Linear Time Sinkhorn Divergences using Positive Features

Meyer Scetbon
CREST, ENSAE,
Institut Polytechnique de Paris,
meyer.scetbon@ensae.fr

Marco Cuturi
Google Brain,
CREST, ENSAE,
cuturi@google.com

Abstract

Although Sinkhorn divergences are now routinely used in data sciences to compare probability distributions, the computational effort required to compute them remains expensive, growing in general quadratically in the size n of the support of these distributions. Indeed, solving optimal transport (OT) with an entropic regularization requires computing a $n \times n$ kernel matrix (the neg-exponential of a $n \times n$ pairwise ground cost matrix) that is repeatedly applied to a vector. We propose to use instead ground costs of the form $c(x, y) = -\log\langle\varphi(x), \varphi(y)\rangle$ where φ is a map from the ground space onto the positive orthant $\mathbb{R}^r_+$, with $r \ll n$. This choice yields, equivalently, a kernel $k(x, y) = \langle\varphi(x), \varphi(y)\rangle$, and ensures that the cost of Sinkhorn iterations scales as $O(nr)$. We show that usual cost functions can be approximated using this form. Additionaly, we take advantage of the fact that our approach yields approximation that remain fully differentiable with respect to input distributions, as opposed to previously proposed adaptive low-rank approximations of the kernel matrix, to train a faster variant of OT-GAN [49].

1 Introduction

Optimal transport (OT) theory [56] plays an increasingly important role in machine learning to compare probability distributions, notably point clouds, discrete measures or histograms [43]. As a result, OT is now often used in graphics [11, 44, 45], neuroimaging [33], to align word embeddings [4, 1, 30], reconstruct cell trajectories [32, 50, 58], domain adaptation [14, 15] or estimation of generative models [5, 49, 26]. Yet, in their original form, as proposed by Kantorovich [34], OT distances are not a natural fit for applied problems: they minimize a network flow problem, with a supercubic complexity $(n^3 \log n)$ [55] that results in an output that is *not* differentiable with respect to the measures' locations or weights [10, §5]; they suffer from the curse of dimensionality [18, 22] and are therefore likely to be meaningless when used on samples from high-dimensional densities.

Because of these statistical and computational hurdles, all of the works quoted above do rely on some form of regularization to smooth the OT problem, and some more specific uses of an entropic penalty, to recover so called Sinkhorn divergences [16]. These divergences are cheaper to compute than regular OT [12, 24], smooth and programmatically differentiable in their inputs [11, 32], and have a better sample complexity [28] while still defining convex and definite pseudometrics [21]. While Sinkhorn divergences do lower OT costs from supercubic down to an embarassingly parallel quadratic cost, using them to compare measures that have more than a few tens of thousands of points in forward mode (less obviously if backward execution is also needed) remains a challenge.

Entropic regularization: starting from ground costs. The definition of Sinkhorn divergences usually starts from that of the ground cost on observations. That cost is often chosen by default to be a q-norm between vectors, or a shortest-path distance on a graph when considering geometric domains [29, 52, 53, 33]. Given two measures supported respectively on n and m points, regularized

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

OT instantiates first a $n \times m$ pairwise matrix of costs C, to solve a linear program penalized by the coupling's entropy. This can be rewritten as a Kullback-Leibler minimization:

$$\min_{\text{couplings } \mathbf{P}} \langle \mathbf{C}, \mathbf{P} \rangle - \varepsilon H(\mathbf{P}) = \varepsilon \min_{\text{couplings } \mathbf{P}} \mathrm{KL}(\mathbf{P} \| \mathbf{K}), \tag{1}$$

where matrix K appearing in Eq. (1) is defined as $\mathbf{K} := \exp(-\mathbf{C}/\varepsilon)$, the elementiwe neg-exponential of a rescaled cost $\mathbf{C}$.As described in more detail in §2, this problem can then be solved using Sinkhorn's algorithm, which only requires applying repeatedly kernel $\mathbf{K}$ to vectors. While faster optimization schemes to compute regularized OT have been been investigated [2, 19, 37], the Sinkhorn algorithm remains, because of its robustness and simplicity of its parallelism, the workhorse of choice to solve entropic OT. Since Sinkhorn's algorithm cost is driven by the cost of applying $\mathbf{K}$ to a vector, speeding up that evaluation is the most impactful way to speedup Sinkhorn's algorithm. This is the case when using separable costs on grids (applying $\mathbf{K}$ boils down to carrying out a convolution at cost $(n^{1+1/d})$ [43, Remark 4.17]) or when using shortest path metrics on graph in which case applying $\mathbf{K}$ can be approximated using a heat-kernel [54]. While it is tempting to use low-rank matrix factorization, using them within Sinkhorn iterations requires that the application of the approximated kernel guarantees the positiveness of the output. As shown by [3] this can only be guaranteed, when using the Nyström method, when regularization is high and tolerance very low.

Starting instead from the Kernel. Because regularized OT can be carried out using only the definition of a kernel $\mathbf{K}$, we focus instead on kernels $\mathbf{K}$ that are guaranteed to have positive entries by design. Indeed, rather than choosing a cost to define a kernel next, we consider instead ground costs of the form $c(x, y) = -\varepsilon \log\langle \varphi(x), \varphi(y) \rangle$ where φ is a map from the ground space onto the positive orthant in $\mathbb{R}^r$. This choice ensures that both the Sinkhorn algorithm itself (which can approximate optimal primal and dual variables for the OT problem) and the evaluation of Sinkhorn divergences can be computed exactly with an effort scaling linearly in r and in the number of points, opening new perspectives to apply OT at scale.

Our contributions are two fold: *(i)* We show that kernels built from positive features can be used to approximate some usual cost functions including the square Euclidean distance using random expansions. *(ii)* We illustrate the versatility of our approach by extending previously proposed OT-GAN approaches [49, 28], that focused on learning adversarially cost functions c_θ and incurred therefore a quadratic cost, to a new approach that learns instead adversarially a kernel k_θ induced from a positive feature map φ_θ. We leverage here the fact that our approach is fully differentiable in the feature map to train a GAN at scale, with linear time iterations.

Notations. Let $\mathcal{X}$ be a compact space endowed with a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and denote $D = \sup_{(x,y) \in \mathcal{X} \times \mathcal{X}} \|(x, y)\|_2$. We denote $\mathcal{P}(\mathcal{X})$ the set of probability measures on $\mathcal{X}$. For all $n \geq 1$, we denote by Δ_n all vectors in $\mathbb{R}^n_+$ with positive entries and summing to 1. We denote $f \in \mathcal{O}(g)$ if $f \leq Cg$ for a universal constant C and $f \in \Omega(g)$ if $g \leq Qf$ for a universal constant Q.

2 Regularized Optimal Transport

Sinkhorn Divergence. Let $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$ be two discrete probability measures. The Sinkhorn divergence [48, 27, 49] between μ and ν is, given a constant $\varepsilon > 0$, equal to

$$\overline{W}_{\varepsilon,c}(\mu, \nu) := W_{\varepsilon,c}(\mu, \nu) - \frac{1}{2}\left(W_{\varepsilon,c}(\mu, \mu) + W_{\varepsilon,c}(\nu, \nu)\right), \text{ where} \tag{2}$$

$$W_{\varepsilon,c}(\mu, \nu) := \min_{\substack{P \in \mathbb{R}^{n \times m}_+ \\ P\mathbf{1}_m = a, P^T \mathbf{1}_n = b}} \langle P, C \rangle - \varepsilon H(P) + \varepsilon. \tag{3}$$

Here $\mathbf{C} := [c(x_i, y_j)]_{ij}$ and H is the Shannon entropy, $H(\mathbf{P}) := -\sum_{ij} P_{ij}(\log P_{ij} - 1)$. Because computing and differentiating $\overline{W}_{\varepsilon,c}$ is equivalent to doing so for three evaluations of $W_{\varepsilon,c}$ (neglecting the third term in the case where only μ is a variable) [43, §4], we focus on $W_{\varepsilon,c}$ in what follows.

Primal Formulation. Problem (3) is ε-strongly convex and admits therefore a unique solution $\mathbf{P}^\star$ which, writing first order conditions for problem (3), admits the following factorization:

$$\exists u^\star \in \mathbb{R}^n_+, v^\star \in \mathbb{R}^m_+ \text{ s.t. } \mathbf{P}^\star = \mathrm{diag}(u^\star)\mathbf{K}\mathrm{diag}(v^\star), \text{ where } \mathbf{K} := \exp(-\mathbf{C}/\varepsilon). \tag{4}$$

These *scalings* $u^\star, v^\star$ can be computed using Sinkhorn's algorithm, which consists in initializing u to any arbitrary positive vector in $\mathbb{R}^m$, to apply then fixed point iteration described in Alg. 1.

These two iterations require together $2nm$ operations if $\mathbf{K}$ is stored as a matrix and applied directly. The number of Sinkhorn iterations needed to converge to a precision δ (monitored by the difference between the column-sum of $\text{diag}(u)\mathbf{K}\text{diag}(v)$ and b) is controlled by the scale of elements in C relative to ε [23]. That convergence deteriorates with smaller ε, as studied in more detail by [57, 20].

Algorithm 1 Sinkhorn

Inputs: $\mathbf{K}, a, b, \delta, u$ **repeat**
$\quad v \leftarrow b/\mathbf{K}^T u,\ u \leftarrow a/\mathbf{K}v$
until $\|v \circ \mathbf{K}^T u - b\|_1 < \delta$;
Result: u, v

Dual Formulation. The dual of (3) plays an important role in our analysis [43, §4.4]:

$$W_{\varepsilon,c}(\mu,\nu) = \max_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^m} a^T\alpha + b^T\beta - \varepsilon(e^{\alpha/\varepsilon})^T\mathbf{K}e^{\beta/\varepsilon} + \varepsilon = \varepsilon\left(a^T\log u^\star + b^T\log v^\star\right) \quad (5)$$

where we have introduced, next to its definition, its evaluation using optimal scalings $u^\star$ and $v^\star$ described above. This equality comes from that fact that *(i)* one can show that $\alpha^\star := \varepsilon\log u^\star,\ \beta^\star := \varepsilon\log v^\star$, *(ii)* the term $(e^{\alpha/\varepsilon})^T\mathbf{K}e^{\beta/\varepsilon} = u^T\mathbf{K}v$ is equal to 1, whenever the Sinkhorn loop has been applied even just once, since these sums describe the sum of a coupling (a probability distribution of size $n\times m$). As a result, given the outputs u, v of Alg. 1 we estimate (3) using

$$\widehat{W}_{\varepsilon,c}(\mu,\nu) = \varepsilon\left(a^T\log u + b^T\log v\right). \quad (6)$$

Approximating $W_{\varepsilon,c}(\mu,\nu)$ can be therefore carried using exclusively calls to the Sinkhorn algorithm, which requires instantiating kernel $\mathbf{K}$, in addition to computing inner product between vectors, which can be computed in $\mathcal{O}(n+m)$ algebraic operations; the instantiation of $\mathbf{C}$ is never needed, as long as $\mathbf{K}$ is given. Using this dual formulation(3) we can now focus on kernels that can be evaluated with a linear cost to achieve linear time Sinkhorn divergences.

3 Linear Sinkhorn with Positive Features

The usual flow in transport dictates to choose a cost first $c(x,y)$ to define a kernel $k(x,y) := \exp(-c(x,y)/\varepsilon)$ next, and adjust the temperature ε depending on the level of regularization that is adequate for the task. We propose in this work to do exactly the opposite, by choosing instead parameterized feature maps $\varphi_\theta : \mathcal{X} \mapsto (\mathbb{R}_+^*)^r$ which associate to any point in $\mathcal{X}$ a vector in the positive orthant. With such maps, we can therefore build the corresponding positive-definite kernel k_θ as $k_\theta(x,y) := \varphi_\theta(x)^T\varphi_\theta(y)$ which is a positive function. Therefore as a by-product and by positivity of the feature map, we can define for all $(x,y)\in\mathcal{X}\times\mathcal{X}$ the following cost function

$$c_\theta(x,y) := -\varepsilon\log\varphi_\theta(x)^T\varphi_\theta(y). \quad (7)$$

Remark 1 (Transport on the Positive Sphere.)**.** *Defining a cost as the log of a dot-product as described in* (7) *has already played a role in the recent OT literature. In [42], the author defines a cost c on the sphere $\mathbb{S}^d$, as $c(x,y) = -\log x^T y$, if $x^T y > 0$, and ∞ otherwise. The cost is therefore finite whenever two normal vectors share the same halfspace, and infinite otherwise. When restricted to the the positive sphere, the kernel associated to this cost is the linear kernel. See App. C for an illustration.*

More generally, the above procedure allows us to build cost functions on any cartesian product spaces $\mathcal{X}\times\mathcal{Y}$ by defining $c_{\theta,\gamma}(x,y) := -\varepsilon\log\varphi_\theta(x)^T\psi_\gamma(y)$ where $\psi_\gamma : \mathcal{Y}\mapsto(\mathbb{R}_+^*)^r$ is a parametrized function which associates to any point $\mathcal{Y}$ also a vector in the same positive orthant as the image space of φ_θ but this is out of the scope of this paper.

