Debiased Sinkhorn barycenters

Hicham Janati [1 2] **Marco Cuturi** [3 2] **Alexandre Gramfort** [2]

Abstract

Entropy regularization in optimal transport (OT) has been the driver of many recent interests for Wasserstein metrics and barycenters in machine learning. It allows to keep the appealing geometrical properties of the unregularized Wasserstein distance while having a significantly lower complexity thanks to Sinkhorn's algorithm. However, entropy brings some inherent *smoothing bias*, resulting for example in blurred barycenters. This side effect has prompted an increasing temptation in the community to settle for a slower algorithm such as log-domain stabilized Sinkhorn which breaks the parallel structure that can be leveraged on GPUs, or even go back to unregularized OT. Here we show how this bias is tightly linked to the reference measure that defines the entropy regularizer and propose debiased Wasserstein barycenters that preserve the best of both worlds: fast Sinkhorn-like iterations without entropy smoothing. Theoretically, we prove that the entropic OT barycenter of univariate Gaussians is a Gaussian and quantify its variance bias. This result is obtained by extending the differentiability and convexity of entropic OT to sub-Gaussian measures with unbounded supports. Empirically, we illustrate the reduced blurring and the computational advantage on various applications.

1. Introduction

Comparing, interpolating or averaging probability distributions is an ubiquitous problem in machine learning. Optimal transport (OT) offers an efficient way to do exactly that while taking into account the geometry of the space they live in (Peyré & Cuturi, 2018). Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of probability measures on $\mathbb{R}^d$. Given some divergence $F : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ and weights $(w_k)_k$ such that $\sum_{k=1}^K w_k = 1$, the weighted barycenter of a set of probability measures $(\alpha_k)_k$ can be defined as the Fréchet mean:

$$\alpha_F \stackrel{\text{def}}{=} \arg\min_{\alpha \in \mathcal{P}(\mathbb{R}^d)} \sum_{k=1}^K w_k F(\alpha_k, \alpha) \ . \qquad (1)$$

Here α_F can be thought as a weighted average of distributions. While the $(\alpha_k)_k$ may have a fixed support or known finite supports when working in machine learning applications, the support of α_F may or may not be known. When the latter is unknown a priori, *free support methods* are needed to jointly minimize the objective with respect to both the support and the mass of the distribution (Cuturi & Doucet, 2014). Otherwise, *fixed support methods*, which only optimize weights on known supports, are employed (Benamou et al., 2014). While free support methods are more general and memory efficient, fixed support ones are faster in practice. In this paper, we focus on fixed support methods.

Using the Wasserstein distance as a divergence F, Li & Wang (2006) were the first to propose the Fréchet mean (1) for a clustering application in computer vision. This idea was later adopted by Agueh & Carlier (2011) to formally define Optimal Transport (OT) barycenters. However, the Wasserstein distance is defined through a linear programming problem which does not scale to large datasets. To address this computational issue, some form of regularization is mandatory: either regularize the measures themselves using sliced projections for instances or regularize the OT problem using ℓ_2 (Blondel et al., 2018) or entropy (Cuturi, 2013). While ℓ_2 preserves some of the sparsity of the non-regularized optimal transportation plan, entropy regularization leads to an approximation of the Wasserstein distance that can be solved using a fast and parallelizable GPU-friendly algorithm: the celebrated Sinkhorn's algorithm (Cuturi, 2013). In the rest of this paper, we will focus on entropic OT. Let C be a non-negative cost function on $\mathbb{R}^d \times \mathbb{R}^d$ such that $\mathrm{C}(x, y) = 0 \Leftrightarrow x = y$. For instance, a usual choice is $\mathrm{C}(x, y) = \|x - y\|^2$. Entropy regularized OT between $\alpha, \beta \in \mathcal{P}(\mathbb{R}^d)$ with the reference measures $m_1, m_2 \in \mathcal{P}(\mathbb{R}^d)$ is defined as:

$$\mathrm{OT}_\varepsilon^{m_1, m_2}(\alpha, \beta) \stackrel{\text{def}}{=}$$
$$\min_{\substack{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \\ \pi_{\#1} = \alpha, \pi_{\#2} = \beta}} \int_{\mathbb{R}^{d \times d}} \mathrm{C} \, \mathrm{d}\pi + \varepsilon \, \mathrm{KL}(\pi | m_1 \otimes m_2) \ , \qquad (2)$$

[1]Inria Saclay, France [2]CREST-ENSAE, France [3]Google Research, Brain team, France. Correspondence to: Hicham Janati <hicham.janati@inria.fr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

where $\varepsilon > 0$, $\pi_{\#1}, \pi_{\#2}$ denote the left and right marginals of π respectively, $m_1 \otimes m_2$ is the product measure of m_1 and m_2, and the relative entropy is defined as:

$$\mathrm{KL}(\pi | m_1 \otimes m_2) \stackrel{\text{def}}{=} \int_{\mathbb{R}^d \times \mathbb{R}^d} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}(m_1 \otimes m_2)}\right) \mathrm{d}\pi \ . \tag{3}$$

Naturally, in the discrete case, Benamou et al. (2014) proposed to compute OT barycenters of discrete measures using $F = \mathrm{OT}_\varepsilon^{m_1,m_2}$ with $m_1 = m_2 = \mathcal{U}$, the uniform measure over the finite set on which the measures are defined. Doing so, they showed that the barycenter problem is equivalent to Iterative Bregman Projections (IBP) which are similar to Sinkhorn's scaling operations. However, entropy regularization leads to an undesirable blurring of the barycenter. While using a very small regularization may appear as an obvious solution, it leads to numerical instabilities that can only be mitigated using log-domain stabilization or full log-domain 'logsumexp' operations (Schmitzer, 2016). This however considerably slows down Sinkhorn's iterations.

To reduce this entropy bias, several divergences F have been proposed. For instance, Solomon et al. (2015) proposed to modify the IBP algorithm by adding a maximum entropy constraint they called *entropy sharpening*. This leads to a non-convex constraint which does not fit within the IBP framework. Luise et al. (2018) proposed to compute the entropy regularized solution $\pi^\star$ and to evaluate the OT loss (2) without the entropy term KL. This indeed leads to sharper barycenters but can only be estimated via gradient descent, thus requiring a full Sinkhorn loop at each iteration and setting a pre-defined learning rate which can be cumbersome in practice. Amari et al. (2019) proposed a modified entropy regularized divergence OT that can still leverage the fast IBP algorithm of Benamou et al. (2014) but requires a final deconvolution step with the kernel $\exp(-\frac{\mathrm{C}}{\varepsilon})$, which is only feasible when ε is small. With this same objective of non-blurred solutions, Ge et al. (2019) even called for a return to the original non-regularized Wasserstein barycenter and proposed an accelerated interior point methods algorithm.

Our main contributions Except (Ge et al., 2019), all the works proposed above employ the uniform measure as reference, i.e they use $\mathrm{OT}_\varepsilon^{\mathcal{U}} \stackrel{\text{def}}{=} \mathrm{OT}_\varepsilon^{m_1,m_2}$ with $m_1 = m_2 = \mathcal{U}$. The purpose of this paper is to highlight a direct link between the already known entropy bias of the OT barycenter and this particular choice of m_1 and m_2. This link is illustrated by showing how the choice of m_1 and m_2 impact the barycenter of univariate Gaussians in $\mathbb{R}^d$. Following (Ramdas et al., 2017; Genevay et al., 2018; Feydy et al., 2018; Luise et al., 2019), we advocate for using the following Sinkhorn divergence which can be defined without specifying m_1 and m_2 for arbitrary measures $\alpha, \beta \in \mathcal{P}(\mathbb{R}^d)$:

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

The choice of the reference measures m_1 and m_2 has led to different formulations of regularized OT. The main contributions of this paper are twofold. (1) theoretical: we quantify the entropy bias of usual reference measures for univariate Gaussians. Precisely, while the Lebesgue measure ($m_1 = m_2 = \mathcal{L}$) induces a blurring bias and the product measure ($m_1 = \alpha, m_2 = \beta$) induces a shrinking bias, S_ε is actually debiased. (2) empirical: we propose a fast iterative algorithm similar to IBP to compute debiased barycenters. Unlike other gradient-based methods, this fixed point algorithm can be efficiently differentiated with respect to the barycentric weights via backpropagation. This allows one to carry out Wasserstein barycentric projections without entropy blurring. This will be illustrated in the experiments.

