ON THE COMPUTATION OF CONSTRAINED WASSERSTEIN BARYCENTERS

Daniel Mimouni
Mines Paris, Université PSL, Centre de Mathématiques Appliquées (CMA), Sophia Antipolis, France
Department de Mathématiques Appliquées, IFP Energies nouvelles, Rueil-Malmaison, France

Welington de Oliveira[1], Gregorio M. Sempere
Mines Paris PSL, CMA

Abstract: This work presents two optimization methods to compute, subject to constraints, a Wasserstein barycenter (WB) of finitely many empirical probability measures. The new measure, denoted by constrained Wasserstein barycenter, extends the applicability of the standard WB to pre-required geometrical or statistical constraints. Our first approach is an extension of the Method of Averaged Marginals (Mimouni et al., 2024) to compute WBs subject to convex constraints. In the nonconvex setting, we propose an optimization model whose necessary optimality conditions are written as a linkage problem with non-elicitable monotonicity. To solve such a linkage problem, we combine the Progressive Decoupling Algorithm (Rockafellar, 2019) with Difference-of-Convex programming techniques. We give the mathematical properties of our approaches and evaluate their numerical performances in two applications, demonstrating both their computational efficiency and the practical relevance of constrained Wasserstein barycenters.

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

Dedicated to Professor R. Tyrrell Rockafellar on the occasion of his 90th birthday

1 Introduction

Let $\mathcal{P}(\mathbb{R}^d)$ be the set of Borel probability measures on $\mathbb{R}^d$. A *Wasserstein barycenter* (WB) 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

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \frac{1}{M} \sum_{m=1}^{M} W_2^2(\mu, \nu^m)\,, \tag{1}$$

where $W_2(\mu, \nu)$ is the (quadratic) 2-Wasserstein distance between two measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$; see (3) below. Informally, a WB is a measure in $\mathcal{P}(\mathbb{R}^d)$ such that the total cost for transporting it to all ν^m is minimal concerning the 2-Wasserstein distance. A WB exists in generality and, if one of the ν^m vanishes on all Borel subsets of Hausdorff dimension $d-1$, then it is also unique [1]. Due to its ability to aggregate and summarize probability measures while preserving spatial characteristics, the concept of Wasserstein barycenter has gained prominence across diverse applications, ranging from applied probability, passing through imaging to machine learning [8].

When the measures to be summarized have finite supports, i.e., empirical measures, problem (1) can be reformulated as a linear programming (LP) problem. However, the size of this LP problem increases dramatically, scaling exponentially with the number of measures. As a result, it can quickly exceed the capabilities of standard LP solvers, even when dealing with a small number of measures [2, 7]. Therefore, specialized methods exploiting the problem's structure must come into play. While exact techniques usually build upon linear programming techniques [23, 6, 18], inexact approaches tackle (1) via reformulations based on an entropic regularization [10, 17, 5, 21]. A new technique leveraging the Douglas-Rachford splitting method is proposed in [20], and asymptotically computes an exact solution to (1).

While the dedicated literature on numerical methods for computing WBs focuses mostly on the unconstrained setting, that is, μ can be freely chosen in the space $\mathcal{P}(\mathbb{R}^d)$, many practical situations require the

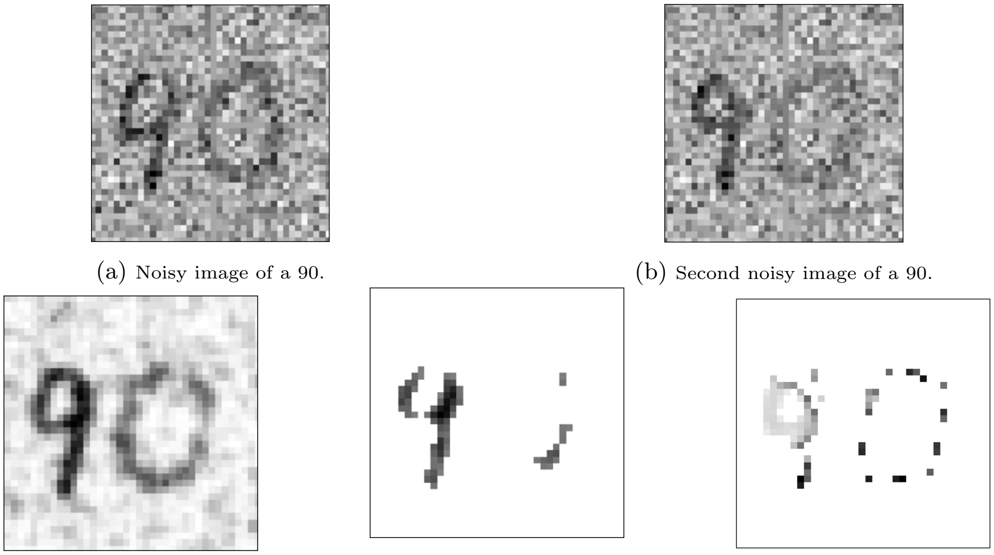
[1] Corresponding author: welington.oliveira@minesparis.psl.eu

target barycenter to satisfy certain constraints, ensuring it belongs to a predefined closed set $\mathfrak{X}$. Mathematically, the problem of computing a *constrained Wasserstein barycenter* (CWB) can be formulated as

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

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 feasibility. Hence, the optimization problem (2) offers greater modeling flexibility to tackle WB applications. For instance, the work [27] employs CWBs in image morphing applications. The authors consider variants of problem (2) with $\mathfrak{X}$ being manifolds, modeling sparsity or generative adversarial network (GAN)-based representations. They show that when compared unconstrained WBs, model (2) provides superior results for tasks like natural image morphing, offering smooth, visually plausible transitions without introducing artifacts. By leveraging priors like GANs, [27] handles nonconvex constraints with a heuristic inspired by the ADMM algorithm.

Another application where a constrained barycenter is sought arises when summarizing images by restricting the total number of pixels that can have nonzero mass. In such an application, a black-and-white image can be associated with a probability measure: the pixels (positions) constitute the measure's support, while the intensity of each pixel represents the measure's (probability) mass. Figure 1c shows an unconstrained WB of the two top images 1a and 1b, computed by solving the unconstrained problem (1). Those images are of dimension 40 × 40, i.e., they have 1600 pixels. To compute a summarized picture with at most 80 pixels with non-zero mass, one may think of defining $\mathfrak{X}$ as being the set of 40 × 40 images with such a property and project the WB of Figure 1c onto $\mathfrak{X}$. Such a naive approach gives the image in Figure 1d, which does not keep the entire relevant information. On the other hand, by considering (a model for) the constrained problem (2) with such a set $\mathfrak{X}$ and applying the algorithm we present in Section 4, we get the image depicted in Figure 1e.



(a) Noisy image of a 90.

(b) Second noisy image of a 90.

(c) Unconstrained barycenter.

(d) Unconstrained barycenter projected onto $\mathfrak{X}$.

(e) Constrained barycenter.

Figure 1: Unconstrained WB, projection of the unconstrained WB, and sparse barycenter of the two noisy images 1a and 1b.

We note that (2) is not the only way to model restrictions in WB applications. Some authors have studied constrains on the transport plans to address specific applications, such as those arising in finance, martingale transport problems, and other domains (see, for instance, [16, 13, 4] and [22, §4.20 and §10.12]). In these contexts, methods based on iterative Bregman projections, such as those discussed in [22], have been adapted to handle the additional constraints by alternating projections. In either of these two constrained Wasserstein barycenter models, convexity of the additional constraints evidently plays an important role in numerical optimization. Many algorithms assume convexity, thereby simplifying the optimization model but reducing applicability in real-world tasks where nonconvex constraints are common. Furthermore, scalability and interpretability of solutions remain open challenges when dealing with high-dimensional or structured data.

1.1 Contributions and Organization

This work presents two approaches to tackle constrained Wasserstein barycenter problems. Our first contribution is the extension of the Method of Averaged Marginals (MAM) introduced in [20] to tackle problem (2). We show that our extension of MAM asymptotically computes an exact solution to (2) provided the constraints are convex. As the original method of [20], our algorithm copes with scalability issues and is memory efficient. In the nonconvex setting, our approach becomes an heuristic that works notably well in some applications, as evidenced by our numerical results.

