Sinkhorn Barycenters with Free Support via Frank Wolfe algorithm

Giulia Luise[1], Saverio Salzo[2], Massimiliano Pontil[1,2], Carlo Ciliberto[3]

[1] Department of Computer Science, University College London, UK
[2] CSML, Istituto Italiano di Tecnologia, Genova, Italy
[3] Department of Electrical and Electronic Engineering, Imperial College London, UK

Outline

Goal and contributions

We propose a novel method to compute the barycenter of a set of probability distributions with respect to the Sinkhorn divergence that:

- does not fix the support beforehand
- handles both discrete and continuous measures
- admits convergence analysis.

Goal and contributions

Our analyais hinges on the following contributions:

- We show that *the gradient of the Sinkhorn divergence is Lipschitz continuous* on the space of probability measures with respect to the Total Variation.
- We characterize the *sample complexity* of an emprical estimator approximating the Sinkhorn gradients.
- A byproduct of our analysis is the generalization of the Frank-Wolfe algorithm to settings where the objective functional is defined only on *a set with empty interior, which is the case for Sinkhorn divergence barycenter problem*.

Setting and Notation

$\mathcal{X} \subset \mathbb{R}^d$ is a compact set

c: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric cost function, e.g. $\mathsf{c}(\cdot, \cdot) = \|\cdot - \cdot\|_2^2$

$\mathcal{M}_1^+(\mathcal{X})$ is the space of probability measures on $\mathcal{X}$.

$\mathcal{M}(\mathcal{X})$ is the Banach space of finite signed measures on $\mathcal{X}$.

Entropic Regularized Optimal Transport

For any $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$, the Optimal Transport problem with entropic regularization is defined as follow

$$\mathsf{OT}_\varepsilon(\alpha,\beta) = \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}^2} \mathsf{c}(x,y)\, d\pi(x,y) + \varepsilon \mathsf{KL}(\pi|\alpha\otimes\beta), \qquad \varepsilon \geq 0 \quad (1)$$

where:

$\mathsf{KL}(\pi|\alpha\otimes\beta)$ is the *Kullback-Leibler divergence* between transport plan π and the product distribution $\alpha\otimes\beta$

$\Pi(\alpha,\beta) = \{\pi \in \mathcal{M}_+^1(\mathcal{X}^2) \colon \mathsf{P}_{1\#}\pi = \alpha,\ \mathsf{P}_{2\#}\pi = \beta\}$ is the transport polytope (with $\mathsf{P}_i \colon \mathcal{X}\times\mathcal{X} \to \mathcal{X}$ the projector onto the i-th component and $\#$ the push-forward)

Sinkhorn Divergences

To remove the bias induced by the KL, [Genevay et al., 2018] proposed to remove the autocorrelation terms $-\frac{1}{2}\mathrm{OT}_\varepsilon(\alpha,\alpha)$, $-\frac{1}{2}\mathrm{OT}_\varepsilon(\beta,\beta)$ from $\mathrm{OT}_\varepsilon(\alpha,\beta)$ in order to get a *divergence*

$$\mathrm{S}_\varepsilon(\alpha,\beta) = \mathrm{OT}_\varepsilon(\alpha,\beta) - \frac{1}{2}\mathrm{OT}_\varepsilon(\alpha,\alpha) - \frac{1}{2}\mathrm{OT}_\varepsilon(\beta,\beta), \tag{2}$$

which is nonnegative, convex and metrizes the weak convergence (see [Feydy et al., 2019]).

In the following we study barycenter problem with this Sinkhorn divergence.

Barycenter Problem

Barycenters of probabilities are useful in a range of applications, as texture mixing, Bayesian inference, imaging.

The barycenter problem w.r.t. Sinkhorn divergence is formulated as follows:

given $\beta_1, \dots \beta_m \in \mathcal{M}_1^+(\mathcal{X})$ input measures, and $\omega_1, \dots, \omega_m \geq 0$ a set of weights such that $\sum_{j=1}^m \omega_j = 1$, solve

$$\min_{\alpha \in \mathcal{M}_1^+(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha), \qquad \text{with} \qquad \mathsf{B}_\varepsilon(\alpha) = \sum_{j=1}^m \omega_j \, \mathsf{S}_\varepsilon(\alpha, \beta_j). \tag{3}$$

Approach: Frank-Wolfe algorithm

Classic methods to approach barycenter problem:

1. fix the support of the barycenter beforehand and optimize the weights only (convergence analysis available)

   OR

2. alternately optimize on weights and support points (no convergence guarantees)

Our approach via Frank-Wolfe:

- It iteratively populates the target barycenter, one point at the time;
- It does not require the support to be fixed beforehand;
- There is no hyperparameter tuning.