In the following section we discuss the different choices of m_1 and m_2 and quantify their induced entropy bias upon the barycenters of univariate Gaussians. In Section 3, we show some useful properties of S_ε (differentiability, convexity) when defined on sub-Gaussian measures with unbounded supports in $\mathbb{R}^d$ which are necessary to prove the theorems of section 2. Next, in Section 4 we turn to computational aspects and provide a fast Sinkhorn-like algorithm for debiased barycenters. We conclude with numerical experiments in Section 5.

2. Reference measure and entropy bias

Notation We denote by $\mathbb{1}$ the vector of ones in $\mathbb{R}^n$. On matrices, $\log$, $\exp$ and the division operator are applied element-wise. We use $\odot$ for the element-wise multiplication between matrices or vectors. On vectors and matrices, the same notation denotes the usual scalar products: for $x, y \in \mathbb{R}^n$, $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$; and for matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n,n}$, $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i,j=1}^n \mathbf{A}_{ij} \mathbf{B}_{ij}$.

Uniform reference and IBP Let $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ and consider two discrete measures $\alpha = \sum_{i=1}^n \alpha_i \delta_{x_i}$ and $\beta = \sum_{i=1}^n \beta_i \delta_{x_i}$. One can identify α and β with their weights α_i and β_i where $\alpha^\top \mathbb{1} = \beta^\top \mathbb{1}$. Let $\mathbf{C} \in \mathbb{R}_+^{n \times n}$ be the matrix such that $\mathbf{C}_{ij} = \mathrm{C}(x_i, x_j)$. The definition of $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ in (2) becomes:

$$\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha, \beta) = \min_{\substack{\pi \in \mathbb{R}_+^{n \times n} \\ \pi \mathbb{1} = \alpha, \pi^\top \mathbb{1} = \beta}} \langle \mathbf{C}, \pi \rangle + \varepsilon \, \mathrm{KL}(\pi | \mathbf{U}) \ , \tag{4}$$

where $\mathbf{U}$ is the uniform measure on $\mathcal{X}^2$ given by $\frac{\mathbb{1}\mathbb{1}^\top}{n^2}$. Let $\mathbf{K}$ be the element-wise exponentiated kernel $\exp\left(-\frac{\mathbf{C}}{\varepsilon}\right)$. By adopting the definition $\widetilde{\mathrm{KL}}(\mathbf{A}, \mathbf{B}) = \sum_{i,j}^n \mathbf{A}_{ij} \log\left(\frac{\mathbf{A}_{ij}}{\mathbf{B}_{ij}}\right) + \mathbf{B}_{ij} - \mathbf{A}_{ij}$ for $\mathbf{A}, \mathbf{B} \in \mathbb{R}_+^{n \times n}$, Benamou et al. (2014) noticed that (4) is equivalent to a Kullback-Leibler projection up to

an additive constant:

$$\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha,\beta) = \min_{\substack{\pi\in\mathbb{R}_+^{n\times n}\\ \pi\mathbb{1}=\alpha,\pi^\top\mathbb{1}=\beta}} \varepsilon\widetilde{\mathrm{KL}}(\pi|\mathbf{K}) \tag{5}$$

and proposed the Iterative Bregman Projections (IBP) algorithm to solve the equivalent barycenter problem:

$$\min_{\substack{\pi_1,\dots,\pi_K\\ \pi_k\in\mathcal{C}_k\cap\mathcal{C}'}} \sum_{k=1}^K w_k\widetilde{\mathrm{KL}}(\pi^k|\mathbf{K}) \ , \tag{6}$$

where $\mathcal{C}_k = \{\pi \in \mathbb{R}_+^{n\times n} | \pi\mathbb{1} = \alpha_k\}$ and $\mathcal{C}' = \{\pi \in \mathbb{R}_+^{n\times n} | \exists\alpha \in \Delta_n,\ \pi_k^\top\mathbb{1} = \alpha,\ \forall k = 1\dots K\}$. The IBP algorithm amounts to performing iterative minimization on one constraint set at a time. Each step can be solved in closed form, leading to Sinkhorn-like iterations, see supplementary section D for details on IBP.

Lebesgue reference and smoothing bias As discussed in the introduction, the obtained barycenter $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}$ suffers from entropy blurring. To quantify this blur, we turn to Lebesgue continuous measures and consider the Lebesgue measure as a reference by setting $m_1 = m_2 = \mathcal{L}$. We argue that by considering normalized histograms, the discrete formulation (5) provides an approximation of $\mathrm{OT}_\varepsilon^{\mathcal{L}}$ when the number of histogram bins tends to $+\infty$. Indeed, since $\mathrm{OT}_\varepsilon^{\mathcal{L}}$ is defined on Lebesgue-continuous measures, one can identify α, β and π with their density functions. Moreover, if the density functions are positive, the same KL factorization (5) is possible for $\mathrm{OT}_\varepsilon^{\mathcal{L}}$. The following theorem shows that the weighted barycenter of univariate Gaussians is Gaussian with an increased variance. Figure 1 illustrates this smoothing bias using discrete histograms with a grid of 500 bins.

Theorem 1 (Blurring bias of $\mathrm{OT}_\varepsilon^{\mathcal{L}}$)**.** *Let* $\mathrm{C}(x,y) = (x-y)^2$ $\varepsilon > 0$ *and* $\varepsilon = 2\varepsilon'^2$. *Let* $(w_k)_k$ *be positive weights that sum to 1. Let* $\mathcal{N}$ *denote the Gaussian distribution. Assume* $\alpha_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ *and let* $\bar{\mu} = \sum_k w_k\mu_k$,

then:

(i) $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{L}}} \sim \mathcal{N}(\bar{\mu}, S^2)$ *where* S *is a positive solution of the equation:* $\sum_{k=1} w_k\sqrt{\varepsilon'^4 + 4\sigma_k^2S^2} = -\varepsilon'^2 + 2S^2$.

(ii) In particular, if all σ_k *are equal to some* $\sigma > 0$,

then $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{L}}} \sim \mathcal{N}(\bar{\mu}, \sigma^2 + \varepsilon'^2)$.

PROOF. See section C.3

The product measure and shrinking bias Besides the smoothing bias of the uniform measure, $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ cannot be generalized to a general OT definition for any arbitrary distributions that are non-discrete or non-Lebesgue continuous measures. To go beyond this binary classification

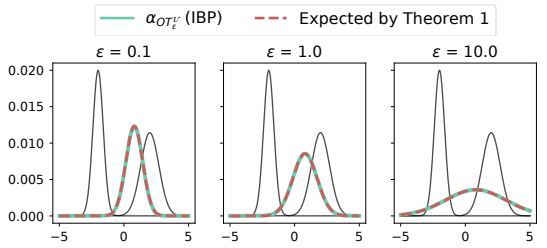

Figure 1. Illustration of theorem 1 with $\mathcal{N}(-2, 0.4)$ and $\mathcal{N}(2, 0.7)$ shown in black, and $(w_1, w_2) = (0.4, 0.6)$. The barycenter $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ matches theoretical expectations and is biased towards blurred distributions.

of probability measures, several authors (Ramdas et al., 2017; Genevay et al., 2018; Feydy et al., 2018) proposed the generic references $m_1 = \alpha$, $m_2 = \beta$. Indeed, the marginal constraints $\pi_1 = \alpha$, $\pi_2 = \beta$ imply that the support of π is included in that of $\alpha \otimes \beta$ and the KL term is always well-defined regardless of the nature of α and β. For the sake of convenience, we denote $\mathrm{OT}_\varepsilon^{\otimes} \overset{\text{def}}{=} \mathrm{OT}_\varepsilon^{\alpha,\beta}$. Di Marino & Gerolin (2019) made the following key observation that characterizes the change of reference. For discrete measures α, β:

$$\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha,\beta) = \mathrm{OT}_\varepsilon^{\otimes}(\alpha,\beta) + \varepsilon\,\mathrm{KL}(\alpha|\mathcal{U}) + \varepsilon\,\mathrm{KL}(\beta|\mathcal{U}) \ . \tag{7}$$