To cope with nonconvexity in a mathematically sound approach, we propose a relaxed model for (2) based on the penalization of the squared distance from the nonconvex set. As we show in Section 4, our model consists of minimizing a nonsmooth Difference-of-Convex (DC) function over a linear subspace. Its necessary optimality conditions can be written as a linkage problem. However, since convexity cannot be elicited at any level, Rockafellar's compelling *Progressive Decoupling Algorithm* [24] (see also [25] and [29]) cannot be directly applied. Based on the recent work [28], we specialize the progressive decoupling strategy to the constrained WB setting to design an algorithm with convergence guarantees to solve such a linkage problem, computing thus points satisfying certain necessary optimality condition to our relaxed (penalized) model for (2). As a further contribution, we conduct experiments on several data sets from the literature to demonstrate the computational efficiency and accuracy of the new approaches.

This work is structured as follows. Section 2 provides some background material on Wasserstein barycenter problems. Section 3 extends the Method of Averaged Marginals (MAM) of [20] to tackle (2) when $\mathfrak{X}$ is convex. Then, numerical results are shown where our approach is experimented with convex and nonconvex sets. The precise case of nonconvex sets is addressed in Section 4, where we propose a relaxed model for (2) and present a progressive decoupling strategy with convergence guarantees. Finally, a numerical illustration comparing the approaches closes the work.

Notation. Given a matrix $A \in \mathbb{R}^{R\times S}$, we denote by $A_{:s} \in \mathbb{R}^R$ its s^{th} column. Given $R \in \mathbb{N}$ and $\tau > 0$, let $\mathbf{1}_R$ be the column vector of all ones of size R, $\Delta_R(\tau) := \{y \in \mathbb{R}^N_+ : y^\top \mathbf{1}_R = \tau\}$, and $\Delta_R := \Delta_R(1)$ the $(R-1)$-simplex. Given a vector $y \in \mathbb{R}^R$, the Euclidean projection of y onto a closed set $X \subset \mathbb{R}^R$ is denoted by $\mathtt{Proj}_X(y)$, which is the set of solutions to the problem $\mathtt{dist}^2_X(y) := \min_{x\in X} \frac{1}{2}\|x-y\|^2$. The indicator function of X is denoted by $\mathtt{i}_X(\cdot)$. 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:

$$W_2(\mu,\nu) := \left(\inf_{\pi\in U(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\xi-\zeta\|^2 d\pi(\xi,\zeta)\right)^{1/2}, \tag{3}$$

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., $W_2^2(\mu,\nu) := (W_2(\mu,\nu))^2$. As already mentioned in the Introduction, a Wasserstein barycenter of a set of M measures $\{\nu^1,\dots,\nu^M\}$ in $\mathcal{P}(\mathbb{R}^d)$ is a solution to problem (1). In this work, we are concerned 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

$$\mathtt{supp}(\nu^m) := \{\zeta^m_1,\dots,\zeta^m_{S^m}\}\,, \ \text{ probability mass by } \ q^m \in \Delta_{S^m}, \ \text{ and thus } \ \nu^m = \sum_{s=1}^{S^m} q^m_s \delta_{\zeta^m_s}. \tag{4}$$

It follows from the definition of the support of a measure ν^m that $\nu^m(\zeta^m_s) = q^m_s > 0$ for all $s = 1,\dots,S^m$. As computing a WB 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 [2] asserts that every solution $\bar{\mu}$ to (1) has support satisfying the following key inclusion:

$$\begin{aligned} \texttt{supp}(\bar{\mu}) \subset \Xi &:= \left\{\xi^1, \dots, \xi^R\right\} \qquad (5) \\ &:= \left\{ \frac{1}{M} \sum_{m=1}^{M} \zeta^m \,:\, \zeta^m \in \texttt{supp}(\nu^m),\ m = 1, \dots, M \right\}. \end{aligned}$$

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

$$\bar{\mu} = \sum_{r \in \left\{j : \bar{p}_j > 0\right\}} \bar{p}_r \delta_{\xi_r}. \qquad (6)$$

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

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

Hence, problem (1) can be reformulated as the following finite-dimensional LP:

$$\begin{cases} \displaystyle\min_{p \in \mathbb{R}^n,\ \pi \geq 0} & \displaystyle\sum_{m=1}^{M} \frac{1}{M} \sum_{r=1}^{R} \sum_{s=1}^{S^m} \|\xi_r - \zeta_s^m\|^2 \pi_{rs}^m \\ \text{s.t.} & (\pi^m)^\top \mathbf{1}_R = q^m, \quad m = 1, \dots, M \\ & \pi^m\, \mathbf{1}_{S^m} = p, \quad m = 1, \dots, M. \end{cases} \qquad (7)$$

Note that if a pair (p, π) is feasible to the above problem, then $p \in \Delta_R$ due to the fact that $\pi \geq 0$ and $q^m \in \Delta_{S^m}$ for all $m = 1, \dots, M$.

Our second observation is related to the constrained Wasserstein barycenter problem (2). Thanks to (5), imposing a constraint of the type $\mu \in \mathfrak{X}$ can be done by restricting p in (7) to a certain set $X \subset \mathbb{R}^R$ related to $\mathfrak{X}$. In other words, in the empirical setting, the constrained WB problem (2) can be alternatively written as follows, for a set $X \subset \mathbb{R}^R$ associated to $\mathfrak{X} \subset \mathcal{P}(\mathbb{R}^d)$:

$$\begin{cases} \displaystyle\min_{p \in X,\ \pi \geq 0} & \displaystyle\sum_{m=1}^{M} \frac{1}{M} \sum_{r=1}^{R} \sum_{s=1}^{S^m} \|\xi_r - \zeta_s^m\|^2 \pi_{rs}^m \\ \text{s.t.} & (\pi^m)^\top \mathbf{1}_R = q^m, \quad m = 1, \dots, M \\ & \pi^m\, \mathbf{1}_{S^m} = p, \quad m = 1, \dots, M. \end{cases} \qquad (8)$$

We highlight that this problem is solvable as long as X is closed and intersects the simplex Δ_R. Observe further that while (7) is always an LP, problem (8) can be a nonconvex nonlinear optimization problem depending on X. To give examples of how the set of constraints $\mathfrak{X}$ on the measure relates to the set of constraints X on the probability vector, suppose we impose Wasserstein barycenters to have expected value equal to a given $\bar{\xi} \in \mathbb{R}^d$. In this case,

$$\mathfrak{X} = \left\{\mu \in \mathcal{P}(\mathbb{R}^d) : \mathbb{E}_\mu[\xi] = \bar{\xi}\right\} \text{ and the corresponding set is } X = \left\{ p \in \mathbb{R}^R : \sum_{r=1}^{R} p_r \xi_r = \bar{\xi} \right\}.$$

Such a setting finds applications, for instance, in scenario tree reduction where one might be interested in assigning probabilities to a smaller scenario tree in order to minimize the sum of Wasserstein distances while preserving, in every subtree issued by a node, certain theoretical expected value; see [12, §3.1] and [19]. If, instead, we require μ to have a support size of at most $\mathbf{n} \geq 1$ atoms, then

$$\mathfrak{X} = \left\{\mu \in \mathcal{P}(\mathbb{R}^d) : |\texttt{supp}(\mu)| \leq \mathbf{n}\right\} \text{ and the corresponding set is } X = \left\{p \in \mathbb{R}^R : \|p\|_0 \leq \mathbf{n}\right\},$$

where $\|p\|_0$ counts the number of nonzero components of the vector p. This is the setting considered in Figure 1e.

3 Constrained Wasserstein Barycenter

In the unconstrained setting, the Method of Averaged Marginals (MAM) proposed in [20] solves the LP problem (7) by exploiting its particular structure and applying the Douglas-Rachford splitting method (DR)

[14, 15]. The resulting algorithm is memory efficiently, can run in a deterministic or randomized fashion, copes with scalability issues, and has convergence guarantees. It updates transportation plans by projecting (in parallel) several vectors of dimension R onto sets of the form of $\Delta_R(\tau) = \{y \in \mathbb{R}^R_+ : y^\top \mathbf{1}_R = \tau\}$, with given scalar $\tau > 0$. This task can be performed *exactly* and efficiently using specialized methods [9]. Once the transportation plans are updated, its marginals p^m, $m = 1, \ldots, M$ are easily computed and an estimation of the probability vector yielding a WB is computed by averaging these marginals. We refer the interested reader to [20, §5] to a thorough discussion about the method and its convergence analysis. In what follows, our goal is to extend MAM to compute constrained WBs.

