Interpolating between Optimal Transport and MMD using Sinkhorn Divergences
Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, Gabriel Peyré

To cite this version:
Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, et al.. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. 2018. <hal-01898858>
Interpolating between Optimal Transport and MMD using Sinkhorn Divergences

Jean Feydy  DMA, École Normale Supérieure
CMLA, ENS Paris-Saclay

Thibault Séjourné DMA, École Normale Supérieure

François-Xavier Vialard LIGM, UPEM

Shun-ichi Amari Brain Science Institute, RIKEN

Alain Trouvé CMLA, ENS Paris-Saclay

Gabriel Peyré DMA, École Normale Supérieure

Abstract

Comparing probability distributions is a fundamental problem in data sciences. Simple norms and divergences such as the total variation and the relative entropy only compare densities in a point-wise manner and fail to capture the geometric nature of the problem. In sharp contrast, Maximum Mean Discrepancies (MMD) and Optimal Transport distances (OT) are two classes of distances between measures that take into account the geometry of the underlying space and metrize the convergence in law.

This paper studies the Sinkhorn divergences, a family of geometric divergences that interpolates between MMD and OT. Relying on a new notion of geometric entropy, we provide theoretical guarantees for these divergences: positivity, convexity and metrization of the convergence in law. On the practical side, we detail a numerical scheme that enables the large scale application of these divergences for machine learning: on the GPU, gradients of the Sinkhorn loss can be computed for batches of a million samples.

1 Introduction

Countless methods in machine learning and imaging sciences rely on comparisons between probability distributions. With applications ranging from shape matching (Vaillant and Glaunès, 2005; Kaltenmark et al., 2017) to classification (Frogner et al., 2015) and generative model training (Goodfellow et al., 2014), a common setting is that of measure fitting: given a unit-mass, positive empirical distribution $\beta \in M_1^+(X)$ on a feature space $X$, a loss function $L : M_1^+(X) \times M_1^+(X) \to \mathbb{R}$ and a model distribution $\alpha_\theta \in M_1^+(X)$ parameterized by a vector $\theta$, we strive to minimize $\theta \mapsto L(\alpha_\theta, \beta)$ through gradient descent. Numerous papers focus on the construction of suitable models $\theta \mapsto \alpha_\theta$. But which loss function $L$ should we use? If $X$ is endowed with a ground distance $d : X \times X \to \mathbb{R}_+$, taking it into account can make sense and help descent algorithm to overcome spurious local minima.

Geometric divergences for Machine Learning. Unfortunately, simple dissimilarities such as the Total Variation norm or the Kullback-Leibler relative entropy do not take into account the distance $d$ on the feature space $X$. As a result, they do not metrize the convergence in law (aka. the weak* topology of measures) and are unstable with respect to deformations of the distributions’ supports. We recall that if $X$ is compact, $\alpha_n \rightharpoonup \alpha$ if for all continuous test functions $f \in C(X)$, $(\alpha_n, f) \to (\alpha, f)$ where $(\alpha, f) \overset{\text{def.}}{=} \int_X f d\alpha = E[f(X)]$ for any random vector $X$ with law $\alpha$.

The two main classes of losses $L(\alpha, \beta)$ which avoid these shortcomings are Optimal Transport distances and Maximum Mean Discrepancies: they are continuous with respect to the convergence in law and metrize its topology. That is, $\alpha_n \to \alpha \iff L(\alpha_n, \alpha) \to 0$. The main purpose of this paper is to study the theoretical properties of a new class of geometric divergences which interpolates between these two families and thus offers an extra degree of freedom through a parameter $\varepsilon$ that can be cross-validated in typical learning scenarios.

1.1 Previous works

OT distances and entropic regularization. A first class of geometric distances between measures is that of Optimal Transportation (OT) costs, which are computed as solutions of a linear program (Kantorovich, 1942) (see (1) below in the special case $\varepsilon = 0$). Enjoying many theoretical properties, these costs allow us to lift a “ground metric” on the fea-
ture space $\mathcal{X}$ towards a metric on the space $\mathcal{M}_1^+(\mathcal{X})$ of probability distributions (Santambrogio, 2015). OT distances (sometimes referred to as Earth Mover’s Distances (Rubner et al., 2000)) are progressively being adopted as an effective tool in a wide range of situations, from computer graphics (Bonneel et al., 2016) to supervised learning (Frogner et al., 2015), unsupervised density fitting (Bassetti et al., 2006) and generative model learning (Montavon et al., 2016; Arjovsky et al., 2017; Salimans et al., 2018; Genevay et al., 2018; Sanjabi et al., 2018). However, in practice, solving the linear problem required to compute these OT distances is a challenging issue; many algorithms that leverage the properties of the underlying feature space $(\mathcal{X}, d)$ have thus been designed to accelerate the computations, see (Peyré and Cuturi, 2017) for an overview.

Out of this collection of methods, entropic regularization has recently emerged as a computationally efficient way of approximating OT costs. For $\varepsilon > 0$, we define

$$\text{OT}_\varepsilon(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi_1 = \alpha, \pi_2 = \beta} \int_{\mathcal{X}^2} C d\pi + \varepsilon \text{KL}(\pi|\alpha \otimes \beta)$$

where $\text{KL}(\pi|\alpha \otimes \beta) \overset{\text{def.}}{=} \int_{\mathcal{X}^2} \log \left( \frac{d\pi}{\alpha \otimes \beta} \right) d\pi$, where $C(x, y)$ is some symmetric positive cost function (we assume here that $C(x, x) = 0$) and where the minimization is performed over coupling measures $\pi \in \mathcal{M}_1^+(\mathcal{X}^2)$ as $\pi_1, \pi_2$ denotes the two marginals of $\pi$. Typically, $C(x, y) = \|x - y\|^p$ on $\mathcal{X} \subset \mathbb{R}^D$ and setting $\varepsilon = 0$ in (1) allows us to retrieve the Earth Mover ($p = 1$) or the quadratic Wasserstein ($p = 2$) distances.

The idea of adding an entropic barrier $\text{KL}(\cdot|\alpha \otimes \beta)$ to the original linear OT program can be traced back to Schrödinger’s problem (Léonard, 2013) and has been used for instance in social sciences (Galichon and Salanié, 2010). Crucially, as highlighted in (Cuturi, 2013), the smooth problem (1) can be solved efficiently on the GPU as soon as $\varepsilon > 0$: the celebrated Sinkhorn algorithm (detailed in Section 3) allows us to compute efficiently a smooth, geometric loss $\text{OT}_\varepsilon$ between sampled measures.

**MMD norms.** Still, to define geometry-aware distances between measures, a simpler approach is to integrate a positive definite kernel $k(x, y)$ on the feature space $\mathcal{X}$. On a Euclidean feature space $\mathcal{X} \subset \mathbb{R}^D$, we typically use RBF kernels such as the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/2\sigma^2)$ or the energy distance (conditionally positive) kernel $k(x, y) = -\|x - y\|$. The kernel loss is then defined, for $\xi = \alpha - \beta$,

$$L_k(\alpha, \beta) \overset{\text{def.}}{=} \frac{1}{2} \|\xi\|^2_k \overset{\text{def.}}{=} \frac{1}{2} \int_{\mathcal{X}^2} k(x, y) d\xi(x)d\xi(y).$$

If $k$ is universal (Micchelli et al., 2006) (i.e. if the linear space spanned by functions $k(x, \cdot)$ is dense in $C(\mathcal{X})$) we know that $\|\cdot\|_k$ metrizes the convergence in law. Such Euclidean norms, introduced for shape matching in (Glaunes et al., 2004), are often referred to as “Maximum Mean Discrepancies” (MMD) (Gretton et al., 2007). They have been extensively used for generative model (GANs) fitting in machine learning (Li et al., 2015; Dziugaite et al., 2015). MMD norms are cheaper to compute than OT and have a smaller sample complexity – i.e. approximation error when sampling a distribution.

### 1.2 Interpolating between OT and MMD using Sinkhorn divergences

Unfortunately though, the “flat” geometry that MMDs induce on the space of probability measures $\mathcal{M}_1^+(\mathcal{X})$ does not faithfully lift the ground distance on $\mathcal{X}$. For instance, on $\mathcal{X} = \mathbb{R}^D$, let us denote by $\alpha_\tau$ the translation of $\alpha$ by $\tau \in \mathbb{R}^D$, defined through $\langle \alpha_\tau, f \rangle = \langle \alpha, f(\cdot + \tau) \rangle$ for continuous functions $f \in C(\mathbb{R}^D)$. Wasserstein distance discrepancies defined for $C(x, y) = \|x - y\|^p$ are such that $\text{OT}_0(\alpha, \alpha_\tau)^2 = \|\tau\|^2$.

