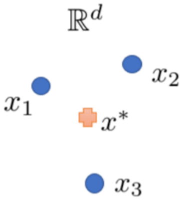
Sinkhorn Barycenters with Free Support via Frank Wolfe algorithm

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

10/12/2019, Vancouver, NeurIPS

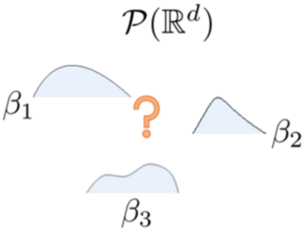
[1] Department of Computer Science, University College London, UK
[2] CSML, Istituto Italiano di Tecnologia, Genova, Italy
[3] Department of Electrical and Electronic Engineering, Imperial College London, UK

Barycenters



Aritmetic mean

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



Barycenter of probability measures

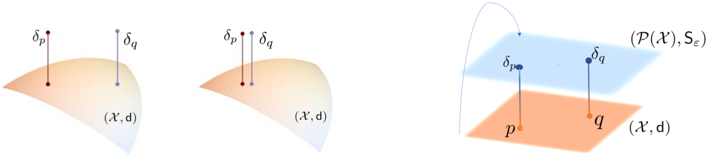
$\alpha^* = \text{argmin}_{\alpha \in \mathcal{P}(\mathbb{R}^d)} \sum_{i=1}^{3} \mathsf{d}(\alpha, \beta_i)$

We use the **Sinkhorn Divergence** (Entropic Optimal Transport) as a measure of similarity between probability distributions.

A glimpse on Sinkhorn divergence S_ε

S_ε **is used to compare probability measures**. Properties:

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



ii) it is well defined also for measures with non-overlapping support

Barycenter problem

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

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

$\hookrightarrow$ Optimization problem over the space of measures

Approach: Frank-Wolfe algorithm

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

1. fixed support methods: the support $\{x_i\}_{i=1}^{N}$ is fixed a priori and the optimization occurs on the weights only. [Benamou et al., 2015, Dvurechenskii et al., 2018] Well understood convergence analysis.

OR

2. free support methods: a standard approach is to use alternating minimization on on weights and support points [Cuturi and Doucet, 2014] (no convergence guarantees). Different approach? Theoretical guarantees?

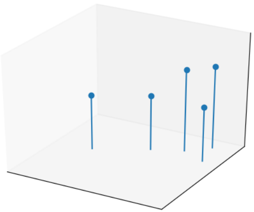
Free support barycenter algorithm

We use a **Frank-Wolfe** approach to minimize B_ε.

Start from $\alpha_0 := \delta_{x_0}$ with $x_0 \in \mathcal{X}$. For each iteration $k = 1, 2, \ldots$:

1. compute $\nabla \mathrm{B}_\varepsilon(\alpha_k)$ with SINKHORN ALGORITHM ($\nabla \mathrm{B}_\varepsilon$ is a smooth function on $\mathcal{X}$)
2. find $x_{k+1} \in \operatorname{argmin}_{x \in \mathcal{X}} \nabla \mathrm{B}_\varepsilon(\alpha_k)(x)$
3. update

$$\alpha_{k+1} = \frac{k}{k+2}\alpha_k + \frac{2}{k+2}\delta_{x_{k+1}}$$

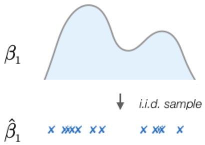


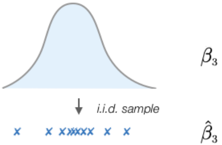
Convergence analysis in the general setting

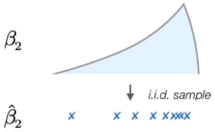
The input measures may be **continuous** . . .

Convergence analysis in the general setting

The input measures may be **continuous** . . .

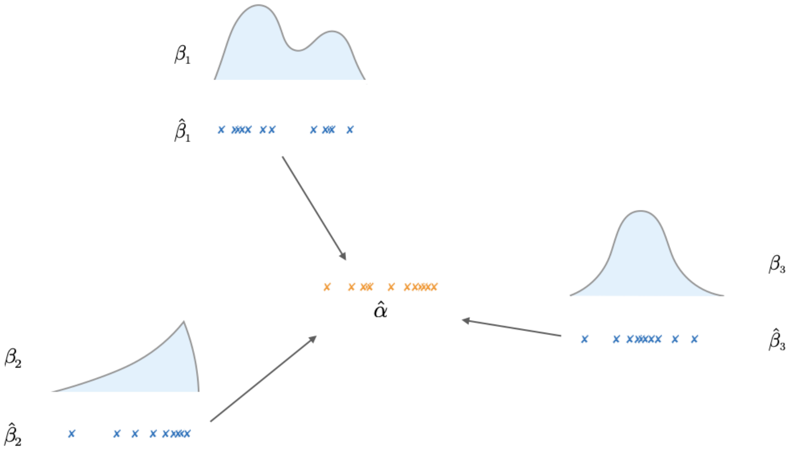






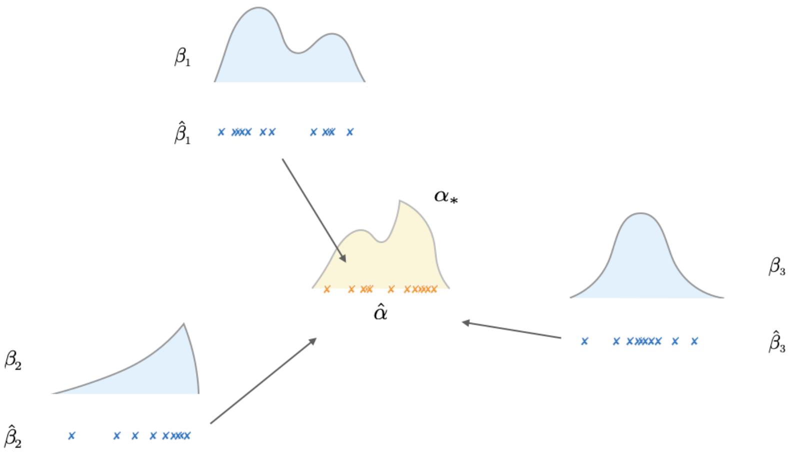
Convergence analysis in the general setting

$\longrightarrow$ we only have access to **samples**. . .



Convergence analysis in the general setting

Question: how close is the estimated $\hat{\alpha}$ to the **ideal** barycenter α_*?



Convergence analysis in the general setting

General setting: continuous input measures $\beta_j \longrightarrow$ we have only access to samples.

We obtain $\hat{\beta}_1, \ldots, \hat{\beta}_m$ 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 $\hat{\beta}_1, \ldots, \hat{\beta}_m$.

Theorem. for any $\tau \in (0, 1]$, the following holds with probability larger than $1 - \tau$

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

To wrap up: contributions

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

- it does not fix the support beforehand (free support method)
- it handles both discrete and continuous measures
- we provide convergence analysis.

The approach builds upon the Frank-Wolfe algorithm. In order to apply FW, we prove new smoothness results on Sinkhhorn divergence.

To see experiments and all the details,

Poster #113

References i

Benamou, J., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015). Iterative bregman projections for regularized transportation problems. *SIAM J. Scientific Computing*, 37(2).

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.

Dvurechenskii, P., Dvinskikh, D., Gasnikov, A., Uribe, C., and Nedich, A. (2018). Decentralize and randomize: Faster algorithm for wasserstein barycenters. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 10760–10770. Curran Associates, Inc.