Spherical Sliced-Wasserstein

Clément Bonet[1], Paul Berg[2], Nicolas Courty[2], François Septier[1], Lucas Drumetz[3], Minh-Tan Pham[2]

[1]Université Bretagne Sud, LMBA
[2]Université Bretagne Sud, IRISA
[3]IMT Atlantique, Lab-STICC

CAp
03/07/2023

Motivation

Optimal Transport widely use nowadays in Machine Learning

- Domain Adaptation [Courty et al., 2016]
- Generative Models (*e.g.* WGAN [Arjovsky et al., 2017])
- Document Classification [Kusner et al., 2015]
- ...

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 on the Sphere

- Sphere: $S^{d-1} = \{x \in \mathbb{R}^d,\ \|x\|_2 = 1\}$
- Geodesic distance: $\forall x, y \in S^{d-1},\ d(x,y) = \arccos(\langle x, y\rangle)$

Definition (Wasserstein distance)

Let $p \geq 1$, $\mu, \nu \in \mathcal{P}_p(S^{d-1})$, then

$$W_p^p(\mu,\nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int d(x,y)^p \ \mathrm{d}\gamma(x,y), \tag{1}$$

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

Wasserstein Distance on the Sphere

Let $\mu, \nu \in \mathcal{P}_p(S^{d-1})$, $x_1, \dots, x_n \sim \mu$, $y_1, \dots, y_n \sim \nu$, $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$.

Numerical computation with plug-in estimator: Linear program

$$W_p^p(\hat{\mu}_n, \hat{\nu}_n) = \min_{\gamma \in \Pi(\hat{\mu}_n, \hat{\nu}_n)} \langle C, \gamma \rangle, \tag{2}$$

with $C = \big(d(x_i, y_j)\big)_{i,j}$.

Complexity: $O(n^3 \log n)$ [Peyré et al., 2019]

Wasserstein Distance on the Sphere

Let $\mu, \nu \in \mathcal{P}_p(S^{d-1})$, $x_1, \dots, x_n \sim \mu$, $y_1, \dots, y_n \sim \nu$, $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$.

Numerical computation with plug-in estimator: Linear program

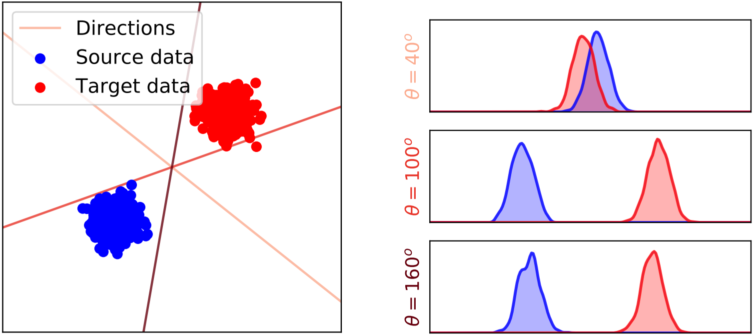
$$W_p^p(\hat{\mu}_n, \hat{\nu}_n) = \min_{\gamma \in \Pi(\hat{\mu}_n, \hat{\nu}_n)} \langle C, \gamma \rangle, \tag{2}$$

with $C = \big(d(x_i, y_j)\big)_{i,j}$.

Complexity: $O(n^3 \log n)$ [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



(a) Samples and directions

(b) One dimensional densities

Figure: Illustration of the projection of distributions on different lines.

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 $\mathbb{R}^d$

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
- Topologically equivalent to the Wasserstein distance [Nadjahi et al., 2019]
- Monte-Carlo approximation in $O\big(Ln(\log n + d)\big)$

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

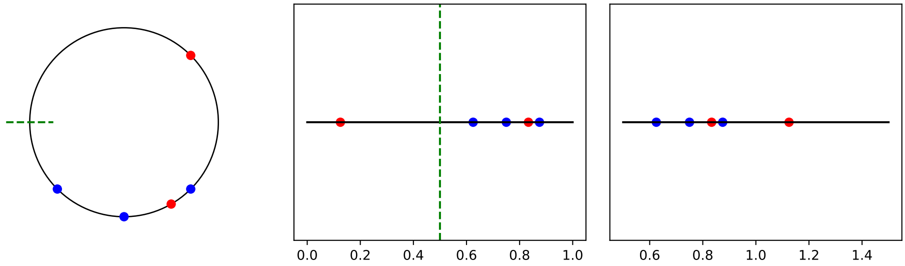
Wasserstein on the Circle

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

- Parametrize S^1 by $[0, 1[$
- $\forall x, y \in [0, 1[,\ d_{S^1}(x, y) = \min(|x - y|, 1 - |x - y|)$
- $\forall \mu, \nu \in \mathcal{P}(S^1)$, [Rabin et al., 2011a]

$$W_p^p(\mu, \nu) = \inf_{\alpha \in \mathbb{R}} \int_0^1 |F_\mu^{-1}(t) - (F_\nu - \alpha)^{-1}(t)|^p \ \mathrm{d}t. \quad (5)$$