In sharp contrast, MMD norms rely on convolutions that weigh the frequencies of $\xi = \alpha - \beta$ independently, according to the Fourier transform of the kernel function. For instance, up to a multiplicative constant of $D$, the energy distance $L_{\|\cdot\|}(\alpha, \alpha_\tau)$ is given by $\int_{\mathbb{R}^D} (1 - \cos((\omega, \tau))) |\hat{\alpha}(\omega)|^2/\|\omega\|^{D+1} d\omega$ (Székely and Rizzo (2004), Lemma 1), where $\hat{\alpha}(\omega)$ is the Fourier transform of the probability measure $\alpha$. Except for the trivial case of a Dirac mass $\alpha = \delta_{x_0}$ for some $x_0 \in \mathbb{R}^D$, we thus always have $L_{\|\cdot\|}(\alpha, \alpha_\tau) < \|\tau\|$ with a value that strongly depends on the smoothness of the reference measure $\alpha$. In practice, as evidenced in Figure 5, this theoretical shortcoming of MMD losses is reflected by vanishing gradients (or similar artifacts) next to the extreme points of the measures’ supports.

**Sinkhorn divergences.** On the one hand, OT losses have appealing geometric properties; on the other hand, cheap MMD norms scales up to large batches with a low sample complexity. Why not interpolate between them to get the best of both worlds?

Following (Genevay et al., 2018) (see also (Ramdas et al., 2017; Salimans et al., 2018; Sanjabi et al., 2018)) we consider a new cost built from $\text{OT}_\varepsilon$ that we call a Sinkhorn divergence:

$$S(\alpha, \beta) \overset{\text{def.}}{=} \text{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2} \text{OT}_\varepsilon(\alpha, \alpha) - \frac{1}{2} \text{OT}_\varepsilon(\beta, \beta).$$

Such a formula satisfies $S(\beta, \beta) = 0$ and interpolates between OT and MMD (Ramdas et al., 2017):

$$\text{OT}_0(\alpha, \beta) \overset{\text{def.}}{=} S(\alpha, \beta) \overset{\text{def.}}{=} S(\alpha, \beta) \overset{\text{def.}}{=} \frac{1}{2} \|\alpha - \beta\|^2_C. \quad (4)$$
The entropic bias. Why bother with the autocorrelation terms OT_ε(α, α) and OT_ε(β, β)? For positive values of ε, in general, OT_ε(β, β) ̸= 0 so that minimizing OT_ε(α, β) with respect to α results in a biased solution: as evidenced by Figure 1, the gradient of OT_ε drives α towards a shrink measure whose support is smaller than that of the target measure β. This is most evident as ε tends to infinity: OT_ε(α, β) → \( \int \int C(x, y) \, d\alpha(x) \, d\beta(y) \), a quantity that is minimized if α is a Dirac distribution located at the median (resp. the mean) value of β if \( C(x, y) = \| x - y \| \) (resp. \( \| x - y \|^2 \)).

In the literature, the formula (3) has been introduced more or less empirically to fix the entropic bias present in the OT_ε cost: with a structure that mimicks that of a squared kernel norm (2), it was assumed or conjectured that S_ε would define a positive definite loss function, suitable for applications in ML. This paper is all about proving that this is indeed what happens.

1.3 Contributions

The purpose of this paper is to show that the Sinkhorn divergences are convex, smooth, positive definite loss functions that metrize the convergence in law. Our purpose of this paper is to show that the Sinkhorn divergences are convex, smooth, positive definite loss functions that metrize the convergence in law. Our purpose of this paper is to show that the Sinkhorn divergences are convex, smooth, positive definite loss functions that metrize the convergence in law. Our purpose of this paper is to show that the Sinkhorn divergences are convex, smooth, positive definite loss functions that metrize the convergence in law. Our purpose of this paper is to show that the Sinkhorn divergences are convex, smooth, positive definite loss functions that metrize the convergence in law. Our purpose of this paper is to show that the Sinkhorn divergences are convex, smooth, positive definite loss functions that metrize the convergence in law. Our purpose of this paper is to show that the Sinkhorn divergences are convex, smooth, positive definite loss functions that metrize the convergence in law. Our purpose of this paper is to show that the Sinkhorn divergences are convex, smooth, positive definite loss functions that metrize the convergence in law.

\[ \text{Theorem 1.} \quad \text{Let } X \text{ be a compact metric space with a Lipschitz cost function } C(x, y) \text{ that induces, for } \varepsilon > 0, \text{ a positive universal kernel } k_\varepsilon(x, y) \overset{\text{def}}{=} \exp(-C(x, y)/\varepsilon). \text{ Then, } S_\varepsilon \text{ defines a symmetric positive definite, smooth loss function that is convex in each of its input variables. It also metrizates the convergence in law: for all probability Radon measures } \alpha \text{ and } \beta \in M^+_1(X), \]

\[
0 = S_\varepsilon(\beta, \beta) \leq S_\varepsilon(\alpha, \beta), \\
\alpha = \beta \iff S_\varepsilon(\alpha, \beta) = 0, \\
\alpha_n \rightharpoonup \alpha \iff S_\varepsilon(\alpha_n, \alpha) \to 0. \tag{7}
\]

Notably, these results also hold for measures with bounded support on a Euclidean space \( X = \mathbb{R}^D \) endowed with ground cost functions \( C(x, y) = \| x - y \| \) or \( C(x, y) = \| x - y \|^2 \) which induce Laplacian and Gaussian kernels respectively.

This theorem legitimizes the use of the unbiased Sinkhorn divergences \( S_\varepsilon \) instead of \( \text{OT}_\varepsilon \) in model-fitting applications. Indeed, computing \( S_\varepsilon \) is roughly as expensive as \( \text{OT}_\varepsilon \) (the computation of the corrective factors being cheap, as detailed in Section 3) and the “debiasing” formula (3) allows us to guarantee that the unique minimizer of \( \alpha \mapsto S_\varepsilon(\alpha, \beta) \) is the target distribution \( \beta \) (see Figure 1). Section 3 details how to implement these divergences efficiently: our algorithms scale up to millions of samples thanks to freely available GPU routines. To conclude, we showcase in Section 4 the typical behavior of \( S_\varepsilon \) compared with \( \text{OT}_\varepsilon \) and standard MMD losses.

2 Proof of Theorem 1

We now give the proof of Theorem 1. Our argument relies on a new Bregman divergence derived from a weak* continuous entropy that we call the Sinkhorn entropy (see Section 2.2). We believe this (convex) entropy function to be of independent interest. Note that all this section is written under the assumptions of Theorem 1; the proof of some intermediate results can be found in the appendix.

2.1 Properties of the OT_ε loss

First, let us recall some standard results of regularized OT theory (Peyré and Cuturi, 2017). Thanks to the Fenchel-Rockafellar theorem, we can rewrite Cuturi’s loss (1) as

\[
\text{OT}_\varepsilon(\alpha, \beta) \overset{\text{def}}{=} \max_{(f, g) \in C(X)^2} \langle \alpha, f \rangle + \langle \beta, g \rangle - \varepsilon \langle \alpha \otimes \beta, \exp(\frac{1}{2}(f \oplus g - C)) \rangle - 1, \tag{8}
\]

where \( f \oplus g \) is the tensor sum \( (x, y) \in X^2 \mapsto f(x) + g(y) \). The primal-dual relationship linking an optimal transport plan \( \pi \) solving (1) to an optimal dual pair \( (f, g) \) that solves (8) is

\[
\pi = \exp\left(\frac{1}{2}(f \oplus g - C)\right) \cdot (\alpha \otimes \beta). \tag{9}
\]

Crucially, the first order optimality conditions for the dual variables are equivalent to the primal’s marginal constraints \( (\pi_1 = \alpha, \pi_2 = \beta) \) on (9). They read

\[
f = T(\beta, g) \text{ a.e. and } g = T(\alpha, f) \text{ } \beta\text{-a.e.,} \tag{10}
\]

Figure 1 – Removing the entropic bias. Solution \( \alpha \) (in red) of the fitting problem \( \min_\alpha L(\alpha, \beta) \) for some \( \beta \) shown in blue. Here, \( C(x, y) = \| x - y \| \) on the unit square \( X \) in \( \mathbb{R}^2 \) and \( \varepsilon = .1 \). The positions of the red dots were optimized by gradient descent, starting from a normal Gaussian sample.
where the “Sinkhorn mapping” $T : \mathcal{M}_1^+(\mathcal{X}) \times \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ is defined through

$$T : (\alpha, f) \mapsto \left( y \in \mathcal{X} \mapsto \min_{x \sim \alpha} \left[ C(x, y) - f(x) \right] \right),$$  \hspace{1cm} (11)

with a SoftMin operator of strength $\varepsilon$ defined through

$$\min_{x \sim \alpha} \varphi(x) = -\varepsilon \log \int_x \exp \left( -\frac{1}{\varepsilon} \varphi(x) \right) d\alpha(x).$$  \hspace{1cm} (12)

**Dual potentials.** The following proposition recalls some important properties of $\text{OT}_\varepsilon$ and the associated dual potentials. Its proof can be found in Section B.1.

