Computing Wasserstein Barycenter via operator splitting: the method of averaged marginals

Daniel Mimouni[1,2,*], Paul Malisani[1], Jiamin Zhu[1], Welington de Oliveira[2]

[1]IFP Energies nouvelles
[2]Mines Paris, Université PSL, Centre de Mathématiques Appliquées (CMA)
*Corresponding author: daniel.mimouni@ifpen.fr

Motivation and Objectives

The Wasserstein barycenter (WB) is an important tool for summarizing sets of probabilities.

Applications : Applied probability, clustering, image processing, stochastic optimization etc., in data science in general when comparing summarize or reduce dimensions is at stake.
Challenge : Computing a WB (large linear optimization problem) generally exceeds standard solvers' capabilities. Therefore, the WB problem is often replaced with a simpler approximated optimization model.
Contribution : We introduce an exact method for computing Wasserstein barycenters in the case of finite and fixed support data and provide an effective algorithm that competes with state-of-the art methods.

1. Model problem and parametrization

We consider empirical measures of the form:

$$\mu = \sum_{r=1}^R p_r \delta_{\xi_r} \quad \text{and } \nu = \sum_{s=1}^S q_s \delta_{\zeta_s},$$

With support defined as finitely many R scenarios $\{\xi_1,\dots,\xi_R\}$ for ξ and S scenarios $\{\zeta_1,\dots,\zeta_S\}$ for ζ. With δ_u the Dirac unit mass on $u\in\Omega$, $p\in\Delta_R$, and $q\in\Delta_S$.

Definition 1 (Discrete Wasserstein Distances) The 2-Wasserstein distance $W_2(\mu,\nu)$ of two empirical measures μ and ν is the root squared of the optimal value of the following LP, known as *transportation problem*

$$\mathtt{OT}(p,q) := \begin{cases} \min_{\pi\geq 0} \sum_{r=1}^R\sum_{s=1}^S \mathsf{d}(\xi_r,\zeta_s)^2\pi_{rs} \\ \text{s.t. } \sum_{r=1}^R \pi_{rs} = q_s, \quad s=1,\dots,S \\ \qquad \sum_{s=1}^S \pi_{rs} = p_r, \quad r=1,\dots,R \end{cases}$$

Definition 2 (Discrete Wassertein Barycenter - WB) A Wassertein barycenter of a set of M empirical probabilities measures $\nu^{(m)}$, is a solution to the following optimization problem

$$\min_{p\in\Delta_R} \sum_{m=1}^M \frac{1}{M}\mathtt{OT}(p,q^{(m)})$$

For empirical measure the WB problem can be written as:

$$\begin{cases} \min_{p,\pi} \frac{1}{M}\sum_{r=1}^R\sum_{s=1}^{S^{(1)}} \mathsf{d}^{(1)}_{rs}\pi^{(1)}_{rs} + \cdots + \frac{1}{M}\sum_{r=1}^R\sum_{s=1}^{S^{(M)}} \mathsf{d}^{(M)}_{rs}\pi^{(M)}_{rs} \\ \text{s.t. } \sum_{r=1}^R \pi^{(1)}_{rs} \qquad\qquad = q^{(1)}_s, \quad s=1,\dots,S^{(1)} \\ \qquad \ddots \qquad\qquad\qquad \vdots \\ \qquad\qquad \sum_{r=1}^R \pi^{(M)}_{rs} = q^{(M)}_s, \quad s=1,\dots,S^{(M)} \\ \sum_{s=1}^{S^{(1)}} \pi^{(1)}_{rs} \qquad\qquad = p_r, \quad r=1,\dots,R \\ \qquad \ddots \qquad\qquad\qquad \vdots \\ \qquad\qquad \sum_{s=1}^{S^{(M)}} \pi^{(M)}_{rs} = p_r, \quad r=1,\dots,R \\ p\in\Delta_R, \pi^{(1)}\geq 0 \quad \cdots \quad \pi^{(M)}\geq 0 \end{cases}$$

2. How to solve such a huge scale LP ?

Method of Averaged Marginals - MAM :

Step 1: Given a multi-transportation plan θ^k
- Marginals $p^{(m),k} = \theta^{(m),k}\mathbb{1}, m=1,\dots,M$
- p^k is a weighted average of $\{p^{(1),k},\dots,p^{(M),k}\}$

Step 2: Given θ^k, p^k and distance matrices
- Compute a multi-transportation plan π^k by performing $\sum_{m=1}^M S^{(m)}$ independant projections onto the simplex Δ_R

Step 3: Givent θ^k, p^k and π^k
- Compute θ^{k+1} by a straightforward operation
- Set $k=k+1$ and repeat

Theorem : The sequence $\{p^k\}$ produced by MAM converges to a WB.

MAM is a new variant of the Douglas-Rachford operator splitting method. This method has for instance been used to derive ADMM or progressive hedging methods.

