Geometric Data Analysis Project Report

Interpolating between Optimal Transport and MMD using Sinkhorn Divergences

Eugène Berta
Télécom Paris, IP Paris
MVA, ENS Paris-Saclay
eugene.berta@gmail.com

Yannis Cattan
Mines Paris, Université PSL
MVA, ENS Paris-Saclay
cattan.yannis@gmail.com

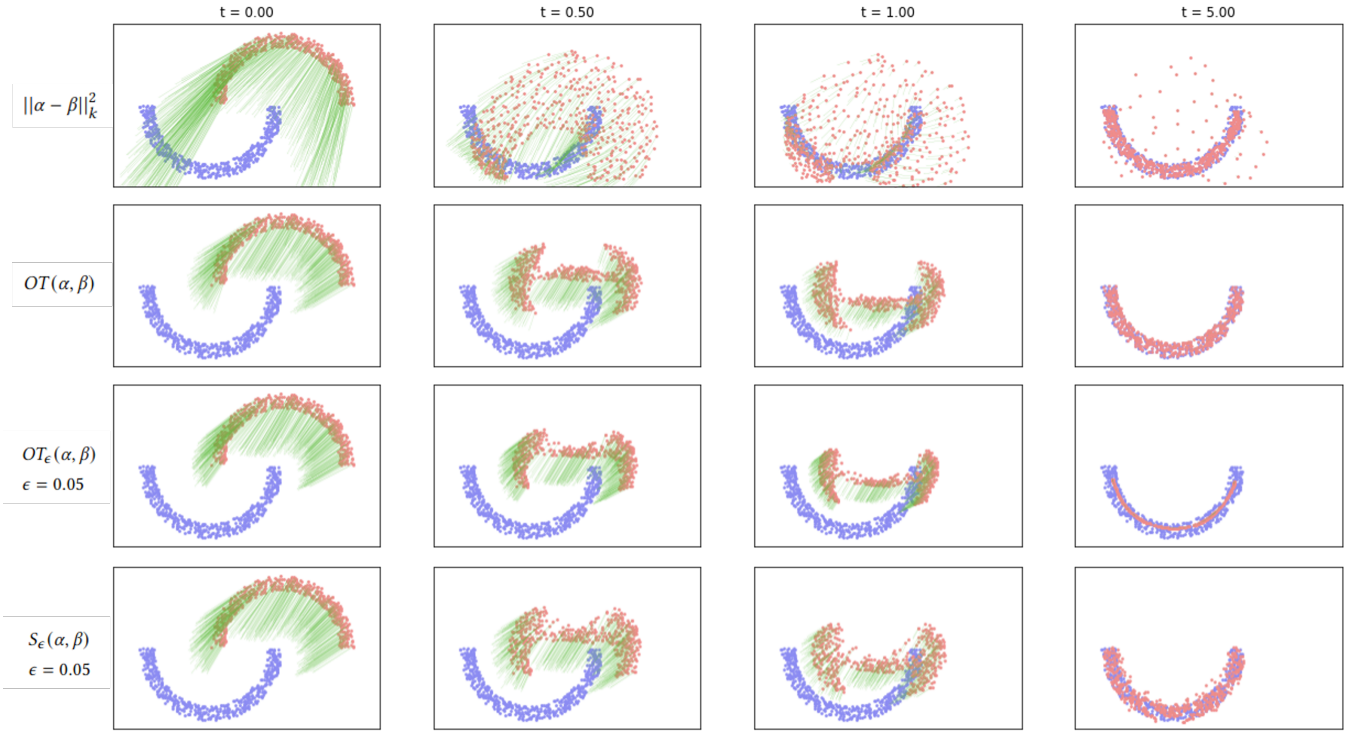

Figure 1: Gradient flows for different geometry-aware divergences.

INTRODUCTION

Our work studies the scientific paper [1] which *(i)* proves new theoretical results regarding the Sinkhorn divergence - a cost used to fit probability distributions - and *(ii)* introduces an adaptation of the existing Sinkhorn algorithm used to compute this divergence and its gradients.

We split our report in 3 parts. First, we will set the scientific context of this paper by describing well-known methods used to compare probability distributions. At a high-level, probability distribution fitting is done by minimizing a loss depending on the current model distribution α_θ and the target distribution β. We would like to stress the fact that each of these losses comes with benefits and drawbacks that we will discuss. In particular, we will explain how the Sinkhorn divergence mitigates the drawbacks of regularized Optimal Transport.

In a second part, we will focus on the main content of the studied paper which is the proof of nice properties for the Sinkhorn divergence loss (positivity, convexity, metrization of the convergence in law) and the introduction of a computational scheme to compute this loss and its gradients.

In a third part, we will discuss the limitations of the paper, make a few personal remarks, and present experiments we conducted to propose a complement on a point we judged disregarded in the paper.

For the sake of simplicity, let us recall a few notations which will be used throughout this report:

- $\mathcal{X}$ denotes a compact feature space. We call "geometry-aware divergences" distances between probability distributions that leverage a distance on this feature space.
- $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$ denote respectively the model and the target probability distributions (i.e. we want to fit α on β) where $\mathcal{M}_1^+(\mathcal{X})$ is the set of positive, unit-mass probability measures on the feature space $\mathcal{X}$.
- $\pi \in \mathcal{M}_1^+(\mathcal{X}^2)$ denotes a coupling measure between α and β. It can be interpreted as a continuous transport mapping from the support of α to the support of β.