**Proposition 1** (Properties of $\text{OT}_\varepsilon$). The optimal potentials $(f, g)$ exist and are unique $(\alpha, \beta)$-a.e. up to an additive constant, i.e. $\forall K \in \mathbb{R}$, $(f + K, g - K)$ is also optimal. At optimality, we get

$$\text{OT}_\varepsilon(\alpha, \beta) = \langle \alpha, f \rangle + \langle \beta, g \rangle.$$  \hspace{1cm} (13)

We recall that a function $F : \mathcal{M}_1^+(\mathcal{X}) \to \mathbb{R}$ is said to be differentiable if there exists $\nabla F(\alpha) \in \mathcal{C}(\mathcal{X})$ such that for any displacement $\xi = \beta - \beta'$ with $(\beta, \beta') \in \mathcal{M}_1^+(\mathcal{X})^2$, we have

$$F(\alpha + t\xi) = F(\alpha) + t\xi, \nabla F(\alpha) + o(t).$$  \hspace{1cm} (14)

The following proposition, whose proof is detailed in Section B.2, shows that the dual potentials are the gradients of $\text{OT}_\varepsilon$.

**Proposition 2.** $\text{OT}_\varepsilon$ is weak* continuous and differentiable. Its gradient reads

$$\nabla \text{OT}_\varepsilon(\alpha, \beta) = (f, g)$$  \hspace{1cm} (15)

where $(f, g)$ satisfies $f = T(\beta, g)$ and $g = T(\alpha, f)$ on the whole domain $\mathcal{X}$ and $T$ is the Sinkhorn mapping (11).

Let us stress that even though the solutions of the dual problem (8) are defined $(\alpha, \beta)$-a.e., the gradient (15) is defined on the whole domain $\mathcal{X}$. Fortunately, an optimal dual pair $(f_0, g_0)$ defined $(\alpha, \beta)$-a.e. satisfies the optimality condition (10) and can be extended in a canonical way: to compute the “gradient” pair $(f, g) \in \mathcal{C}(\mathcal{X})^2$ associated to a pair of measures $(\alpha, \beta)$, using $f = T(\beta, g_0)$ and $g = T(\alpha, f_0)$ is enough.

### 2.2 Sinkhorn and Haussdorff divergences

Having recalled some standard properties of $\text{OT}_\varepsilon$, let us now state a few original facts about the corrective, symmetric term $-\frac{1}{2} \text{OT}_\varepsilon(\alpha, \alpha)$ used in (3). We still suppose that $(\mathcal{X}, d)$ is a compact set endowed with a symmetric, Lipschitz cost function $C(x, y)$. For $\varepsilon > 0$, the associated Gibbs kernel is defined through

$$k_\varepsilon : (x, y) \in \mathcal{X} \times \mathcal{X} \mapsto \exp \left( -\frac{C(x, y)}{\varepsilon} \right).$$  \hspace{1cm} (16)

Crucially, we now assume that $k_\varepsilon$ is a positive universal kernel on the space of signed Radon measures.

**Definition 1** (Sinkhorn negentropy). Under the assumptions above, we define the Sinkhorn negentropy of a probability Radon measure $\alpha \in \mathcal{M}_1^+(\mathcal{X})$ through

$$F_\varepsilon(\alpha) \overset{\text{def.}}{=} -\frac{1}{2} \text{OT}_\varepsilon(\alpha, \alpha).$$  \hspace{1cm} (17)

The following proposition is the cornerstone of our approach to prove the positivity of $S_\varepsilon$, providing an alternative expression of $F_\varepsilon$. Its proof relies on a change of variables $\mu = \exp(f/\varepsilon) \alpha$ in (8) that is detailed in the Section B.3 of the appendix.

**Proposition 3.** Let $(\mathcal{X}, d)$ be a compact set endowed with a symmetric, Lipschitz cost function $C(x, y)$ that induces a positive kernel $k_\varepsilon$. Then, for $\varepsilon > 0$ and $\alpha \in \mathcal{M}_1^+(\mathcal{X})$, one has

$$\frac{1}{2} F_\varepsilon(\alpha) + \frac{1}{2} \mu \overset{\text{def.}}{=} \min_{\mu \in \mathcal{M}_1^+(\mathcal{X})} \langle \mu, \log \frac{d\mu}{d\alpha} \rangle + \frac{1}{2} \|\mu\|_{\text{TV}}^2.$$  \hspace{1cm} (18)

The following proposition, whose proof can be found in the Section B.4 of the appendix, leverages the alternative expression (18) to ensure the convexity of $F_\varepsilon$.

**Proposition 4.** Under the same hypotheses as Proposition 3, $F_\varepsilon$ is a strictly convex functional on $\mathcal{M}_1^+(\mathcal{X})$.

We now define an auxiliary “Hausdorff” divergence that can be interpreted as an $\text{OT}_\varepsilon$ loss with decoupled dual potentials.

**Definition 2** (Hausdorff divergence). Thanks to Proposition 2, the Sinkhorn negentropy $F_\varepsilon$ is differentiable in the sense of (14). For any probability measures $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$ and regularization strength $\varepsilon > 0$, we can thus define

$$H_\varepsilon(\alpha, \beta) \overset{\text{def.}}{=} \frac{1}{2} \langle \alpha - \beta, \nabla F_\varepsilon(\alpha) - \nabla F_\varepsilon(\beta) \rangle \geq 0.$$  \hspace{1cm} (18)

It is the symmetric Bregman divergence induced by the strictly convex functional $F_\varepsilon$ (Bregman, 1967) and is therefore a positive definite quantity.

### 2.3 Proof of the Theorem

We are now ready to conclude. First, remark that the dual expression (8) of $\text{OT}_\varepsilon(\alpha, \beta)$ as a maximization of linear forms ensures that $\text{OT}_\varepsilon(\alpha, \beta)$ is convex with respect to $\alpha$ and with respect to $\beta$ (but not jointly convex if $\varepsilon > 0$). $S_\varepsilon$ is thus convex with respect to both inputs $\alpha$ and $\beta$ as a sum of the functions $\text{OT}_\varepsilon$ and $F_\varepsilon$ – see Proposition 4.

Convexity also implies that,

$$\text{OT}_\varepsilon(\alpha, \alpha) + \langle \beta - \alpha, \nabla_2 \text{OT}_\varepsilon(\alpha, \alpha) \rangle \leq \text{OT}_\varepsilon(\alpha, \beta),$$

$$\text{OT}_\varepsilon(\beta, \beta) + \langle \alpha - \beta, \nabla_1 \text{OT}_\varepsilon(\beta, \beta) \rangle \leq \text{OT}_\varepsilon(\alpha, \beta).$$

Using (15) to get $\nabla_2 \text{OT}_\varepsilon(\alpha, \alpha) = -\nabla F_\varepsilon(\alpha)$,

$$\nabla_1 \text{OT}_\varepsilon(\beta, \beta) = -\nabla F_\varepsilon(\beta)$$

and summing the above inequalities, we show that $H_\varepsilon \leq S_\varepsilon$, which implies (5).
To prove (6), note that \( S_{\varepsilon}(\alpha, \beta) = 0 \Rightarrow H_{\varepsilon}(\alpha, \beta) = 0 \), which implies that \( \alpha = \beta \) since \( F_{\varepsilon} \) is a strictly convex functional.

Finally, we show that \( S_{\varepsilon} \) metrizes the convergence in law (7) in the Section B.5 of the appendix.

## 3 Computational scheme

We have shown that Sinkhorn divergences (3) are positive definite, convex loss functions on the space of probability measures. Let us now detail their implementation on modern hardware.

**Encoding measures.** For the sake of simplicity, we focus on discrete, sampled measures on a Euclidean feature space \( \mathcal{X} \subset \mathbb{R}^D \). Our input measures \( \alpha \) and \( \beta \in \mathcal{M}_1^+(\mathcal{X}) \) are represented as sums of weighted Dirac atoms

\[
\alpha = \sum_{i=1}^N \alpha_i \delta_{x_i}, \quad \beta = \sum_{j=1}^M \beta_j \delta_{y_j} \tag{19}
\]

and encoded as two pairs \((\alpha, x)\) and \((\beta, y)\) of float arrays. Here, \( \alpha \in \mathbb{R}^N \) and \( \beta \in \mathbb{R}^M \) are non-negative vectors of shapes \([N]\) and \([M]\) that sum up to 1, whereas \( x \in \mathbb{R}^N \) and \( y \in \mathbb{R}^M \) are real-valued tensors of shapes \([N,D]\) and \([M,D]\) — if we follow python’s convention.

### 3.1 The Sinkhorn algorithm(s)

**Working with dual vectors.** Proposition 1 is key to the modern theory of regularized Optimal Transport: it allows us to compute the \( OT_{\varepsilon} \) cost — and thus the Sinkhorn divergence \( S_{\varepsilon} \), thanks to (3) — using dual variables that have the same memory footprint as the input measures: solving (8) in our discrete setting, we only need to store the sampled values of the dual potentials \( f \) and \( g \) on the measures’ supports.