Similarly, the same identity holds for Lebesgue-continuous measures in $\mathcal{P}(\mathbb{R}^d)$:

$$\mathrm{OT}_\varepsilon^{\mathcal{L}}(\alpha,\beta) = \mathrm{OT}_\varepsilon^{\otimes}(\alpha,\beta) + \varepsilon\,\mathrm{KL}(\alpha|\mathcal{L}) + \varepsilon\,\mathrm{KL}(\beta|\mathcal{L}) \ . \tag{8}$$

The identity (7) unveils another merit of $\mathrm{OT}_\varepsilon^{\otimes}$ over $\mathrm{OT}_\varepsilon^{\mathcal{U}}$: its corresponding barycenter problem is equivalent to a regularized $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ barycenter with a negative KL penalty. Interestingly, even though '$-\,\mathrm{KL}$' is concave, $\mathrm{OT}_\varepsilon^{\otimes}$ remains convex with respect to one of its arguments (Feydy et al., 2018). However, $\mathrm{OT}_\varepsilon^{\otimes}$ yet suffers from some limitations: (1) $\mathrm{OT}_\varepsilon^{\otimes}$ cannot be written as a KL projection, thus the fast IBP algorithm is lost; (2) the barycenter $\alpha_{\mathrm{OT}_\varepsilon^{\otimes}}$ of Gaussians can be a degenerate Gaussian, as demonstrated by Theorem 2 which shows that if ε is large, the barycenter collapses to a Dirac (cf. Figure 3). This phenomenon can however be leveraged as a deconvolution technique: Rigollet & Weed (2018) showed that minimizing $\mathrm{OT}_\varepsilon^{\otimes}$ is equivalent to maximum-likelihood deconvolution of an additive Gaussian-noise model.

Theorem 2 (Shrinking bias of $\mathrm{OT}_\varepsilon^{\otimes}$)**.** *Let* $\mathrm{C}(x,y) = (x-y)^2$, $\varepsilon > 0$ *and* $\varepsilon = 2\varepsilon'^2$. *Let* $(w_k)_k$ *be positive weights that sum to 1. Let* $\mathcal{N}$ *denote the Gaussian distribution. Assume that* $\alpha_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ *and let* $\bar{\mu} = \sum_k w_k\mu_k$, $\bar{\sigma}^2 = \sum_{k=1} w_k\sigma_k^2$:

(i) if $\varepsilon'^2 < \bar{\sigma}^2$ *then* $\alpha_{\mathrm{OT}_\varepsilon^{\otimes}} \sim \mathcal{N}(\bar{\mu}, S^2)$ *where* S *is a positive solution of the equation:* $\sum_{k=1} w_k\sqrt{\varepsilon'^4 + 4\sigma_k^2S^2} = \varepsilon'^2 +$

$2S^2$. *In particular, if all* σ_k *are equal to some* $\sigma > 0$, *then* $\alpha_{\mathrm{OT}_\varepsilon^\otimes} \sim \mathcal{N}(\bar{\mu}, \sigma^2 - \varepsilon'^2)$.

(ii) if $\varepsilon'^2 \geq \bar{\sigma}^2$ *then* $\alpha_{\mathrm{OT}_\varepsilon^\otimes}$ *is a Dirac located at* $\bar{\mu}$.

PROOF. See section 3.

Debiased barycenters Interestingly, these limitations and significant differences between $\mathrm{OT}_\varepsilon^{\mathcal{U}}$, $\mathrm{OT}_\varepsilon^{\mathcal{L}}$ and $\mathrm{OT}_\varepsilon^\otimes$ disappear when considering the following Sinkhorn divergences:

$$\mathrm{S}_\varepsilon^m(\alpha,\beta) \stackrel{\text{def}}{=} \mathrm{OT}_\varepsilon^m(\alpha,\beta) - \frac{\mathrm{OT}_\varepsilon^m(\alpha,\alpha) + \mathrm{OT}_\varepsilon^m(\beta,\beta)}{2} \ ,$$

$$\mathrm{S}_\varepsilon(\alpha,\beta) \stackrel{\text{def}}{=} \mathrm{OT}_\varepsilon^\otimes(\alpha,\beta) - \frac{\mathrm{OT}_\varepsilon^\otimes(\alpha,\alpha) + \mathrm{OT}_\varepsilon^\otimes(\beta,\beta)}{2} \ .$$

Using (7) and (8) it holds:

$$\mathrm{S}_\varepsilon(\alpha,\beta) = \mathrm{S}_\varepsilon^m(\alpha,\beta) \ , \tag{9}$$

where m is either $\mathcal{U}$ or $\mathcal{L}$ depending on the nature of α and β. Therefore, S_ε is defined on arbitrary probability measures which can be mixtures of continuous measures and Dirac masses. Moreover, Feydy et al. (2018) showed that when the support of the measures is compact and with the additional assumption that C is negative semi-definite, S_ε is differentiable and convex with respect to one of its arguments. In the following section, we generalize the aforementioned statements for measures with unbounded supports in $\mathbb{R}^d$. The negativity assumption on C holds for instance if $C(x,y) = \|x-y\|^d$ with $0 < d \leq 2$ (Berg et al., 1984, Chapter 3, Cor 3.3) and is the only (cheap) price to pay for a debiased OT divergence. These convexity and differentiability results are essential to prove the debiasing of S_ε stated in Theorem 3 and illustrated in Figure 2.

Theorem 3 (Debiasing of S_ε)**.** *Let* $\mathrm{C}(x,y) = (x-y)^2$ *and* $0 < \varepsilon < +\infty$ *and* $\varepsilon = 2\varepsilon'^2$. *Let* $(w_k)_k$ *be positive weights that sum to 1. Let* $\mathcal{N}$ *denote the Gaussian distribution. Assume that* $\alpha_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ *and let* $\bar{\mu} = \sum_k w_k \mu_k$ *then:*

(i) $\alpha_{\mathrm{S}_\varepsilon} \sim \mathcal{N}(\bar{\mu}, S^2)$ *where* S *is a positive solution* $S^\star$ *of the equation:*

$\sum_{k=1} w_k \sqrt{\varepsilon'^4 + 4\sigma_k^2 S^2} = \sqrt{\varepsilon'^4 + 4S^4}$. *Moreover, given a sorted sequence* $\sigma_{(1)} \leq \cdots \leq \sigma_{(K)}$, *it holds* $S^\star \in (\sigma_{(0)}, \sigma_{(K)})$.

(ii) In particular, if all σ_k *are equal to some* $\sigma > 0$, *then* $\alpha_{\mathrm{S}_\varepsilon} \sim \mathcal{N}(\bar{\mu}, \sigma^2)$.

PROOF. See section 3.

Figure 3 shows a comparison of the three barycenters discussed in this section. We intentionally chose Gaussians with equal variances to emphasize two observations: (1) the debiasing of S_ε: the barycenter $\alpha_{\mathrm{S}_\varepsilon}$ has the same variance of the input measures for all ε; (2) the shrinking bias of $\mathrm{OT}_\varepsilon^\otimes$ is significant even for small values of ε.

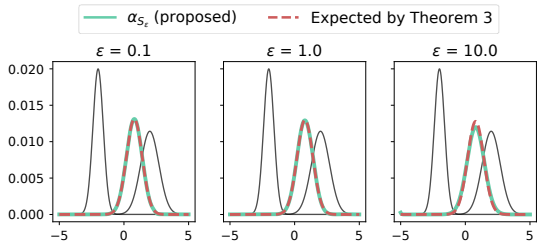

Figure 2. Illustration of theorem 3. Unlike with the uniform measure (Figure 1), the debiased barycenter remains unscathed when increasing ε.