Frank-Wolfe Algorithm on Banach spaces

$\mathcal{W}$ Banach space, $\mathcal{W}^*$ topological dual and $\mathcal{D} \subset \mathcal{W}^*$ nonempty, convex, closed, bounded set.
$\mathsf{G} : \mathcal{D} \to \mathbb{R}$ convex + some smoothness properties

Algorithm 1 FRANK-WOLFE IN DUAL BANACH SPACES

Input: initial $w_0 \in \mathcal{D}$, precision $(\Delta_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}_{++}$, such that $\Delta_k(k+2)$ is nondecreasing.

For $k = 0, 1, \ldots$

Take z_{k+1} such that $\mathsf{G}'(w_k, z_{k+1} - w_k) \leq \min_{z\in\mathcal{D}} \mathsf{G}'(w_k, z - w_k) + \frac{\Delta_k}{2}$

$w_{k+1} = w_k + \frac{2}{k+2}(z_{k+1} - w_k)$

Theorem

Suppose in addition that $\nabla\mathsf{G}$ is L-Lipschitz continuous with $L > 0$. Let $(w_k)_{k\in\mathbb{N}}$ be obtained according to Alg 1. Then, for every integer $k \geq 1$,

$$\mathsf{G}(w_k) - \min \mathsf{G} \leq \frac{2}{k+2} L \left(\mathrm{diam}\mathcal{D}\right)^2 + \Delta_k. \tag{4}$$

Can Frank-Wolfe be applied?

Optimization domain. $\mathcal{M}_1^+(\mathcal{X})$ is convex, closed, and bounded in the Banach space $\mathcal{M}(\mathcal{X})$: ✔

Objective functional. The objective functional B_ε is convex since it is a convex combination of $\mathsf{S}_\varepsilon(\cdot, \beta_j)$, with $j = 1 \dots m$. ✔

Lipschitz continuity of the gradient. This is the most critical condition.

Lipschitz continuity of Sinkhorn potentials

This is one of the main contributions of the paper.

Theorem

The gradient $\nabla \mathsf{S}_\varepsilon$ is Lipschitz continuous, i.e. for all $\alpha, \alpha', \beta, \beta' \in \mathcal{M}_1^+(\mathcal{X})$,

$$\left\|\nabla \mathsf{S}_\varepsilon(\alpha,\beta) - \nabla \mathsf{S}_\varepsilon(\alpha',\beta')\right\|_\infty \lesssim \left(\left\|\alpha-\alpha'\right\|_{TV} + \left\|\beta-\beta'\right\|_{TV}\right). \quad (5)$$

It follows that $\nabla \mathsf{B}_\varepsilon$ is also Lipschitz continuous and hence our framework is suitable to apply FW algorithm.

How the algorithm works - I

The inner step in FW algorithm amounts to:

$$\mu_{k+1} \in \operatorname*{argmin}_{\mu \in \mathcal{M}_1^+(\mathcal{X})} \sum_{j=1}^{m} \omega_j \ \langle \nabla \mathsf{S}_\varepsilon[(\cdot, \beta_j)](\alpha_k), \mu \rangle \,. \tag{6}$$

Note that:

- by Bauer maximum principle $\rightarrow$ solutions of (6) are achieved at the extreme points of the optimization domain;
- extreme points of $\mathcal{M}_1^+(\mathcal{X})$ are Dirac deltas.

Hence (6) is equivalent to

$$\mu_{k+1} = \delta_{x_{k+1}} \qquad \text{with} \qquad x_{k+1} \in \operatorname*{argmin}_{x \in \mathcal{X}} \sum_{j=1}^{m} \omega_j \ \big(\nabla \mathsf{S}_\varepsilon[(\cdot, \beta_j)](\alpha_k)(x)\big). \tag{7}$$

How the algorithm works - II

Once the new support point x_{k+1} has been obtained, FW update corresponds to

$$\alpha_{k+1} = \alpha_k + \frac{2}{k+2}(\delta_{x_{k+1}} - \alpha_k) = \frac{k}{k+2}\alpha_k + \frac{2}{k+2}\delta_{x_{k+1}}. \tag{8}$$

Weights and support points are updated simultaneously at each iteration.

Convergence analysis-finite case

Theorem

Suppose that $\beta_1, \dots \beta_m \in \mathcal{M}_1^+(\mathcal{X})$ *have finite support and let* α_k *be the* k*-th iterate of our algorithm. Then,*

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\alpha \in \mathcal{M}_1^+(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha) \leq \frac{C_\varepsilon}{k+2}, \tag{9}$$

where C_ε *is a constant depending on* ε *and on the domain* $\mathcal{X}$*.*

What if the input measures $\beta_1, \dots \beta_m \in \mathcal{M}_1^+(\mathcal{X})$ are continuous and we only have access to samples?

