Sample Complexity of Sinkhorn Divergences

Aude Genevay
DMA,
ENS Paris

Lénaic Chizat
Inria

Francis Bach
Inria and
DI, ENS Paris

Marco Cuturi
Google and
CREST ENSAE

Gabriel Peyré
CNRS and DMA,
ENS Paris

Abstract

Optimal transport (OT) distances and maximum mean discrepancies (MMD) are now routinely used in machine learning to compare probability measures. Our focus in this paper is on *Sinkhorn divergences* (SDs), a regularized variant of OT distances that can interpolate, depending on the regularization strength ε, between OT ($\varepsilon = 0$) and MMD ($\varepsilon = \infty$). Although the tradeoff induced by that regularization is by now well understood computationally (OT, SDs and MMD require respectively $O(n^3 \log n)$, $O(n^2)$ and n^2 operations to compare two samples of size n), much less is known in terms of the *sample complexity* of SDs, namely bounding the gap between the evaluation of SDs on two densities *vs.* samples from these densities. That complexity for OT and MMD stand at two extremes: $O(1/n^{1/d})$ for OT in dimension d and $O(1/\sqrt{n})$ for MMD. that for SDs has only been studied empirically. In this paper, we *(i)* derive a bound on the approximation error made with SDs when approximating OT as a function of the regularizer ε, *(ii)* prove that the optimizers of regularized OT are bounded in a Sobolev (RKHS) ball independent of the two measures and *(iii)* reformulate SDs as a maximization problem in a RKHS to obtain a sample complexity in $1/\sqrt{n}$ (as in MMD), with a constant that depends however on ε, making the bridge between OT and MMD complete.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

1 Introduction

Optimal Transport (OT) has emerged in recent years as a powerful tool to compare probability distributions. Indeed, Wasserstein distances endow the space of probability measures with a rich Riemannian structure (Ambrosio et al., 2006), one that is able to capture meaningful geometric features between measures even when their supports do not overlap. OT has been, however, long neglected in data sciences for two main reasons, which could be loosely described as *computational*—computing OT is costly since it usually requires solving a network flow problem—and *statistical*—OT suffers from the curse-of-dimensionality, since has recalled later in this paper, the Wasserstein distance computed between two samples converges only very slowly to its population counterpart.

Recent years have witnessed significant advances on those computational aspects that hindered the application of OT. A recent wave of works have exploited entropic regularization, both to compare discrete measures with finite support (Cuturi, 2013) or measures that can be sampled from (Genevay et al., 2016). Among the many learning tasks performed with this regularization, one may cite domain adaptation (Courty et al., 2014), text retrieval (Huang et al., 2016) or multilabel classification (Frogner et al., 2015). The ability of OT to compare probability distributions with disjoint supports (as opposed to the Kullback-Leibler divergence) has also made it popular as a loss function to learn generative models (Genevay et al., 2018; Salimans et al., 2018; Beaumont et al., 2002) as an alternative to the approximation considered in (Arjovsky et al., 2017).

At the other end of the spectrum, the maximum mean discrepancy (MMD) (Gretton et al., 2006) is an integral probability metric (Sriperumbudur et al., 2012) on a reproducing kernel Hilbert space (RKHS) of test functions. The MMD is easy to compute, and has also been used in a very wide variety of applications, notably to estimate of generative models (Li et al., 2015; Dziugaite et al., 2015; Li et al., 2017).

OT and MMD differ, however, on a fundamental aspect: their sample complexity. The definition of sample complexity that we choose here is the convergence rate of a given metric between a measure and its empirical counterpart, as a function of the number of samples. This notion is crucial in machine learning, as bad sample complexity implies overfitting and high gradient variance when using these divergences for parameter estimation. In that context, it is well known that the sample complexity of MMD is independent of the dimension, scaling as $\frac{1}{\sqrt{n}}$ (Gretton et al., 2006) where n is the number of samples. In contrast, it is well known that standard OT suffers from the curse of dimensionality (Dudley, 1969): Its sample complexity is exponential in the dimension of the ambient space. Although it was recently shown that these results can be refined by considering the implicit dimension of data (Weed and Bach, 2017) or the densitie properties' Weed and Berthet (2019), the sample complexity of OT appears now to be the major bottleneck for the use of OT in high-dimensional machine learning problems.

A remedy to this problem may lie, again, in regularization. Divergences defined through regularized OT, known as Sinkhorn divergences, seem to be indeed less prone to over-fitting. Indeed, a certain amount of regularization seems to improve performance in simple learning tasks (Cuturi, 2013). Additionally, recent papers (Ramdas et al., 2017; Genevay et al., 2018) have pointed out the fact that Sinkhorn divergences are in fact interpolating between OT (when regularization goes to zero) and MMD (when regularization goes to infinity). However, aside from a recent central limit theorem in the case of measures supported on discrete spaces (Bigot et al., 2017), the convergence of empirical Sinkhorn divergences, and more generally their sample complexity, remains an open question.

Contributions. This paper provides three main contributions, which all exhibit theoretical properties of Sinkhorn divergences. Our first result is a bound on the speed of convergence of regularized OT to standard OT as a function of the regularization parameter, in the case of continuous measures. The second theorem proves that the optimizers of the regularized optimal transport problem lie in a Sobolev ball which is independent of the measures. This allows us to rewrite the Sinkhorn divergence as an expectation maximization problem in a RKHS ball and thus justify the use of kernel-SGD for regularized OT as advocated in (Genevay et al., 2016). As a consequence of this reformulation, we provide as our third contribution a sample complexity result. We focus on how the sample size and the regularization parameter affect the convergence of the empirical Sinkhorn divergence (i.e., computed from samples of two continuous measures) to the continuous Sinkhorn divergence. We show that the Sinkhorn divergence benefits from the same sample complexity as MMD, scaling in $\frac{1}{\sqrt{n}}$ but with a constant that depends on the inverse of the regularization parameter. Thus sample complexity worsens when getting closer to standard OT, and there is therefore a tradeoff between a good approximation of OT (small regularization parameter) and fast convergence in terms of sample size (larger regularization parameter). We conclude this paper with a few numerical experiments to asses the dependence of the sample complexity on ε and d in very simple cases.

Notations. We consider $\mathcal{X}$ and $\mathcal{Y}$ two bounded subsets of $\mathbb{R}^d$ and we denote by $|\mathcal{X}|$ and $|\mathcal{Y}|$ their respective diameter $\sup\{\|x - x'\| | x, x' \in \mathcal{X}(resp. \mathcal{Y})\}$. The space of positive Radon measures of mass 1 on $\mathcal{X}$ is denoted $\mathcal{M}_+^1(\mathcal{X})$ and we use upper cases X, Y to denote random variables in these spaces. We use the notation $\varphi = O(1 + x^k)$ to say that $\varphi \in \mathbb{R}$ is bounded by a polynomial of order k in x with positive coefficients.

2 Reminders on Sinkhorn Divergences

We consider two probability measures $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and β on $\mathcal{M}_+^1(\mathcal{Y})$. The Kantorovich formulation (1942) of optimal transport between α and β is defined by

$$W(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y), \qquad (\mathcal{P})$$

where the feasible set is composed of probability distributions over the product space $\mathcal{X} \times \mathcal{Y}$ with fixed marginals α, β:

$$\Pi(\alpha, \beta) \overset{\text{def.}}{=} \left\{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) \; ; \; P_{1\sharp}\pi = \alpha, P_{2\sharp}\pi = \beta\right\},$$

where $P_{1\sharp}\pi$ (resp. $P_{2\sharp}\pi$) is the marginal distribution of π for the first (resp. second) variable, using the projection maps $P_1(x, y) = x; P_2(x, y) = y$ along with the push-forward operator $\sharp$.

The cost function c represents the cost to move a unit of mass from x to y. Through this paper, we will assume this function to be $\mathcal{C}^\infty$ (more specifically, we need it to be $\mathcal{C}^{\frac{d}{2}+1}$). When $\mathcal{X} = \mathcal{Y}$ is endowed with a distance $d_\mathcal{X}$, choosing $c(x, y) = d_\mathcal{X}(x, y)^p$ where $p \geqslant 1$ yields the p-Wasserstein distance between probability measures.

We introduce regularized optimal transport, which consists in adding an entropic regularization to the optimal transport problem, as proposed in (Cuturi, 2013). Here we use the relative entropy of the transport plan with respect to the product measure $\alpha \otimes \beta$ following (Genevay et al., 2016):

$$W_\varepsilon(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon H(\pi \mid \alpha\otimes\beta), \qquad (\mathcal{P}_\varepsilon)$$

where

$$H(\pi \mid \alpha\otimes\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right)\mathrm{d}\pi(x,y). \tag{1}$$

Choosing the relative entropy as a regularizer allows to express the dual formulation of regularized OT as the maximization of an expectation problem, as shown in (Genevay et al., 2016)

$$\begin{aligned} W_\varepsilon(\alpha,\beta) &= \max_{u\in\mathcal{C}(\mathcal{X}),v\in\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\beta(y) \\ &\quad - \varepsilon\int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon \\ &= \max_{u\in\mathcal{C}(X),v\in\mathcal{C}(Y)} \mathbb{E}_{\alpha\otimes\beta}\left[f_\varepsilon^{XY}(u,v)\right] + \varepsilon \end{aligned}$$

where $f_\varepsilon^{xy}(u,v) = u(x)+v(y)-\varepsilon e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}$. This reformulation as the maximum of an expectation will prove crucial to obtain sample complexity results. The existence of optimal dual potentials (u,v) is proved in the appendix. They are unique $\alpha-$ and $\beta-$a.e. up to an additive constant.