1 CONTEXT

1.1 How to compare probability distributions ?

We will first present different distances used in the literature and their respective limitations that encouraged the introduction of the Sinkhorn divergence.

Total Variation.

$$TV(\alpha,\beta) \overset{\text{def.}}{=} 2 \sup_{A\in\mathcal{P}(\mathcal{X})} |\alpha(A)-\beta(A)| \tag{1}$$

The Total Variation (TV) maxes out at 2 when the two probability distributions have disjoint supports. It is a non-geometric divergence in the sense that it pays no attention to the distance on the original feature space $\mathcal{X}$. This is well illustrated by the fact that two disjoint distributions α and β will verify $TV(\alpha,\beta)=2$ whatever the distance that separate their support.

Relative entropy.

$$KL(\beta,\alpha) \overset{\text{def.}}{=} \int_{\mathcal{X}} \log\left(\frac{d\beta}{d\alpha}\right) d\beta \tag{2}$$

The Relative entropy, also called Kullback-Leibler divergence (KL) has a natural connection to the Maximum Likelihood Estimator in statistics and has thus become ubiquitous in data science applications. For some applications however, Relative Entropy should not be considered as the best option. Like the Total Variation divergence, it is not faithful to the geometry of the feature space $\mathcal{X}$. Taking a similar example as before, if the support of β is not included in the support of α, $KL(\beta,\alpha)=+\infty$ whatever the distributions.

We illustrate the behavior of different distances on a toy example in Figure 2 in the Appendix. You can see on this figure that TV and KL have a "constant regime" even though the distributions are getting closer. This illustrates the fact that they are not geometry faithful.

Maximum Mean Discrepancies. Maximum Mean Discrepancies (MMD) divergences is a first class of geometry-aware divergences. Since it overcomes the limitations of the previous divergences, it can be used in applications where the geometry of the feature space matters.

$$L_k(\alpha,\beta) \overset{\text{def.}}{=} \frac{1}{2}||\xi||_k^2 = \frac{1}{2}\int_{\mathcal{X}^2} k(x,y)d\xi(x)d\xi(y) \tag{3}$$

where $k(x,y)$ is a positive definite kernel on $\mathcal{X}$ and $\xi=\alpha-\beta$.

MMD divergences are simple to compute and faithful to the geometry of the feature space (the kernel usually depends on the distance $||x-y||$). This is most importantly illustrated by the fact that, unlike Relative entropy and Total Variation, they **metrize the convergence in law**, i.e. for a given loss L:

$$\alpha_n \rightharpoonup \alpha \iff L(\alpha_n,\alpha)\to 0 \tag{4}$$

Where $\rightharpoonup$ is the weak convergence of measures i.e.:

$$\mathbb{E}_{\alpha_n}[f]\to\mathbb{E}_{\alpha}[f]$$

for any bounded, continuous function f.

This is of paramount interest for geometric applications and this is why MMD distances have been widely used to fit distributions on latent space of generative models. However, MMD behave very differently with the choice of kernel and they come with theoretical shortcomings that induce vanishing gradients on the boundaries of the distributions support. This issue is visible on the gradient flow plot (Figure 1, line 1). We observe that some points never reach the target distribution as the gradients are too small to push them towards the right direction.

Optimal Transport. Another class of geometric divergences that is gaining traction in data science is Optimal Transport (OT). OT directly lifts a metric on the feature space $\mathcal{X}$ to build a metric on the probability measure space $\mathcal{M}_1^+(\mathcal{X})$:

$$OT(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi_1=\alpha,\pi_2=\beta} \int_{\mathcal{X}^2} C d\pi \tag{5}$$

where $C(x,y)$ is a symmetric positive cost function (in practice we often use $||x-y||^2$ which induces the Wasserstein distance on the probability measures space).

OT is geometrical by construction and like MMD divergences, it metrizes the convergence in law. Moreover, OT comes with very strong theoretical guarantees and the gradient flow plot (Figure 1, line 2) shows that it overcomes the apparent limitations of MMD divergences. However, solving the OT problem in the general case is very expensive computationally. In practice, this is the main limitation to a wider use of OT.

Regularized Optimal Transport. To overcome the computational burden of Optimal Transport, a regularized version of the original cost has been introduced together with the GPU compatible *Sinkhorn algorithm* [2] to solve the problem at scale:

$$OT_\epsilon(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi_1=\alpha,\pi_2=\beta} \int_{\mathcal{X}^2} C d\pi + \epsilon KL(\pi,\alpha\otimes\beta) \tag{6}$$

Regularized OT lifted the computational burden of OT and thus helped democratize its applications. However, it is clear from the equation above that the optimal solution is different from the solution of the original problem. The regularization term introduces an **entropic bias** in the optimal transport problem, as we can see on the gradient flow (Figure 1, line 3). Unlike the classical OT solution, the regularized OT minimizer does not cover the full support of the target distribution. This shows that OT_ϵ loses in part the geometric advantages of OT.

