MAM - Revisited

DANIEL MIMOUNI AND WELINGTON DE OLIVEIRA[1]
Mines Paris, Université PSL, Centre de Mathématiques Appliquées (CMA), Sophia Antipolis, France

Abstract: The Method of Averaged Marginals (MAM) has recently been proposed as an exact approach for computing Wasserstein barycenters of empirical measures. It is memory-efficient, can operate in either deterministic or randomized modes, addresses scalability challenges, and comes with convergence guarantees. However, its main computational bottleneck lies in the repeated projections onto the $(R-1)$-dimensional simplex required at each iteration. This work overcomes that limitation by introducing a variant of MAM that eliminates the need for such projections. Crucially, for all other steps, the proposed variant preserves the original method's computational complexity and memory footprint. This modification enables MAM to be executed on GPUs, as the algorithm primarily relies on simple yet highly parallelizable matrix operations at each iteration. Moreover, the method can accommodate additional constraints, provided that projecting onto the set defined by these constraints, intersected with a hyperplane, is computationally efficient.

Keywords: Wasserstein barycenter, Splitting methods, DC Programming, Linkage problem
Mathematics Subject Classification: 90C26, 65K05, 49J52, 49J53

January 5, 2026
From wlo to Juan Nicolás Mendoza Roncancio

---

1 Introduction

Juan, read this text assuming that $X = \mathbb{R}^R$ is the entire space, i.e., there is no constraint on the barycenter. In your project the constraint will appear on the transportation plans.

Let $\mathcal{P}(\mathbb{R}^d)$ be the set of Borel probability measures on $\mathbb{R}^d$ and $\mathfrak{X} \subset \mathcal{P}(\mathbb{R}^d)$ a closed set. A *Constrained Wasserstein barycenter* (CWB) of a set of M measures $\nu^m \in \mathcal{P}(\mathbb{R}^d)$, $m = 1, \ldots, M$, is a solution to the following optimization problem

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \frac{1}{M} \sum_{m=1}^{M} W_2^2(\mu, \nu^m) \quad \text{s.t.} \quad \mu \in \mathfrak{X}\,, \tag{1}$$

where $W_2(\mu, \nu)$ is the (quadratic) 2-Wasserstein distance between two measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$; see (2) below. The additional constraint $\mu \in \mathfrak{X}$ is particularly relevant in applications where the barycenter must adhere to specific structural or operational requirements, or align with prior knowledge about the desired properties of μ. The set $\mathfrak{X}$ can influence the geometry of the barycenter, its statistical properties, its support, or its physical suitability.

Notation. Given $R \in \mathbb{N}$ and $\tau > 0$, let $\mathbf{1}_R$ be the column vector of all ones of size R, $\Delta_R := \{y \in \mathbb{R}^R_+ : y^\top \mathbf{1}_R = 1\}$is the $(R-1)$-simplex and $H = \{y \in \mathbb{R}^R : y^\top \mathbf{1}_R = 1\}$. The Euclidean projection of $y \in \mathbb{R}^R$ onto a closed set $X \subset \mathbb{R}^R$ is denoted by $\mathtt{Proj}_X(y)$. Also, δ_ξ denotes the Dirac unit mass on a given point $\xi \in \mathbb{R}^d$.

2 Background Material

Let ξ and ζ be two random vectors having probability measures μ and ν in $\mathcal{P}(\mathbb{R}^d)$, that is, $\xi \sim \mu$ and $\zeta \sim \nu$. Their 2-Wasserstein distance is given by:

$$W_2(\mu, \nu) := \left(\inf_{\pi \in U(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\xi - \zeta\|^2 d\pi(\xi, \zeta) \right)^{1/2}, \tag{2}$$

where $U(\mu, \nu)$ is the set of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ having marginals μ and ν. We denote by $W_2^2(\mu, \nu)$ the squared 2-Wassserstein distance, i.e., $W_2^2(\mu, \nu) := (W_2(\mu, \nu))^2$. In this work, we are concerned

[1]Corresponding author: welington.oliveira@minesparis.psl.eu

with empirical (discrete) measures ν^m having finite support sets: for all $m = 1, \dots, M$, the number of atoms of ν^m is denoted by S^m, its support by

$$\mathtt{supp}(\nu^m) := \{\zeta_1^m, \dots, \zeta_{S^m}^m\}, \quad \text{probability mass by } \; q^m \in \Delta_{S^m}, \quad \text{and thus } \; \nu^m = \sum_{s=1}^{S^m} q_s^m \delta_{\zeta_s^m}. \tag{3}$$

It follows from the definition of the support of a measure ν^m that $\nu^m(\zeta_s^m) = q_s^m > 0$ for all $s = 1, \dots, S^m$. As computing a CWB of M measures ν^m amounts to determine a new measure $\bar{\mu}$ solving (1), it turns out that such a task consists of choosing simultaneously a support $\mathtt{supp}(\bar{\mu})$ and a probability vector $\bar{p}$ minimizing the (weighted) Wasserstein distance to all M measures. As for the decision on the support, Proposition 1 in [1] asserts that every solution $\bar{\mu}$ to a unconstrained version of (1) (i.e., $\mathfrak{X} = \mathcal{P}(\mathbb{R}^d)$)) has support satisfying the following key inclusion:

$$\begin{aligned} \mathtt{supp}(\bar{\mu}) \subset \Xi &:= \left\{\xi^1, \dots, \xi^R\right\} \\ &:= \left\{ \frac{1}{M} \sum_{m=1}^{M} \zeta^m : \; \zeta^m \in \mathtt{supp}(\nu^m), \; m = 1, \dots, M \right\}. \end{aligned} \tag{4}$$