To correct for the fact that $W_\varepsilon(\alpha,\alpha)\neq 0$, (Genevay et al., 2018) propose Sinkhorn divergences, a natural normalization of that quantity defined as

$$\bar{W}_\varepsilon(\alpha,\beta) = W_\varepsilon(\alpha,\beta) - \frac{1}{2}(W_\varepsilon(\alpha,\alpha) + W_\varepsilon(\beta,\beta)). \tag{2}$$

This normalization ensures that $\bar{W}_\varepsilon(\alpha,\alpha)=0$, but also has a noticeable asymptotic behavior as mentioned in (Genevay et al., 2018). Indeed, when $\varepsilon\to 0$ one recovers the original (unregularized) OT problem, while choosing $\varepsilon\to+\infty$ yields the maximum mean discrepancy associated to the kernel $k=-c/2$, where MMD is defined by:

$$MMD_k(\alpha,\beta) = \mathbb{E}_{\alpha\otimes\alpha}[k(X,X')] + \mathbb{E}_{\beta\otimes\beta}[k(Y,Y')] - 2\mathbb{E}_{\alpha\otimes\beta}[k(X,Y)].$$

In the context of this paper, we study in detail the sample complexity of $W_\varepsilon(\alpha,\beta)$, knowing that these results can be extended to $\bar{W}_\varepsilon(\alpha,\beta)$.

3 Approximating Optimal Transport with Sinkhorn Divergences

In the present section, we are interested in bounding the error made when approximating $W(\alpha,\beta)$ with $W_\varepsilon(\alpha,\beta)$.

Theorem 1. *Let α and β be probability measures on $\mathcal{X}$ and $\mathcal{Y}$ subsets of $\mathbb{R}^d$ such that $|\mathcal{X}|=|\mathcal{Y}|\leqslant D$ and assume that c is L-Lipschitz w.r.t. x and y. It holds*

$$\begin{aligned} 0 \leqslant W_\varepsilon(\alpha,\beta) - W(\alpha,\beta) &\leqslant 2\varepsilon d\log\left(\tfrac{e^2\cdot L\cdot D}{\sqrt{d}\cdot\varepsilon}\right) & (3)\\ &\sim_{\varepsilon\to 0} 2\varepsilon d\log(1/\varepsilon). & (4) \end{aligned}$$

Proof. For a probability measure π on $\mathcal{X}\times\mathcal{Y}$, we denote by $C(\pi)=\int c\,\mathrm{d}\pi$ the associated transport cost and by $H(\pi)$ its relative entropy with respect to the product measure $\alpha\otimes\beta$ as defined in (1). Choosing π_0 a minimizer of $\min_{\pi\in\Pi(\alpha,\beta)} C(\pi)$, we will build our upper bounds using a family of transport plans with finite entropy that approximate π_0. The simplest approach consists in considering block approximation. In contrast to the work of Carlier et al. (2017), who also considered this technique, our focus here is on quantitative bounds.

Definition 1 (Block approximation)**.** *For a resolution $\Delta>0$, we consider the block partition of $\mathbb{R}^d$ in hypercubes of side Δ defined as*

$$\{Q_k^\Delta = [k_1\cdot\Delta,(k_1+1)\cdot\Delta[\times\dots[k_d\cdot\Delta,(k_d+1)\cdot\Delta[\,; \quad k=(k_1,\dots,k_d)\in\mathbb{Z}^d\}.$$

To simplify notations, we introduce $Q_{ij}^\Delta \overset{\text{def.}}{=} Q_i^\Delta\times Q_j^\Delta$, $\boldsymbol{\alpha}_i^\Delta \overset{\text{def.}}{=} \alpha(Q_i^\Delta)$, $\boldsymbol{\beta}_j^\Delta \overset{\text{def.}}{=} \beta(Q_j^\Delta)$. The block approximation of π_0 of resolution Δ is the measure $\pi^\Delta\in\Pi(\alpha,\beta)$ characterized by

$$\pi^\Delta|_{Q_{ij}^\Delta} = \frac{\pi_0(Q_{ij}^\Delta)}{\boldsymbol{\alpha}_i^\Delta\cdot\boldsymbol{\beta}_j^\Delta}(\alpha|_{Q_i^\Delta}\otimes\beta|_{Q_j^\Delta})$$

for all $(i,j)\in(\mathbb{Z}^d)^2$, with the convention $0/0=0$.

π^Δ is nonnegative by construction. Observe also that for any Borel set $B\subset\mathbb{R}^d$, one has

$$\begin{aligned} \pi^\Delta(B\times\mathbb{R}^d) &= \sum_{(i,j)\in(\mathbb{Z}^d)^2} \frac{\pi_0(Q_{ij}^\Delta)}{\boldsymbol{\alpha}_i^\Delta\cdot\boldsymbol{\beta}_j^\Delta}\cdot\alpha(B\cap Q_i^\Delta)\cdot\boldsymbol{\beta}_j^\Delta \\ &= \sum_{i\in\mathbb{Z}^d}\alpha(B\cap Q_i^\Delta) = \alpha(B), \end{aligned}$$

which proves, using the symmetric result in β, that π^Δ belongs to $\Pi(\alpha,\beta)$. As a consequence, for any $\varepsilon>0$ one has $W_\varepsilon(\alpha,\beta)\leqslant C(\pi^\Delta)+\varepsilon H(\pi^\Delta)$. Recalling also that the relative entropy H is nonnegative over the set of probability measures, we have the bound

$$0\leqslant W_\varepsilon(\alpha,\beta)-W(\alpha,\beta)\leqslant (C(\pi^\Delta)-C(\pi_0))+\varepsilon H(\pi^\Delta).$$

We can now bound the terms in the right-hand side, and choose a value for Δ that minimizes these bounds.

The bound on $C(\pi^\Delta) - C(\pi_0)$ relies on the Lipschitz regularity of the cost function. Using the fact that $\pi^\Delta(Q_{ij}^\Delta) = \pi_0(Q_{ij}^\Delta)$ for all i, j, it holds

$$C(\pi^\Delta) - C(\pi_0) = \sum_{(i,j)\in(\mathbb{Z}^d)^2} \pi_0(Q_{ij}^\Delta)\Big(\sup_{x,y\in Q_{ij}^\Delta} c(x,y) - \inf_{x,y\in Q_{ij}^\Delta} c(x,y)\Big) \leqslant 2L\Delta\sqrt{d},$$

where L is the Lipschitz constant of the cost (separately in x and y) and $\Delta\sqrt{d}$ is the diameter of each set Q_i^Δ.

As for the bound on $H(\pi^\Delta)$, using the fact that $\pi_0(Q_{ij}^\Delta) \leqslant 1$ we get

$$\begin{aligned} H(\pi^\Delta) &= \sum_{(i,j)\in(\mathbb{Z}^d)^2} \log\left(\frac{\pi_0(Q_{ij}^\Delta)}{\boldsymbol{\alpha}_i^\Delta \cdot \boldsymbol{\beta}_j^\Delta}\right)\pi_0(Q_{ij}^\Delta) \\ &\leqslant \sum_{(i,j)\in(\mathbb{Z}^d)^2} \Big(\log(1/\boldsymbol{\alpha}_i^\Delta) + \log(1/\boldsymbol{\beta}_j^\Delta)\Big)\pi_0(Q_{ij}^\Delta) \\ &= -H^\Delta(\alpha) - H^\Delta(\beta), \end{aligned}$$

where we have defined $H^\Delta(\alpha) = \sum_{i\in\mathbb{Z}^d} \boldsymbol{\alpha}_i^\Delta \log(\boldsymbol{\alpha}_i^\Delta)$ and similarly for β. Note that in case α is a discrete measure with finite support, $H^\Delta(\alpha)$ is equal to (minus) the discrete entropy of α as long as Δ is smaller than the minimum separation between atoms of α. However, if α is not discrete then $H^\Delta(\alpha)$ blows up to $-\infty$ as Δ goes to 0 and we need to control how fast it does so. Considering α^Δ the block approximation of α with constant density $\boldsymbol{\alpha}_i^\Delta/\Delta^d$ on each block Q_i^Δ and (minus) its differential entropy $H_{\mathcal{L}^d}(\alpha^\Delta) = \int_{\mathbb{R}^d} \alpha^\Delta(x)\log\alpha^\Delta(x)dx$, it holds $H^\Delta(\alpha) = H_{\mathcal{L}^d}(\alpha^\Delta) - d\cdot\log(1/\Delta)$. Moreover, using the convexity of $H_{\mathcal{L}^d}$, this can be compared with the differential entropy of the uniform probability on a hypercube containing $\mathcal{X}$ of size $2D$. Thus it holds $H_{\mathcal{L}^d}(\alpha^\Delta) \geqslant -d\log(2D)$ and thus $H^\Delta(\alpha) \geqslant -d\cdot\log(2D/\Delta)$.