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



Particular Cases

- For $p = 1$, [Hundrieser et al., 2021]

$$W_1(\mu,\nu) = \int_0^1 |F_\mu(t) - F_\nu(t) - \mathrm{LevMed}(F_\mu - F_\nu)| \,\mathrm{d}t, \tag{6}$$

where

$$\mathrm{LevMed}(f) = \inf\left\{t \in \mathbb{R},\ \mathrm{Leb}(\{x \in [0,1[,\ f(x) \leq t\}) \geq \frac{1}{2}\right\}. \tag{7}$$

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

$$W_2^2(\mu,\nu) = \int_0^1 |F_\mu^{-1}(t) - t - \hat{\alpha}|^2 \,\mathrm{d}t \quad \text{with} \quad \hat{\alpha} = \int x \,\mathrm{d}\mu(x) - \frac{1}{2}. \tag{8}$$

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

$$W_2^2(\mu_n,\nu) = \frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2 + \frac{1}{n^2}\sum_{i=1}^n (n+1-2i)x_i + \frac{1}{12}. \tag{9}$$

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

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

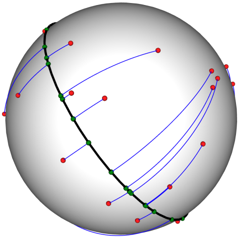


Figure: Illustration of the geodesic projections on a great circle (in black). In red, random points sampled on the sphere. In green the projections and in blue the trajectories.

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,

$$SSW_p^p(\mu,\nu) = \int_{\mathbb{V}_{d,2}} W_p^p(P^U_\# \mu, P^U_\# \nu)\ \mathrm{d}\sigma(U), \tag{10}$$

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

Is SSW a Distance?

Question: Is SSW a distance?

Proposition

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

- Lacking property (for now): indiscernibility property, *i.e.* $SSW_p(\mu,\nu) = 0 \implies \mu = \nu$.
- Need to show that $P^U_\# \mu = P^U_\# \nu$ for σ-ae $U \in \mathbb{V}_{d,2}$ implies $\mu = \nu$.
- Idea: relate P^U to a well chosen (injective) Radon transform which integrates along $\{x \in S^{d-1},\ P^U(x) = z\}$ for $U \in \mathbb{V}_{d,2}$ and $z \in S^1$.

Projections Sets

Proposition

Let $U \in \mathbb{V}_{d,2}$*,* $z \in S^1$*. The projection set on* $z \in S^1$ *is*

$$\{x \in S^{d-1},\ P^U(x) = z\} = \{x \in F \cap S^{d-1},\ \langle x, Uz \rangle > 0\}, \tag{11}$$

where $F = \mathrm{span}(UU^T)^\perp \oplus \mathrm{span}(Uz)$*.*

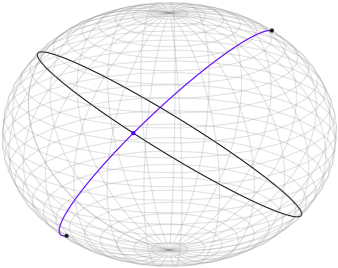


Figure: The set of projection on the blue point $Uz \in \mathrm{span}(UU^T) \cap S^{d-1}$ is plotted in blue.

A Spherical Radon Transform

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

$$\forall z \in S^1,\ \forall U \in \mathbb{V}_{d,2},\ \tilde{R}f(z,U) = \int_{S^{d-1}} f(x)\ \mathrm{d}\sigma_d^z(x), \tag{12}$$

with σ_d^z a suitable measure on $\{x \in S^{d-1},\ P^U(x) = z\}$.

Results on the injectivity of $\tilde{R}$ so far:

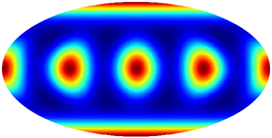
- In our work: linked it with the Hemispherical Radon transform studied in [Rubin, 1999]
- In [Quellmalz et al., 2023]: showed that it a distance on S^2

Gradient Flows

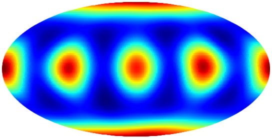
Goal:

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

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 500 particles

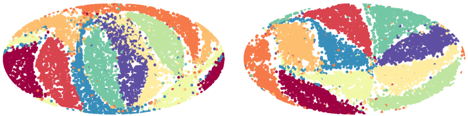
Figure: Minimization of SSW with respect to a mixture of vMF.