3.1 Achieving linear time Sinkhorn iterations with Positive Features

Choosing a cost function c_θ as in (7) greatly simplifies computations, by design, since one has, writing for the matrices of features for two set of points $x_1,\dots,x_n$ and $y_1,\dots,y_m$

$$\boldsymbol{\xi} := [\varphi_\theta(x_1),\dots,\varphi_\theta(x_n)] \in (\mathbb{R}_+^*)^{r\times n}, \qquad \boldsymbol{\zeta} := [\varphi_\theta(y_1),\dots,\varphi_\theta(y_m)] \in (\mathbb{R}_+^*)^{r\times m},$$

that the resulting sample kernel matrix $\mathbf{K}_\theta$ corresponding to the cost c_θ is $\mathbf{K}_\theta = \left[e^{-c_\theta(x_i,y_j)/\varepsilon}\right]_{i,j} = \boldsymbol{\xi}^T\boldsymbol{\zeta}$. Moreover thanks to the positivity of the entries of the kernel matrix $\mathbf{K}_\theta$ there is no duality gap and we obtain that

$$W_{\varepsilon,c_\theta}(\mu,\nu) = \max_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^m} a^T\alpha + b^T\beta - \varepsilon(\boldsymbol{\xi}e^{\alpha/\varepsilon})^T\boldsymbol{\zeta}e^{\beta/\varepsilon} + \varepsilon. \quad (8)$$

Therefore the Sinkhorn iterations in Alg. 1 can be carried out in exactly $r(n+m)$ operations. The main question remains on how to choose the mapping φ_θ. In the following, we show that, for some well chosen mappings φ_θ, we can approximate the ROT distance for some classical costs in linear time.

3.2 Approximation properties of Positive Features

Let $\mathcal{U}$ be a metric space and ρ a probability measure on $\mathcal{U}$. We consider kernels on $\mathcal{X}$ of the form:

$$\text{for } (x,y) \in \mathcal{X}^2,\ k(x,y) = \int_{u\in\mathcal{U}} \varphi(x,u)^T \varphi(y,u) d\rho(u). \tag{9}$$

Here $\varphi : \mathcal{X} \times \mathcal{U} \to (\mathbb{R}_+^*)^p$ is such that for all $x \in \mathcal{X}$, $u \in \mathcal{U} \to \|\varphi(x,u)\|_2$ is square integrable (for the measure $d\rho$). Given such kernel and a regularization ε we define the cost function $c(x,y) := -\varepsilon \log(k(x,y))$. In fact, we will see in the following that for some usual cost functions $\tilde{c}$, e.g. the square Euclidean cost, the Gibbs kernel associated $\tilde{k}(x,y) = \exp(-\varepsilon^{-1}\tilde{c}(x,y))$ admits a decomposition of the form Eq.(9). To obtain a finite-dimensional representation, one can approximate the integral with a weighted finite sum. Let $r \geq 1$ and $\theta := (u_1, ..., u_r) \in \mathcal{U}^r$ from which we define the following positive feature map

$$\varphi_\theta(x) := \frac{1}{\sqrt{r}} \left(\varphi(x,u_1), ..., \varphi(x,u_r)\right) \in \mathbb{R}^{p\times r}$$

and a new kernel as $k_\theta(x,y) := \langle \varphi_\theta(x), \varphi_\theta(y)\rangle$. When the $(u_i)_{1\leq i\leq r}$ are sampled independently from ρ, k_θ may approximates the kernel k arbitrary well if the number of random features r is sufficiently large. For that purpose let us now introduce some assumptions on the kernel k.

Assumption 1. *There exists a constant $\psi > 0$ such that for all $x, y \in \mathcal{X}$:*

$$|\varphi(x,u)^T \varphi(y,u)/k(x,y)| \leq \psi \tag{10}$$

Assumption 2. *There exists a $\kappa > 0$ such that for ally $x, y \in \mathcal{X}$, $k(x,y) \geq \kappa > 0$ and φ is differentiable there exists $V > 0$ such that:*

$$\sup_{x\in\mathcal{X}} \mathbf{E}_\rho \left(\|\nabla_x \varphi(x,u)\|^2\right) \leq V \tag{11}$$

We can now present our main result on our proposed approximation scheme of $W_{\varepsilon,c}$ which is obtained in linear time with high probability. See Appendix A.1 for the proof.

Theorem 3.1. *Let $\delta > 0$ and $r \geq 1$. Then the Sinkhorn Alg. 1 with inputs $\mathbf{K}_\theta$, a and b outputs (u_θ, v_θ) such that $|W_{\varepsilon,c_\theta} - \widehat{W}_{\varepsilon,c_\theta}| \leq \frac{\delta}{2}$ in $\mathcal{O}\left(\frac{n\varepsilon r}{\delta}\left[Q_\theta - \log\min_{i,j}(a_i,b_j)\right]^2\right)$ algebric operations where $Q_\theta = -\log\min_{i,j} k_\theta(x_i,y_j)$. Moreover if Assumptions 1 and 2 hold then for $\tau > 0$,*

$$r \in \Omega\left(\frac{\psi^2}{\delta^2}\left[\min\left(d\varepsilon^{-1}\|\mathbf{C}\|_\infty^2 + d\log\left(\frac{\psi V D}{\tau\delta}\right), \log\left(\frac{n}{\tau}\right)\right)\right]\right) \tag{12}$$

and $u_1, ..., u_r$ drawn independently from ρ, with a probability $1-\tau$, $Q_\theta \leq \varepsilon^{-1}\|\mathbf{C}\|_\infty^2 + \log\left(2+\delta\varepsilon^{-1}\right)$ and it holds

$$|W_{\varepsilon,c} - \widehat{W}_{\varepsilon,c_\theta}| \leq \delta \tag{13}$$

Therefore with a probability $1-\tau$, Sinkhorn Alg. 1 with inputs $\mathbf{K}_\theta$, a and b output a δ-approximation of the ROT distance in $\tilde{\mathcal{O}}\left(\frac{n}{\varepsilon\delta^3}\|\mathbf{C}\|_\infty^4\psi^2\right)$ algebraic operation where the notation $\tilde{\mathcal{O}}(.)$ omits polylogarithmic factors depending on R, D, ε, n and δ.

It worth noting that for every $r \geq 1$ and θ, Sinkhorn Alg. 1 using kernel matrix $\mathbf{K}_\theta$ will converge towards an approximate solution of the ROT problem associated with the cost function c_θ in linear time thanks to the positivity of the feature maps used. Moreover, to ensure with high probability that the solution obtained approximate an optimal solution for the ROT problem associated with the cost function c, we need, if the features are chosen randomly, to ensure a minimum number of them. In constrast such result is not possible in [3]. Indeed in their works, the number of random features r cannot be chosen arbitrarily as they need to ensure the positiveness of the all the coefficients of the approximated kernel matrix obtained by the Nyström algorithm of [40] to run the Sinkhorn iterations and therefore need a very high precision which requires a certain number of random features r.

Remark 2 (Acceleration.)**.** *It is worth noting that our method can also be applied in combination with the accelerated version of the Sinkhorn algorithm proposed in [31]. Indeed for $\tau > 0$, applying our approximation scheme to their algorithm leads with a probability $1-\tau$ to a $\delta/2$-approximation of $W_{\varepsilon,c}$ in $\mathcal{O}\left(\frac{nr}{\sqrt{\delta}}[\sqrt{\varepsilon^{-1}}A_\theta]\right)$ algebraic operations where $A_\theta = \inf_{(\alpha,\beta)\in\Theta_\theta} \|(\alpha,\beta)\|_2$, Θ_θ is the set of optimal dual solutions of (8) and r satisfying Eq.(12). See the full statement and the proof in Appendix A.2.*

The number of random features prescribed in Theorem 3.1 ensures with high probability that $\widehat{W_{\varepsilon,c_\theta}}$ approximates $W_{\varepsilon,c}$ well when $u_1,\dots,u_r$ are drawn independently from ρ. Indeed, to control the error due to the approximation made through the Sinkhorn iterations, we need to control the error of the approximation of $\mathbf{K}$ by $\mathbf{K}_\theta$ relatively to $\mathbf{K}$. In the next proposition we show with high probability that for all $(x,y)\in\mathcal{X}\times\mathcal{X}$,

$$(1-\delta)k(x,y) \le k_\theta(x,y) \le (1+\delta)k(x,y) \tag{14}$$

for an arbitrary $\delta > 0$ as soon as the number of random features r is large enough. See Appendix A.3 for the proof.

Proposition 3.1. *Let $\mathcal{X}\subset\mathbb{R}^d$ compact, $n\ge 1$, $\mathbf{X}=\{x_1,...,x_n\}$ and $\mathbf{Y}=\{y_1,...,y_n\}$ such that $\mathbf{X},\mathbf{Y}\subset\mathcal{X}$, $\delta>0$. If $u_1,...,u_r$ are drawn independently from ρ then under Assumption 1 we have*

$$\mathbb{P}\left(\sup_{(x,y)\in\mathbf{X}\times\mathbf{Y}}\left|\frac{k_\theta(x,y)}{k(x,y)}-1\right|\ge\delta\right)\le 2n^2\exp\left(-\frac{r\delta^2}{2\psi^2}\right)$$

Moreover if in addition Assumption 2 holds then we have

$$\mathbb{P}\left(\sup_{(x,y)\in\mathcal{X}\times\mathcal{X}}\left|\frac{k_\theta(x,y)}{k(x,y)}-1\right|\ge\delta\right)\le\frac{(\kappa^{-1}D)^2C_{\psi,V,r}}{\delta^2}\exp\left(-\frac{r\delta^2}{2\psi^2(d+1)}\right)$$

where $C_{\psi,V,r}=2^9\psi(4+\psi^2/r)V\sup_{x\in\mathcal{X}}k(x,x)$ and $D=\sup_{(x,y)\in\mathcal{X}\times\mathcal{X}}\|(x,y)\|_2$.

Remark 3 (Ratio Approximation.)**.** *The uniform bound obtained here to control the ratio gives naturally a control of the form Eq.(14). In comparison, in [47], the authors obtain a uniform bound on their difference which leads with high probability to a uniform control of the form*

$$k(x,y)-\tau\le k_\theta(x,y)\le k(x,y)+\tau \tag{15}$$

where τ is a decreasing function with respect to r the number of random features required. To be able to recover Eq.(14) from the above control, one may consider the case when $\tau=\inf_{x,y\in\mathbf{X}\times\mathbf{Y}}k(x,y)\delta$ which can considerably increases the number of of random features r needed to ensure the result with at least the same probability. For example if the kernel is the Gibbs kernel associated to a cost function c, then $\inf_{x,y\in\mathbf{X}\times\mathbf{Y}}k(x,y)=\exp(-\|\mathbf{C}\|_\infty/\varepsilon)$. More details are left in Appendix A.3.

In the following, we provides examples of some usual kernels k that admits a decomposition of the form Eq.(9), satisfy Assumptions 1 and 2 and hence for which Theorem 3.1 can be applied.

Arc-cosine Kernels. Arc-cosine kernels have been considered in several works, starting notably from [51], [13] and [6]. The main idea behind arc-cosine kernels is that they can be written using positive maps for vectors x,y in $\mathbb{R}^d$ and the signs (or higher exponent) of random projections $\mu=\mathcal{N}(0,I_d)$

$$k_s(x,y)=\int_{\mathbb{R}^d}\Theta_s(u^Tx)\Theta_s(u^Ty)d\mu(u)$$

where $\Theta_s(w)=\sqrt{2}\max(0,w)^s$ is a rectified polynomial function. In fact from these formulations, we build a perturbed version of k_s which admits a decomposition of the form Eq.(9) that satisfies the required assumptions. See Appendix A.5 for the full statement and the proof.

Gaussian kernel. The Gaussian kernel is in fact an important example as it is both a very widely used kernel on its own and its cost function associated is the square Euclidean metric. A decomposition of the form (9) has been obtained in ([39]) for the Gaussian kernel but it does not satisfies the required assumptions. In the following lemma, we built a feature map of the Gaussian kernel that satisfies them. See Appendix A.4 for the proof.

