Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance

Giulia Luise [1] Alessandro Rudi [2] Massimiliano Pontil [1,3] Carlo Ciliberto [1]

[1]Department of Computer Science, University College London, London, UK.
[2]INRIA- Sierra-Project team ENS, Paris, France.
[3]Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genova, Italy.

Abstract

Applications of optimal transport have recently received remarkable attention thanks to the computational advantages of entropic regularization. However, in most situations the Sinkhorn approximation of the Wasserstein distance is replaced by a regularized version that is less accurate but easy to differentiate. In this work we characterize the differential properties of the original Sinkhorn distance, proving that it enjoys the same smoothness as its regularized version and we explicitly provide an efficient algorithm to compute its gradient. We show that this result benefits both theory and applications: on one hand, high order smoothness confers statistical guarantees to learning with Wasserstein approximations. On the other hand, the gradient formula allows us to efficiently solve learning and optimization problems in practice.

1. BACKGROUND: Entropic regularizations of Wasserstein distance

Optimal transport theory investigates how to compare probability measures over a metric space X.

Discrete Setting: discrete probability measures of the form $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ with $(x_i)_{i=1}^n$ and the vector weight $a = (a_1, \dots, a_n)^\top \in \Delta_n$ in the n-dimensional simplex $\Delta_n = \left\{ p \in \mathbb{R}_+^n \;\middle|\; p^\top \mathbb{1}_n = 1 \right\}$. Given $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$, the Wasserstein distance between μ and ν is defined as

$$\mathrm{W}_p^p(\mu,\nu) = \min_{T \in \Pi(a,b)} \langle T, M \rangle \tag{1}$$

where $M \in \mathbb{R}^{n \times m}$ is the *cost matrix* with entries $M_{ij} = \mathsf{d}(x_i, y_j)^p$ and $\Pi(a,b)$ denotes the *transportation polytope*

$$\Pi(a,b) = \left\{ T \in \mathbb{R}_+^{n \times m} \;\middle|\; T\mathbb{1}_m = a, \quad T^\top \mathbb{1}_n = b \right\}.$$

Regularization of Wasserstein distance

Set $h(T) := -\sum_{i,j=1}^{n,m} T_{ij}(\log T_{ij} - 1)$ and $T_\lambda = \operatorname{argmin}_{T \in \Pi(a,b)} \langle T, M \rangle - \frac{1}{\lambda} h(T)$

Definition Given μ and ν as above, entropic regularizations of the Wasserstein distance, referred to as Sinkhorn distances [2] are defined as

$$\widetilde{\mathrm{S}}_\lambda(a,b) = \langle T_\lambda, M \rangle - \frac{1}{\lambda} h(T_\lambda) \quad \text{and} \quad \mathrm{S}_\lambda = \langle T_\lambda, M \rangle. \tag{2}$$

Proposition *Let $\lambda > 0$. For any pair of discrete measures $\mu, \nu \in \mathcal{P}(X)$ with respective weights $a \in \Delta_n$ and $b \in \Delta_m$, we have*

$$\left| \mathrm{S}_\lambda(\mu,\nu) - W(\mu,\nu) \right| \le c_1 e^{-\lambda} \qquad \left| \widetilde{\mathrm{S}}_\lambda(\mu,\nu) - W(\mu,\nu) \right| \le c_2 \lambda^{-1},$$

where c_1, c_2 are constants independent of λ, depending on the support of μ and ν.

Question: Is S_λ a more natural approximation of the Wasserstein distance W?

