

Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance

Giulia Luise, Alessandro Rudi, Massimiliano Pontil, Carlo Ciliberto

▶ **To cite this version:**

Giulia Luise, Alessandro Rudi, Massimiliano Pontil, Carlo Ciliberto. Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance. NIPS 2018 - Advances in Neural Information Processing Systems, Dec 2018, Montreal, Canada. pp.5864-5874. hal-01958887

HAL Id: hal-01958887

https://inria.hal.science/hal-01958887

Submitted on 19 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance

Giulia Luise[1] *g.luise.16@ucl.ac.uk*
Alessandro Rudi[2] *alessandro.rudi@inria.fr*
Massimiliano Pontil[1,3] *m.pontil@cs.ucl.ac.uk*
Carlo Ciliberto[1] *c.ciliberto@ucl.ac.uk*

December 18, 2018

Abstract

Applications of optimal transport have recently gained 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. Promising preliminary experiments complement our analysis.

1 Introduction

Applications of optimal transport have been gaining increasing momentum in machine learning. This success is mainly due to the recent introduction of Sinkhorn approximation [15,28], which offers an efficient alternative to the heavy cost of evaluating the Wasserstein distance directly. The computational advantages have motivated recent applications in optimization and learning over the space of probability distributions, where the Wasserstein distance is a natural metric. However, in these settings adopting Sinkhorn approximation requires solving a further optimization problem *with respect to* the corresponding distance rather than only evaluating it in a point. This consists in a bi-level problem [18] for which it is challenging to derive an optimization approach [22]. As a consequence, a regularized version of the Sinkhorn distance is usually considered [5, 14, 16, 21, 29], for which it is possible to efficiently compute a gradient and thus employ it in first-order optimization

[1]University College London, WC1E 6BT London, United Kingdom
[2]INRIA - Sierra Project-team, École Normale Supérieure, Paris, 75012 Paris, France.
[3]Computational Statistics and Machine Learning - Istituto Italiano di Tecnologia, 16100 Genova, Italy

methods [16]. More recently, also efficient automatic differentiation strategies have been proposed [9], with applications ranging from dictionary learning [30] to GANs [22] and discriminat analysis [20]. A natural question is whether the easier tractability of this regularization is paid in terms of accuracy. Indeed, while as a direct consequence of [13] it can be shown that the original Sinkhorn approach provides a sharp approximation to the Wasserstein distance [13], the same is not guaranteed for its regularized version.

In this work we recall both theoretically and empirically that in optimization problems the original Sinkhorn distances is significantly more favorable than its regularized counterpart, which has been indeed noticed to have a tendency to find over-smooth solutions [39]. We take this as a motivation to study the differential properties of the sharp Sinkhorn with the goal of deriving a strategy to address optimization and learning problems over probability distributions. The principal contributions of this work are threefold. Firstly, we show that both Sinkhorn distances are smooth functions. Secondly, we provide an explicit formula to efficiently compute the gradient of the sharp Sinkhorn. As intended, this latter result allows us to adopt this function in applications such as approximating Wasserstein barycenters [16], which to the best of our knowledge has not been investigated in this setting so far.

As a third main contribution, we provide a novel sound approach to the challenging problem of *learning with Sinkhorn loss*, recently considered in [21]. In particular, we leverage the smoothness of the Sinkhorn distance to study the generalization properties of a structured prediction estimator adapted from [11] to this setting, proving consistency and finite sample bounds. Explicit knowledge of the gradient allows to solve the learning problem in practice. We provide preliminary empirical evidence of the effectiveness of the proposed approach.

2 Background: Optimal Transport and Wasserstein Distance

We provide here a brief overview of the notions used in this work. Given our interest in the computational aspects of optimal transport metrics we refer the reader to [28] for a more in depth introduction to the topic.

Optimal transport theory investigates how to compare probability measures over a domain X. Given a distance function $d : X \times X \to \mathbb{R}$ between points on X (e.g. the Euclidean distance on $X = \mathbb{R}^d$), the goal of optimal transport is to "translate" (or lift) it to distances between probability distributions *over* X. This allows to equip the space $\mathcal{P}(X)$ of probability measures on X with a metric referred to as *Wasserstein* distance, which, for any $\mu, \nu \in \mathcal{P}(X)$ and $p \geq 1$ is defined (see [36]) as

$$W_p^p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d^p(x, y)\, d\pi(x, y), \tag{1}$$

where W_p^p denotes the p-th power of W_p and where $\Pi(\mu, \nu)$ is the set of probability measures on the product space $X \times X$ whose marginals coincide with μ and ν; namely

$$\Pi(\mu, \nu) = \Big\{ \pi \in \mathcal{P}(X \times X) \;\Big|\; P_1 \sharp \pi = \mu, \quad P_2 \sharp \pi = \nu \Big\}, \tag{2}$$

with $P_i(x_1, x_2) = x_i$ the projection operators for $i = 1, 2$ and $P_i \sharp \boldsymbol{\pi}$ the push-forward of $\boldsymbol{\pi}$ [36], namely $P_1 \sharp \boldsymbol{\pi} = \boldsymbol{\pi}(\cdot, X)$ and $P_2 \sharp \boldsymbol{\pi} = \boldsymbol{\pi}(X, \cdot)$.

Wasserstein Distance on Discrete Measures. In the following we focus on measures with discrete support. In particular, we consider distributions $\mu, \nu \in \mathcal{P}(X)$ that can be written as linear combinations $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$ of Dirac's deltas centered on a finite number n and m of points $(x_i)_{i=1}^{n}$ and $(y_j)_{j=1}^{m}$ in X. In order for μ and ν to be probabilities, the vector weights $a = (a_1, \ldots, a_n)^\top \in \Delta_n$ and $b = (b_1, \ldots, b_m)^\top \in \Delta_m$ must belong respectively to the n and m-dimensional simplex, defined as

$$\Delta_n = \left\{ p \in \mathbb{R}_+^n \;\middle|\; p^\top \mathbb{1}_n = 1 \right\} \tag{3}$$

where $\mathbb{R}_+^n$ is the set of vectors $p \in \mathbb{R}^n$ with non-negative entries and $\mathbb{1}_n \in \mathbb{R}^n$ denotes the vector of all ones, so that $p^\top \mathbb{1}_n = \sum_{i=1}^{n} p_i$ for any $p \in \mathbb{R}^n$. In this setting, the evaluation of the Wasserstein distance corresponds to solving a network flow problem [8] in terms of the weight vectors a and b

$$W_p^p(\mu, \nu) = \min_{T \in \Pi(a,b)} \langle T, M \rangle \tag{4}$$

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

$$\Pi(a, b) = \{ T \in \mathbb{R}_+^{n \times m} \;\middle|\; T\mathbb{1}_m = a, \quad T^\top \mathbb{1}_n = b \}, \tag{5}$$

which specializes $\Pi(\mu, \nu)$ in Eq. (2) to this setting and contains all possible joint probabilities with marginals "corresponding" to a and b. In the following, with some abuse of notation, we will denote by $W_p(a, b)$ the Wasserstein distance between the two discrete measures μ and ν with corresponding weight vectors a and b.

An Efficient Approximation of the Wasserstein Distance. Solving the optimization in Eq. (4) is computationally very expensive [15]. To overcome the issue, the following regularized version of the problem is considered,

$$\widetilde{S}_\lambda(a, b) = \min_{T \in \Pi(a,b)} \langle T, M \rangle - \frac{1}{\lambda} h(T) \qquad \text{with} \qquad h(T) = -\sum_{i,j=1}^{n,m} T_{ij}(\log T_{ij} - 1) \tag{6}$$

where $\lambda > 0$ is a regularization parameter. Indeed, as observed in [15], the addition of the entropy h makes the problem significantly more amenable to computations. In particular, the optimization in Eq. (6) can be solved efficiently via Sinkhorn's matrix scaling algorithm [31]. We refer to the function $\widetilde{S}_\lambda$ as the *regularized Sinkhorn* distance.

In contrast to the Wasserstein distance, the regularized Sinkhorn distance is differentiable (actually smooth, see Thm. 2) with respect to both entries a and b, hence particularly appealing for practical applications where the goal is to solve a minimization over probability spaces. Indeed, this distance has been recently used with success in settings related to *barycenter estimation* [1, 5, 16], supervised learning [21] and dictionary learning [29].

3 Motivation: Better Approximation of the Wasserstein Distance

The computational benefit provided by the regularized Sinkhorn distance is paid in terms of the approximation with respect to the Wasserstein distance. Indeed, the entropic term in Eq. (6) perturbs the value of the original functional in Eq. (4) by a term proportional to $1/\lambda$, leading to potentially very different behaviours of the two functions (Example 1 illustrates this effect in practice). In this sense, a natural candidate for a better approximation is

$$S_\lambda(a,b) \;=\; \langle\, T_\lambda \,,\, M \,\rangle \qquad \text{with} \qquad T_\lambda \;=\; \operatorname*{argmin}_{T\in\Pi(a,b)} \langle\, T \,,\, M \,\rangle - \frac{1}{\lambda} h(T) \tag{7}$$

that corresponds to eliminating the contribution of the entropic regularizer $h(T_\lambda)$ from $\widetilde{S}_\lambda$ *after* the transport plan T_λ has been obtained. The function S_λ was originally introduced in [15] as the Sinkhorn distance, although recent literature on the topic has often adopted this name for the regularized version Eq. (6). To avoid confusion, in the following we will refer to S_λ as the *sharp Sinkhorn* distance. Note that we will interchangeably use the notations $S_\lambda(a,b)$ and $S_\lambda(\mu,\nu)$ where clear from the context.