Lemma 1. *Let $d \geq 1$, $\varepsilon > 0$ and k be the kernel on $\mathbb{R}^d$ such that for all $x, y \in \mathbb{R}^d$, $k(x,y) = e^{-\|x-y\|_2^2/\varepsilon}$. Let $R > 0$, $q = \frac{R^2}{2\varepsilon d W_0(R^2/\varepsilon d)}$ where W_0 is the Lambert function, $\sigma^2 = q\varepsilon/4$, $\rho = \mathcal{N}\left(0, \sigma^2 Id\right)$ and let us define for all $x, u \in \mathbb{R}^d$ the following map*

$$\varphi(x,u) = (2q)^{d/4} \exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right) \exp\left(\frac{\varepsilon^{-1}\|u\|_2^2}{\frac{1}{2} + \varepsilon^{-1}R^2}\right)$$

Then for any $x, y \in \mathbb{R}^d$ we have $k(x,y) = \int_{u \in \mathbb{R}^d} \varphi(x,u)\varphi(y,u)d\rho(u)$. Moreover if $x, y \in \mathcal{B}(0,R)$ and $u \in \mathbb{R}^d$ we have $k(x,y) \geq \exp(-4\varepsilon^{-1}R^2) > 0$,

$$|\varphi(x,u)\varphi(y,u)/k(x,y)| \leq 2^{d/2+1}q^{d/2} \quad \text{and} \quad \sup_{x \in \mathcal{B}(0,R)} \mathbf{E}(\|\nabla_x \varphi\|_2^2) \leq 2^{d/2+3}q^{d/2}\left[(R/\varepsilon)^2 + \frac{q}{4\varepsilon}\right].$$

3.3 Constructive approach to Designing Positive Features: Differentiability

In this section we consider a constructive way of building feature map φ_θ which may be chosen arbitrary, or learned accordingly to an objective defined as a function of the ROT distance, e.g. OT-GAN objectives [49, 25]. For that purpose, we want to be able to compute the gradient of $W_{\varepsilon,c_\theta}(\mu,\nu)$ with respect to the kernel $\mathbf{K}_\theta$, or more specifically with respect to the parameter θ and the locations of the input measures. In the next proposition we show that the ROT distance is differentiable with respect to the kernel matrix. See Appendix B for the proof.

Proposition 3.2. *Let $\epsilon > 0$, $(a,b) \in \Delta_n \times \Delta_m$ and let us also define for any $\mathbf{K} \in (\mathbb{R}_+^*)^{n \times m}$ with positive entries the following function:*

$$G(\mathbf{K}) := \sup_{(\alpha,\beta) \in \mathbb{R}^n \times \mathbb{R}^m} \langle \alpha, a \rangle + \langle \beta, a \rangle - \varepsilon (e^{\alpha/\varepsilon})^T \mathbf{K} e^{\beta/\varepsilon}. \tag{16}$$

Then G is differentiable on $(\mathbb{R}_+^*)^{n \times m}$ and its gradient is given by

$$\nabla G(\mathbf{K}) = -\varepsilon e^{\alpha^*/\varepsilon}(e^{\beta^*/\varepsilon})^T \tag{17}$$

where (α^*, β^*) are optimal solutions of Eq.(16).

Note that when c is the square euclidean metric, the differentiability of the above objective has been obtained in [17]. We can now provide the formula for the gradients of interest. For all $\mathbf{X} := [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$, we denote $\mu(\mathbf{X}) = \sum_{i=1}^n a_i \delta_{x_i}$ and $W_{\varepsilon,c_\theta} = W_{\varepsilon,c_\theta}(\mu(\mathbf{X}), \nu)$. Assume that θ is a M-dimensional vector for simplicity and that $(x,\theta) \in \mathbb{R}^d \times \mathbb{R}^M \to \varphi_\theta(x) \in (\mathbb{R}_+^*)^r$ is a differentiable map. Then from proposition 3.2 and by applying the chain rule theorem, we obtain that

$$\nabla_\theta W_{\varepsilon,c_\theta} = -\varepsilon \left(\left(\frac{\partial \boldsymbol{\xi}}{\partial \theta}\right)^T u_\theta^\star (\boldsymbol{\zeta} v_\theta^\star)^T + \left(\frac{\partial \boldsymbol{\zeta}}{\partial \theta}\right)^T v_\theta^\star (\boldsymbol{\xi} u_\theta^\star)^T \right), \quad \nabla_X W_{\varepsilon,c_\theta} = -\varepsilon \left(\frac{\partial \boldsymbol{\xi}}{\partial X}\right)^T u_\theta^\star (\boldsymbol{\zeta} v_\theta^\star)^T$$

where (u_θ^*, v_θ^*) are optimal solutions of (5) associated to the kernel matrix $\mathbf{K}_\theta$. Note that $\left(\frac{\partial \boldsymbol{\xi}}{\partial \theta}\right)^T$, $\left(\frac{\partial \boldsymbol{\zeta}}{\partial \theta}\right)^T$ and $\left(\frac{\partial \boldsymbol{\xi}}{\partial X}\right)^T$ can be evaluated using simple differentiation if φ_θ is a simple random feature, or, more generally, using automatic differentiation if φ_θ is the output of a neural network.

Discussion. Our proposed method defines a kernel matrix $\mathbf{K}_\theta$ and a parametrized ROT distance W_{ε,c_θ} which are differentiable with respect to the input measures and the parameter θ. These proprieties are important and used in many applications, e.g. GANs. However such operations may not be allowed when using a data-dependent method to approximate the kernel matrix such as the Nyström method used in [3]. Indeed there, the approximated kernel $\widetilde{\mathbf{K}}$ and the ROT distance $W_{\varepsilon,\widetilde{c}}$ associated are not well defined on a neighbourhood of the locations of the inputs measures and therefore are not differentiable.

4 Experiments

Efficiency vs. Approximation trade-off using positive features. In Figures 1,3 we plot the deviation from ground truth, defined as $\mathrm{D} := 100 \times \frac{\mathrm{ROT} - \widetilde{\mathrm{ROT}}}{|\mathrm{ROT}|} + 100$, and show the time-accuracy

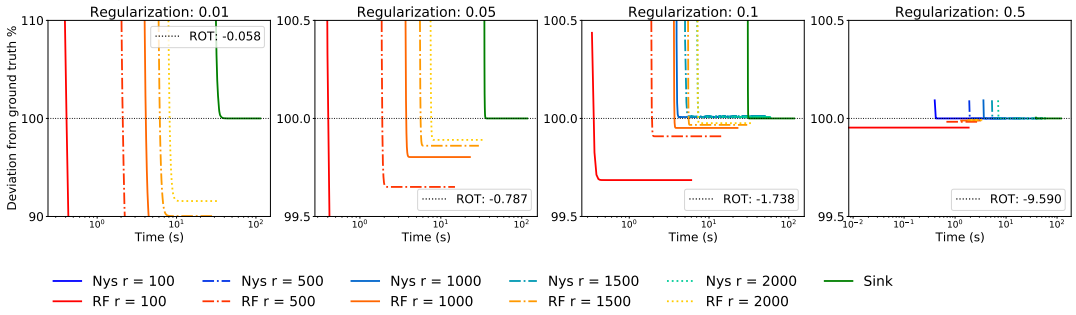

Figure 1: In this experiment, we draw 40000 samples from two normal distributions and we plot the deviation from ground truth for different regularizations. These two normal distributions are in $\mathbb{R}^2$. One of them has mean $(1,1)^T$ and identity covariance matrix I_2. The other has 0 mean and covariance $0.1 \times I_2$. We compare the results obtained for our proposed method (**RF**) with the one proposed in [3] (**Nys**) and with the Sinkhorn algorithm (**Sin**) proposed in [16]. The cost function considered here is the square Euclidean metric and the feature map used is that presented in Lemma 1. The number of random features (or rank) chosen varies from 100 to 2000. We repeat for each problem 50 times the experiment. Note that curves in the plot start at different points corresponding to the time required for initialization. *Right*: when the regularization is sufficiently large both **Nys** and **RF** methods obtain very high accuracy with order of magnitude faster than **Sin**. *Middle right, middle left*: **Nys** fails to converge while **RF** works for any given random features and provides very high accuracy of the ROT cost with order of magnitude faster than **Sin**. *Left*: when the regularization is too small all the methods failed as the Nystrom method cannot be computed, the accuracy of the **RF** method is of order of 10% and Sinkhorn algorithm may be too costly.

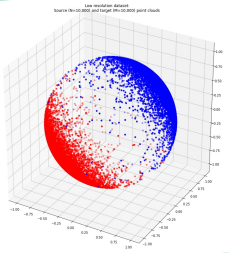

Figure 2: Here we show the two distributions considered in the experiment presented in Figure 3 to compare the time-accuracy tradeoff between the different methods. All the points are drawn on the unit sphere in $\mathbb{R}^3$, and uniform distributions are considered respectively on the red dots and on the blue dots. There are 10000 samples for each distribution.

tradeoff for our proposed method **RF**, Nystrom **Nys** [3] and Sinkhorn **Sin** [16], for a range of regularization parameters ε (each corresponding to a different ground truth $W_{\varepsilon,c}$) and approximation with r random features in two settings. In particular, we show that our method obtains very high accuracy with order of magnitude faster than **Sin** in a larger regime of regularizations than **Nys**. In Figure 5 in Appendix C, we also show the time-accuracy tradeoff in the high dimensional setting.

Using positive features to learn adversarial kernels in GANs. Let P_X a given distribution on $\mathcal{X} \subset \mathbb{R}^D$, $(\mathcal{Z}, \mathcal{A}, \zeta)$ an arbitrary probability space and let $g_\rho : \mathcal{Z} \to \mathcal{X}$ a parametric function where the parameter ρ lives in a topological space $\mathcal{O}$. The function g_ρ allows to generate a distribution on $\mathcal{X}$ by considering the push forward operation through g_ρ. Indeed $g_{\rho\sharp}\zeta$ is a distribution on $\mathcal{X}$ and if the function space $\mathcal{F} = \{g_\rho \colon \rho \in \mathcal{O}\}$ is large enough, we may be able to recover P_X for a well chosen ρ. The goal is to learn ρ^* such that $g_{\rho^*\sharp}\zeta$ is the closest possible to P_X according to a specific metric on the space of distributions. Here we consider the Sinkhorn distance as introduced in Eq.(2). One difficulty when using such metric is to define a well behaved cost to measure the distance between distributions in the ground space. We decide to learn an adversarial cost by embedding the native space $\mathcal{X}$ into a low-dimensional subspace of $\mathbb{R}^d$ thanks to a parametric function f_γ. Therefore by defining $h_\gamma(x,y) := (f_\gamma(x), f_\gamma(y))$ and given a fixed cost function c on $\mathbb{R}^d$, we can define a parametric cost function on $\mathcal{X}$ as $c \circ h_\gamma(x,y) := c(f_\gamma(x), f_\gamma(y))$. To train a Generative Adversarial Network (GAN), one may therefore optimizes the following objective:

$$\min_\rho \max_\gamma \overline{W}_{\varepsilon, c\circ h_\gamma}(g_{\rho\#}\zeta, P_X)$$

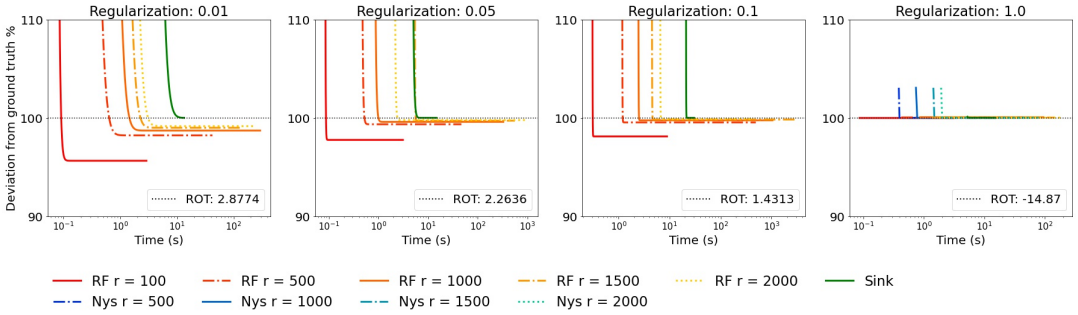

Figure 3: In this experiment, we draw 20000 samples from two distributions on the sphere (see Figure 4) and we plot the deviation from ground truth for different regularizations. We compare the results obtained for our proposed method (**RF**) with the one proposed in [3] (**Nys**) and with the Sinkhorn algorithm (**Sin**) proposed in [16]. The cost function considered here is the square Euclidean metric and the feature map used is that presented in Lemma 1. The number of random features (or rank) chosen varies from 100 to 2000. We repeat for each problem 10 times the experiment. Note that curves in the plot start at different points corresponding to the time required for initialization. *Right*: when the regularization is sufficiently large both **Nys** and **RF** methods obtain very high accuracy with order of magnitude faster than **Sin**. *Middle right, middle left, left*: **Nys** fails to converge while **RF** works for any given random features and provides very high accuracy of the ROT cost with order of magnitude faster than **Sin**.