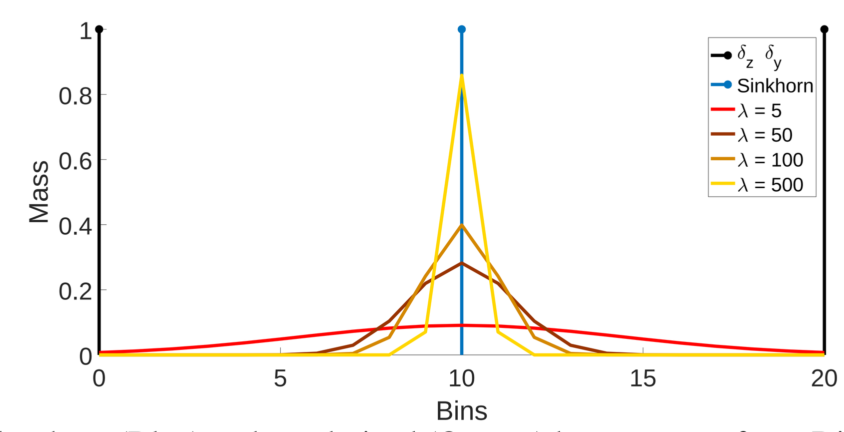

Figure: Comparison of the sharp (Blue) and regularized (Orange) barycenters of two Dirac's deltas (Black) centered in 0 and 20 for different values of λ.

2. DIFFERENTIAL PROPERTIES

Sinkhorn maps are C^∞ smooth in the interior of Δ_n.
The results are obtained leveraging the Implicit Function Theorem via a proof technique analogous to that in [3].

Theorem *For any $\lambda > 0$, the Sinkhorn distances $\widetilde{\mathrm{S}}_\lambda$ and $\mathrm{S}_\lambda : \Delta_n \times \Delta_n \to \mathbb{R}$ are C^∞ in the interior of their domain.*

Sketch of the proof. *Step 1:* since S_λ and $\widetilde{\mathrm{S}}_\lambda$ are smooth as functions of T^λ, it is enough to show that T^λ is a smooth function of a, b.
Step 2: by Sinkhorn's scaling theorem, $T^\lambda = \operatorname{diag}(e^{\lambda\alpha^\lambda}) e^{-\lambda M} \operatorname{diag}(e^{\lambda\beta^\lambda})$. Then, T^λ is smooth if $(\alpha^\lambda, \beta^\lambda)$ is smooth as a function of (a,b).
Step 3: let us set

$$\mathcal{L}(a,b;\alpha,\beta) = \alpha^\top a + \beta^\top b - \frac{1}{\lambda} \sum_{i,j=1}^{n,m} e^{-\lambda(M_{ij} - \alpha_i - \beta_j)} \tag{3}$$

and recall that $(\alpha^\lambda, \beta^\lambda) = \operatorname{argmax}_{\alpha,\beta} \mathcal{L}(a,b;\alpha,\beta)$. The smoothness of $(\alpha^\lambda, \beta^\lambda)$ is proved using the Implicit Function theorem and follows from the smoothness and strong convexity in α, β of the function $\mathcal{L}$. ■

The Implicit Function Theorem also provides a formula for the gradient of S_λ:

Input: $a \in \Delta_n$, $b \in \Delta_m$, cost matrix $M \in \mathbb{R}_+^{n,m}$, $\lambda > 0$.

$T = \textsc{Sinkhorn}(a, b, M, \lambda), \qquad \bar{T} = T_{1:n,1:(m-1)}$
$L = T \odot M, \qquad \bar{L} = L_{1:n,1:(m-1)}$
$D_1 = \operatorname{diag}(T\mathbb{1}_m), \quad D_2 = \operatorname{diag}(\bar{T}^\top \mathbb{1}_n)^{-1}$
$H = D_1 - \bar{T} D_2 \bar{T}^\top, \quad \mathrm{f} = -L\mathbb{1}_m + \bar{T} D_2 \bar{L}^\top \mathbb{1}_n,$
$\mathrm{g} = H^{-1} \mathrm{f}$

Return: $\mathrm{g} - \mathbb{1}_n(\mathrm{g}^\top \mathbb{1}_n)$

Algorithm 1: Gradient of S_λ

The routine for the gradient is used to implement optimization problems with S_λ as loss. While solutions of optimization with $\widetilde{\mathrm{S}}_\lambda$ are often 'blurry', S_λ preserves the *sharpness* of the data.

In the following, the barycenters computed with S_λ and $\widetilde{\mathrm{S}}_\lambda$ constitute an example of how S_λ is not affected by the same oversmoothing effect as $\widetilde{\mathrm{S}}_\lambda$.