Summing up, we have for all $\Delta > 0$

$$W_\varepsilon(\alpha,\beta) - W(\alpha,\beta) \leqslant 2L\Delta\sqrt{d} + 2\varepsilon d\cdot\log(2D/\Delta).$$

The above bound is convex in Δ, minimized with $\Delta = 2\sqrt{d}\cdot\varepsilon/L$. This yields

$$W_\varepsilon(\alpha,\beta) - W(\alpha,\beta) \leqslant 4\varepsilon d + 2\varepsilon d\log\left(\frac{L\cdot D}{\sqrt{d}\cdot\varepsilon}\right). \quad \square$$

4 Properties of Sinkhorn Potentials

We prove in this section that Sinkhorn potentials are bounded in the Sobolev space $\mathbf{H}^s(\mathbb{R}^d)$ regardless of the marginals α and β. For $s > \frac{d}{2}$, $\mathbf{H}^s(\mathbb{R}^d)$ is a reproducing kernel Hilbert space (RKHS): This property will be crucial to establish sample complexity results later on, using standard tools from RKHS theory.

Definition 2. *The Sobolev space $\mathbf{H}^s(\mathcal{X})$, for $s \in \mathbb{N}^*$, is the space of functions $\varphi : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ such that for every multi-index k with $|k| \leqslant s$ the mixed partial derivative $\varphi^{(k)}$ exists and belongs to $L^2(\mathcal{X})$. It is endowed with the following inner-product*

$$\langle\varphi,\psi\rangle_{\mathbf{H}^s(\mathcal{X})} = \sum_{|k|\leqslant s}\int_{\mathcal{X}} \varphi^{(k)}(x)\psi^{(k)}(x)\mathrm{d}x. \tag{5}$$

Theorem 2. *When $\mathcal{X}$ and $\mathcal{Y}$ are two compact sets of $\mathbb{R}^d$ and the cost c is $\mathcal{C}^\infty$, then the Sinkhorn potentials (u, v) are uniformly bounded in the Sobolev space $\mathbf{H}^s(\mathbb{R}^d)$ and their norms satisfy*

$$\|u\|_{\mathbf{H}^s} = O\left(1 + \frac{1}{\varepsilon^{s-1}}\right) \text{ and } \|v\|_{\mathbf{H}^s} = O\left(1 + \frac{1}{\varepsilon^{s-1}}\right),$$

with constants that only depend on $|\mathcal{X}|$ (or $|\mathcal{Y}|$ for v),d, and $\left\|c^{(k)}\right\|_\infty$ for $k = 0, \dots, s$. In particular, we get the following asymptotic behavior in ε: $\|u\|_{\mathbf{H}^s} = O(1)$ as $\varepsilon \to +\infty$ and $\|u\|_{\mathbf{H}^s} = O(\frac{1}{\varepsilon^{s-1}})$ as $\varepsilon \to 0$.

To prove this theorem, we first need to state some regularity properties of the Sinkhorn potentials.

Proposition 1. *If $\mathcal{X}$ and $\mathcal{Y}$ are two compact sets of $\mathbb{R}^d$ and the cost c is $\mathcal{C}^\infty$, then*

- *$u(x) \in [\min_y v(y) - c(x,y), \max_y v(y) - c(x,y)]$ for all $x \in \mathcal{X}$*
- *u is L-Lipschitz, where L is the Lipschitz constant of c*
- *$u \in \mathcal{C}^\infty(\mathcal{X})$ and $\left\|u^{(k)}\right\|_\infty = O(1 + \frac{1}{\varepsilon^{k-1}})$*

and the same results also stand for v (inverting u and v in the first item, and replacing $\mathcal{X}$ by $\mathcal{Y}$).

Proof. The proofs of all three claims exploit the optimality condition of the dual problem:

$$\exp\left(\frac{-u(x)}{\varepsilon}\right) = \int \exp\left(\frac{v(y) - c(x,y)}{\varepsilon}\right)\beta(y)\mathrm{d}y. \tag{6}$$

Since β is a probability measure, $e^{\frac{-u(x)}{\varepsilon}}$ is a convex combination of $\varphi : x \mapsto e^{\frac{v(x)-c(x,y)}{\varepsilon}}$ and thus $e^{\frac{-u(x)}{\varepsilon}} \in [\min_y \varphi(y), \max_y \varphi(y)]$. We get the desired bounds by taking the logarithm. The two other points use the following lemmas:

Lemma 1. *The derivatives of the potentials are given by the following recurrence*

$$u^{(n)}(x) = \int g_n(x,y)\gamma_\varepsilon(x,y)\beta(y)\mathrm{d}y, \tag{7}$$

where

$$g_{n+1}(x,y) = g_n'(x,y) + \frac{u'(x) - c'(x,y)}{\varepsilon} g_n(x,y),$$

$g_1(x,y) = c'(x,y)$ *and* $\gamma_\varepsilon(x,y) = \exp(\frac{u(x)+v(y)-c(x,y)}{\varepsilon})$.

Lemma 2. *The sequence of auxiliary functions* $(g_k)_{k=0...}$ *verifies* $\left\|u^{(k)}\right\|_\infty \leqslant \|g_k\|_\infty$. *Besides, for all* $j = 0, \dots, k$, *for all* $k = 0, \dots, n-2$, $\left\|g_{n-k}^{(j)}\right\|_\infty$ *is bounded by a polynomial in* $\frac{1}{\varepsilon}$ *of order* $n - k + j - 1$.

The detailed proofs of the lemmas can be found in the appendix. We give here a sketch in the case where $d = 1$. Lemma 1 is obtained by a simple recurrence, consisting in differentiating both sides of the dual optimality condition. Differentiating under the integral is justified with the usual domination theorem, bounding the integrand thanks to the Lipschitz assumption on c, and this bound is integrable thanks to the marginal constraint. Differentiating once and rearranging terms gives:

$$u'(x) = \int c'(x,y)\gamma_\varepsilon(x,y)\beta(y)dy. \tag{8}$$

where γ_ε is defined in Lemma 1. One can easily see that $\gamma'_\varepsilon(x,y) = \frac{u'(x)-c'(x,y)}{\varepsilon}\gamma_\varepsilon(x,y)$ and this allows to conclude the recurrence, by differentiating both sides of the equality. From the primal constaint, we have that $\int_\mathcal{Y} \gamma_\varepsilon(x,y)\beta(y)\mathrm{d}y = 1$. Thus thanks to Lemma 1 we immediately get that $\left\|u^{(n)}\right\|_\infty \leqslant \|g_n\|_\infty$. For $n = 1$, since $g_1 = c'$ we get that $\|u'\|_\infty = \|c'\|_\infty = L$ and this proves the second point of Proposition 1. For higher values of n, we need the result from Lemma 2. This property is also proved by recurrence, but requires a bit more work. To prove the induction step, we need to go from bounds on $g_{n-k}^{(i)}$, for $k = 0, \dots, n-2$ and $i = 0, \dots, k$ to bounds on $g_{n+1-k}^{(i)}$, for $k = 0, \dots, n-1$ and $i = 0, \dots, k$. Hence only new quantities that we need to bound are $g_{n+1-k}^{(k)}, k = 0, \dots, n-1$. This is done by another (backwards) recurrence on k which involves some tedious computations, based on Leibniz formula, that are detailed in the appendix. □

Combining the bounds of the derivatives of the potentials with the definition of the norm in $\mathbf{H}^s$, is enough to complete the proof of Theorem 2.

Proof. (Theorem 2) The norm of u in $\mathbf{H}^s(\mathcal{X})$ is

$$\|u\|_{\mathbf{H}^s} = \left(\sum_{|k|\leqslant s}\int_\mathcal{X}(u^{(k)})^2\right)^{\frac{1}{2}} \leqslant |\mathcal{X}|\left(\sum_{|k|\leqslant s}\left\|u^{(k)}\right\|_\infty^2\right)^{\frac{1}{2}}.$$

From Proposition 1 we have that $\forall k, \left\|u^{(k)}\right\|_\infty = O(1 + \frac{1}{\varepsilon^{k-1}})$ and thus we get that $\|u\|_{\mathbf{H}^s} = O(1 + \frac{1}{\varepsilon^{s-1}})$. We just proved the bound in $\mathbf{H}^s(\mathcal{X})$ but we actually want to have a bound on $\mathbf{H}^s(\mathbb{R}^d)$. This is immediate thanks to the Sobolev extension theorem (Calderón, 1961) which guarantees that $\|u\|_{\mathbf{H}^s(\mathbb{R}^d)} \leqslant C\|u\|_{\mathbf{H}^s(\mathcal{X})}$ under the assumption that $\mathcal{X}$ is a bounded Lipschitz domain. □

This result, aside from proving useful in the next section to obtain sample complexity results on the Sinkhorn divergence, also proves that kernel-SGD can be used to solve continuous regularized OT. This idea introduced in Genevay et al. (2016) consists in assuming the potentials are in the ball of a certain RKHS, to write them as a linear combination of kernel functions and then perform stochastic gradient descent on these coefficients. Knowing the radius of the ball and the kernel associated with the RKHS (here the Sobolev or Matérn kernel) is crucial to obtain good numerical performance and ensure the convergence of the algorithm.