Indeed, taking the max of the Sinkhorn distance according to γ allows to learn a discriminative cost $c \circ h_\gamma$ [25, 49]. However in practice, we do not have access to the distribution of the data P_X, but only to its empirical version $\widehat{P}_X$, where $\widehat{P}_X := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\mathbf{X} := \{x_1, ..., x_n\}$ are the n i.i.d samples drawn from P_X. By sampling independently n samples $\mathbf{Z} := \{z_1, ..., z_n\}$ from ζ and denoting $\widehat{\zeta} := \frac{1}{q}\sum_{i=1}^q \delta_{z_i}$ we obtain the following approximation:

$$\min_\rho \max_\gamma \overline{W}_{\varepsilon, c\circ h_\gamma}(g_{\rho\#}\widehat{\zeta}, \widehat{P}_X)$$

However as soon as n gets too large, the above objective, using the classic Sinkhorn Alg. 1 is very costly to compute as the cost of each iteration of Sinkhorn is quadratic in the number of samples. Therefore one may instead split the data and consider $B \geq 1$ mini-batches $\mathbf{Z} = (\mathbf{Z}^b)_{b=1}^B$ and $\mathbf{X} = (\mathbf{X}^b)_{b=1}^B$ of size $s = \frac{n}{B}$, and obtain instead the following optimisation problem:

$$\min_\rho \max_\gamma \frac{1}{B}\sum_{b=1}^B \overline{W}_{\varepsilon, c\circ h_\gamma}(g_{\rho\#}\widehat{\zeta}^b, \widehat{P}_X^b)$$

where $\widehat{\zeta}^b := \frac{1}{s}\sum_{i=1}^s \delta_{z_i^b}$ and $\widehat{P}_X^b := \frac{1}{s}\sum_{i=1}^s \delta_{x_i^b}$. However the smaller the batches are, the less precise the approximation of the objective is. To overcome this issue we propose to apply our method and replace the cost function c by an approximation defined as $c_\theta(x, y) = -\epsilon \log \varphi_\theta(x)^T \varphi_\theta(y)$ and consider instead the following optimisation problem:

$$\min_\rho \max_\gamma \frac{1}{B}\sum_{b=1}^B \overline{W}_{\varepsilon, c_\theta\circ h_\gamma}(g_{\rho\#}\widehat{\zeta}^b, \widehat{P}_X^b).$$

Indeed in that case, the Gibbs kernel associated to the cost function $c_\theta \circ h_\gamma$ is still factorizafable as we have $c_\theta \circ h_\gamma(x, y) = -\epsilon \log \varphi_\theta(f_\gamma(x))^T \varphi_\theta(f_\gamma(y))$. Such procedure allows us to compute the objective in linear time and therefore to largely increase the size of the batches. Note that we keep the batch formulation as we still need it because of memory limitation on GPUs. Moreover, we may either consider a random approximation by drawing θ randomly for a well chosen distribution or we could learn the random features θ. In the following we decide to learn the features θ in order to obtain a cost function $c_\theta \circ h_\gamma$ even more discriminative. Finally our objective is:

$$\min_\rho \max_{\gamma,\theta} \frac{1}{B}\sum_{b=1}^B \overline{W}_{\varepsilon, c_\theta\circ h_\gamma}(g_{\rho\#}\widehat{\zeta}^b, \widehat{P}_X^b) \tag{18}$$

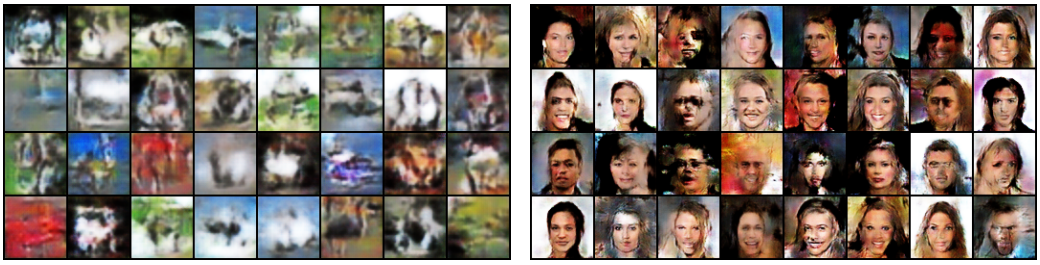

Figure 4: Images generated by two learned generative models trained by optimizing the objective (18) where we set the number of batches $s = 7000$, the regularization $\varepsilon = 1$, and the number of features $r = 600$. *Left, right:* samples obtained from the proposed generative model trained on respectively CIFAR-10 [35] and celebA [38].

Therefore here we aim to learn an embedding from the input space into the feature space thanks to two operations. The first one consists in taking a sample and embedding it into a latent space thanks to the mapping f_γ and the second one is an embedding of this latent space into the feature space thanks to the feature map φ_θ. From now on we assume that g_ρ and f_γ are neural networks. More precisely we take the exact same functions used in [46, 36] to define g_ρ and f_γ. Moreover, φ_θ is the feature map associated to the Gaussian kernel defined in Lemma 1 where θ is initialised with a normal distribution. The number of random features considered has been fixed to be $r = 600$ in the following. The training procedure is the same as [27, 36] and consists in alterning n_c optimisation steps to train the cost function $c_\theta \circ h_\gamma$ and an optimisation step to train the generator g_ρ. The code is available at github.com/meyerscetbon/LinearSinkhorn.

$k_\theta(f_\gamma(x), f_\gamma(z))$	Image x	Noise z
Image x	$1802 \times 1e12$	$2961 \times 1e5$
Noise z	$2961 \times 1e5$	48.65

Table 1: Comparison of the learned kernel k_θ, trained on CIFAR-10 by optimizing the objective (18), between images taken from CIFAR-10 and random noises sampled in the native of space of images. The values shown are averages obtained between 5 noise and/or image samples. As we can see the cost learned has well captured the structure of the image space.

Optimisation. Thanks to proposition 3.2, the objective is differentiable with respect to θ, γ and ρ. We obtain the gradient by computing an approximation of the gradient thanks to the approximate dual variables obtained by the Sinkhorn algorithm. We refers to section 3.3 for the expression of the gradient. This strategy leads to two benefits. First it is memory efficient as the computation of the gradient at this stage does not require to keep track of the computations involved in the Sinkhorn algorithm. Second it allows, for a given regularization, to compute with very high accuracy the Sinkhorn distance. Therefore, our method may be applied also for small regularization.

Results. We train our GAN models on a Tesla K80 GPU for 84 hours on two different datasets, namely CIFAR-10 dataset [35] and CelebA dataset [38] and learn both the proposed generative model and the adversarial cost function c_θ derived from the adversarial kernel k_θ. Figure 4 illustrates the generated samples and Table 1 displays the geometry captured by the learned kernel.

Discussion. Our proposed method has mainly two advantages compared to the other Wasserstein GANs (W-GANs) proposed in the literature. First, the computation of the Sinkhorn divergence is linear with respect to the number of samples which allow to largely increase the batch size when training a W-GAN and obtain a better approximation of the true Sinkhorn divergence. Second, our approach is fully differentiable and therefore we can directly compute the gradient of the Sinhkorn divergence with respect the parameters of the network. In [49] the authors do not differentiate through the Wasserstein cost to train their network. In [25] the authors do differentiate through the iterations of the Sinkhorn algorithm but this strategy require to keep track of the computation involved in the Sinkhorn algorithm and can be applied only for large regularizations as the number of iterations cannot be too large.

Acknowledgements

This work was funded by a "Chaire d'excellence de l'IDEX Paris Saclay".

Broader Impact

Optimal Transport (OT) has gained interest last years in machine learning with many applications in neuroimaging, generative models, supervised learning, word embeddings, reconstruction cell trajectories or adversarial examples. This work brings new applications to OT in the high dimensional setting as it provides a linear time method to compute an approximation of the OT cost and gives a constructive method to learn an adapted kernel or equivalently an adapted cost function depending on the problem considered.

References

[1] Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. Unsupervised hyper-alignment for multilingual word embeddings. In *International Conference on Learning Representations*, 2019.

[2] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *arXiv preprint arXiv:1705.09634*, 2017.

[3] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Weed. Massively scalable sinkhorn distances via the nystr\" om method. *arXiv preprint arXiv:1812.05189*, 2018.

[4] David Alvarez-Melis and Tommi Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, 2018.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:214–223, 2017.

[6] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[7] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5:4308, 2014.

[8] Tamer Başar and Pierre Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

[9] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[10] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

[11] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):71:1–71:10, 2016.

[12] Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Math. Comput.*, 87(314):2563–2609, 2018.

[13] Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 342–350. Curran Associates, Inc., 2009.

[14] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

[15] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9): 1853–1865, 2017.

[16] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

[17] Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *Proceedings of ICML*, volume 32, pages 685–693, 2014.

[18] Richard M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[19] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. *arXiv preprint arXiv:1802.04367*, 2018.

[20] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376. PMLR, 10–15 Jul 2018.

[21] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

[22] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[23] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114:717–735, 1989.

[24] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

[25] Aude Genevay, Gabriel Peyré, and Marco Cuturi. GAN and VAE from an optimal transport point of view. (arXiv preprint arXiv:1706.01807), 2017.

[26] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. *arXiv preprint arXiv:1810.02733*, 2018.

[27] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of AISTATS*, pages 1608–1617, 2018.

[28] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. 2019.

[29] Alexandre Gramfort, Gabriel Peyré, and Marco Cuturi. Fast optimal transport averaging of neuroimaging data. In *Information Processing in Medical Imaging - 24th International Conference, IPMI 2015*, pages 261–272, 2015.

[30] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890, 2019.

[31] Sergey Guminov, Pavel Dvurechensky, Nazarii Tupitsa, and Alexander Gasnikov. Accelerated alternating minimization, accelerated sinkhorn's algorithm and accelerated iterative bregman projections, 2019.

[32] Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning population-level diffusions with generative RNNs. In *International Conference on Machine Learning*, pages 2417–2426, 2016.

[33] Hicham Janati, Thomas Bazeille, Bertrand Thirion, Marco Cuturi, and Alexandre Gramfort. Multi-subject meg/eeg source imaging with sparse multi-task regression. *NeuroImage*, page 116847, 2020.

[34] Leonid Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2): 227–229, 1942.

[35] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

[36] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. *arXiv preprint arXiv:1705.08584*, 2017.

[37] Tianyi Lin, Nhat Ho, and Michael Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3982–3991, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/lin19a.html`.

[38] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[39] Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. Convolutional kernel networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2627–2635. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5348-convolutional-kernel-networks.pdf`.

[40] Cameron Musco and Christopher Musco. Recursive sampling for the nyström method, 2016.

[41] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103 (1):127–152, 2005.

[42] Vladimir Oliker. Embedding S_n into $\mathbb{R}^{n+1}$ with given integral gauss curvature and optimal mass transport on S_n. *Advances in Mathematics*, 213(2):600 – 620, 2007.

[43] Gabriel Peyré and Marco Cuturi. Metric learning: a survey. *Foundations and Trends in Machine Learning*, 11(5-6), 2019.

[44] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672, 2016.

[45] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5099–5108. Curran Associates, Inc., 2017.

[46] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015.

[47] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[48] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[49] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rkQkBnJAb`.

[50] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176 (4):928–943, 2019.

[51] Alex J. Smola, Zoltán L. Óvári, and Robert C Williamson. Regularization with dot-product kernels. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 308–314. MIT Press, 2001.

[52] Justin Solomon, Leonidas Guibas, and Adrian Butscher. Dirichlet energy for analysis and synthesis of soft maps. In *Computer Graphics Forum*, volume 32, pages 197–206. Wiley Online Library, 2013.

[53] Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Earth mover's distances on discrete surfaces. *Transaction on Graphics*, 33(4), 2014.

[54] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4):66:1–66:11, 2015.

[55] Robert E. Tarjan. Dynamic trees as search trees via euler tours, applied to the network simplex algorithm. *Mathematical Programming*, 78(2):169–177, 1997.

[56] Cedric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, 2003. ISBN 9780821833124.

[57] Jonathan Weed, Francis Bach, et al. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

[58] Karren Dai Yang, Karthik Damodaran, Saradha Venkatachalapathy, Ali C Soylemezoglu, GV Shivashankar, and Caroline Uhler. Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16(4):e1007828, 2020.

Supplementary materials

Outline. In Sec. A we provide the proofs related to the approximation proprieties of our proposed method. In Sec. B we show the differentiability of the constructive approach. Finally in Sec. C we add more experiments and illustrations of our proposed method.

A Approximation via Random Fourier Features

A.1 Proof of Theorem 3.1

In the following we denote $\mathbf{K} = (k(x_i, y_j))_{i,j=1}^n$ $\mathbf{K}_\theta = (k_\theta(x_i, y_j))_{i,j=1}^n$ the two gram matrices associated with k and k_θ respectively. By duality and from these two matrices we can define the two objectives to maximize to obtain $W_{\varepsilon,c}$ and W_{ε,c_θ}:

$$W_{\varepsilon,c} = \max_{\alpha,\beta} f(\alpha,\beta) := \langle \alpha, a\rangle + \langle \beta, b\rangle - \varepsilon\langle e^{\alpha/\varepsilon}, \mathbf{K}e^{\beta/\varepsilon}\rangle$$
$$W_{\varepsilon,c_\theta} = \max_{\alpha,\beta} f_\theta(\alpha,\beta) := \langle \alpha, a\rangle + \langle \beta, b\rangle - \varepsilon\langle e^{\alpha/\varepsilon}, \mathbf{K}_\theta e^{\beta/\varepsilon}\rangle$$

Moreover as k and φ are assumed to be positive, there exists unique (up to a scalar translation) (α^*, β^*) and $(\alpha_\theta^*, \beta_\theta^*)$ respectively solutions of $\max_{\alpha,\beta} f(\alpha,\beta)$ and $\max_{\alpha,\beta} f_\theta(\alpha,\beta)$.