The function S_λ defined in Eq. (7) is nonnegative and safisfies the triangular inequality. However, $S_\lambda(a,a) \neq 0$, and hence S_λ is not -strictly speaking- a distance on Δ_n. As shown in [15](Thm. 1), it sufficies to multiply $S_\lambda(a,b)$ by $\mathbf{1}_{a\neq b}$ to recover an actual distance which satisfies all the axioms. Despite this fact, with some sloppiness we will refer to S_λ itself as Sinkhorn *distance*.

As the intuition suggests, the absence of the entropic term $h(T_\lambda)$ is reflected in a faster rate at approximating the Wasserstein distance. The following result makes this point precise.

Proposition 1. *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*

$$\left|\, S_\lambda(\mu,\nu) - W(\mu,\nu) \,\right| \leq c_1\, e^{-\lambda} \qquad\qquad \left|\, \widetilde{S}_\lambda(\mu,\nu) - W(\mu,\nu) \,\right| \leq c_2 \lambda^{-1}, \tag{8}$$

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

The proof of Prop. 1 is a direct consequence of the result in [13] (Prop. 5.1), which proves the exponential convergence of T_λ in Eq. (7) to the optimal plan of W with maximum entropy, namely

$$T_\lambda \to T^* = \operatorname{argmax} \left\{\, h(T) \;\middle|\; T \in \Pi(a,b) \quad \langle T, M\rangle = W(\mu,\nu) \,\right\} \tag{9}$$

as $\lambda \to +\infty$. While the sharp Sinkhorn distance S_λ preserves the rate of converge of T_λ, the extra term $\lambda^{-1} h(T_\lambda)$ in the definition of the regularized Sinkhorn distance $\widetilde{S}_\lambda$ causes the slower rate. In particular, Eq. (8) (Right) is a direct consequence of [17] (Prop. 2.1). In the appendix we provide more context on the derivation of the two inequalities.

Prop. 1 suggests that, given a fixed λ, the sharp Sinkhorn distance can offer a more accurate approximation of the Wasserstein distance. This intuition is further supported by Example 1 where we compare the behaviour of the two approximations on the problem of finding an optimal transport barycenter of probability distributions.

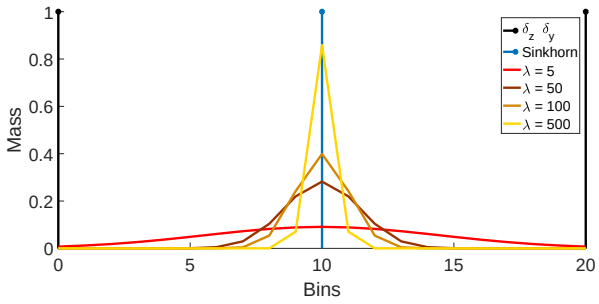


Figure 1: Comparison of the sharp (Blue) and regularized (Orange) barycenters of two Dirac's deltas (Black) centered in 0 and 20 for different values of λ.

Wasserstein Barycenters. Finding the barycenter of a set of discrete probability measures $\mathcal{D} = (\nu_i)_{i=1}^{\ell}$ is a challenging problem in applied optimal transport settings [16]. The *Wasserstein barycenter* is defined as

$$\mu_W^* = \operatorname*{argmin}_{\mu} \mathcal{B}_W(\mu, \mathcal{D}), \qquad \mathcal{B}_W(\mu, \mathcal{D}) = \sum_{i=1}^{\ell} \alpha_i \, W(\mu, \nu_i), \tag{10}$$

namely the point μ_W^* minimizing the weighted average distance between all distributions in the set $\mathcal{D}$, with α_i scalar weights. Finding the Wasserstein barycenter is computationally very expensive and the typical approach is to approximate it with the barycenter $\tilde{\mu}_\lambda^*$, obtained by substituting the Wasserstein distance W with the regularized Sinkhorn distance $\widetilde{S}_\lambda$ in the the objective functional of Eq. (10). However, in light of the result in Prop. 1, it is natural to ask whether the corresponding baricenter μ_λ^* of the sharp Sinkhorn distance S_λ could provide a better estimate of the Wasserstein one. While we defer an empirical comparison of the two barycenters to Sec. 6, here we consider a simple scenario in which the sharp Sinkhorn can be proved to be a significantly better approximation of the Wasserstein distance.

Example 1 (Barycenter of two Deltas)**.** *We consider the problem of estimating the barycenter of two Dirac's deltas $\mu_1 = \delta_z, \mu_2 = \delta_y$ centered at $z = 0$ and $y = n$ with $z, y \in \mathbb{R}$ and n an even integer. Let $X = \{x_0, \dots, x_n\} \subset \mathbb{R}$ be the set of all integers between 0 and n and M the cost matrix with squared Euclidean distances. Assuming uniform weights $\alpha_1 = \alpha_2$, it is well-known that the Wasserstein barycenter is the delta centered on the euclidean mean of z and y, $\mu_W^* = \delta_{\frac{z+y}{2}}$. A direct calculation (see Appendix A) shows instead that the* regularized *Sinkhorn barycenter $\tilde{\mu}_\lambda^* = \sum_{i=0}^{n} a_i \delta_{x_i}$ tends to spread the mass across all $x_i \in X$, accordingly to the amount of regularization,*

$$a_i \propto e^{-\lambda((z-x_i)^2+(y-x_i)^2)/2} \qquad i = 0, \dots, n, \tag{11}$$

behaving similarly to a (discretized) Gaussian with standard deviation of the same order of the regularization λ^{-1}. On the contrary, the sharp *Sinkhorn barycenter equals the Wasserstein one, namely $\mu_\lambda^* = \mu_W^*$ for every $\lambda > 0$. An example of this behavior is reported in Fig. 1.*

Main Challenges of the Sharp Sinkhorn. The example above, together with Prop. 1, provides a strong argument in support of adopting the sharp Sinkhorn distance over its

regularized version. However, while the gradient of the regularized Sinkhorn distance can be easily computed (see [16] or Sec. 4) and therefore it is possible to address optimization problems such as the barycenter in Eq. (10) with first-order methods (e.g. gradient descent), an explicit form for the gradient of the sharp Sinkhorn distance has not been considered. Also, approaches based on automatic differentiation have been recently adopted to compute the gradient of a variant of S_λ, where the plan T_λ is the one obtained after a fixed number L of iterations [20, 22, 22, 30]. These methods have been observed to be both computationally very efficient and also very effective in practice on a number of machine learning applications. However, in this work we are interested in investigating the analytic properties of the gradient of the sharp Sinkhorn distance, for which we provide an explicit algorithm in the following.

4 Differential Properties of Sinkhorn Distances

In this section we present two main results of this work, namely a proof of the smoothness of the two Sinkhorn distances introduced above, and the explicit derivation of a formula for the gradient S_λ. These results will be key to employ the sharp Sinkhorn distance in practical applications. The results are obtained leveraging the Implicit Function Theorem [19] via a proof technique analogous to that in [6, 10, 20] which we outline in this section and discuss in detail in the appendix.

Theorem 2. *For any $\lambda > 0$, the Sinkhorn distances $\widetilde{S}_\lambda$ and $S_\lambda : \Delta_n \times \Delta_n \to \mathbb{R}$ are C^∞ in the interior of their domain.*

Thm. 2 guarantees both Sinkhorn distances to be infinitely differentiable. In Sec. 5 this result will allow us to derive an estimator for supervised learning with Sinkhorn loss and characterize its corresponding statistical properties (i.e. universal consistency and learning rates). The proof of Thm. 2 is instrumental to derive a formula for the gradient of S_λ. We discuss here its main elements and steps while referring to the supplementary material for the complete proof.

Sketch of the proof. The proof of Thm. 2 hinges on the characterization of the (Lagrangian) dual problem of the regularized Sinkhorn distance in Eq. (6). This can be formulated (see e.g. [15]) as

$$\max_{\alpha,\beta} \mathcal{L}_{a,b}(\alpha,\beta), \qquad \mathcal{L}_{a,b}(\alpha,\beta) = \alpha^\top a + \beta^\top b - \frac{1}{\lambda}\sum_{i,j=1}^{n,m} e^{-\lambda(M_{ij}-\alpha_i-\beta_j)} \tag{12}$$

with dual variables $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^m$.
By Sinkhorn's scaling theorem [31], the optimal primal solution T_λ in Eq. (7) can be obtained from the dual solution (α_*, β_*) of Eq. (12) as

$$T_\lambda = \operatorname{diag}(e^{\lambda\alpha_*})\, e^{-\lambda M} \operatorname{diag}(e^{\lambda\beta_*}), \tag{13}$$

where for any $v \in \mathbb{R}^n$, the vector $e^v \in \mathbb{R}^n$ denotes the element-wise exponentiation of v (analogously for matrices) and $\operatorname{diag}(v) \in \mathbb{R}^{n\times n}$ is the diagonal matrix with diagonal

corresponding to v.