Thanks to this result, we can work with the fixed set Ξ having finitely many R atoms[1] and optimize only with respect to the probability vector: once $\bar{p}$ is determined, we can recover a WB measure by setting

$$\bar{\mu} = \sum_{r \in \{j : \bar{p}_j > 0\}} \bar{p}_r \delta_{\xi_r}. \tag{5}$$

Accordingly, two observations arise. First, with two empirical distributions μ and ν^m, the squared 2-Wasserstein distance simplifies to the following transportation problem:

$$W_2^2(\mu, \nu^m) = \min_{\pi \in \mathbb{R}_+^{R \times S^m}} \sum_{r=1}^{R} \sum_{s=1}^{S^m} \|\xi_r - \zeta_s^m\|^2 \pi_{rs} \quad \text{s.t.} \quad (\pi)^\top \mathbf{1}_R = q^m \text{ and } \pi \, \mathbf{1}_{S^m} = p.$$

Second, thanks to (4), imposing a constraint of the type $\mu \in \mathfrak{X}$ can be done by restricting p to a certain set $X \subset \mathbb{R}^R$ related to $\mathfrak{X}$. In other words, in the empirical setting, the constrained WB problem (1) can be alternatively written as follows, for a set $X \subset \mathbb{R}^R$ associated to $\mathfrak{X} \subset \mathcal{P}(\mathbb{R}^d)$:

$$\left\{ \begin{array}{ll} \min\limits_{p \in X, \; \pi \geq 0} & \displaystyle\sum_{m=1}^{M} \frac{1}{M} \sum_{r=1}^{R} \sum_{s=1}^{S^m} \|\xi_r - \zeta_s^m\|^2 \pi_{rs}^m \\ \text{s.t.} & (\pi^m)^\top \mathbf{1}_R = q^m, \quad m = 1, \dots, M \\ & \pi^m \, \mathbf{1}_{S^m} = p, \quad m = 1, \dots, M. \end{array} \right. \tag{6}$$

We highlight that this problem is solvable as long as X is closed and intersects the simplex Δ_R.

3 Constrained Wasserstein Barycenter

The Method of Averaged Marginals (MAM), originally proposed in [7] for computing unconstrained Wasserstein barycenters (WBs) and later extended in [6] to handle constrained WBs, solves the linear program (6) by leveraging its specific structure and applying the Douglas–Rachford splitting method (DR) [4, 5]. The resulting algorithm is memory efficiently, can run in a deterministic or randomized fashion, copes with scalability issues, and has convergence guarantees.

At every iteration, MAM updates transportation plans by projecting (in parallel) $\sum_{m=1}^{M} S^m$ vectors of dimension R onto the simplex Δ_R. This step constitutes the primary computational bottleneck of the algorithm. In what follows, we introduce a variant of MAM that eliminates the need for such projections. Importantly, all other steps retain the original method's computational complexity and memory efficiency. This modification enables MAM to be executed on GPUs, as the algorithm primarily relies on simple yet highly parallelizable matrix operations at each iteration. Throughout this work we make the following assumption (recall that Δ_R stands for the simplex in $\mathbb{R}^R$ and $H = \{y \in \mathbb{R}^R : y^\top \mathbf{1}_R = 1\}$).

Assumption 1. *The set $X \subset \mathbb{R}^R$ in (6) is closed, and satisfies $X \cap \Delta_R \neq \emptyset$. Furthermore, the Euclidean projection onto $X \cap H$ is convenient to execute.*

[1] The WB's support size R is determined in function of the number M of measures ν^m and their support sizes S^m, $m = 1, \dots, M$; see [1].

We recall that closeness of X and condition $X \cap \Delta_R \neq \emptyset$ are enough to ensure that (6) is solvable. Next, by denoting

$$c^m_{rs} := \frac{1}{M}\|\xi_r - \zeta^m_s\|^2 \ \ \forall\, r,\ s,\ m, \qquad \text{and inner product} \quad \langle c, \pi\rangle := \sum_{r,s,m} c^m_{rs}\pi^m_{rs}, \tag{7}$$

we drop the decision variable p in (6) and rewrite the problem in the following compact form:

$$\min_{\pi \geq 0} \langle c, \pi\rangle \quad \text{s.t.} \quad \pi \in \mathcal{K} \tag{8a}$$

where

$$\mathcal{K} := \left\{ \pi = (\pi^1, \ldots, \pi^M) \middle| \begin{array}{l} \pi^1\, \mathbf{1}_{S^1} = \pi^2\, \mathbf{1}_{S^2} = \cdots = \pi^M\, \mathbf{1}_{S^M} \in X \cap H \\ (\pi^m)^\top\, \mathbf{1}_R = q^m,\ m = 1, \ldots, M \end{array} \right\}. \tag{8b}$$

Thus, once problem (8) is solved, we can easily recover a p-solution to problem (6) and, as a consequence, a constrained WB measure $\bar{\mu}$.

