in an optimal way

$$W_c(\mu,\nu) = \min_{T_\#\mu=\nu} \int c(x,T(x))\ \mathrm{d}\mu(x) \tag{1}$$

- Coupling
- Value of the min

Widely use nowadays in Machine Learning

- Generative Models (*e.g.* WGAN [Arjovsky et al., 2017])
- Domain Adaptation [Courty et al., 2016]
- ...

Data generally lie on manifolds, *e.g.* on the sphere $S^{d-1} = \{x \in \mathbb{R}^d,\ \|x\|_2 = 1\}$:

- Directional data, meteorology, cosmology...
- Also used as embeddings for VAEs, Self-supervised learning...

Wasserstein Distance

Definition (Wasserstein distance)

Let M be a Riemannian manifold endowed with the Riemannian distance d, $p \geq 1$, $\mu, \nu \in \mathcal{P}_p(M)$, then

$$W_p^p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int d^p(x, y) \, \mathrm{d}\gamma(x, y), \tag{2}$$

where $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(M \times M), \ \pi^1_\# \gamma = \mu, \pi^2_\# \gamma = \nu\}$ and $\pi^1(x, y) = x$, $\pi^2(x, y) = y$, $\pi^1_\# \gamma = \gamma \circ (\pi^1)^{-1}$.

Numerical approximation: Linear program $O(n^3 \log n)$ [Peyré et al., 2019]

Proposed Solutions:

- Entropic regularization + Sinkhorn $O(n^2)$ [Cuturi, 2013]
- Minibatch estimator [Fatras et al., 2020]
- Sliced-Wasserstein [Rabin et al., 2011b, Bonnotte, 2013] but only on Euclidean spaces

Sliced-Wassertein on $\mathbb{R}^d$

Wasserstein on $\mathbb{R}$:

$$\forall p \geq 1, \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}),\ W_p^p(\mu,\nu) = \int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^p \ \mathrm{d}u \tag{3}$$

Sliced-Wassertein on $\mathbb{R}^d$

Wasserstein on $\mathbb{R}$:

$$\forall p \geq 1, \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}),\ W_p^p(\mu,\nu) = \int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^p \ \mathrm{d}u \tag{3}$$

Definition (Sliced-Wasserstein [Rabin et al., 2011b])

Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$,

$$SW_p^p(\mu,\nu) = \int_{S^{d-1}} W_p^p(P_\#^\theta \mu, P_\#^\theta \nu) \ \mathrm{d}\lambda(\theta), \tag{4}$$

where $P^\theta(x) = \langle x, \theta \rangle$, λ uniform measure on S^{d-1}.

Sliced-Wassertein on $\mathbb{R}^d$

Wasserstein on $\mathbb{R}$:

$$\forall p \geq 1, \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}),\ W_p^p(\mu,\nu) = \int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^p \,\mathrm{d}u \tag{3}$$

Definition (Sliced-Wasserstein [Rabin et al., 2011b])

Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$,

$$SW_p^p(\mu,\nu) = \int_{S^{d-1}} W_p^p(P^\theta_\# \mu, P^\theta_\# \nu) \,\mathrm{d}\lambda(\theta), \tag{4}$$

where $P^\theta(x) = \langle x, \theta \rangle$, λ uniform measure on S^{d-1}.

Properties:

- Distance
- Topogically equivalent to the Wasserstein distance
- Monte-Carlo approximation in $O(Ln \log n)$

Radon Transform

Definition (Radon Transform)

Let $f \in L^1(\mathbb{R}^d)$, then the Radon transform $R : L^1(\mathbb{R}^d) \to L^1(\mathbb{R} \times S^{d-1})$ is defined as

$$\forall \theta \in S^{d-1},\ \forall t \in \mathbb{R},\ Rf(t, \theta) = \int_{\mathbb{R}^d} f(x) \mathbb{1}_{\{\langle x, \theta \rangle = t\}} \mathrm{d}x. \tag{5}$$