1.2 The Sinkhorn divergence

To circumvent the drawback of OT_ϵ - that is the entropic bias - a new cost called *Sinkhorn divergence* has been introduced in the literature:

$$S_\epsilon(\alpha,\beta) \overset{\text{def.}}{=} OT_\epsilon(\alpha,\beta) - \frac{1}{2}OT_\epsilon(\alpha,\alpha) - \frac{1}{2}OT_\epsilon(\beta,\beta) \tag{7}$$

This loss can be computed using the Sinkhorn algorithm. It has been shown to give convincing results empirically (Figure 1, line 4). However, before the studied paper, it was only assumed that S_ϵ would define a positive definite loss function. The core contribution of the studied paper is to prove that the Sinkhorn divergence is indeed a positive-definite and strictly convex loss function that metrizes the convergence in law (see (4)).

Additionally the Sinkhorn divergence interpolates between OT and MMD with an additional degree of freedom ϵ [3]:

$$OT_0(\alpha,\beta) \xleftarrow{0 \leftarrow \epsilon} S_\epsilon(\alpha,\beta) \xrightarrow{\epsilon \to +\infty} \frac{1}{2}||\alpha-\beta||^2_{-C} \tag{8}$$

Hence the title of the studied paper.

2 MAIN CONTENT

This section focuses on the main contributions of the paper : the proof of the theoretical guarantees for the Sinkhorn divergence and the adaptation of the computational scheme to compute the loss and its gradient.

2.1 Main theorem and explanation of its proof

Theorem 1. Let $\mathcal{X}$ be a compact metric space with a Lipschitz symmetric cost function C(x,y) that induces, for $\epsilon > 0$, a positive universal kernel $k_\epsilon(x,y) \overset{\text{def.}}{=} e^{-C(x,y)/\epsilon}$. Then, S_ϵ is a positive-definite and strictly convex loss function which metrizes the convergence in law. For all probability Radon measures α and $\beta \in \mathcal{M}_1^+(\mathcal{X})$ we have:

$$0 = S_\epsilon(\beta,\beta) \leq S_\epsilon(\alpha,\beta) \tag{9}$$

$$\alpha = \beta \iff S_\epsilon(\alpha,\beta) = 0 \tag{10}$$

$$S_\epsilon(\alpha,\beta) \text{ is convex w.r.t both } \alpha \text{ and } \beta \tag{11}$$

$$\alpha_n \rightharpoonup \alpha \iff S_\epsilon(\alpha_n,\alpha) \to 0 \tag{12}$$

Let us explain the original proof of the paper in our own words, by emphasizing the parts which we found more complicated to understand.

The Sinkhorn negentropy. The proof of the Theorem 1 uses extensively the *Sinkhorn negentropy* introduced in the paper:

$$F_\epsilon(\alpha) \overset{\text{def.}}{=} -\frac{1}{2} OT_\epsilon(\alpha,\alpha) \tag{13}$$

In (6), we defined the regularized OT as the minimum of an optimisation problem on the coupling measure space $\mathcal{M}_1^+(\mathcal{X}^2)$. By using Fenchel-Rockafellar duality theorem, one can rewrite OT_ϵ as the solution of an optimisation problem on the space of the continuous bounded functions $C(\mathcal{X})^2$:

$$\begin{aligned} OT_\epsilon(\alpha,\beta) \overset{\text{def.}}{=} & \max_{(f,g)\in C(\mathcal{X})^2} \langle \alpha, f\rangle + \langle \beta, g\rangle \\ & - \epsilon \langle \alpha\otimes\beta, \exp\big(\frac{1}{\epsilon}(f\oplus g - C)\big) - 1\rangle \end{aligned} \tag{14}$$

Starting from this new definition of the OT_ϵ loss (14) and by using a change of variable, one can rewrite the negentropy as:

$$F_\epsilon(\alpha) = \epsilon \min_{\mu\in\mathcal{M}_1^+(\mathcal{X})} \langle \alpha, \log\frac{d\alpha}{d\mu}\rangle + \frac{1}{2}||\mu||^2_{k_\epsilon} - \frac{1}{2} \tag{15}$$

We refer the reader to the Appendix A.1 should they want more details on the calculus to go from (14) to (15). In this Appendix, we try to detail the calculus a bit more than in the original paper.

Convexity of the negentropy. With this new expression of the Sinkhorn negentropy, we can show that it is a strictly convex functional on $\mathcal{M}_1^+(\mathcal{X})$. This result will help us show that the Sinkhorn divergence is also convex w.r.t both its inputs α and β. To show that F_ϵ is strictly convex, let us define:

$$\begin{aligned} E_\epsilon(\alpha,\mu) &\overset{\text{def.}}{=} \langle \alpha, \log\frac{d\alpha}{d\mu}\rangle + \frac{1}{2}\langle \mu, k_\epsilon \circledast \mu\rangle - \frac{1}{2} \\ &= KL(\alpha,\mu) + \langle \alpha-\mu, 1\rangle + \frac{1}{2}||\mu||^2_{k_\epsilon} - \frac{1}{2} \end{aligned} \tag{16}$$

We would like to warn the readers that the definition of the above function suffers from a lack of consistency in the original paper, as the term $-\frac{1}{2}$ was forgotten in the definition of $E_\epsilon(\alpha,\beta)$ (see proof of Proposition 14 in the Appendix B.3 of the original paper).