Proof. *Let us first show the following proposition:*

Proposition 1. *Let $\delta > 0$ and $r \geq 1$. Assume that for all $(x, y) \in \mathbf{X} \times \mathbf{Y}$,*

$$\left|\frac{k(x,y) - k_\theta(x,y)}{k(x,y)}\right| \leq \frac{\delta\varepsilon^{-1}}{2 + \delta\varepsilon^{-1}} \tag{19}$$

then Sinkhorn Alg. 1 with inputs a, b, K_θ outputs $(\alpha_\theta, \beta_\theta)$ in

$$\mathcal{O}\left(\frac{nr}{\delta\varepsilon^{-1}}\left[\log\left(\frac{1}{\iota}\right) + \log\left(2 + \delta\varepsilon^{-1}\right) + \varepsilon^{-1}R^2\right]^2\right)$$

where

$$\iota = \min_{i,j}(a_i, b_j) \quad \textit{and} \quad R = \max_{(x,y)\in\mathbf{X}\times\mathbf{Y}} c(x,y). \tag{20}$$

such that:

$$|W_{\varepsilon,c} - f_\theta(\alpha_\theta, \beta_\theta)| \leq \delta$$

Proof. *We remark that:*

$$\begin{aligned}|f(\alpha^*,\beta^*) - f_\theta(\alpha_\theta,\beta_\theta)| \leq\ & |f(\alpha^*,\beta^*) - f(\alpha_\theta^*,\beta_\theta^*)| \\ & + |f(\alpha_\theta^*,\beta_\theta^*) - f_\theta(\alpha_\theta^*,\beta_\theta^*)| \\ & + |f_\theta(\alpha_\theta^*,\beta_\theta^*) - f_\theta(\alpha_\theta,\beta_\theta)|\end{aligned}$$

Moreover we have that:

$$\begin{aligned}|f(\alpha^*,\beta^*) - f(\alpha_\theta^*,\beta_\theta^*)| &= f(\alpha^*,\beta^*) - f(\alpha_\theta^*,\beta_\theta^*) \\ &= f(\alpha^*,\beta^*) - f_\theta(\alpha_\theta^*,\beta_\theta^*) + f_\theta(\alpha_\theta^*,\beta_\theta^*) - f(\alpha_\theta^*,\beta_\theta^*) \\ &\leq |f(\alpha^*,\beta^*) - f_\theta(\alpha^*,\beta^*)| + |f_\theta(\alpha_\theta^*,\beta_\theta^*) - f(\alpha_\theta^*,\beta_\theta^*)|\end{aligned}$$

Therefore we obtain that:

$$\begin{aligned}|f(\alpha^*,\beta^*) - f_\theta(\alpha_\theta,\beta_\theta)| \leq\ & 2|f(\alpha_\theta^*,\beta_\theta^*) - f_\theta(\alpha_\theta^*,\beta_\theta^*)| + |f(\alpha^*,\beta^*) - f_\theta(\alpha^*,\beta^*)| \\ & + |f_\theta(\alpha_\theta^*,\beta_\theta^*) - f_\theta(\alpha_\theta,\beta_\theta)|\end{aligned}$$

Let us now introduce the following lemma:

Lemma 2. *Let $1 > \tau > 0$ and let us assume that for all $(x, y) \in \mathbf{X} \times \mathbf{Y}$,*

$$\left|\frac{k(x,y) - k_\theta(x,y)}{k(x,y)}\right| \leq \tau$$

then for any $\alpha, \beta \in \mathbb{R}^n$ it holds

$$|f(\alpha,\beta) - f_\theta(\alpha,\beta)| \leq \varepsilon\tau[\langle e^{\varepsilon^{-1}\alpha}, \mathbf{K}e^{\varepsilon^{-1}\beta}\rangle] \tag{21}$$

and

$$|f(\alpha,\beta) - f_\theta(\alpha,\beta)| \leq \varepsilon\frac{\tau}{1-\tau}[\langle e^{\varepsilon^{-1}\alpha}, \mathbf{K}_\theta e^{\varepsilon^{-1}\beta}\rangle] \tag{22}$$

Proof. *Let $\alpha, \beta \in \mathbb{R}^n$. We remarks that:*

$$f(\alpha,\beta) - f_\theta(\alpha,\beta) = \varepsilon[\langle e^{\varepsilon^{-1}\alpha}, (\mathbf{K}_\theta - \mathbf{K})e^{\varepsilon^{-1}\beta}\rangle] \tag{23}$$

Therefore we obtain that:

$$|f(\alpha,\beta) - f_\theta(\alpha,\beta)| \leq \varepsilon \sum_{i,j=1}^{n} e^{\varepsilon^{-1}\alpha_i} e^{\varepsilon^{-1}\beta_j} |[\mathbf{K}_\theta]_{i,j} - \mathbf{K}_{i,j}| \tag{24}$$

And the first inequality follows from the fact that $|[\mathbf{K}_\theta]_{i,j} - \mathbf{K}_{i,j}| \leq \tau|\mathbf{K}_{i,j}|$ for all $i, j \in \{1, ..., n\}$ and that k is positive. Moreover from the same inequality we obtain that:

$$|[\mathbf{K}_\theta]_{i,j} - \mathbf{K}_{i,j}| \leq \frac{\tau}{1-\tau}[\mathbf{K}_\theta]_{i,j}$$

Therefore the second inequality follows.

Therefore thanks to lemma 2, we obtain that:

$$|f(\alpha_\theta^*,\beta_\theta^*) - f_\theta(\alpha_\theta^*,\beta_\theta^*)| \leq \varepsilon\frac{\tau}{1-\tau}[\langle e^{\varepsilon^{-1}\alpha_\theta^*}, \mathbf{K}_\theta e^{\varepsilon^{-1}\beta_\theta^*}\rangle] \tag{25}$$

But as $(\alpha_\theta^*, \beta_\theta^*)$ is the optimum of f_θ, the first order conditions give us that $\langle e^{\varepsilon^{-1}\alpha_\theta^*}, \mathbf{K}_\theta e^{\varepsilon^{-1}\beta_\theta^*}\rangle = 1$ and finally we have:

$$|f(\alpha_\theta^*,\beta_\theta^*) - f_\theta(\alpha_\theta^*,\beta_\theta^*)| \leq \varepsilon\frac{\tau}{1-\tau} \tag{26}$$

Thanks to lemma 2, we also deduce that:

$$|f(\alpha^*,\beta^*) - f_\theta(\alpha^*,\beta^*)| \leq \varepsilon\tau \tag{27}$$

Let us now introduce the following theorem:

Theorem A.1. *([19]) Given $\mathbf{K}_\theta \in \mathbb{R}^{n\times n}$ with positive entries and $a, b \in \Delta_n$ the Sinkhorn Alg. 1 computes $(\alpha_\theta, \beta_\theta)$ such that*

$$|f_\theta(\alpha_\theta^*,\beta_\theta^*) - f_\theta(\alpha_\theta,\beta_\theta)| \leq \frac{\delta}{2}$$

in $\mathcal{O}\left(\delta^{-1}\varepsilon \log\left(\frac{1}{\iota \min_{i,j}[K_\theta]_{i,j}}\right)^2\right)$ iterations where $\iota = \min_{i,j}(a_i, b_j)$ and each of which requires $\mathcal{O}(1)$ matrix-vector products with K_θ and $\mathcal{O}(n)$ additional processing time.

Moreover from Eq. (19) we have that

$$[\mathbf{K}_\theta]_{i,j} \geq (1-\tau)\mathbf{K}_{i,j} \tag{28}$$

where $\tau = \frac{\delta\varepsilon^{-1}}{2+\delta\varepsilon^{-1}}$, therefore $\log\left(\frac{1}{\min_{i,j}[\mathbf{K}_\theta]_{i,j}}\right) \leq \log\left(\frac{1}{(1-\tau)\min_{i,j}\mathbf{K}_{i,j}}\right) \leq \log\left(\frac{1}{1-\tau}\right) + \varepsilon^{-1}R^2$ where $R = \max_{(x,y)\in\mathbf{X}\times\mathbf{Y}} c(x,y)$ and we obtain that

$$|f(\alpha^*,\beta^*) - f_\theta(\alpha_\theta,\beta_\theta)| \leq 2\varepsilon\frac{\tau}{1-\tau} + \varepsilon\tau + \frac{\delta}{2} \tag{29}$$

By replacing τ by its value, we obtain the desired result.

We are now ready to prove the theorem. Let $r \geq 1$. From theorem A.1, we obtain directly that:

$$|f(\alpha^*, \beta^*) - f_\theta(\alpha_\theta, \beta_\theta)| \leq \frac{\delta}{2} \tag{30}$$

in $\mathcal{O}\left(\frac{nr}{\delta}\left[\log\left(\frac{1}{\iota}\right) + Q_\theta\right]^2\right)$ algebric operations. Moreover let $\tau > 0$ and

$$r \in \Omega\left(\frac{\psi^2}{\delta^2}\left[\min\left(d\varepsilon^{-1}R^2 + d\log\left(\frac{\psi VD}{\tau\delta}\right), \log\left(\frac{n}{\tau}\right)\right)\right]\right)$$

and $u_1, ..., u_r$ drawn independently from ρ. Then from Proposition 3.1 we obtain that with a probability of at least $1-\delta$ it holds for all $(x, y) \in \mathbf{X} \times \mathbf{Y}$,

$$\left|\frac{k(x,y) - k_\theta(x,y)}{k(x,y)}\right| \leq \frac{\delta\varepsilon^{-1}}{2 + \delta\varepsilon^{-1}} \tag{31}$$

and the result follows from Proposition 1.

A.2 Accelerated Version

[31] show that one can accelarated the Sinkhorn algorithm (see Alg. 2) and obtain a δ-approximation of the ROT distance. For that purpose, [31] introduce a reformulation of the dual problem (8) and obtain

$$W_{\varepsilon,c_\theta} = \sup_{\eta_1,\eta_2} F_\theta(\eta_1, \eta_2) := \varepsilon\left[\langle\eta_1, a\rangle + \langle\eta_2, b\rangle - \log\left(\langle\mathbf{K}_\theta e^{\eta_2}\rangle\right)\right] \tag{32}$$

which can be shown to be an L-smooth function ([41]) where $L \leq 2\varepsilon^{-1}$. Let us now present our result using the accelarated Sinkhorn algorithm.

Theorem A.2. *Let $\delta > 0$ and $r \geq 1$. Then the Accelerated Sinkhorn Alg. 2 with inputs $\mathbf{K}_\theta$, a and b outputs $(\alpha_\theta, \beta_\theta)$ such that*

$$|W_{\varepsilon,c_\theta} - F_\theta(\alpha_\theta, \beta_\theta)| \leq \frac{\delta}{2}$$

in $\mathcal{O}\left(\frac{nr}{\sqrt{\delta}}[\sqrt{\varepsilon^{-1}}A_\theta]\right)$ algebraic operations where $A_\theta = \inf_{(\alpha,\beta)\in\Theta_\theta} \|(\alpha, \beta)\|_2$ and Θ_θ is the set of optimal dual solutions of (8). Moreover let $\tau > 0$,

$$r \in \Omega\left(\frac{\psi^2}{\delta^2}\left[\min\left(d\varepsilon^{-1}\|C\|_\infty^2 + d\log\left(\frac{\psi VD}{\delta\delta}\right), \log\left(\frac{n}{\delta}\right)\right)\right]\right) \tag{33}$$

and $u_1, ..., u_r$ drawn independently from ρ, then with a probability $1-\tau$ it holds

$$|W_{\varepsilon,c} - F_\theta(\alpha_\theta, \beta_\theta)| \leq \delta \tag{34}$$

Proof. *Let us first introduce the theorem presented in [31]:*

Theorem A.3. *Given $\mathbf{K}_\theta \in \mathbb{R}^{n\times n}$ with positive entries and $a, b \in \Delta_n$ the Accelerated Sinkhorn Alg. (2) computes $(\alpha_\theta, \beta_\theta)$ such that*

$$|W_{\varepsilon,c_\theta} - F_\theta(\alpha_\theta, \beta_\theta)| \leq \delta$$

in $\mathcal{O}\left(\sqrt{\frac{\eta}{\delta}}A_\theta\right)$ iterations where $A_\theta = \inf_{(\alpha_\theta^*,\beta_\theta^*)\in\Theta^*} \|(\alpha_\theta^*, \beta_\theta^*)\|_2$ and Θ^* is the set of optimal dual solutions. Moreover each of which requires $\mathcal{O}(1)$ matrix-vector products with $\mathbf{K}_\theta$ and $\mathcal{O}(n)$.