5 Approximation from Samples

In practice, measures α and β are only known through a finite number of samples. Thus, what can be actually computed in practice is the Sinkhorn divergence between the empirical measures $\hat{\alpha}_n \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\beta}_n \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, where $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ are n-samples from α and β, that is

$$\begin{aligned} W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n) &= \max_{u,v} \sum_{i=1}^n u(X_i) + \sum_{i=1}^n v(Y_i) \\ &- \varepsilon \sum_{i=1}^n \exp\left(\frac{u(X_i) + v(Y_i) - c(X_i, Y_i)}{\varepsilon}\right) + \varepsilon \\ &= \max_{u,v} \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u,v) + \varepsilon, \end{aligned}$$

where $(X_i, Y_i)_{i=1}^n$ are i.i.d random variables distributed according to $\alpha \otimes \beta$. On actual samples, these quantities can be computed using Sinkhorn's algorithm (Cuturi, 2013).

Our goal is to quantify the error that is made by approximating α, β by their empirical counterparts $\hat{\alpha}_n, \hat{\beta}_n$, that is bounding the following quantity:

$$\begin{aligned} &|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)| = \\ &\qquad |\mathbb{E}f_\varepsilon^{XY}(u^*, v^*) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\hat{u}, \hat{v})|, \end{aligned} \tag{9}$$

where (u^*, v^*) are the optimal Sinkhorn potentials associated with (α, β) and $(\hat{u}, \hat{v})$ are their empirical counterparts.

Theorem 3. *Consider the Sinkhorn divergence between two measures* α *and* β *on* $\mathcal{X}$ *and* $\mathcal{Y}$ *two bounded subsets of* $\mathbb{R}^d$, *with a* $\mathcal{C}^\infty$, *L-Lipschitz cost c. One has*

$$\mathbb{E}|W_\varepsilon(\alpha, \beta) - W_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)| = O\left(\frac{e^{\frac{\kappa}{\varepsilon}}}{\sqrt{n}}\left(1 + \frac{1}{\varepsilon^{\lfloor d/2 \rfloor}}\right)\right)$$

where $\kappa = 2L|\mathcal{X}| + \|c\|_\infty$ *and constants only depend on* $|\mathcal{X}|$,$|\mathcal{Y}|$,d, *and* $\left\|c^{(k)}\right\|_\infty$ *for* $k = 0 \dots \lfloor d/2 \rfloor$. *In particular, we get the following asymptotic behavior in*

ε:

$$\mathbb{E}|W_\varepsilon(\alpha,\beta) - W_\varepsilon(\hat\alpha_n,\hat\beta_n)| = O\left(\frac{e^{\frac{\kappa}{\varepsilon}}}{\varepsilon^{\lfloor d/2\rfloor}\sqrt{n}}\right) \text{ as } \varepsilon \to 0$$

$$\mathbb{E}|W_\varepsilon(\alpha,\beta) - W_\varepsilon(\hat\alpha_n,\hat\beta_n)| = O\left(\frac{1}{\sqrt{n}}\right) \text{ as } \varepsilon \to +\infty.$$

An interesting feature from this theorem is the fact when ε is large enough, the convergence rate does not depend on ε anymore. This means that at some point, increasing ε will not substantially improve convergence. However, for small values of ε the dependence is critical.

We prove this result in the rest of this section. The main idea is to exploit standard results from PAC-learning in RKHS. Our theorem is an application of the following result from Bartlett and Mendelson (2002) (combining Theorem 12,4) and Lemma 22 in their paper):

Proposition 2. *(Bartlett-Mendelson '02) Consider α a probability distribution, ℓ a B-lipschitz loss and $\mathcal{G}$ a given class of functions. Then*

$$\mathbb{E}_\alpha\left[\sup_{g\in\mathcal{G}} \mathbb{E}_\alpha \ell(g,X) - \frac{1}{n}\sum_{i=1}^n \ell(g,X_i)\right] \leqslant 2B\mathbb{E}_\alpha \mathcal{R}(\mathcal{G}(X_1^n))$$

where $\mathcal{R}(\mathcal{G}(X_1^n))$ is the Rademacher complexity of class $\mathcal{G}$ defined by $\mathcal{R}(\mathcal{G}(X_1^n)) = \sup_{g\in\mathcal{G}} \mathbb{E}_\sigma \frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)$ where $(\sigma_i)_i$ are iid Rademacher random variables. Besides, when $\mathcal{G}$ is a ball of radius λ in a RKHS with kernel k the Rademacher complexity is bounded by

$$\mathcal{R}(\mathcal{G}_\lambda(X_1^n)) \leqslant \frac{\lambda}{n}\sqrt{\sum_{i=1}^n k(X_i,X_i)}.$$

Our problem falls in this framework thanks to the following lemma:

Lemma 3. *Let $\mathcal{H}_\lambda^s \stackrel{\text{def.}}{=} \{u \in \mathbf{H}^s(\mathbb{R}^d) \mid \|u\|_{\mathbf{H}^s(\mathbb{R}^d)} \leqslant \lambda\}$, then there exists λ such that:*

$$|W_\varepsilon(\alpha,\beta) - W_\varepsilon(\hat\alpha_n,\hat\beta_n)| \leqslant 3\sup_{(u,v)\in(\mathcal{H}_\lambda^s)^2} |\mathbb{E}f_\varepsilon^{XY}(u,v) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u,v)|.$$

Proof. Inserting $\mathbb{E}f_\varepsilon^{XY}(\hat u,\hat v)$ and using the triangle inequality in (9) gives

$$|W_\varepsilon(\alpha,\beta)-W_\varepsilon(\hat\alpha_n,\hat\beta_n)| \leqslant |\mathbb{E}f_\varepsilon^{XY}(u^*,v^*)-\mathbb{E}f_\varepsilon^{XY}(\hat u,\hat v)| + |\mathbb{E}f_\varepsilon^{XY}(\hat u,\hat v) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\hat u,\hat v)|.$$

From Theorem 2, we know that the all the dual potentials are bounded in $\mathbf{H}^s(\mathbb{R}^d)$ by a constant λ which doesn't depend on the measures. Thus the second term is bounded by $\sup_{(u,v)\in(\mathcal{H}_\lambda^s)^2} |\mathbb{E}f_\varepsilon(u,v) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon(u,v)|$.

The first quantity needs to be broken down further. Notice that it is non-negative since (u^*,v^*) is the maximizer of $\mathbb{E}f_\varepsilon(\cdot,\cdot)$ so we can leave out the absolute value. We have:

$$\begin{aligned}\mathbb{E}f_\varepsilon^{XY}(u^*,v^*) - \mathbb{E}f_\varepsilon^{XY}(\hat u,\hat v) \leqslant\\ \mathbb{E}f_\varepsilon^{XY}(u^*,v^*) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u^*,v^*) \quad (10)\\ + \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u^*,v^*) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\hat u,\hat v) \quad (11)\\ + \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\hat u,\hat v) - \mathbb{E}f_\varepsilon^{XY}(\hat u,\hat v) \quad (12)\end{aligned}$$

Both (10) and (12) can be bounded by $\sup_{(u,v)\in(\mathcal{H}_\lambda^s)^2} |\mathbb{E}f_\varepsilon^{XY}(u,v) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u,v)|$ while (11) is non-positive since $(\hat u,\hat v)$ is the maximizer of $\frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\cdot,\cdot)$. ☐

To apply Proposition 2 to Sinkhorn divergences we need to prove that *(a)* the optimal potentials are in a RKHS and *(b)* our loss function f_ε is Lipschitz in the potentials.

The first point has already been proved in the previous section. The RKHS we are considering is $\mathbf{H}^s(\mathbb{R}^d)$ with $s = \lfloor\frac{d}{2}\rfloor + 1$. It remains to prove that f_ε is Lipschitz in (u,v) on a certain subspace that contains the optimal potentials.

Lemma 4. *Let $\mathcal{A} = \{(u,v) \mid u \oplus v \leqslant 2L|\mathcal{X}| + \|c\|_\infty\}$. We have:*

(i) the pairs of optimal potentials (u^*,v^*) such that $u^*(0) = 0$ belong to $\mathcal{A}$,

(ii) f_ε is B-Lipschitz in (u,v) on $\mathcal{A}$ with $B \leqslant 1 + \exp(2\frac{L|\mathcal{X}|+\|c\|_\infty}{\varepsilon})$.

Proof. Let us prove that we can restrict ourselves to a subspace on which f_ε is Lipschitz in (u,v).

$$f_\varepsilon^{xy}(u,v) = u(x)+v(y)-\varepsilon\exp\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right)$$

$$\nabla f_\varepsilon(u,v) = 1 - \exp\left(\frac{u+v-c}{\varepsilon}\right).$$

To ensure that f_ε is Lipschitz, we simply need to ensure that the quantity inside the exponential is upper-bounded at optimality and then restrict the function to all (u,v) that satisfy that bound.