We can thus work with dual vectors \( f \in \mathbb{R}^N \) and \( g \in \mathbb{R}^M \), defined through \( f_i = f(x_i) \) and \( g_j = g(y_j) \), which encode an implicit transport plan \( \pi \) from \( \alpha \) to \( \beta \) (9).

Crucially, the optimality condition (10) now reads:

\[
\forall i \in [1,N], \forall j \in [1,M],
\]

\[
f_i = -\varepsilon \text{LSE}_{k=1}^M \left( \log(\beta_k) + \frac{1}{\varepsilon} g_k - \frac{1}{\varepsilon} C(x_i, y_k) \right) \tag{20}
\]

\[
g_j = -\varepsilon \text{LSE}_{k=1}^N \left( \log(\alpha_k) + \frac{1}{\varepsilon} f_k - \frac{1}{\varepsilon} C(x_k, y_j) \right) \tag{21}
\]

where

\[
\text{LSE}_{k=1}^N(V_k) = \log \sum_{k=1}^N \exp(V_k) \tag{22}
\]

denotes a (stabilized) log-sum-exp reduction.

If \((f, g)\) is an optimal pair of dual vectors that satisfies Equations (20-21), we deduce from (13) that

\[
\text{OT}_{\varepsilon}(\alpha_i, x_i, \beta_j, y_j) = \sum_{i=1}^N \alpha_i f_i + \sum_{j=1}^M \beta_j g_j. \tag{23}
\]

But how can we solve this coupled system of equations given \( \alpha, x, \beta \) and \( y \) as input data?

**The Sinkhorn algorithm.** One simple answer: by enforcing (20) and (21) alternatively, updating the vectors \( f \) and \( g \) until convergence (Cuturi, 2013). Starting from null potentials \( f_i = 0 = g_j \), this numerical scheme is nothing but a block-coordinate ascent on the dual problem (8). One step after another, we are enforcing null derivatives on the dual cost with respect to the \( f_i \)'s and the \( g_j \)'s.

**Convergence.** The “Sinkhorn loop” converges quickly towards its unique optimal value: it enjoys a linear convergence rate (Peyré and Cuturi, 2017) that can be improved with some heuristics (Thibault et al., 2017). When computed through the dual expression (23), \( OT_{\varepsilon} \) and its gradients (26-27) are robust to small perturbations of the values of \( f \) and \( g \): monitoring convergence through the \( L^1 \) norm of the updates on \( f \) and breaking the loop as we reach a set tolerance level is thus a sensible stopping criterion. In practice, if \( \varepsilon \) is large enough — say, \( \varepsilon \geq 0.05 \) on the unit square with an Earth Mover’s cost \( C(x,y) = \|x-y\| \) — waiting for 10 or 20 iterations is more than enough.

**Symmetic \( OT_{\varepsilon} \) problems.** All in all, the baseline Sinkhorn loop provides an efficient way of solving the discrete problem \( OT_{\varepsilon}(\alpha, \beta) \) for generic input measures. But in the specific case of the (symmetric) corrective terms \( OT_{\varepsilon}(\alpha, \alpha) \) and \( OT_{\varepsilon}(\beta, \beta) \) introduced in (3), we can do better.

The key here is to remark that if \( \alpha = \beta \), the dual problem (8) becomes a concave maximization problem that is symmetric with respect to its two variables \( f \) and \( g \). Hence, there exists a (unique) optimal dual pair \((f,g=f)\) on the diagonal which is characterized in the discrete setting by the symmetric optimality condition:

\[
\forall i \in [1,N],
\]

\[
f_i = -\varepsilon \text{LSE}_{k=1}^N \left[ \log(\alpha_k) + \frac{1}{\varepsilon} f_k - \frac{1}{\varepsilon} C(x_i, x_i) \right]. \tag{24}
\]

Fortunately, given \( \alpha \) and \( x \), the optimal vector \( f \) that solves this equation can be computed by iterating a well-conditioned fixed-point update:

\[
f_i \leftarrow \frac{1}{2} \left( f_i - \varepsilon \text{LSE}_{k=1}^N \left[ \log(\alpha_k) + \frac{1}{\varepsilon} f_k - \frac{1}{\varepsilon} C(x_i, x_i) \right] \right). \tag{25}
\]

This symmetric variant of the Sinkhorn algorithm can be shown to converge much faster than the standard
3.2 Computing the Sinkhorn divergence and its gradients

Given two pairs \((\alpha, x)\) and \((\beta, y)\) of float arrays that encode the probability measures \(\alpha\) and \(\beta\) (19), we can now implement the Sinkhorn divergence \(S_{\epsilon}(\alpha, \beta)\): The cross-correlation dual vectors \(f \in \mathbb{R}^N\) and \(g \in \mathbb{R}^M\) associated to the discrete problem \(\text{OT}_{\epsilon}(\alpha, \beta)\) can be computed using the Sinkhorn iterations (20-21).

The autocorrelation dual vectors \(p \in \mathbb{R}^N\) and \(q \in \mathbb{R}^M\), respectively associated to the symmetric problems \(\text{OT}_{\epsilon}(\alpha, \alpha)\) and \(\text{OT}_{\epsilon}(\beta, \beta)\), can be computed using the symmetric Sinkhorn update (25).

The Sinkhorn loss can be computed using (3) and (23):

\[
S_{\epsilon}(\alpha_i, x_i, \beta_j, y_j) = \sum_{i=1}^{N} \alpha_i (f_i - p_i) + \sum_{j=1}^{M} \beta_j (g_j - q_j).
\]

What about the gradients? In this day and age, we could be tempted to rely on the automatic differentiation engines provided by modern libraries, which let us differentiate the result of twenty or so Sinkhorn iterations as a mere composition of elementary operations (Genevay et al., 2018). But beware: this loop has a lot more structure than a generic feed forward network. Taking advantage of it is key to a x2-x3 gain in performances, as we now describe.

Crucially, we must remember that the Sinkhorn loop is a fixed point iterative solver: at convergence, its solution satisfies an equation given by the implicit function theorem. Thanks to (15), using the very definition of gradients in the space of probability measures (14) and the intermediate variables in the computation of \(S_{\epsilon}(\alpha, \beta)\), we get that

\[
\partial_\alpha S_{\epsilon}(\alpha_i, x_i, \beta_j, y_j) = f_i - p_i \quad (26)
\]

and

\[
\partial_x S_{\epsilon}(\alpha_i, x_i, \beta_j, y_j) = \nabla \varphi(x_i), \quad (27)
\]

where \(\varphi: \mathcal{X} \to \mathbb{R}\) is equal to \(f_i - p_i\) on the \(x_i\)'s and is defined through

\[
\varphi(x) = -\epsilon \log \sum_{j=1}^{M} \exp \left[ \log(\beta_j) + \frac{1}{\epsilon} g_j - \frac{1}{\epsilon} C(x, y_j) \right] + \epsilon \log \sum_{i=1}^{N} \exp \left[ \log(\alpha_i) + \frac{1}{\epsilon} p_i - \frac{1}{\epsilon} C(x, x_i) \right].
\]

Graph surgery with PyTorch. Assuming convergence in the Sinkhorn loops, it is thus possible to compute the gradients of \(S_{\epsilon}\) without having to backprop through the twenty or so iterations of the Sinkhorn algorithm: we only have to differentiate the expression above with respect to \(x\). But does it mean that we should differentiate \(C\) or the log-sum-exp operation by hand? Fortunately, no!

Modern libraries such as PyTorch (Paszke et al., 2017) are flexible enough to let us "hack" the naive autograd algorithm, and act as though the optimal dual vectors \(f, p, g\) and \(q\) did not depend on the input variables of \(S_{\epsilon}\). As documented in our reference code,

\begin{verbatim}
github.com/jeanfeydy/global-divergences,
\end{verbatim}

an appropriate use of the .detach() method in PyTorch is enough to get the best of both worlds: an automatic differentiation engine that computes our gradients using the formula at convergence instead of the baseline backpropagation algorithm. All in all, as evidenced by the benchmarks provided Figure 3, this trick allows us to divide by a factor 2-3 the time needed to compute a Sinkhorn divergence and its gradient with respect to the \(x_i\)'s.

3.3 Scaling up to large datasets

The Sinkhorn iterations rely on a single non-trivial operation: the log-sum-exp reduction (22). In the ML literature, this SoftMax operator is often understood as a row- or column-wise reduction that acts on \([N, M]\) matrices. But as we strive to implement the update rules (20-21) and (25) on the GPU, we can go further.

Batch computation. First, if the number of samples \(N\) and \(M\) in both measures is small enough, we can optimize the GPU usage by computing Sinkhorn divergences by batches of size \(B\). In practice, this can be achieved by encoding the cost function \(C\) as a 3D tensor of size \([B, N, M]\) made up of stacked matrices \((C(x_i, y_j))_{i,j}\), while \(f\) and \(g\) become \([B, N]\) and \([B, M]\) tensors, respectively. Thanks to the broadcasting syntax supported by modern libraries, we can then seamlessly compute, in parallel, loss values \(S_{\epsilon}(\alpha_k, \beta_k)\) for \(k\) in \([1, B]\).

