Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance

Giulia Luise [1] Alessandro Rudi [2] Massimiliano Pontil [1,3] Carlo Ciliberto [1,4]

[1]Department of Computer Science, University College of London, London, UK. [2]INRIA - Sierra-Project team, École Normale Supérieure, PSL Research, Paris, France
[3]Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genova, Italy. [4]Department of Electrical and Electronic Engineering, Imperial College London, UK.

CONTRIBUTIONS IN A NUTSHELL

Goal: In a supervised learning setting, learn functions $f : \mathcal{X} \to \mathcal{Y}$ where the **output** space $\mathcal{Y}$ is a set of discrete **probability distributions**.

Tool: Optimal Transport
Approach: Use Sinkhorn approximations as loss functions

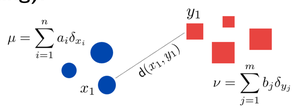
Contributions:
- Characterise the **differential properties** of Sinkhorn approximations.
- Provide learning bounds for learning with Sinkhorn loss(es), adopting a **structured prediction** perspective.

BACKGROUND

Optimal transport theory compares probability measures over a metric space.

Wasserstein distance (discrete setting):

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



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

$$\Pi(a,b) = \{T \in \mathbb{R}_+^{n\times m} \;\Big|\; T1_m = a, \quad T^\top 1_n = b\Big\}.$$

Regularization of Wasserstein distance

Definition Given μ and ν as above, entropic regularizations of the Wasserstein distance, referred to as Sinkhorn distances [1] are defined as

$$\bar{\mathsf{S}}_\lambda(a,b) \;=\; \langle T_\lambda, M\rangle - \frac{1}{\lambda}h(T_\lambda) \quad \text{and} \quad \mathsf{S}_\lambda(a,b) \;=\; \langle T_\lambda, M\rangle\,,$$

where

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

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*

$$\big|\, \mathsf{S}_\lambda(\mu,\nu) - \mathsf{W}(\mu,\nu)\,\big| \le c_1\, e^{-\lambda} \qquad \big|\, \bar{\mathsf{S}}_\lambda(\mu,\nu) - \mathsf{W}(\mu,\nu)\,\big| \le c_2/lambda,$$

with c_1, c_2 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?

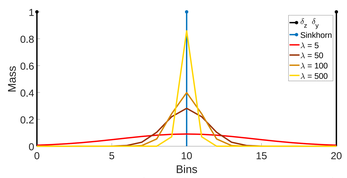


Figure: Comparison of the sharp (Blue) and regularized (Orange) barycenters of two Dirac's deltas (Black) centered in 0 and 20 for different values of λ.

DIFFERENTIAL PROPERTIES

We characterise regularity properties of Sinkhorn maps.

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

Proof (sketch). The proof is organized in the following steps:
Step 1: S_λ and $\bar{\mathsf{S}}_\lambda$ are smooth as functions of $T^\lambda \to$ sufficient to show that T^λ is smooth in a, b.
Step 2: Set $(\alpha^\star, \beta^\star) = \mathrm{argmax}_{\alpha,\beta}\, \mathcal{L}(a, b; \alpha, \beta)$, with

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

By Sinkhorn's scaling theorem, $T^\lambda = \mathrm{diag}(\mathrm{e}^{\lambda\alpha^\star})\mathrm{e}^{-\lambda M}\mathrm{diag}(\mathrm{e}^{\lambda\beta^\star}) \to T^\lambda$ is smooth if $(\alpha^\star, \beta^\star)$ is smooth as a function of (a, b).
Step 3: The smoothness of $(\alpha^\star, \beta^\star)$ 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 $M \in \mathbb{R}_+^{n,m}$, $\lambda > 0$.
$T = \textsc{Sinkhorn}(a, b, M, \lambda), \qquad \bar{T} = T_{1:n,1:(m-1)}$
$L = T \odot M, \qquad \bar{L} = L_{1:n,1:(m-1)}$
$D_1 = \mathrm{diag}(T1_m), \quad D_2 = \mathrm{diag}(\bar{T}^\top 1_n)^{-1}$
$H = D_1 - \bar{T}D_2\bar{T}^\top, \quad \mathsf{f} = -L1_m + \bar{T}D_2\bar{L}^\top 1_n,$
$\mathsf{g} = H^{-1}\,\mathsf{f}$
Return: $\mathsf{g} - 1_n(\,\mathsf{g}^\top 1_n)$

Algorithm 1: Gradient of S_λ

Synthetic experiment. Find the barycenter of nested ellipses.