To solve (8) we employ the DR algorithm, which asymptotically computes a solution by repeating the following steps, with $k = 0, 1, \ldots$, given initial point $\theta^0 = (\theta^{1,0}, \ldots, \theta^{M,0})$, and prox-parameter $\rho > 0$:

$$\left\{ \begin{array}{lll} \pi^{k+1} & = & \texttt{Proj}_{\mathcal{K}}(\theta^k) \\ \hat{\pi}^{k+1} & = & \arg\min_{\pi \geq 0} \langle c, \pi\rangle + \frac{\rho}{2}\|\pi - (2\pi^{k+1} - \theta^k)\|^2 \\ \theta^{k+1} & = & \theta^k + \hat{\pi}^{k+1} - \pi^{k+1}. \end{array} \right. \tag{9}$$

Assumption 1 ensures that the functions above are proper, convex, lower-semicontinuous, and problem (8) is solvable. Convergence of the above scheme follows from Theorem 25.6 and Corollary 27.4 of [2].

The DR algorithm is attractive when the two first steps in (9) are convenient to execute, which is the case in our setting. Indeed, the second step is merely the projection of $2\pi^{k+1} - \theta^k - c/\rho$ onto the positive orthant:

$$\hat{\pi}^{k+1} = \max\left\{2\pi^{k+1} - \theta^k - c/\rho, 0\right\}.$$

Juan, in your project we wish to have constraints on π^m. We need to find applications and constraints such that the second step (9) has an explicit expression. This will be the crucial point for your project. If you find situations in which this step is easy, then you can update Algorithm 1 below to accommodate such type of constraints. Note that the constraint $\pi \geq 0$ must be present in this step, which may complicate things...

The following original result, which does not require X to be convex, shows that the first step in (9) is also simple provided the projection onto $X \cap H$ is convenient to execute. Juan, this proposition is the main contribution of this ongoing work. It is based on Romero's formula.

Proposition 1. *Let* $\theta = (\theta^1, \ldots, \theta^M) \in \mathbb{R}^{R\times(S^1+\cdots+S^M)}$,

$$a_m := \frac{\frac{1}{S^m}}{\frac{1}{S^1}+\cdots+\frac{1}{S^M}}, \quad \mathfrak{p}^m := \theta^m\, \mathbf{1}_{S^m}, \quad \gamma^m := (\mathfrak{p}^m)^\top\, \mathbf{1}_R - 1, \quad \text{and} \quad \mathfrak{q}^m := (\theta^m)^\top\, \mathbf{1}_R, \quad m = 1, \ldots, M.$$

Given a nonempty and closed set $X \cap H \neq \emptyset$, *let* $p \in \texttt{Proj}_{X\cap H}(\sum_{m=1}^M a_m p^m)$ *and* $\mathcal{K}$ *given in* (8b). *Then, an element* $\pi \in \texttt{Proj}_{\mathcal{K}}(\theta)$ *has the form*

$$\pi^m_{rs} = \theta^m_{rs} + \frac{p^m_r - \mathfrak{p}^m_r}{S^m} + \frac{q^m_s - \mathfrak{q}^m_s}{R} + \frac{\gamma^m}{RS^m} \quad \forall\, m, r, s. \tag{10}$$

Proof. Given an arbitrary $w \in H$, let us define the set

$$\mathcal{K}_w := \left\{ \pi = (\pi^1, \ldots, \pi^M) \middle| \begin{array}{l} \pi^1\, \mathbf{1}_{S^1} = \pi^2\, \mathbf{1}_{S^2} = \cdots = \pi^M\, \mathbf{1}_{S^M} = w \\ (\pi^m)^\top\, \mathbf{1}_R = q^m,\ m = 1, \ldots, M \end{array} \right\}.$$

Observe that $\mathcal{K}$ is nonempty[2] and can be written as $\mathcal{K} = \cup_{w\in X\cap H}\mathcal{K}_w$. Therefore, computing a point in $\texttt{Proj}_{\mathcal{K}}(\theta)$ can be done by solving

$$\min_{y\in\mathcal{K}} \|y - \theta\|^2 = \min_{w\in X\cap H,\, y\in\mathcal{K}_w} \|y - \theta\|^2 = \min_{w\in X\cap H} \left\{ \min_{y\in\mathcal{K}_w} \|y - \theta\|^2 \right\}.$$

[2]Because p and all q^m have the same mass.

The inner problem above can be written as

$$z(w) := \texttt{Proj}_{\mathcal{K}_w}(\theta) \;=\; \begin{cases} \arg\min\limits_{y} & \sum\limits_{m=1}^{M} \|y^m - \theta^m\|^2 \\ \text{s.t.} & y^m\, \mathbf{1}_{S^m} \qquad = w, \qquad m = 1, \dots, M \\ & (y^m)^\top \mathbf{1}_R \qquad = q^m \qquad m = 1, \dots, M. \end{cases}$$

Since w is fixed, this problem can be decomposed into M subproblems:

$$\begin{cases} \min\limits_{y^m} & \|y^m - \theta^m\|^2 \\ \text{s.t.} & y^m\, \mathbf{1}_{S^m} \qquad = w, \\ & (y^m)^\top \mathbf{1}_R \qquad = q^m, \end{cases}$$

and, because $\mathbf{1}_R^\top w = \mathbf{1}_R^\top q^m = 1$, each one of them has an explicit solution given by the Romero's formula [3, Remark 3.5]:

$$z^m(w) = \theta^m_{rs} + \frac{w_r - \mathfrak{p}^m_r}{S^m} + \frac{q^m_s - \mathfrak{q}^m_s}{R} + \frac{\gamma^m}{RS^m} \quad \forall\, m, r, s. \tag{11}$$