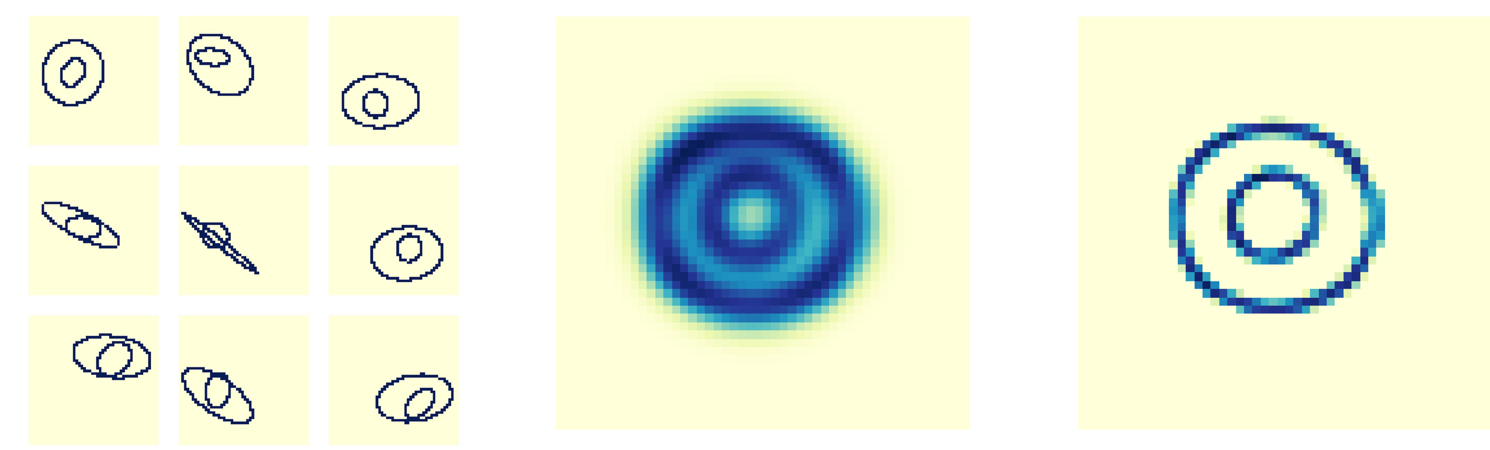

Figure: Barycenter of Nested Ellipses: (Left) Sample input data. (Middle) Barycenter with $\widetilde{\mathrm{S}}_\lambda$. (Right) Barycenter with S_λ.

3. LEARNING WITH SINKHORN LOSS: SETTING

Problem Setting: $\mathcal{X}$ input space, $\mathcal{Y} = \Delta_n$ a set of histograms (output space).
Goal: approximate a minimizer of the *expected risk*

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f), \qquad \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{S}(f(x), y)\, d\rho(x,y) \tag{4}$$

given a finite number of training points $(x_i, y_i)_{i=1}^\ell$ independently sampled from the unknown distribution ρ on $\mathcal{X} \times \mathcal{Y}$. The loss function $\mathcal{S} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ measures prediction errors and in our setting corresponds to either S_λ or $\widetilde{\mathrm{S}}_\lambda$.

Structured Prediction Estimator. Given a training set $(x_i, y_i)_{i=1}^\ell$, we consider $\hat{f} : \mathcal{X} \to \mathcal{Y}$ the structured prediction estimator proposed in [1], defined as

$$\hat{f}(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \sum_{i=1}^{\ell} \alpha_i(x)\, \mathcal{S}(y, y_i) \tag{5}$$

for any $x \in \mathcal{X}$.

The weights $\alpha_i(x)$ are learned from the data and can be interpreted as scores suggesting the candidate output distribution y to be close to a specific output distribution y_i observed in training *according to the metric $\mathcal{S}$*.

