On the Computation of Wasserstein barycenters

Giovanni Puccetti[1], Ludger Rüschendorf [2], and Steven Vanduffel [3]

[1] *Department of Economics, Management and Quantitative Methods, University of Milano, Italy.*
[2] *Department of Mathematical Stochastics, University of Freiburg, Germany*
[3] *Department of Economics, Vrije Universiteit Brussels, Belgium*

November 4, 2019

Abstract

The Wasserstein barycenter is an important notion in the analysis of high dimensional data with a broad range of applications in applied probability, economics, statistics, and in particular to clustering and image processing. In this paper, we state a general version of the equivalence of the Wasserstein barycenter problem to the n-coupling problem. As a consequence, the coupling to the sum principle (characterizing solutions to the n-coupling problem) provides a novel criterion for the explicit characterization of barycenters. Based on this criterion, we provide as a main contribution the simple to implement iterative swapping algorithm (ISA) for computing barycenters. The ISA is a completely non-parametric algorithm which provides a sharp image of the support of the barycenter and has a quadratic time complexity which is comparable to other well established algorithms designed to compute barycenters. The algorithm can also be applied to more complex optimization problems like the k-barycenter problem.

Keywords: Wasserstein barycenter, swapping algorithm, optimal transportation, k-means clustering, image processing.

1 Wasserstein barycenters and optimal n-couplings

The computation of Wasserstein barycenters has recently raised a lot of interest in the literature due to a broad range of applications in applied probability, statistics, economics, and in particular image processing. We refer to Solomon et al. (2015), Bonneel et al. (2015), Anderes et al. (2016), and Peyré and Cuturi (2019) for a comprehensive list of recent activity.

In this paper, we develop a novel algorithm for computing the barycenter of n probability measures $\mu_1, \dots, \mu_n \in P_2(\mathbb{R}^d)$ with finite second moments, and using the L^2-Wasserstein metric W_2 defined as

$$W_2^2(\mu, \nu) := \inf \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} ||x-y||^2 \, dP(x,y) : P \in M(\mu, \nu) \right\},$$

for measuring distances. Here, $M(\mu, \nu)$ denotes the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\mu, \nu \in P_2(\mathbb{R}^d)$. In what follows, any couple (X, Y) of random vectors $X \sim \mu$ and $Y \sim \nu$ is called an *optimal coupling* of μ and ν if $\mathbb{E}||X-Y||^2 = W_2^2(\mu, \nu)$. By definition, a (Wasserstein) barycenter μ of (μ_i) is any solution of

$$\inf \left\{ \sum_{i=1}^n W_2^2(\mu_i, \mu) : \mu \in P_2(\mathbb{R}^d) \right\}. \tag{1.1}$$

A barycenter μ exists in generality and, if one of the μ_i vanishes on all Borel subsets of Hausdorff dimension $d-1$ (e.g., if μ_i is absolutely continuous), then it is also unique; see Agueh and Carlier (2011), Kim and Pass (2014, Ex. 3.3), and Álvarez-Esteban et al. (2016).

A main starting point of our paper is the insight that the barycenter problem can be cast as the optimal n-coupling problem, as in Rüschendorf and Uckelmann (1997, 2002). *Optimal n-couplings* $(X_i) = (X_1, \dots, X_n)$ are defined as solutions of

$$\sup\left\{\mathbb{E}\Big|\Big|\sum_{i=1}^n X_i\Big|\Big|^2; X_i\sim\mu_i, 1\le i\le n\right\}. \tag{1.2}$$

The existence of a solution of (1.2) follows from classical arguments (Rachev and Rüschendorf, 1998).

The connection between the barycenter problem (1.1) and the n-coupling problem (1.2) was stated in Agueh and Carlier (2011, Proposition 4.2) under a continuity assumption on μ_i (smallness on small sets) and also in Anderes et al. (2016, Proposition 1) for discrete measures. As compared to the above references, the following simple proof is valid for general measures, in particular it is also valid for discrete distributions as used in the remainder of this paper.

Proposition 1.1. *(X_i) is an optimal n-coupling if and only if the distribution $\overline{\mu}_n$ of $\overline{S}_n := \sum_{i=1}^n X_i/n$ is a barycenter of (μ_i).*

Proof. The proof is a consequence of the following well known inequality holding for any random vectors $X_1,\dots,X_n$, Z:

$$\sum_{i=1}^n ||X_i - Z||^2 \ge \sum_{i=1}^n ||X_i - \overline{S}_n||^2. \tag{1.3}$$

If $\mu\in P_2(\mathbb{R}^d)$ is a candidate for the barycenter and if (X_i,Z) is an optimal coupling of μ_i and μ, for all $1\le i\le n$, then define $\overline{\mu}_n$ as the law of $\overline{S}_n = \sum_{i=1}^n X_i/n$. By (1.3), we obtain

$$\sum_{i=1}^n W_2^2(\mu_i,\mu) = \mathbb{E}\sum_{i=1}^n ||X_i - Z||^2 \ge \mathbb{E}\sum_{i=1}^n ||X_i-\overline{S}_n||^2 \ge \sum_{i=1}^n W_2^2(\mu_i,\overline{\mu}_n).$$

Thus, $\overline{\mu}_n$ is an improvement over μ. As a consequence, we obtain

$$\inf\left\{\sum_{i=1}^n W_2^2(\mu_i,\mu):\mu\in P_2(\mathbb{R}^d)\right\} = \inf\left\{\mathbb{E}\sum_{i=1}^n \Big|\Big|X_i-\overline{S}_n\Big|\Big|^2; X_i\sim\mu_i, 1\le i\le n\right\}. \tag{1.4}$$

By elementary arguments it is easy to see that (X_i) solves (1.4) if and only if (X_i) is an optimal n-coupling and hence the distribution $\overline{\mu}_n$ of $\overline{S}_n$ is a barycenter of (μ_i). □

Remark 1.1 (Weighted barycenters)**.** If one considers the more general barycenter problem

$$\inf\left\{\sum_{i=1}^n \lambda_i\, W_2^2(\mu_i,\mu);\mu\in P_2(\mathbb{R}^d)\right\},$$

where $\lambda_1,\dots,\lambda_n$ are positive weights summing to 1, then Proposition 1.1 continues to hold but now with the arithmetic mean $\overline{S}_n := \sum_{i=1}^n X_i/n$ replaced by the weighted mean $\tilde{S}_n := \sum_{i=1}^n \lambda_i\, X_i$. For the ease of notation, in what follows we describe our algorithm for the case $\lambda_i = 1/n$. In the general case of a weighted barycenter problem, one applies the algorithm to the variables $\lambda_i\, X_i$; see also our application in Section 4.1. □

As a consequence of Proposition 1.1, Wasserstein barycenters result from solutions of the n-coupling problem and characterizations of optimal n-couplings directly imply corresponding characterizations of barycenters. In Rüschendorf and Uckelmann (2002) it has been shown that optimal coupling of all X_i to the sum $S_n := \sum_{i=1}^n X_i$ as well as to $S_{(i)} := \sum_{j\neq i} X_j$ is a necessary condition for an optimal n-coupling. Moreover, if P^{S_n} is Lebesgue-continuous, then any of these optimal coupling conditions is also sufficient for an optimal n-coupling. This provides relevant sufficient conditions for obtaining an optimal n-coupling and a barycenter, respectively. In particular, by the proof of Theorem 2.4 in Rüschendorf and Uckelmann (2002), this continuity holds if any of the μ_i is Lebesgue-continuous.

Agueh and Carlier (2011) state in their formula (4.10) a variational characterization of barycenters which is a consequence of the dual representation in Gangbo and Święch (1998) (see Theorem 4.1 in Agueh and Carlier, 2011). For related results see also Rüschendorf and Uckelmann (1997). This variational characterization, however, needs for its application the explicit knowledge of the solutions to the dual problem. On the other hand, the results in Rüschendorf and Uckelmann (2002) for the characterization of optimal n-couplings and thus also for the barycenter problem are constructive. They lead in particular to the following general sufficient criterion for the explicit construction of optimal n-couplings resp. barycenters, which makes it possible to determine the explicit solution for some classes of distributions.

Proposition 1.2. *Let $f_i : \mathbb{R}^d \to \mathbb{R}, 1 \leq i \leq n$, be convex, lower semicontinuous real functions such that*

$$\sum_{i=1}^{n} f_i(x) = ||x||^2/2 + \text{const}$$

holds almost everywhere (w.r.t. the Lebesgue measure λ^d). Let $\mu \in P_2(\mathbb{R}^d)$ be Lebesgue-continuous, then the image measure $\mu^{\nabla f_i}$ (the push-forward of μ through ∇f_i) is well defined. Under the assumption that $\mu_i = \mu^{\nabla f_i}, 1 \leq i \leq n$, it holds by letting $S \sim \mu$ and $X_i := \nabla f_i(S)$ that (X_i) is an optimal n-coupling of the (μ_i) and $P^{S/n}$ is a barycenter of (μ_i).