From the above result and applying an analogue proof of Theorem A.1, we obtain the desired result.

A.3 Proof of Proposition 3.1

Proof. *The proof is given for $p = 1$ but it hold also for any $p \geq 1$ after making some simple modifications. To obtain the first inequality we remarks that*

$$\mathbb{P}\left(\sup_{(x,y)\in\mathcal{X}\times\mathcal{X}}\left|\frac{k_\theta(x,y)}{k(x,y)} - 1\right| \geq \delta\right) \leq \sum_{(x,y)\in\mathbf{X}\times\mathbf{Y}} \mathbb{P}\left(\left|\frac{k_\theta(x,y)}{k(x,y)} - 1\right| \geq \delta\right) \tag{35}$$

Moreover as $\mathbf{E}_\rho\left(\frac{\varphi(x,u)\varphi(y,u)}{k(x,y)}\right) = 1$, the result follows by applying Hoeffding's inequality.

Algorithm 2 Accelerated Sinkhorn Algorithm.

Input: Initial estimate of the Lipschitz constant L_0, a, b, and $\mathbf{K}$
Init: $A_0 = \alpha_0 = 0$, $\eta^0 = \zeta^0 = \lambda^0 = 0$.
for $k \geq 0$ **do**
  $L_{k+1} = L_k/2$
  while *True* **do**
    Set $L_{k+1} = L_k/2$
    Set $a_{k+1} = \frac{1}{2L_{k+1}} + \sqrt{\frac{1}{4L_{k+1}^2} + a_k^2 \frac{L_k}{L_{k+1}}}$
    Set $\tau_k = \frac{1}{a_{k+1}L_{k+1}}$
    Set $\lambda^k = \tau_k \zeta^k + (1-\tau_k)\zeta^k$
    Choose $i_k = \underset{i \in \{1,2\}}{\operatorname{argmax}} \|\nabla_i \phi(\lambda^k)\|_2$
    if $i_k = 1$ **then**
      $\eta_1^{k+1} = \lambda_1^k + \log(a) - \log(e^{\lambda_1^k} \circ \mathbf{K} e^{\lambda_2^k})$
      $\eta_2^{k+1} = \lambda_2^{k+1}$
      else
        $\eta_1^{k+1} = \lambda_1^{k+1}$
        $\eta_2^{k+1} = \lambda_2^k + \log(b) - \log(e^{\lambda_2^k} \circ \mathbf{K}^T e^{\lambda_1^k})$
      end
    end
    Set $\zeta^{k+1} = \zeta^k - a_{k+1} \nabla F_\theta(\lambda^k)$
    if $\phi(\eta^k + 1) \leq \phi(\lambda^k) - \frac{\|\nabla F_\theta(\lambda^k)\|^2}{2L_{k+1}}$ **then**
      Set $z = \text{Diag}(e^{\lambda_1^k}) \circ \mathbf{K} \circ \text{Diag}(e^{\lambda_2^k})$
      Set $c = \langle e^{\lambda_1^k}, \mathbf{K} e^{\lambda_2^k} \rangle$
      Set $\hat{x}^{k+1} = \frac{a_{k+1} c^{-1} z + L_k a_k^2 \hat{x^k}}{L_{k+1} a_{k+1}^2}$
      Break
    end
    Set $L_{k+1} = 2L_{k+1}$
  end
end
Result: Transport Plan $\hat{x}^{k+1}$ and dual points $\eta^{k+1} = (\eta_1^{k+1}, \eta_2^{k+1})^T$

To show the second inequality, we follow the same strategy adopted in [47]. Let us denote $f(x,y) = \frac{k_\theta(x,y)}{k(x,y)} - 1$ *and* $\mathcal{M} := \mathcal{X} \times \mathcal{X}$. *First we remarks that* $|f(x,y)| \leq K+1$ *and* $\mathbf{E}_\rho(f) = 0$. *As* $\mathcal{M}$ *is a compact, we can find an* μ*-net that covers* $\mathcal{M}$ *with* $\mathcal{N}(\mathcal{M}, \mu) = \left(\frac{4R}{\mu}\right)^{2d}$ *where* $R = \sup_{(x,y)} \|(x,y)\|_2$ *balls of radius* δ. *Let us denote* $z_1, ..., z_{\mathcal{N}(\mathcal{M},\mu)} \in \mathcal{M}$ *the centers of these balls, and let* L_f *denote the Lipschitz constant of* f. *As* f *is differentiable We have therefore* $L_f = \sup_{z \in \mathcal{M}} \|\nabla f(z)\|_2$. *Moreover we have:*

$$\nabla f(z) = \frac{\nabla k_\theta(z)}{k(z)} - \frac{k_\theta(z)}{k(z)} \nabla k(z) \tag{36}$$

$$= \frac{1}{k(z)} \left[(\nabla k_\theta(z) - \nabla k(z)) + \nabla k(z) \left(1 - \frac{k_\theta(z)}{k(z)}\right) \right] \tag{37}$$

Therefore we have

$$\mathbf{E}(\|\nabla f(z)\|^2) \leq \frac{2}{k(z)^2} \left[\mathbf{E}(\|\nabla k_\theta(z) - \nabla k(z)\|^2) + \|\nabla k(z)\|^2 \mathbf{E} \left(1 - \frac{k_\theta(z)}{k(z)}\right)^2 \right] \tag{38}$$

But for any $z \in \mathcal{M}$ we have from Eq. (15) :

$$\mathbf{E}\left(1 - \frac{k_\theta(z)}{k(z)}\right)^2 = \int_{t \geq 0} \mathbb{P}\left(\left(1 - \frac{k_\theta(z)}{k(z)}\right)^2 \geq t\right) \tag{39}$$

$$\leq \frac{K^2}{r} \tag{40}$$

Moreover, we have:

$$\nabla k_\theta(z) = \frac{1}{r}\sum_{i=1}^{r} \nabla_x \varphi(x, u_i)\varphi(y, u_i) + \varphi(x, u_i)\nabla_y \varphi(y, u_i) \tag{41}$$

Therefore we have:

$$\begin{aligned}\|\nabla k_\theta(z)\|^2 = &\frac{1}{r^2}\sum_{i,j=1}^{r} \langle \nabla_x \varphi(x, u_i), \nabla_x \varphi(x, u_j)\rangle \varphi(y, u_i)\varphi(y, u_j) \\ &+ \frac{1}{r^2}\sum_{i,j=1}^{r} \nabla_y \varphi(y, u_i), \nabla_y \varphi(y, u_j)\rangle \varphi(x, u_i)\varphi(x, u_j) \\ &+ \frac{2}{r^2}\sum_{i,j=1}^{r} \nabla_x \varphi(x, u_i), \nabla_y \varphi(x, u_j)\rangle \varphi(y, u_i)\varphi(x, u_j)\end{aligned}$$

Moreover as:

$$|\varphi(y, u_i)\varphi(x, u_j)| \leq \frac{\varphi(y, u_i)^2 + \varphi(x, u_j)^2}{2} \tag{42}$$

$$\leq K \sup_{x \in \mathcal{X}} k(x, x) \tag{43}$$

And:

$$|\langle \nabla_x \varphi(x, u_i), \nabla_y \varphi(y, u_j)\rangle| \leq \|\nabla_x \varphi(x, u_i)\| \|\nabla_y \varphi(y, u_j)\| \tag{44}$$

$$\leq \frac{\|\nabla_x \varphi(x, u_i)\|^2 + \|\nabla_y \varphi(y, u_j)\|^2}{2} \tag{45}$$

And by denoting:

$$V := \sup_{x \in \mathcal{X}} \mathbf{E}_\rho\left(\|\nabla_x \varphi(x, u)\|^2\right) \tag{46}$$

Therefore we have:

$$\mathbf{E}\left(|\langle \nabla_x \varphi(x, u_i), \nabla_y \varphi(y, u_j)\rangle|\right) \leq V \tag{47}$$

We can now derive the following upper bound:

$$\mathbf{E}(\|\nabla k_\theta(z) - \nabla k(z)\|^2) = \mathbf{E}(\|\nabla k_\theta(z)\|^2) - \|\nabla k(z)\|^2 \leq 4VK \sup_{x \in \mathcal{X}} k(x, x) \tag{48}$$

Moreover by convexity of the ℓ_2 square norm, we also obtain that:

$$\|\nabla k(z)\|^2 \leq VK \sup_{x \in \mathcal{X}} k(x, x) \tag{49}$$

Therefore we have

$$\mathbf{E}(\|\nabla f(z)\|^2) \leq 2\kappa^{-2} VK \sup_{x \in \mathcal{X}} k(x, x)\left[4 + \frac{K^2}{r}\right] \tag{50}$$

Then by applying Markov inequality we obtain that:

$$\mathbb{P}\left(L_f \geq \frac{\delta}{2\mu}\right) \leq 2\kappa^{-2} VK \sup_{x \in \mathcal{X}} k(x, x)\left[4 + \frac{K^2}{r}\right]\left(\frac{2\mu}{\delta}\right)^2 \tag{51}$$

Moreover, the union bound followed by Hoeffding's inequality applied to the anchors in the μ-net gives

$$\mathbb{P}\left(\cup_{i=1}^{\mathcal{N}(\mathcal{M},\mu)}|f(z_i)| \geq \delta\right) \leq 2\mathcal{N}(\mathcal{M},\mu)\exp\left(-\frac{r\delta^2}{2K^2}\right) \tag{52}$$

Then by combining Eq. (51) and Eq.(52) we obtain that:

$$\mathbb{P}\left(\sup_{z\in\mathcal{M}}|f(z)| \geq \delta\right) \leq 2\left(\frac{4R}{\mu}\right)^{2d}\exp\left(-\frac{r\delta^2}{2K^2}\right) + 2\kappa^{-2}VK\sup_{x\in\mathcal{X}}k(x,x)\left[4+\frac{K^2}{r}\right]\left(\frac{2\mu}{\delta}\right)^2$$

Therefore by denoting

$$A_1 := 2\left(4R\right)^{2d}\exp\left(-\frac{r\delta^2}{2K^2}\right) \tag{53}$$

$$A_2 := 2\kappa^{-2}VK\sup_{x\in\mathcal{X}}k(x,x)\left[4+\frac{K^2}{r}\right]\left(\frac{2}{\delta}\right)^2 \tag{54}$$

and by choosing $\mu = \frac{A_1}{A_2}^{\frac{1}{2d+2}}$, we obtain that:

$$\mathbb{P}\left(\sup_{z\in\mathcal{M}}|f(z)| \geq \delta\right) \leq 2^9\left[\frac{\kappa^{-2}KV\sup_{x\in\mathcal{X}}k(x,x)\left[4+\frac{K^2}{r}\right]R^2}{\delta^2}\right]\exp\left(-\frac{r\delta^2}{2K^2(d+1)}\right)$$

Ratio Approximation. Let us assume here that $p = 1$ for simplicity. The uniform bound obtained on the ratio gives naturally a control of the form Eq.(14) with a prescribed number of random features r. This result allows to control the error when using the kernel matrix $\mathbf{K}_\theta$ instead of the true kernel matrix $\mathbf{K}$ in the Sinkhorn iterations. In the proposition above, we obtain such a result with a probability of at least $1 - 2n^2\exp\left(-\frac{r\delta^2}{2\psi^2}\right)$ where r is the number of random features and ψ is defined as

$$\psi := \sup_{u\in\mathcal{U}}\sup_{(x,y)\in\mathbf{X}\times\mathbf{Y}}\left|\frac{\varphi(x,u)\varphi(y,u)}{k(x,y)}\right|.$$