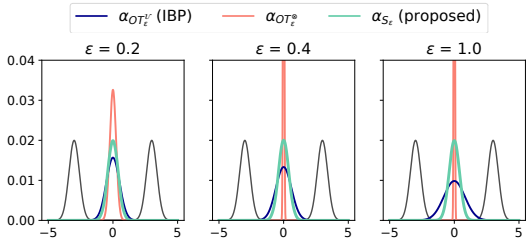

Figure 3. Illustration of the three theorems with $\mathcal{N}(-3, 0.4)$ and $\mathcal{N}(3, 0.4)$ shown in black using uniform weights. Entropy regularization causes a smoothing bias (blue) and a shrinking bias (red). Debiasing with S_ε (cyan) is perfect and independent of ε.

Besides debiasing, the barycenter $\alpha_{\mathrm{S}_\varepsilon}$ also comes with a computational advantage. Using the identity (9), we *bypass* the technical difficulties of the product measure in S_ε and derive an algorithm similar to IBP to compute $\alpha_{\mathrm{S}_\varepsilon}$ which will be the subject of section 4.

3. S_ε is convex and differentiable on sub-Gaussian measures with unbounded supports

Notation The set of continuous function on $\mathbb{R}^d$ is denoted by $\mathcal{C}(\mathbb{R}^d)$. The set of probability measures with a second order moment is denoted by $\mathcal{P}_2(\mathbb{R}^d)$. For $\alpha \in \mathcal{P}(\mathbb{R}^d)$, $\mathcal{L}_p(\mathbb{R}^d, \alpha)$ denotes the set of continuous functions $\mathbb{R}^d \to \mathbb{R}$ such that $\int |f|^p \, \mathrm{d}\alpha < +\infty$. Let $f \in \mathcal{L}_1(\mathbb{R}^d, \alpha)$, $g \in \mathcal{L}_1(\mathbb{R}^d, \beta)$ and denote $\langle \alpha, f \rangle = \int_{\mathbb{R}^d} f \, \mathrm{d}\alpha$. The tensor operators $\otimes$ and $\oplus$ denote respectively the mappings $f \otimes g : (x,y) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto f(x).g(y)$ and $f \oplus g : (x,y) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto f(x) + g(y)$.

To prove theorems 2 and 3, we characterize the optimality condition of the barycenter problem. First, we show that $\mathrm{OT}_\varepsilon^\otimes$ and S_ε are convex (w.r.t. one variable) and differentiable. Our differentiability proof is inspired from that of Feydy et al. (2018) where the compactness assumption of the whole $\mathcal{X}$ is replaced with a sub-Gaussian tails assump-

tion on the measures that allows one to apply Lebesgue's dominated convergence theorem on $\mathbb{R}^d$. The convexity proof is however novel and is solely based on the dual problem of $\mathrm{OT}_\varepsilon^\otimes$. Proving theorem 1 requires studying $\mathrm{OT}_\varepsilon^\mathcal{L}$ which involves a slightly different dual problem. Since the differences are purely technical, we defer the proof of theorem 1 in the appendix and focus in this section on the product measure $\mathrm{OT}_\varepsilon^\otimes$ and S_ε for the sake of clarity.

Dual problem In this section, we set $\mathrm{C}(x,y) = \|x-y\|^2$ with its associated Gaussian kernel $\mathrm{K}(x,y) = e^{-\frac{\|x-y\|^2}{\varepsilon}}$. Let $\alpha, \beta \in \mathcal{P}(\mathbb{R}^d)$. We define the linear operators on $\mathcal{K}$ and $\mathcal{K}^\top$ such that $\mathcal{K}(\mu) = \int_{\mathbb{R}^d} \mathrm{K}(x,y)\,\mathrm{d}\mu(y)$ and $\mathcal{K}^\top(\mu) = \int_{\mathbb{R}^d} K^\top(x,y)\,\mathrm{d}\mu(x)$ for any non-negative measure $\mu \in \mathcal{M}_+(\mathbb{R}^d)$. Problem (2) has a dual formulation given by:

$$\begin{aligned}\mathrm{OT}_\varepsilon^\otimes(\alpha,\beta) = \sup_{\substack{f\in\mathcal{L}_1(\mathbb{R}^d,\alpha)\\ g\in\mathcal{L}_1(\mathbb{R}^d,\beta)}} \int_{\mathbb{R}^d} f\,\mathrm{d}\alpha + \int_{\mathbb{R}^d} g\,\mathrm{d}\beta \\ -\varepsilon\int_{\mathbb{R}^d\times\mathbb{R}^d} \exp\left(\frac{f\oplus g - C}{\varepsilon}\right)\mathrm{d}\alpha\,\mathrm{d}\beta + \varepsilon\ .\end{aligned} \tag{10}$$

if α and β have finite second moments, (10) is well defined and a couple of dual potentials (f,g) are optimal if and only if they are solutions of Sinkhorn's equations (Mena & Weed, 2019):

$$\begin{aligned} e^{\frac{f}{\varepsilon}}.\mathcal{K}(e^{\frac{g}{\varepsilon}}.\beta) = 1,\ \ \alpha-a.e\ , \\ e^{\frac{g}{\varepsilon}}.\mathcal{K}^\top(e^{\frac{f}{\varepsilon}}.\alpha) = 1,\ \ \beta-a.e\ . \end{aligned} \tag{11}$$

and the optimal transport plan π is given by: $\pi = \exp\left(\frac{f\oplus g - C}{\varepsilon}\right).(\alpha\otimes\beta)$

Thus, at optimality the integral over $\mathbb{R}^d\times\mathbb{R}^d$ sums to 1 and:

$$\mathrm{OT}_\varepsilon^\otimes(\alpha,\beta) = \int_{\mathbb{R}^d} f\,\mathrm{d}\alpha + \int_{\mathbb{R}^d} g\,\mathrm{d}\beta \tag{12}$$

Symmetric terms $\mathrm{OT}_\varepsilon^\otimes(\alpha,\alpha)$ When $\alpha=\beta$, the symmetry of the problem leads to the existence of a symmetric pair of potentials (h,h). Indeed, if (f,g) is optimal (g,f) is also optimal. Moreover, since C is symmetric, the optimal transport plan π is also symmetric which leads to $f=g$. Thus the following proposition holds.

Proposition 1. *Let $\alpha\in\mathcal{P}_2(\mathbb{R}^d)$, it holds:*

$$\begin{aligned}\mathrm{OT}_\varepsilon^\otimes(\alpha,\alpha) = \sup_{h\in\mathcal{L}_1(\mathbb{R}^d,\alpha)} 2\int_{\mathbb{R}^d} h\,\mathrm{d}\alpha \\ -\varepsilon\int_{\mathbb{R}^d\times\mathbb{R}^d}\exp\left(\frac{h\oplus h - C}{\varepsilon}\right)\mathrm{d}^2\alpha + \varepsilon\ ,\end{aligned} \tag{13}$$

Moreover, the supremum is attained at the unique (by strong concavity; since C is definite negative) autocorrelation potential $h\in\mathcal{L}_1(\mathbb{R}^d,\alpha)$ if and only if h is a solution of $e^{\frac{fh}{\varepsilon}}.\mathcal{K}(e^{\frac{h}{\varepsilon}}.\alpha) = 1,\ \ \alpha-a.e$, and at optimality it holds: $\frac{1}{2}\mathrm{OT}_\varepsilon^\otimes(\alpha,\alpha) = \int_{\mathbb{R}^d} h\,\mathrm{d}\alpha$.