Definition (Back-projection operator)

The back-projection operator $R^* : C_0(\mathbb{R} \times S^{d-1}) \to C_0(\mathbb{R}^d)$ is defined as

$$\forall g \in C_0(\mathbb{R} \times S^{d-1}), \forall x \in \mathbb{R}^d,\ R^* g(x) = \int_{S^{d-1}} g(\langle x, \theta \rangle, \theta)\ \mathrm{d}\theta. \tag{6}$$

$\forall f, g,\ \langle Rf, g \rangle_{\mathbb{R} \times S^{d-1}} = \langle f, R^* g \rangle_{\mathbb{R}^d}$

Radon Transform of Measures and Link with SW

- Radon transform of a measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ defined as the measure $R\mu$ such that $\langle R\mu, g\rangle_{\mathbb{R}\times S^{d-1}} = \langle \mu, R^* g\rangle_{\mathbb{R}^d}$.
- Disintegration of $R\mu \in \mathcal{P}(\mathbb{R} \times S^{d-1})$ *w.r.t.* λ: $R\mu = \lambda \otimes K$
- [Bonneel et al., 2015, Proposition 6]: For λ-ae $\theta \in S^{d-1}$, $K(\theta,\cdot) = P^\theta_\# \mu$

$$\forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}^d),\ SW_p^p(\mu,\nu) = \int_{S^{d-1}} W_p^p\big((R\mu)^\theta, (R\mu)^\theta\big)\ \mathrm{d}\lambda(\theta). \tag{7}$$

Interest: If R injective, SW is a distance.

SW on the Sphere

Goal: defining SW discrepancy on the sphere taking care of geometry of the manifold

	SW	SSW
Closed-form of W	Line	?
Projection	$P^\theta(x) = \langle x, \theta \rangle$	?
Integration	S^{d-1}	?

Table: SW to SSW

SW on the Sphere

Goal: defining SW discrepancy on the sphere taking care of geometry of the manifold

	SW	SSW
Closed-form of W	Line	?
Projection	$P^\theta(x) = \langle x, \theta \rangle$	?
Integration	S^{d-1}	?

Table: SW to SSW

- Generalization of straight lines on manifolds: geodesics
- On S^{d-1}, geodesics = great circles

Wasserstein on the Circle

Let $\mu, \nu \in \mathcal{P}(S^1)$ where $S^1 = \mathbb{R}/\mathbb{Z}$.

- Parametrize S^1 by $[0, 1[$
- $\forall x, y \in [0, 1[,\ d_{S^1}(x, y) = \min(|x - y|, 1 - |x - y|)$
- For a cost function $c(x, y) = h(d_{S^1}(x, y))$ with $h : \mathbb{R} \to \mathbb{R}^+$ increasing and convex
- $\forall \mu, \nu \in \mathcal{P}(S^1)$, [Rabin et al., 2011a]

$$W_c(\mu, \nu) = \inf_{\alpha \in \mathbb{R}} \int_0^1 h(|F_\mu^{-1}(t) - (F_\nu - \alpha)^{-1}(t)|)\ \mathrm{d}t. \tag{8}$$

- To find α: binary search [Delon et al., 2010]

Particular Cases

- For $h = \mathrm{Id}$, [Hundrieser et al., 2021]

$$W_1(\mu,\nu) = \int_0^1 |F_\mu(t) - F_\nu(t) - \mathrm{LevMed}(F_\mu - F_\nu)| \, \mathrm{d}t, \tag{9}$$

where

$$\mathrm{LevMed}(f) = \inf \left\{ t \in \mathbb{R}, \ \beta(\{x \in [0,1[, \ f(x) \leq t\}) \geq \frac{1}{2} \right\}. \tag{10}$$