The KL divergence is a jointly convex function on $\mathcal{M}_1^+(\mathcal{X}) \times \mathcal{M}_1^+(\mathcal{X})$. Additionally, assuming that k_ϵ is a positive universal kernel, $\mu \to ||\mu||^2_{k_\epsilon}$ is a stricly convex MMD norm. Altogether, we have that E_ϵ is a jointly convex functional on $\mathcal{M}_1^+(\mathcal{X}) \times \mathcal{M}_1^+(\mathcal{X})$. Given this result, one can show that F_ϵ is a strictly convex functional (see Appendix B.4 of the original paper).

We can now conclude regarding the convexity of the Sinkhorn divergence: by noticing that $\text{OT}_\epsilon(\alpha,\beta)$ can be written as a maximization of linear forms (14), we know it is convex w.r.t α and β. And by using the convexity of F_ϵ, we conclude that S_ϵ is convex w.r.t both inputs α and β as a sum of the functions OT_ϵ and F_ϵ (recalling (13)).

Positive definiteness of S_ϵ. The proof of the positive definiteness of the Sinkhorn divergence involves the symmetric Bregman divergence induced by the strictly convex functional F_ϵ. It is therefore a positive definite quantity:

$$H_\epsilon(\alpha,\beta) \overset{\text{def.}}{=} \frac{1}{2}\langle \alpha-\beta, \nabla F_\epsilon(\alpha) - \nabla F_\epsilon(\beta)\rangle \geq 0 \tag{17}$$

We refer the reader to the Proposition 2 of the orignal paper regarding the differentiability of F_ϵ. Furthermore, convexity of S_ϵ implies:

$$OT_\epsilon(\alpha,\alpha) + \langle \beta-\alpha, \nabla_2 OT_\epsilon(\alpha,\alpha)\rangle \leq OT_\epsilon(\alpha,\beta) \tag{18}$$

$$OT_\epsilon(\beta,\beta) + \langle \alpha-\beta, \nabla_1 OT_\epsilon(\beta,\beta)\rangle \leq OT_\epsilon(\alpha,\beta) \tag{19}$$

By definition of F_ϵ and by using the symmetry of $\text{OT}_\epsilon(\alpha,\alpha)$, we have that:

$$\begin{aligned} -\nabla_\alpha F_\epsilon(\alpha) &= \frac{1}{2}\nabla_\alpha OT_\epsilon(\alpha,\alpha) \\ &= \frac{1}{2}\big(\nabla_1 OT_\epsilon(\alpha,\alpha) + \nabla_2 OT_\epsilon(\alpha,\alpha)\big) \\ &= \nabla_2 OT_\epsilon(\alpha,\alpha) \\ &= \nabla_1 OT_\epsilon(\alpha,\alpha) \end{aligned}$$

Hence, by summing the lines (18) and (19), we show that $H_\epsilon \leq S_\epsilon$ which concludes the positive definiteness of S_ϵ since H_ϵ is positive definite itself.

Metrization of the convergence in law. To show $\implies$ in (12), we first use the weak continuity of the regularized OT cost paired with the uniform convergence for dual-potentials. Using definition (14) of the OT cost, we deduce that S_ϵ is weakly continuous. Using, the definiteness of S_ϵ, we conclude that $S_\epsilon(\alpha_n,\alpha)$ converges towards 0 when $\alpha_n \rightharpoonup \alpha$.
To show $\impliedby$ in (12), let us assume that $S_\epsilon(\alpha_n,\alpha) \to 0$. Using the

Prohorov theorem, we know that any subsequence $(\alpha_{n_k})_k$ converges towards a limit α_{n_∞}. And by using the weakly continuity of S_ϵ, we have that $S_\epsilon(\alpha_{n_\infty}, \alpha) = 0$. The positive definiteness of S_ϵ gives us that $\alpha_{n_\infty} = \alpha$. Then, since $\mathcal{X}$ is compact, the set of probability Radon measures $\mathcal{M}_1^+(\mathcal{X})$ is sequentially compact for the weak* topology. α_n is thus a compact sequence with a unique adherence value: it converges weakly towards α.

2.2 Adaptation of the computational scheme

The Sinkhorn divergence can naively be computed by separately using the Sinkhorn algorithm on $OT_\epsilon(\alpha, \beta)$, $OT_\epsilon(\alpha, \alpha)$ and $OT_\epsilon(\beta, \beta)$ (see 7). The studied paper proposes an adaptation of the algorithm enabling a faster computation of the loss by leveraging the symmetry of $p \to OT_\epsilon(p, p)$.

We refer the reader to the section 3 of the original paper to get the details of the adaptation of the existing Sinkhorn algorithm. We made this choice because *(i)* we think this contribution is slightly less important than the theoretical guarantees provided by Theorem 1 and mostly because *(ii)* this adaptation is easier to understand than the proof of Theorem 1.

3 LIMITATIONS AND EXTENSIONS

In this section, we first discuss the tightness of the hypothesis of the main theorem. We then give our comments on the paper before discussing the entropic bias in regularized OT more extensively.

3.1 Limitations of the current scope of hypothesis

The first hypothesis we can discuss concerns the computational scheme. To extend the Sinkhorn algorithm to the Sinkhorn divergence, the paper assumes that α and β are discrete distributions. Of course, this can be generalized to semi-discrete or continuous distance computation assuming we can discretize the distributions involved. However, this introduces a discretization error in the computation of the distance that is detrimental to the final result. The computational scheme proposed could be extended to semi-discrete and continuous distances as first proposed in [4] for OT and OT_ϵ .