In comparison, in [47], the authors obtain a uniform bound on their difference and by denoting

$$\phi = \sup_{u\in\mathcal{U}}\sup_{(x,y)\in\mathbf{X}\times\mathbf{Y}}\left|\varphi(x,u)\varphi(y,u)\right|,$$

one obtains that with a probability of at least $1 - 2n^2\exp\left(-\frac{r\tau^2}{2\phi^2}\right)$ for all $(x,y) \in \mathbf{X}\times\mathbf{Y}$

$$k(x,y) - \tau \leq k_\theta(x,y) \leq k(x,y) + \tau \tag{55}$$

To be able to recover Eq.(14) from the above control, we need to take $\tau = \inf_{x,y\in\mathbf{X}\times\mathbf{Y}}k(x,y)\delta$ and by denoting $\phi' = \frac{\phi}{\inf_{x,y\in\mathbf{X}\times\mathbf{Y}}k(x,y)}$ we obtain that with a probability of at least $1 - 2n^2\exp\left(-\frac{r\delta^2}{2\phi'^2}\right)$ for all $(x,y) \in \mathbf{X}\times\mathbf{Y}$

$$(1-\delta)k(x,y) \leq k_\theta(x,y) \leq (1+\delta)k(x,y)$$

Therefore the number of random features needed to guarantee Eq.(14) from a control between the difference of the two kernels with at least a probability $1-\delta$ has to be larger than $\left(\frac{\phi'}{\psi}\right)^2$ times the number of random features needed from the control of Proposition 3.1 to guarantee Eq.(14) with at least the same probability $1-\delta$. But we always have that

$$\psi = \sup_{u\in\mathcal{U}}\sup_{(x,y)\in\mathbf{X}\times\mathbf{Y}}\left|\frac{\varphi(x,u)\varphi(y,u)}{k(x,y)}\right| \leq \frac{\sup_{u\in\mathcal{U}}\sup_{(x,y)\in\mathbf{X}\times\mathbf{Y}}|\varphi(x,u)\varphi(y,u)|}{\inf_{x,y\in\mathbf{X}\times\mathbf{Y}}k(x,y)} = \phi'$$

and in some cases the ratio $\left(\frac{\phi'}{\psi}\right)^2$ can be huge. Indeed, as we will see in the following, for the Gaussian kernel,

$$k(x,y) = \exp(-\varepsilon^{-1}\|x-y\|_2^2)$$

there exists φ and $\mathcal{U}$ such that for all x, y and $u \in \mathcal{U}$:

$$\varphi(x,u)\varphi(y,u) = k(x,y)h(u,x,y)$$

where for all $(x_0, y_0) \in \mathbf{X} \times \mathbf{Y}$,

$$\sup_{u\in\mathcal{U}} |h(u,x_0,y_0)| = \sup_{u\in\mathcal{U}} \sup_{(x,y)\in\mathbf{X}\times\mathbf{Y}} |h(u,x,y)|.$$

Therefore by denoting $M = \sup_{(x,y)\in\mathbf{X}\times\mathbf{Y}} \|x-y\|_2$ and $m = \inf_{(x,y)\in\mathbf{X}\times\mathbf{Y}} \|x-y\|_2$, we obtain that

$$\left(\frac{\phi'}{\psi}\right)^2 = \left(\frac{\sup_{x,y\in\mathbf{X}\times\mathbf{Y}} k(x,y)}{\inf_{x,y\in\mathbf{X}\times\mathbf{Y}} k(x,y)}\right)^2 = \exp\left(2\varepsilon^{-1}[M^2 - m^2]\right)$$

A.4 Proof of Lemma 1

Proof. *Let $\varepsilon > 0$ and $x, y \in \mathbb{R}^d$. We have that:*

$$\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|y-u\|_2^2\right) = \exp\left(-\varepsilon^{-1}\|x-y\|_2^2\right)\exp\left(-4\varepsilon^{-1}\left\|u-\left(\frac{x+y}{2}\right)\right\|_2^2\right) \tag{56}$$

And as the LHS is integrable we have:

$$\int_{u\in\mathbb{R}^d} \exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|y-u\|_2^2\right) du = \int_{u\in\mathbb{R}^d} e^{-\varepsilon^{-1}\|x-y\|_2^2}\exp\left(-4\varepsilon^{-1}\left\|u-\left(\frac{x+y}{2}\right)\right\|_2^2\right) du$$

Therefore we obtain that:

$$e^{-\varepsilon^{-1}\|x-y\|_2^2} = \left(\frac{4}{\pi\varepsilon}\right)^{d/2}\int_{u\in\mathbb{R}^d} \exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|y-u\|_2^2\right) du \tag{57}$$

Now we want to transform the above expression as the one stated in 9. To do so, let $q > 0$ and let us denote f_q the probability density function associated with the multivariate Gaussian distribution $\rho_q \sim \mathcal{N}\left(0, \frac{q}{4\varepsilon^{-1}}Id\right)$. We can rewrite the RHS of Eq. (57) as the following:

$$\begin{aligned}
&\left(\frac{4}{\pi\varepsilon}\right)^{d/2}\int_{u\in\mathbb{R}^d} \exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right) du \\
&= \left(\frac{4}{\pi\varepsilon}\right)^{d/2}\int_{u\in\mathbb{R}^d} \exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\frac{f_q(u)}{f_q(u)}d(u) \\
&= \left(\frac{4}{\pi\varepsilon}\right)^{d/2}\int_{u\in\mathbb{R}^d} \exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\left[\left(2\pi\frac{q}{4\varepsilon^{-1}}\right)^{d/2} e^{\frac{2\varepsilon^{-1}\|u\|_2^2}{q}}\right] d\rho_q(u) \\
&= (2q)^{d/2}\int_{u\in\mathbb{R}^d} \exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right) e^{\frac{2\varepsilon^{-1}\|u\|_2^2}{q}} d\rho_q(u)
\end{aligned}$$

Therefore for each $q > 0$, we obtain a feature map of k in $L^2(d\rho_q)$ which is defined as:

$$\varphi(x,u) = (2q)^{d/4}\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right) e^{\frac{\varepsilon^{-1}\|u\|_2^2}{q}}.$$

Moreover thanks to Eq. (56) we have also:

$$\begin{aligned}
\varphi(x,u)\varphi(y,u) &= (2q)^{d/2}\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|y-u\|_2^2\right) e^{\frac{2\varepsilon^{-1}\|u\|_2^2}{q}} \\
&= (2q)^{d/2}\exp\left(-\varepsilon^{-1}\|x-y\|_2^2\right)\exp\left(-4\varepsilon^{-1}\left\|u-\left(\frac{x+y}{2}\right)\right\|_2^2\right) e^{\frac{2\varepsilon^{-1}\|u\|_2^2}{q}}
\end{aligned}$$

Therefore we have:

$$\frac{\varphi(x,u)\varphi(y,u)}{k(x,y)} = (2q)^{d/2}\exp\left(-4\varepsilon^{-1}\left\|u-\left(\frac{x+y}{2}\right)\right\|_2^2\right)e^{\frac{2\varepsilon^{-1}\|u\|_2^2}{q}}$$
$$= (2q)^{d/2}\exp\left(-4\varepsilon^{-1}\left(1-\frac{1}{2q}\right)\left\|u-\left(1-\frac{1}{2q}\right)\left(\frac{x+y}{2}\right)\right\|_2^2\right)$$
$$\exp\left(\frac{4\varepsilon^{-1}}{2q-1}\left\|\left(\frac{x+y}{2}\right)\right\|_2^2\right)$$

Finally by choosing

$$q = \frac{\varepsilon^{-1}R^2}{2dW\left(\frac{\varepsilon^{-1}R^2}{d}\right)}$$

where W is the positive real branch of the Lambert function, we obtain that for any $x, y \in \mathcal{B}(0,R)$:

$$0 \leq \frac{\varphi(x,u)\varphi(y,u)}{k(x,y)} \leq 2\times(2q)^{d/2} \tag{58}$$

Moreover we have:

$$\varphi(x,u) = (2q)^{d/4}\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)e^{\frac{\varepsilon^{-1}\|u\|_2^2}{q}}$$

Therefore φ is differentiable with respect to x and we have:

$$\|\nabla_x\varphi\|_2^2 = 4\varepsilon^{-2}\|x-u\|_2^2\varphi(x,u)^2 \tag{59}$$
$$\leq 4\varepsilon^{-2}\psi\sup_{x\in\mathcal{X}}k(x,x)\|x-u\|_2^2 \tag{60}$$

where $\psi = 2\times(2q)^{d/2}$. But by definition of the kernel we have $\sup_{x\in\mathcal{B}(0,R)}k(x,x)=1$ and finally we have that for all $x\in\mathcal{B}(0,R)$:

$$\mathbf{E}(\|\nabla_x\varphi\|_2^2) \leq 4\varepsilon^{-2}\psi\left[R^2+\frac{q}{4\varepsilon^{-1}}\right] \tag{61}$$

A.5 Another example: Arc-cosine kernel

Lemma 3. *Let $d\geq 1$, $s\geq 0$, $\kappa>0$ and $k_{s,\kappa}$ be the perturbed arc-cosine kernel on $\mathbb{R}^d$ defined as for all $x,y\in\mathbb{R}^d$, $k_{s,\kappa}(x,y)=k_s(x,y)+\kappa$. Let also $\sigma>1$, $\rho=\mathcal{N}\left(0,\sigma^2 Id\right)$ and let us define for all $x,u\in\mathbb{R}^d$ the following map:*

$$\varphi(x,u) = \left(\sigma^{d/2}\sqrt{2}\max(0,u^Tx)^s\exp\left(-\frac{\|u\|^2}{4}\left[1-\frac{1}{\sigma^2}\right]\right),\sqrt{\kappa}\right)^T$$

Then for any $x,y\in\mathbb{R}^d$ we have:

$$k_{s,\kappa}(x,y) = \int_{u\in\mathbb{R}^d}\varphi(x,u)^T\varphi(y,u)d\rho(u)$$

Moreover we have for all $x,y\in\mathbb{R}^d$ $k_{s,\kappa}(x,y)\geq\kappa>0$ and for any compact $\mathcal{X}\subset\mathbb{R}^d$ we have:

$$\sup_{u\in\mathbb{R}^d}\sup_{(x,y)\in\mathcal{X}\times\mathcal{X}}\left|\frac{\varphi(x,u)\varphi(y,u)}{k(x,y)}\right| < +\infty \quad \textit{and} \quad \sup_{x\in\mathcal{X}}\mathbf{E}(\|\nabla_x\varphi\|_2^2) < +\infty$$

Proof. *Let $s\geq 0$. From [13], we have that:*

$$k_s(x,y) = \int_{\mathbb{R}^d}\Theta_s(u^Tx)\Theta_s(u^Ty)\frac{e^{-\frac{\|u\|_2^2}{2}}}{(2\pi)^{d/2}}du$$

where $\Theta_s(w) = \max(0, w)^s$. Let $\sigma > 1$ and f_σ the probability density function associated with the distribution $\mathcal{N}(0, \sigma^2 Id)$. Therefore we have that

$$k_s(x, y) = \int_{\mathbb{R}^d} \Theta_s(u^T x)\Theta_s(u^T y) \frac{e^{-\frac{\|u\|_2^2}{2}}}{(2\pi)^{d/2}} \frac{f_\sigma(u)}{f_\sigma(u)} du \tag{62}$$

$$= \sigma^d \int_{\mathbb{R}^d} \Theta_s(u^T x)\Theta_s(u^T y) \exp\left(-\frac{\|u\|^2}{2}\left[1 - \frac{1}{\sigma^2}\right]\right) d\rho(u) \tag{63}$$

where $\rho = \mathcal{N}(0, \sigma^2 Id)$. And by defining for all $x, u \in \mathbb{R}^d$ the following map:

$$\varphi(x, u) = \left(\sigma^{d/2}\sqrt{2}\max(0, u^T x)^s \exp\left(-\frac{\|u\|^2}{4}\left[1 - \frac{1}{\sigma^2}\right]\right), \sqrt{\kappa}\right)^T$$

we obtain that any $x, y \in \mathbb{R}^d$:

$$\begin{aligned}\int_{u\in\mathbb{R}^d} \varphi(x, u)^T \varphi(y, u) d\rho(u) &= \kappa + \sigma^d \int_{\mathbb{R}^d} \Theta_s(u^T x)\Theta_s(u^T y) \exp\left(-\frac{\|u\|^2}{2}\left[1 - \frac{1}{\sigma^2}\right]\right) d\rho(u) \\ &= \kappa + k_s(x, y) \\ &= k_{s,\kappa}(x, y)\end{aligned}$$

Moreover from the definition of the feature map φ, it is clear that $k_{s,\kappa} \geq \kappa > 0$,

$$\sup_{u\in\mathbb{R}^d} \sup_{(x,y)\in\mathcal{X}\times\mathcal{X}} \left|\frac{\varphi(x, u)\varphi(y, u)}{k(x, y)}\right| < +\infty \quad \textit{and} \quad \sup_{x\in\mathcal{X}} \mathbf{E}(\|\nabla_x \varphi\|_2^2) < +\infty.$$

B Constructive Method: Differentiability

B.1 Proof of Proposition 3.2

Proof. *Let us first introduce the following Lemma:*

Lemma 4. *Let (α^*, β^*) solution of (5), then we have*

$$\max_i \alpha_i^* - \min_i \alpha_i^* \leq \varepsilon R(\mathbf{K})$$
$$\max_j \beta_j^* - \min_j \beta_j^* \leq \varepsilon R(\mathbf{K})$$

where $R(\mathbf{K}) = -\log\left(\iota \frac{\min_{i,j} \mathbf{K}_{i,j}}{\max_{i,j} \mathbf{K}_{i,j}}\right)$ with $\iota := \min_{i,j}(a_i, b_j)$.