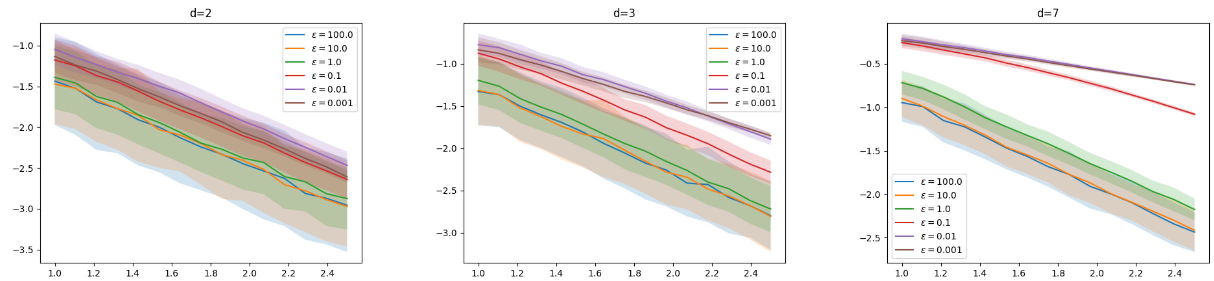

Figure 1: $\bar{W}_\varepsilon(\hat\alpha_n, \hat\alpha'_n)$ as a function of n in log-log space : Influence of ε for fixed d on two uniform distributions on the hypercube with quadratic cost.

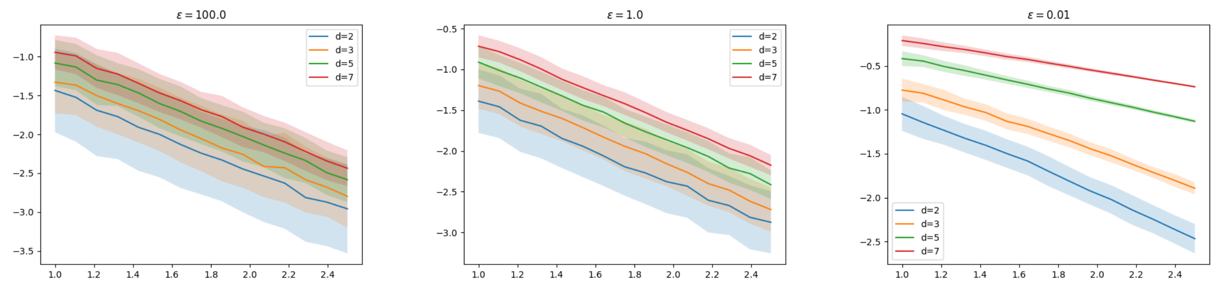

Figure 2: $\bar{W}_\varepsilon(\hat\alpha_n, \hat\alpha'_n)$ as a function of n in log-log space : Influence of d for fixed ε on two uniform distributions on the hypercube with quadratic cost.

Recall the bounds on the optimal potentials from Proposition 1. We have that $\forall x \in \mathcal{X}, y \in \mathcal{Y}$,

$$u(x) \leqslant L|x| \quad \text{and} \quad v(y) \leqslant \max_x u(x) - c(x,y).$$

Since we assumed $\mathcal{X}$ to be a bounded set, denoting by $|\mathcal{X}|$ the diameter of the space we get that at optimality $\forall x \in \mathcal{X}, y \in \mathcal{Y}$

$$u(x) + v(y) \leqslant 2L|\mathcal{X}| + \|c\|_\infty .$$

Let us denote $\mathcal{A} = \{(u,v) \in (\mathbf{H}^s(\mathbb{R}^d))^2 \mid u \oplus v \leqslant 2L|\mathcal{X}| + \|c\|_\infty\}$, we have that $\forall (u,v) \in \mathcal{A}$,

$$|\nabla f_\varepsilon(u,v)| \leqslant 1 + \exp(2\frac{L|\mathcal{X}| + \|c\|_\infty}{\varepsilon}). \qquad \square$$

We now have all the required elements to prove our sample complexity result on the Sinkhorn loss, by applying Proposition 2.

Proof. (Theorem 3) Since f_ε is Lipschitz and we are optimizing over $\mathbf{H}^s(\mathbb{R}^d)$ which is a RKHS, we can apply Proposition 2 to bound the sup in Lemma 3. We get:

$$\mathbb{E}|W_\varepsilon(\alpha,\beta) - W_\varepsilon(\hat\alpha_n, \hat\beta_n)| \leqslant 3\frac{2B\lambda}{n}\mathbb{E}\sqrt{\sum_{i=1}^n k(X_i,X_i)}$$

where $B \leqslant 1 + \exp(2\frac{L|\mathcal{X}|+\|c\|_\infty}{\varepsilon})$ (Lemma 4), $\lambda = O(\max(1, \frac{1}{\varepsilon^{d/2}}))$ (Theorem 2). We can further bound $\sqrt{\sum_{i=1}^n k(X_i,X_i)}$ by $\sqrt{n \max_{x\in\mathcal{X}} k(x,x)}$ where k is the kernel associated to $H^s(\mathbb{R}^d)$ (usually called Matern or Sobolev kernel) and thus $\max_{x\in\mathcal{X}} k(x,x) = k(0,0) := K$ which doesn't depend on n or ε. Combining all these bounds, we get the convergence rate in $\frac{1}{\sqrt{n}}$ with different asymptotic behaviors in ε when it is large or small. $\square$

Using similar arguments, we can also derive a concentration result:

Corollary 1. *With probability at least $1-\delta$,*

$$|W_\varepsilon(\alpha,\beta) - W_\varepsilon(\hat\alpha_n, \hat\beta_n)| \leqslant 6B\frac{\lambda K}{\sqrt{n}} + C\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$$

where B, λ, K are defined in the proof above, and $C = \kappa + \varepsilon\exp(\frac{\kappa}{\varepsilon})$ with $\kappa = 2L|\mathcal{X}| + \|c\|_\infty$.

Proof. We apply the bounded differences (Mc Diarmid) inequality to $g : (x_1,\dots,x_n) \mapsto \sup_{u,v\in\mathcal{H}^s_\lambda}(\mathbb{E}f_\varepsilon^{XY} - \frac{1}{n}f_\varepsilon^{X_i,Y_i})$. From Lemma 4 we get that $\forall x,y$, $f_\varepsilon^{xy}(u,v) \leqslant \kappa + \varepsilon e^{\kappa/\varepsilon} \stackrel{\text{def.}}{=} C$, and thus, changing one of the variables in g changes the value of the function by at most $2C/n$. Thus the bounded differences inequality gives

$$\mathbb{P}\left(|g(X_1,\dots,X_n) - \mathbb{E}g(X_1,\dots,X_n)| > t\right) \leqslant 2\exp(\frac{t^2 n}{2C^2})$$

Choosing $t = C\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$ yields that with probability at least $1-\delta$

$$g(X_1,\ldots,X_n) \leqslant \mathbb{E}g(X_1,\ldots,X_n) + C\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$$

and from Theorem 3 we already have

$$\mathbb{E}g(X_1,\ldots,X_n) = \mathbb{E}\sup_{u,v\in\mathcal{H}^s_\lambda}(\mathbb{E}f^{XY}_\varepsilon - \frac{1}{n}f^{X_i,Y_i}_\varepsilon) \leqslant \frac{2B\lambda K}{\sqrt{n}}.$$

□

6 Experiments

We conclude with some numerical experiments on the sample complexity of Sinkhorn Divergences. Since there are no explicit formulas for W_ε in general, we consider $\bar{W}_\varepsilon(\hat{\alpha}_n, \hat{\alpha}'_n)$ where $\hat{\alpha}_n \stackrel{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, $\hat{\alpha}'_n \stackrel{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{X_i'}$ and $(X_1,\ldots,X_n)$ and $(X'_1,\ldots,X'_n)$ are two independent n-samples from α. Note that we use in this section the normalized Sinkhorn Divergence as defined in (2), since we know that $\bar{W}_\varepsilon(\alpha,\alpha) = 0$ and thus $\bar{W}_\varepsilon(\hat{\alpha}_n, \hat{\alpha}'_n) \to 0$ as $n \to +\infty$.

Each of the experiments is run 300 times, and we plot the average of $\bar{W}_\varepsilon(\hat{\alpha}_n, \hat{\alpha}'_n)$ as a function of n in log-log space, with shaded standard deviation bars.

First, we consider the uniform distribution over a hypercube with the standard quadratic cost $c(x,y) = \|x-y\|_2^2$, which falls within our framework, as we are dealing with a $\mathcal{C}^\infty$ cost on a bounded domain. Figure 1 shows the influence of the dimension d on the convergence, while Figure 2 shows the influence of the regularization ε on the convergence for a given dimension. The influence of ε on the convergence rate increases with the dimension: the curves are almost parallel for all values of ε in dimension 2 but they get further apart as dimension increases. As expected from our bound, there is a cutoff which happens here at $\varepsilon = 1$. All values of $\varepsilon \geqslant 1$ have similar convergence rates, and the dependence on $\frac{1}{\varepsilon}$ becomes clear for smaller values. The same cutoff appears when looking at the influence of the dimension on the convergence rate for a fixed ε. The curves are parallel for all dimensions for $\varepsilon \geqslant 1$ but they have very different slopes for smaller ε.