3.1 The Method of Averaged Marginals for Constrained WB

This section features our first contribution: the extension of MAM to compute a solution $\bar{\mu}$ to problem (2). As discussed in Section 2, this amounts to compute a p-part solution to problem (8) and recover $\bar{\mu}$ as in (6). To this end, we make the following assumption.

Assumption 1. *The set $X \subset \mathbb{R}^R$ in* (8) *is closed, convex, and satisfies $X \cap \Delta_R \neq \emptyset$. Furthermore, the Euclidean projection onto it is convenient to execute.*

We recall that closeness of X and condition $X \cap \Delta_R \neq \emptyset$ are enough to ensure that (8) is solvable. Next, by denoting

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

we drop the decision variable p in (8) and rewrite the problem in the following compact form:

$$\min_{\pi \in \mathcal{B}_X} \; \langle c, \pi\rangle \quad \text{s.t.} \quad \pi^m \in \Pi^m, \quad m = 1, \ldots, M, \tag{9a}$$

where

$$\Pi^m := \Big\{\pi^m \geq 0 : \, (\pi^m)^\top \mathbf{1}_R = q^m\Big\}, \; m = 1, \ldots, M, \tag{9b}$$

and

$$\mathcal{B}_X := \Big\{\pi = (\pi^1, \ldots, \pi^M) : \; \pi^1 \mathbf{1}_{S^1} = \pi^2 \mathbf{1}_{S^2} = \cdots = \pi^M \mathbf{1}_{S^M} \in X\Big\}. \tag{9c}$$

Thus, once problem (9) is solved, we can easily recover a p-solution to problem (8) and, as a consequence, a constrained WB measure $\bar{\mu}$.

To solve (9), we follow the lead of [20] and employ the DR algorithm, which asymptotically computes a solution by repeating the following steps, with $k = 0, 1, \ldots$ and given initial point $\theta^0 = (\theta^{1,0}, \ldots, \theta^{M,0})$ and prox-parameter $\rho > 0$:

$$\left\{\begin{array}{lcl} \pi^{k+1} & = & \mathtt{Proj}_{\mathcal{B}_X}(\theta^k) \\ \hat{\pi}^{k+1} & = & \arg\min\limits_{\pi \in \Pi} \langle c, \pi\rangle + \dfrac{\rho}{2}\|\pi - (2\pi^{k+1} - \theta^k)\|^2 \\ \theta^{k+1} & = & \theta^k + \hat{\pi}^{k+1} - \pi^{k+1}, \end{array}\right. \tag{10}$$

with $\Pi := \Pi^1 \times \ldots \times \Pi^M$. Assumption 1 ensures that the functions above are proper, convex, lower-semicontinuous functions and problem (9) is solvable. The following is a direct consequence of Theorem 25.6 and Corollary 27.4 of [3].

Theorem 1. *Under Assumption 1, the sequence $\{\theta^k\}$ produced by the DR algorithm* (10) *converges to a point $\bar{\theta}$, and the following holds: $\bar{\pi} := \mathtt{Proj}_{\mathcal{B}_X}(\bar{\theta})$ solves* (9)*, and $\{\pi^k\}$ and $\{\hat{\pi}^k\}$ converge to $\bar{\pi}$.*

The DR algorithm is attractive when the two first steps in (10) are convenient to execute, which is the case in our setting. Indeed, the following result extracted from [20, Prop. 5.2] asserts that the second step amounts to perform projections onto $\Delta_R(q^m_s)$, for $s = 1, \ldots, S^m$ and $m = 1, \ldots, M$.

Proposition 2 ([20], Prop. 5.2)**.** *The minimization $\hat{\pi} := \arg\min_{\pi \in \Pi} \langle c, \pi\rangle + \frac{\rho}{2}\|\pi - y\|^2$ can be performed exactly, in parallel along the columns of each transport plan y^m, as*

$$\hat{\pi}^m_{:s} = \mathtt{Proj}_{\Delta_R(q^m_s)}\Big(y^m_{:s} - \frac{1}{\rho}c^m_{:s}\Big), \tag{11}$$

for all $m = 1, \ldots, M$ and $s = 1, \ldots, S^m$.

The following original result, which does not require X to be convex, shows that the first step in (10) is simple provided the projection onto X is convenient to execute.

Proposition 3. *Let $\theta \in \mathbb{R}^{R\times(S^1+\cdots+S^M)}$, $a_m := \frac{\frac{1}{S^m}}{\frac{1}{S^1}+\cdots+\frac{1}{S^M}}$, and $p^m := \theta^m \mathbf{1}_{S^m}$, $m = 1,\dots,M$. Given a nonempty and closed set $X \subset \mathbb{R}^R$, let $p \in \mathtt{Proj}_X(\sum_{m=1}^M a_m p^m)$ and $\mathcal{B}_X$ given in (9c). Then, an element $\pi \in \mathtt{Proj}_{\mathcal{B}_X}(\theta)$ has the form*

$$\pi^m_{:s} = \theta^m_{:s} + \frac{(p - p^m)}{S^m}, \tag{12}$$

for all $m = 1,\dots,M$ and $s = 1,\dots,S^m$.

Proof. Given an arbitrary $w \in \mathbb{R}^R$, let us define the set

$$\mathcal{B}_w := \left\{\pi = (\pi^1,\dots,\pi^M) :\ \pi^1 \mathbf{1}_{S^1} = \pi^2 \mathbf{1}_{S^2} = \cdots = \pi^M \mathbf{1}_{S^M} = w\right\}.$$

Observe that $\mathcal{B}_w$ is nonempty[1] and $\mathcal{B}_X$ in (9c) can be written as $\mathcal{B}_X = \cup_{w\in X}\mathcal{B}_w$. Therefore, computing a point in $\mathtt{Proj}_{\mathcal{B}_X}(\theta)$ can be done by solving

$$\min_{y\in\mathcal{B}_X} \frac{1}{2}\|y-\theta\|^2 = \min_{w\in X,\, y\in\mathcal{B}_w} \frac{1}{2}\|y-\theta\|^2 = \min_{w\in X}\left\{\min_{y\in\mathcal{B}_w} \frac{1}{2}\|y-\theta\|^2\right\}.$$

The inner problem above is nothing but the projection onto $\mathcal{B}_w$. It can be written as

$$z = \mathtt{Proj}_{\mathcal{B}_w}(\theta) \ = \ \begin{cases} \arg\min\limits_{y} & \frac{1}{2}\sum\limits_{m=1}^{M} \|y^m - \theta^m\|^2 \\ \text{s.t.} & \sum_{s=1}^{S^m} y^m_{rs} - w_r = 0, \qquad r = 1,\dots,R,\ m = 1,\dots,M. \end{cases}$$

Being a solvable strongly convex quadratic program problem, the existence of Lagrange multipliers is ensured. As a result, the optimality conditions for this problem read as

$$(z^m_{rs} - \theta^m_{rs}) + \lambda^m_r = 0,\ \forall\, r,s,m \tag{13}$$

$$\sum_{s=1}^{S^m} z^m_{rs} = w_r,\ \forall\, r,m. \tag{14}$$

Note that summing equation (13) over $s = 1,\dots,S^m$, with r and m fixed, gives

$$\lambda^m_r = \frac{\sum_{s=1}^{S^m} z^m_{rs} - \sum_{s=1}^{S^m}\theta^m_{rs}}{S^m} = \frac{w_r - p^m_r}{S^m},\ \forall\, r,$$

where the last equality follows by (14) and definition of p^m. As a result, we conclude that $z = \mathtt{Proj}_{\mathcal{B}_w}(\theta)$ is given by

$$z^m_{rs} = \theta^m_{rs} + \frac{w_r - p^m_r}{S^m},\ \forall\, r,s,m. \tag{15}$$