Proof. By the classical optimal transportation results in Rüschendorf and Rachev (1990) and Brenier (1991) the functions ∇f_i are optimal transport maps of μ to μ_i. Furthermore, by assumption $\sum_{i=1}^{n} \nabla f_i = \text{Id}$ almost everywhere, i.e., $\sum_{i=1}^{n} X_i = S$ almost surely. Thus, all X_i are optimally coupled to their sum S and, as a consequence of the characterization in Rüschendorf and Uckelmann (2002), (X_i) is an optimal n-coupling. By Proposition 1.1 above, this implies that $P^{S/n}$ is a barycenter of (μ_i). □

Remark 1.2 (Construction of optimal n-couplings and barycenters)**.**

a) The condition in Proposition 1.2 that the distribution of the optimal n-coupling sum $S = S_n$ is Lebesgue-continuous is satisfied if at least one of the μ_i is Lebesgue-continuous. One can formulate a corresponding sufficient condition for general non-continuous distributions in terms of subgradients (see the proof of Theorem 2.3 in Rüschendorf and Uckelmann, 2002).

b) The conditions of Proposition 1.2 include for instance the case of multivariate elliptical distributions and in particular the case of multivariate normal distributions (as dealt with in Section 3). For this case consider $f_i(x) = x^\top A_i x$, $1 \leq i \leq n$, with some positive semidefinite matrices A_i. The condition $\sum_{i=1}^{n} \nabla f_i(x) = 2\big(\sum_{i=1}^{n} A_i\big)x = x$ then yields the crucial equation (3.1) for the involved covariance matrix Σ_0 of μ.
The more general case in which $f_i(x) = R_i(x^\top A_i x)$, $1 \leq i \leq n$, where the R_i are increasing convex, leads to optimal n-couplings of the form $\nabla f_i(x) = 2r_i(x^\top A_i x)A_i x$ with $r_i = R_i' \geq 0$ and increasing. This case allows to deal with elliptical distributions with different radial parts. In general the equation $\sum_{i=1}^{n} \nabla f_i = \text{Id}$ leads to nonlinear equations that need to be solved numerically. □

The optimal n-coupling problem (1.2) can also be equivalently rewritten as

$$\begin{gathered}\sup\{\mathbb{E}[f(X_1, \ldots, X_n)]; X_i \sim \mu_i, 1 \leq i \leq n\}, \text{ with} \\ f(x_1, \ldots, x_n) = \sum_{i=1}^{n} \langle x_i, \sum_{j \neq i} x_j \rangle.\end{gathered} \tag{1.5}$$

This formulation suggests to solve the barycenter problem by an iterative sequence of 2-coupling problems, i.e., by iteratively calculating the Wasserstein distance of X_i and $\sum_{j \neq i} X_j$. This can be done by applying an iterative version of the so-called swapping algorithm, which was investigated in detail in Puccetti (2017). As a result, we obtain an approximation of the optimal n-coupling and, as a consequence of Proposition 1.1, thus also of the barycenter of (μ_i). The swapping algorithm in fact is grounded on the basic characterization of optimal couplings by cyclically monotone support in Rüschendorf and Rachev (1990) and was motivated in this connection already in that paper.

The remainder of the paper is organized as follows. In Section 2, we propose the iterative swapping algorithm (ISA) for computing barycenters of discrete measures. In Section 3, we test its performance against the benchmark of Gaussian measures (approximated by empirical counterparts) for which analytical results are available. In Section 4, we provide applications to the visualization of perturbed images and the so-called k-barycenter problem. We summarize our conclusions and provide possible extensions in Section 5.

2 The Iterative Swapping Algorithm (ISA)

Apart from the case $d = 1$ (which allows for a fast solution, see Rabin et al., 2012), there are no analytical formulas available for the solution of (1.1). By means of Proposition 1.2, however, one can give an explicit solution for some classes of examples like elliptical measures (as studied in Section 3).

The barycenter of general measures can be approximated by taking consistent empirical versions of the input measures. In fact, under weak assumptions, the sequence of barycenters of empirical versions converges to the true barycenter; see Theorem 3, Corollary 5, and Proposition 6 in Le Gouic and Loubes (2017).

For discrete measures, the computation of a Wasserstein barycenter is in principle obtained through the solution of a finite(high)-dimensional linear program; see (LP) below. However, if each of the n measures is represented by k points, this LP has k^n variables, which quickly becomes intractable for linear programming software and this even for moderate values of k. This is the main motivation why alternative techniques recently emerged in the literature. Different algorithms to compute Wasserstein barycenters of discrete measures (point clouds) have been described amongst others:

- in Carlier et al. (2015), where a simple linear programming reformulation leads to an LP which scales linearly with the number of marginals. Such reduction (similarly obtained in Anderes et al., 2016), together with a bound on the support of the unknown barycenter, makes the problem more tractable but still suffers from heavy computation time and memory consumption. A second algorithm introduced in Carlier et al. (2015) uses the dual formulation of the problem (1.1).

- in Rabin et al. (2012), where the authors introduce and use the so-called *sliced Wasserstein distance* between projections of the input measures on the line, in which case problem (1.1) can be efficiently solved. This method shows to be effective in lower dimensions $d = 2, 3$, and has been further developed in Bonneel et al. (2015).

- in Benamou et al. (2015), where an entropic regularization of the initial linear program is proposed, which makes it possible to use a simple iteration scheme for computing its solution.

Through various examples and applications, we will compare in the remainder of the paper the performance of our proposed algorithm with those of the above referenced techniques.

As described in Section 1, an approximate determination of Wasserstein barycenters can be obtained by iteratively solving 2-coupling problems. For discrete measures with a finite number of support points the characterization of optimal couplings by cyclical monotonicity of the support then suggests an iterative use of the swapping algorithm, used in Puccetti (2017) for the calculation of the Wasserstein distance; see also Rüschendorf and Rachev (1990). An alternative motivation for using iterative swapping consists in formulating the barycenter problem as a linear programming problem and to approximate it by a multi-index assignment problem.

We denote by n the number of pre-assigned probability measures, k the number of atoms of empirical measures, and d the dimensionality of the space $\mathbb{R}^d$ where they take values. For any $x \in \mathbb{R}^d$, δ_x denotes the Dirac unit mass on x. From now onwards, we consider n discrete, k-atomic measures of the form

$$\mu_i = \sum_{j=1}^{k} \frac{1}{k} \delta_{x_j^i}\,, \quad 1 \le i \le n, \tag{2.1}$$

where $x_1^i, \ldots, x_k^i \in \mathbb{R}^d$ are the k atoms in $\mathbb{R}^d$ of the i-th measure, each one having probability mass $1/k$.

Let $\mathcal{J} = \{1, \ldots, k\}^n$ and $\mathcal{J}_j^i = \{(u_1, \ldots, u_n) \in \mathcal{J} : u_i = j\}$. Taking as marginals the k-atomic measures in (2.1), problem (1.5) becomes the finite-dimensional linear program

$$\begin{aligned} \mathrm{LP} = \max_{p_u \ge 0} \sum_{u \in \mathcal{J}} f_u\, p_u, \quad \text{s.t.} \\ \sum_{u \in \mathcal{J}_j^i} p_u = 1/k, \quad 1 \le j \le k, \quad 1 \le i \le n, \end{aligned} \tag{LP}$$

where $f_u = f(x^1_{u_1}, \dots, x^n_{u_n})$, for $u = (u_1, \dots, u_n) \in \mathcal{J}$. Any solution of (LP) is a discrete probability measure on $\mathbb{R}^{d \times n}$ having marginals (μ_i). Consequently, the barycenter of (μ_i) is finitely supported on the set $B = \{(x^1_{u_1} + \cdots + x^n_{u_n})/n, (u_1, \dots, u_n) \in \mathcal{J}\}$.

In principle, the cardinality of B can attain the upper bound k^n. However, it has been noted in Theorem 2 in Anderes et al. (2016) that there always exists a barycenter of discrete measures whose support has at most $(nk - n + 1)$ points. This result encourages the practice of assuming a smaller number of support points in a barycenter as an approximation to the true solution; see for instance Oberman and Ruan (2015) and Schmitzer (2016). Motivated by the sparseness of at least one barycenter, we add to (LP) the constraints $p_u \in \{0, 1/k\}, u \in \mathcal{J}$, and obtain the following multi-index assignment problem

$$\text{AS} = \max \left\{ \frac{1}{k} \sum_{j=1}^{k} f(x^1_{\sigma_1(j)}, \dots, x^n_{\sigma_n(j)}); \sigma_1, \dots, \sigma_n \in \Sigma_k \right\}, \tag{AS}$$

in which Σ_k denotes the set of all the permutations of $\{1, \dots, k\}$.

We obviously have that LP $\geq$ AS, whereas LP $=$ AS holds in general by Birkhoff's theorem (Birkhoff, 1946) only in case $n = 2$ (when a barycenter is then theoretically guaranteed to be supported on k points). Problem (AS) is known as an axial n-index assignment problem, which for $n = 2$ reduces to the classical assignment problem which can be generally solved by the Hungarian algorithm or more refined techniques in roughly $O(k^3)$; see Puccetti (2017) for more precise computational details. For $n \geq 3$, Problem (AS) is $\mathcal{NP}$-hard and, a few special cases apart, only (meta)heuristic and enumerative methods are known for its exact solution; see Burkard et al. (2009, Ch. 10).