Sample complexity of Sinkhorn Potentials

FW can be applied when only an *approximation* of the gradient is available.

Hence we need *quantify* the approximation error between $\nabla \mathsf{S}_\varepsilon(\cdot, \beta)$ and $\nabla \mathsf{S}_\varepsilon(\cdot, \hat{\beta})$ in terms of the sample size of $\hat{\beta}$:

Theorem (Sample Complexity of Sinkhorn Potentials)

Suppose that c *is smooth. Then, for any* $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$ *and any empirical measure* $\hat{\beta}$ *of a set of* n *points independently sampled from* β*, we have, for every* $\tau \in (0, 1]$

$$\|\nabla_1 \mathsf{S}_\varepsilon(\alpha, \beta) - \nabla_1 \mathsf{S}_\varepsilon(\alpha, \hat{\beta})\|_\infty \leq \frac{C_\varepsilon \log \frac{3}{\tau}}{\sqrt{n}} \tag{10}$$

with probability at least $1 - \tau$.

Convergence analysis-general case

Using the sample complexity of Sinkhorn gradient, we are able to characterize the convergence analysis of our algorithm in the general setting.

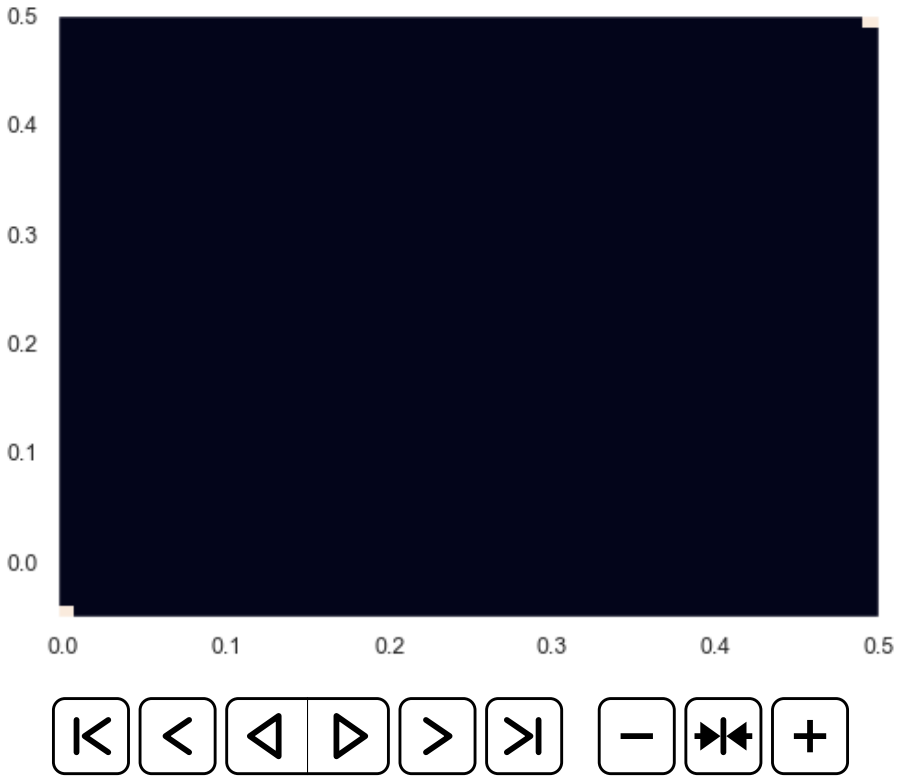
Theorem

Suppose that c *is smooth. Let* $n \in \mathbb{N}$ *and* $\hat{\beta}_1, \ldots, \hat{\beta}_m$ *be empirical distributions with* n *support points, each independently sampled from* $\beta_1, \ldots, \beta_m$. *Let* α_k *be the* k*-th iterate of our algorithm applied to* $\hat{\beta}_1, \ldots, \hat{\beta}_m$. *Then for any* $\tau \in (0, 1]$, *the following holds with probability larger than* $1 - \tau$

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\alpha \in \mathcal{M}_1^+(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha) \leq \frac{C_\varepsilon \log \frac{3m}{\tau}}{\min(k, \sqrt{n})}. \tag{11}$$