Proof B.1. *Indeed at optimality, the primal-dual relationship between optimal variables gives us that for all $i = 1, ..., n$:*

$$e^{\alpha_i^*/\varepsilon} \langle \mathbf{K}_{i,:}, e^{\beta^*/\varepsilon} \rangle = a_i \leq 1$$

Moreover we have that

$$\min_{i,j} \mathbf{K}_{i,j} \langle \mathbf{1}, e^{\beta^*/\varepsilon} \rangle \leq \langle \mathbf{K}_{i,:}, e^{\beta^*/\varepsilon} \rangle \leq \max_{i,j} \mathbf{K}_{i,j} \langle \mathbf{1}, e^{\beta^*/\varepsilon} \rangle$$

Therefore we obtain that

$$\max_i \alpha_i^* \leq \varepsilon \log\left(\frac{1}{\min_{i,j} \mathbf{K}_{i,j} \langle \mathbf{1}, e^{\beta^*/\varepsilon} \rangle}\right)$$

and

$$\min_i \alpha_i^* \geq \varepsilon \log\left(\frac{\iota}{\langle \mathbf{1}, e^{\beta^*/\varepsilon} \rangle \max_{i,j} \mathbf{K}_{i,j}}\right)$$

Therefore we obtain that

$$\max_i \alpha_i^* - \min_i \alpha_i^* \geq -\varepsilon \log\left(\iota \frac{\min_{i,j} \mathbf{K}_{i,j}}{\max_{i,j} \mathbf{K}_{i,j}}\right)$$

An analogue proof for β^* leads to similar result.

Let us now define for any $\mathbf{K} \in (\mathbb{R}_+^*)^{n\times m}$ with positive entries the following objective function:

$$F(\mathbf{K}, \alpha, \beta) := \langle \alpha, a\rangle + \langle \beta, a\rangle - \varepsilon (e^{\alpha/\varepsilon})^T \mathbf{K} e^{\beta/\varepsilon}.$$

Let us first show that

$$G(\mathbf{K}) := \sup_{(\alpha,\beta)\in\mathbb{R}^n\times\mathbb{R}^m} F(\mathbf{K}, \alpha, \beta) \tag{64}$$

is differentiable on $(\mathbb{R}_+^*)^{n\times m}$. For that purpose let us introduce for any $\gamma_1, \gamma_2 > 0$, the following objective function:

$$G_{\gamma_1,\gamma_2}(\mathbf{K}) := \sup_{\substack{(\alpha,\beta)\in B_\infty^n(0,\gamma_1)\times B_\infty^m(0,\gamma_2)\\ \alpha^T e_1 = 0}} F(\mathbf{K}, \alpha, \beta)$$

where $B_\infty^n(0,\gamma)$ denote the ball of radius γ according to the infinite norm and $e_1 = (1, 0, ...0)^T \in \mathbb{R}^n$. In the following we denote by

$$S_{\gamma_1,\gamma_2} := \left\{(\alpha,\beta) \in B_\infty^n(0,\gamma_1)\times B_\infty^m(0,\gamma_2) \quad : \quad \alpha^T e_1 = 0\right\}.$$

Let us now introduce the following Lemma:

Lemma 5. *Let $\varepsilon > 0$, $(a, b) \in \Delta_n \times \Delta_m$, $K \in (\mathbb{R}_+^*)^{n\times m}$ with positive entries. Then*

$$\max_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^m} a^T\alpha + b^T\beta - \varepsilon (e^{\alpha/\varepsilon})^T \mathbf{K} e^{\beta/\varepsilon}$$

admits a unique solution (α^*, β^*) such that $\alpha^T e_1 = 0$, $\|\alpha^*\|_\infty \leq \varepsilon R_1(\mathbf{K})$, and , $\|\beta^*\|_\infty \leq \varepsilon[R_1(\mathbf{K}) + R_2(\mathbf{K})]$ where $R_1(\mathbf{K}) = -\log\left(\iota \frac{\min_{i,j} K_{i,j}}{\max_{i,j} K_{i,j}}\right)$, $R_2(\mathbf{K}) = \log\left(n \frac{\max_{i,j} K_{i,j}}{\iota}\right)$ and $\iota := \min_{i,j}(a_i, b_j)$.

Proof B.2. *In fact the existence and uncity up to a scalar transformation is a well known result. See for example [16]. Therefore there is a unique solution (α^0, β^0) such that $(\alpha^0)^T e_1 = 0$. Moreover thanks to Lemma 4, we have that for any (α^*, β^*) optimal solution that*

$$\max_i \alpha_i^* - \min_i \alpha_i^* \leq \varepsilon R(\mathbf{K}) \tag{65}$$

$$\max_j \beta_j^* - \min_j \beta_j^* \leq \varepsilon R(\mathbf{K}) \tag{66}$$

Therefore we have $\|\alpha^0\|_\infty \leq \max_i \alpha_i^0 - \min_i \alpha_i^0 \leq \varepsilon R(\mathbf{K})$. Moreover, the first order optimality conditions for the dual variables (α, β) implies that for all $j = 1, .., m$

$$\beta_j^0 = -\varepsilon \log\left(\sum_{i=1}^n \frac{\mathbf{K}_{i,j}}{b_j} \exp\left(\frac{\alpha_i^0}{\varepsilon}\right)\right)$$

Therefore we have that:

$$\|\beta^0\|_\infty \leq \|\alpha^0\|_\infty + \varepsilon \log\left(n \frac{\max_{i,j} \mathbf{K}_{i,j}}{\iota}\right)$$

and the result follows.

Let $\mathbf{K}_0 \in (\mathbb{R}_+^*)^{n\times m}$, and let us denote $M_0 = \max_{i,j} \mathbf{K}_0[i,j]$, $m_0 = \min_{i,j} \mathbf{K}_0[i,j]$ and

$$A_\omega := \left\{\mathbf{K} \in (\mathbb{R}_+^*)^{n\times m} \quad \textit{such that} \quad \|\mathbf{K} - \mathbf{K}_0\|_\infty < \omega\right\}$$

By considering $\omega_0 = \frac{m_0}{2}$, we obtain that for any $K \in A_{\omega_0}$,

$$R_1(\mathbf{K}) \leq \log\left(\frac{1}{\iota}\frac{2M_0 + m_0}{m_0}\right)$$

$$R_2(\mathbf{K}) \leq \log\left(n \frac{2M_0 + m_0}{2\iota}\right)$$

Therefore by denoting

$$\gamma_1^0 = \varepsilon \log\left(\frac{1}{\iota}\frac{2M_0+m_0}{m_0}\right)$$
$$\gamma_2^0 = \varepsilon\left[\log\left(\frac{1}{\iota}\frac{2M_0+m_0}{m_0}\right) + \log\left(n\frac{2M_0+m_0}{2\iota}\right)\right]$$

Therefore, from Lemma 5, we have that for all $\mathbf{K} \in A_{\omega_0}$ *there exists a unique optimal solution* $(\alpha,\beta) \in B_\infty^n(0,\gamma_1^0) \times B_\infty^m(0,\gamma_2^0)$ *satisfying* $\alpha^T e_1 = 0$. *Therefore we have first that for all* $K \in A_{\omega_0}$

$$G_{\gamma_1^0,\gamma_2^0}(\mathbf{K}) = G(\mathbf{K}) \tag{67}$$

and moreover for all $\mathbf{K} \in A_{\omega_0}$, *the following set*

$$Z_{\mathbf{K}} := \left\{ (\alpha,\beta) \in S_{\gamma_1^0,\gamma_2^0} \quad \textit{such that} \quad F(\mathbf{K},\alpha,\beta) = \sup_{(\alpha,\beta)\in S_{\gamma_1^0,\gamma_2^0}} F(\mathbf{K},\alpha,\beta) \right\}$$

is a singleton. Let us now consider the restriction of F *on* $A_{\omega_0} \times S_{\gamma_1^0,\gamma_2^0}$ *denoted* F_0. *It is clear from their definition that* A_{ω_0} *is an open convex set, and* $S_{\gamma_1^0,\gamma_2^0}$ *is compact. Moreover* F_0 *is clearly continuous, and for any* $(\alpha,\beta) \in S_{\gamma_1^0,\gamma_2^0}$, $F_0(\cdot,\alpha,\beta)$ *is convex. Moreover for any* $\mathbf{K} \in A_{\omega_0}$ *the set* $Z_{\mathbf{K}}$ *is a singleton, therefore from Danskin theorem [8], we deduce that* $G_{\gamma_1^0,\gamma_2^0}$ *is convex and differentiable on* A_{ω_0} *and we have for all* $K \in A_{\omega_0}$

$$\nabla G_{\gamma_1^0,\gamma_2^0}(\mathbf{K}) = -\varepsilon e^{\alpha^*/\varepsilon}(e^{\beta^*/\varepsilon})^T \tag{68}$$

where $(\alpha^*,\beta^*) \in Z_K$. *Note that any solutions of Eq.(64) can be used to evaluated* $\nabla G_{\gamma_1^0,\gamma_2^0}(\mathbf{K})$. *Moreover thanks to Eq.(67), we deduce also that* G *is also differentiable on* A_{ω_0}. *Finally the reasoning hold for any* $\mathbf{K}_0 \in (\mathbb{R}_+^*)^{n\times m}$, *therefore* G *is differentiable and we have:*

$$\nabla G(\mathbf{K}) = -\varepsilon e^{\alpha^*/\varepsilon}(e^{\beta^*/\varepsilon})^T \tag{69}$$

C Illustrations and Experiments

In Figure 5, we show the time-accuracy tradeoff in the high dimensional setting. Here the samples are taken from the higgs dataset[1] [7] where the sample lives in $\mathbb{R}^{28}$. This dataset contains two class of signals: a signal process which produces Higgs bosons and a background process which does not. We take randomly 5000 samples from each of these two distributions.

In Figure 6, we consider a discretization of the positive sphere using $50^2 = 2,500$ points and generate three simple histograms of blurred pixels located in the three corners of the simplex.

[1] https://archive.ics.uci.edu/ml/datasets/HIGGS

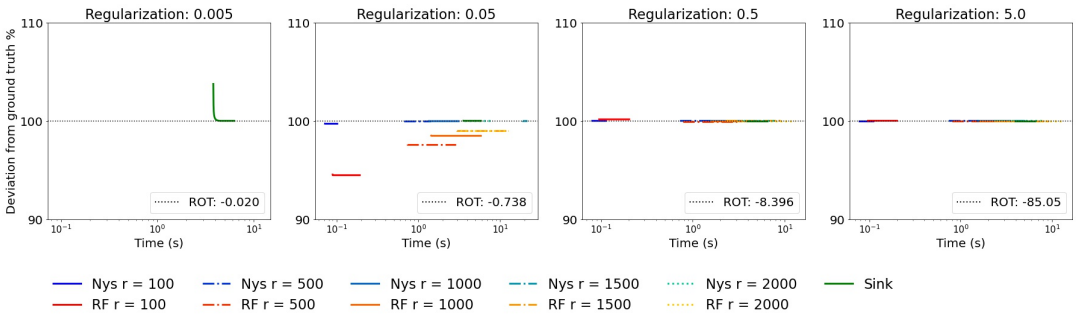

Figure 5: In this experiment, we take randomly 10000 samples from the two distributions of the higgs dataset and we plot the deviation from ground truth for different regularizations. We compare the results obtained for our proposed method (**RF**) with the one proposed in [3] (**Nys**) and with the Sinkhorn algorithm (**Sin**) proposed in [16]. The cost function considered here is the square Euclidean metric and the feature map used is that presented in Lemma 1. The number of random features (or rank) chosen varies from 100 to 2000. We repeat for each problem 10 times the experiment. Note that curves in the plot start at different points corresponding to the time required for initialization. *Right, middle right*: when the regularization is sufficiently large both **Nys** and **RF** methods obtain very high accuracy with order of magnitude faster than **Sin**. *Middle left*: both methods manage to obtain high accuracy of the ROT with order of magnitude faster than **Sin**. Note that **Nys** performs better in this setting than our proposed method. *Left*: both methods fail to obtain a good approximation of the ROT.

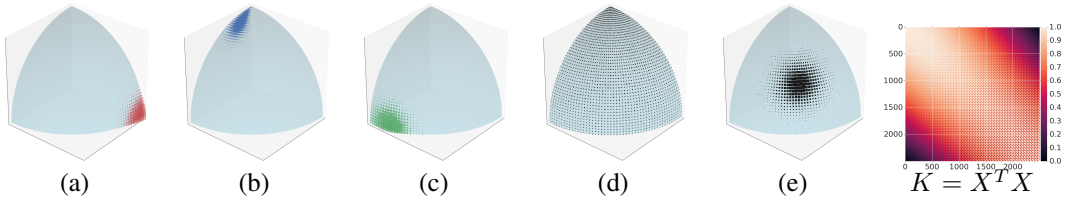

Figure 6: Using a discretization of the positive sphere with $50^2 = 2,500$ points we generate three simple histograms (a,b,c) located in the three corners of the simplex. (d) Wasserstein barycenter with a cost $c(x, y) = -\log(x^T y)$ using the method by [9]. (e) Soft-max with temperature 1000 of that barycenter (strongly increasing the relative influence of peaks) reveals that mass is concentrated in areas that would make sense from the more usual $c(x, y) = \arccos x^T y$ distance on the sphere. The kernel corresponding to that cost, here the simple outer product of a matrix X of dimsension 3×2500.