Approximating the Wasserstein barycenter calculation by a multi-index assignment problem, one can iteratively use the swapping algorithm in Puccetti (2017) as an heuristic to approximate an optimal assignment in (AS) with a quadratic cost $O(nk^2)$.

For the function f in (1.5), the basic idea behind the algorithm consists in assessing, for any index $i \in \{1, \dots, n\}$, and any positions $1 \leq k_1 < k_2 \leq k$, whether

$$\langle x^i_{\sigma_i(k_1)}, \sum_{j \neq i} x^j_{\sigma_j(k_1)} \rangle + \langle x^i_{\sigma_i(k_2)}, \sum_{j \neq i} x^j_{\sigma_j(k_2)} \rangle < \langle x^i_{\sigma_i(k_2)}, \sum_{j \neq i} x^j_{\sigma_j(k_1)} \rangle + \langle x^i_{\sigma_i(k_1)}, \sum_{j \neq i} x^j_{\sigma_j(k_2)} \rangle. \tag{2.2}$$

If condition (2.2) is satisfied, one swaps $\sigma_i(k_1)$ and $\sigma_i(k_2)$. As the number of possible permutations is finite, the swapping procedure terminates after a finite number of swaps and provides a new multi-index assignment, delivering (at each step) a strictly bigger value of f in (AS).

Iterated Swapping Algorithm (ISA):

1. Fix k-atomic measures as in (2.1) and let $\sigma_i = \text{Id}, 1 \leq i \leq n$.
2. For all possible pairs (k_1, k_2) with $1 \leq k_1 < k_2 \leq k$, and for all $i \in \{1, \dots, n\}$, if the swapping condition (2.2) holds, then swap $\sigma_i(k_1)$ and $\sigma_i(k_2)$. A new multi-index assignment $\{\sigma'_1, \dots, \sigma'_n\}$ is found.
3. Repeat 2. with $\sigma = \sigma'$ until no further swaps are possible. The algorithm terminates after a finite number of iterations of step 2. and outputs the final assignment $\{\widehat{\sigma}_1, \dots, \widehat{\sigma}_n\}$.

The Iterated Swapping Algorithm (ISA in what follows) thus outputs a *pairwise optimal assignment* $\{\widehat{\sigma}_1, \dots, \widehat{\sigma}_n\}$, i.e., a set of permutations for which the function f in (AS) cannot be improved by iterative swaps of only two positions. Denote by $(\widehat{X}_1, \dots, \widehat{X}_n) \sim \{\widehat{\sigma}_1, \dots, \widehat{\sigma}_n\}$ a random vector having probability measure (with marginals μ_i) uniformly distributed on the k atoms $\{(x^1_{\widehat{\sigma}_1(j)}, \dots, x^n_{\widehat{\sigma}_n(j)}), 1 \leq j \leq k\}$. In the following we show that $(\widehat{X}_1, \dots, \widehat{X}_n)$ is an approximate optimal n-coupling and, therefore, the distribution of $\sum_{i=1}^n \widehat{X}_i/n$ provides an approximation of the barycenter of (μ_i).

3 The Gaussian case

In this section, we test the accuracy of the ISA against the benchmark of Gaussian measures, where the problem has a computable solution. In fact, the barycenter of n Gaussian measures $N(\mathbf{0}, \mathbf{\Sigma}_i), 1 \leq i \leq n$, with non-singular covariance matrices Σ_i, is equal to $\mu = N(\mathbf{0}, \mathbf{\Sigma}_0/n^2)$, in which $\mathbf{\Sigma}_0$ is the unique positive definite solution of

$$\mathbf{\Sigma}_0 = \sum_{i=1}^{n} \left(\mathbf{\Sigma}_0^{1/2} \mathbf{\Sigma}_i \mathbf{\Sigma}_0^{1/2} \right)^{1/2}. \tag{3.1}$$

The sufficiency of condition (3.1) to optimality in the Gaussian case was already noted in Knott and Smith (1994) in the case $n = 3$. Rüschendorf and Uckelmann (2002) proved the necessity and sufficiency of (3.1) and the existence of a solution (and hence of a solution to the n-coupling problem for Gaussian measures). Agueh and Carlier (2011) established the uniqueness of such a solution. As mentioned before, the non-trivial matrix equation (3.1) follows from Proposition 1.2 by some simple algebra and the fact that in the Gaussian case the optimal coupling of $N(0, \mathbf{\Sigma}_i)$ to $N(0, \mathbf{\Sigma}_0)$ is linear (Dowson and Landau, 1982). For non-centered Gaussian measures, the mean vector of the barycenter is simply the average of all the means.

The matrix $\mathbf{\Sigma}_0$ in (3.1) can be computed using the intuitive iterative procedure

$$\mathbf{K}_0^{(t+1)} = \left(\sum_{i=1}^{n} \left(\mathbf{K}_0^{(t)} \mathbf{\Sigma}_i \mathbf{K}_0^{(t)} \right)^{1/2} \right)^{1/2}.$$

We found $\lim_{t\to\infty} \mathbf{K}_0^{(t+1)} = \mathbf{\Sigma}_0^{1/2}$ componentwise in any dimension d when taking as initial condition $\mathbf{K}_0^{(0)} = \sum_{i=1}^{n} \mathbf{\Sigma}_i/n$. For the uniqueness of a solution of (3.1) and more details on the efficient computation of $\mathbf{\Sigma}_0$, we refer to Álvarez-Esteban et al. (2016).

Now, let μ_i^k be the empirical distribution associated to a set of k independent simulations from $N(\mathbf{0}, \mathbf{\Sigma}_i)$, for $1 \leq i \leq n$. Let (X_i) be a n-optimal coupling and recall that $S_n = \sum_{i=1}^{n} X_i$. We have that

$$\begin{aligned} \mathbf{\Sigma}_0 = \mathbb{E}\left[S_n\, S_n^T\right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i \right) \left(\sum_{i=1}^{n} X_i \right)^T \right] & \\ = \sum_{i=1}^{n} \mathbb{E}\Big[X_i\, X_i^T\Big] + \sum_{i=1}^{n} \mathbb{E}\left[X_i \Big(\sum_{j\neq i} X_i \Big)^T \right] &= \sum_{i=1}^{n} \mathbf{\Sigma}_i + \sum_{i=1}^{n} \mathbb{E}\left[X_i \Big(\sum_{j\neq i} X_i \Big)^T \right]. \end{aligned}$$

Based on this last equality, when the algorithm terminates, the matrix $\mathbf{\Sigma}_0$ can be estimated by $\widehat{\mathbf{\Sigma}}$ given as

$$\widehat{\mathbf{\Sigma}} = \sum_{i=1}^{n} \mathbf{\Sigma}_i + \sum_{i=1}^{n} \mathbb{E}\left[\widehat{X}_i \Big(\sum_{j\neq i} \widehat{X}_i \Big)^T \right], \tag{3.2}$$

where $(\widehat{X}_1, \ldots, \widehat{X}_n) \sim \{\widehat{\sigma}_1, \ldots, \widehat{\sigma}_n\}$ corresponds to the final assignment $\{\widehat{\sigma}_1, \ldots, \widehat{\sigma}_n\}$, as found by the algorithm for the empirical marginals $\mu_1^k, \ldots, \mu_n^k$.

Example 3.1. We compute ISA estimates of the optimal covariance matrix $\mathbf{\Sigma}_0/n^2$ of the barycenter of $n = 3$ Gaussian distributions having null mean and equicorrelation matrices

$$\mathbf{\Sigma}_i = \begin{pmatrix} 1 & \sigma_i & \ldots & \sigma_i \\ \sigma_i & 1 & \ldots & \sigma_i \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_i & \sigma_i & \ldots & 1 \end{pmatrix}, \tag{3.3}$$

with $\sigma_1 = 0, \sigma_2 = 0.4, \sigma_3 = -0.15$, for dimensions $d = 2, 3, 4$. For equicorrelation matrices (3.3), also the covariance matrix of the barycenter has equal correlations. Figure 1 illustrates the accuracy of the empirical barycenters found, the number of iterations of the algorithm and the computation times. As compared to the similar benchmark study carried out in Carlier et al. (2015), where the authors test an optimization algorithm based on the dual formulation of (1.1), the ISA can handle a larger number k of sample points and a larger number n of measures resulting in a higher accuracy; see Table 1.