Next, we show that when w is an element of $\texttt{Proj}_{X\cap H}(\sum_{m=1}^M a_m \mathfrak{p}^m)$, then $z(w)$ above belongs to the set $\texttt{Proj}_{\mathcal{K}}(\theta)$. To this end, note that

$$\begin{aligned} \min_{y\in\mathcal{K}} \|y-\theta\|^2 &= \min_{w\in X\cap H} \left\{ \min_{y\in\mathcal{K}_w} \|y-\theta\|^2 \right\} \\ &= \arg\min_{w\in X\cap H} \|z(w)-\theta\|^2 \\ &= \min_{w\in X\cap H} \sum_{m=1}^{M}\sum_{r=1}^{R}\sum_{s=1}^{S^m} (z^m_{rs}(w) - \theta^m_{rs})^2 \\ &= \min_{w\in X\cap H} \sum_{m=1}^{M}\sum_{r=1}^{R}\sum_{s=1}^{S^m} \Big(\frac{w_r - \mathfrak{p}^m_r}{S^m} + \frac{q^m_s - \mathfrak{q}^m_s}{R} + \frac{\gamma^m}{RS^m} \Big)^2 \\ &= \min_{w\in X\cap H} \sum_{m=1}^{M}\sum_{r=1}^{R}\sum_{s=1}^{S^m} \Big(\frac{w_r}{S^m} - \big(\frac{\mathfrak{p}^m_r}{S^m} - \frac{q^m_s - \mathfrak{q}^m_s}{R} - \frac{\gamma^m}{RS^m}\big) \Big)^2. \end{aligned}$$

Recall that discarding constants in the objective function does not alter the set of optimal solutions. Thus, we can safely remove any additive constant terms without affecting the minimization problem:

$$\begin{aligned} \arg\min_{w\in X\cap H} \left\{ \min_{y\in\mathcal{K}_w} \|y-\theta\|^2 \right\} &= \arg\min_{w\in X\cap H} \sum_{m=1}^{M}\sum_{r=1}^{R}\sum_{s=1}^{S^m} \Big[\Big(\frac{w_r}{S^m}\Big)^2 - 2\Big(\frac{w_r}{S^m}\Big)\Big(\frac{\mathfrak{p}^m_r}{S^m} - \frac{q^m_s - \mathfrak{q}^m_s}{R} - \frac{\gamma^m}{RS^m}\Big) \Big] \\ &= \arg\min_{w\in X\cap H} \sum_{m=1}^{M}\sum_{r=1}^{R} \Big[S^m \Big(\frac{w_r}{S^m}\Big)^2 - 2\Big(\frac{w_r}{S^m}\Big)\Big(S^m \frac{\mathfrak{p}^m_r}{S^m} - \frac{\sum_{s=1}^{S^m}(q^m_s - \mathfrak{q}^m_s)}{R} - S^m \frac{\gamma^m}{RS^m}\Big) \Big] \\ &= \arg\min_{w\in X\cap H} \sum_{m=1}^{M}\sum_{r=1}^{R} \frac{1}{S^m} \Big[w_r^2 - 2w_r \Big(\mathfrak{p}^m_r - \frac{-\gamma^m}{R} - \frac{\gamma^m}{R} \Big) \Big] \end{aligned}$$

because

$$\sum_{s=1}^{S^m} (q^m_s - \mathfrak{q}^m_s) = 1 - \sum_{s=1}^{S^m} \mathfrak{q}^m_s = 1 - \sum_{s=1}^{S^m}\sum_{r=1}^{R} \theta^m_{rs} = 1 - \sum_{r=1}^{R}\sum_{s=1}^{S^m} \theta^m_{rs} = 1 - \sum_{r=1}^{R} \mathfrak{p}^m_r = -\gamma^m.$$

Now, dividing the objective function by the positive constant $\sum_{j=1}^{M} \frac{1}{S^j}$ does not alter the solution set:

$$
\begin{aligned}
\arg \min_{w \in X \cap H} \left\{ \min_{y \in \mathcal{K}_w} \|y - \theta\|^2 \right\} &= \arg \min_{w \in X \cap H} \sum_{m=1}^{M} \sum_{r=1}^{R} \frac{1}{\frac{S^m}{\sum_{j=1}^{M} \frac{1}{S^j}}} (w_r^2 - 2 w_r \mathfrak{p}_r^m) \\
&= \arg \min_{w \in X \cap H} \sum_{m=1}^{M} a_m \sum_{r=1}^{R} (w_r^2 - 2 w_r \mathfrak{p}_r^m) \\
&= \arg \min_{w \in X \cap H} \sum_{r=1}^{R} \left[w_r^2 \Big(\sum_{m=1}^{M} a_m \Big) - 2 w_r \Big(\sum_{m=1}^{M} a_m \mathfrak{p}_r^m \Big) \right] \\
&= \arg \min_{w \in X \cap H} \sum_{r=1}^{R} \left[w_r^2 - 2 w_r \Big(\sum_{m=1}^{M} a_m \mathfrak{p}_r^m \Big) \right]
\end{aligned}
$$

because $\sum_{m=1}^{M} a_m = 1$. By adding constants to complete the square, we get

$$
\begin{aligned}
\arg \min_{w \in X \cap H} \left\{ \min_{y \in \mathcal{K}_w} \|y - \theta\|^2 \right\} &= \arg \min_{w \in X \cap H} \sum_{r=1}^{R} \left[w_r - \Big(\sum_{m=1}^{M} a_m \mathfrak{p}_r^m \Big) \right]^2 \\
&= \arg \min_{w \in X \cap H} \left\| w - \Big(\sum_{m=1}^{M} a_m \mathfrak{p}^m \Big) \right\|^2 \\
&= \mathtt{Proj}_{X \cap H} \Big(\sum_{m=1}^{M} a_m \mathfrak{p}^m \Big).
\end{aligned}
$$