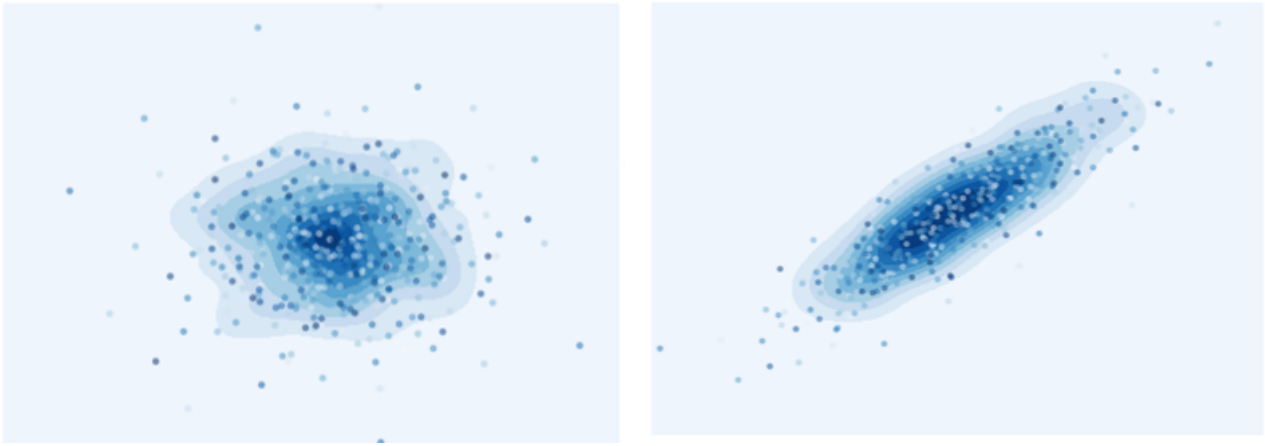
Barycenter of nested ellipses

Barycenter of 30 randomly generated nested ellipses on a 50×50 grid similarly to [Cuturi and Doucet, 2014]. Each image is interpreted as a probability distribution in 2D.



Barycenters of continuous measures

Barycenter of 5 Gaussian distributions with mean and covariance randomly generated.



scatter plot: output of our method

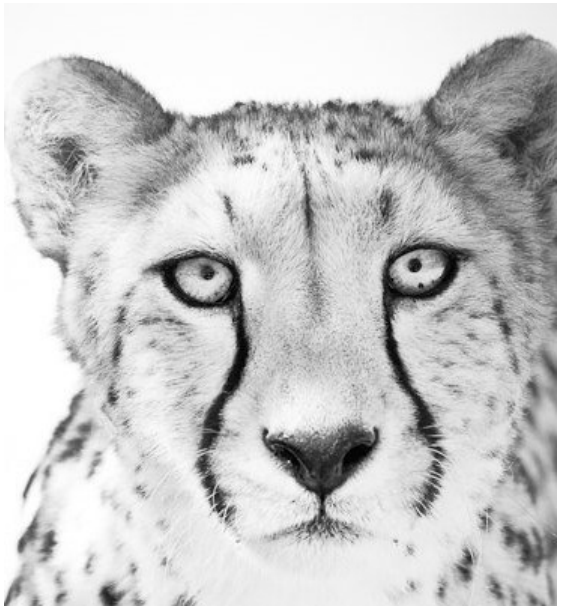
level sets of its density: true Wasserstein barycenter

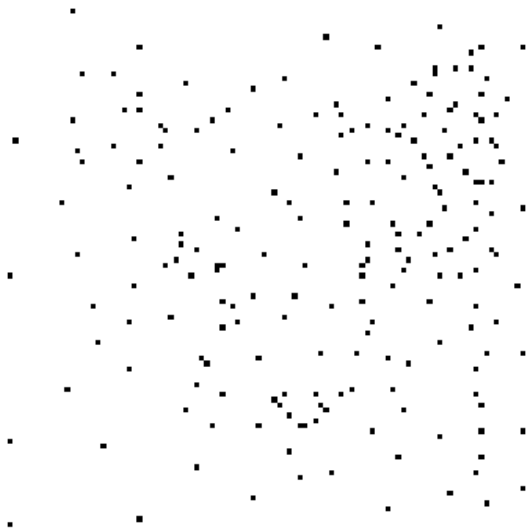
FW recovers both the mean and covariance of the target barycenter.

Matching of a distribution

"Barycenter" of a single measure $\beta \in \mathcal{M}_+^1(\mathcal{X})$.

Solution of this problem is β itself $\rightarrow$ we can interpret the intermediate iterates as compressed version of the original measure.





FW prioritizes the support points with higher weight.

Summary

- We proposed a novel method to compute Sinkhorn barycenter with free supports via Frank-Wolfe algorithm.
- We proved convergence rate both in case of finite and continuous measures.
- We proved two new results on Sinkhorn divergences- Lipschitz continuity and sample complexity of the gradient- instrumental for the convergence analysis of the method.

References I

Cuturi, M. and Doucet, A. (2014). Fast computation of wasserstein barycenters. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China. PMLR.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-I., Trouvé, A., and Peyré, G. (2019). Interpolating between optimal transport and mmd using sinkhorn divergences. *International Conference on Artificial Intelligence and Statistics (AIStats)*.

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