Sinkhorn Barycenters with Free Support via Frank Wolfe algorithm

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

November 28, 2019

[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 of the paper

Setting and problem statement

Approach

Convergence analysis

Experiments

Goal and contributions of the paper

Goal

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 analysis hinges on the following contributions:

- We show that *the gradient of the Sinkhorn divergence is Lipschitz continuous*
- 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 problem statement

Setting and Notation

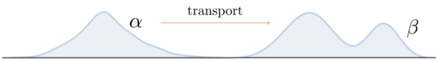
- $\mathcal{X} \subset \mathbb{R}^d$ is a compact set
- c: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric cost function, e.g. $\mathsf{c}(\cdot,\cdot) = \|\cdot - \cdot\|_2^2$
- $\mathcal{P}(\mathcal{X})$ is the space of probability measures on $\mathcal{X}$.
- $\mathcal{M}(\mathcal{X})$ is the Banach space of finite signed measures on $\mathcal{X}$.

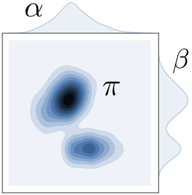
Entropic Regularized Optimal Transport

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

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

where $\Pi(\alpha,\beta) = \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{X}) \;\text{ s.t. }\; \mathsf{Proj}^1_\#\pi = \alpha, \mathsf{Proj}^2_\#\pi = \beta\}$.

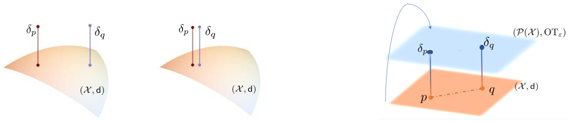




Properties of OT_ε

OT_ε **is used to compare probability measures**:

i) geometric flavour, lifting of the distance on $\mathcal{X}$ to $\mathcal{P}(\mathcal{X})$



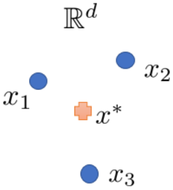
ii) meaningful for measures with non-overlapping support

Sinkhorn divergence [Genevay et al., 2018] is a small variant of OT_ε:

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

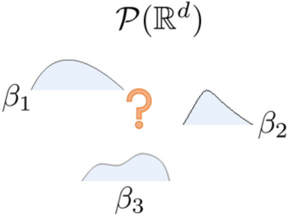
S_ε is nonnegative, convex (see [Feydy et al., 2019]).

Barycenters



aritmetic mean

$x^* = \mathrm{argmin}_{x \in \mathbb{R}^d} \sum_{i=1}^{3} \|x - x_i\|^2$



Sinkhorn barycenter

$\alpha^* = \mathrm{argmin}_{\alpha \in \mathcal{P}(\mathbb{R}^d)} \sum_{i=1}^{3} \mathsf{S}_\varepsilon(\alpha, \beta_i)$

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

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

Approach: Frank-Wolfe algorithm

Classic methods to approach barycenter problem:
assume $\alpha^* = \sum_{i=1}^{N} \mathsf{a}_i \delta_{x_i}$

1. fixed support methods: the support $\{x_i\}_{i=1}^{N}$ is fixed a priori and the optimization occurs on the weights only. E.g.: Iterative Bregman projections. Well understood convergence analysis.

OR

2. free support methods: a standard approach is to use alternating minimization on on weights and support points (no convergence guarantees). Different approach? Theoretical guarantees?

Our approach

Our approach via Frank-Wolfe:

- There is no alternation in optimizing wrt weights and wrt support points;
- It iteratively populates the barycenter, adding one point to the support at each iteration;
- It has no hyperparameter tuning.

Approach

Frank-Wolfe algorithm

- $\mathcal{W}$ is a real Banach space, with dual $\mathcal{W}^*$
- $\mathcal{D} \subset \mathcal{W}^*$ nonempty, convex, closed, bounded set
- $\mathsf{G} : \mathcal{D} \to \mathbb{R}$ convex function with $\nabla\mathsf{G} : \mathcal{D} \to \mathcal{W}$ Lipschitz

Algorithm 1 Frank-Wolfe

input: initial $w_0 \in \mathcal{D}$, threshold Δ_k s.t. $\Delta_k(k+2)$ is nondecreasing

For $k = 1, 2, \ldots$

take z_{k+1} s.t. $\langle \nabla\mathsf{G}(w_k), z_{k+1} - w_k \rangle \leq \min_{z\in\mathcal{D}} \langle \nabla\mathsf{G}(w_k), z - w_k \rangle + \frac{\Delta_k}{2}$