The KeOps library. Unfortunately though, tensor-centric methods such as the one presented above cannot scale to measures sampled with large numbers \(N\) and \(M\) of Dirac atoms: as these numbers exceed 10,000, huge \([N, M]\) matrices stop fitting into GPU memories. To alleviate this problem, we leveraged the KeOps library (Charlier et al., 2018) that provides online map-reduce routines on the GPU with full PyTorch integration. Performing online log-sum-exp reductions with a running maximum, the KeOps primitives allow us to compute Sinkhorn divergences with a linear memory footprint. As evidenced by the
of the measure $\alpha$ a "non-parametric" version of the data fitting problem $\rightarrow$ model Cuturi’s original cost OT divergence $S$.

We now describe the guarantees on top of a comprehensive implementation.

4 Numerical illustration

Then a matter of seconds. A Sinkhorn loss with 100,000 samples per measure is much faster than the baseline MMD – even though the explicit gradient formula (26-27) lets us win a factor 2-3 compared with a baseline, tensorized, implementation of $L_{\alpha}$. Deformations of the support that it favors.

In 1D, the optimal transport problem can be solved with $N$ points – we display the segment $[0,1]$. The legend on the left indicates the energy distance. Experiments performed on $\mathcal{X} = \mathbb{R}^3$ with a cheap laptop’s GPU (GTX 960M).

Figure 2 – The KeOps library allows us to break the memory bottleneck. Using CUDA routines that sum kernel values without storing them in memory, we can outperform baseline, tensorized, implementations of the energy distance. Experiments performed on $\mathcal{X} = \mathbb{R}^3$ with a cheap laptop’s GPU (GTX 960M).

Figure 3 – Sinkhorn divergences scale up to finely sampled distributions. As a rule of thumb, Sinkhorn divergences take 20-50 times as long to compute as a baseline MMD – even though the explicit gradient formula (26-27) lets us win a factor 2-3 compared with a naive autograd implementation. For the sake of this benchmark, we ran the Sinkhorn and symmetric Sinkhorn loops with fixed numbers of iterations: 20 and 3 respectively, which is more than enough for measures on the unit (hyper)cubef if $\varepsilon \geq 0.05$ – here, we work in $\mathbb{R}^3$.

As a rule of thumb, Sinkhorn divergences take 20-50 times as long to compute as a baseline MMD – even though the explicit gradient formula (26-27) lets us win a factor 2-3 compared with a naive autograd implementation. For the sake of this benchmark, we ran the Sinkhorn and symmetric Sinkhorn loops with fixed numbers of iterations: 20 and 3 respectively, which is more than enough for measures on the unit (hyper)cube if $\varepsilon \geq 0.05$ – here, we work in $\mathbb{R}^3$.

In the previous sections, we have provided theoretical guarantees on top of a comprehensive implementation guide for the family of Sinkhorn divergences $S_\varepsilon$. Let us now describe the geometry induced by these new loss functions on the space of probability measures.

Gradient flows. To compare MMD losses $L_\varepsilon$ with Cuturi’s original cost $OT_\varepsilon$ and the de-biased Sinkhorn divergence $S_\varepsilon$, a simple yet relevant experiment is to let a model distribution $\alpha(t)$ flow with time $t$ along the “Wasserstein-2” gradient flow of a loss functional $\alpha \mapsto L(\alpha, \beta)$ that drives it towards a target distribution $\beta$ (Santambrogio, 2015). This corresponds to the “non-parametric” version of the data fitting problem evoked in Section 1, where the parameter $\theta$ is nothing but the vector of positions $\mathbf{x}$ that encodes the support of a measure $\alpha = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}$. Understood as a “model free” idealization of fitting problems in machine learning, this experiment allows us to grasp the typical behavior of the loss function as we discover the deformations of the support that it favors.

Figure 4 – Gradient flows for 1-D measures sampled with $N = M = 5000$ points – we display $\alpha(t)$ (in red) and $\beta$ (in blue) through kernel density estimations on the segment $[0,1]$. The legend on the left indicates the energy distance. Experiments performed on $\mathcal{X} = \mathbb{R}^3$ with a cheap laptop’s GPU (GTX 960M).
Interpolating between Optimal Transport and MMD using Sinkhorn Divergences

\[ \|\alpha - \beta\|_2^2 + \infty \]

\[ S_\varepsilon(\alpha, \beta) \]

\[ \varepsilon = 0.1 \]

\[ \varepsilon = 0.01 \]

\[ OT_\varepsilon(\alpha, \beta) \]

\[ \varepsilon = 0.1 \]

\[ t = 0. \quad t = .25 \quad t = .50 \quad t = 1.00 \quad t = 5.00 \]

Figure 5 – Gradient flows for 2-D measures. The setting is the same as in Figure 4, but the measures \( \alpha(t) \) (in red) and \( \beta \) (in blue) are now directly displayed as point clouds of \( N = M = 500 \) points. The evolution of the support of \( \alpha(t) \) is thus made apparent, and we display as a green vector field the descent direction \( -\nabla_x L(\alpha, \beta) \).

In Figures 4 and 5, \( \beta = \frac{1}{M} \sum_{j=1}^{M} \delta_{y_j} \) is a fixed target measure while \( \alpha = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i(t)} \) is parameterized by a time-varying point cloud \( x(t) = (x_i(t))_{i=1}^{N} \in (\mathbb{R}^D)^N \) in dimension \( D = 1 \) or \( 2 \). Starting from a set initial condition at time \( t = 0 \), we simply integrate the ODE

\[ \dot{x}(t) = -N \nabla_x [L(\frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \beta)](x_i(t)) \]

with a Euler scheme and display the evolution of \( \alpha(t) \) up to time \( t = 5 \).

Interpretation. In both figures, the fourth line highlights the entropic bias that is present in the OT\( _\varepsilon \) loss: \( \alpha(t) \) is driven towards a minimizer that is a “shrunk” version of \( \beta \). As showed in Theorem 1, the de-biased loss \( S_\varepsilon \) does not suffer from this issue: just like MMD norms, it can be used as a reliable, positive-definite divergence.

Going further, the dynamics induced by the Sinkhorn divergence interpolates between that of an MMD (\( \varepsilon = +\infty \)) and Optimal Transport (\( \varepsilon = 0 \)), as shown in (4). Here, \( C(x, y) = \|x - y\| \) and we can indeed remark that the second and third lines bridge the gap between the flow of the energy distance \( L_{-\|\cdot\|} \) (in the first line) and that of the Earth Mover’s cost \( OT_0 \) which moves particles according to an optimal transport plan.

Please note that in both experiments, the gradient of the energy distance with respect to the \( x_i \)’s vanishes at the extreme points of \( \alpha \)’s support. Crucially, for small enough values of \( \varepsilon \), \( S_\varepsilon \) recovers the translation-aware geometry of OT and we observe a clean convergence of \( \alpha(t) \) to \( \beta \) as no sample lags behind.

5 Conclusion

Recently introduced in the ML literature, the Sinkhorn divergences were designed to interpolate between MMD and OT. We have now shown that they also come with a bunch of desirable properties: positivity, convexity, metrization of the convergence in law and scalability to large datasets.

To the best of our knowledge, it is the first time that a loss derived from the theory of entropic Optimal Transport is shown to stand on such a firm ground. As the foundations of this theory are progressively being settled, we now hope that researchers will be free to focus on one of the major open problems in the field: the interaction of geometric loss functions with concrete machine learning models.
Bibliography


A Standard results

Before detailing our proofs, we first recall some well-known results regarding the Kullback-Leibler divergence and the SoftMin operator defined in (12).

A.1 The Kullback-Leibler divergence

First properties. For any pair of Radon measures \(\alpha, \beta \in \mathcal{M}^+(\mathcal{X})\) on the compact metric set \((\mathcal{X}, d)\), the Kullback-Leibler divergence is defined through

\[
\text{KL}(\alpha, \beta) \overset{\text{def.}}{=} \begin{cases} 
\langle \alpha, \log \frac{d\alpha}{d\beta} \rangle - 1 + \langle \beta, 1 \rangle & \text{if } \alpha \ll \beta \\
+\infty & \text{otherwise}
\end{cases}
\]

It can be rewritten as an \(f\)-divergence associated to \(\psi : x \in \mathbb{R}_{\geq 0} \mapsto x \log(x) - x + 1 \in \mathbb{R}_{\geq 0}\),

with \(0 \cdot \log(0) = 0\), as

\[
\text{KL}(\alpha, \beta) = \begin{cases} 
\langle \beta, \psi\left(\frac{d\alpha}{d\beta}\right) \rangle & \text{if } \alpha \ll \beta \\
+\infty & \text{otherwise}
\end{cases}
\]

(28)

Since \(\psi\) is a strictly convex function with a unique global minimum at \(\psi(1) = 0\), we thus get that \(\text{KL}(\alpha, \beta) \geq 0\) with equality iff. \(\alpha = \beta\).