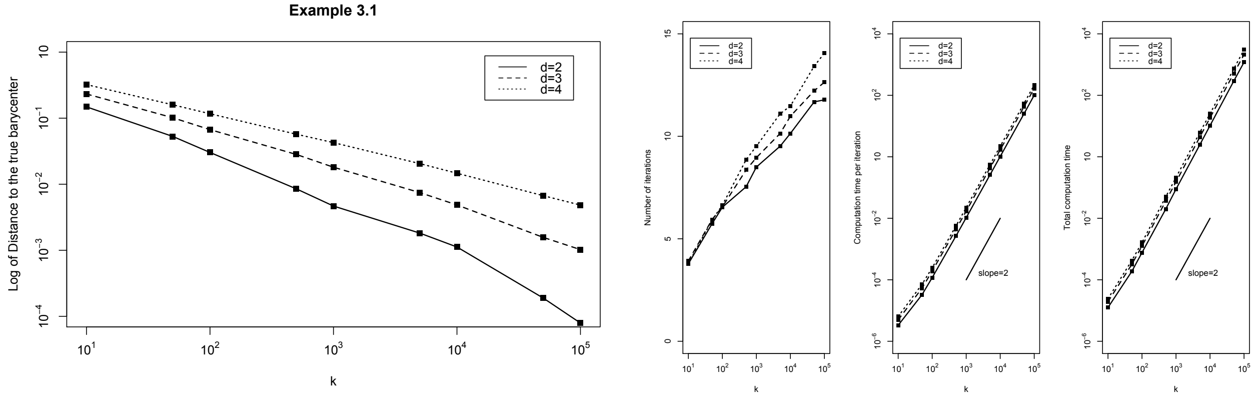

Figure 1: **(Left)** Log-Wasserstein distance between the empirical barycenter computed by ISA and the true barycenter of $n = 3$ Gaussian distributions as described in Example 3.1, for different values of k and d. **(Right, 1)** Number of iterations of Step 2. of the algorithm, i.e., the number of times the swapping condition (2.2) is checked for all $i \in \{1, \dots, n\}$ and indexes $1 \leq k_1 < k_2 \leq k$, before a pairwise optimal allocation is attained. **(Right, 2-3)** Log-computation times, per iteration and total. All estimates are averages evaluated over 50 different initial random samples. In all the applications described in this paper the ISA is compiled in C++ and has run on an Apple Mac mini (3.2 GHz Intel Core i7, 16 GB RAM).

	$d = 2$	$d = 3$	$d = 4$
$\widehat{\sigma}_{ii} - \sigma_{ii}$	1.11e-04 (0.01%)	1.15e-03 (0.12%)	4.72e-03 (0.49%)
$\widehat{\sigma}_{ij} - \sigma_{ij}$	1.49e-05 (0.02%)	3.25e-05 (0.04%)	2.78e-05 (0.04%)

Table 1: Absolute (relative) difference between the empirical estimates $\widehat{\mathbf{\Sigma}}/n^2$ (see (3.2)) and the true covariance matrix $\mathbf{\Sigma}_0/n^2$ of the barycenter of $n = 3$ Gaussian distributions as described in Example 3.1. Empirical estimates are computed via ISA for dimensions $d = 2, 3, 4$ and $k = 10^5$ sample points. Error estimates are averages evaluated over 50 different initial random samples.

In Figure 2, we compare the ISA barycenter of the same three Gaussian measures with the so-called sliced Wasserstein barycenter (SWB in the following) as described in Bonneel et al. (2015). Built on the notion of sliced Wasserstein distance (Rabin et al., 2012), this approach approximates barycenters of measures using easily computable 1-d Wasserstein distances along radial projections of the input measures. The sliced Wasserstein barycenter turns out to be the solution of an optimization problem which integrates the distances of all projections.

While the SWB is more accurate for lower numbers of discretization points $k \leq 10^3$, we experienced non-convergence of the underlying iterative algorithm (reaching the fixed maximum number of possible iterations) for higher number of points. On the other hand, the ISA is always more accurate for higher levels of k but at the cost of an higher computation time.

For increasing dimensions, requiring the ISA to attain an exactly pairwise optimal assignment might be onerous or might not deliver a worthwhile extra accuracy. This effect was studied in detail in the case of calculating Wasserstein distances in Puccetti (2017). Figure 3 shows how most of the increase of the objective function (1.5) is gained trough the first iterations of the algorithm, where the large majority of swaps is performed. In particular we notice that the first round of swaps accounts for more than 99% of the relative increase if one starts from a random initial configuration. Based on this observation, it might be convenient to introduce a stopping rule based on an

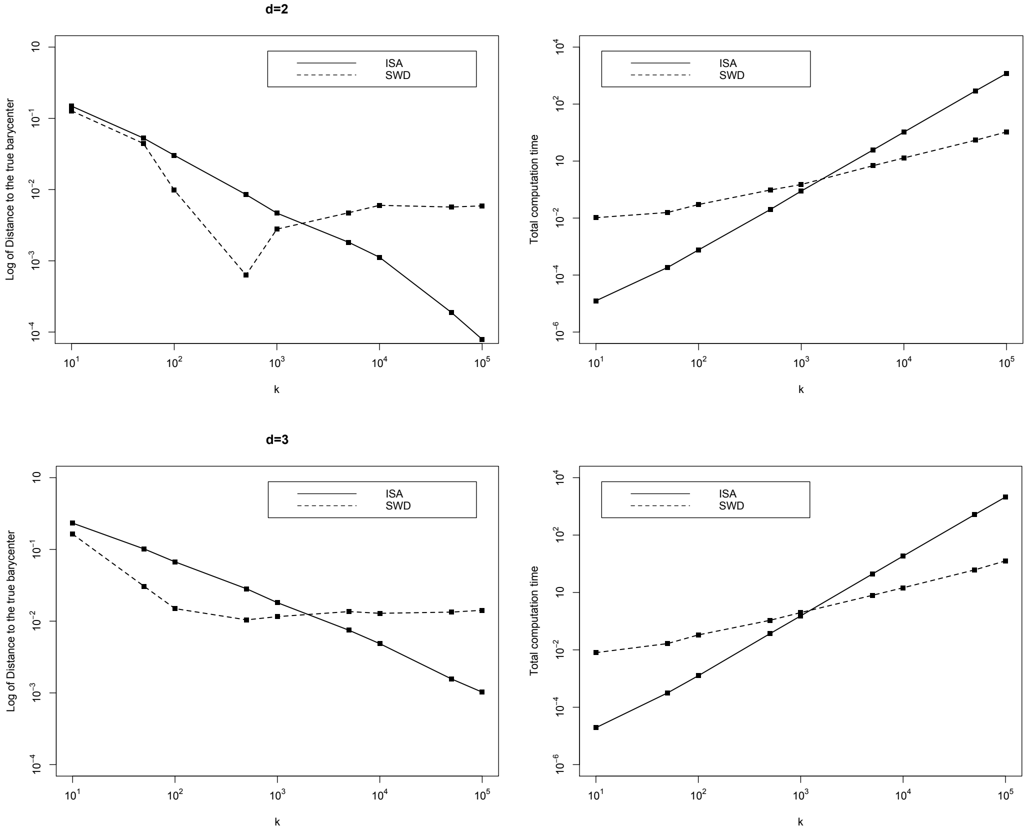

Figure 2: **(Left)** Log-Wasserstein distances between the empirical barycenter produced by ISA resp. SWB, and the true barycenter of $n = 3$ Gaussian distributions as described in Example 3.1, for $d = 2$ (*top*) and $d = 3$ (*bottom*). **(Right)** Corresponding total Log-computation times. The SWB is produced via a maximum number of 100 possible iterations and 10 projections. Doubling the number of iterations and/or projections in this example do not yield an extra accuracy. All estimates are averages evaluated over 50 different initial random samples. In all the applications described in this paper the SWB is coded in MATLAB via the script available at `https://github.com/gpeyre/2014-JMIV-SlicedTransport.`

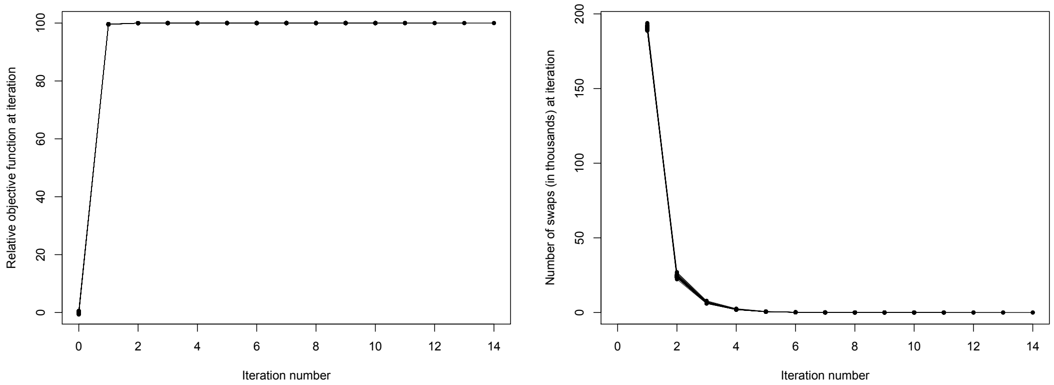

Figure 3: **(Left)** Relative increase of the objective function (1.5) at each iteration of the ISA for the measures described in Example 3.1, for $d = 2$ and $k = 10^4$ discretization points. **(Right)** Number of swaps (in thousands) performed at each iteration. Even if broadly overlapping, each plot shows the lines for 50 different initial random samples.

accuracy condition or, equivalently, on a maximum number of iteration of step 2. to be performed. In the following example, we require that the algorithm stops when the objective function computed at two consecutive iterations of step 2. does not vary above a fixed level of accuracy $\xi > 0$, that is when

$$\frac{1}{k}\left|\sum_{j=1}^{k} f(x^1_{\sigma_1(i)},\dots,x^n_{\sigma_n(i)}) - f(x^1_{\sigma'_1(i)},\dots,x^n_{\sigma'_n(i)})\right| < \xi. \tag{3.4}$$

As an alternative or additional stopping condition, one can (super)impose a maximum number of iterations to be performed.