Next, we show that when $w \in X$ is an element of $\mathtt{Proj}_X(\sum_{m=1}^M a_m p^m)$, then z above belongs to the set $\mathtt{Proj}_{\mathcal{B}_X}(\theta)$. Indeed,

$$\|z-\theta\|^2 = \sum_{m=1}^{M}\|z^m - \theta^m\|^2 = \sum_{m=1}^{M} S^m \|\frac{w-p^m}{S^m}\|^2 = \sum_{m=1}^{M}\frac{1}{S^m}\|w-p^m\|^2,$$

and thus

$$\begin{aligned}
\arg\min_{w\in X}\left\{\min_{y\in\mathcal{B}_w}\frac{1}{2}\|y-\theta\|^2\right\} &= \arg\min_{w\in X}\frac{1}{2}\sum_{m=1}^{M}\frac{1}{S^m}\|w-p^m\|^2 \\
&= \arg\min_{w\in X}\frac{1}{2}\sum_{m=1}^{M}\frac{\frac{1}{S^m}\|w-p^m\|^2}{\frac{1}{\sum_{j=1}^{M}\frac{1}{S^j}}} \\
&= \arg\min_{w\in X}\frac{1}{2}\sum_{m=1}^{M} a_m\|w-p^m\|^2 \\
&= \arg\min_{w\in X}\frac{1}{2}\sum_{m=1}^{M}\left\{a_m\|w\|^2 - 2a_m w^\top p^m + a_m\|p^m\|^2\right\} \\
&= \arg\min_{w\in X}\frac{1}{2}\left\{\|w\|^2\sum_{m=1}^{M}a_m - 2w^\top\left(\sum_{m=1}^{M}a_m p^m\right)\right\} \\
&= \mathtt{Proj}_X\left(\sum_{m=1}^{M}a_m p^m\right),
\end{aligned}$$

[1] The plans $\pi^m_{:1} = w$ and $\pi^m_{:s} = 0$ for $s \neq 1$ compose a point π that belongs to $\mathcal{B}_w$.

because $\sum_{m=1}^M a_m = 1$. This shows that the minimum value of $\min_{w\in X}\min_{y\in\mathcal{B}_w}\|y-\theta\|^2$ is reached at z given in (15) and $w \in \texttt{Proj}_X(\sum_{m=1}^M a_m p^m)$. The proof is thus complete. □

Recall that this proposition does not assume convexity of X. Hence, whether convexity is present or not, projecting onto $\mathcal{B}_X \subset \mathbb{R}^{R\times(S^1+\cdots+S^M)}$ is simple as long as the projection onto $X \subset \mathbb{R}^R$ is easy to perform. By relying on Propositions 2 and 3, we now gather and simplify the three steps of the DR Algorithm 10 to provide our extension of the MAM algorithm of [20] to the constrained setting. We start by invoking Proposition 2 that allows us to decompose the second step of DR Algorithm 10 into $\sum_{m=1}^M S^m$ simple projections: every column of $\hat{\pi}^{m,k+1}$ is given by

$$\hat{\pi}_{:s}^{m,k+1} = \texttt{Proj}_{\Delta_R(q_s^m)}\Big(2\pi_{:s}^{m,k+1} - \theta_{:s}^{m,k} - \frac{1}{\rho}c_{:s}^m\Big)\ \forall\, s, m.$$

It follows from Proposition 3, with $p^k \in \texttt{Proj}_X(\sum_{m=1}^m a_m p^{m,k})$, and $p^{m,k} = \theta^{m,k}\mathbf{1}_{S^m}$, that $2\pi_{:s}^{m,k+1} - \theta_{:s}^{m,k} = 2(\theta_{:s}^{m,k} + \frac{p^k - p^{m,k}}{S^m}) - \theta_{:s}^{m,k} = \theta_{:s}^{m,k} + 2\frac{p^k-p^{m,k}}{S^m}$. Therefore,

$$\hat{\pi}_{:s}^{m,k+1} = \texttt{Proj}_{\Delta_R(q_s^m)}\Big(\theta_{:s}^{m,k} + 2\frac{p^k - p^{m,k}}{S^m} - \frac{1}{\rho}c_{:s}^m\Big)\ \forall\, s, m.$$

Furthermore, the step $\theta^{k+1} = \theta^k + \hat{\pi}^{k+1} - \pi^{k+1}$ in the DR algorithm boils down to

$$\begin{aligned}\theta_{:s}^{m,k+1} &= \theta_{:s}^{m,k} + \texttt{Proj}_{\Delta_R(q_s^m)}\Big(\theta_{:s}^{m,k} + 2\frac{p^k - p^{m,k}}{S^m} - \frac{1}{\rho}c_{:s}^m\Big) - \theta_{:s}^{m,k} - \frac{p^k - p^{m,k}}{S^m}\\ &= \texttt{Proj}_{\Delta_R(q_s^m)}\Big(\theta_{:s}^{m,k} + 2\frac{p^k - p^{m,k}}{S^m} - \frac{1}{\rho}c_{:s}^m\Big) - \frac{p^k - p^{m,k}}{S^m},\ \forall\, s, m.\end{aligned}$$

Putting all together and removing the variables π^k and $\hat{\pi}^k$ we get the extension of MAM presented in Algorithm 1.

Algorithm 1 Method of Averaged Marginals for Constrained WBs

```
 1: Input: $M$ empirical probability measures $\nu^m \in \mathcal{P}(\mathbb{R}^d)$, initial plan $\theta^m \in \mathbb{R}^{R\times S^m}$, set $X$ of constraints,
    and a scalar $\rho > 0$

 2: Define $c_{rs}^m := \frac{1}{M}\|\xi_r - \zeta_s^m\|^2$, for all $r = 1,\ldots,R$, $s = 1,\ldots,S^m$ and $m = 1,\ldots,M$
 3: Set $a_m := (\frac{1}{S^m})/(\sum_{j=1}^M \frac{1}{S^j})$ and $p^m := \sum_{s=1}^{S^m}\theta_{rs}^m$, $m = 1,\ldots,M$

 4: while not converged do

 5:     $p \leftarrow \texttt{Proj}_X(\sum_{m=1}^M a_m p^m)$

 6:     for $m = 1,\ldots,M$ do
 7:         for $s = 1,\ldots,S^m$ do
 8:             $\theta_{:s}^m \leftarrow \texttt{Proj}_{\Delta_R(q_s^m)}\Big(\theta_{:s}^m + 2\frac{p-p^m}{S^m} - \frac{1}{\rho}c_{:s}^m\Big) - \frac{p-p^m}{S^m}$
 9:         end for
10:         $p^m \leftarrow \sum_{s=1}^{S^m}\theta_{:s}^m$
11:     end for

12: end while
13: return $p$
```

A possible manner to stop Algorithm 1 is when the barycenter estimate p stabilizes. As commented in [20, §5], this should be seen as a heuristic stopping test. In the next result, we index the variables p and θ in Algorithm 1 by k, which represents a pass between lines 4 and 12.

Theorem 4. *Under Assumption 1, the sequence $\{(p^k,\theta^k)\}$ produced by the Algorithm 1 converges to a point $(\bar{p},\bar{\theta})$, with $(\bar{p},\texttt{Proj}_{\mathcal{B}_X}(\bar{\theta}))$ solving* (8).