Restriction on sub-Gaussians To derive theorems 2 and 3, we show that both $\mathrm{OT}_\varepsilon^\otimes$ and S_ε are convex and differentiable and provide a solution of the first order optimality condition. Notice that the convexity of $\mathrm{OT}_\varepsilon^\otimes$ with respect to α and with respect to β follows immediately from (10) since it corresponds to a supremum of linear functionals. Feydy et al. (2018) showed the differentiability of $\mathrm{OT}_\varepsilon^\otimes$ and the convexity of S_ε on measures with compact supports. On $\mathbb{R}^d$, more assumptions on α and β are required. Throughout this section we restrict $\mathrm{OT}_\varepsilon^\otimes$ and S_ε to the convex set of sub-Gaussian probability measures:

Assumption 1. *We set $C(x,y) = \|x-y\|^2$ and restrict $\mathrm{OT}_\varepsilon^\otimes$ and S_ε to the set of sub-Gaussian probability measures $\mathcal{G}(\mathbb{R}^d) \overset{\text{def}}{=} \{\mu | \exists q>0,\ \mathbb{E}_\mu(e^{\frac{\|X\|^2}{2dq^2}}) \le 2\}$.*

Mena & Weed (2019) showed that if $\alpha,\beta\in\mathcal{G}(\mathbb{R}^d)$, there exists a pair of potentials (f,g) verifying the fixed point equations (11) on the whole space $\mathbb{R}^d$ that are bounded by quadratic functions. This result is key to show the differentiability of $\mathrm{OT}_\varepsilon^\otimes$ on $\mathcal{G}(\mathbb{R}^d)$.

Proposition 2 (Mena & Weed (2019), Prop. 6)**.** *Let $\alpha,\beta\in\mathcal{G}(\mathbb{R}^d)$. There exists a pair of smooth functions (f,g) such that (11) holds on $\mathbb{R}^d$ and $\forall x,y\in\mathbb{R}^d$:*

$$\begin{aligned} -dq^2(1+\frac{1}{2}(\|x\|+\sqrt{2dq})^2) \le \frac{f(x)}{\varepsilon} \le \frac{1}{2}(\|x\|+\sqrt{2dq})^2 \\ -dq^2(1+\frac{1}{2}(\|y\|+\sqrt{2dq})^2) \le \frac{g(y)}{\varepsilon} \le \frac{1}{2}(\|y\|+\sqrt{2dq})^2 \end{aligned} \tag{14}$$

Differentiability In the rest of this section, (f,g) denotes a pair of potentials defined by Proposition 2. We say that a function $F:\mathcal{G}(\mathbb{R}^d)\to\mathbb{R}$ is differentiable at α if there exists $\nabla F(\alpha)\in\mathcal{C}(\mathbb{R}^d)$ such that for any displacement $t\delta\alpha$ with $t>0$ and $\delta\alpha=\alpha_1-\alpha_2$ with $\alpha_1,\alpha_2\in\mathcal{G}(\mathbb{R}^d)$, and:

$$F(\alpha+t\delta\alpha) = F(\alpha) + t\langle\delta\alpha,\nabla F(\alpha)\rangle + o(t)\ , \tag{15}$$

where $\langle\delta\alpha,\nabla F(\alpha)\rangle = \int_{\mathbb{R}^d}\nabla F(\alpha)\,\mathrm{d}\delta\alpha$.

Proposition 3. *Let $\alpha,\beta\in\mathcal{G}(\mathbb{R}^d)$, and (f,g) their associated pair of dual potentials given by proposition 2. $\mathrm{OT}_\varepsilon^\otimes(\alpha,.)$ is differentiable on sub-Gaussian measures with unbounded supports and its gradient is given by:*

$$\nabla_\beta\,\mathrm{OT}_\varepsilon^\otimes(\alpha,\beta) = g\ . \tag{16}$$

SKETCH OF PROOF. The proof is inspired from Feydy et al. (2018) in the case of measures with compact supports. The

difference arises when taking the limit of integrals of the potentials. Thanks to assumption 1, proposition 2 provides an upper bound that allows to conclude by dominated convergence. The full proof is provided in the appendix.

The differentiability of S_ε follows immediately:

Corollary 1. *Let $\alpha, \beta \in \mathcal{G}(\mathbb{R}^d)$, and (f, g) their associated pair of dual potentials given by proposition 2 and h_β the autocorrelation potential associated with β. $\mathrm{S}_\varepsilon^\otimes(\alpha, .)$ is differentiable on sub-Gaussian measures with unbounded supports and its gradient is given by:*

$$\nabla_\beta \mathrm{S}_\varepsilon^\otimes(\alpha, \beta) = g - h_\beta \ . \tag{17}$$

Remark 1. It is important to keep in mind that the notion of differentiability (and gradient) of the functions $\mathrm{OT}_\varepsilon^\otimes$ and S_ε differ from the usual Fréchet differentiability. Indeed, the space of probability measures $\mathcal{P}(\mathbb{R}^d)$ has an empty interior in the space of signed Radon measures $\mathcal{M}(\mathbb{R}^d)$. The definition adopted here defines derivatives along feasible directions in $\mathcal{P}(\mathbb{R}^d)$. This is however sufficient to characterize the convexity of S_ε and its stationary points (see appendix A for details).

Convexity Now we turn to showing that S_ε is convex with respect to either one of its arguments separately. To do so, we prove the first order characterization of convexity of a differentiable function $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ given by:

$$F(\alpha) \geq F(\alpha') + \langle \alpha - \alpha', \nabla F(\alpha') \rangle \ , \tag{18}$$

As shown by the proof of the following Lemma, the positivity of K plays a key role in proving the convexity of S_ε.

Lemma 1. *Let $\alpha, \alpha' \in \mathcal{G}(\mathbb{R}^d)$ and let $h_\alpha, h_{\alpha'}$ denote their respective autocorrelation potentials given by proposition 1. Then if $\mathrm{K}(x, y) = e^{-\frac{\|x-y\|^2}{\varepsilon}}$:*

$$\int e^{\frac{h_\alpha(x)}{\varepsilon}} \mathrm{K}(x, y) e^{\frac{h_{\alpha'}(y)}{\varepsilon}} \, \mathrm{d}\alpha(x) \, \mathrm{d}\alpha'(y) \leq 1 \tag{19}$$

Proposition 4. *Under assumption (1), S_ε is convex on sub-Gaussian measures with respect to either of its arguments.*

PROOF. Let $\beta \in \mathcal{G}(\mathbb{R}^d)$. Let $\alpha, \alpha' \in \mathcal{G}(\mathbb{R}^d)$. Let (f, g) and (f', g') denote the pair of potentials associated with $\mathrm{OT}_\varepsilon^\otimes(\alpha, \beta)$ and $\mathrm{OT}_\varepsilon^\otimes(\alpha', \beta)$ respectively and for any $\mu \in \mathcal{G}(\mathbb{R}^d)$, let h_μ denote the autocorrelation potential associated with $\mathrm{OT}_\varepsilon^\otimes(\mu, \mu)$. The first order inequality (18) applied to $F = S_\varepsilon(., \beta)$ is equivalent to:

$$\begin{aligned}
&(18) \Leftrightarrow \langle \alpha, f - h_\alpha \rangle + \langle \beta, g - h_\beta \rangle \geq \\
&\langle \alpha', f' - h_{\alpha'} \rangle + \langle \beta, g' - h_\beta \rangle + \langle \alpha - \alpha', f' - h_{\alpha'} \rangle \\
&\Leftrightarrow \langle \alpha, f - h_\alpha \rangle + \langle \beta, g \rangle \geq \langle \beta, g' \rangle + \langle \alpha, f' - h_{\alpha'} \rangle \\
&\Leftrightarrow \langle \alpha, f \rangle + \langle \beta, g \rangle \geq \langle \beta, g' \rangle + \langle \alpha, f' - h_{\alpha'} + h_\alpha \rangle \\
&\Leftrightarrow \mathrm{OT}_\varepsilon^\otimes(\alpha, \beta) \geq \langle \alpha, f' - h_{\alpha'} + h_\alpha \rangle + \langle \beta, g' \rangle
\end{aligned} \tag{20}$$

To show the last inequality we use the definition of the dual problem (10) and evaluate the dual function at the suboptimal potentials $(f' - h_{\alpha'} + h_\alpha, g')$. Doing so leads to:

$$\begin{aligned}
&\mathrm{OT}_\varepsilon^\otimes(\alpha, \beta) \geq \langle \alpha, f' - h_{\alpha'} + h_\alpha \rangle + \langle \beta, g' \rangle + \varepsilon \\
&- \varepsilon \int_{\mathbb{R}^d \times \mathbb{R}^d} \exp\left(\frac{(f' - h_{\alpha'} + h_\alpha) \oplus g' - C}{\varepsilon} \right) \mathrm{d}\alpha \, \mathrm{d}\beta \ .
\end{aligned}$$