The proof is thus complete. □

Formula (10) yields a plan satisfying all the marginals $p, q^1, \ldots, q^M$. Indeed, this plan is known to be the projection of θ onto the set of plans that satisfies such marginals. This is a result by Romero [8], and is known as the Romero's formula [3, Remark 3.5].

Recall that this proposition does not assume convexity of X. Hence, whether convexity is present or not, projecting onto $\mathcal{K} \subset \mathbb{R}^{R \times (S^1 + \cdots + S^M)}$ is simple as long as the projection onto $X \cap H \subset \mathbb{R}^R$ is easy to perform. Here is an important difference with [6], which only assumes that projection onto X is simple.

4 The algorithm

Putting all together and making simplifications we get the extension of MAM presented in Algorithm 1.

Some comments on the algorithm are in order.

1. The cost matrix in Algorithm 1 can be arbitrary, for example as defined in (7).
2. The auxiliary variables θ^m are initialized by projecting the zero matrix onto the set of matrices whose column sums equal p^0 and row sums equal q^m.
3. If $X = \mathbb{R}^R$, then the projection of $y \in \mathbb{R}^R$ onto $X \cap H$ is simply $y - \frac{(\mathbf{1}_R^\top y - 1)}{R} \mathbf{1}_R$.
4. Observe that γ^m as defined in the algorithm coincides with the definition in Proposition 1. Indeed, $(\mathfrak{q}^m)^\top \mathbf{1}_{S^m} = \sum_{s=1}^{S^m} \mathfrak{q}_s^m = \sum_{s=1}^{S^m} \sum_{r=1}^{R} \theta_{rs}^m = \sum_{r=1}^{R} \sum_{s=1}^{S^m} \theta_{rs}^m = \sum_{r=1}^{R} \mathfrak{p}_r^m = (\mathfrak{p}^m)^\top \mathbf{1}_R$.
5. Each iteration of the algorithm relies on simple, highly parallelizable matrix operations, which can be efficiently executed on GPUs for improved numerical performance.
6. If necessary, after convergence the transportation plans can be computed as

$$
\pi_{rs}^m := \theta_{rs}^m + \frac{p_r - \mathfrak{p}_r^m}{S^m} + \frac{q_s^m - \mathfrak{q}_s^m}{R} + \frac{\gamma^m}{R S^m} \quad \forall\, m, r, s.
$$

Algorithm 1 Method of Averaged Marginals - Revisited

Input: Let initial point $p^0 \in H$, scalar $\rho > 0$, probability vectors $q^m \in \Delta_{S^m}$, and cost matrices $c^m \in \mathbb{R}^{R \times S^m}$, $m = 1, \dots, M$, be given

Initialization: Define $a_m := (\frac{1}{S^m})/(\sum_{j=1}^M \frac{1}{S^j})$ and set $\mathfrak{p}^m \leftarrow p^0$ for all $m = 1, \dots, M$. Initialize $\theta^m_{rs} \leftarrow \frac{p^0_r}{S^m} + \frac{q^m_s}{R} - \frac{1}{RS^m}$ for all m, r, s

while not converged **do**

$\quad p \leftarrow \mathtt{Proj}_{X \cap H}(\sum_{m=1}^M a_m \mathfrak{p}^m)$

$\quad$ **for** $m = 1, \dots, M$ **do**

$\quad\quad \mathfrak{q}^m \leftarrow (\theta^m)^\top \mathbf{1}_R$

$\quad\quad \gamma^m \leftarrow (\mathfrak{q}^m)^\top \mathbf{1}_{S^m} - 1$

$\quad\quad$ **for** $r = 1 \dots, R$ **do**

$\quad\quad\quad$ **for** $s = 1 \dots, S^m$ **do**

$\quad\quad\quad\quad \beta \leftarrow \frac{p_r - \mathfrak{p}^m_r}{S^m} + \frac{q^m_s - \mathfrak{q}^m_s}{R} + \frac{\gamma^m}{RS^m}$

$\quad\quad\quad\quad \theta^m_{rs} \leftarrow \max\{\theta^m_{rs} + \beta - c^m_{rs}/\rho, -\beta\}$

$\quad\quad\quad$ **end for**

$\quad\quad$ **end for**

$\quad\quad \mathfrak{p}^m \leftarrow \theta^m \mathbf{1}_{S^m}$

$\quad$ **end for**

end while

return p

5 Heuristic

Juan, don't you worry about this section right now. Later on we may revisit this part with constraints on π^m.