Since both Sinkhorn distances are smooth functions of T_λ, it is sufficient to show that $T_\lambda(a, b)$ itself is smooth as a function of a and b. Given the characterization of Eq. (13) in terms of the dual solution, this amounts to prove that $\alpha_*(a, b)$ and $\beta_*(a, b)$ are smooth with respect to (a, b), which is the most technical step of the proof and can be shown leveraging the Implicit Function Theorem [19]. Indeed, the dual variables $\alpha_*(a, b)$ and $\beta_*(a, b)$ are obtained as argmax of the strictly convex function $\mathcal{L}$ Eq. (12). The argmax corresponds to a stationary point of the gradient, which means $\nabla_{\alpha,\beta}\mathcal{L}(\alpha_*, \beta_*) = 0$. The last part of the proof consists in applying the Implicit Function Theorem to the function $\nabla_{\alpha,\beta}\mathcal{L}$. Note that the theorem can be applied thanks to the strict convexity of $\mathcal{L}$, which guarantess that the Jacobian of $\nabla_{\alpha,\beta}\mathcal{L}$, which is the Hessian of $\mathcal{L}$ is invertible. All the details are discussed at length in the Appendix. □

The gradient of Sinkhorn distances. We now discuss how to derive the gradient of Sinkhorn distances with respect to one of the two variables. In both cases, the dual problem introduced in Eq. (12) plays a fundamental role. In particular, as pointed out in [16], the gradient of the regularized Sinkhorn distance can be obtained directly from the dual solution as $\nabla_a \widetilde{S}_\lambda(a, b) = \alpha_*(a, b)$, for any $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. This characterization is possible because of well-known properties of primal and dual optimization problems [8]. The sharp Sinkhorn distance does not have a formulation in terms of a dual problem and therefore a similar argument does not apply. Nevertheless, we show here that it is still possible to obtain its gradient in closed form in terms of the dual solution.

Theorem 3. *Let $M \in \mathbb{R}^{n\times m}$ be a cost matrix, $a \in \Delta_n$, $b \in \Delta_m$ and $\lambda > 0$. Let $\mathcal{L}_{a,b}(\alpha, \beta)$ be defined as in Eq. (12), with argmax in (α_*, β_*). Let T_λ be defined as in Eq. (13). Then,*

$$\nabla_a S_\lambda(a, b) = P_{T\Delta_n}(A\ L\mathbb{1}_m + B\ \bar{L}^\top \mathbb{1}_n) \tag{14}$$

where $L = T_\lambda \odot M \in \mathbb{R}^{n\times m}$ is the entry-wise multiplication between T_λ and M and $\bar{L} \in \mathbb{R}^{n\times m-1}$ corresponds to L with the last column removed. The terms $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m-1}$ are

$$[A\ B] = -\lambda\ D\ \left[\ \nabla^2_{(\alpha,\beta)}\mathcal{L}_{a,b}(\alpha_*, \beta_*)\ \right]^{-1}, \tag{15}$$

with $D = [I\ \mathbf{0}]$ the matrix concatenating the $n\times n$ identity matrix I and the matrix $\mathbf{0} \in \mathbb{R}^{n\times m-1}$ with all entries equal to zero. The operator $P_{T\Delta_n}$ denotes the projection onto the tangent plane $T\Delta_n = \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 0\}$ to the simplex Δ_n.

The proof of Thm. 3 can be found in the supplementary material (Sec. C). The result is obtained by first noting that the gradient of S_λ is characterized (via the chain rule) in terms of the the gradients $\nabla_a \alpha_*(a, b)$ and $\nabla_a \beta_*(a, b)$ of the dual solutions. The main technical step of the proof is to show that these gradients correspond respectively to the terms A and B defined in Eq. (15).

To obtain the gradient of S_λ in practice, it is necessary to compute the Hessian $\nabla^2_{(\alpha,\beta)}\mathcal{L}_{a,b}(\alpha_*, \beta_*)$ of the dual functional. A direct calculation shows that this corresponds to the matrix

$$\nabla^2_{(\alpha,\beta)}\mathcal{L}(\alpha_*, \beta_*) = \begin{bmatrix} \operatorname{diag}(a) & \bar{T}_\lambda \\ \bar{T}_\lambda^\top & \operatorname{diag}(\bar{b}) \end{bmatrix}, \tag{16}$$

Algorithm 1 Computation of $\nabla_a S_\lambda(a, b)$

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

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

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

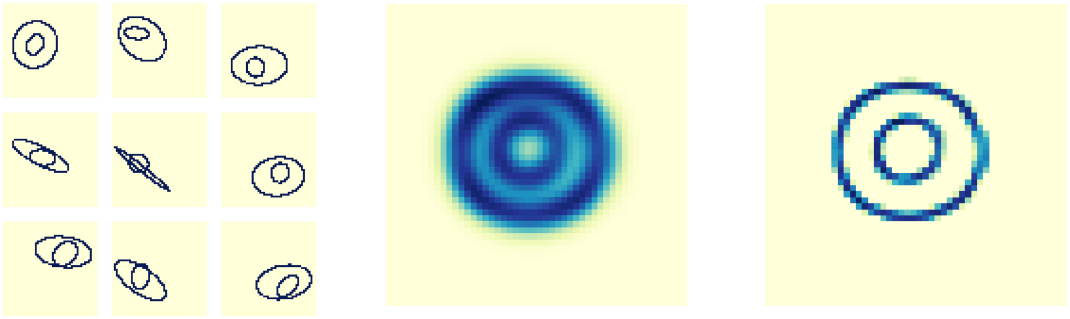


Figure 2: Nested Ellipses: (Left) Sample input data. (Middle) Regularized (Right) sharp Sinkhorn barycenters.

where $\bar{T}_\lambda$ (equivalently $\bar{b}$) corresponds to T_λ (respectively b) with the last column (element) removed. See the supplementary material (Sec. C) for the details of this derivation.

From the discussion above, it follows that the gradient of S_λ can be obtained in closed form in terms of the transport plan T_λ. Alg. 1 reports an efficient approach to perform this operation. The algorithm can be derived by simple algebraic manipulation of Eq. (14), given the characterization of the Hessian in Eq. (16). We refer to the supplementary material for the detailed derivation of the algorithm.

Barycenters with the sharp Sinkhorn. Using Alg. 1 we can now apply the accelerated gradient descent approach proposed in [16] to find barycenters with respect to the sharp Sinkhorn distance. Fig. 2 reports a qualitative experiment inspired by the one in [16], with the goal of comparing the two Sinkhorn barycenters. We considered 30 images of random nested ellipses on a 50×50 grid. We interpret each image as a distribution with support on pixels. The cost matrix is given by the squared Euclidean distances between pixels. Fig. 2 shows some examples images in the dataset and the corresponding barycenters of the two Sinkhorn distances. While the barycenter $\tilde{\mu}_\lambda^*$ of $\widetilde{S}_\lambda$ suffers a blurry effect, the S_λ barycenter μ_λ^* is very sharp, suggesting a better estimate of the ideal one.

We conclude this section with a computational consideration on the two methods.

Remark 1 (Computations)**.** *We compare the computational complexity of evaluating the*

gradients of $\widetilde{S}_\lambda$ and S_λ. Both gradients rely on the solution of the Sinkhorn problem in Eq. (6), which requires $O(nm\epsilon^{-2}\lambda)$ operations to achieve an ϵ-accurate solution (this is easily derived from [1], see supplementary material). While the gradient of $\widetilde{S}_\lambda$ does not require further operations, the gradient of S_λ requires the inversion of an $n \times n$ matrix as stated in Alg. 1. However, since the matrix to be inverted is a sum of a diagonal matrix and a low-rank matrix, the inversion requires $O(nm^2)$ operations (e.g. using Woodbury matrix identity), for a total cost of the gradient equal to $O(nm(\epsilon^{-2}\lambda + m))$. Even for very large n, Alg. 1 is still efficient in all those settings where $m << n$.

Moreover, note that the most expensive additional operations consist of matrix multiplications and the inversion of a positive definite matrix, which are very efficiently implemented on modern machines. Indeed, in our experiments the Sinkhorn algorithm was always the most expensive component of the computation. It is important to notice however that in practical applications both routines can be parallelized, and several ideas can be exploited to lower the computational costs of either algorithms depending on the problem structure (see for instance the convolutional Wasserstein distance *in [32]). Therefore, depending on the setting, the computation of the gradient of the sharp Sinkhorn could be comparable or significantly slower than the regularized Sinkhorn or the automatic differentiation considered in [22].*

5 Learning with Sinkhorn Loss Functions

Given the characterization of smoothness for both Sinkhorn distances, in this section we focus on a specific application: supervised learning with a Sinkhorn loss function. Indeed, the result of Thm. 2 will allow to characterize the statistical guarantees of an estimator devised for this problem in terms of its universal consistency and learning rates. Supervised learning with the (regularized) Sinkhorn loss was originally considered in [21], where an empirical risk minimization approach was adopted. In this work we take a structured prediction perspective [4]. This will allow us to study a learning algorithm with strong theoretical guarantees that can be efficiently applied in practice.

Problem Setting. Let $\mathcal{X}$ be an input space and $\mathcal{Y} = \Delta_n$ a set of histograms. As it is standard in supervised learning settings, the goal is to approximate a minimizer of the *expected risk*

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f), \qquad \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{S}(f(x), y)\, d\rho(x, y) \tag{17}$$

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 $\widetilde{S}_\lambda$.

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 [11], defined such that

$$\widehat{f}(x) = \operatorname*{argmin}_{y\in\mathcal{Y}} \sum_{i=1}^{\ell} \alpha_i(x)\, \mathcal{S}(y, y_i) \tag{18}$$

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* $\mathcal{S}$. While different learning strategies can be adopted to learn the α scores, we consider the kernel-based approach in [11]. In particular, given a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ [3], we have

$$\alpha(x) = (\alpha_1(x), \dots, \alpha(x))^\top = (K + \gamma \ell I)^{-1} K_x \tag{19}$$

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 $i, j = 1, \dots, \ell$.