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

Convergence rate $O(1/k)$

[Jaggi, 2013]

Our setting

Recall: $\mathsf{B}_\varepsilon(\alpha) = \sum_{j=1}^m w_j \mathsf{S}_\varepsilon(\alpha, \beta_j)$.

$$\begin{array}{lcl}
\mathcal{W}^* & \longrightarrow & \mathcal{M}(\mathcal{X}) \\
\mathcal{D} \subset \mathcal{W}^* & \longrightarrow & \mathcal{P}(\mathcal{X}) \subset \mathcal{M}(\mathcal{X}) \\
\mathcal{W} & \longrightarrow & \mathcal{C}(\mathcal{X}) \\
\mathsf{G} : \mathcal{D} \longrightarrow \mathbb{R} & \longrightarrow & \mathsf{B}_\varepsilon : \mathcal{P}(\mathcal{X}) \longrightarrow \mathbb{R}
\end{array}$$

Optimization domain: $\mathcal{P}(\mathcal{X})$ is closed, convex and bounded in $\mathcal{M}(\mathcal{X})$ ✓

Objective functional:

- convexity ✓
- Lipschitzness of the gradient ?

Lipschitz continuity of Sinkhorn potentials

This is one of the main contributions of the paper.

Theorem

The gradient $\nabla \mathsf{S}_\varepsilon : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{X})$ *is Lipschitz continuous, i.e. for all* $\alpha, \alpha', \beta, \beta' \in \mathcal{P}(\mathcal{X})$*,*

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

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

Our setting

Recall: $\mathsf{B}_\varepsilon(\alpha) = \sum_{j=1}^m w_j \mathsf{S}_\varepsilon(\alpha, \beta_j)$.

$$
\begin{array}{lcl}
\mathcal{W}^* & \longrightarrow & \mathcal{M}(\mathcal{X}) \\
\mathcal{D} \subset \mathcal{W}^* & \longrightarrow & \mathcal{P}(\mathcal{X}) \subset \mathcal{M}(\mathcal{X}) \\
\mathcal{W} & \longrightarrow & \mathcal{C}(\mathcal{X}) \\
\mathsf{G} : \mathcal{D} \longrightarrow \mathbb{R} & \longrightarrow & \mathsf{B}_\varepsilon : \mathcal{P}(\mathcal{X}) \longrightarrow \mathbb{R}
\end{array}
$$

Optimization domain: $\mathcal{P}(\mathcal{X})$ is closed, convex and bounded in $\mathcal{M}(\mathcal{X})$ ✓

Objective functional:

- convexity ✓
- Lipschitzness of the gradient ?

Our setting

Recall: $\mathsf{B}_\varepsilon(\alpha) = \sum_{j=1}^m w_j \mathsf{S}_\varepsilon(\alpha, \beta_j)$.

$\mathcal{W}^*$	$\longrightarrow$	$\mathcal{M}(\mathcal{X})$
$\mathcal{D} \subset \mathcal{W}^*$	$\longrightarrow$	$\mathcal{P}(\mathcal{X}) \subset \mathcal{M}(\mathcal{X})$
$\mathcal{W}$	$\longrightarrow$	$\mathcal{C}(\mathcal{X})$
$\mathsf{G} : \mathcal{D} \longrightarrow \mathbb{R}$	$\longrightarrow$	$\mathsf{B}_\varepsilon : \mathcal{P}(\mathcal{X}) \longrightarrow \mathbb{R}$

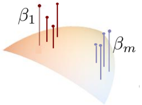
Optimization domain: $\mathcal{P}(\mathcal{X})$ is closed, convex and bounded in $\mathcal{M}(\mathcal{X})$ ✓

Objective functional:

- convexity ✓
- Lipschitzness of the gradient ✓

Convergence analysis

Convergence analysis-finite case



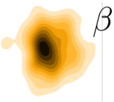
Theorem

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

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

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

Convergence analysis for a free-support method.



What if β_j are not finite and we only have access to samples?

Frank-Wolfe algorithm allows to use approximations of the gradient rather than the real gradient.