- For $h(x) = x^2$ and $\nu = \mathrm{Unif}(S^1)$,

$$W_2^2(\mu,\nu) = \int_0^1 |F_\mu^{-1}(t) - t - \hat{\alpha}|^2 \, \mathrm{d}t \quad \text{with} \quad \hat{\alpha} = \int x \, \mathrm{d}\mu(x) - \frac{1}{2}. \tag{11}$$

In particular, if $x_1 < \cdots < x_n$ and $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, then

$$W_2^2(\mu_n,\nu) = \frac{1}{n}\sum_{i=1}^n x_i^2 - \Big(\frac{1}{n}\sum_{i=1}^n x_i\Big)^2 + \frac{1}{n^2}\sum_{i=1}^n (n+1-2i)x_i + \frac{1}{12}. \tag{12}$$

Sliced-Wasserstein on the Sphere

- Great circle: Intersection between 2-plane and S^{d-1}
- Parametrize 2-plane by the Stiefel manifold

$$\mathbb{V}_{d,2} = \{U \in \mathbb{R}^{d\times 2},\ U^T U = I_2\}$$

- Projection on great circle C: For a.e. $x \in S^{d-1}$,

$$P^C(x) = \operatorname*{argmin}_{y\in C}\ d_{S^{d-1}}(x,y),$$

where $d_{S^{d-1}}(x,y) = \arccos(\langle x, y\rangle)$.

- For $U \in \mathbb{V}_{d,2}$, $C = \mathrm{span}(UU^T) \cap S^{d-1}$,

$$P^U(x) = U^T \operatorname*{argmin}_{y\in C} d_{S^{d-1}}(x,y) = \frac{U^T x}{\|U^T x\|_2}.$$

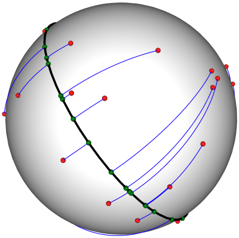


Figure: Illustration of the geodesic projections on a great circle (in black). In red, random points sampled on the sphere. In green the projections and in blue the trajectories.

Spherical Sliced-Wasserstein

Definition (Spherical Sliced-Wasserstein)

Let $p \geq 1$, $\mu, \nu \in \mathcal{P}_p(S^{d-1})$ absolutely continuous *w.r.t.* Lebesgue measure,

$$SSW_p^p(\mu, \nu) = \int_{\mathbb{V}_{d,2}} W_p^p(P^U_\# \mu, P^U_\# \nu)\ \mathrm{d}\sigma(U), \tag{13}$$

with σ the uniform distribution over $\mathbb{V}_{d,2}$.

	SW	SSW
Closed-form of W	Line	(Great)-Circle
Projection	$P^\theta(x) = \langle x, \theta \rangle$	$P^U(x) = \frac{U^T x}{\|U^T x\|_2}$
Integration	S^{d-1}	$\mathbb{V}_{d,2}$

Table: Comparison SW-SSW

A New Spherical Radon Transform

Question: Is SSW a distance?

Definition (Spherical Radon Transform)

Let $f \in L^1(S^{d-1})$, then we define a Spherical Radon transform $\tilde{R} : L^1(S^{d-1}) \to L^1(S^1 \times \mathbb{V}_{d,2})$ as

$$\forall z \in S^1,\ \forall U \in \mathbb{V}_{d,2},\ \tilde{R}f(z,U) = \int_{S^{d-1}} f(x)\mathbb{1}_{\{z=P^U(x)\}}\mathrm{d}x. \tag{14}$$

Definition (Back-projection operator)

The back-projection operator $R^* : C_0(S^1 \times \mathbb{V}_{d,2}) \to C_b(S^{d-1})$ is defined as for a.e. $x \in S^{d-1}$,

$$\tilde{R}^* g(x) = \int_{\mathbb{V}_{d,2}} g(P^U(x), U)\ \mathrm{d}\sigma(U). \tag{15}$$