Remark 2 (Structured Prediction and Differentiability of Sinkhorn)**.** *The current work provides both a* theoretical *and* practical *contribution to the problem of learning with Sinkhorn distances. On one hand, the smoothness guaranteed by Thm. 2 will allow us to characterize the generalization properties of the estimator (see below). On the other hand, Thm. 3 provides an efficient approach to* solve *the problem in Eq. (18). Indeed note that this optimization corresponds to solving a barycenter problem in the form of Eq. (10). Given the gradient estimation algorithm in Alg. 1, this work allows to solve it by adopting first order methods such as gradient descent.*

Universal Consistency of $\widehat{f}$. We now characterize the theoretical properties of the estimator introduced in Eq. (18). We start by showing $\widehat{f}$ is *universally consistent*, namely that it achieves minimum expected risk as the number ℓ of training points increases. To avoid technical issues on the boundary, in the following we will require $\mathcal{Y} = \Delta_n^\epsilon$ for some $\epsilon > 0$ to be the set of points $p \in \Delta_n$ with $p_i \geq \epsilon$ for any $i = 1, \dots, n$. The main technical step in this context is to show that for any smooth loss function on $\mathcal{Y}$, the estimator in Eq. (18) is consistent. In this sense, the characterization of smoothness in Thm. 2 is key to prove the following result, in combination with Thm. 4 in [11].

Theorem 4 (Universal Consistency)**.** *Let $\mathcal{Y} = \Delta_n^\epsilon$, $\lambda > 0$ and $\mathcal{S}$ be either $\widetilde{S}_\lambda$ or S_λ. Let k be a bounded continuous universal*[1] *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 Eq. (18) trained with $(x_i, y_i)_{i=1}^\ell$ points independently sampled from ρ and $\gamma_\ell = \ell^{-1/4}$. Then*

$$\lim_{\ell \to \infty} \mathcal{E}(\widehat{f}_\ell) \quad = \quad \min_{f : \mathcal{X} \to \mathcal{Y}} \mathcal{E}(f) \quad \textit{with probability } 1. \tag{20}$$

The proof of Thm. 4 is reported in Appendix D. A result analogous to the one above was originally proved in [11] (Thm. 4) for a wide family of functions referred to as *Structure Encoding Loss Function (SELF)* (see [12] or the appendix of this work). While several loss functions used in structured prediction have been observed to satisfy this SELF definition, such characterization was not available for the Sinkhorn distances. The main technical step in the proof of Thm. 4 in this sense is to prove that any smooth function on $\mathcal{Y}$ satisfies the definition of SELF (see Def. 6 and Thm. 7 in the Appendix). Combining this result with Thm.4 in [11], we obtain that *for every smooth loss function $\mathcal{S}$ on $\mathcal{Y}$ the corresponding*

[1] This is a standard assumptions for universal consistency (see [35]). Example: $k(x, x') = e^{-\|x - x'\|^2 / \sigma}$.

estimator $\widehat{f}$ *in Eq. (18) is universally consistent*. The universal consistency of the Sikhorn distances is therefore guaranteed by the smoothness result of Thm. 2.

Thm. 4 guarantees $\widehat{f}$ to be a valid estimator for the learning problem. To our knowledge, this is the first result characterizing the universal consistency of an estimator minimizing (an approximation to) the Wasserstein distance.

Learning Rates. By imposing standard regularity conditions on the learning problem, it is possible to provide also excess risk bounds for $\widehat{f}$. Since these conditions are quite technical, we provide here a brief overview while deferring an in-depth discussion to Appendix D. We start from the observation (see e.g. Lemma 6 in [11]) that the solution $f^* : \mathcal{X} \to \mathcal{Y}$ of the learning problem introduced in Eq. (17) is such that

$$f^*(x) = \operatorname*{argmin}_{z \in \mathcal{Y}} \int_{\mathcal{Y}} \mathcal{S}(z, y) \, d\rho(y|x) \tag{21}$$

almost surely on $\mathcal{X}$. In particular $f^*(x)$ corresponds to the minimizer of the *conditional expectation* $\mathbb{E}_{y|x}\mathcal{S}(z, y)$ of the loss $\mathcal{S}(z, y)$ with respect to y given $x \in \mathcal{X}$. As it is standard in statistical learning theory, in order to obtain generalization bounds for estimating f^* we will impose regularity assumptions on the conditional distribution $\rho(\cdot|x)$ or, more precisely, on its corresponding *conditional mean embedding* ([33, 34]) with respect to a suitable space of functions.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the kernel used for the estimation of the weights α in Eq. (19) and let $\mathcal{F}$ be the associated reproducing kernel Hilbert spaces (RKHS) (see [3] for a definition). Let $h : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be the kernel $h(y, y') = e^{-\|y-y'\|}$ on the output set $\mathcal{Y}$. The RKHS associated to h is $\mathcal{H} = W_2^{(n+1)/2}(\mathcal{Y})$, the Sobolev space of square integrable functions with smoothness $\frac{n+1}{2}$ (see e.g. [37]). We consider a function $g^* : \mathcal{X} \to \mathcal{H}$ such that

$$g^*(x) = \int_{\mathcal{Y}} h(y, \cdot) d\rho(y|x) \tag{22}$$

almost surely on $\mathcal{X}$. For any $x \in X$, the quantity $g^*(x)$ is known as the *conditional mean embedding* of $\rho(\cdot|x)$ in $\mathcal{H}$, originally introduced in [33, 34]. In particular, in [34] it was shown that in order to obtain learning rates for an estimator approximating g^*, a key assumption is that g^* belongs to $\mathcal{H} \otimes \mathcal{F}$, the tensor product between the space $\mathcal{H}$ on the output and the space $\mathcal{F}$ on the input. In this work we will require the same assumption. It can be verified that $\mathcal{H} \otimes \mathcal{F}$ is a RKHS for vector-valued functions [2, 25, 26] and that by asking $g^* \in \mathcal{H} \otimes \mathcal{F}$ we are requiring the conditional mean embedding of $\rho(\cdot|x)$ to be sufficiently regular *as a function on* $\mathcal{X}$. We are now ready to report our result on the statistical performance of $\widehat{f}$.

Theorem 5 (Learning Rates)**.** *Let* $\mathcal{Y} = \Delta_n^\varepsilon$, $\lambda > 0$ *and* $\mathcal{S}$ *be either* $\widetilde{S}_\lambda$ *or* S_λ. *Let* $\mathcal{H} = W_2^{(n+1)/2}(\mathcal{Y})$ *and* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *be a bounded continuous reproducing kernel on* $\mathcal{X}$ *with associated RKHS* $\mathcal{F}$. *Let* $\widehat{f}_\ell : \mathcal{X} \to \mathcal{Y}$ *be the estimator in Eq. (18) trained with* ℓ *training points independently sampled from* ρ *and with* $\gamma = \ell^{-1/2}$. *If* g^* *defined in Eq. (22) is such that* $g^* \in \mathcal{H} \otimes \mathcal{F}$, *then*

$$\mathcal{E}(f) - \min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f) \leq c \, \tau^2 \ell^{-1/4} \tag{23}$$

holds with probability $1 - 8e^{-\tau}$ *for any* $\tau > 0$, *with* c *a constant independent of* ℓ *and* τ.

	Support			
Improvement	1%	2%	10%	50%
$\mathcal{B}_W(\tilde{\mu}_\lambda^*) - \mathcal{B}_W(\mu_\lambda^*)$	14.914 ± 0.076	12.482 ± 0.135	2.736 ± 0.569	0.258 ± 0.012

Table 1: Average absolute improvement in terms of the ideal Wasserstein barycenter functional $\mathcal{B}_W$ in Eq. (10) of *sharp* vs *regularized* Sinkhorn, for barycenters of random measures with sparse support.

The proof of Thm. 5 requires to combine our characterization of the Sinkhorn distances (or more generally smooth functions on $\mathcal{Y}$) as structure encoding loss functions (see Thm. 7) with Thm. 5 in [11] where a result analogous to the one above is reported for SELF loss functions. See Appendix D for a detailed proof.

Remark 3. *A relevant question is whether the Wasserstein distance could be similarly framed in the setting of structured prediction. However, the argument used to address Sinkhorn distances relies on their smoothness properties and cannot be extended to the Wasserstein distance, which is not differentiable. A completely different approach may still be successful and we will investigate this question in future work.*

We conclude this section with a note on previous work. We recall that [21] has provided the first *generalization bounds* for an estimator minimizing the regularized Sinkhorn loss. In Thm. 5 however we characterize the *excess risk bounds* of the estimator in Eq. (18). The two approaches and analysis are based on different assumptions on the problem. Therefore, a comparison of the corresponding learning rates is outside the scope of this work and is left for future research.

6 Experiments

We present here some experiments that compare the two Sinkhorn distances empirically. Optimization was performed with the accelerated gradient from [16] for S_λ and Bregman projections [5] for $\widetilde{S}_\lambda$. The computation of barycenters with Bregman iteration is extremely fast compared to gradient descent. We have then used the barycenter of $\widetilde{S}_\lambda$ computed with Bregman projections as *initial datum* for gradient descent with S_λ: this has a positive influence on the number of iterations needed to converge and in this light the optimization with respect to the sharp Sinkhorn distance acts as a *refinement* of the solution with respect to $\widetilde{S}_\lambda$.

Barycenters with Sinkhorn Distances. We compared the quality of Sinkhorn barycenters in terms of their approximation of the (ideal) Wasserstein barycenter. We considered discrete distributions on 100 bins, corresponding to the integers from 1 to 100 and a squared Euclidean cost matrix M. We generated datasets of 10 measures each, where only $k = 1, 2, 10, 50$ (randomly chosen) consecutive bins are different from zero, with the non-zero entries sampled uniformly at random between 0 and 1 (and then normalized to sum up to 1). We empirically chose the Sinkhorn regularization parameter λ to be the smallest value such that the output T_λ of the Sinkhorn algorithm would be within 10^{-6} from the transport polytope in 1000 iterations. Tab. 1 reports the absolute improvement of the barycenter of the sharp Sinkhorn distance with respect to the one obtained with the regularized Sinkhorn, averaged over 10 independent dataset generation for each support

# Classes	Reconstruction Error (%) S_λ	$\widetilde{S}_\lambda$	Hell [12]	KDE [38]
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

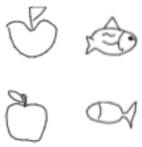


Figure 3: Average reconstruction errors of the Sinkhorn, Hellinger, and KDE estimators on the Google QuickDraw reconstruction problem. On the right, a mini-sample of the dataset.

size k. As can be noticed, the sharp Sinkhorn consitstently outperforms its regularized counterpart. Interestingly, this improvement is more evident for measures with sparse support and tends to reduce as the support increases. This is in line with the remark in example Example 1 and the fact that the regularization term in $\widetilde{S}_\lambda$ tends to encourage oversmoothed solutions.