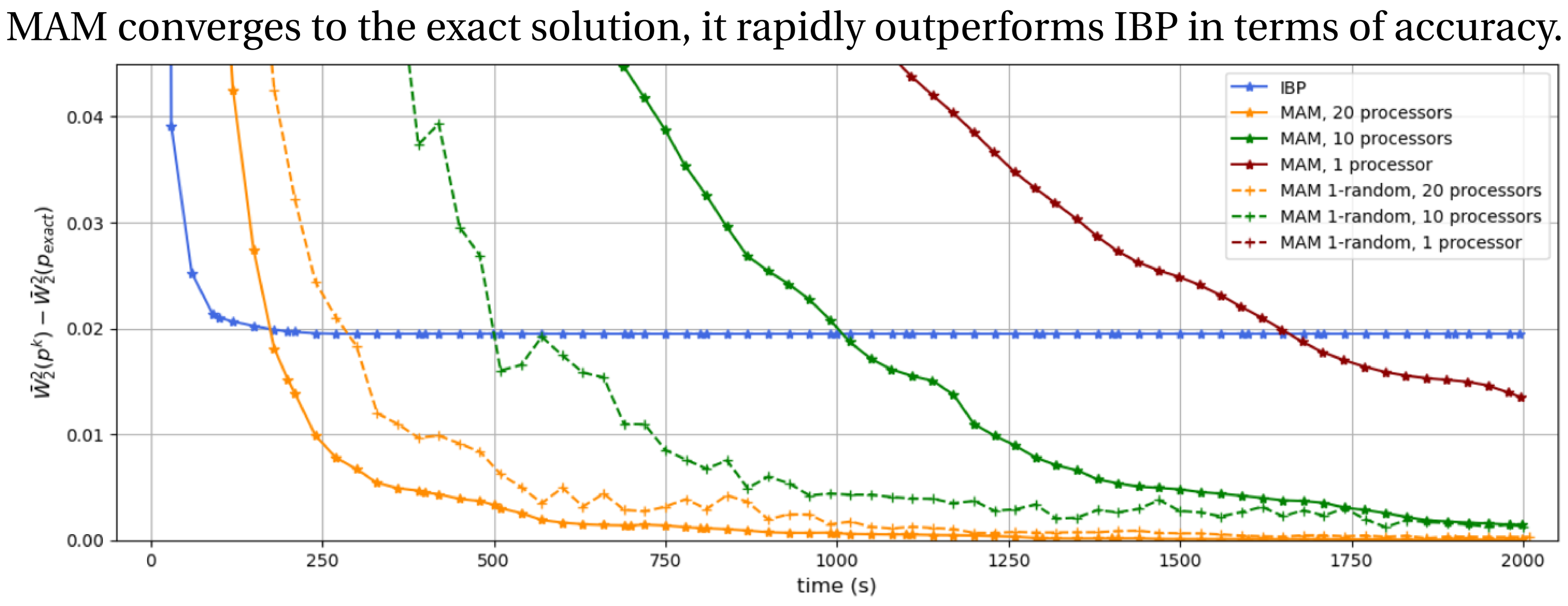
3. Qualitative study : MAM vs. IBP

- MAM: Exact algorithm
- IBP: Iterative Bregman Projection is a state-of-the-art algorithm for WB. IBP is based on an entropic regularization of the problem, thus it computes an inexact WB.

(top) For each digit, 36 out of the 100 scaled, translated and rotated images considered for each barycenter. (bottom) Barycenters after $t = 10, 50, 500, 1000, 2000$ seconds, where the left-hand-side is IBP evolution of its barycenter approximation, the middle panel is MAM evolutions using 10 processors (CPU) and the right-hand-side is the exact solution computed with Gurobi.

4. Quantitative study : MAM vs. IBP

IBP, MAM and randomized MAM are compared. IBP computes the exact solution of an inexact problem tuned through a bounded hyperparameter, therefore it is natural to witness IBP converging to a solution close but not equal to an exact WB.
MAM converges to the exact solution, it rapidly outperforms IBP in terms of accuracy.



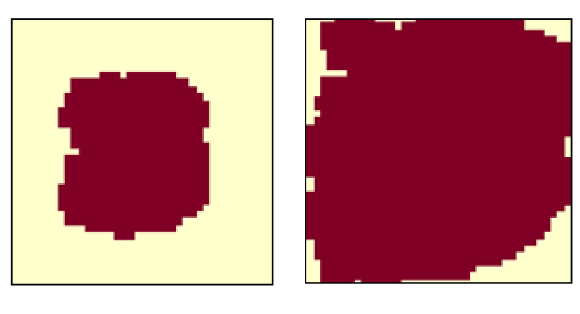
Evolution w.r.t time, of the difference between the Wasserstein barycenter distance of an approximation and the Wasserstein barycentric distance of the exact solution given by the LP, defined as:

$$\bar{W}_2^2(\mu^k) - \bar{W}_2^2(\mu_{exact}) := \sum_{m=1}^M \frac{1}{M}\mathtt{OT}(p^k,q^{(m)}) - \sum_{m=1}^M \frac{1}{M}\mathtt{OT}(p_{exact},q^{(m)})$$

60 images are used for the digit '3' from the (40x40) centered MNIST databased.

5. Influence of the support

The larger is the support size, the more inaccurate IBP becomes. Indeed, the greater is the support size, the more restrained is the choice of the hyperparameter for IBP, due to a double-precison overflow error. Being an exact method, MAM is insensitive to support size.



40×40 pixel grid, where the red represents the pixels which are in the union of the dataset support composed by 60 distributions. (left) for the classical MNIST, (right) for the randomly translated and rotated MNIST.

LaTeX TikZposter