To conclude, all we need to show is that,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \exp\left(\frac{(f' - h_{\alpha'} + h_\alpha) \oplus g' - C}{\varepsilon} \right) \mathrm{d}\alpha \, \mathrm{d}\beta \leq 1 \tag{21}$$

By the Fubini-Tonelli theorem, the order of integration is irrelevant. First integrating with respect to β, we use the optimality conditions (11) on the pair (f', g') then on $h_{\alpha'}$:

$$\begin{aligned}
B &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \exp\left(\frac{(f' - h_{\alpha'} + h_\alpha) \oplus g' - C}{\varepsilon} \right) \mathrm{d}\alpha \, \mathrm{d}\beta \\
&= \int_{\mathbb{R}^d} \exp\left(\frac{h_\alpha - h_{\alpha'}}{\varepsilon} \right) \mathrm{d}\alpha \\
&= \int_{\mathbb{R}^d \times \mathbb{R}^d} \exp\left(\frac{h_\alpha \oplus h_{\alpha'} - C}{\varepsilon} \right) \mathrm{d}\alpha \, \mathrm{d}\alpha'
\end{aligned}$$

Thus, Lemma 1 applies and we have $B \leq 1$. □

Barycenter of sub-Gaussian distributions. We have shown that $\mathrm{OT}_\varepsilon^\otimes$ and S_ε are convex and differentiable, thus the weighted barycenters $\alpha_{\mathrm{OT}_\varepsilon^\otimes}$ and $\alpha_{\mathrm{S}_\varepsilon}$ can be characterized by the first order optimality condition as follows. Let (f_k, g_k) denote the potentials associated with $\mathrm{OT}_\varepsilon^\otimes(\alpha_k, \alpha)$ and h_α the autocorrelation potential associated with $\mathrm{OT}_\varepsilon^\otimes(\alpha, \alpha)$. Using the first order characterization of convexity (18), $\alpha^\star$ is a global minimizer of the barycenter loss of $\mathrm{OT}_\varepsilon^\otimes$ if and only if for any direction $\beta \in \mathcal{G}(\mathbb{R}^d)$, $\langle \sum_{k=1}^K w_k \nabla_{\alpha^\star} \mathrm{OT}_\varepsilon^\otimes(\alpha_k, \alpha^\star), \beta - \alpha^\star \rangle \geq 0$. This is equivalent to $\sum_{k=1}^K w_k \langle g_k, \beta - \alpha^\star \rangle \geq 0$. Similarly, for $\alpha_{\mathrm{S}_\varepsilon}$ we get the optimality condition $\sum_{k=1}^K w_k \langle g_k - h_{\alpha^\star}, \beta - \alpha^\star \rangle \geq 0$. We are now ready to summarize the different steps of the proofs of the theorems. For S_ε, we provide solutions of the optimality conditions by considering quadratic potentials and Gaussian barycenters $\alpha_{\mathrm{S}_\varepsilon}$. We proceed by identification of the coefficients of the polynomials and the parameters of the barycenters and show that the obtained solutions verify the optimality condition. For $\mathrm{OT}_\varepsilon^\otimes$, we proceed similarly for $2\varepsilon'^2 < \sigma^2$. For $2\varepsilon'^2 \geq 2\sigma^2$, we show directly that for the Dirac measure $\alpha^\star = \delta_{\bar{\mu}}$, there exists a set of potentials that verify the optimality condition alongside Sinkhorn's equations. The detailed derivations are provided in the supplementary materials.

4. Fast Sinkhorn-like algorithm

Discrete measures on a finite space The purpose of this section is to derive a fast Sinkhorn-like algorithm to compute $\alpha_{\mathrm{S}_\varepsilon}$ on a fixed support. Let $\mathcal{X} = \{x_1, \dots, x_n\}$ be a finite grid of size n. With images for instance, each x_i would correspond to a pixel. We identify a probability measure $\alpha = \sum_{i=1}^n \alpha_i \delta_{x_i} \in \mathcal{P}(\mathcal{X})$ with its weights vector $(\alpha_i) \in \mathbb{R}^n_{++}$ such that $\sum_{i=1} \alpha_i = 1$. In the rest of this paper, OT_ε and S_ε can be seen as functions operating on the interior of the probability simplex of $\mathbb{R}^n$ denoted by $\Delta_n = \{x \in \mathbb{R}^n_{++} | \sum_{i=1} x_i = 1\}$. We assume that the cost matrix $\mathbf{C} \in \mathbb{R}^{n\times n}_+$ is symmetric negative semi-definite (or equivalently, its associated kernel $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$ is positive semi-definite). This assumption holds for instance if $\mathbf{C}_{ij} = \|x_i - x_j\|^p$ with $p \in]0, 2]$ (see (Berg et al., 1984, 3, Thm 2.2, Cor 3.3) for both claims)

Debiased barycenters To obtain a fast iterative algorithm for the debiased barycenters $\alpha_{\mathrm{S}_\varepsilon}$, we are going to leverage the IBP algorithm through the uniform measure on $\mathcal{X}$ as follows. First, the identity (9) ensures that S_ε is independent of the reference measures. Thus, one can write:

$$\mathrm{S}_\varepsilon(\alpha, \beta) = \mathrm{OT}^{\mathcal{U}}_\varepsilon(\alpha, \beta) - \frac{\mathrm{OT}^{\mathcal{U}}_\varepsilon(\alpha, \alpha) + \mathrm{OT}^{\mathcal{U}}_\varepsilon(\beta, \beta)}{2} \quad .$$

Using (5), one can write $\mathrm{OT}^{\mathcal{U}}_\varepsilon(\alpha, \beta)$ as a KL projection. The remaining autocorrelation terms can be replaced by their dual problems to obtain the following proposition. A detailed derivation is provided in appendix E.

Proposition 5. *Let $\alpha_1, \dots, \alpha_K \in \Delta_n$ and $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$. Let π denote a sequence $\pi_1, \dots, \pi_K$ of transport plans in $\mathbb{R}^{n\times n}_+$ and the constraint sets $\mathcal{H}_1 = \{\pi | \forall k,\ \pi_k \mathbb{1} = \alpha_k\}$, and $\mathcal{H}_2 = \{\pi | \forall k\, \forall k',\ \pi_k^\top \mathbb{1} = \pi_{k'} \mathbb{1}\}$. The barycenter problem $\min_{\alpha\in\Delta_n} \sum_{k=1}^K w_k \mathrm{S}_\varepsilon(\alpha_k, \alpha)$ is equivalent to:*

$$\min_{\substack{\pi\in\mathcal{H}_1\cap\mathcal{H}_2 \\ d\in\mathbb{R}^n_+}} \left[\varepsilon \sum_{k=1}^K w_k \widetilde{\mathrm{KL}}(\pi_k | \mathbf{K}\,\mathrm{diag}(d)) + \frac{\varepsilon}{2} \langle d - \mathbb{1}, \mathbf{K}(d - \mathbb{1}) \rangle \right] \quad . \tag{22}$$

where $\widetilde{\mathrm{KL}}(\mathbf{A}, \mathbf{B}) = \sum_{i,j}^n \mathbf{A}_{ij} \log\left(\frac{\mathbf{A}_{ij}}{\mathbf{B}_{ij}}\right) + \mathbf{B}_{ij} - \mathbf{A}_{ij}$.