We relax next some of the assumptions needed in our theorem to see how the Sinkhorn divergence behaves empirically. First we relax the regularity assumption on the cost, using $c(x,y) = \|x-y\|_1$. As seen on the two left images in figure 3 the behavior is very similar to the quadratic cost but with a more pronounced influence of ε, even for small dimensions. The fact that the convergence rate gets slower as ε gets smaller is already very clear in dimension 2, which wasn't the case for the quadratic cost. The influence of the dimension for a given value of ε is not any different however.

We also relax the bounded domain assumption, considering a standard normal distribution over $\mathbb{R}^d$ with a quadratic cost. While the influence of ε on the convergence rate is still obvious, the influence of the dimension is less clear. There is also a higher variance, which can be expected as the concentration bound from Corollary 1 depends on the diameter of the domain.

For all curves, we observe that d and ε impact variance, with much smaller variance for small values of ε and high dimensions. From the concentration bound, the dependency on ε coming from the uniform bound on f_ε is of the form $\varepsilon\exp(\kappa/\varepsilon)$, suggesting higher variance for small values of ε. This could indicate that our uniform bound on f_ε is not tight, and we should consider other methods to get tighter bounds in further work.

7 Conclusion

We have presented two convergence theorems for SDs: a bound on the approximation error of OT and a sample complexity bound for empirical Sinkhorn divergences. The $1/\sqrt{n}$ convergence rate is similar to MMD, but with a constant that depends on the inverse of the regularization parameter, which nicely complements the interpolation property of SDs pointed out in recent papers. Furthermore, the reformulation of SDs as the maximization of an expectation in a RKHS ball also opens the door to a better use of kernel-SGD for the computation of SDs.

Our numerical experiments suggest some open problems. It seems that the convergence rate still holds for unbounded domains and non-smooth cost functions. Besides, getting tighter bounds in our theorem might allow us to derive a sharp estimate on the optimal ε to approximate OT for a given n, by combining our two convergence theorems together.

References

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Springer, 2006.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3 (Nov):463–482, 2002.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approx-

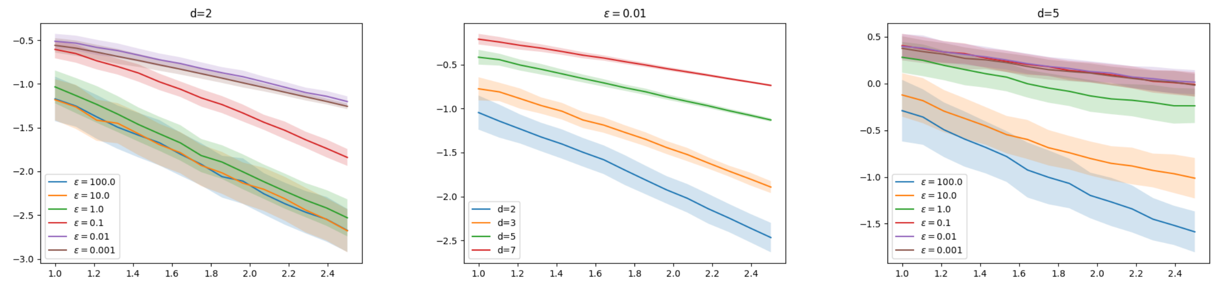

Figure 3: $\bar{W}_\varepsilon(\hat{\alpha}_n, \hat{\alpha}'_n)$ as a function of n in log-log space - cost $c(x,y) = \|x-y\|_1$ with uniform distributions (two leftmost figures) and quadratic cost $c(x,y) = \|x-y\|_2^2$ with standard normal distributions (right figure).

imate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

J. Bigot, E. Cazelles, and N. Papadakis. Central limit theorems for sinkhorn divergence between probability distributions on finite spaces and statistical applications. *arXiv preprint arXiv:1711.08947*, 2017.

P. J. Bushell. Hilbert's metric and positive contraction mappings in a Banach space. *Archive for Rational Mechanics and Analysis*, 52(4):330–338, 1973.

A. Calderón. Lebesgue spaces of differentiable functions. In *Proc. Sympos. Pure Math*, volume 4, pages 33–49, 1961.

G. Carlier, V. Duval, G. Peyré, and B. Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

Y. Chen, T. Georgiou, and M. Pavon. Entropic and displacement interpolation: a computational approach using the hilbert metric. *SIAM Journal on Applied Mathematics*, 76(6):2375–2396, 2016.

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Adv. in Neural Information Processing Systems*, pages 2292–2300, 2013.

R. Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1): 40–50, 1969.

G. Dziugaite, D. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence-Proceedings of the 31st Conference, UAI 2015*, pages 258–267, 2015.

J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. Poggio. Learning with a Wasserstein loss. In *Adv. in Neural Information Processing Systems*, pages 2044–2052, 2015.

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In D. D. Lee, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Proc. NIPS'16*, pages 3432–3440. Curran Associates, Inc., 2016.

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

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In *Adv. in Neural Information Processing Systems*, pages 513–520, 2006.

G. Huang, C. Guo, M. J. Kusner, Y. Sun, F. Sha, and K. Q. Weinberger. Supervised word mover's distance. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4862–4870. Curran Associates, Inc., 2016.

L. Kantorovich. On the transfer of masses (in Russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. MMD GAN: Towards deeper understanding of moment matching network. *arXiv preprint arXiv:1705.08584*, 2017.

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015.

A. Ramdas, N. G. Trillos, and M. Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018.

B. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schoelkopf, and G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

J. Weed and Q. Berthet. Estimation of smooth densities in wasserstein distance. *arXiv preprint arXiv:1902.01778*, 2019.

Appendix

1 Proof of Lemma 1 and Lemma 2

Lemma 1. *The derivatives of the potentials are given by the following recurrence*

$$u^{(n)}(x) = \int g_n(x,y)\gamma_\varepsilon(x,y)\beta(y)\mathrm{d}y, \tag{13}$$

where

$$g_{n+1}(x,y) = g_n'(x,y) + \frac{u'(x) - c'(x,y)}{\varepsilon} g_n(x,y),$$

$g_1(x,y) = c'(x,y)$ *and* $\gamma_\varepsilon(x,y) = \exp(\frac{u(x)+v(y)-c(x,y)}{\varepsilon})$.

Proof. (Lemma 1) For better clarity, we carry out the computations in dimension 1 but all the arguments are valid in higher dimension and we will clarify delicate points throughout the proof.

Differentiating both sides of the optimality condition (6) and rearranging yields

$$u'(x) = \int c'(x,y)\gamma_\varepsilon(x,y)\beta(y)dy. \tag{14}$$

Notice that $\gamma_\varepsilon'(x,y) = \frac{u'(x)-c'(x,y)}{\varepsilon}\gamma_\varepsilon(x,y)$. Thus by immediate recurrence (differentiating both sides of the equality again) we get that

$$u^{(n)}(x) = \int g_n(x,y)\gamma_\varepsilon(x,y)\beta(y)dy, \tag{15}$$

where $g_{n+1}(x,y) = g_n'(x,y) + \frac{u'(x)-c'(x,y)}{\varepsilon} g_n(x,y)$ and $g_1(x,y) = c'(x,y)$

To extend this first lemma to the d-dimensional case, we need to consider the sequence of indexes $\sigma = (\sigma_1, \sigma_2, \dots) \in \{1, \dots, d\}^{\mathbb{N}}$ which corresponds to the axis along which we successively differentiate. Using the same reasoning as above, it is straightforward to check that

$$\frac{\partial^k u}{\partial x_{\sigma_1} \dots \partial x_{\sigma_k}} = \int g_{\sigma,k}\gamma_\varepsilon$$

where $g_{\sigma,1} = \frac{\partial c}{\partial x_{\sigma_1}}$ and $g_{\sigma,k+1} = \frac{\partial g_{\sigma,k+1}}{\partial x_{\sigma_{k+1}}} + \frac{1}{\varepsilon}\left(\frac{\partial u}{\partial x_{\sigma_{k+1}}} - \frac{\partial c}{\partial x_{\sigma_{k+1}}}\right) g_{\sigma,k+1}$

□

Lemma 2. *The sequence of auxiliary functions* $(g_k)_{k=0\dots}$ *verifies* $\left\|u^{(k)}\right\|_\infty \leqslant \|g_k\|_\infty$. *Besides, for all* $j = 0, \dots, k$, *for all* $k = 0, \dots, n-2$, $\left\|g_{n-k}^{(j)}\right\|_\infty$ *is bounded by a polynomial in* $\frac{1}{\varepsilon}$ *of order* $n-k+j-1$.

Proof. (Lemma 2) The proof is made by recurrence on the following property :
P_n : *For all* $j = 0, \dots, k$, *for all* $k = 0, \dots, n-2$, $\left\|g_{n-k}^{(j)}\right\|_\infty$ *is bounded by a polynomial in* $\frac{1}{\varepsilon}$ *of order* $n-k+j-1$.
Let us initialize the recurrence with $n = 2$

$$g_2 \quad = \quad g_1' + \frac{u' - c'}{\varepsilon} g_1 \tag{16}$$

$$\|g_2\|_\infty \quad \leqslant \quad \|g_1'\|_\infty + \frac{\|u'\|_\infty + \|c'\|_\infty}{\varepsilon}\|g_1\|_\infty \tag{17}$$

Recall that $\|u'\|_\infty = \|g_1\|_\infty = \|c'\|_\infty$. Let $C = \max_k \left\|c^{(k)}\right\|_\infty$, we get that $\|g_2\|_\infty \leqslant C + \frac{C+C}{\varepsilon}C$ which is of the required form.