Dual formulation. The convex conjugate of \(\psi\) is defined for \(u \in \mathbb{R}\) by

\[
\psi^*(u) \overset{\text{def.}}{=} \sup_{x > 0} (xu - \psi(x)) = e^u - 1,
\]

and we have \(\psi(x) + \psi^*(u) \geq xu\) (29)

for all \((x, u) \in \mathbb{R}_{\geq 0} \times \mathbb{R}\), with equality if \(x > 0\) and \(u = \log(x)\). This allows us to rewrite the Kullback-Leibler divergence as the solution of a dual concave problem:

**Proposition 5** (Dual formulation of KL). Under the assumptions above,

\[
\text{KL}(\alpha, \beta) = \sup_{h \in \mathcal{F}_b(\mathcal{X}, \mathbb{R})} \langle \alpha, h \rangle - \langle \beta, e^h - 1 \rangle
\]

(30)

where \(\mathcal{F}_b(\mathcal{X}, \mathbb{R})\) is the space of bounded measurable functions from \(\mathcal{X}\) to \(\mathbb{R}\).

**Proof.** Lower bound on the sup. If \(\alpha\) is not absolutely continuous with respect to \(\beta\), there exists a Borel set \(A\) such that \(\alpha(A) > 0\) and \(\beta(A) = 0\). Consequently, for \(h = \lambda 1_A\),

\[
\langle \alpha, h \rangle - \langle \beta, e^h - 1 \rangle = \lambda \alpha(A) \xrightarrow{\lambda \to +\infty} +\infty.
\]

Otherwise, if \(\alpha \ll \beta\), we define \(h_* = \log \frac{d\alpha}{d\beta}\) and see that

\[
\langle \alpha, h_* \rangle - \langle \beta, e^{h_*} - 1 \rangle = \text{KL}(\alpha, \beta).
\]

If \(h_n = \log(\frac{d\alpha}{d\beta}) 1_{\lambda \leq \alpha/d\beta \leq n} \in \mathcal{F}_b(\mathcal{X}, \mathbb{R})\), the monotone and dominated convergence theorems then allow us to show that

\[
\langle \alpha, h_n \rangle - \langle \beta, e^{h_n} - 1 \rangle \xrightarrow{n \to +\infty} \text{KL}(\alpha, \beta).
\]

Upper bound on the sup. If \(h \in \mathcal{F}_b(\mathcal{X}, \mathbb{R})\) and \(\alpha \ll \beta\), combining (28) and (29) allow us to show that

\[
\text{KL}(\alpha, \beta) = \langle \alpha, h \rangle + \langle \beta, e^h - 1 \rangle = \langle \beta, \psi\left(\frac{d\alpha}{d\beta}\right) + \psi^*(h) - h \frac{d\alpha}{d\beta} \rangle \geq 0.
\]

The optimal value of \(\langle \alpha, h \rangle - \langle \beta, e^h - 1 \rangle\) is bounded above and below by \(\text{KL}(\alpha, \beta)\): we get (30). \(\square\)

Since \(\langle \alpha, h \rangle - \langle \beta, e^h - 1 \rangle\) is a convex function of \((\alpha, \beta)\), taking the supremum over test functions \(h \in \mathcal{F}_b(\mathcal{X}, \mathbb{R})\) defines a convex divergence:

**Proposition 6.** The KL divergence is a (jointly) convex function on \(\mathcal{M}^+(\mathcal{X}) \times \mathcal{M}^+(\mathcal{X})\).

Going further, the density of continuous functions in the space of bounded measurable functions allows us to restrict the optimization domain:

**Proposition 7.** Under the same assumptions,

\[
\text{KL}(\alpha, \beta) = \sup_{h \in \mathcal{C}(\mathcal{X}, \mathbb{R})} \langle \alpha, h \rangle - \langle \beta, e^h - 1 \rangle
\]

(31)

where \(\mathcal{C}(\mathcal{X}, \mathbb{R})\) is the space of (bounded) continuous functions on the compact set \(\mathcal{X}\).

**Proof.** Let \(h = \sum_{i \in I} h_i 1_{A_i}\) be a simple Borel function on \(\mathcal{X}\), and let us choose some error margin \(\delta > 0\). Since \(\alpha\) and \(\beta\) are Radon measures, for any \(i\) in the finite set of indices \(I\), there exists a compact set \(K_i\) and an open set \(V_i\) such that \(K_i \subset A_i \subset V_i\) and

\[
\sum_{i \in I} \max\{\alpha(V_i\setminus K_i), \beta(V_i\setminus K_i)\} \leq \delta.
\]

Moreover, for any \(i \in I\), there exists a continuous function \(\varphi_i\) such that \(1_{K_i} \leq \varphi_i \leq 1_{V_i}\). The continuous function \(g = \sum_{i \in I} h_i \varphi_i\) is then such that

\[
|\langle \alpha, g - h \rangle| \leq ||h||_{\infty} \delta \quad \text{and} \quad |\langle \beta, e^g - e^h \rangle| \leq ||e^h||_{\infty} \delta
\]

so that

\[
|\langle \alpha, h \rangle - \langle \beta, e^h - 1 \rangle| - |\langle \alpha, g \rangle - \langle \beta, e^g - 1 \rangle| \leq (||h||_{\infty} + ||e^h||_{\infty}) \delta.
\]

As we let our simple function approach any measurable function in \(\mathcal{F}_b(\mathcal{X}, \mathbb{R})\), choosing \(\delta\) arbitrarily small, we then get (31) through (30). \(\square\)
We can then show that the Kullback-Leibler divergence is weakly lower semi-continuous:

**Proposition 8.** If \( \alpha_n \to \alpha \) and \( \beta_n \to \beta \) are weakly converging sequences in \( \mathcal{M}^+(\mathcal{X}) \), we get

\[
\liminf_{n \to +\infty} \text{KL}(\alpha_n, \beta_n) \geq \text{KL}(\alpha, \beta).
\]

**Proof.** According to (31), the KL divergence is defined as a pointwise supremum of weakly continuous applications

\[
\varphi_h : (\alpha, \beta) \mapsto \langle \alpha, h \rangle - \langle \beta, e^h - 1 \rangle,
\]

for \( h \in \mathcal{C}(\mathcal{X}, \mathbb{R}) \). It is thus lower semi-continuous for the convergence in law. \( \square \)

### A.2 SoftMin Operator

**Proposition 9** (The SoftMin interpolates between a minimum and a sum). Under the assumptions of the definition (12), we get that

\[
\min_{x \sim \alpha} \varphi(x) \xrightarrow{\varepsilon \to 0} \min_{x \in \text{Supp}(\alpha)} \varphi(x) \xrightarrow{\varepsilon \to +\infty} \langle \alpha, \varphi \rangle.
\]

If \( \varphi \) and \( \psi \) are two continuous functions in \( \mathcal{C}(\mathcal{X}) \) such that \( \varphi \leq \psi \),

\[
\min_{x \sim \alpha} \varphi(x) \leq \min_{x \sim \alpha} \psi(x). \tag{32}
\]

Finally, if \( K \in \mathbb{R} \) is constant with respect to \( x \), we have that

\[
\min_{x \sim \alpha} [K + \varphi(x)] = K + \min_{x \sim \alpha} \varphi(x). \tag{33}
\]

**Proposition 10** (The SoftMin operator is continuous). Let \( (\alpha_n) \) be a sequence of probability measures converging weakly towards \( \alpha \), and \( (\varphi_n) \) be a sequence of continuous functions that converges uniformly towards \( \varphi \). Then, for \( \varepsilon > 0 \), the SoftMin of the values of \( \varphi_n \) on \( \alpha_n \) converges towards the SoftMin of the values of \( \varphi \) on \( \alpha \), i.e.

\[
(\alpha_n \to \alpha, \varphi_n \xrightarrow{\|\|_{L^\infty}} \varphi) \implies \min_{x \sim \alpha_n} \varphi_n(x) \to \min_{x \sim \alpha} \varphi(x).
\]

### B Proofs

#### B.1 Dual Potentials

We first state some important properties of solutions \((f, g)\) to the dual problem (8). Please note that these results hold under the assumption that \((\mathcal{X}, d)\) is a compact metric space, endowed with a ground cost function \( C : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) that is \( \kappa \)-Lipschitz with respect to both of its input variables.

The existence of an optimal pair \((f, g)\) of potentials that reaches the maximal value of the dual objective is proved using the contractance of the Sinkhorn map \( T \), defined in (11), for the Hilbert projective metric (Franklin and Lorenz, 1989).