Proof. As presented, Algorithm 1 is the DR algorithm (10) applied to problem (9). As $X \cap \Delta_R \neq \emptyset$ in view of Assumption 1, problem (9) has a solution, and thus Theorem 1 ensures that: $\{\theta^k\}$ converges to a point $\bar{\theta}$, with $\texttt{Proj}_{\mathcal{B}_X}(\bar{\theta})$ a solution to (9). Proposition 3, with the additional assumption of convexity of X, asserts that

$$\bar{\pi}^m_{:s} = \bar{\theta}^m_{:s} + \frac{\bar{p} - \bar{p}^m}{S^m} \; \forall\, s, m,$$

solves (9), where $\bar{p}^m = \sum_{s=1}^{S^m} \bar{\theta}^m_{:s}$. As $\sum_{s=1}^{S^m} \bar{\pi}^m_{:s} = \bar{p}$, for all m, the relation between problems (9) ensures that $(\bar{p}, \bar{\pi})$ solves (8). □

Remark 1. *If $X = \mathbb{R}^R$, then Algorithm 1 boils down to the Method of Averaged Marginals of [20]. As in that paper, we can also randomize our algorithm by performing, at every iteration, the projections on line 8 only for a single measure m chosen at random. In this randomized setting, the convergence results above hold almost surely. This follows directly from the analysis in [20, Thm. 5.5].*

While Algorithm 1 provides an exact manner for computing a constrained Wasserstein barycenter under Assumption 1, its simplicity and decentralized nature encourage its application in the nonconvex setting. The key insight is that if one can efficiently compute projections onto X, our approach could be (heuristically) applied to problem (9) even when X is nonconvex. The following section provides numerical insights into what can be obtained by Algorithm 1 in the convex and nonconvex settings.

3.2 Some Numerical Insights

This subsection presents some practical cases of why constrained WBs are worth considering. We compare unconstrained WB, convex and nonconvex constrained WBs, all computed by Algorithm 1 with three different choices for X (being the first choice $X = \mathbb{R}^R$). In the first example, we consider a simple case of a transportation problem with limited storage locations. In the second instance, a toy problem of image processing is presented. The goal is to obtain a sparse barycenter that accurately represents the initial data.

3.2.1 Demand and Location Storage Localization

We consider a localization problem for demand and storage optimization. The dataset comprises demand maps of a certain product for Paris over a 12-month period (one map per month). Figure 2 shows a sample of nine out of the twelve months, with each map illustrating the aggregated product demand in red. The objective is to determine optimal storage locations to minimize distribution costs.

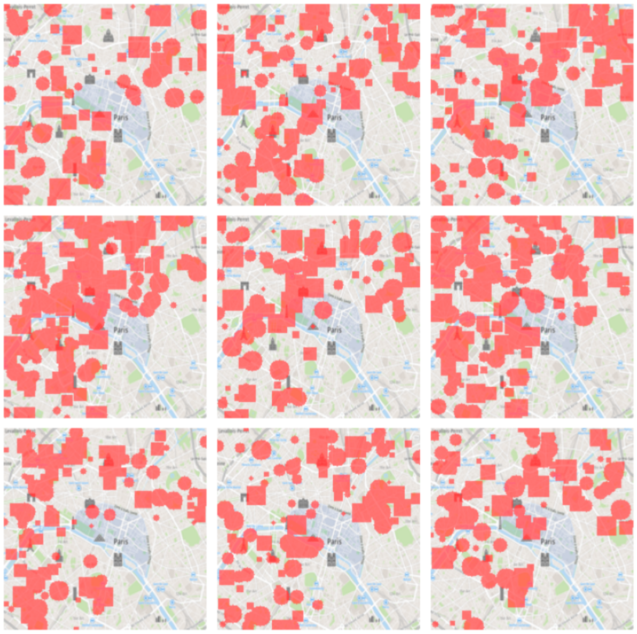
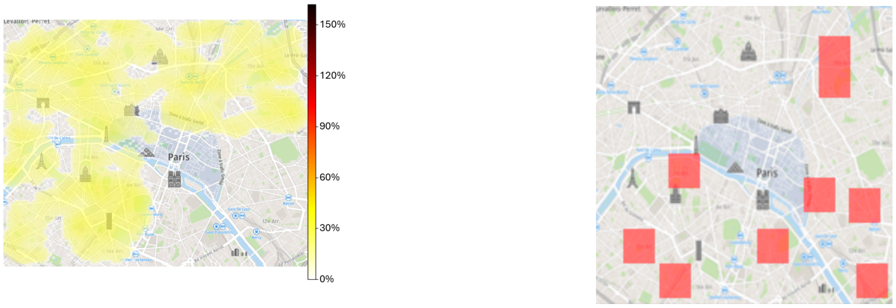


Figure 2: Sample of 9 (out of 12) months of collected demand data: the aggregated product demand is represented in red.

First, let us compute an unconstrained Wasserstein barycenter of the demand maps. Such a barycenter, presented in Figure 3a, suggests the need to rent 5726 storage facilities, which is impractical due to high costs in certain areas. Therefore, we restrict the storage to eight affordable locations $\mathcal{L}_\ell$ (marked in red in Figure 3b) with capacities $u_r > 0$ for all $r \in \left\{j = 1, \dots, R : \xi^j \in \cup_{\ell=1,\dots,8}\mathcal{L}_\ell\right\}$. As such restrictions yield probability (in this case demand) $p_r = 0$ for locations ξ^r outside the locations $\mathcal{L}_\ell$, the problem can be recast as a smaller Wasserstein problem with bound constraints given by the convex set:



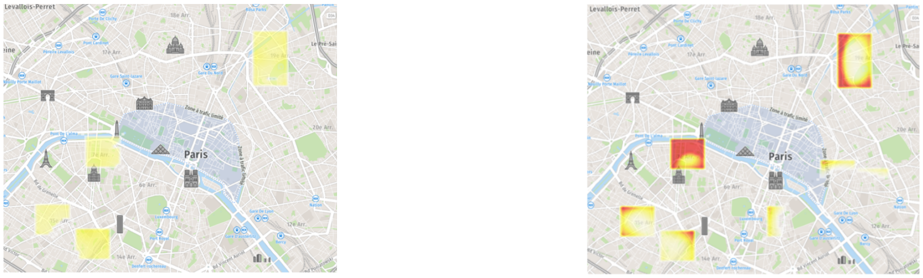
(a) Unconstrained WB with storage occupancy colorbar: 100% accounts for maximal u_r capacity being reached.

(b) Affordable areas are shown in red.

Figure 3: Unconstrained Barycenter of the demand maps and affordable storage locations.

$$X = \left\{ p:\ p_r \leq u_r, \quad \forall r \text{ s.t. } \xi^r \in \bigcup_{\ell=1,\dots,8} \mathcal{L}_\ell \right\}.$$

Projecting the unconstrained barycenter of Figure 3a (probability vector p^u) onto the affordable areas depicted in Figure 3b results in the map shown in Figure 4a. Unfortunately, this projection violates the probabilistic nature of the barycenter, as the projected solution satisfies only 7% of the total demand: $\texttt{Proj}_X(p^u)^\top \mathbf{1}_R = 0.07$. To address this mismatch, we integrate the affordability constraint directly into the optimization problem to define problem (8), whose solution computed by Algorithm 1 is shown in Figure 4b. This solution achieves 100% ($\sum_{r=1}^R p_r = 1$) demand fulfillment while requiring only 625 storage facilities. Notably, it identifies new high-utilization locations (e.g., three sites in the bottom-right corner with approximately 40% occupancy).



(a) Unconstrained WB projected onto the set X.

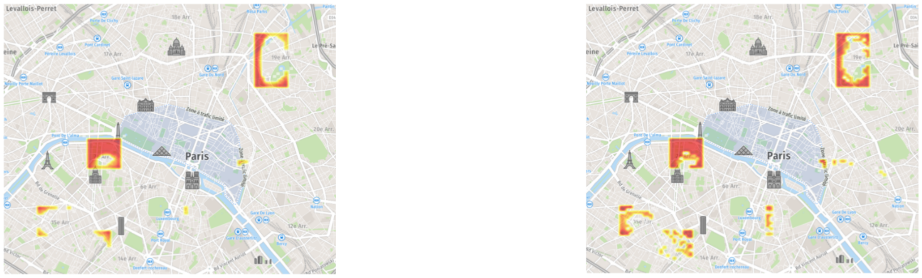
(b) Constrained WB computed by Algorithm 1.

Figure 4: Projected (unconstrained) WB versus constrained WB.