Example 3.2. We compute ISA estimates of the optimal covariance matrix $\mathbf{\Sigma}_0/n^2$ of the barycenter of $n = 3, 5, 10$ Gaussian distributions having zero mean and equicorrelation matrices (3.3) with the randomly drawn correlation shown in Table 2. We use the stopping condition (3.4) with $\xi = 0.001$ and a maximum number of allowed iterations equal to 20. Accuracy of the empirical barycenters, number of iterations of the algorithm and computation times are illustrated in Figure 4.

	1	2	3	4	5	6	7	8	9	10
σ_i	-0.990608	0.845346	-0.273613	-0.619842	0.320685	-0.248976	-0.546461	-0.36234	0.155687	0.631475

Table 2: Randomly drawn correlation parameters used in Example 3.2.

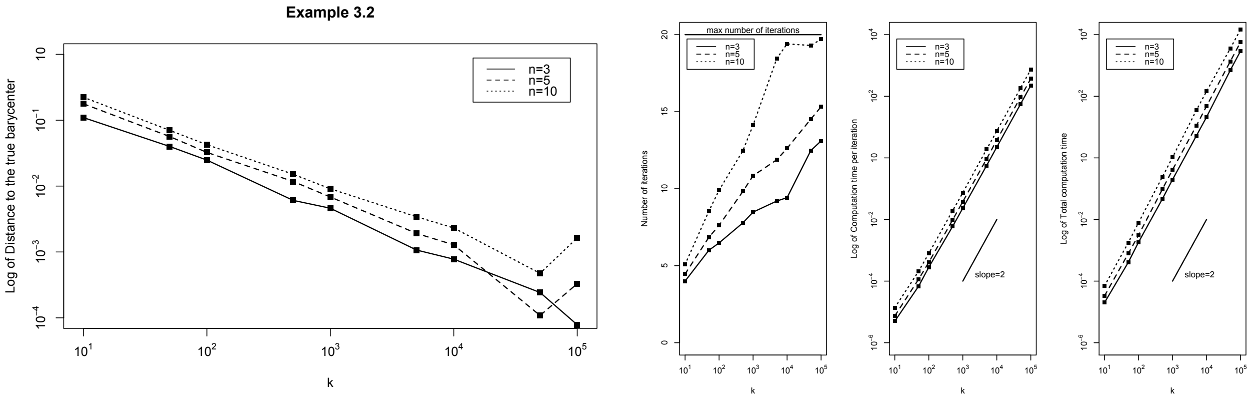

Figure 4: The same as Figure 1 for $n = 3, 5, 10$ bivariate ($d = 2$) Gaussian distributions as described in Example 3.2, for different values of k. All estimates are averages evaluated over 50 different initial random samples (20 for $n = 10$ in the cases $k = 5 \times 10^4, 10^5$).

Figures 1 and 4 show convergence of the ISA algorithm and an experimental run time of $O(nk^2)$ per iteration, which is coherent with the fact that the algorithm checks condition (2.2) $nk(k-1)/2$ times. The number of total iterations of the algorithm turns out to be bounded in n for dimensions d relevant to applications or in general can be bounded (by a terminating condition (3.4) based on accuracy). This quadratic time complexity is comparable to other widely used algorithms; see Section 4. Figure 1 also shows that accuracy decreases when the dimension d is increasing, and similarly to the other algorithms designed to compute barycenters, the ISA is most effective in lower dimensions $d = 2, 3, 4$.

4 Applications

Wasserstein barycenters and general optimal transportation problems have several applications in image processing and computer graphics, for which we refer for instance to Li and Wang (2008), Bonneel et al. (2015) and Rabin et al. (2012). Motivations for using Wasserstein barycenters are summarized in Ye et al. (2017) and references therein. A clear and exhaustive overview of computational optimal transport is given in Peyré and Cuturi (2019).

4.1 Visualization of perturbed images

Similarly to what is done in Figure 1 in Cuturi and Doucet (2014), in Figure 5 we compute the Wasserstein barycenter of $n = 36$ couples of ellipses via three different methods. Each image is rendered as a discrete measure of $k = 400$ atoms on a grid of 65×65 pixels. If the pixel's color is black, then the probability mass at that point is $1/k$. If it is white, it is set to be equal to 0.

The barycenter computed by the ISA is compared to the one produced over the same set of input measures by the algorithm described in Benamou et al. (2015) and then to the solution of the LP described in Carlier et al. (2015, eq. 2.15).

The algorithm in Benamou et al. (2015) operates by adding an entropic regularization penalty to the original transportation problem depending on a regularization parameter $\gamma > 0$, and then uses Iterative Bregman Projections (IBP for simplicity in the following) to solve it. This scheme translates into iterations that are simple matrix-vector products which, in the case of the squared Euclidean distance, only require iterative convolutions of vectors against a discrete diffusion kernel. Entropic regularized optimal transports find its pedigree in Cuturi (2013) and Cuturi and Doucet (2014), who solve the regularized barycenter problem using a gradient descent scheme. Benamou et al. (2015) state that the IBP converges orders of magnitude faster than gradient descent, and it has been further exploited to shape data in 2-D and 3-D; see Solomon et al. (2015). The IBP algorithm is also freely available with guideline and examples at `http://www.numerical-tours.com/matlab/optimaltransp_5_entropic/#56`.

The LP described in Carlier et al. (2015) provides an exact representation of the barycenter at the cost of solving a linear problem with $n \times k \times res^2$ variables and $n(k \times res^2 + res^2 + k)$ constraints, where res is the resolution of the obtained barycenter (the number of pixels used to represent it on a square grid). Notice that in this example ($n = 36, k = 400, res = 65$) one would have roughly 60 million variables and 60 million constraints. By pre-localizing the support of the barycenter, we were able to reduce the LP to $res = 45$ (29 million variables/constraints). Using CVX with the powerful LP solver Gurobi, it takes almost 11 *hours* to compute the barycenter. Even considering that the solution of the LP is exact (once provided the pre-localization of the barycenter) this computation time appears huge if compared to the 0.52 and, respectively, 1.75 *seconds* that ISA and IBP take to approximate the barycenter for the same set of marginals. Most importantly, solving the LP for $n = 36$ marginals almost exploited the total memory of our computer, thus imposing serious limitations on the number of marginals that can be dealt with.

The ISA, IBP and LP algorithms are profoundly different and provide different approximations of Wasserstein barycenters. Entropic regularization enables scalable computations, but the regularization parameter γ adds a slight amount of smoothing in the computed approximation of the barycenter. The ISA does not modify the original transportation cost, but treats discrete measures uniformly distributed on the same number of points and only attains pairwise optimal assignments. The LP gives an exact solution, but its applicability is limited to low dimensions/resolutions.

These differences are evident in the three different outcomes of the algorithms: the ISA produces a visually sharp image of the support of the barycenter, the IBP outcomes a slightly blurred image as a result of the regularization of the original problem, whereas the LP barycenter properly allows for exact gradients of gray (probability).

When one aims at higher dimensions, computing the barycenter via an LP is out of reach. In Figure 6 we compare ISA and IBP barycenters for $n = 1296$ couples of ellipses. Again, each couple still consists of $k = 400$ atoms on a grid of 65×65 pixels. Also in this case, the barycenter computed by IBP shows some probability mass between two circles which the ISA renders as sparse points. By superimposing the two different barycenters (Figure 6, bottom-right), it is evident that the two different methodologies are producing different approximations of the same optimal measure.

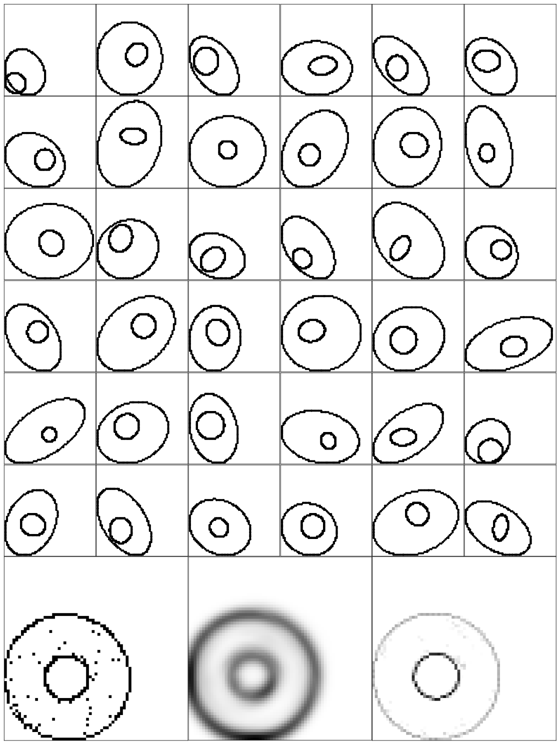

Figure 5: **(Top)** 36 artificial images of couples of ellipses. **(Bottom, left)** Barycenter as computed by the ISA. In this example it takes 12 iterations to reach a pairwise-optimal assignment. Each iteration takes approximately 0.0437 sec in C++ (total comp. time of 0.52 sec.). **(Bottom, center)** Barycenter as computed by the IBP with 97 iterations. For the IBP we took a stopping condition based on the Frobenius norm to produce almost indistinguishable subsequent barycenters. Each IBP iteration takes on average 0.0181 sec in MatLab (total comp. time of 1.75 sec.). **(Bottom, right)** Barycenter as computed by the LP in Carlier et al. (2015) with prelocalization of the barycenter. The LP has been solved by CVX using Gurobi solver in about 11 hours. For the application of the ISA, since the k sample points of each image are chosen deterministically, the initial assignment in step 1. is chosen according to Remark 2 in Section 5.