A second remark is that the paper is written assuming that α and β are unit-mass distributions. Under this assumption, the proof and the proposed computational scheme are not valid for unbalanced optimal transport, where there is no constraint on the total mass of α and β. Another paper proposed a method to overcome this limitation [5].

3.2 Personal remarks on the paper

As subjective readers of the paper, we were positively impressed by its scientific contributions. However, we have identified what we consider as a few drawbacks in its structure. We think they are detrimental to making the context and contributions understandable by the largest possible scientific public.

Target public of the paper. As said in the introduction and conclusion of the original paper, we believe that the results exposed are of great importance for the machine learning community, especially for researchers in need of an efficient geometry-aware loss function. However, our opinion is that there is a gap between the large span of scientists that could benefit from this paper, and the (growing but still restrained) community of researchers that are familiar with Optimal Transport, which we believe is a requirement to properly understand this paper. We think the paper is not thorough enough in its introduction to OT and we believe this is detrimental to its impact.

Structure of the introduction. After spending some time trying to understand the scientific context of the paper, we came to disagree with the way the authors introduced the problem and their contribution. Indeed, they chose to introduce the Sinkhorn divergence as an interpolation between OT and MMD to "get the best of both worlds". We find this sentence a bit confusing : it does not seem obvious to us that the interpolation will benefit from the nice geometric properties of the first distance ($OT(\alpha, \beta)$) and the reduced computation cost of the second one ($||\alpha - \beta||_k^2$). As for us, the proper way to introduce Sinkhorn divergence is different: we see Sinkhorn divergence as a way to counterbalance the entropic bias.

For us, the added value of this paper is to guarantee that what was at first an empirical fix of OT_ϵ is actually a well defined mathematical distance. For this reason, we think the introductory part on OT lacks a thorough discussion about the entropic bias in OT_ϵ and the way the regularization term in S_ϵ fixes this bias.

The author's interpretation of Sinkhorn divergence plays a crucial role in the paper as it is what they chose to emphasize in the title. We are questioning the relevance of this choice as we do not consider the interpolation between OT and MMD to be the most important aspect of Sinkhorn divergences in this context.

Of course, this alternative point of view is certainly due to the fact that we are new to the field, thus lacking the historical timeline of the scientific progresses that led to Sinkhorn divergence. However, we believe our ordering adds some clarity to the contribution of the paper.

3.3 Further discussing the entropic bias in regularized Optimal Transport.

As explained in details previously, we regard the Sinkhorn divergence as a way to fix the entropic bias introduced in regularized OT. Moreover, we argue that the paper lacks an in-depth discussion on entropic bias. In this section, we try to illustrate empirically this bias and the way Sinkhorn divergence "fixes" the problem.

Let us start by recalling that, as ϵ tends towards $+\infty$, the term $KL(\pi, \alpha \otimes \beta)$ in (6) gains in importance in the minimization problem. To the limit, this forces $\pi = \alpha \otimes \beta$ and thus :

$$OT_{+\infty} = \int_{\mathcal{X}^2} C(x, y) d\alpha(x) d\beta(y) + 0$$

This term is responsible for the "shrinking" effect that we observe on the gradient flows (Figure 1, line 3), and that the Sinkhorn divergence compensates for. To illustrate this empirically, we conducted two experiments.

Experiment 1. In Figure 3, we plot $OT(\alpha_\theta, \beta)$ and $OT_\epsilon(\alpha_\theta, \beta)$ where α_θ and β are normal distributions, and the parameter θ is modified throughout the experiment to shrink the distribution

α_θ by reducing its variance linearly. We observe that OT_ϵ is minimized later than classical OT in the timeline of the experiment. This illustrates that the OT_ϵ divergence satisfies with a shrunk version of the true minimizer of OT. This is the entropic bias that Sinkhorn divergence aims at solving.

Experiment 2. In Figures 4 and 5 we tried to fit a normal distribution α on a second normal distribution β using regularized optimal transport with a large ϵ (= 0.2) to exaggerate the shrinking effect. We plot the evolution of $\mathrm{OT}(\alpha, \beta)$ and $S_\epsilon(\alpha, \beta)$ throughout the experiment. The results confirm two things.

First, the Sinkhorn divergence mimics the true OT with a lot of precision. It is hard to make the difference between the two on the full loss plot (bottom-left graph of Figure 4) and we see that the auto-correlation term $-\frac{1}{2}OT_\epsilon(\alpha, \alpha) - \frac{1}{2}OT_\epsilon(\beta, \beta)$ compensates for the regularization term $KL(\pi, \alpha \otimes \beta)$ almost perfectly (Figure 5).

Secondly, at the end of the experiment (what we can assume to be the "shrinking" regime) OT and S_ϵ are increasing. This illustrates that while the regularization in OT_ϵ is detrimental to the approximation of OT, the Sinkhorn divergence manages to mimic the true behaviour of OT.

4 CONCLUSION

The studied paper provides *(i)* strong theoretical guarantees for the Sinkhorn divergence - a cost function used to alleviate the entropic bias induced by regularized OT - and *(ii)* an adaptation of the existing Sinkhorn algorithm used to compute this cost and its gradient without loosing too much computational efficiency.

Even though the paper has strict hypotheses, its contribution paves the way for a democratized usage of geometric loss functions in concrete machine learning applications.