Let $f(\pi) = \langle c, \pi \rangle + \mathbf{i}_{\mathbb{R}_+^{R \times S}}(\pi)$ and observe that the second step in the Douglas-Rachford scheme (9) reads as

$$\hat{\pi}^{k+1} = \mathtt{Prox}_{\frac{1}{\rho}f}(2\mathtt{Proj}(\theta^k) - \theta^k) = \mathtt{Prox}_{\frac{1}{\rho}f}(2(\theta^k + \beta^k) - \theta^k) = \mathtt{Prox}_{\frac{1}{\rho}f}(\theta^k + 2\beta^k),$$

with

$$\beta^k = \mathtt{Proj}(\theta^k) - \theta^k = -(\theta^k - \mathtt{Proj}(\theta^k)) = -\frac{1}{2}\nabla \mathtt{d}^2(\theta^k, \mathcal{K}),$$

where $\mathtt{d}^2(x, \mathcal{K}) = \min_{y \in \mathcal{K}} \|y - x\|^2$ is the squared distance from x to $\mathcal{K}$. Therefore,

$$\hat{\pi}^{k+1} = \mathtt{Prox}_{\frac{1}{\rho}f}(\theta^k - \nabla \mathtt{d}^2(\theta^k, \mathcal{K})) = \mathtt{Prox}_{\frac{1}{\rho}f}\left(\theta^k - \frac{1}{\rho}\nabla[\rho \mathtt{d}^2(\theta^k, \mathcal{K})]\right),$$

which is a proximal-gradient step (with parameter $t = \frac{1}{\rho}$) applied to the problem $\min_\pi f(\pi) + \rho \mathtt{d}^2(\theta, \mathcal{K})$. To accelerate the convergence of Algorithm 1 in practice, we propose increasing this step size while keeping ρ fixed. Specifically, for any $\lambda \geq 1$, we define $\hat{\pi}^{k+1}_\lambda$ as a proximal-gradient step with parameter $t = \frac{\lambda}{\rho}$:

$$\begin{aligned}\hat{\pi}^{k+1}_\lambda &= \mathtt{Prox}_{\frac{\lambda}{\rho}f}\left(\theta^k - \frac{\lambda}{\rho}\nabla[\rho \mathtt{d}^2(\theta^k, \mathcal{K})]\right) = \mathtt{Prox}_{\frac{\lambda}{\rho}f}(\theta^k - \lambda \nabla \mathtt{d}^2(\theta^k, \mathcal{K})) \\ &= \mathtt{Prox}_{\frac{\lambda}{\rho}f}(\theta^k + 2\lambda\beta^k).\end{aligned}$$

Given the definition of f, we have that

$$\hat{\pi}^{k+1}_\lambda = \max\left\{0,\ \theta^k + 2\lambda\beta^k - \frac{\lambda}{\rho}c\right\},$$

where the maximum is to be understood componentwise. Having computed this point, we update the consecutive iterate as follows

$$\theta^{k+1} = \theta^k + \hat{\pi}^{k+1}_\lambda - \mathtt{Proj}(\theta^k), \tag{12}$$

which boils down to the last step in the Douglas-Rachford scheme (9) provided $\lambda = 1$. By expanding the above equation we get

$$\begin{aligned}\theta^{k+1} &= \theta^k + \max\left\{0,\ \theta^k + 2\lambda\beta^k - \frac{\lambda}{\rho}c\right\} - (\theta^k + \beta^k)\\ &= \max\left\{\theta^k + (2\lambda - 1)\beta^k - \frac{\lambda}{\rho}c,\ -\beta^k\right\}\\ &= \max\left\{\theta^k + \frac{(1-2\lambda)}{2}\nabla \mathtt{d}^2(\theta^k, \mathcal{K}) - \frac{\lambda}{\rho}c,\ \frac{1}{2}\nabla \mathtt{d}^2(\theta^k, \mathcal{K})\right\}.\end{aligned}$$

Again, the choice $\lambda = 1$ gives Algorithm 1, which is the DR scheme (9) after simplifications. Choosing $\lambda > 1$ can precludes convergence (at least for every initial point θ^0), meaning that the above fixed-point iteration may never converges (the underlying operator fails to be non-expansive). However, if this scheme converges to a fixed point $\bar\theta$, then $\mathtt{Proj}(\bar\theta)$ solves (8).

Proposition 2. *Let $\lambda \geq 1$ and consider the fixed-point iteration* (12)*. Suppose that the sequence $\{\theta^k\}$ converges to a fixed point $\bar\theta$ (which is certainly the case when $\lambda = 1$). Then $\mathtt{Proj}(\bar\theta)$ solves* (8)*.*

Proof. We start by noting that for all k, the optimality condition of $\pi^{k+1} = \mathtt{Proj}_{\mathcal{K}}(\theta^k)$ gives $0 \in \pi^{k+1} - \theta^k + N_{\mathcal{K}}(\pi^{k+1})$, i.e.,

$$\frac{\nabla \mathtt{d}^2(\theta^k)}{2} = \theta^k - \pi^{k+1} \in N_{\mathcal{K}}(\pi^{k+1}) = N_{\mathcal{K}}(\mathtt{Proj}_{\mathcal{K}}(\theta^k)).$$