α are obtained via a kernel-based approach: given a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we have

$$\alpha(x) = (\alpha_1(x), \dots, \alpha(x))^\top = (K + \gamma\ell I)^{-1} K_x \tag{6}$$

where $\gamma > 0$ is a regularization parameter while $K \in \mathbb{R}^{\ell\times\ell}$ and $K_x \in \mathbb{R}^n$ are respectively the empirical kernel matrix with entries $K_{ij} = k(x_i, x_j)$ and the evaluation vector with entries $(K_x)_i = k(x, x_i)$, for any $i, j = 1, \dots, \ell$.

4. LEARNING WITH SINKHORN LOSS: STATISTICAL ANALYSIS

Thanks to the smoothness of S_λ we can prove consistency and learning rates of the estimator. Thanks to the gradient we can solve the problem in practice!

Theorem[Universal Consistency] *Let $\mathcal{Y} = \Delta_n^\epsilon$, $\lambda > 0$ and $\mathcal{S}$ be either $\widetilde{\mathrm{S}}_\lambda$ or S_λ. Let k be a bounded continuous universal kernel on $\mathcal{X}$. For any $\ell \in \mathbb{N}$ and any distribution ρ on $\mathcal{X}\times\mathcal{Y}$ let $\hat{f}_\ell : \mathcal{X} \to \mathcal{Y}$ be the estimator in (5) trained with $(x_i, y_i)_{i=1}^\ell$ points independently sampled from ρ and $\gamma_\ell = \ell^{-1/4}$. Then*

$$\lim_{\ell\to\infty} \mathcal{E}(\hat{f}_\ell) = \min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f) \quad \textit{with probability } 1. \tag{7}$$

Theorem[Learning Rates](informal) *Let $\mathcal{Y} = \Delta_n^\epsilon$, $\lambda > 0$ and $\mathcal{S}$ be either $\widetilde{\mathrm{S}}_\lambda$ or S_λ. Under suitable regularity assumption on ρ, the estimator $\hat{f}_\ell : \mathcal{X} \to \mathcal{Y}$ with ℓ training points independently sampled from ρ and with $\gamma = \ell^{-1/2}$*

$$\mathcal{E}(f) - \min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f) = O(\ell^{-1/4}) \tag{8}$$

holds with high probability with respect to the sampling of training data.

Role of the smoothness in the statistical analysis: the proof is technical but a key role is played by the smoothness of Sinkhorn maps shown before. This is the first universal consistency result for learning with Sinkhorn loss!

5. EXPERIMENTS

We evaluated the Sinkhorn distances with the estimator (5) in an image reconstruction problem: given an image depicting a drawing, the goal is to learn how to reconstruct the lower half of the image (output) given the upper half (input).

Reconstruction Error on QuickDraw (%)

# Classes	S_λ	$\widetilde{\mathrm{S}}_\lambda$	Hell	KDE
2	$\mathbf{3.7 \pm 0.6}$	4.9 ± 0.9	8.0 ± 2.4	12.0 ± 4.1
4	$\mathbf{22.2 \pm 0.9}$	31.8 ± 1.1	29.2 ± 0.8	40.8 ± 4.2
10	$\mathbf{38.9 \pm 0.9}$	44.9 ± 2.5	48.3 ± 2.4	64.9 ± 1.4

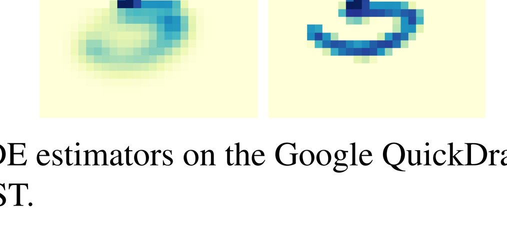

Figure: Average reconstruction errors of the Sinkhorn, Hellinger, and KDE estimators on the Google QuickDraw reconstruction problem. On the right, example of reconstruction on MNIST.

REFERENCES

[1] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. "A Consistent Regularization Approach for Structured Prediction". In: *Advances in Neural Information Processing Systems 29*.

[2] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., 2013, pp. 2292–2300.

[3] Rémi Flamary et al. "Wasserstein discriminant analysis". In: *Machine Learning* (2018).