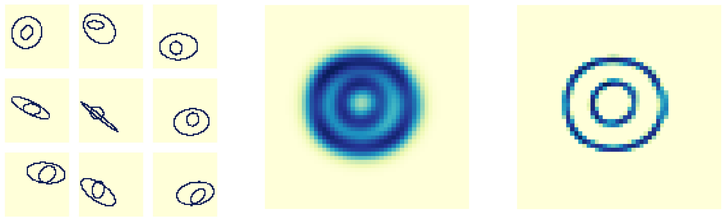


Figure: (Left) Sample input data. (Middle) Barycenter with $\bar{\mathsf{S}}_\lambda$. (Right) Barycenter with S_λ. While solutions of optimization with $\bar{\mathsf{S}}_\lambda$ are often 'blurry', S_λ preserves the *sharpness* of the data.

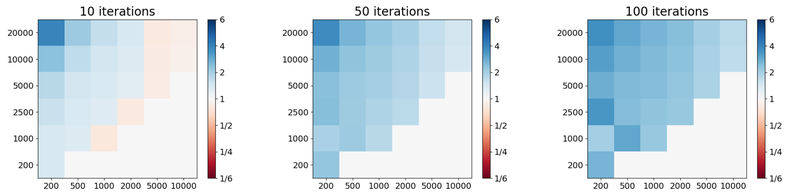


Figure: Ratio of *time*(autodiff) / *time*(Alg.1) for 10, 50, and 100 iterations of Sinkhorn algorithm

LEARNING WITH SINKHORN LOSS: SETTING

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

$$\mathcal{E}(f) = \textstyle\int_{\mathcal{X}\times\mathcal{Y}} \mathcal{S}(f(x), y)\; d\rho(x, y)$$

given a training set $(x_i, y_i)_{i=1}^\ell$ independently sampled from ρ. The loss function $\mathcal{S} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ in our setting is either S_λ or $\bar{\mathsf{S}}_\lambda$.

Structured Prediction Estimator. Given a training set $(x_i, y_i)_{i=1}^\ell$, we consider $\widehat{f} : \mathcal{X} \to \mathcal{Y}$ the structured prediction estimator proposed in [2], defined as

$$\hat{f}(x) = \operatorname*{argmin}_{y\in\mathcal{Y}} \sum_{i=1}^\ell \alpha_i(x)\; \mathcal{S}(y, y_i), \qquad \text{for any } x \in \mathcal{X}. \tag{1}$$

The weights $\alpha_i(x)$ $\to$ Are scores measuring similarity of test point and training points
$\to$ Are obtained via Kernel Ridge Regression

STATISTICAL ANALYSIS

We use the smoothness of S_λ to prove consistency and learning rates of the estimator

Theorem (Universal Consistency) *Let $\mathcal{Y} = \Delta_n^\epsilon$, $\lambda > 0$ and $\mathcal{S}$ be either $\bar{\mathsf{S}}_\lambda$ 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 $\widehat{f}_\ell : \mathcal{X} \to \mathcal{Y}$ be the estimator in* (1) *trained with ℓ points sampled from ρ. Then*

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

Theorem (Learning Rates -informal) *Let $\mathcal{Y} = \Delta_n^\epsilon$, $\lambda > 0$ and $\mathcal{S}$ and $\widehat{f}_\ell$ as above. Then,*

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

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

Role of the smoothness : the proof is technical but essentially allows to embed the problem into a Hilbert setting. This is the first universal consistency result for learning with Sinkhorn loss!

EXPERIMENTS

Image Reconstruction

Goal: given the upper half of Google QuickDraw images, predict their bottom half.

Reconstruction Error (%)

# Cls.	S_λ	$\bar{\mathsf{S}}_\lambda$	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: (Left) Reconstruction error of Sinkhorn, Hellinger and KDE. Misclassification rate of the base SVM classifier: $0.02, 0.07, 0.17$. (Right) Examples of training and reconstructed data.

REFERENCES

M. Cuturi, *Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances*, NIPS 2013
C. Ciliberto et al, *A Consistent Regularization Approach for Structured Prediction*, NIPS 2016
R. Flamary et al, Wasserstein Discriminative Analisys, Journal of Machine Learning 2018