Now assume that P_n is true for some $n \geqslant 2$. This means we have bounds on $g_{n-k}^{(i)}$, for $k = 0, \dots, n-2$ and $i = 0, \dots, k$. To prove the property at rank $n+1$ we want bounds on $g_{n+1-k}^{(i)}$, for $k = 0, \dots, n-1$ and $i = 0, \dots, k$.

The only new quantity that we need to bound are $g^{(k)}_{n+1-k}, k = 0, \ldots, n-1$. Let us start by bounding $g_2^{(n-1)}$ which corresponds to $k = n-1$ and we will do a backward recurrence on k. By applying Leibniz formula for the successive derivatives of a product of functions, we get

$$g_2 = g_1' + \frac{u' - c'}{\varepsilon} g_1 \tag{18}$$

$$g_2^{(n-1)} = g_1^{(n)} + \sum_{p=0}^{n-1} \binom{n-1}{p} \frac{u^{(p+1)} - c^{(p+1)}}{\varepsilon} g_1^{(n-1-p)} \tag{19}$$

$$\left\| g_2^{(n-1)} \right\|_\infty \leqslant \left\| g_1^{(n)} \right\|_\infty + \sum_{p=0}^{n-1} \binom{n-1}{p} \frac{\left\| u^{(p+1)} \right\|_\infty + \left\| c^{(p+1)} \right\|_\infty}{\varepsilon} \left\| g_1^{(n-1-p)} \right\|_\infty \tag{20}$$

$$\leqslant C + \sum_{p=0}^{n-1} \binom{n-1}{p} \frac{\| g_{p+1} \|_\infty + C}{\varepsilon} C \tag{21}$$

Thanks to P_n we have that $\|g_p\|_\infty \leqslant \sum_{i=0}^{p} a_{i,p} \frac{1}{\varepsilon^i}, p = 1, \ldots, n$ so the highest order term in ε in the above inequality is $\frac{1}{\varepsilon^n}$. Thus we get $\left\| g_2^{(n-1)} \right\|_\infty \leqslant \sum_{i=0}^{n+1} a_{i,2,n-1} \frac{1}{\varepsilon^i}$ which is of the expected order

Now assume $g^{(j)}_{n+1-j}$ are bounded with the appropriate polynomials for $j < k \leqslant n-1$. Let us bound $g^{(k)}_{n+1-k}$

$$\left\| g^{(k)}_{n+1-k} \right\|_\infty \leqslant \left\| g^{(k+1)}_{n-k} \right\|_\infty + \sum_{p=0}^{k} \binom{k}{p} \frac{\left\| u^{(p+1)} \right\|_\infty + \left\| c^{(p+1)} \right\|_\infty}{\varepsilon} \left\| g^{(k-p)}_{n-k} \right\|_\infty \tag{22}$$

$$\leqslant \left\| g^{(k+1)}_{n-k} \right\|_\infty + \sum_{p=0}^{k} \binom{k}{p} \frac{\| g_{p+1} \|_\infty + C}{\varepsilon} \left\| g^{(k-p)}_{n-k} \right\|_\infty \tag{23}$$

The first term $\left\| g^{(k+1)}_{n-k} \right\|_\infty$ is bounded with a polynomial of order $\frac{1}{\varepsilon^{n+1}}$ by recurrence assumption. Regarding the terms in the sum, they also have all been bounded and

$$\| g_{p+1} \|_\infty \left\| g^{(k-p)}_{n-k} \right\|_\infty \leqslant \left(\sum_{i=0}^{p} a_{i,p+1} \frac{1}{\varepsilon^i} \right) \left(\sum_{i=0}^{n-p} a_{i,n-k,k-p} \frac{1}{\varepsilon^i} \right) \leqslant \sum_{i=0}^{n} \tilde{a}_i \frac{1}{\varepsilon^i}$$

So $\left\| g^{(k)}_{n+1-k} \right\|_\infty \leqslant \sum_{i=0}^{n+1} a_{i,n+1-k,k} \frac{1}{\varepsilon^i}$

To extend the result in $\mathbb{R}^d$, the recurrence is made on the the following property

$$\left\| g^{(j)}_{\sigma,n-k} \right\|_\infty \leqslant \sum_{i=0}^{n-k+|j|-1} a_{i,n-k,j,\sigma} \frac{1}{\varepsilon^i} \qquad \forall j \mid |j| = 0, \ldots, k \quad \forall k = 0, \ldots, n-2 \quad \forall \sigma \in \{1, \ldots, d\}^{\mathbb{N}} \tag{24}$$

where j is a multi-index since we are dealing with multi-variate functions, and $g_{\sigma,n-k}$ is defined at the end of the previous proof. The computations can be carried out in the same way as above, using the multivariate version of Leibniz formula in (19) since we are now dealing with multi-indexes. □

2 Existence of the Dual Potentials

We prove the following theorem, which guarantees the existence of solutions to the dual problem, in a general setting. This proof is based on the same idea from that of the existence of a solution to Schrodinger's system (which shares strong links with regularized OT) in (Chen et al., 2016), inspired from the original proof of (Franklin and Lorenz, 1989) which deals with discrete regularized OT.

Theorem 4. ***(Existence of a dual solution)*** *Consider the dual of entropy-regularized OT, with marginals $\alpha, \beta \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y})$ supported on two subsets of $\mathbb{R}^d$, and with a cost function c bounded on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{L}^\infty(\alpha) \overset{\text{def.}}{=} \{f : \mathcal{X} \to \mathbb{R} | \exists C > 0$ such that $f(x) \leqslant C$ α-a.e.$\}$ and $L^\infty(\alpha)$ the quotient set of $\mathcal{L}^\infty(\alpha)$ by the equivalence relation 'being equal α-a.e.'. Then the dual problem has solutions $(u^*, v^*) \in L^\infty(\alpha) \times L^\infty(\beta)$ which are unique $\alpha-$ and $\beta-$a.e. up to an additive constant.*

It is straightforward to see that for any solution (u^*, v^*) to the dual problem, the pair $(u^* + k, v^* - k)$ for $k \in \mathbb{R}$ is also a solution to the dual problem. Besides, modifying the values of u^* and v^* outside of the support of the measures does not have any effect on the value of the problem.

The proof of existence of a solution to the dual problem essentially amounts to rewriting the optimality condition as a fixed point equation, and proving that a fixed point exists. To do so, we show that the operator in the fixed point equation is a contraction for a certain metric, called the Hilbert metric. We prove the existence of potentials in a general framework, as we consider arbitrary measures α and β and any bounded regular cost function c.

The dual problem is unconstrained, and it is jointly concave in both variables. Thus, we can fix one and optimize over the other, and the first order condition for u gives:

$$u(x) = -\varepsilon \log \left(\int_{\mathcal{Y}} e^{\frac{v(y)-c(x,y)}{\varepsilon}} \mathrm{d}\beta(y) \right) \quad \text{for a.e. } x \in \mathcal{X}, \tag{25}$$

and similarly for v:

$$v(y) = -\varepsilon \log \left(\int_{\mathcal{X}} e^{\frac{u(x)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(y) \right) \quad \text{for a.e. } y \in \mathcal{Y}. \tag{26}$$

Remark 1. Although the optimality conditions (25) and (26) only fix the value of the optimal potentials (u^*, v^*) on the supports of α and β respectively, they allow to extrapolate the values of the potentials outside of this support.

2.1 Hilbert Metric

We start with a few definitions and properties of the Hilbert metric, which will be useful later on. Proof of these results can be found in (Bushell, 1973).

Definition 3. ***(Hilbert metric)*** *Consider $\mathcal{K}$ a closed solid cone on a real Banach space $\mathcal{B}$ i.e. $\mathcal{K}$ satisfies the 4 following properties:*

1. *the interior of $\mathcal{K}$ is not empty,*
2. *$\mathcal{K} + \mathcal{K} \subseteq \mathcal{K}$,*
3. *$\alpha\mathcal{K} \subseteq \mathcal{K} \, \forall \alpha \geqslant 0$,*
4. *$\mathcal{K} \cap -\mathcal{K} = \{0\}$.*

We use the partial order induced by the cone, meaning $x \leqslant y \Leftrightarrow y - x \in \mathcal{K}$, and define the following quantities

$$M(a,b) \stackrel{\text{def.}}{=} \inf\{\lambda | a \leqslant \lambda b\} \quad \textit{and} \quad m(a,b) \stackrel{\text{def.}}{=} \sup\{\lambda | a \leqslant \lambda b\} \quad \textit{for } a, b \in \mathcal{K}^+ \stackrel{\text{def.}}{=} \mathcal{K} \setminus \{0\}.$$

Then the Hilbert metric d_H on $\mathcal{K}^+$ is given by

$$d_H(a,b) \stackrel{\text{def.}}{=} \log \frac{M(a,b)}{m(a,b)}. \tag{27}$$

Note that the Hilbert metric is projective, meaning that it is invariant by multiplication by a positive factor: $d_H(a,b) = d_H(\alpha a, b) = d_H(a, \alpha b)$, $\forall\, \alpha > 0$.