For all $f \in L^1(S^{d-1})$, $g \in C_0(S^1 \times \mathbb{V}_{d,2})$,

$$\langle \tilde{R}f, g\rangle_{S^1 \times \mathbb{V}_{d,2}} = \langle f, \tilde{R}^* g\rangle_{S^{d-1}}. \tag{16}$$

Spherical Radon Transform of Measures

Define the Radon transform of $\mu \in \mathcal{M}(S^{d-1})$ as $\tilde{R}\mu$ such that

$$\forall g \in C_0(S^1 \times \mathbb{V}_{d,2}),\ \int_{S^1 \times \mathbb{V}_{d,2}} g(z,U)\ \mathrm{d}(\tilde{R}\mu)(z,U) = \int_{S^{d-1}} \tilde{R}^* g(x)\ \mathrm{d}\mu(x). \quad (17)$$

Disintegration: $\tilde{R}\mu = \sigma \otimes (\tilde{R}\mu)^U$

Proposition

Let $\mu \in \mathcal{M}(S^{d-1})$, then for σ-almost every $U \in \mathbb{V}_{d,2}$, $(\tilde{R}\mu)^U = P^U_{\#}\mu$.

$$\forall \mu, \nu \in \mathcal{P}_{p,ac}(S^{d-1}),\ SSW_p^p(\mu,\nu) = \int_{\mathbb{V}_{d,2}} W_p^p\big((\tilde{R}\mu)^U, (\tilde{R}\nu)^U\big)\ \mathrm{d}\sigma(U). \quad (18)$$

SSW distance?

- $\forall \mu, \nu \in \mathcal{P}_p(S^{d-1})$, $SSW_p(\mu, \nu) \geq 0$, $SSW_p(\mu, \nu) = SSW_p(\nu, \mu)$, $\mu = \nu \implies SSW_p(\mu, \nu) = 0$
- Triangular inequality: $\forall \mu, \nu, \alpha \in \mathcal{P}_p(S^{d-1})$,

$$\begin{aligned} SSW_p(\mu, \nu) &= \left(\int_{\mathbb{V}_{d,2}} W_p^p(P^U_\# \mu, P^U_\# \nu) \ \mathrm{d}\sigma(U) \right)^{\frac{1}{p}} \\ &\leq \left(\int_{\mathbb{V}_{d,2}} \big(W_p(P^U_\# \mu, P^U_\# \alpha) + W_p(P^U_\# \alpha, P^U_\# \nu) \big)^p \ \mathrm{d}\sigma(U) \right)^{\frac{1}{p}} \\ &\leq \left(\int_{\mathbb{V}_{d,2}} W_p^p(P^U_\# \mu, P^U_\# \alpha) \ \mathrm{d}\sigma(U) \right)^{\frac{1}{p}} \\ &+ \left(\int_{\mathbb{V}_{d,2}} W_p^p(P^U_\# \alpha, P^U_\# \nu) \ \mathrm{d}\sigma(U) \right)^{\frac{1}{p}} \\ &= SSW_p(\mu, \alpha) + SSW_p(\alpha, \nu). \end{aligned} \tag{19}$$

- Distance if $SSW_p(\mu, \nu) = 0 \implies \mu = \nu$, *i.e.* for σ-ae U, $(\tilde{R}\mu)^U = (\tilde{R}\nu)^U \implies \mu = \nu$, *i.e.* $\tilde{R}$ injective

Properties of the Spherical Radon Transform

- $\tilde{R}$ related to Hemispherical transform $\mathcal{H}$ [Rubin, 2003] on S^{d-2}, where for $f \in L^1(S^{d-2})$,

$$\forall x \in S^{d-2}, \mathcal{H}f(x) = \int_{S^{d-2}} f(y)\mathbb{1}_{\{\langle x,y\rangle > 0\}}\,\mathrm{d}y. \tag{20}$$

Proposition

$$\ker(\tilde{R}) = \{\mu \in \mathcal{M}_{\mathrm{even}}(S^{d-1}),\ \forall H \in \mathcal{G}_{d,d-1},\ \mu(H \cap S^{d-1}) = 0\},$$

where $\mu \in \mathcal{M}_{\mathrm{even}}$ if for all $f \in C(S^{d-1})$, $\langle \mu, f\rangle = \langle \mu, f_+\rangle$ with for all x, $f_+(x) = (f(x) + f(-x))/2$.