Since $\widetilde{\mathrm{KL}}$ is jointly convex and $\mathbf{K}$ is assumed positive-definite, the objective (22) is convex. Minimizing (22) with respect to π leads to the barycenter problem $\alpha_{\mathrm{OT}^{\mathcal{U}}_\varepsilon}$ (6) with the modified kernel $\mathbf{K}\,\mathrm{diag}(d)$. This problem can be solved via the fast IBP algorithm. Minimizing with respect to d leads to the Sinkhorn fixed point equation $d = \frac{\sum_{w_k} \pi_k^\top \mathbb{1}}{\mathbf{K}d}$ for which there exists a converging sequence $d_{n+1} \leftarrow \sqrt{\frac{d_n \odot \sum_{w_k} \pi_k^\top \mathbb{1}}{\mathbf{K}d}}$ $(\star)$ (Knight et al., 2014). Given

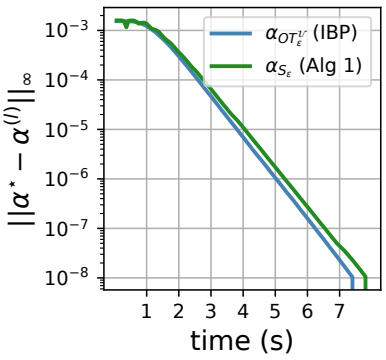

Figure 4. Convergence to the true barycenters of univariate Gaussians $\mathcal{N}(-0.5, 0.1)$ and $\mathcal{N}(0.5, 0.1)$. Algorithm 1 is as fast as IBP with a linear convergence rate.

Algorithm 1 Debiased Sinkhorn Barycenter

Input: $\alpha_1, \dots, \alpha_K$, $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$
Output: $\alpha_{\mathrm{S}_\varepsilon}$
Initialize all scalings (b_k), d to $\mathbb{1}$,
repeat
  for $k = 1$ **to** K **do**
    $a_k \leftarrow \left(\frac{\alpha_k}{\mathbf{K}b_k}\right)$
  end for
  $\alpha \leftarrow d \odot \prod_{k=1}^K (\mathbf{K}^\top a_k)^{w_k}$
  for $k = 1$ **to** K **do**
    $b_k \leftarrow \left(\frac{\alpha}{\mathbf{K}^\top a_k}\right)$
  end for
  $d \leftarrow \sqrt{d \odot \left(\frac{\alpha}{\mathbf{K}d}\right)}$
until convergence

that (22) is smooth and convex, alternate minimization – which amounts to perform IBP and (*) iterations – converges towards its minimum. However, we notice that in practice, either taking one iteration or fully optimizing the subproblems produces the same minimizer. We thus propose to combine one IBP iteration with the update $(\star)$, which leads to Algorithm 1 (see the appendix for further details on the IBP algorithm). Using the theoretical barycenters of Gaussians given by theorems 1 and 3, we can monitor the convergence to the ground truth (Figure 4). Theoretically, both IBP and algorithm 1 have a $\mathcal{O}(Kn^2)$ complexity per iteration. A convergence proof of IBP can obtained using alternating Bregman projections (See (Benamou et al., 2014) and the references theirein). For Algorithm 1 however, similar techniques were not successful. Proving its convergence will be pursued in future work.

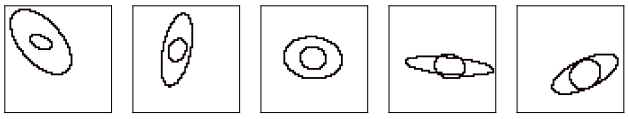

Figure 5. 5 examples of random nested ellipses of size (60×60) used to compute the barycenters of Figure 6.

5. Applications

Now we turn to showing the practical benefits of debiased barycenters in terms of accuracy, speed and performance. Python code can be found at https://github.com/hichamjanati/debiased-ot-barycenters.

Benchmarks In addition to $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}$, $\alpha_{\mathrm{OT}_\varepsilon^{\otimes}}$, we evaluate the performance of the following barycenters:

- α_{A_ε}: Sharp barycenters introduced by Luise et al. (2018), where A_ε is defined as: $A_\varepsilon(\alpha, \beta) = \langle \mathbf{C}, \pi_\varepsilon^\star(\alpha, \beta)\rangle$. Here $\pi_\varepsilon^\star(\alpha, \beta)$ is the primal minimizer of the regularized problem $\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha, \beta)$, computed via accelerated gradient descent.
- $\alpha_{\mathrm{S}_\varepsilon}^F$: Free support barycenters introduced by Luise et al. (2019) that uses the same debiased divergence S_ε, and deals with the free support problem by adding / removing a Dirac particle with Frank-Wolf's algorithm.
- α_W: The original non-regularized Wasserstein problem solved with interior point methods - using the accelerated MAAIPM algorithm of Ge et al. (2019).

Debiased barycenters of ellipses To demonstrate how debiased barycenters $\alpha_{\mathrm{S}_\varepsilon}$ reduce smoothing and are computationally competitive with $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}$, we compare the barycenters of 10 randomly generated nested ellipses displayed in Figure 5. We set the cost matrix $\mathbf{C}$ to the squared Euclidean distance on the unit square and set $\varepsilon = 0.002$. We use the same termination criterion for all methods based on a maximum relative change of the barycenters set to 10^{-5}.

For $\alpha_{\mathrm{S}_\varepsilon}, \alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}, \alpha_{\mathrm{OT}_\varepsilon^{\otimes}}, \alpha_{A_\varepsilon}$, we use the convolution trick introduced by Solomon et al. (2015) which amounts to computing the kernel operation $\mathbf{K}a$ on a vectorized image a by applying a Gaussian convolution on the rows and the columns of a, thereby reducing the complexity of one Debiased / IBP iteration from $O(n^2)$ to $O(n^{\frac{3}{2}})$.

Figure 5 shows that even though α_{A_ε} and α_W are not blurred compared to $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}$, they cannot compete computationally with Sinkhorn-like algorithms. The debiased barycenter is sharp and runs in about the same time as $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}$. Besides, the shrinking bias of $\mathrm{OT}_\varepsilon^{\otimes}$ unfolded by theorem 2 is illustrated in the degeneracy of the ellipse $\alpha_{\mathrm{OT}_\varepsilon^{\otimes}}$.

Barycenters of 3D shapes To visually illustrate the impact of the reduced smoothing bias of S_ε, we computed

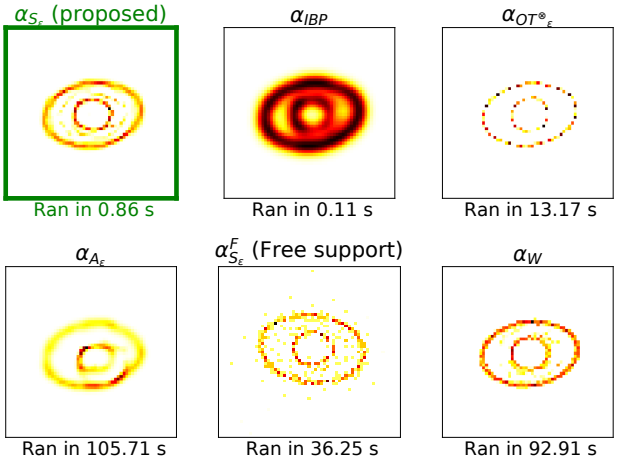

Figure 6. Barycenters of the 10 nested ellipses shown in Figure 5. Results illustrate the reduced blurring of the proposed approach and running times presented below each image demonstrate the computational efficiency. All 6 barycenters were computed on a laptop with an Intel Core i5 3.1 GHz Processor.

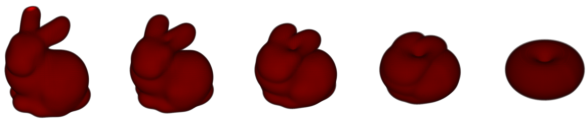

Figure 7. Interpolation of two 3D shapes a $(200)^3$ uniform grid with IBP illustrating a clear blurring bias of $\mathrm{OT}_\varepsilon^{\mathcal{U}}$.

Figure 8. Interpolation of two 3D shapes on a $(200)^3$ uniform grid with the proposed Debiased Sinkhorn (Alg 1). The interpolation is sharper and completes in about the same time as figure 7 (5 seconds on a GPU).

a barycentric interpolation of shapes discretized in a 3D grid of $200 \times 200 \times 200$ voxels. The different interpolations correspond to weights $(w, 1-w)$ where $w \in [0, 0.25, 0.5, 0.75, 1]$. We set the cost matric $\mathbf{C}$ to the squared Euclidean distance on the unit cube and set $\varepsilon = 0.01$. Results presented in Figures 7 and 8 using $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ and S_ε qualitatively demonstrate that S_ε leads to sharper edges, while in both cases it takes a few seconds to compute on a GPU. Again, the kernel operation $\mathbf{K}a$ on a vectorized 3D grid a can be computed via a sequence of 3 Gaussian convolutions on each axis (x, y, z) which reduces the complexity of one Debiased / IBP iteration from $O(n^2)$ to $O(n^{\frac{4}{3}})$.