Yet, we ask ourselves if this paper would have had a broader impact should it have been more thorough on the introduction of Optimal Transport and a bit more direct when explaining the advantages of the Sinkhorn divergence.

REPRODUCIBILITY

You can reproduce all our figures using the notebook publicly available at : https://github.com/eugeneberta/MVA/blob/main/S1_Geometric_Data_Analysis/Project/reproduce_project_figures.ipynb

ACKNOWLEDGMENTS

We would like to thank the authors of the studied paper for their interesting work.

REFERENCES

[1] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences, 2018.
[2] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. 2013.
[3] Aaditya Ramdas, Nicolas Garcia, and Marco Cuturi. On wasserstein two sample testing and related families of nonparametric tests, 2015.
[4] Genevay Aude, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport, 2016.
[5] Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport, 2019.

A APPENDIX

A.1 Rewriting the Sinkhorn negentropy

Using (14), we have:

$$\begin{aligned} OT_\epsilon(\alpha,\alpha) &\overset{\text{def.}}{=} \max_{(f,g)\in C(\mathcal{X})^2} \langle \alpha, f+g\rangle \\ &\quad - \epsilon\langle \alpha\otimes\alpha, \exp\big(\frac{1}{\epsilon}(f\oplus g - C)\big)-1\rangle \end{aligned} \tag{20}$$

The above problem is a symmetric concave problem with respect to the variables f and g. This means that there exists an optimal solution $(f,g)=f$ and we have:

$$\begin{aligned} OT_\epsilon(\alpha,\alpha) &\overset{\text{def.}}{=} \max_{f\in C(\mathcal{X})} 2\langle \alpha, f\rangle \\ &\quad - \epsilon\langle \alpha\otimes\alpha, \exp\big(\frac{1}{\epsilon}(f\oplus f - C)\big)-1\rangle \end{aligned}$$

Using the density of continuous functions in the set of simple measurable functions, we show that this maximization can be done in the full set of measurable functions $\mathcal{F}_b(\mathcal{X},\mathbb{R})$:

$$\begin{aligned} OT_\epsilon(\alpha,\alpha) &\overset{\text{def.}}{=} \max_{f\in\mathcal{F}_b(\mathcal{X},\mathbb{R})} 2\langle \alpha, f\rangle \\ &\quad - \epsilon\langle \alpha\otimes\alpha, \exp\big(\frac{1}{\epsilon}(f\oplus f - C)\big)-1\rangle \\ &= \max_{f\in\mathcal{F}_b(\mathcal{X},\mathbb{R})} 2\langle \alpha, f\rangle \\ &\quad - \epsilon\langle \exp(f/\epsilon)\alpha, k_\epsilon \circledast \exp(f/\epsilon)\alpha\rangle + \epsilon \end{aligned}$$

where $\circledast$ denotes the convolution operator defined as :

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

for $k\in C(\mathcal{X}\times\mathcal{X})$ and $\mu\in\mathcal{M}_1^+(\mathcal{X})$. Below, we give further details on how this convolution operator appeared:

$$\begin{aligned} A &= -\epsilon\langle \alpha\otimes\alpha, \exp\big(\frac{1}{\epsilon}(f\oplus f - C)\big)-1\rangle \\ &= -\epsilon\left[\int_{\mathcal{X}^2} \exp\left(\frac{f(x)+f(y)-C(x,y)}{\epsilon}\right)d\alpha(x)d\alpha(y) \right. \\ &\quad \left. - \int_{\mathcal{X}^2} d\alpha(x)d\alpha(y)\right] \\ &= -\epsilon\int_{\mathcal{X}^2} \exp\left(\frac{f(x)+f(y)-C(x,y)}{\epsilon}\right)d\alpha(x)d\alpha(y) + \epsilon \\ &= -\epsilon\int_{\mathcal{X}^2} \mathrm{e}^{-C(x,y)/\epsilon}\,\mathrm{e}^{f(y)/\epsilon}\,\mathrm{e}^{f(x)/\epsilon}d\alpha(x)d\alpha(y) + \epsilon \\ &= -\epsilon\int_{\mathcal{X}} \mathrm{e}^{f(y)/\epsilon}\underbrace{\left(\int_{\mathcal{X}} \mathrm{e}^{-C(x,y)/\epsilon}\mathrm{e}^{f(x)/\epsilon}d\alpha(x)\right)}_{(k_\epsilon\circledast\exp(f/\epsilon)\alpha)(y)}d\alpha(y) + \epsilon \\ &= -\epsilon\,\langle \exp(f/\epsilon)\alpha, k_\epsilon\circledast\exp(f/\epsilon)\alpha\rangle + \epsilon \end{aligned}$$

Then, by using the following change of variable:

$$\mu = \exp(f/\epsilon)\,\alpha \quad \text{i.e.} \quad f = \epsilon\,\log(\frac{d\mu}{d\alpha})$$

we get:

$$-\frac{1}{2}OT_\epsilon(\alpha,\alpha) = \epsilon \min_{\mu\in\mathcal{M}^+(\mathcal{X}),\ \alpha<<\mu<<\alpha} \langle \alpha, \log\frac{d\alpha}{d\mu}\rangle + \frac{1}{2}\langle \mu, k_\epsilon\circledast\mu\rangle - \frac{1}{2}$$