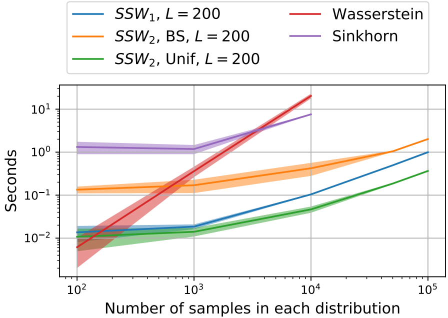
Proposition

Let $p \geq 1$, SSW_p is a pseudo-distance on $\mathcal{P}_p(S^{d-1})$.

Runtime Comparisons

Method	Complexity
Wasserstein + LP	$O(n^3 \log n)$
Sinkhorn	$O(n^2)$
SSW_2 + BS	$O(L(n+m)(d+\log(\frac{1}{\epsilon})) + Ln\log n + Lm\log m)$
SSW_1	$O(L(n+m)(d+\log(n+m)))$
SSW_2+Unif	$O(Ln(d+\log n))$

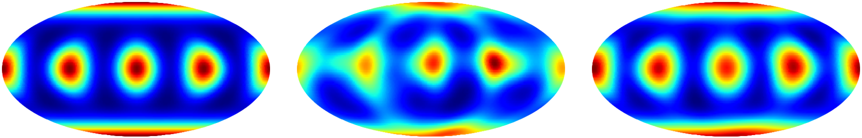
Table: Complexity



Goal:

$$\underset{\mu}{\operatorname{argmin}}\ SSW_2^2(\mu,\nu),$$

where we have access to ν through samples, *i.e.* $\hat{\nu}_m = \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$ with $(y_j)_j$ i.i.d samples of ν.



(a) Target: Mixture of vMF (b) KDE estimate of the MLP (c) KDE estimate of 500 particles

Figure: Minimization of SSW with respect to a mixture of vMF.

Variational Inference

Goal:

$$\underset{\mu}{\operatorname{argmin}}\ SSW_2^2(\mu,\nu),$$

where we know the density of ν up to a constant.

Algorithm SWVI [Yi and Liu, 2021]

Input: V a potential, K the number of iterations of SWVI, N the batch size, ℓ the number of MCMC steps
Initialization: Choose q_θ a sampler
for $k=1$ **to** K **do**
  Sample $(z_i^0)_{i=1}^N \sim q_\theta$
  Run ℓ MCMC steps starting from $(z_i^0)_{i=1}^N$ to get $(z_j^\ell)_{j=1}^N$
  // Denote $\hat{\mu}_0 = \frac{1}{N}\sum_{j=1}^N \delta_{z_j^0}$ and $\hat{\mu}_\ell = \frac{1}{N}\sum_{j=1}^N \delta_{z_j^\ell}$
  Compute $J = SW_2^2(\hat{\mu}_0, \hat{\mu}_\ell)$
  Backpropagate through J *w.r.t.* θ
  Perform a gradient step
end for

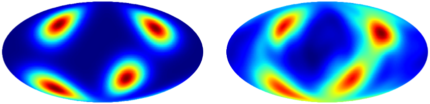
Variational Inference

Goal:

$$\underset{\mu}{\operatorname{argmin}} \; SSW_2^2(\mu,\nu),$$

where we know the density of ν up to a constant.

- Use SSW instead of SW
- Use Normalizing flows + MCMC on the sphere



(a) Target distribution (b) Density learned

Figure: Amortized SSWVI with a normalizing flow *w.r.t.* a mixture of vMF.