Learning with Wasserstein loss. We evaluated the Sinkhorn distances in an image reconstruction problem similar to the one considered in [38] for structured prediction. 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). Similarly to [16] we interpret each (half) image as an histogram with mass corresponding to the gray levels (normalized to sum up to 1). For all experiments, according to [11], we evaluated the performance of the reconstruction in terms of the classification accuracy of an image recognition SVM classifier trained on a separate dataset. To train the structured prediction estimator in Eq. (18) we used a Gaussian kernel with bandwith σ and regularization parameter γ selected by cross-validation.

Google QuickDraw. We compared the performance of the two estimators on a challenging dataset. We selected $c = 2, 4, 10$ classes from the Google QuickDraw dataset [23] which consists in images of size 28×28 pixels. We trained the structured prediction estimators on 1000 images per class and tested on other 1000 images. We repeated these experiments 5 times, each time randomly sampling a different training and test dataset. Fig. 3 reports the reconstruction error (i.e. the classification error of the SVM classifier) over images reconstructed by the Sinkhorn estimators, the structured prediction estimator with Hellinger loss [11] and the Kernel Dependency Estimator (KDE) [38]. As can be noticed, both Sinkhorn estimators perform significantly better than their competitors (except the Hellinger distance outperforming $\widetilde{S}_\lambda$ on 4 classes). This is in line with the intuition that optimal transport metrics respect the way the mass is distributed on images [15, 16]. Moreover, it is interesting to note that the estimator of the sharp Sinkhorn distance provides always better reconstructions than its regularized counterpart, supporting the idea that it is more suited to settings where Wasserstein distance should be used.

The experiments above are a preliminary assessment of the potential of sharp Sinkhorn distance in barycenters and learning settings. More extensive experiments on real data will be matter of future work.

7 Discussion

In this paper we investigated the differential properties of Sinkhorn distances. We proved the smoothness of the two functions and derived an explicit algorithm to efficiently compute the gradient of the sharp Sinkhorn distance. Our result allows to employ the sharp Sinkhorn distance in applications that rely on first order optimization methods, such as in approximating Wasserstein barycenters and supervised learning on probability distributions. In this latter setting, our characterization of smoothness allowed to study the statistical properties of the Sinkhorn distance as loss function. In particular we considered a structured prediction estimator for which we proved universal consistency and generalization bounds. Future work will focus on further applications and a more extensive comparison with the existing literature.

References

[1] J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *NIPS*, pages 1961–1971, 2017.

[2] M. A. Alvarez, L. Rosasco, N. D. Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

[3] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[4] G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. Vishwanathan. Predicting structured data. neural information processing, 2007.

[5] J. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative bregman projections for regularized transportation problems. *SIAM J. Scientific Computing*, 37(2), 2015.

[6] Y. Bengio. Gradient-based optimization of hyperparameters. *Neural computation*, 12(8):1889–1900, 2000.

[7] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[8] D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1st edition, 1997.

[9] N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4):71–1, 2016.

[10] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine learning*, 46(1-3):131–159, 2002.

[11] C. Ciliberto, L. Rosasco, and A. Rudi. A consistent regularization approach for structured prediction. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and

R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4412–4420. Curran Associates, Inc., 2016.

[12] C. Ciliberto, A. Rudi, L. Rosasco, and M. Pontil. Consistent multitask learning with nonlinear output relations. In *Advances in Neural Information Processing Systems*, pages 1983–1993, 2017.

[13] R. Cominetti and J. S. Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1):169–187, Oct 1994.

[14] N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *ECML/PKDD 2014*, LNCS, pages 1–16, Nancy, France, Sept. 2014.

[15] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

[16] M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR.

[17] M. Cuturi and G. Peyré. A smoothed dual approach for variational wasserstein problems. *SIAM J. Imaging Sciences*, 9(1):320–343, 2016.

[18] S. Dempe. *Foundations of bilevel programming*. Springer Science & Business Media, 2002.

[19] C. Edwards. *Advanced Calculus of Several Variables*. Dover Books on Mathematics. Dover Publications, 2012.

[20] R. Flamary, M. Cuturi, N. Courty, and A. Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, May 2018.

[21] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio. Learning with a wasserstein loss. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2053–2061, Cambridge, MA, USA, 2015. MIT Press.

[22] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

[23] I. Google. *QuickDraw Dataset*. 2017. Available on line.

[24] T. Kollo and D. von Rosen. *Advanced Multivariate Statistics with Matrices*. Mathematics and Its Applications. Springer Netherlands, 2006.

[25] G. Lever, L. Baldassarre, S. Patterson, A. Gretton, M. Pontil, and S. Grünewälder. Conditional mean embeddings as regressors. In *International Conference on Machine Learing (ICML)*, volume 5, 2012.

[26] C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

[27] V. Moretti. *Spectral Theory and Quantum Mechanics: With an Introduction to the Algebraic Formulation*. UNITEXT. Springer Milan, 2013.

[28] G. Peyré, M. Cuturi, et al. Computational optimal transport. Technical report, 2017.

[29] A. Rolet, M. Cuturi, and G. Peyré. Fast dictionary learning with a smoothed wasserstein loss. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 630–638, Cadiz, Spain, 09–11 May 2016. PMLR.

[30] M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngole, D. Coeurjolly, M. Cuturi, G. Peyré, and J.-L. Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

[31] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21(2):343–348, 1967.

[32] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

[33] L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Process. Mag.*, 30(4):98–111, 2013.

[34] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968. ACM, 2009.

[35] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[36] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

[37] H. Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[38] J. Weston, O. Chapelle, A. Elisseeff, B. Schölkopf, and V. Vapnik. Kernel dependency estimation. In *Advances in Neural Information Processing Systems 15*, pages 873–880, Cambridge, MA, USA, Oct. 2003. Max-Planck-Gesellschaft, MIT Press.

[39] J. Ye, P. Wu, J. Z. Wang, and J. Li. Fast discrete distribution clustering using wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332, May 2017.

Supplementary Material

A Barycenter of Dirac Deltas

Wasserstein barycenter problems can be divided into two main classes: problems in which the support is free (and must be computed, generating a nonconvex problem [16]) and problems where the support is fixed. In some cases, the latter is the only valid choice: for instance, when the geometric domain is a space of symbols and the cost matrix M contains the symbol-to-symbol dissimilarities, no extra information of the symbol space is available and the support of the barycenter will have to lie on a pre-determined set in order to be meaningful. A concrete example is the following: when dealing with histograms on words, the barycenter will optimize how to spread the mass among a set of known words that are used to build the matrix M, through a word2vec operation. In the following we carry out the computation of the barycenter of two Dirac deltas with regularized Sinkhorn and Sinkhorn distances, in order to prove what stated in example 1.

Barycenter with $\widetilde{S}_\lambda$: Let $\mu = \delta_z$ be the Dirac delta centered at $z \in \mathbb{R}^d$ and $\nu = \delta_y$ the Dirac delta centered at $y \in \mathbb{R}^d$. We fix the set of admissible support of the barycenter $X = \{x_1, \dots, x_n\}$, where $x_i \in \mathbb{R}^d$ for any i. For the sake of simplicity let us assume that X contains the point $(y+z)/2$. The cost matrices with mutual distances between z and X and y and X will be

$$M^z = \{d(z, x_i)\}_{i=1}^n \in \mathbb{R}^n, \qquad M^y = \{d(y, x_i)\}_{i=1}^n.$$

Since the support is fixed, only the masses $a = (a_1, \dots, a_n)$ of the barycenter $\tilde{\mu}_\lambda = \sum_{i=1}^n a_i \delta_{x_i}$ are to be computed. Vector a is the minimizer of the following functional

$$\Delta_n \ni a \longrightarrow \mathcal{B}_{\widetilde{S}_\lambda}(a) = \frac{1}{2}\widetilde{S}_\lambda(a, \delta_z) + \frac{1}{2}\widetilde{S}_\lambda(a, \delta_y).$$

Note that since Dirac delta has mass 1 concentrated at a point, the transport polytope corresponding to a and a Dirac delta is $\Pi(a, 1)$. The elements in $\Pi(a, 1)$ are those matrices $T \in \mathbb{R}^{n\times 1}$ such that $T\mathbb{1}_1 = a$ and $T^\top \mathbb{1}_n = 1$. Thus,

$$\begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_n \end{pmatrix} \begin{pmatrix} 1 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \tag{24}$$

which implies $T_1 = a_1, \dots, T_n = a_n$. In this case, $\Pi(a, 1)$ contains only one matrix, which coincides with $a^\top$. The distance $\widetilde{S}_\lambda(a, \delta_z)$ is given by $\langle a^\top, M^z\rangle - \frac{1}{\lambda}h(a)$ and, similarly, $\widetilde{S}_\lambda(a, \delta_y) = \langle a^\top, M^y\rangle - \frac{1}{\lambda}h(a)$. Then, the goal is to minimize

$$a \longrightarrow \frac{1}{2}\langle a, M^z\rangle + \frac{1}{2}\langle a, M^y\rangle + \frac{1}{\lambda}\sum_{i=1}^n a_i(\log a_i - 1)$$

with the constraint that $a \in \Delta_n$. The partial derivative with respect to a_i is given by

$$\frac{\partial \mathcal{B}_{\widetilde{S}_\lambda}}{\partial a_i} = \frac{1}{2}(M_i^z + M_i^y) + \frac{1}{\lambda}\log a_i$$

Setting it equal to zero, it yields $a_i = e^{-\lambda(M^z_i + M^y_i)/2}$. The constraint $a \in \Delta_n$ leads to

$$a_i = \frac{e^{-\lambda(M^z_i + M^y_i)/2}}{\sum_{j=1}^n e^{-\lambda(M^z_j + M^y_j)/2}}.$$

Then the barycenter $\tilde{\mu}^*_\lambda$ has masses $(a_1, \dots, a_n)$ where each a_i is strictly positive, with maximum at the entry corresponding to the point x_i which realizes the minimum distance from z and y, i.e. $(z+y)/2$. The sparsity of the initial deltas is lost.