The (convex) constrained barycenter depicted in Figure 4b requires few storages in some locations $\mathcal{L}_\ell$. To maximize profitability, we introduce a nonconvex constraint that mandates a minimum storage utilization of 40%. The new constraint set is as follows:

$$X = \left\{ p:\ \sum_{r\in\{j:\xi^j\in\mathcal{L}_\ell\}} p_r \in \{0\} \cup [0.4u_\ell,\, u_\ell] \quad \ell = 1, \dots, 8 \right\}.$$

Projecting the computed (convex constrained) barycenter onto this nonconvex set results in the allocation shown in Figure 5a, fulfilling only 64% of the demand (the projected barycenter is not a probability measure). However, integrating this nonconvex constraint directly into the optimization problem yields a point (Figure 5b) that satisfies 100% of the demand using only 308 storage facilities, compared to the 625 required previously.



(a) Convex constrained WB projected onto the nonconvex set X.

(b) Nonconvex constrained barycenter computed by Algorithm 1.

Figure 5: Projected convex constrained WB versus nonconvex constrained WB.

In summary, in this application, integrating constraints directly into the optimization process consistently produces better results than applying constraints post hoc. By incorporating both convex and nonconvex constraints, we achieve a practical solution that balances demand fulfillment and storage cost efficiency. We remark that for the dataset composed by 12 images of size 100 × 100, Algorithm 1 computed each one of the above (unconstrained, convex and nonconvex constrained) WBs within a couple of minutes, with none taking significantly longer than the others.

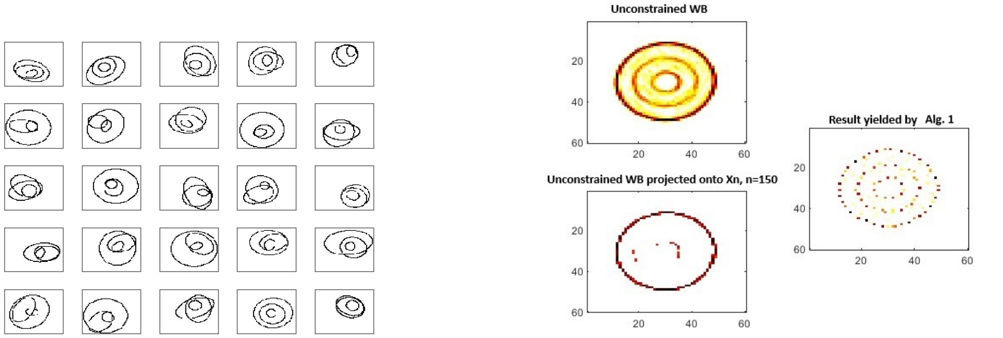
3.3 Sparse Barycenter to a Set of Images

This subsection continues the discussion on the experiments presented in Figure 5 and evaluates Algorithm 1 in the context of nonconvex, constrained Wasserstein barycenter problems. We consider two test cases. In the first case, we show that our algorithm performs effectively, even though it is a heuristic in the nonconvex setting. The second test case illustrates a situation where the outcome produced by our approach is not satisfactory. In both cases, we work with the nonconvex set forcing sparsity:

$$X_{\mathtt{n}} = \{p : \|p\|_0 \leq \mathtt{n}\} \, .$$

Here, $\mathtt{n}$ is a given natural number.

Ellipses: We consider a sample of $M = 100$ images of size $R = 60 \times 60$ with three nested ellipses. Figure 6a shows 25 of these images. A naive strategy to obtain a sparse image summarizing the dataset is first computing a unconstrained WB, and then projecting it onto $X_{\mathtt{n}}$. This is exemplified in Figure 6b (the two leftmost images) with $\mathtt{n} = 150$. This strategy is clearly unsatisfactory. On the other hand, Algorithm 1 applied with $X = X_{150}$ provides the rightmost image in Figure 6b, which clearly depicts three nested ellipses. In this test, problem (9) has the order of 10^7 decision variables, and we let Algorithm 1 run 5000 iterations, which took about five minutes on a personal computer.



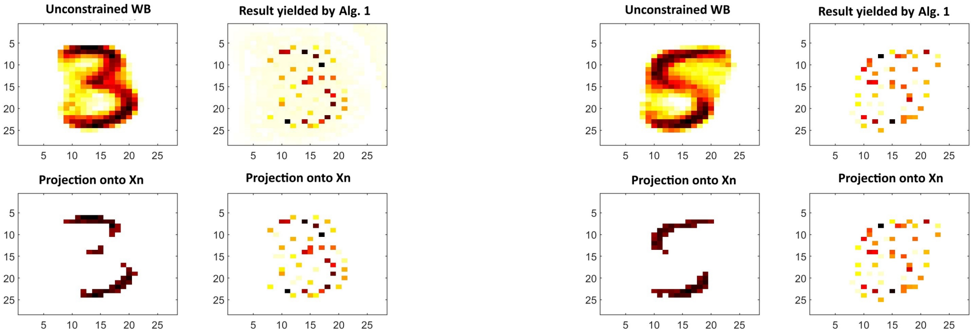
(a) Sample of 25 out of $M = 100$ images.

(b) Unconstrained WB, projection of the Unconstrained WB onto X_{150}, and image provided by Algorithm 1.

Figure 6: Sparse WB to a set of $M = 100$ images. The level of sparsity is chosen to be $\mathtt{n} = 150$.

MNIST: In this experiment, we use as initial input the well-known MNIST database, which includes $R = 28 \times 28$ images of handwritten numbers. By considering two samples of $M = 100$ images for the numbers three and five, we try to compute a sparse barycenter for each of the samples.

As in the previous example, we compare the computed unconstrained WB, its projection onto $X_{\mathtt{n}}$, and the point provided by Algorithm 1 with $X = X_{\mathtt{n}}$. In this experiment, problem (9) has the order of 10^6 decision variables, and we let Algorithm 1 run 5000 iterations, which took only a couple of minutes. While the unconstrained WB computed by MAM is meaningful, the sparse points provided by our approach fail to be probability vectors, and their quality in terms of visual meaning is poor. As shown in Figure 7, the projections of such points (including the unconstrained WB) onto $X_{\mathtt{n}}$ are not useful.



(a) Results for the number 3.

(b) Results for the number 5.

Figure 7: Unconstrained, sparse, and projection onto the nonconvex set $X_{\mathtt{n}}$, with $\mathtt{n} = 45$ for the digit three and $\mathtt{n} = 47$ for the digit five.

These experiments demonstrate that while Algorithm 1 converges in a convex setting, its use as a heuristic does not always yield satisfactory results. This observation supports the need for the mathematically sound approach of the next section.

4 A Progressive Decoupling Approach

Algorithm 1 is (asymptotically) exact when the set X is convex. However, the lack of convergence guarantees in the nonconvex case and the numerical illustration of the last section lead to question its use in some applications where a nonconvex set X is deemed indispensable. For this reason, we investigate in this section a penalized model for the nonconvex constrained Wasserstein problem (9) and a tailored solving methodology based on the *Progressive Decoupling Algorithm* (PDA) of Rockafellar [24]; see also [25, 29, 28].

Our initial, and unfruitful tentative, was to consider the following relaxed version of (9), with $\eta > 0$ a penalty parameter:

$$\min_{\pi} \langle c, \pi\rangle + \frac{\eta}{2}\mathtt{dist}^2_{\mathcal{B}_X}(\pi) \quad \text{s.t.} \quad \pi^m \in \Pi^m, \quad m = 1, \dots, M,$$

with $\mathtt{dist}^2_{\mathcal{B}_X}(\pi)$ the squared distance of π from $\mathcal{B}_X$. More precisely,

$$\mathtt{dist}^2_{\mathcal{B}_X}(\pi) = \min_{\theta\in\mathcal{B}_X} \frac{1}{2}\|\theta - \pi\|^2 = \frac{1}{2}\|\pi\|^2 - \max_{\theta\in\mathcal{B}_X}\left\{\langle\theta, \pi\rangle - \frac{1}{2}\|\theta\|^2\right\}$$

has a difference-of-convex (DC) structure: the two rightmost functions above are convex on variable π. As a result, the above penalized model is a DC programming problem for which specialized algorithms exist [11, 26]. However, in our experiments, the results provided by such a DC model were not particularly appealing. This is why we propose to add the convex constraints

$$\pi \in \mathcal{B} := \mathcal{B}_{\mathbb{R}^R} \quad \text{(see Eq. (9c))}$$

to our DC model. The resulting and more involving optimization problem reads as follows:

$$\min_{\pi\in\mathcal{B}} \langle c, \pi\rangle + \frac{\eta}{2}\mathtt{dist}^2_{\mathcal{B}_X}(\pi) \quad \text{s.t.} \quad \pi^m \in \Pi^m, \quad m = 1, \dots, M. \tag{16}$$

Observe that this problem consists of minimizing a DC function over a linear subspace $\mathcal{B}$. Hence, any solution $\bar\pi$ to (16) is accompanied with a dual variable $\bar y$ solving the *linkage problem* [24]:

$$\text{find } \bar\pi \in \mathcal{B} \text{ and } \bar y \in \mathcal{B}^{\perp} \text{ such that } \bar y \in T(\bar\pi), \tag{17}$$

with $T(\pi) := \partial[\langle c, x\rangle + \sum_m \mathtt{i}_{\Pi}(\pi^m)] + \eta\partial^{\mathrm{C}}\mathtt{dist}^2_{\mathcal{B}_X}(\pi)$, and ∂^{C} denoting the Clarke subdifferential. (This linkage problem is nothing but an alternative way to write the Lagrange system yielding a Clarke stationary point to (16).)