While optimal potentials are only defined \((\alpha, \beta)\)-a.e., as highlighted in Proposition 1, they are extended to the whole domain \( \mathcal{X} \) by imposing, similarly to the classical theory of OT (Santambrogio, 2015, Remark 1.13), that they satisfy

\[
f = T(\beta, g) \quad \text{and} \quad g = T(\alpha, f), \tag{34}
\]

with \( T \) defined in (11). We thus assume in the following that this condition holds. The following propositions studies the uniqueness and the smoothness (with respect to the spatial position and with respect to the input measures) of these functions \((f, g)\) defined on the whole space.

**Proposition 11** (Uniqueness of the dual potentials up to an additive constant). Let \((f_0, g_0)\) and \((f_1, g_1)\) be two optimal pairs of dual potentials for a problem \(\text{OT}_\varepsilon(\alpha, \beta)\) that satisfy (34). Then, there exists a constant \( K \in \mathbb{R} \) such that

\[
f_0 = f_1 + K \quad \text{and} \quad g_0 = g_1 - K. \tag{35}
\]

**Proof.** For \( t \in [0, 1] \), let us define \( f_t = f_0 + t(f_1 - f_0), \)

\[
g_t = g_0 + t(g_1 - g_0) \quad \text{and} \quad \varphi(t) = \langle \alpha, f_t \rangle + \langle \beta, g_t \rangle - \varepsilon \langle \alpha \otimes \beta, \exp \left( \frac{1}{\varepsilon} (f_t \otimes g_t - C) \right) - 1 \rangle,
\]

the value of the dual objective between the two optimal pairs. As \( \varphi \) is a concave function bounded above by \( \varphi(0) = \varphi(1) = \text{OT}_\varepsilon(\alpha, \beta) \), it is constant with respect to \( t \). Hence, for all \( t \in [0, 1] \),

\[
0 = \varphi''(t) = -\frac{1}{\varepsilon} \langle \alpha \otimes \beta, \exp(f_t \otimes g_t - C) \rangle - (f_1 - f_0)^2 + (g_1 - g_0)^2.
\]

This is only possible if, \( \alpha \otimes \beta \)-a.e. in \((x, y)\),

\[
(f_1(x) - f_0(x) + g_1(y) - g_0(y))^2 = 0,
\]

i.e. there exists a constant \( K \in \mathbb{R} \) such that

\[
f_1(x) - f_0(x) = +K \quad \text{a.a.e.},
\]

\[
g_1(y) - g_0(y) = -K \quad \beta \text{-a.e.}
\]

As we extend the potentials through (34), the SoftMin operator commutes with the addition of \( K \) (33) and lets our result hold on the whole feature space. \( \square \)

**Proposition 12** (Lipschitz property). The optimal potentials \((f, g)\) of the dual problem (8) are both \( \kappa \)-Lipschitz functions on the feature space \((\mathcal{X}, d)\).
Now, according to Proposition 11, such a limit pair of optimal potentials for OT\(_\varepsilon\) can be computed through a straightforward, continuous expression that does not depend on the value of the optimal dual potentials \((f, g)\) at the anchor point \(x_o\):

\[
\text{OT}\varepsilon(\alpha, \beta) = \langle \alpha, f \rangle + \langle \beta, g \rangle.
\]

Combining this equation with Proposition 13 (that guarantees the uniform convergence of potentials for weakly converging sequences of probability measures) allows us to conclude.

Lower bound. First, let us remark that \((f, g)\) is a suboptimal pair of dual potentials for OT\(_\varepsilon\). Hence,

\[
\text{OT}\varepsilon(\alpha_t, \beta_t) \geq \langle \alpha_t, f \rangle + \langle \beta_t, g \rangle
\]

and thus, since

\[
\text{OT}\varepsilon(\alpha, \beta) = \langle \alpha, f \rangle + \langle \beta, g \rangle
\]

one has

\[
\Delta_t \geq \langle \delta\alpha, f \rangle + \langle \delta\beta, g \rangle
\]

\[
-\varepsilon\langle \delta\alpha \otimes \beta + \alpha \otimes \delta\beta, \exp(\frac{1}{\varepsilon}(f \oplus g - C)) \rangle + o(1)
\]

\[
\geq \langle \delta\alpha, f - \varepsilon \rangle + \langle \delta\beta, g - \varepsilon \rangle + o(1),
\]

since \(g\) and \(f\) satisfy the optimality equations (10).

Upper bound. Conversely, let us denote by \((g_t, f_t)\) the optimal pair of potentials for OT\(_\varepsilon\) satisfying \(g_t(x_o) = 0\) for some arbitrary anchor point \(x_o \in \mathcal{X}\). As \((f_t, g_t)\) are suboptimal potentials for OT\(_\varepsilon\), we get that

\[
\text{OT}\varepsilon(\alpha, \beta) \geq \langle \alpha_t, f_t \rangle + \langle \beta_t, g_t \rangle
\]

and thus, since

\[
\text{OT}\varepsilon(\alpha_t, \beta_t) = \langle \alpha_t, f_t \rangle + \langle \beta_t, g_t \rangle
\]

\[
-\varepsilon\langle \alpha_t \otimes \beta_t, \exp(\frac{1}{\varepsilon}(f_t \oplus g_t - C)) \rangle - 1
\]

and

\[
\Delta_t \leq \langle \delta\alpha, f_t \rangle + \langle \delta\beta, g_t \rangle
\]

\[
-\varepsilon\langle \delta\alpha \otimes \beta + \alpha \otimes \delta\beta, \exp(\frac{1}{\varepsilon}(f_t \oplus g_t - C)) \rangle + o(1)
\]

\[
\leq \langle \delta\alpha, f_t - \varepsilon \rangle + \langle \delta\beta, g_t - \varepsilon \rangle + o(1).
\]

Conclusion. Now, let us remark that as \(t\) goes to 0

\[
\alpha + t\delta\alpha \rightarrow \alpha \quad \text{and} \quad \beta + t\delta\beta \rightarrow \beta.
\]

Thanks to Proposition 13, we thus know that \(f_t\) and \(g_t\) converge uniformly towards \(f\) and \(g\). Combining the lower and upper bound, we get

\[
\Delta_t \xrightarrow{t \to 0} \langle \delta\alpha, f - \varepsilon \rangle + \langle \delta\beta, g - \varepsilon \rangle = \langle \delta\alpha, f \rangle + \langle \delta\beta, g \rangle
\]

since \(\delta\alpha\) and \(\delta\beta\) both have an overall mass that sums up to zero.
B.3 Proof of Proposition 3

The definition of \( \text{OT}_\varepsilon(\alpha, \alpha) \) is that

\[
\text{OT}_\varepsilon(\alpha, \alpha) = \max_{(f, g) \in C(\mathcal{X})^2} \langle \alpha, f + g \rangle - \varepsilon (\alpha \times \alpha, e^{(f \pm g - C)/\varepsilon} - 1).
\]

Reduction of the problem. Thanks to the symmetry of this concave problem with respect to the variables \( f \) and \( g \), we know that there exists a pair \( (f, g = f) \) of optimal potentials on the diagonal, and

\[
\text{OT}_\varepsilon(\alpha, \alpha) = \max_{f \in C(\mathcal{X})} 2\langle \alpha, f \rangle - \varepsilon (\alpha \times \alpha, e^{(f \pm f - C)/\varepsilon} - 1).
\]

Thanks to the density of continuous functions in the set of simple measurable functions, just as in the proof of Proposition 7, we show that this maximization can be done in the full set of measurable functions \( \mathcal{F}_b(\mathcal{X}, \mathbb{R}) \):

\[
\text{OT}_\varepsilon(\alpha, \alpha) = \max_{f \in \mathcal{F}_b(\mathcal{X}, \mathbb{R})} 2\langle \alpha, f \rangle - \varepsilon (\alpha \times \alpha, e^{(f \pm f - C)/\varepsilon} - 1).
\]

where \( \ast \) denotes the smoothing (convolution) operator defined through

\[
[k \ast \mu](x) = \int_{\mathcal{X}} k(x, y) \, d\mu(y)
\]

for \( k \in C(\mathcal{X} \times \mathcal{X}) \) and \( \mu \in \mathcal{M}^+(\mathcal{X}) \).

Optimizing on measures. Through a change of variables

\[
\mu = e^{\langle f/\varepsilon \rangle} \alpha \quad \text{i.e.} \quad f = \varepsilon \log \frac{d\mu}{d\alpha},
\]

keeping in mind that \( \alpha \) is a probability measure, we then get that

\[
\text{OT}_\varepsilon(\alpha, \alpha) = \varepsilon \max_{\mu \in \mathcal{M}^+(\mathcal{X}), \alpha \ll \mu \ll \alpha} 2\langle \alpha, \log \frac{d\mu}{d\alpha} \rangle - \varepsilon (\mu, k \ast \mu) + 1
\]

\[
-\frac{1}{2} \text{OT}_\varepsilon(\alpha, \alpha) = \varepsilon \min_{\mu \in \mathcal{M}^+(\mathcal{X}), \alpha \ll \mu \ll \alpha} \langle \alpha, \log \frac{d\mu}{d\alpha} \rangle + \frac{1}{2} \langle \mu, k \ast \mu \rangle - \frac{1}{2}
\]

where we optimize on positive measures \( \mu \in \mathcal{M}^+(\mathcal{X}) \) such that \( \alpha \ll \mu \) and \( \mu \ll \alpha \).

