Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance

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

1Department of Computer Science, University College London, London, UK.
2INRIA - Sierra Project Team - ENSI, Paris, France.
3Computational 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 easier 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 propose 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 $\mu = \sum_i a_i \delta_{x_i}$ and the vector $\nu = \sum_i b_i \delta_{y_i}$ in the $d$-dimensional simplex $\Delta^n = \{\mu \in \mathbb{R}^n : \sum_i \mu_i = 1\}$. Given $\mu \in \mathbb{R}^n$ and $\nu \in \mathbb{R}^n$, the Wasserstein distance between $\mu$ and $\nu$ is defined as

$$W_2(\mu, \nu) = \inf_{\Pi} \mathbb{E}_{(x,y) \sim \Pi} d(x,y)^2$$

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

$$\Pi(a,b) = \{T \in \mathbb{R}^{n \times n}_{+} : T_{1:n} = a, T_{::n} = b\}.$$ 

Regularization of Wasserstein Distance

Set $\lambda \geq 0$ and $T(a,b) = \sum_{i,j} T_{ij} |x_i-y_j|^\lambda$. Then $\argmin_{\Pi} T(a,b)$ is obtained via a kernel-based approach: given a positive definite kernel $k: X \times X \rightarrow \mathbb{R}$, we consider $k(x,y) = \exp(-\gamma d(x,y)^2)$ and the evaluation vector with entries $\kappa_k(i,j) = k(x_i,x_j)$ and the empirical kernel matrix with entries $\kappa_k(i,j) = k(x_i,x_j)$ for any $i,j = 1, \ldots, n$. Under suitable regularity assumptions on $\kappa_k$, the estimator $\hat{f} : X \rightarrow \mathbb{R}$ with training points independently sampled from $\mu$ and $\gamma > 0$ is a regularization parameter where $K \in \mathbb{R}^{n \times n}$ and $\kappa_k$ are respectively the empirical kernel matrix with entries $\kappa_k(i,j) = k(x_i,x_j)$ and the evaluation vector with entries $\kappa_k(i,j) = k(x_i,x_j)$. Because for any $i,j = 1, \ldots, n$, $\kappa_k(i,j) \rightarrow k(x_i,x_j)$ as $n \rightarrow \infty$.

Solving the problem in practice:

$$\alpha(x) = \argmin_{\gamma > 0} \left(\gamma + k(x,x) \right)$$

where $\alpha(x)$ are obtained via a kernel-based approach: given a positive definite kernel $k: X \times X \rightarrow \mathbb{R}$, we have $\alpha(x) = \argmin_{\gamma > 0} \left(\gamma + k(x,x) \right)$. The weights $\alpha(x)$ 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 $\hat{y}$. Then the problem is given a positive definite kernel $k: X \times X \rightarrow \mathbb{R}$, we have $K \in \mathbb{R}^{n \times n}$ and $\kappa_k$ are respectively the empirical kernel matrix with entries $\kappa_k(i,j) = k(x_i,x_j)$ and the evaluation vector with entries $\kappa_k(i,j) = k(x_i,x_j)$. Because for any $i,j = 1, \ldots, n$, $\kappa_k(i,j) \rightarrow k(x_i,x_j)$ as $n \rightarrow \infty$.

3. LEARNING WITH SINKHORN LOSS: STATISTICAL ANALYSIS

Thanks to the smoothness of $S$, 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 $Y = \gamma \alpha > 0$ and $b$ be either $S$ or $\hat{S}$. Let $f$ be a bounded continuous universal kernel $K$. For any $\alpha \in (0,1)$ and any distribution $\mu \times \nu$ on $X \times Y \rightarrow \mathbb{R}$, the estimator $f : X \rightarrow \mathbb{R}$ is the estimator $\hat{f}$ that minimizes $\sum_{i=1}^n \mathbb{E}\left[f(x_i) - Y_i\right]^2$ with probability $1 - \delta$. Then

$$\lim_{n \rightarrow \infty} \mathbb{E}\left[f(x) - Y\right]^2 = \mathbb{E}\left[f(x) - \hat{f}(x)\right]^2$$

with high probability with respect to the sampling of training data. The 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). The routine for the gradient is used to implement optimization problems with $S$ as loss. While solutions of optimization with $S$ are often ‘blurry’, $S_i$ preserves the sharpness of the data.

In the following, the barycenters computed with $S_1$ and $S_2$ constitute an example of how $S_1$ is not affected by the same oversmoothing effect as $S_2$.

Figure: Barycenter of Nested Ellipses: (Left) Sample input data. (Middle) Barycenter with $S_2$. (Right) Barycenter with $S_1$.

REFERENCES