We denote by $x \sim y$ the equivalence relation induced by the Hilbert metric, i.e. $x \sim y \Leftrightarrow d_H(x,y) = 0$ .The Hilbert metric is a pseudo-metric on the interior of the cone $\mathring{\mathcal{K}}$, and a metric on the quotient of $\mathring{\mathcal{K}}$ to the unit sphere:

Proposition 3. *(Bushell, 1973) $(\mathring{\mathcal{K}}, d_H)$ is a pseudo-metric space and $(\mathring{\mathcal{K}}/\sim, d_H)$ is a metric space. Besides if the norm induced by the cone is monotonic, i.e. $0 \leqslant x \leqslant y \Rightarrow \|x\| \leqslant \|y\|$, then $(\mathring{\mathcal{K}}/\sim, d_H)$ is a complete metric space.*

To use Banach's fixed point theorem on $(\mathring{\mathcal{K}}/\sim, d_H)$, we need to introduce the notion of contraction ratio:

Definition 4. *We say that an operator* $\mathcal{E}$ *is a* positive map *in the cone if* $\mathcal{E}(\mathring{\mathcal{K}}) \subset \mathring{\mathcal{K}}$*. For a positive map* $\mathcal{E}$*, we denote its* projective diameter *by*

$$\Delta(\mathcal{E}) \overset{\text{def.}}{=} \sup\{d_H(\mathcal{E}(a), \mathcal{E}(b)) \mid a, b \in \mathring{\mathcal{K}}\},$$

and its contraction ratio

$$\kappa(\mathcal{E}) \overset{\text{def.}}{=} \inf\{\lambda \mid d_H(\mathcal{E}(a), \mathcal{E}(b)) \leqslant \lambda d_H(x, y) \forall x, y \in \mathring{\mathcal{K}}\}.$$

In the case where the mapping is linear, we have a relation between the contraction ratio and the projective diameter.

Proposition 4. *Consider a* linear *positive map* $\mathcal{E}$ *on* $\mathring{\mathcal{K}}$*, then*

$$\kappa(\mathcal{E}) \leqslant \tanh\left(\frac{1}{4}\Delta(\mathcal{E})\right),$$

and $\Delta(\mathcal{E}) \leqslant 2\sup_a\{d_H(\mathcal{E}(a), 1)) \mid a \in \mathcal{K}^+\}$.

Since $|\tanh(x)| < 1$ for $|x| < +\infty$, this means that if the projective diameter of a positive mapping is finite, then it is a contraction. The proof of the first inequality is given in (Bushell, 1973) while the second is a direct application of the triangle inequality.

2.2 Fixed Point Theorem

Now let us rewrite the optimality condition as a fixed point equation. We consider the *exponential scalings* (a, b) of the dual variables (u, v). At optimality we have that

$$a(x) = \left(\int_{\mathcal{Y}} b(y) e^{\frac{-c(x,y)}{\varepsilon}} \mathrm{d}\beta(y)\right)^{-1} \quad \text{and} \quad b(y) = \left(\int_{\mathcal{X}} a(x) e^{\frac{-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\right)^{-1}. \tag{28}$$

We define the operators $\varphi^{\varepsilon,\alpha}$ and $\varphi^{\varepsilon,\beta}$ such that

$$\varphi^{\varepsilon,\alpha}(f) \overset{\text{def.}}{=} \int_{\mathcal{X}} f(x) e^{\frac{-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x) \quad \text{and} \quad \varphi^{\varepsilon,\beta}(f) \overset{\text{def.}}{=} \int_{\mathcal{Y}} f(y) e^{\frac{-c(x,y)}{\varepsilon}} \mathrm{d}\beta(y), \tag{29}$$

and we denote by $\mathcal{E}$ the operator such that $\mathcal{E}(a) \overset{\text{def.}}{=} 1/a$.

Proposition 5. *The optimal exponential scalings* (a^*, b^*) *satisfy the following fixed-point equations:*

$$a^* = \Phi(a^*) \quad \textit{where} \quad \Phi \overset{\text{def.}}{=} \mathcal{E} \circ \varphi^{\varepsilon,\beta} \circ \mathcal{E} \circ \varphi^{\varepsilon,\alpha}, \tag{30}$$

and

$$b^* = \tilde{\Phi}(b^*) \quad \textit{where} \quad \tilde{\Phi} \overset{\text{def.}}{=} \mathcal{E} \circ \varphi^{\varepsilon,\alpha} \circ \mathcal{E} \circ \varphi^{\varepsilon,\beta}. \tag{31}$$

To prove the existence of solutions to (28) we first need to prove the following lemma

Lemma 3. *Consider the operators* Φ *defined in* (30) *and* $\tilde{\Phi}$ *defined in* (31)*, and let* $L^\infty_+(\alpha) \overset{\text{def.}}{=} \{a \in L^\infty(\alpha) \mid a(x) > 0,\ \alpha - a.e.\}$*. Then* Φ *and* $\tilde{\Phi}$ *are contractions on* $L^\infty_+(\alpha)$ *and* $L^\infty_+(\beta)$ *respectively with contraction ratio* $\Delta(\Phi) \leqslant \tanh\left(\frac{1}{4}\frac{2\|c\|_\infty}{\varepsilon}\right) < 1$.

Proof. (Lemma 3) We consider the space of positive bounded functions $L^\infty_+(\alpha) \overset{\text{def.}}{=} \{f \in L^\infty(\alpha) | f(x) > 0 \forall x \in \mathcal{X}\}$ and $L^\infty_+(\beta)$. It is easy to check that it is a cone with non-empty interior and we can thus endow $L^\infty_+(\alpha)$ and $L^\infty_+(\beta)$ with Hilbert's metric. We also have that $\mathcal{E}$, $\varphi^{\varepsilon,\alpha}$ and $\varphi^{\varepsilon,\beta}$ are positive maps mapping L^∞_+ to itself, $L^\infty_+(\alpha)$ to $L^\infty_+(\beta)$ and $L^\infty_+(\beta)$ to $L^\infty_+(\alpha)$ respectively.

To compute the contraction ratio of the composition Φ, we can simply compute the contraction ratio of each of the composing functions and multiply them to get the whole contraction ratio.

The inversion operator $\mathcal{E}$ is an isometry for Hilbert's metric:

$$d_H(\mathcal{E}(a),\mathcal{E}(b)) = \frac{\inf\{\lambda|1/a \leqslant \lambda 1/b\}}{\sup\{\lambda|1/a \leqslant \lambda 1/b\}} = \frac{\inf\{\lambda|a \leqslant \lambda b\}}{\sup\{\lambda|b \leqslant \lambda a\}} = d_H(b,a) = d_H(a,b).$$

We are left with computing the contraction ratio of $\varphi^{\varepsilon,\alpha}$ and $\varphi^{\varepsilon,\beta}$. Since they are both linear maps, we can instead consider the quantity $\sup_a\{d_H(\varphi^\varepsilon(a),1) \mid a \in \mathcal{K}^+\}$ thanks to proposition 4. We focus on $\varphi^{\varepsilon,\alpha}$ as $\varphi^{\varepsilon,\beta}$ behaves the same way. We have that $\forall a \in L^\infty(\alpha)$

$$e^{\frac{-\|c\|_\infty}{\varepsilon}}\int_X a(x)\mathrm{d}\alpha(x) \leqslant \int_X a(x)e^{\frac{-c(x,y)}{\varepsilon}}\mathrm{d}\alpha(x) \leqslant e^{\frac{\|c\|_\infty}{\varepsilon}}\int_X a(x)\mathrm{d}\alpha(x),$$

and thus

$$\Delta(\varphi^{\varepsilon,\alpha}) \leqslant 2\sup_a\left(\log\frac{\sup\varphi^{\varepsilon,\alpha}(a)}{\inf\varphi^{\varepsilon,\alpha}(a)}\right) \leqslant 2\log\left(e^{\frac{2\|c\|_\infty}{\varepsilon}}\right) < \infty.$$

Combining all contraction ratios, we get $\Delta(\Phi) \leqslant \tanh\left(\frac{1}{4}\log\left(e^{\frac{2\|c\|_\infty}{\varepsilon}}\right)\right) < 1$ and thus Φ is a contraction for the Hilbert metric. □

Proof. (Theorem 4) The norm induced by the cone $L^\infty_+(\alpha)$ is monotonic, as one can check that $x \leqslant y \Rightarrow \|x\|_\infty \leqslant \|y\|_\infty$. Thus, $(L^\infty_+(\alpha)/\sim, d_H)$ is a complete metric space according to Proposition 3. Thanks to Lemma 3 and Proposition 5, we can conclude with Banach's fixed point theorem that Φ and $\tilde{\Phi}$ admit a unique fixed point in $L^\infty_+(\mathcal{X})/\sim$. Since dual potentials are the log of these exponential scalings, we therefore have unicity of the potential scalings, up to an additive constant, instead of a multiplicative constant for the exponential scalings. □