Weighted barycenters of sample images. The ISA can also be applied on sample images when one generally needs a different number of Dirac masses for each measure. One can then apply the ISA by sampling the same (high) number of points for each measure, in a similar way as done for the Gaussian application in Section 3. In Figure 7 we show the barycenters computed via ISA, IBP, and SWB, for varying weights corresponding to a bilinear interpolation inside a square.

At this point we stress that while our procedure is completely non-parametric (the end-user does not have to calibrate it to the specific measures under study nor to choose any parameter), the IBP algorithm is sensitive to the choice of the regularization parameter γ used to introduce the entropy penalization, and also to the sharpness of the kernel discretization (Solomon et al., 2015). While theoretically for $\gamma \to 0$ one retrieves the solution for the non-penalized original cost function, in practice, the IBP algorithm is competitive in a range where the regularization term is not too small to prohibit computational tractability nor too large to make the iteration scheme converge to a maximum entropy solution (which is the limiting case for $\gamma \to \infty$). For increasing levels of γ the barycenter becomes less and less sparse, an effect that is well illustrated in Peyré and Cuturi (2019).

The convergence of the IBP algorithm is fast if the regularization parameter and the width of the convolution kernel are chosen appropriately to the measures under study, but might deviate from the correct solution if the parameters are misspecified, as illustrated in Figure 7. In contrast, the ISA algorithm is completely non-parametric and is always more accurate (at the cost of a higher computation time) if one allows for an increasing number of iterations and/or number of initial sampling points. The sliced Wasserstein barycenters, analogously to Figure 6 and 7 in Bonneel et al. (2015), show artifacts.

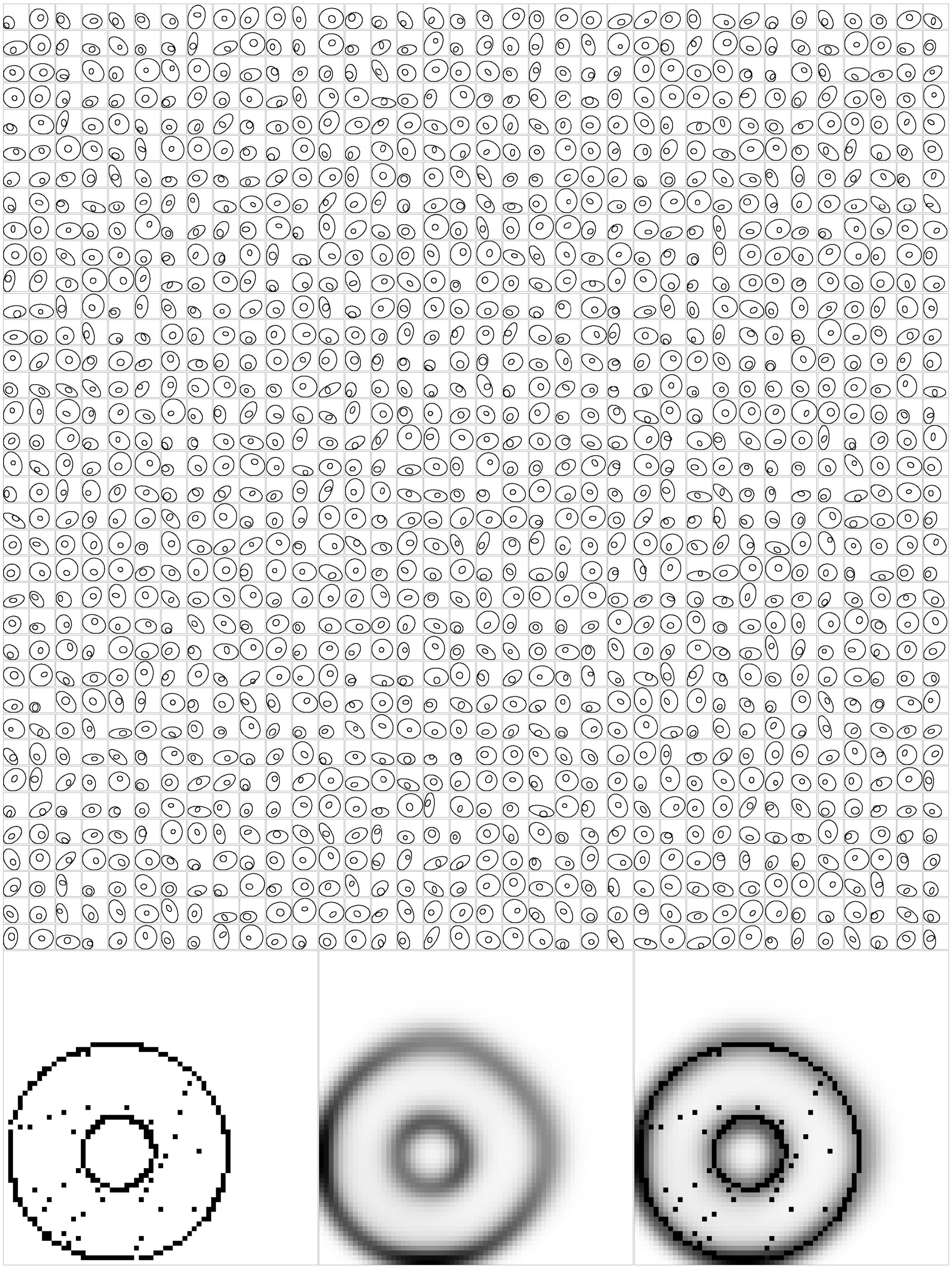

Figure 6: **(Top)** 1296 artificial images of couples of ellipses. **(Bottom, left)** Barycenter as computed by the ISA. In this example it takes 11 iterations to reach a pairwise-optimal assignment. Each iteration takes approximately 2.6 sec. in C++ (total comp. time of 28.6 sec.). **(Bottom, center)** Barycenter as computed by the IBP with 89 iterations. Each IBP iteration takes on average 0.28 sec in MatLab (total comp. time of 24.9 sec.). **(Bottom, right)** Superposition of the ISA and IBP barycenters. Applications of ISA and IBP are the same as in Figure 5. We thank Marco Cuturi for providing the code to produce the starting images of ellipses.