Barycenter with S_λ: On the other hand, let us compute the barycenter between μ and ν with respect to the Sinkhorn distance recalled in Eq. (7). The very same considerations on $\Pi(a, 1)$ still hold, so $\Pi(a, 1)$ contains $T = a^\top$ only. Hence, in this case the Sinkhorn barycenter functional $\mathcal{B}_{S_\lambda}$ coincides with the Wasserstein barycenter functional $\mathcal{B}_W$, since $S_\lambda(a, \delta_j) = \langle a^\top, M^j \rangle = W(a, \delta_j)$, for $j = z, y$. This trivially implies that $\mu^*_\lambda = \mu^*_W$.

B Proof of Proposition 1 in section 3

Proposition 1. *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*

$$\left| S_\lambda(\mu, \nu) - W(\mu, \nu) \right| \leq c_1 e^{-\lambda} \qquad \left| \widetilde{S}_\lambda(\mu, \nu) - W(\mu, \nu) \right| \leq c_2 \lambda^{-1}, \tag{8}$$

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

Proof. As shown in [13](Prop.5.1), the sequence T_λ converges to an optimal plan of W as λ goes to infinity. More precisely,

$$T_\lambda \to T^* = \operatorname{argmax}_{T \in \Pi(a,b)} \{h(T);\ \langle T, M \rangle = W(\mu, \nu)\}$$

exponentially fast, that is $\|T_\lambda - T^*\|_{\mathbb{R}^{nm}} \leq c\, e^{-\lambda}$. Thus,

$$|S_\lambda(\mu, \nu) - W(\mu, \nu)| = |\langle T_\lambda, M \rangle - \langle T^*, M \rangle| \leq \|T_\lambda - T^*\| \|M\| \leq c\, e^{-\lambda} \|M\| =: c_1 e^{-\lambda}.$$

As for the second part, let T^* be the $\operatorname{argmax}_{T \in \Pi(a,b)} \{h(T);\ \langle T, M \rangle = W(\mu, \nu)\}$. By optimality of T_λ and T^* for their optimization problems, it holds

$$0 \leq \langle T_\lambda, M \rangle - \langle T^*, M \rangle \leq \lambda^{-1} (h(T_\lambda) - h(T^*));$$

Indeed, since T_λ is the optimum, it attains the minimum and hence

$$\langle T_\lambda, M \rangle - \lambda^{-1} h(T_\lambda) \leq \langle T, M \rangle - \lambda^{-1} h(T)$$

for any other T, including T^*. By definition of $\widetilde{S}_\lambda$ and W, the inequalities above can be rewritten as

$$0 \leq \widetilde{S}_\lambda(\mu, \nu) - W(\mu, \nu) \leq \lambda^{-1} h(T^*) =: c_2 \lambda^{-1}$$

which goes to 0 with speed λ^{-1} as λ goes to infinity. □

C Proofs on differential properties and formula of the gradient

In this section we go over all the details of the proofs sketched in section 4.

Theorem 2. *For any $\lambda > 0$, the Sinkhorn distances $\widetilde{S}_\lambda$ and $S_\lambda : \Delta_n \times \Delta_n \to \mathbb{R}$ are C^∞ in the interior of their domain.*

Proof. Let us show the proof for S_λ first. We organize it in three steps:
Step 1. S_λ is smooth when T_λ is: when considering histograms, S_λ depends on its argument a and b through the optimal coupling $T_\lambda(a, b)$, being the cost matrix M fixed. Thus, since S_λ is a smooth function of T_λ (being the Frobenius product of T_λ with a constant matrix), showing that S_λ is smooth in a, b amounts to showing that T_λ is smooth.
Step 2. T_λ is smooth when (α_*, β_*) is: By Sinkhorn's scaling theorem [31], the optimal plan T_λ is characterized as follows

$$T_\lambda = \mathrm{diag}(e^{\lambda\alpha_*})e^{-\lambda M}\mathrm{diag}(e^{\lambda\beta_*}.) \tag{25}$$

Being the exponential a smooth function, $T_\lambda(a, b)$ is smooth in a and b if the dual optima $\alpha_*(a, b)$ and $\beta_*(a, b)$ are. Our goal is then showing smoothness with respect to a and b of the dual optima.
Step 3. (α_*, β_*) is smooth in a, b: this is the most technical part of the proof. First of all, let us stress that one among the $n + m$ rows/columns constraints of $\Pi(a, b)$ is *redundant*: the standard dual problem recalled in Eq. (12) has an extra dual variable, and this degree of freedom is clear noticing that if (α, β) is feasible, than the pair $(\alpha + t\mathbb{1}_n, \beta - t\mathbb{1}_m)$ is also feasible. In the following, we get rid of the redundancy removing one of the dual variables. Hence, let us set

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

where $\bar{b}$ corresponds to b with the last element removed.
To avoid cumbersome notation, from now on we denote $x = (a, b)$ and $\gamma = (\alpha, \beta)$. The function $\mathcal{L}$ is smooth and strictly convex in γ: hence, for every fixed x in the interior of $\Delta_n \times \Delta_n$ there exist $\gamma^*(x)$ such that $\mathcal{L}(x; \gamma^*(x)) = \min_\gamma \mathcal{L}(x; \gamma)$. We now fix x_0 and show that $x \mapsto \gamma^*(x)$ is C^k on a neighbourhood of x_0. Set $\Psi(x; \gamma) := \nabla_\gamma \mathcal{L}(x; \gamma)$; the smoothness of $\mathcal{L}$ ensures that $\Psi \in C^k$. Fix $(x_0; \gamma_0)$ such that $\Psi(x_0; \gamma_0) = 0$. Since $\nabla_\gamma \Psi(x; \gamma) = \nabla^2_\gamma \mathcal{L}(x; \gamma)$ and $\mathcal{L}$ is strictly convex, $\nabla_\gamma \Psi(x_0; \gamma_0)$ is invertible. Then, by the implicit function theorem, there exist a subset $U_{x_0} \subset \Delta_n \times \Delta_n$ and a function $\phi : U_{x_0} \to \Delta_n \times \Delta_n$ such that

i) $\phi(x_0) = \gamma_0$

ii) $\Psi(x, \phi(x)) = 0, \qquad \forall x \in U_{x_0}$

iii) $\phi \in C^k(U_{x_0})$.

For each x in U_{x_0}, since $\phi(x)$ is a stationary point for $\mathcal{L}$ and $\mathcal{L}$ is strictly convex, then $\phi(x) = \gamma^*(x)$, which is- recalling the notation set before- (α_*, β_*). By a standard covering argument, (α_*, β_*) is C^k on the interior of $\Delta_n \times \Delta_n$. As this holds true for any k, the optima

(α_*, β_*), and hence S_λ, are C^∞ on the interior of $\Delta_n \times \Delta_n$.

Let us now focus on the smoothness of $\widetilde{S}_\lambda$. Note that when a, b belong to the interior of the simplex, all components are strictly positive. From the characterization of T_λ recalled in Eq. (25), we know $T_{\lambda ij} > 0$ for any $i, j = 1 \dots n, m$. Then, since the logarithm is a smooth function of T_λ, the term $\lambda^{-1} h(T_\lambda)$ is smooth in a and b. This fact combined with the first part of the proof shows the smoothness of $\widetilde{S}_\lambda(a, b) = \langle T_\lambda, M \rangle - \lambda^{-1} h(T_\lambda)$. □

With a similar procedure, the implicit function theorem provides a formula for the gradient of sharp Sinkhorn distance.

Theorem 3. *Let $M \in \mathbb{R}^{n \times m}$ be a cost matrix, $a \in \Delta_n$, $b \in \Delta_m$ and $\lambda > 0$. Let $\mathcal{L}_{a,b}(\alpha, \beta)$ be defined as in Eq. (12), with argmax in (α_*, β_*). Let T_λ be defined as in Eq. (13). Then,*

$$\nabla_a S_\lambda(a, b) = P_{T\Delta_n}(A \; L \mathbb{1}_m + B \; \bar{L}^\top \mathbb{1}_n) \tag{14}$$

where $L = T_\lambda \odot M \in \mathbb{R}^{n \times m}$ is the entry-wise multiplication between T_λ and M and $\bar{L} \in \mathbb{R}^{n \times m-1}$ corresponds to L with the last column removed. The terms $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m-1}$ are

$$[A \; B] = -\lambda \; D \; \left[\; \nabla^2_{(\alpha,\beta)} \mathcal{L}_{a,b}(\alpha_*, \beta_*) \; \right]^{-1}, \tag{15}$$

with $D = [I \; \mathbf{0}]$ the matrix concatenating the $n \times n$ identity matrix I and the matrix $\mathbf{0} \in \mathbb{R}^{n \times m-1}$ with all entries equal to zero. The operator $P_{T\Delta_n}$ denotes the projection onto the tangent plane $T\Delta_n = \{x \in \mathbb{R}^n : \; \sum_{i=1}^n x_i = 0\}$ to the simplex Δ_n.