Algorithm 1 Frank-Wolfe

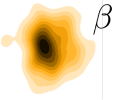
---

input: initial $w_0 \in \mathcal{D}$, treshold Δ_k s.t. $\Delta_k(k+2)$ is nondecreasing

For $k = 1, 2, \ldots$

take z_{k+1} s.t. $\langle \nabla \mathsf{G}(w_k), z_{k+1} - w_k \rangle \leq \min_{z \in \mathcal{D}} \langle \nabla \mathsf{G}(w_k), z - w_k \rangle + \frac{\Delta_k}{2}$

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



What if β_j are not finite and we only have access to samples?

Frank-Wolfe algorithm allows to use approximations of the gradient rather than the real gradient.

We need to control the approximation $\nabla \mathrm{B}_\varepsilon(\cdot, \hat{\beta})$ of $\nabla \mathrm{B}_\varepsilon(\cdot, \beta)$ $\longrightarrow$ it is enough to control the approximation $\nabla \mathrm{S}_\varepsilon(\cdot, \hat{\beta})$ of $\nabla \mathrm{S}_\varepsilon(\cdot, \beta)$. Can we do this?

Algorithm 1 Frank-Wolfe

input: initial $w_0 \in \mathcal{D}$, treshold Δ_k s.t. $\Delta_k(k+2)$ is nondecreasing

For $k = 1, 2, \dots$

take z_{k+1} s.t. $\langle \nabla \mathsf{G}(w_k), z_{k+1} - w_k \rangle \leq \min_{z \in \mathcal{D}} \langle \nabla \mathsf{G}(w_k), z - w_k \rangle + \frac{\Delta_k}{2}$

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

Sample complexity of Sinkhorn Gradients

Theorem (Sample Complexity of Sinkhorn Potentials)

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

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

with probability at least $1 - \tau$*.*

Convergence guarantees in general setting

Setting:

- cost function c smooth
- $\hat{\beta}_1, \ldots, \hat{\beta}_m$ be empirical distributions with $n \in \mathbb{N}$ support points, each independently sampled from $\beta_1, \ldots, \beta_m$.

Let α_k be the k-th iterate of FW applied to $\hat{\beta}_1, \ldots, \hat{\beta}_m$. Then for any $\tau \in (0, 1]$,

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

with probability larger than $1 - \tau$.

Convergence analysis in general setting

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\alpha \in \mathcal{P}(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha) \leq \frac{C_\varepsilon \log \frac{3m}{\tau}}{\min(k, \sqrt{n})} \qquad \text{w.h.p.}$$

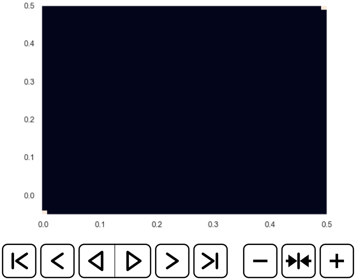
If $\hat{\beta}_j$, $j = 1, \ldots m$, are sampled with $n = k^2$ points at iteration k:
$\longrightarrow$ rate of convergence: $O(\frac{1}{k})$

If $\hat{\beta}_j$, $j = 1, \ldots m$, are sampled with $n = k$ points at iteration k:
$\longrightarrow$ rate of convergence: $O(\frac{1}{\sqrt{k}})$.

Experiments

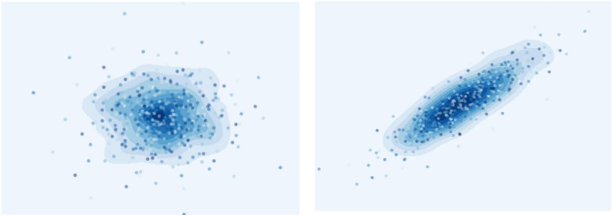
Barycenter of nested ellipses

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.



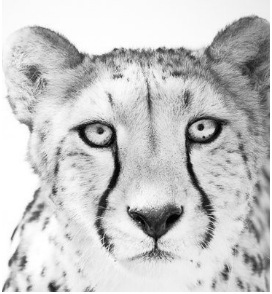
scatter plot: output of our method
level sets of its density: true Wasserstein barycenter

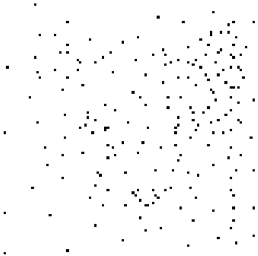
FW recovers both the mean and covariance of the target barycenter.

Matching of a distribution

"Barycenter" of a single measure $\beta \in \mathcal{P}(\mathcal{X})$.

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.

Thank you for your attention!

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.