The work [24] investigates linkage problems and proposes the Progressive Decoupling Algorithm as a solving methodology. PDA solves the linkage problem should monotonicity of T be elicitable at a certain level. See also [29] and [25] for more details. However, being $\mathtt{dist}^2_{\mathcal{B}_X}$ a DC function, monotonicity of T above cannot be elicitable at any level, and thus the PDA of [24] is not directly applicable to our setting. We refer the interest reader to [28, § 2.5] for more details. However, we can exploit the DC structure of problem (16) using the method proposed in [28]. To this end, let us write the objective function of (16) as

$$f(\pi) - h(\pi), \quad \text{with} \quad \begin{cases} f(\pi) := \langle c, \pi\rangle + \frac{\eta}{2}\|\pi\|^2 \\ h(\pi) := \eta \max\limits_{\theta\in\mathcal{B}_X}\left\{\langle\theta, \pi\rangle - \dfrac{1}{2}\|\theta\|^2\right\}. \end{cases}$$

Recall that $\mathcal{B}^{\perp}$ is the normal cone (at every point) of the linear subspace $\mathcal{B}$. Furthermore, note that the subdifferential set $\partial^{\mathrm{C}}h(\pi)$ coincides with the projection $\eta\mathtt{Proj}_{\mathcal{B}_X}(\pi)$. The algorithm proposed in [28] linearizes h at a reference point π^{ℓ_k} (stability center) and defines a new stability center by inexactly solving the convex subproblem (see Algorithm 1 and Section 3.3 of [28]):

$$\min_{\pi\in\mathcal{B}} f(\pi) - \langle g^{\ell_k}, \pi\rangle + \frac{\mu}{2}\|\pi - \pi^{\ell_k}\|^2 \quad \text{s.t.} \quad \pi^m \in \Pi^m, \quad m = 1, \dots, M.$$

Here, $\mu > 0$ is a given parameter. Observe that such a subproblem is a quadratic variant of the unconstrained Wasserstein barycenter problem (7). To get around the practical inconvenience of solving difficult convex subproblems like this per iteration, the work [28] proposes to employ PDA with a safeguard permitting to stop the algorithm as soon as an incumbent point to (16) is found. In this work, the employed safeguard is a descent test accompanied by a penalty function associated to the constraints $\pi^m \in \Pi^m$. More specifically, we apply the PDA to the above subproblem to generate a sequence $\{\pi^k\}$ until a trial point π^{k+1} satisfying the following descent test is computed:

$$F(\pi^{k+1}) \le F(\pi^{\ell_k}) - \frac{\kappa}{2}v_k,$$

with $\kappa \in (0, \frac{1}{2})$,

$$F(\pi) := f(\pi) - h(\pi) + \mathtt{Penalty}\sum_m \mathtt{dist}_{\Pi^m}(\pi^m), \;\; \mathtt{Penalty} > 0,$$

and $v_k \ge 0$ defined in Algorithm 2 below. When such a descent test is satisfied, we halt PDA, set $\pi^{\ell_{k+1}} = \pi^{k+1}$, compute a new subgradient $g^{\ell_{k+1}} \in \eta\mathtt{Proj}_{\mathcal{B}_X}(\pi^{\ell_{k+1}})$ to define the next subproblem, and repeat the process. Such a procedure can also be viewed as an inexact DC algorithm, where the convex subproblem is addressed by PDA but not solved to optimality. Our approach to tackle the nonconvex Wasserstein barycenter model (16) is presented in Algorithm 2.

Algorithm 2 PROGRESSIVE DECOUPLING ALGORITHM FOR NONCONVEX WB PROBLEMS

```
1: **Input**: $M$ empirical probability measures $\nu^m \in \mathcal{P}(\mathbb{R}^d)$, initial primal and dual variables $\pi \in \mathcal{B}$ and
   $y^0 \in \mathcal{B}^{\perp}$, and scalars $\eta, \rho > 0$, $\kappa \in (0, \frac{1}{2})$, $\mu > 2\kappa$ and $\texttt{Penalty} > 0$

2: Define $c^m_{rs} := \frac{1}{M}\|\xi_r - \zeta^m_s\|^2$, for all $r = 1, \dots, R$, $s = 1, \dots, S^m$ and $m = 1, \dots, M$

3: Set $\ell_0 = 0$, $g^0 \in \eta\texttt{Proj}_{B_X}(\pi^0)$ and $a_m := (\frac{1}{S^m})/(\sum_{j=1}^{M} \frac{1}{S^j})$, $m = 1, \dots, M$

4: **while** not converged **do**
5:     $z^k \leftarrow c - y^k - g^{\ell_k} - \rho\pi^k - \mu\pi^{\ell_k}$
6:     **for** $m = 1, \dots, M$ **do**
7:         **for** $s = 1, \dots, S^m$ **do**
8:             $\hat{\pi}^{k,m}_{:s} \leftarrow \texttt{Proj}_{\Delta(q^m_s)}\left[-\frac{1}{\eta+\rho} z^{k,m}_{:s}\right]$        ▷ Projection onto the simplex
9:         **end for**
10:        $p^{k,m} \leftarrow \sum_{s=1}^{S^m} \hat{\pi}^{k,m}_{:s}$
11:    **end for**

12:    $p^{k+1} \leftarrow \sum_{m=1}^{M} a_m p^{k,m}$        ▷ Barycenter of the current iterate

13:    **for** $m = 1, \dots, M$ **do**
14:        **for** $s = 1, \dots, S^m$ **do**
15:            $\pi^{k+1,m}_{:s} \leftarrow \hat{\pi}^{k,m}_{:s} + \frac{p^{k+1} - p^{k,m}}{S^m}$        ▷ Projection onto $\mathcal{B}$
16:        **end for**
17:    **end for**

18:    $y^{k+1} \leftarrow y^k - \rho(\hat{\pi}^k - \pi^k)$
19:    $v_k \leftarrow \max\{\|\pi^{k+1} - \pi^{\ell_k}\|^2, \|y^{k+1} - y^k\|^2, \|\pi^{k+1} - \pi^k\|^2\}$

20:    **if** $F(\pi^{k+1}) \leq F(\pi^{\ell_k}) - \frac{\kappa}{2} v_k$ **then**
21:        $\ell_{k+1} \leftarrow k + 1$        ▷ Serious step
22:        $\bar{p}^{\ell_{k+1}} \leftarrow \texttt{Proj}_X(p^{\ell_{k+1}})$

23:        **for** $m = 1, \dots, M$ **do**
24:            **for** $s = 1, \dots, S^m$ **do**
25:                $g^{\ell_{k+1},m}_{:s} \leftarrow \eta\left(\pi^{\ell_{k+1},m}_{:s} + \frac{\bar{p}^{\ell_{k+1}} - p^{\ell_{k+1}}}{S^m}\right)$        ▷ A subgradient
26:            **end for**
27:        **end for**
28:    **else**
29:        $\ell_{k+1} \leftarrow \ell_k$        ▷ Null step
30:    **end if**

31: **end while**
32: **return** $p^{\ell_{k+1}}$
```

Remark 2. *A few comments on Algorithm 2 are in order.*

- *Although steps 8, 15 and 25 all require equivalent loops over m and s, these loops cannot be merged, as they require the previous computation of p^{k+1} and $\bar{p}^{\ell_{k+1}}$ in steps 12 and 22, respectively. However, all of those can be run in parallel over m and s, as none of the steps depend on a previous iteration of the loop.*
- *If the descent test in step 20 holds for a given iterate $\ell_{k+1} = k + 1$, the method requires computing a subgradient of h (equivalently a projection onto $\mathcal{B}_X$) at $\pi^{\ell_{k+1}}$. This step needs the partial sums over*

s of $\pi^{\ell_{k+1}}$. Note that the following relation holds

$$\sum_{s=1}^{S^m} \pi_{:s}^{\ell_{k+1},m} = \sum_{s=1}^{S^m} \hat{\pi}_{:s}^{k+1,m} + (p^{k+1} - p^{k,m}) = p^{k+1},$$

so the partial sums of $\pi^{\ell_{k+1}}$ (independent over m) are equal to p^{k+1}, which is already computed at line 12.