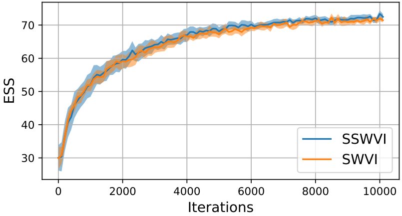


Figure: Comparison of the ESS between SWVI et SSWVI with the mixture target (mean over 10 runs).

Wasserstein Autoencoders

Figure: Autoencoder with spherical latent space.

SSWAE:

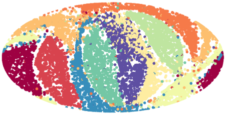
$$\mathcal{L}(f,g) = \int c\big(x, g(f(x))\big)\mathrm{d}\mu(x) + \lambda SSW_2^2(f_\#\mu, p_Z), \tag{21}$$

Much interest in using a spherical latent space [Davidson et al., 2018, Xu and Durrett, 2018], *e.g.* uniform.

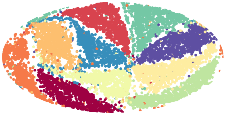
Wasserstein Autoencoders

SSWAE:

$$\mathcal{L}(f,g) = \int c\big(x, g(f(x))\big)\mathrm{d}\mu(x) + \lambda SSW_2^2(f_\#\mu, p_Z), \tag{22}$$



(a) SWAE



(b) SSWAE

Figure: Latent space of SWAE and SSWAE for a uniform prior on S^2.

Table: FID (Lower is better).

Method / Prior	Unif(S^{10})
SSWAE	**14.91 ± 0.32**
SWAE	15.18 ± 0.32
WAE-MMD IMQ	18.12 ± 0.62
WAE-MMD RBF	20.09 ± 1.42
SAE	19.39 ± 0.56
Circular GSWAE	15.01 ± 0.26

Wasserstein Autoencoders

SSWAE:

$$\mathcal{L}(f,g) = \int c\big(x, g(f(x))\big)\mathrm{d}\mu(x) + \lambda SSW_2^2(f_{\#}\mu, p_Z), \tag{23}$$

(a) SSWAE (b) SWAE (c) SAE

Figure: Samples generated with Sliced-Wasserstein Autoencoders with a uniform prior on S^{10}.

Conclusion

- First SW discrepancy on manifolds
- Good performance on ML tasks

Future works

- SW on hyperbolic spaces
- Statistical analysis

Conclusion

- First SW discrepancy on manifolds
- Good performance on ML tasks

Future works

- SW on hyperbolic spaces
- Statistical analysis

Thank you!

References I

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

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

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

References II

Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical variational auto-encoders. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 856–865. AUAI Press, 2018. URL http://auai.org/uai2018/proceedings/papers/309.pdf.

Julie Delon, Julien Salomon, and Andrei Sobolevski. Fast transport optimization for monge costs on the circle. *SIAM Journal on Applied Mathematics*, 70(7): 2239–2258, 2010.

Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein : asymptotic and gradient properties. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2131–2141. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/fatras20a.html.

Shayan Hundrieser, Marcel Klatt, and Axel Munk. The statistics of circular optimal transport. *arXiv preprint arXiv:2103.15426*, 2021.

References III

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Julien Rabin, Julie Delon, and Yann Gousseau. Transportation distances on the circle. *Journal of Mathematical Imaging and Vision*, 41(1):147–167, 2011a.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011b.

Boris Rubin. Notes on radon transforms in integral geometry. *Fractional Calculus and Applied Analysis*, 6(1):25–72, 2003.

Jiacheng Xu and Greg Durrett. Spherical latent spaces for stable variational autoencoders. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4503–4513. Association for Computational Linguistics, 2018.

Mingxuan Yi and Song Liu. Sliced wasserstein variational inference. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2021.