Proof. Let us adopt the same notation as in the previous proof. Since $\Psi = \nabla_{(\alpha,\beta)} \mathcal{L}$, by a direct computation, Ψ can be written as

$$\Psi(a, b; \alpha, \beta) = \begin{pmatrix} a - C\mathbb{1} \\ b - C^\top \mathbb{1} \end{pmatrix},$$

where C is the $n \times m - 1$ matrix given by $\mathrm{diag}(e^{\lambda \alpha_*}) e^{\lambda \bar{M}} \mathrm{diag}(e^{\lambda \beta_*})$ and $\bar{M}$ is the matrix M with m^{th} column removed. In the following, we keep track of the dependence on a only. Being Ψ the gradient of $\mathcal{L}$, and $\gamma^*(a) = (\alpha_*(a), \beta_*(a))$ a stationary point, we have

$$\Psi(a; \gamma^*(a)) = 0. \tag{26}$$

For the sake of clarity, notice that:

i) $a \in \mathbb{R}^n$;

ii) $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{m-1} \longrightarrow \mathbb{R}$, as we are considering it is a function of a, α , β;

iii) $\Psi(a, \gamma(a)) = \nabla_{\alpha,\beta} \mathcal{L}(a, \gamma(a)) \in \mathbb{R}^{n+m-1 \times 1}$;

iv) $\alpha_* : \mathbb{R}^n \to \mathbb{R}^n$, $\beta_* : \mathbb{R}^n \to \mathbb{R}^{m-1}$, thus $\gamma^* : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^{m-1}$.

Our goal is to derive $\nabla_a \gamma^*(a)$: by matrix differentiation rules [24] and Eq. (26),

$$\nabla_a \Psi(a, \gamma^*(a)) = \nabla_1 \Psi(a, \gamma^*(a)) + \nabla_a \gamma^*(a) \nabla_2 \Psi(a, \gamma^*(a)) = 0. \tag{27}$$

Let us analyse each term: $\nabla_1 \Psi(a, \gamma^*(a)) = [I_n, \mathbf{0}_{n,m-1}]$ is $n \times n+m-1$ matrix with identity and zeros block, and $\nabla_2 \Psi(a, \gamma^*(a)) = \nabla^2_{\alpha,\beta} \mathcal{L}(a, \gamma^*(a)) =: H$ is the Hessian of $\mathcal{L}$ evaluated at $(a, \gamma^*(a))$, which is a $n+m-1 \times n+m-1$ matrix. Together with Eq. (27), this yields

$$\nabla_a \gamma^*(a) = [\nabla_a \alpha_*(a), \nabla_a \beta_*(a)] = -DH^{-1}.$$

For the sake of clarity, note that $\nabla_a \alpha_*(a)$ and $\nabla_a \beta_*(a)$ contains the gradients of the components as columns, i.e.

$$\nabla_a \alpha_* = \begin{pmatrix} \nabla_a \alpha_{*1}, & \nabla_a \alpha_{*2}, & \dots, & \nabla_a \alpha_{*n} \end{pmatrix}$$
$$\nabla_a \beta_* = \begin{pmatrix} \nabla_a \beta_{*1}, & \nabla_a \beta_{*2}, & \dots, & \nabla_a \beta_{*m-1} \end{pmatrix}.$$

Now, since $S_\lambda(a, b) = \langle T_\lambda, M \rangle$ and T_λ corresponds to Eq. (25) a straightforward computation shows that

$$\nabla_a S_\lambda(a, b) = \sum_{i,j=1}^{n,m} \nabla_a T_{\lambda ij} M_{ij} = \lambda \sum_{i,j=1}^{n,m} T_{\lambda ij} M_{ij} \nabla_a \alpha_{*i} + \lambda \sum_{i,j=1}^{n,m-1} T_{\lambda ij} M_{ij} \nabla_a \beta_{*j}.$$

Setting $L := T_\lambda \odot M$, then the formula above can be rewritten as

$$\nabla_a S_\lambda(a, b) = \lambda \sum_{i}^{n} \nabla_a \alpha_{*i} \sum_{j=1}^{m} L_{ij} + \lambda \sum_{j=1}^{m-1} \nabla_a \beta_{*j} \sum_{i=1}^{n} L_{ij},$$

which is exactly

$$\nabla_a S_\lambda(a, b) = \lambda (\nabla_a \alpha_* L \mathbb{1}_m + \nabla_a \beta_* \bar{L}^\top \mathbb{1}_n).$$

Since by definition, the gradient belongs to the tangent space of the domain, and $a \in \Delta_n$, we project on the tangent space of the simplex, recovering $P_{T\Delta_n} \lambda (\nabla_a \alpha_* L \mathbb{1}_m + \nabla_a \beta_* \bar{L}^\top \mathbb{1}_n)$. □

C.1 Massaging the gradient to get an algorithmic-friendly form

In the proof of theorem 3 we have derived a formula for the gradient of Sinkhorn distance. In this section we further manipulate it in order to obtain an algorithmic friendly expression that also points out some interesting bits that were hidden in the formula above. All the notation has already been introduced: from now on, we will drop the λ and denote the optimal plan by T to make the notation neater.

An explicit computation of the second derivatives of $\mathcal{L}$ with respect to α_i and β_j for $i = 1, \dots, n$ and $j = 1, \dots, m-1$ leads to the following identity

$$H = \begin{pmatrix} \operatorname{diag}(T\mathbb{1}) & \bar{T} \\ \bar{T}^\top & \operatorname{diag}(\bar{T}^\top \mathbb{1}) \end{pmatrix} = \begin{pmatrix} \operatorname{diag}(a) & \bar{T} \\ \bar{T}^\top & \operatorname{diag}(\bar{b}) \end{pmatrix}.$$

That is, H is a block matrix and each block can be expressed in terms of the plan T. The block structure can be exploited when it comes to compute the inverse: we have shown that the gradient of the dual potentials is given by

$$[\nabla_a \alpha_*, \nabla_a \beta_*] = -DH^{-1}, \qquad D = [I_n, \mathbf{0}_{n,m-1}].$$

Now, the inverse of a block matrix is again a block matrix, say

$$H^{-1} = \begin{pmatrix} K_1 & K_2 \\ K_3 & K_4 \end{pmatrix}.$$

Then, $[\nabla_a \alpha_*, \nabla_a \beta_*] = -[K_1, K_2]$. By the formula of the block inverse, setting

$$\mathcal{K} = \operatorname{diag}(T\mathbb{1}) - \bar{T}\operatorname{diag}(\bar{T}^\top \mathbb{1})^{-1}\bar{T}^\top,$$

the blocks K_1 and K_2 are given by

$$K_1 = \mathcal{K}^{-1}, \qquad K_2 = -\mathcal{K}^{-1}\bar{T}\operatorname{diag}(\bar{T}^\top \mathbb{1})^{-1}.$$

Note that $\mathcal{K}$ is symmetric and so its inverse. Now, we can rewrite $\lambda(\nabla_a \alpha_* L\mathbb{1}_m + \nabla_a \beta_* \bar{L}^\top \mathbb{1}_n)$, with $L = T \odot M$, as

$$\lambda(-\mathcal{K}^{-1} S\mathbb{1}_m + \mathcal{K}^{-1}\bar{T}\operatorname{diag}(\bar{T}^\top \mathbb{1})^{-1}\bar{L}^\top \mathbb{1}_n)$$

and we conclude that

$$\nabla_a S_\lambda(a, b) = \lambda \cdot \operatorname{solve}(\mathcal{K}, -L\mathbb{1}_m + \bar{T}\operatorname{diag}(\bar{T}^\top \mathbb{1})^{-1}\bar{L}^\top \mathbb{1}_n).$$

Comment on Remark 1: In the recent work [1], it has been shown that Sinkhorn-Knopp algorithm outputs a matrix T_λ whose distance $\|T_\lambda \mathbb{1} - a\|_1 + \|T_\lambda^\top \mathbb{1} - b\|_1$ from the transport polytope $\Pi(a, b)$ is smaller than ϵ in $O(\epsilon^{-2}\log(s/\ell))$ iterations, where $s = \sum_{ij} e^{-\lambda M_{ij}}$ and $\ell = \min_{ij} e^{-\lambda M_{ij}}$. Let us denote by $M_{\max}$ and $M_{\min}$ the maximum and minimum elements of M respectively. Then,

$$\frac{s}{\ell} = \sum_{ij} e^{-\lambda(M_{ij} - M_{\max})} \geq e^{-\lambda(M_{\min} - M_{\max})} \geq 1.$$

This yields the lower bound

$$\log\left(\frac{s}{\ell}\right) \geq c\lambda$$

where c is a constant independent of λ. We can then conclude that Sinkhorn-Knopp algorithm returns a matrix T_λ such that

$$\langle T_\lambda, M\rangle \leq W(a, b) + \epsilon$$

in $O(nm\epsilon^{-2}M_{\max}^2\lambda)$.