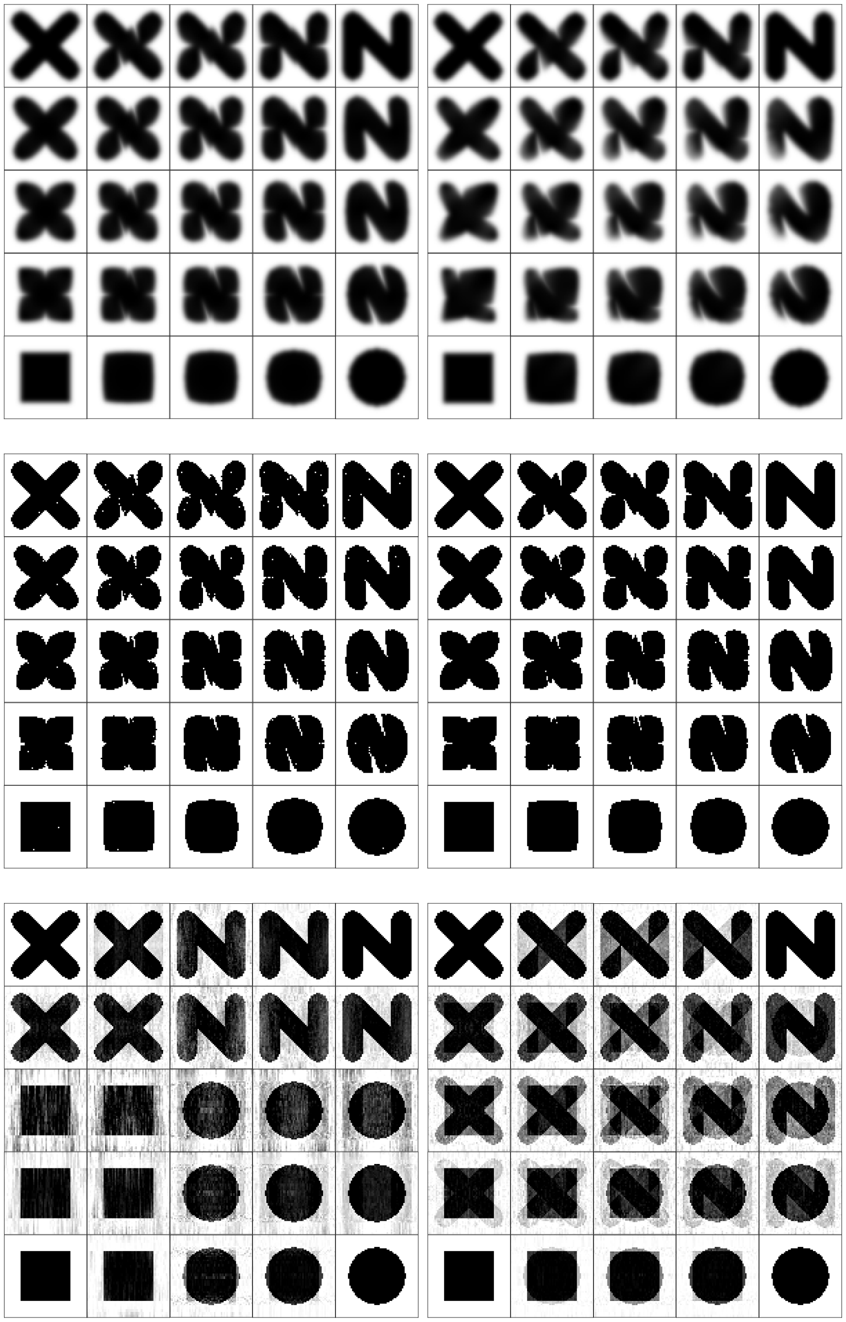

Figure 7: Barycenters of 4 sample images computed for varying weights corresponding to a bilinear interpolation inside a square. **(Top)** IBP barycenters computed with a fixed number of 30 iterations and two different widths of the convolution kernel. Each barycenter is immediately computed. A different discretization of the diffusion kernel might affect the overall convergence of the algorithm, as one can observe from the right picture, where the central barycenter computed for equal weights is not symmetric as it should be. Similarly, symmetry properties for the other barycenters are also affected. **(Center)** ISA barycenters computed with $k = 10000$ (left) and $k = 30000$ (right). On average, the computation of each barycenter takes 12 sec. resp. 118 sec. Despite the higher computation time, one obtains a sharp image of the barycenter by allowing for a higher number of sampling points. **(Bottom)** Sliced Wasserstein barycenters computed with a number of 100 iterations, 100 projections, and two different initial conditions. The computation of each barycenter takes approximately 0.15 sec. Doubling the number of iterations or taking 1000 projections yields a similar picture.

4.2 Clustering and the *k*-barycenter problem

The problem of determining the barycenter of n distributions with respect to the Wasserstein distance has a natural extension in clustering applications to the so-called *k-barycenter problem*. The k-barycenter problem has become popular in computer science since the paper Li and Wang (2008). In this literature barycenters are usually referred to as *centroids* and we borrow this taxonomy in what follows.

Given n probability measures $\mu_1, \dots, \mu_n \in P_2(\mathbb{R}^d)$, the problem is to find k probability measures $\nu_1, \dots, \nu_k$ that are solutions of

$$\inf \left\{ \sum_{i=1}^{k} \sum_{j \in S_i} W_2^2(\mu_j, \nu_i); \nu_1, \dots, \nu_k \in P_2(\mathbb{R}^d) \right\}, \tag{4.1}$$

where, for $1 \leq i \leq k$, we set

$$S_i = \left\{ \mu_j : W_2^2(\mu_j, \nu_i) = \min_{1 \leq r \leq k} W_2^2(\mu_j, \nu_r) \right\}. \tag{4.2}$$

The optimal set of *centroids* (ν_i) minimizes the sum of distances between probability measures and their closest centroid and determines the optimal clusters (S_i) for the corresponding clustering problem.

For the k-barycenter problem one commonly proceeds similarly as Lloyd's k-means clustering for vectors under the Euclidean distance. Given an initial set of centroids $\nu_1^{(1)}, \dots, \nu_k^{(1)}$, one iterates, for $t \geq 1$, the following two steps:

k-means algorithm

1. *Assignment.* Given $\nu_1^{(t)}, \dots, \nu_k^{(t)}$, determine via (4.2) the corresponding clustering $S_1^t, \dots, S_k^t$.

2. *Update.* Calculate the barycenters of $S_1^t, \dots, S_k^t$ to be the set of new centroids $\nu_1^{(t+1)}, \dots, \nu_k^{(t+1)}$.

The above algorithm converges when the clustering in the assignment step no longer changes. In the case of discrete measures, this problem was originally studied in Li and Wang (2008), under the name *D2-clustering*. Recently, a series of approximate algorithms for solving (4.1) have been proposed in the literature; a good summary is given in the introductory part of Ye et al. (2017).

In general, the most onerous step in the k-means algorithm is the computation of the optimal centroid for each cluster at each iteration. For example, Ho et al. (2017) use for this step the algorithm of Cuturi and Doucet (2014), whereas Ye et al. (2017) develop a sophisticated modification of the Bregman alternating direction method of multipliers (B-ADMM) approach for computing the approximate discrete Wasserstein barycenter of large clusters.

Generally speaking, the time complexity per iteration of these methods is $\mathrm{O}(nk^2)$ for the computation of a barycenter of a set of n distributions; see Ye et al. (2017) for more detailed computational details. However, each method based on the k-means algorithm relies on a good starting point for the support of the true centroids; see also Ye and Li (2014) and Irpino et al. (2014).

Thus, we propose to apply the ISA introduced in Section 2 to the k-means algorithm, keeping a quadratic time complexity. We mention that similar extensions can also be given to several related clustering problems like k-means clustering with fixed size, i.e., looking for centroids of the form $\nu = \sum_{i=1}^{k} p_i \delta_{x_i}$, either with p_i fixed (e.g., $p_i = 1/k$) or with free choice of p_i (variable size clusters).

As a first example of the k-barycenter problem, we consider $n = 30$ Gaussian point clouds; see Figure 8. Each cloud consists of $k = 400$ points simulated from a bivariate Gaussian distribution with random correlation and mean chosen with probability $1/2$ as either $(0, 0)'$ (zero mean) or as $(\xi, \xi)'$ (shifted mean). At this point we apply the k-means algorithm illustrated above using the ISA for the assignment and update steps to cluster the point clouds into two clusters (zero and shifted mean). As a starting set of centroids the algorithm selects the two clouds with the furthest sample mean in Euclidean distance. Table 3 shows, for values of the shift ξ between 0 and 1, the number of errors committed by the algorithm in reconstructing the benchmark clustering.

We observe that the ISA allows for a fast construction of clustering also in the case of large clusters. Empirical runs show that the algorithm reaches a stable configuration within 10 iterations. Especially for $\xi < 0.5$, the final clustering might have more errors with respect to the benchmark if compared to the initial assignment. Of course, the accuracy of the algorithm is increasing in ξ, always leading to the correct clustering for $\xi \geq 0.8$.

	number of iterations										
shift	0	1	2	3	4	4	6	7	8	9	10
$\xi = 0.1$	10.20	11.08	11.30	11.38	11.48	11.44	11.42	11.52	11.52	11.52	11.52
$\xi = 0.2$	8.00	9.52	10.06	10.54	10.82	10.96	11.02	11.08	11.14	11.14	11.14
$\xi = 0.3$	5.94	7.04	8.14	8.88	9.26	9.42	9.62	9.68	9.72	9.76	9.76
$\xi = 0.4$	3.82	4.28	5.06	5.18	5.38	5.50	5.56	5.58	5.60	5.60	5.60
$\xi = 0.5$	2.40	2.24	2.26	2.24	2.18	2.16	2.16	2.16	2.16	2.16	2.16
$\xi = 0.6$	1.56	1.46	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42
$\xi = 0.7$	0.92	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
$\xi = 0.8$	0.52	0.22	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\xi = 0.9$	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\xi = 1.0$	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: Average number of errors committed by the ISA in reconstructing the benchmark clustering of 30 Gaussian point clouds with null or shifted mean. Averages are computed over 50 identical runs of the algorithm over a randomized set of point clouds similar to Figure 8. The computation of each number in the table takes on average 9 sec.