Thus, at a fixed point $\bar\theta$, we have that $\frac{\nabla \mathtt{d}^2(\bar\theta)}{2} \in N_{\mathcal{K}}(\mathtt{Proj}_{\mathcal{K}}(\bar\theta))$. Moreover, equation (12) gives $\mathtt{Proj}(\bar\theta) = \hat\pi_\lambda$. As $\bar\beta = \mathtt{Proj}(\bar\theta) - \bar\theta$ and $\hat\pi_\lambda = \mathtt{Prox}_{\frac{\lambda}{\rho}f}(\bar\theta + 2\lambda\bar\beta)$, we have that

$$\begin{aligned}0 &\in c + \frac{\rho}{\lambda}\Big(\hat\pi_\lambda - (\bar\theta + 2\lambda\bar\beta)\Big) + N_{\mathbb{R}_+^{R\times S}}(\hat\pi_\lambda)\\ &= c + \frac{\rho}{\lambda}\Big(\bar\theta + \bar\beta - (\bar\theta + 2\lambda\bar\beta)\Big) + N_{\mathbb{R}_+^{R\times S}}(\hat\pi_\lambda)\\ &= c + \frac{\rho}{\lambda}(1-2\lambda)\bar\beta + N_{\mathbb{R}_+^{R\times S}}(\hat\pi_\lambda)\\ &= c + \frac{\rho}{\lambda}(2\lambda - 1)\frac{\nabla \mathtt{d}^2(\bar\theta, \mathcal{K})}{2} + N_{\mathbb{R}_+^{R\times S}}(\hat\pi_\lambda)\\ &\in c + N_{\mathcal{K}}(\mathtt{Proj}(\bar\theta)) + N_{\mathbb{R}_+^{R\times S}}(\hat\pi_\lambda),\end{aligned}$$

because $\frac{\nabla \mathtt{d}^2(\bar\theta,\mathcal{K})}{2} \in N_{\mathcal{K}}(\mathtt{Proj}(\bar\theta)) = N_{\mathcal{K}}(\bar\pi_\lambda)$. Hence, as both the nonempty polyhedron $\mathcal{K}$ and $\mathbb{R}_+^{R\times S}$ have nonempty relative interiors, we conclude that

$$0 \in c + N_{\mathcal{K}}(\mathtt{Proj}(\bar\theta)) + N_{\mathbb{R}_+^{R\times S}}(\mathtt{Proj}(\bar\theta)) = c + N_{\mathcal{K}\cap\mathbb{R}_+^{R\times S}}(\mathtt{Proj}(\bar\theta)),$$

i.e., $\mathtt{Proj}(\bar\theta)$ solves (8). $\square$

Given this result, we propose to pick up $\lambda \geq 1$ and replace line 11 in Algorithm 1 with

$$\theta^m_{rs} \leftarrow \max\left\{\theta^m_{rs} + (2\lambda - 1)\beta - \frac{\lambda}{\rho}c^m_{rs},\ -\beta\right\}.$$

The resulting scheme is a heuristic for $\lambda > 1$, but works remarkably well in our experiments for moderate values (e.g. $\lambda = 40$ or less).

6 Numerical results

Preliminary results for computing unconstrained WBS on 30 images of size 60×60 are shown in Figure 1. We compare our algorithm (MAM-R) and heuristic (MAM-H) with IBP and MAM. All the four algorithms run up to 300 seconds and use GPUs - MAM has a few operations that can run on GPU. Except for IBP, the other three solvers were initialized with $p^0 = \mathbf{1}_R/R$ and θ^0 was computed as in the initialization step of Algorithm 1.

After running 5 minutes, MAM produced 946 iterations while MAM-R and MAM-H performed around 6500 iterations. Therefore, in terms of iteration time, our approaches are almost 6.9 times faster than MAM. Such a speed clearly reflects on the quality of the results, as presented by evolution along time depicted in Figure 1. Figure 2 presents the results after 5 minutes of processing. The rightmost image in that figure