D Proofs in 5: Learning with Sinkhorn Loss Functions

We recall the main definition and tools from [11] needed to fully understand what discussed in section 5. The structured prediction estimator recalled in Eq. (18) is derived in [11] for a large class of loss functions $\mathcal{S} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that are referred to as *Structure Encoding Loss Functions* (SELF) and satisfy the following assumption:

Definition 6 (SELF)**.** *Let $\mathcal{Y}$ be a set. A function $\mathcal{S} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a* Structure Encoding Loss Function *(SELF) if there exists a separable Hilbert space $\mathcal{H}_\mathcal{Y}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{Y}}$, a continuous map $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ and a bounded linear operator $V : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ such that*

$$\mathcal{S}(y, y') = \langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_\mathcal{Y}} \qquad y,\, y' \in \mathcal{H}_\mathcal{Y}. \tag{28}$$

While in [11] it has been observed that a wide range of commonly used loss functions are SELF, no such result was known for Sinkhorn loss. This work also provides an answer to this question. In this direction, let us show a first result on smooth function, which will be a key tool in the rest of the analysis. Note that we will use the notation H^r for the Sobolev space W_2^r.

Theorem 7. *(Smooth functions are SELF) Let $\mathcal{Y} = \Delta_n$. Any function $\mathcal{S} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ such that $\mathcal{S} \in C^\infty(\mathcal{Y} \times \mathcal{Y})$ is SELF.*

Proof. By assumption $\mathcal{S} \in C^\infty(\mathcal{Y} \times \mathcal{Y})$. Since $\mathcal{Y}$ is compact,

$$C^\infty(\mathcal{Y} \times \mathcal{Y}) = C^\infty(\mathcal{Y}) \otimes C^\infty(\mathcal{Y}) \subset H^r(\mathcal{Y}) \otimes H^r(\mathcal{Y}), \tag{29}$$

for $r = (n+1)/2$.The Sobolev space $H^r(\mathcal{Y})$ is a Reproducing Kernel Hilbert Space (RKHS) [7] and we denote by $k_y = k(y, \cdot) \in H^r(\mathcal{Y})$ the reproducing kernel. The product space $H^r \otimes H^r$ is also an RKHS with reproducing kernel $K((y_1, y_2), (y_1', y_2')) = k(y_1, y_1')k(y_2, y_2')$, i.e. in general $K_{y,y'} = k_y \otimes k_{y'}$. Since $\mathcal{S} \in H^r \otimes H^r$, by reproducing property there exists a function $V \in H^r \otimes H^r$ such that

$$\mathcal{S}(y, y') = \langle V, k_y \otimes k_{y'} \rangle_{H^s \otimes H^s}.$$

By the isometric isomorphism $H^r \otimes H^r \cong HS(H^r, H^r)$ [27], with $HS(H^r, H^r)$ the space of Hilbert-Schmidt operators from H^r to itself, it holds

$$\mathcal{S}(y, y') = \langle V, k_y \otimes k_{y'} \rangle_{H^r \otimes H^r} = \langle V, k_y \otimes k_{y'} \rangle_{HS} = \mathrm{Tr}(V^* k_y \otimes k_{y'}) = \langle k_{y'}, V^* k_y \rangle_{H^r}, \tag{30}$$

where V^* is the adjoint operator of V. To meet the conditions of definition 6 it remains to show that V^* and k_y are bounded. But k_y is bounded in H^r for any $y \in \mathcal{Y}$ by definition of reproducing kernel and the operator norm $\|V^*\|$ is bounded from above by the Hilbert-Schmidt norm $\|V\|_{HS}$ which is trivially bounded since $V \in HS(H^r, H^r)$. □

Corollary 8. *The regularized and sharp Sinkhorn losses $\widetilde{S}_\lambda$ and $S_\lambda : \Delta_n^\epsilon \times \Delta_n^\epsilon \to \mathbb{R}$ are SELF.*

Proof. Since $\Delta_n^\epsilon \subset \Delta_n$ is compact and $\widetilde{S}_\lambda$, S_λ are C^∞ in the interior on $\Delta_n \times \Delta_n$ by Thm. 2, a direct application of the result above shows that $\widetilde{S}_\lambda$ and S_λ are SELF. □

Summing up these elements, the proof of Thm. 4 easily follows:

Theorem 4 (Universal Consistency). *Let $\mathcal{Y} = \Delta_n^\epsilon$, $\lambda > 0$ and $\mathcal{S}$ be either $\widetilde{S}_\lambda$ or S_λ. Let k be a bounded continuous universal*[2] *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 Eq. (18) trained with $(x_i, y_i)_{i=1}^\ell$ points independently sampled from ρ and $\gamma_\ell = \ell^{-1/4}$. Then*

$$\lim_{\ell \to \infty} \mathcal{E}(\widehat{f}_\ell) \quad = \quad \min_{f : \mathcal{X} \to \mathcal{Y}} \mathcal{E}(f) \quad \textit{with probability } 1. \tag{20}$$

Proof. Since $\widetilde{S}_\lambda$, S_λ are SELF function and Δ_n^ϵ is compact, the result follows from Thm. 4 in [11]. □

We conclude the section with some comments on Thm. 5 and its proof. We have shown that $\widetilde{S}_\lambda$ and S_λ are SELF and can be written as

$$S_\lambda(y, y') = \langle k_y, V k_{y'} \rangle_{H^r(\Delta_n^\epsilon)} \tag{31}$$

with k the reproducing kernel of the Sobolev space $H^r(\Delta_n^\epsilon)$.

Theorem 5 (Learning Rates). *Let $\mathcal{Y} = \Delta_n^\epsilon$, $\lambda > 0$ and $\mathcal{S}$ be either $\widetilde{S}_\lambda$ or S_λ. Let $\mathcal{H} = W_2^{(n+1)/2}(\mathcal{Y})$ and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded continuous reproducing kernel on $\mathcal{X}$ with associated RKHS $\mathcal{F}$. Let $\hat{f}_\ell : \mathcal{X} \to \mathcal{Y}$ be the estimator in Eq. (18) trained with ℓ training points independently sampled from ρ and with $\gamma = \ell^{-1/2}$. If g^* defined in Eq. (22) is such that $g^* \in \mathcal{H} \otimes \mathcal{F}$, then*

$$\mathcal{E}(f) - \min_{f : \mathcal{X} \to \mathcal{Y}} \mathcal{E}(f) \leq c\, \tau^2 \ell^{-1/4} \tag{23}$$

holds with probability $1 - 8e^{-\tau}$ for any $\tau > 0$, with c a constant independent of ℓ and τ.

Proof. The proof substantially takes advantage of the fact that $\widetilde{S}_\lambda$ and S_λ are SELF and inherits the generalization bounds proved in Thm. 5 in [11]. □

Remark 4. *A relevant question is whether the Wasserstein distance could be similarly framed in the setting of structured prediction. However, the argument used to address Sinkhorn distances relies on their smoothness properties and cannot be extended to the Wasserstein distance, which is not differentiable. A completely different approach may still be successful and we will investigate this question in future work.*

E Experiment on MNIST

This last section is a short supplement to section 6. We present a small experiment on the MNIST dataset that has the same flavour as the experiment on GoogleQuickDraw dataset but adresses a more specific target: to evaluate better the quality of the prediction rather than the overall quality of the reconstructed image, we train the SVM classifier trained on a separate dataset made of 2000 examples of lower halves of digits $1, 2, 5, 8, 9$. Since the classificator is trained on lower halves only, we have selected a subset of digits with clearly

[2] This is a standard assumptions for universal consistency (see [35]). Example: $k(x, x') = e^{-\|x - x'\|^2/\sigma}$.

#err	$\widetilde{S}_\lambda$	S_λ
$\widetilde{S}_\lambda$	16	11
S_λ	1	6



Figure 4: (Right) Relative error (see text) for the Sinkhorn estimators on the digit reconstruction problem. (Left) Sample predictions for egularized (First image) and sharp Sinkhorn estimators.

diverse shapes, to disregard any legitimate vagueness. This means that any classification errors will be due to a poor prediction of the lower half.

We performed the reconstruction with both $\widetilde{S}_\lambda$ and S_λ loss. We tested the performance of the two estimators on 100 examples. Fig. 4 reports the *performance* of the two estimators, as follows:

i) the terms on the diagonal presents the number of misclassification of the lower half predicted with $\widetilde{S}_\lambda$ and S_λ losses;

ii) the number on the upper diagonal represents the number of errors occurred in the classification of the prediction with $\widetilde{S}_\lambda$ on those examples that were correctly classified when reconstructed with S_λ;

iii) conversely, the number on the lower diagonal represents the number of errors occurred in the prediction with S_λ on those examples that were correctly classified when reconstructed with $\widetilde{S}_\lambda$.

To be more precise, denote by $L(\widetilde{S}_\lambda)$ the vector with labels predicted by the classifier when tested on the halves of digits predicted with $\widetilde{S}_\lambda$ loss and analogously $L(S_\lambda)$ the vector with labels given by the classifier tested on the halves of images predicted with S_λ loss. Vector L is the vector with the true labels of the test set. Consider two vectors $\tilde{e}^\lambda \in \{0,1\}^{100}$ and $e^\lambda \in \{0,1\}^{100}$ defined as follows:

$$\tilde{e}^\lambda_i = \begin{cases} 0 & \text{if } L_i = L(\widetilde{S}_\lambda)_i \\ 1 & \text{otherwise} \end{cases} \qquad e^\lambda{}_i = \begin{cases} 0 & \text{if } L_i = L(S_\lambda)_i \\ 1 & \text{otherwise.} \end{cases}$$

Table in Fig. 4 corresponds to

$$\begin{pmatrix} \sum_i \tilde{e}^\lambda_i & \sum_i \tilde{e}^\lambda_i (1 - e^\lambda_i) \\ \sum_i e^\lambda_i (1 - \tilde{e}^\lambda_i) & \sum_i e^\lambda_i . \end{pmatrix}.$$

What we observed is the following: since the classifier was trained and tested on the lower halves only, the blurriness in the reconstruction performed with $\widetilde{S}_\lambda$ played a substantial role in the misclassification on digit 5 in favour of digit 8. On the other hand, the sharpness of the reconstruction with S_λ is a major advantage for the correct classification.