Optimal transport barycentric embeddings One of the many machine learning applications of OT barycenters

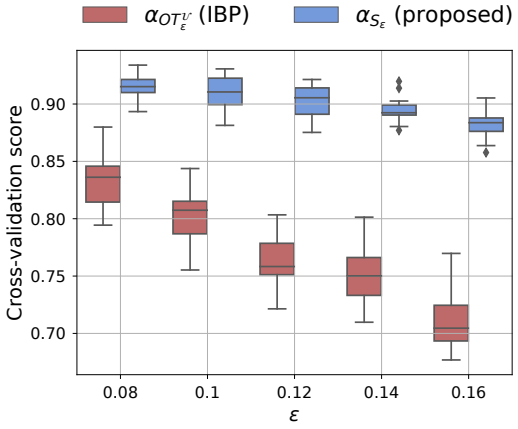

Figure 9. Cross-validation accuracy with 95% confidence intervals obtained on 500 MNIST images using barycentric embedding with S_ε or $\mathrm{OT}_\varepsilon^{\mathcal{U}}$. Debiasing of S_ε improves performance. S_ε is less sensitive to ε.

is to compute low-dimensional barycentric embeddings. Introduced by Bonneel et al. (2016), OT barycentric coordinates are defined as follows. Given a dictionary $\mathcal{A}$ of distributions $\alpha_1, \dots, \alpha_K$ and $w \in \Delta_K$, let $\alpha_F(w) = \arg\min_\alpha \sum_{k=1}^K w_k F(\alpha_k, \alpha)$ for some OT divergence F. The OT coordinates $\hat{w}$ of a distribution β are defined as the weights of the barycenter $\alpha_F(w)$ best approximating β for a given divergence. Using a quadratic divergence, it reads: $\hat{w} = \arg\min_{w \in \Delta_K} \|\alpha_F(w) - \beta\|^2$. To leverage the differentiability of the IBP iterations, Bonneel et al. (2016) used the divergence $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ and proposed to substitute the minimizer $\alpha_F(w)$ with the l-th IBP iterate $\alpha_F^{(l)}(w)$. Differentiating the barycenter nets $\alpha_F^{(l)}(w)$ with respect to w can be done via automatic differentiation, while the full minimization can be done using accelerated gradient descent using a soft-max reparametrization. Here we use the ADAM optimizer of the pyTorch library (Paszke et al., 2017). To evaluate the benefits of debiasing, we take 500 samples of the MNIST dataset (LeCun & Cortes, 2010) with 100 instances of each digit (0-1-2-3-4). We select 10% of the dataset (a subset of 50 images; ergo K=50) at random as our learning dictionary $\mathcal{A}$ and compute the barycentric coordinates of the remaining 90% subset denoted as $\mathcal{D}$. Thus, for each image among the 450 samples of $\mathcal{D}$, we compute the closest (in squared ℓ_2) weighted barycenter of the elements of $\mathcal{A}$ by optimizing over the weights. Thus, each image is represented by a vector of weights $w \in \Delta_K$. Our new embedded dataset is now a table of shape (450×50). We train a random forest classifier using the Scikit-learn library (Pedregosa et al., 2011) on this learned embedding) and compute a 10-fold cross-validation. Figure 9 displays the accuracy scores for $F = \mathrm{OT}_\varepsilon^{\mathcal{U}}$ and $F = \mathrm{S}_\varepsilon$ for 20 different randomized selections of the dictionary $\mathcal{A}$. The debiased S_ε improves accuracy and is less sensitive to the setting of ε.

Conclusion

Entropy regularized OT was previously known to induce a bias that can be mitigated using Sinkhorn divergences. Using OT barycenters of Gaussian distributions, we have shown that this entropy bias can be a blur or a shrink depending on the reference measure defining the relative entropy function. We have also extended the convexity and differentiability properties of OT and the Sinkhorn divergence to measures with non-compact supports.

Acknowledgments

MC and HJ acknowledge the support of a chaire dexcellence de l'IDEX Paris Saclay. AG and HJ were supported by the European Research Council Starting Grant SLAB ERC-YStG-676943. We thank Thibault Séjourné and François-Xavier Vialard for fruitful discussions, in particular for pointing out the identity (8). We thank Zikai Ziong for sharing the matlab code and adapting it to our ellipses experiment.

References

Agueh, M. and Carlier, G. Barycenters in the Wasserstein space. *SIAM*, 43(2):904–924, 2011.

Amari, S.-i., Karakida, R., Oizumi, M., and Cuturi, M. Information geometry for regularized optimal transport and barycenters of patterns. *Neural computation*, 31(5): 827–848, 2019.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative bregman projections for regularized transportation problems. *SIAM J. Scientific Computing*, 37, 2014.

Berg, C., Christensen, J. P. R., and Ressel, P. *Harmonic Analysis on Semigroups*. Springer, Berlin, 1984.

Blondel, M., Seguy, V., and Rolet, A. Smooth and sparse optimal transport. *international conference on artificial intelligence and statistics*, 2018.

Bonneel, N., Peyré, G., and Cuturi, M. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4), July 2016. ISSN 0730-0301. doi: 10.1145/2897824.2925918. URL https://doi.org/10.1145/2897824.2925918.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Scaling Algorithms for Unbalanced Transport Problems. *arXiv:1607.05816 [math.OC]*, 2017.

Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Neural Information Processing Systems*, 2013.

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 685–693, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/cuturi14.html.

Di Marino, S. and Gerolin, A. An Optimal Transport approach for the Schrodinger bridge problem and convergence of Sinkhorn algorithm, 2019.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. Interpolating between optimal transport and mmd using sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 10 2018.

Ge, D., Wang, H., Xiong, Z., and Ye, Y. Interior-point methods strike back: Solving the wasserstein barycenter problem. In *NeurIPS 2019*, 2019.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1608–1617. PMLR, 09–11 Apr 2018.

Ivan Gentil, Christian Léonard, L. R. About the analogy between optimal transport and minimal entropy. *Annales de la Facult des Sciences de Toulouse, Mathmatiques.*, 2017.

Knight, P. A., Ruiz, D., and Uar, B. A symmetry preserving algorithm for matrix scaling. *SIAM Journal on Matrix Analysis and Applications*, 35, 07 2014. doi: 10.1137/110825753.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Li, J. and Wang, J. Z. Real-time computerized annotation of pictures. In *Proceedings of the 14th ACM International Conference on Multimedia*, MM 06, pp. 911920, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595934472. doi: 10.1145/1180639.1180841. URL https://doi.org/10.1145/1180639.1180841.

Luise, G., Rudi, A., Pontil, M., and Ciliberto, C. Differential properties of sinkhorn approximation for learning with wasserstein distance. In *Advances in Neural Information Processing Systems*, pp. 5859–5870, 2018.

Luise, G., Salzo, S., Pontil, M., and Ciliberto, C. Sinkhorn barycenters with free support via frank-wolfe algorithm. In *Advances in Neural Information Processing Systems*, 2019.

Mena, G. and Weed, J. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. 2019.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peyré, G. and Cuturi, M. Computational Optimal Transport. *arXiv e-prints*, March 2018.

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

Rigollet, P. and Weed, J. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11):1228 – 1235, 2018. ISSN 1631-073X. doi: https://doi.org/10.1016/j.crma.2018.10.010. URL http://www.sciencedirect.com/science/article/pii/S1631073X18302802.

Schmitzer, B. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41:A1443–A1481, 2016.

Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, July 2015. ISSN 0730-0301.

Sullivan, C. and Kaszynski, A. Pyvista: 3d plotting and mesh analysis through a streamlined interface for the visualization toolkit (vtk). *Journal of Open Source Software*, 4(37):1450, 2019. doi: 10.21105/joss.01450. URL https://doi.org/10.21105/joss.01450.