Expansion of the problem. As \( k_\varepsilon(x, y) = \exp(-C(x, y)/\varepsilon) \) is positive for all \( x \) and \( y \) in \( \mathcal{X} \), we can remove the \( \mu \ll \alpha \) constraint from the optimization problem:

\[
-\frac{1}{2} \text{OT}_\varepsilon(\alpha, \alpha) = \varepsilon \min_{\mu \in \mathcal{M}^+(\mathcal{X}), \alpha \ll \mu} \langle \alpha, \log \frac{d\mu}{d\alpha} \rangle + \frac{1}{2} \langle \mu, k_\varepsilon \ast \mu \rangle - \frac{1}{2}.
\]

Indeed, restricting a positive measure \( \mu \) to the support of \( \alpha \) lowers the right-hand term \( \langle \mu, k_\varepsilon \ast \mu \rangle \) without having any influence on the density of \( \alpha \) with respect to \( \mu \). Finally, let us remark that the \( \alpha \ll \mu \) constraint is already encoded in the \( \log \frac{d\mu}{d\alpha} \) operator, which blows up to infinity if \( \alpha \) has no density with respect to \( \mu \); all in all, we thus have:

\[
F_\varepsilon(\alpha) = -\frac{1}{2} \text{OT}_\varepsilon(\alpha, \alpha) = \varepsilon \min_{\mu \in \mathcal{M}^+(\mathcal{X})} \langle \alpha, \log \frac{d\mu}{d\alpha} \rangle + \frac{1}{2} \langle \mu, k_\varepsilon \ast \mu \rangle - \frac{1}{2},
\]

which is the desired result.

Existence of the optimal measure \( \mu \). In the expression above, the existence of an optimal \( \mu \) is given as a consequence of the well-known fact from OT theory that optimal dual potentials \( f \) and \( g \) exist, so that the dual OT problem (8) is a max and not a mere supremum. Nevertheless, since this property of \( F_\varepsilon \) is key to the metrization of the convergence in law by Sinkhorn divergences, let us endow it with a direct, alternate proof:

**Proposition 14.** For any \( \alpha \in \mathcal{M}^+_1(\mathcal{X}) \), assuming that \( \mathcal{X} \) is compact, there exists a unique \( \mu_\alpha \in \mathcal{M}^+(\mathcal{X}) \) such that

\[
F_\varepsilon(\alpha) = \varepsilon \left[ \langle \alpha, \log \frac{d\mu_\alpha}{d\alpha} \rangle + \frac{1}{2} \langle \mu_\alpha, k_\varepsilon \ast \mu_\alpha \rangle - \frac{1}{2} \right].
\]

Moreover, \( \alpha \ll \mu_\alpha \ll \alpha \).

**Proof.** Notice that for \( (\alpha, \mu) \in \mathcal{M}^+_1(\mathcal{X}) \times \mathcal{M}^+(\mathcal{X}) \),

\[
E_\varepsilon(\alpha, \mu) \overset{\text{def}}{=} \langle \alpha, \log \frac{d\mu}{d\alpha} \rangle + \frac{1}{2} \langle \mu, k_\varepsilon \ast \mu \rangle
\]

\[
= \text{KL}(\alpha, \mu) + \langle \alpha - \mu, 1 \rangle + \frac{1}{2} \| \mu \|_{k_\varepsilon}^2 - \frac{1}{2}.
\]

Since \( C \) is bounded on the compact set \( \mathcal{X} \times \mathcal{X} \) and \( \alpha \) is a probability measure, we can already say that

\[
\frac{1}{2} F_\varepsilon(\alpha) \leq E_\varepsilon(\alpha, \alpha) - \frac{1}{2} = \frac{1}{2} \langle \alpha \times \alpha, e^{-C/\varepsilon} \rangle - \frac{1}{2} < +\infty.
\]

Upper bound on the mass of \( \mu \). Since \( \mathcal{X} \times \mathcal{X} \) is compact and \( k_\varepsilon(x, y) > 0 \), there exists \( \eta > 0 \) such that \( k(x, y) > \eta \) for all \( x \) and \( y \) in \( \mathcal{X} \). We thus get

\[
\| \mu \|_{k_\varepsilon}^2 \geq \langle \mu, 1 \rangle^2 \eta
\]
and show that
\[ E_\varepsilon(\alpha, \mu) \geq \langle \alpha - \mu, 1 \rangle + \frac{1}{2}\|\mu\|_{k_\varepsilon}^2 - \frac{1}{2} \geq \langle \mu, 1 \rangle(\langle \mu, 1 \rangle - 1) - \frac{1}{2}. \]

As we build a minimizing sequence \((\mu_n)\) for \(F_\varepsilon(\alpha)\), we can thus assume that \((\mu_n, 1)\) is uniformly bounded by some constant \(M > 0\).

**Weak continuity.** Crucially, the Banach-Alaoglu theorem asserts that
\[ \{ \mu \in \mathcal{M}^+(\mathcal{X}) \mid \langle \mu, 1 \rangle \leq M \} \]

is weakly compact; we can thus extract a weakly converging subsequence \(\mu_{n_k} \rightharpoonup \mu_\infty\) from the minimizing sequence \((\mu_n)\). Using Proposition 8 and the fact that \(k_\varepsilon\) is continuous on \(\mathcal{X} \times \mathcal{X}\), we show that \(\mu \mapsto E_\varepsilon(\alpha, \mu)\) is a weakly lower semi-continuous function: \(\mu_\infty = \mu_\alpha\) realizes the minimum of \(E_\varepsilon\) and we get our existence result.

**Uniqueness.** We assumed that our kernel \(k_\varepsilon\) is positive universal. The squared norm \(\mu \mapsto \|\mu\|_{k_\varepsilon}^2\) is thus a strictly convex functional and using Proposition 6, we can show that \(\mu \mapsto E_\varepsilon(\alpha, \mu)\) is strictly convex. This ensures that \(\mu_\alpha\) is uniquely defined.

**B.4 Proof of Proposition 4**

Let us take a pair of measures \(\alpha_0 \neq \alpha_1\) in \(\mathcal{M}^+_1(\mathcal{X})\), and \(t \in (0, 1)\); according to Proposition 14, there exists a pair of measures \(\mu_0, \mu_1\) in \(\mathcal{M}^+(\mathcal{X})\) such that
\[
(1 - t)F_\varepsilon(\alpha_0) + tF_\varepsilon(\alpha_1) = \varepsilon((1 - t)E_\varepsilon(\alpha_0, \mu_0) + tE_\varepsilon(\alpha_1, \mu_1)) > \varepsilon E_\varepsilon((1 - t)\alpha_0 + t\alpha_1, (1 - t)\mu_0 + t\mu_1) \geq F_\varepsilon((1 - t)\alpha_0 + t\alpha_1),
\]
which is enough to conclude. To show the strict inequality, let us remark that
\[
(1 - t)E_\varepsilon(\alpha_0, \mu_0) + tE_\varepsilon(\alpha_1, \mu_1) = E_\varepsilon((1 - t)\alpha_0 + t\alpha_1, (1 - t)\mu_0 + t\mu_1)
\]
would imply that \(\mu_0 = \mu_1\), since \(\mu \mapsto \|\mu\|_{k_\varepsilon}^2\) is strictly convex. As \(\alpha \mapsto \text{KL}(\alpha, \beta)\) is strictly convex on the set of measures \(\alpha\) that are absolutely continuous with respect to \(\beta\), we would then have \(\alpha_0 = \alpha_1\) and a contradiction with our first hypothesis.

**B.5 Proof of the Metrization of the Convergence in Law**

The regularized OT cost is weakly continuous, and the uniform convergence for dual potentials ensures that \(H_\varepsilon\) and \(S_\varepsilon\) are both continuous too. Paired with (6), this property guarantees the convergence towards 0 of the Hausdorff and Sinkhorn divergences, as soon as \(\alpha_n \to \alpha\).

Conversely, let us assume that \(S_\varepsilon(\alpha_n, \alpha) \to 0\) (resp. \(H_\varepsilon(\alpha_n, \alpha)\)). Any weak limit \(\alpha_{n_\infty}\) of a subsequence \((\alpha_{n_k})_k\) is equal to \(\alpha\); since our divergence is weakly continuous, we have \(S_\varepsilon(\alpha_{n_\infty}, \alpha) = 0\) (resp. \(H_\varepsilon(\alpha_{n_\infty}, \alpha)\)), and positive definiteness holds through (6).

In the meantime, since \(\mathcal{X}\) is compact, the set of probability Radon measures \(\mathcal{M}^+_1(\mathcal{X})\) is sequentially compact for the weak* topology. \(\alpha_n\) is thus a compact sequence with a unique adherence value: it converges, towards \(\alpha\).