Wasserstein Autoencoders

Autoencoder with spherical latent space [Davidson et al., 2018, Xu and Durrett, 2018]

SSWAE:

$$\mathcal{L}(f,g) = \int c\big(x, g(f(x))\big)\mathrm{d}\mu(x) + \lambda SSW_2^2(f_\#\mu, p_Z), \quad (13)$$



(a) SWAE (b) SSWAE

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

Table: FID on MNIST (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

Density Estimation

Goal: learn a normalizing flow T such that $T_\# \mu = p_Z$ with $p_Z = \mathrm{Unif}(S^{d-1})$:

$$\underset{T}{\mathrm{argmin}}\ SSW_2^2(T_\# \mu, p_Z), \tag{14}$$

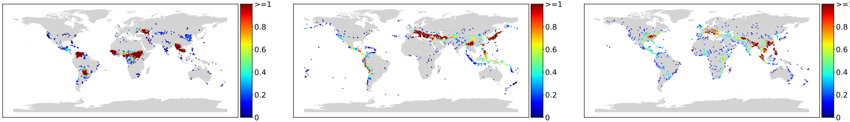
where we have access to μ through samples.

Density:

$$\forall x \in S^{d-1},\ f_\mu(x) = p_Z(T(x)) |\det J_T(x)|. \tag{15}$$

Table: Negative test log likelihood.

	Earthquake	Flood	Fire
SSW	$\mathbf{0.84}_{\pm \mathbf{0.07}}$	$\mathbf{1.26}_{\pm \mathbf{0.05}}$	$\mathbf{0.23}_{\pm \mathbf{0.18}}$
SW	$0.94_{\pm 0.02}$	$1.36_{\pm 0.04}$	$0.54_{\pm 0.37}$
Stereo	$1.91_{\pm 0.1}$	$2.00_{\pm 0.07}$	$1.27_{\pm 0.09}$



(a) Fire (b) Earthquake (c) Flood

Figure: Density estimation of models trained on earth data. We plot the density on the test data.

Conclusion

Conclusion

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

Perspectives and follow-up works:

- Study statistical properties
- Try other Spherical Sliced-Wasserstein discrepancies via other Radon transforms
- Study other Riemannian manifolds: Hyperbolic spaces [Bonet et al., 2022], SPDs [Bonet et al., 2023]
- Implemented in POT [Flamary et al., 2021]

Conclusion

Conclusion

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

Perspectives and follow-up works:

- Study statistical properties
- Try other Spherical Sliced-Wasserstein discrepancies via other Radon transforms
- Study other Riemannian manifolds: Hyperbolic spaces [Bonet et al., 2022], SPDs [Bonet et al., 2023]
- Implemented in POT [Flamary et al., 2021]

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.

Clément Bonet, Laetitia Chapel, Lucas Drumetz, and Nicolas Courty. Hyperbolic sliced-wasserstein via geodesic and horospherical projections. *arXiv preprint arXiv:2211.10066*, 2022.

Clément Bonet, Benoît Malézieux, Alain Rakotomamonjy, Lucas Drumetz, Thomas Moreau, Matthieu Kowalski, and Nicolas Courty. Sliced-wasserstein on symmetric positive definite matrices for meeg signals. 2023.

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.

References III

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *The Journal of Machine Learning Research*, 22(1):3571–3578, 2021.

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

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Michael Quellmalz, Robert Beinert, and Gabriele Steidl. Sliced optimal transport on the sphere. *arXiv preprint arXiv:2304.09092*, 2023.

References IV

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. Inversion and characterization of the hemispherical transform. *Journal d'Analyse Mathématique*, 77(1):105–128, 1999.

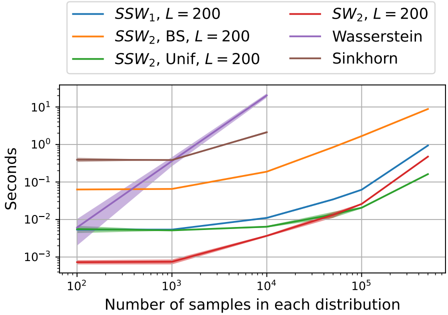
Jiacheng Xu and Greg Durrett. Spherical latent spaces for stable variational autoencoders. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November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

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



Wasserstein Autoencoders

Figure: Autoencoder with spherical latent space.

SSWAE:

$$\mathcal{L}(f,g) = \int c\big(x, g(f(x))\big)\mathrm{d}\mu(x) + \lambda SSW_2^2(f_\#\mu, p_Z), \tag{16}$$

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

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

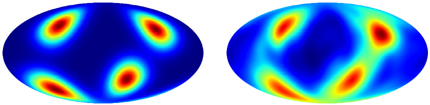
Variational Inference

Goal:

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

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.

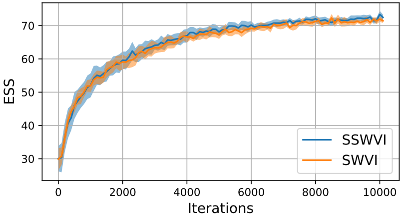


Figure: Comparison of the ESS between SWVI et SSWVI with the mixture target (mean over 10 runs).