By observing the constraints on μ, we realize that we can remove $\mu << \alpha$ since μ will not tend towards $+\infty$ during the minimization as the term $\frac{1}{2}\langle \mu, k_\epsilon\circledast\mu\rangle$ is always positive.
Moreover, we can remove the constraint $\mu >> \alpha$ since this constraint is already "contained" in the term $\log\frac{d\alpha}{d\mu}$. Indeed, this term blows up to infinity if α has no density with respect to μ.
Hence we have:

$$\begin{aligned} -\frac{1}{2}OT_\epsilon(\alpha,\alpha) &= \epsilon \min_{\mu\in\mathcal{M}^+(\mathcal{X})} \langle \alpha, \log\frac{d\alpha}{d\mu}\rangle + \frac{1}{2}\langle \mu, k_\epsilon\circledast\mu\rangle - \frac{1}{2} \\ &= \epsilon \min_{\mu\in\mathcal{M}_1^+(\mathcal{X})} \langle \alpha, \log\frac{d\alpha}{d\mu}\rangle + \frac{1}{2}||\mu||^2_{k_\epsilon} - \frac{1}{2} \end{aligned}$$

A.2 Figures

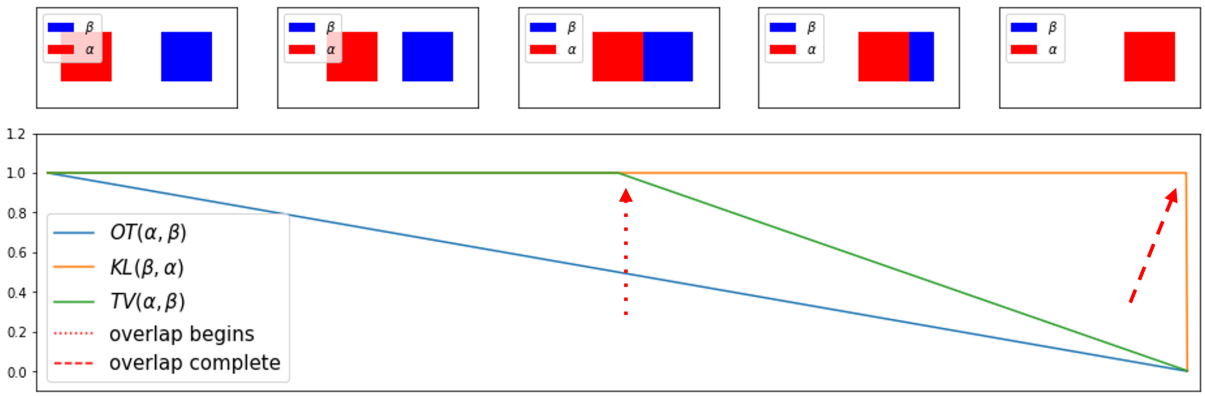

Figure 2: Evolution of different distances during the linear fitting of two uniform squared distributions on $\mathbb{R}^2$. For comparison purposes, the losses are all normalized so that $L(\alpha, \beta)$ is set to 1 at initial position.