- *Algorithm 2 is a specialization of the method of [28] to our setting. In addition to the problem's structure exploitation, we have considered the simplifications discussed in Subsection 3.3 of [28].*

Theorem 5 ([28, Thm. 3.1])**.** *Consider sequences $\{\pi^k\}_k$ and $\{\pi^{\ell_k}\}_k$ computed by Algorithm 2:*

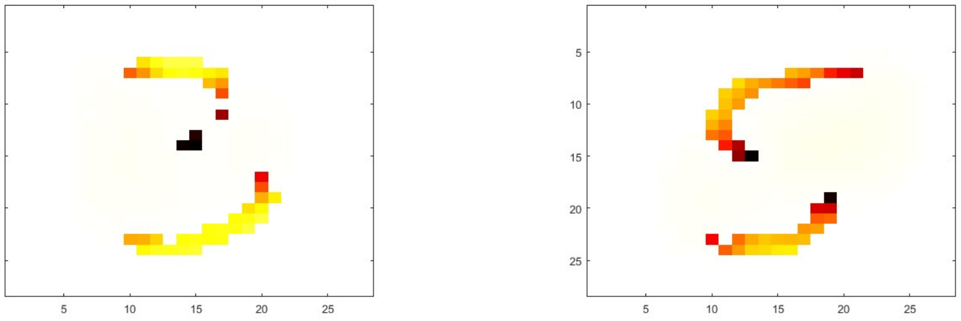
1. *If only $\ell := \ell_k$ serious steps are performed, then $\lim_k \pi^k = \tilde{\pi}$ where $\tilde{\pi}$ solves the problem*

$$\min_{\pi \in \mathcal{B}} f(\pi) - [h(\pi^\ell) + \langle g^\ell, \pi - \pi^\ell \rangle] + \frac{\mu}{2} \|\pi - \pi^\ell\|^2 \,, \quad s.t.\ \pi^m \in \Pi^m \ \forall m = 1, \ldots, M. \qquad (18)$$

Moreover, if `Penalty` *is big enough for* `Penalty` $\sum_m$ $\texttt{dist}_{\Pi^m}$ *to be an exact penalty to problem* (18)*, then $\tilde{\pi} = \pi^\ell$ solves the linkage problem* (17)*.*

2. *If Algorithm 2 performs infinitely many serious steps, then every cluster point of $\{\pi^{\ell_k}\}_k$ solves the linkage problem* (17)

Sparse Barycenter for MNIST: Let us return to the second test problem of Subsection 3.3, i.e, the MNIST dataset. As shown by Figure 7, Algorithm 1 failed to provide a satisfactory sparse WB when the nonconvex constraint $X_{\texttt{n}}$ is considered in (9). By considering the model (16) and applying Algorithm 2 to that test problem we obtain the results depicted in Figure 8.



(a) Sparse WB for the number 3 with $\texttt{n} = 45$.

(b) Sparse WB for the number 5 with $\texttt{n} = 47$.

Figure 8: Examples of two sparse WB computed with Algorithm 2.

Note that Figure 8 shows more satisfactory results than those displayed in Figure 7. However, Algorithm 2 is significantly slower than Algorithm 1: Algorithm 2 performed 5000 iterations (for each example) in about thirty minutes. Furthermore, the memory usage of Algorithm 2 is not as efficient as that of Algorithm 1. These two drawbacks of Algorithm 2 should be addressed in future research.

References

[1] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *Siam Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] 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.

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

[4] M. Beiglböck, P. Henry-Labordere, and F. Penkner. Model-independent bounds for option prices—a mass transport approach. *Finance and Stochastics*, 17:477–501, 2013.

[5] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):1111–1138, 2015.

[6] S. Borgwardt. An LP-based, strongly-polynomial 2-approximation algorithm for sparse Wasserstein barycenters. *Operational Research*, 22(2):1511–1551, Apr 2022.

[7] S. Borgwardt and S. Patterson. On the computational complexity of finding a sparse Wasserstein barycenter. *Journal of Combinatorial Optimization*, 41(3):736–761, Apr 2021.

[8] G. Carlier, A. Oberman, and E. Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1621–1642, nov 2015.

[9] L. Condat. Fast projection onto the simplex and the $\boldsymbol{l_1}$ball. *Mathematical Programming*, 158(1):575–585, Jul 2016.

[10] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR.

[11] W. de Oliveira. The ABC of DC programming. *Set-Valued Var. Anal.*, 28(4):679–706, 2020.

[12] W. de Oliveira, C. Sagastizábal, D. D. J. Penna, M. E. P. Maceira, and J. M. Damázio. Optimal scenario tree reduction for stochastic streamflows in power generation planning problems. *Optimization Methods and Software*, 25(6):917–936, 2010.

[13] Y. Dolinsky and H. M. Soner. Robust hedging with proportional transaction costs. *Finance and Stochastics*, 18:327–347, 2014.

[14] 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.

[15] 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.

[16] A. Galichon, P. Henry-Labordere, and N. Touzi. A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options. 2014.

[17] A. Gramfort, G. Peyré, and M. Cuturi. Fast optimal transport averaging of neuroimaging data. In S. Ourselin, D. C. Alexander, C.-F. Westin, and M. J. Cardoso, editors, *Information Processing in Medical Imaging*, pages 261–272, Cham, 2015. Springer International Publishing.

[18] J. v. Lindheim. Simple approximative algorithms for free-support Wasserstein barycenters. *Computational Optimization and Applications*, 85(1):213–246, May 2023.

[19] D. Mimouni, P. Malisani, J. Zhu, and W. de Oliveira. Scenario tree reduction via Wasserstein barycenters, 2024. https://arxiv.org/abs/2411.14477.

[20] 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.

[21] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[22] G. Peyré and M. Cuturi. Computational optimal transport, 2020.

[23] G. Puccetti, L. Rüschendorf, and S. Vanduffel. On the computation of Wasserstein barycenters. *Journal of Multivariate Analysis*, 176(104581), 2020.

[24] R. T. Rockafellar. Progressive decoupling of linkages in optimization and variational inequalities with elicitable convexity or monotonicity. *Set-Valued and Variational Analysis*, 27:863–893, 2019.

[25] R. T. Rockafellar. Generalizations of the proximal method of multipliers in convex optimization. *Computational Optimization and Applications*, 87(1):219–247, Jan 2024.

[26] G. M. Sempere, W. de Oliveira, and J. O. Royset. An implementable proximal-type method for computing critical points to minimization problems with a nonsmooth and nonconvex constraint, 2025. To appear in Journal of Optimization Theory and Applications.

[27] D. Simon and A. Aberdam. Barycenters of natural images constrained Wasserstein barycenters for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7910–7919, 2020.

[28] J. Souza and W. de Oliveira. A progressive decoupling algorithm for minimizing the difference of convex and weakly convex functions. *Journal of Optimization Theory and Applications*, 2025.

[29] J. Sun and M. Zhang. The elicited progressive decoupling algorithm: A note on the rate of convergence and a preliminary numerical experiment on the choice of parameters. *Set-Valued and Variational Analysis*, 29(4):997–1018, Dec 2021.