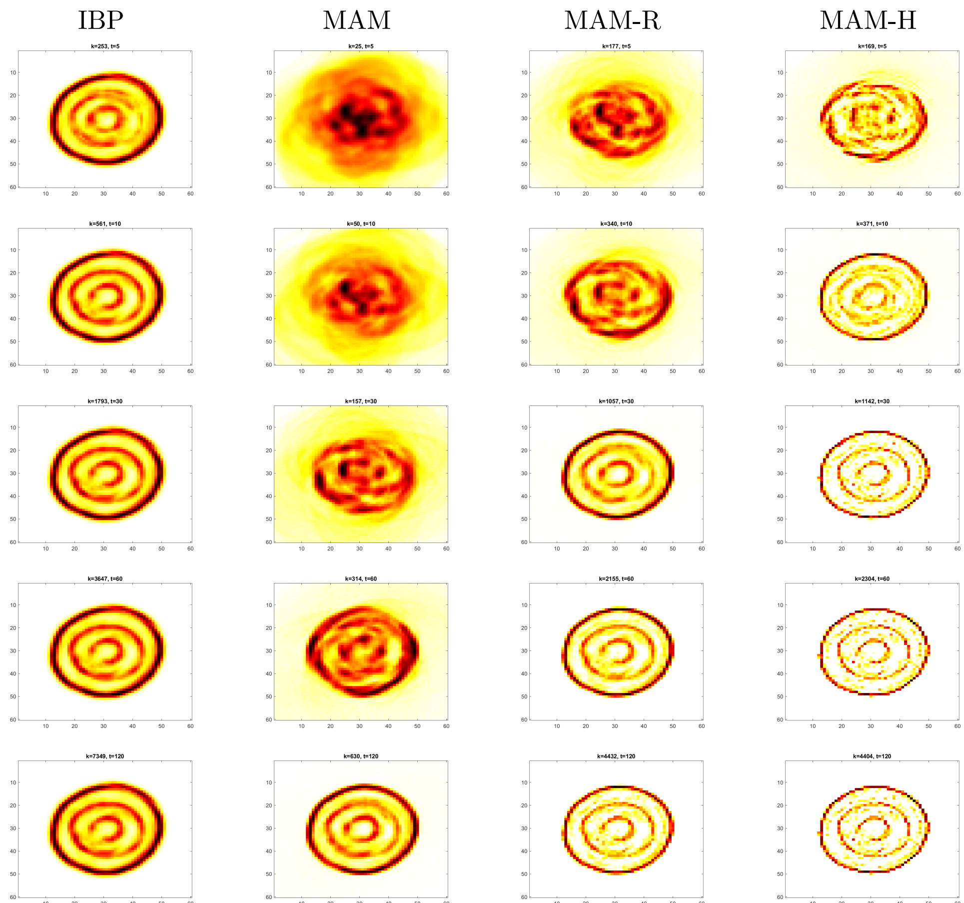

Figure 1: Evolution of computed barycenter approximations along time. IBP with $\lambda = 520$ and MAM-H with $\lambda = 40$. MAM, MAM-R, and MAM-H with $\rho = 350$.

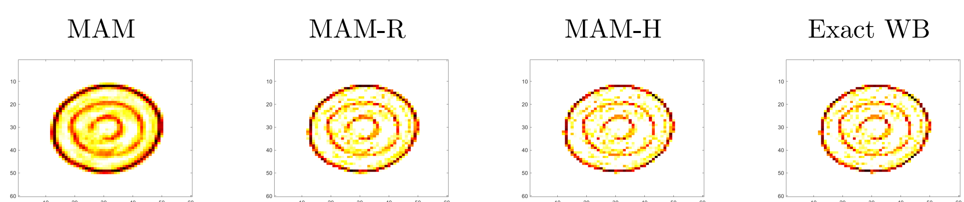

Figure 2: Barycenter provided by the methods MAM, MAM-R, and MAM-H after five minutes of processing. The rightmost figure is an exact solution computed by Gurobi.

corresponds to an exact barycenter, computed by Gurobi. Observe that practically the same solution was computed by our heuristic after 30 seconds. MAM-R computed a comparable solution after 270 seconds, while MAM was still far from that point after 300 seconds.

Figure 3 depicts the result for the free-support test problem of Altschuler XXX. Again, all the solvers were initialized with $p^0 = \mathbf{1}_R/R$. Different from the previous experiments which were performed in GPUs, all the computations in Figure 3 were done in CPUs, as the problem's data exceeds the memory capacity of our available GPUs.

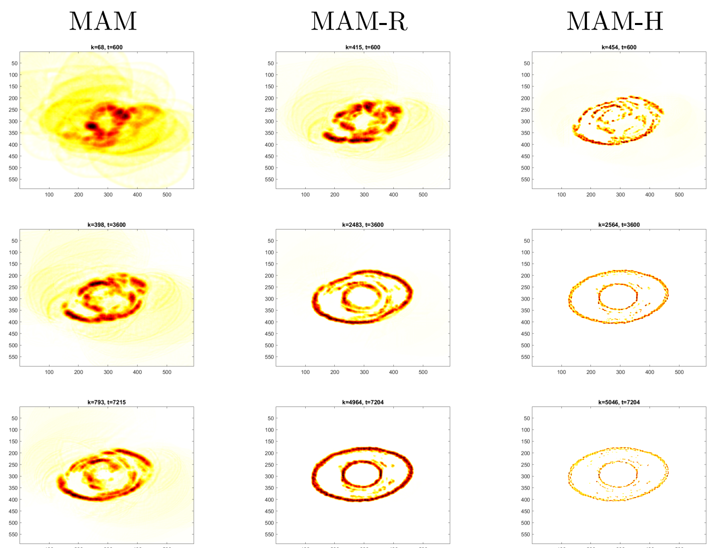

Figure 3: Evolution of computed barycenter approximations along time. The three solvers employ $\rho = 5\,000$, and MAM-H uses $\lambda = 40$.

References

[1] E. Anderes, S. Borgwardt, and J. Miller. Discrete Wasserstein barycenters: optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84(2):389–409, Oct 2016.

[2] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, 2nd edition, 2017.

[3] H. H. Bauschke, S. Singh, and X. Wang. Projecting onto rectangular matrices with prescribed row and column sums. *Fixed Point Theory and Algorithms for Sciences and Engineering*, 2021(1):23, Dec 2021.

[4] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American Mathematical Society*, 82(2):421–439, 1956.

[5] J. Eckstein and D. P. Bertsekas. On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, apr 1992.

[6] D. W. Mimouni, W. de Oliveira, and G. M. Sempere. On the computation of constrained wasserstein barycenters. *Pacific Journal of Optimization*, 21(4):787–808, 2025.

[7] D. W. Mimouni, P. Malisani, J. Zhu, and W. de Oliveira. Computing Wasserstein barycenters via operator splitting: The method of averaged marginals. *SIAM Journal on Mathematics of Data Science*, 6(4):1000–1026, 2024.

[8] D. Romero. Easy transportation-like problems on K-dimensional arrays. *Journal of Optimization Theory and Applications*, 66(1):137–147, Jul 1990.