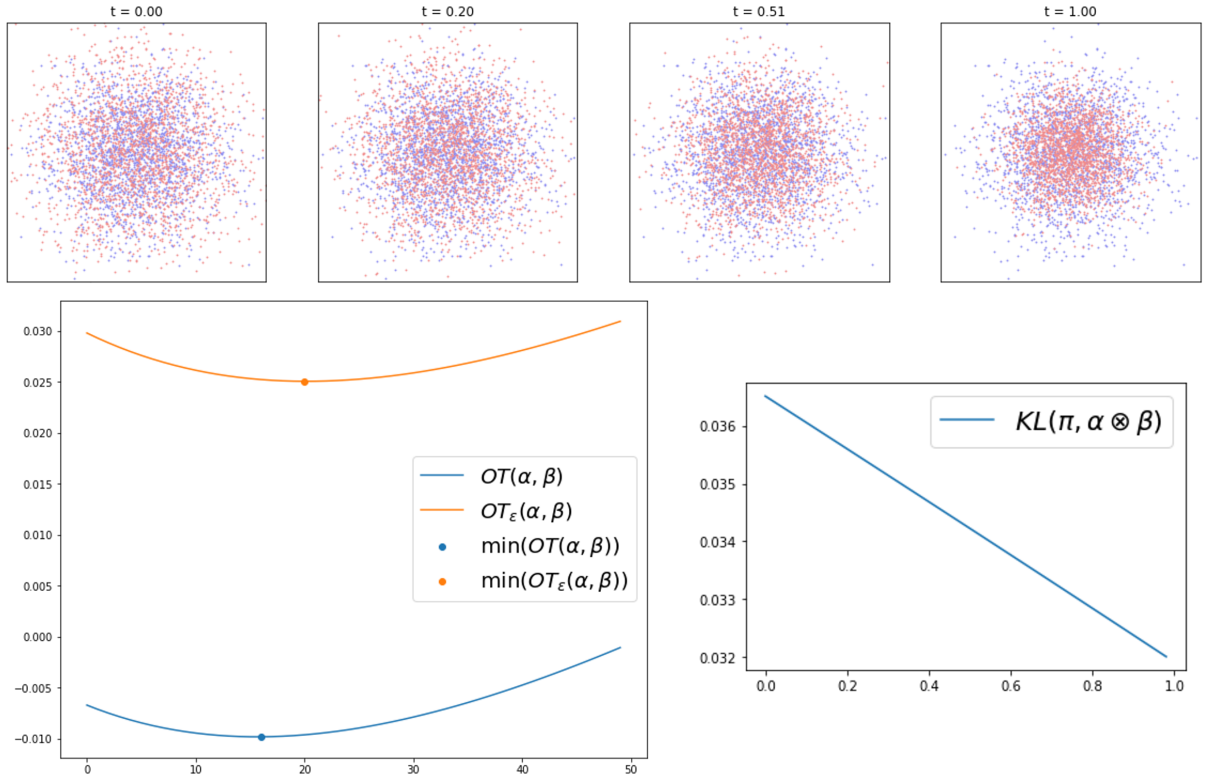

Figure 3: *experiment 1* - Plotting the evolution of OT and OT_ϵ between centered normal distributions α and β when α is shrunk towards zero. Experiment captions on top (α is in red, β is in blue), minimizers on the left, regularization term on the right.

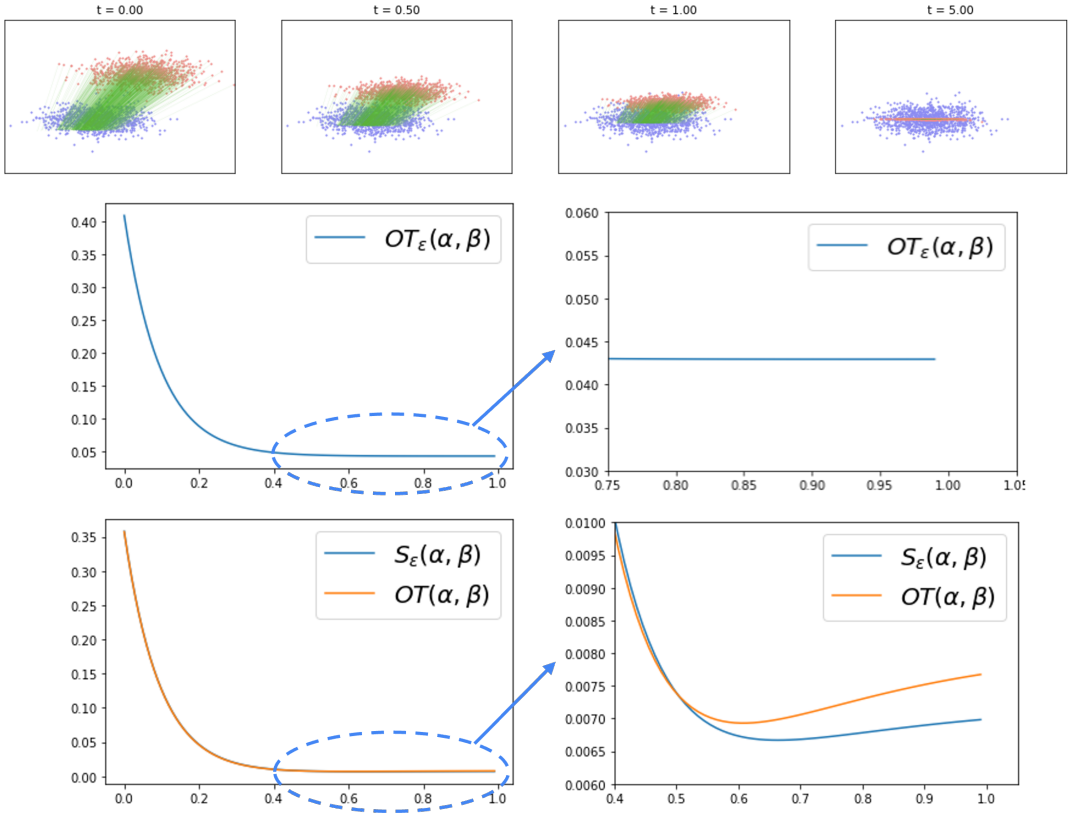

Figure 4: ***experiment 2*** **- Fitting two distributions using OT_ϵ and plotting the evolution of OT and S_ϵ during the experiment. Gradient flows on top, full plots on the left, zooms on the right.**

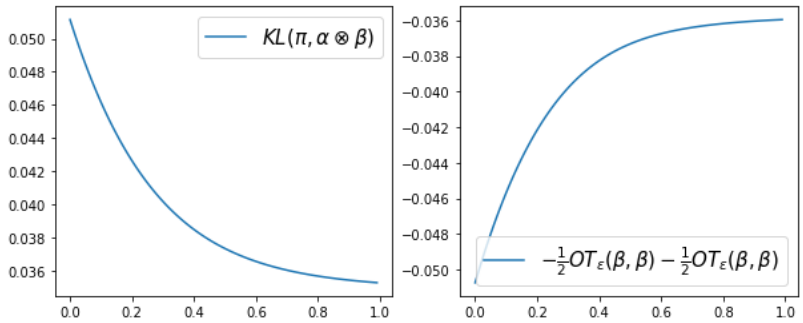

Figure 5: ***experiment 2*** **- Evolution of the two compensating regularization terms $KL(\pi, \alpha \otimes \beta)$ and $-\frac{1}{2}OT_\epsilon(\alpha, \alpha) - \frac{1}{2}OT_\epsilon(\beta, \beta)$ during the experiment.**