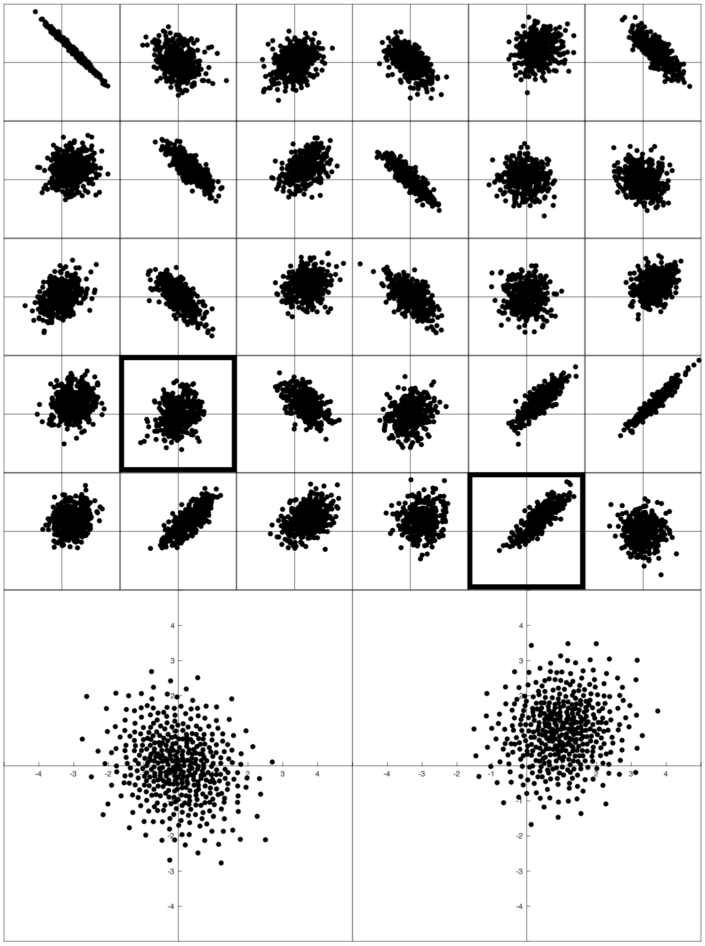

Figure 8: A set of $n = 30$ Gaussian point clouds with random correlation and random shift $\xi = 1$, and the corresponding centroids of the two groups as computed by the ISA. Starting centroids are framed.

As a last example, we explore the use of Wasserstein clusters for clustering of geometrical shapes. We consider the clustering of $n = 300$ shapes, divided into a random number of random ellipses and rectangles; see Figure 9. Each figure is rendered with $k = 400$ points similarly to the ellipses described in Section 4. We apply the ISA with random selection of the initial clusters (centroids). In the recognition of the shapes, the ISA has an average success rate of 79.6% after a fixed number of 10 iterations. The rate is similar if a totally unsupervised initial condition is replaced by taking two different sample shapes as initial centroids (success rate=80.0%), or if the number iterations is halved or doubled (generally after 5 iterations the algorithm converges to a final assignment); see Table 4.

The obtained success rates indicate that the Wasserstein barycenter w.r.t. squared euclidean distance may not be for all geometric structures the best tool for clustering. This is due to the intrinsic structure of the metric used. Depending on the class of shapes/images considered, the Wasserstein distance might deliver better results if each shape/image is first rendered using statistical models based on a set of multivariate feature vectors, as done for instance in Li and Wang (2003, 2008), amongst others. It is difficult to compare these methods with ours as accuracy rates vary a lot depending on the examples considered; see Li and Wang (2003, 2008); Ye et al. (2017) for a comparison. For the kind of geometric problems considered here, we conclude that the ISA is able to very well visually render and distinguish between the different barycenters; see Figure 9.

	n. iterations (comp. time in sec.)			
starting condition	3 (20)	5 (30)	10 (111)	20 (112)
totally random	76.9%	78.9%	79.6%	79.5%
sample shapes	77.7%	79.0%	80.0%	80.1%

Table 4: Average success rate (and computation times) of the ISA in discriminating ellipses from rectangles from a set of 300 random shapes. Averages are computed over 50 identical runs of the algorithm over a randomized set of shapes similar to Figure 9.

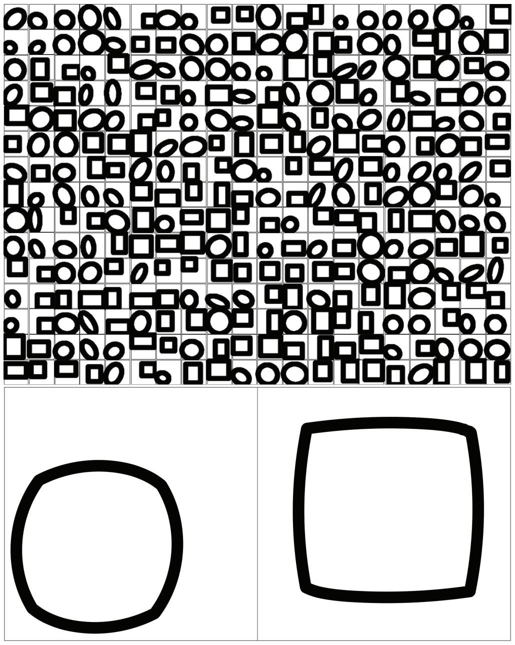

Figure 9: A set of $n = 300$ shapes divided into a random number of random ellipses and rectangles along with the centroids computed by the ISA starting from a random initial configuration. In this example the ISA has a success rate of about 80% of shapes correctly assigned. The time needed for computing the two centroids after 10 iterations is 111 seconds.

5 Final remarks and extensions

In this paper, we introduce and investigate the iterative swapping algorithm (ISA) for computing the Wasserstein barycenter of sums of Dirac masses. The algorithm builds on the equivalence of the barycenter problem to the n-coupling problem (Rüschendorf and Uckelmann, 2002), which we state under general assumptions. For discrete measures, the computation of a Wasserstein barycenter is in principle obtained through the solution of a finite(high)-dimensional linear program. The idea of this paper is to approximate such linear program by a multi-index assignment problem and to use in an iterative manner the swapping algorithm to compute a close-to-optimal solution. The barycenter of general measures can be then approximated by taking sufficiently large finite samples of the measures.

The ISA is based on the easy to evaluate and necessary condition (2.2) providing an appealing quadratic complexity and a sharp image of the support of the barycenter. Remarkably, the ISA is also a completely non-parametric methodology which does not need to be tailored to the specific case study, and always provides more accuracy for increasing resolutions of the marginal inputs. We compare the quality of the barycenter obtained by ISA with the quality obtained using other methods designed to compute Wasserstein barycenters, namely with the sliced Wasserstein barycenters, with the barycenters produced by an entropic regularization of the problem, and with barycenters resulting form linear programming algorithms.

The reduction of complexity of optimization problems by restricting to pairwise swapping steps appears to be applicable to a wider range of problems, such as to clustering and k-barycenter problems in Section 4.2. For 2-coupling problems the quality of this reduction was investigated in Puccetti (2017). We believe that the results in this paper are in general quite promising.

We conclude the paper by pointing out the following remarks.

1. *Different swapping conditions.* One could use different yet equivalent representations of the n-coupling problem to produce different swapping conditions in (2.2). Using the function

$$f(x_1, \dots, x_d) = \Big|\Big| \sum_{i=1}^{n} x_i \Big|\Big|^2.$$

in (1.5), as in the original formulation of the problem, is not computationally efficient due to the presence of the quadratic terms $||x_i||^2$ which are fixed and do not enter the maximization problem.

An equivalent alternative is to maximize the function

$$f(x_1, \dots, x_d) = 2 \sum_{i=1}^{n} \langle x_i, \sum_{j>i} x_j \rangle.$$

Notice that this function induces a swapping condition different than (2.2). Substituting (2.2) with

$$\langle x^i_{\sigma_i(k_1)}, \sum_{j>i} x^j_{\sigma_j(k_1)} \rangle + \langle x^i_{\sigma_i(k_2)}, \sum_{j>i} x^j_{\sigma_j(k_2)} \rangle < \langle x^i_{\sigma_i(k_2)}, \sum_{j>i} x^j_{\sigma_j(k_1)} \rangle + \langle x^i_{\sigma_i(k_1)}, \sum_{j>i} x^j_{\sigma_j(k_2)} \rangle.$$

halves the number of iterations of step 2. delivering slightly less accurate results to the ones presented in the paper.

Alternatively, one could simultaneously compare, for all possible pairs (k_1, k_2) with $1 \leq k_1 < k_2 \leq k$, all 2^{n-1} possible ways of swapping the 2 positions of *each* of the n permutations involved in (2.2) and choose the one delivering the largest function value. This alternative brings slightly more accurate estimates, and less iterations of step 2., but at the cost of a computation time exponentially increasing in n.

In summary, we found that a swapping condition based on the cost function in (1.5) delivers the best trade-off between computation time (to be kept polynomial on n) and accuracy.

2. *Different initial configurations.* In order to decrease the number of iterations of Step 2. of the algorithm, one could start from a given assignment, based on the order of a univariate statistics $g(x^i_j)$, or a random

assignment. For the ellipses applications as illustrated in Section 4, ordering all the k points of each image based on the product of their coordinates nearly halves the number of iterations if compared to a random initial assignment. However, this initial assignment does not deliver any advantage for example in the Gaussian case of Section 3 and we did not find a general rule.

3. *Discrete measures with general probabilities.* In this paper we consider marginal distributions uniformly distributed on a number of points, but the algorithm could also deal with general discrete measures by taking repetitions of points with equal probability. However, this would require an higher number of points for each marginal hence resulting in a considerably higher computation time.

4. *Barycenter of copulas.* It is not always obvious what the barycenter of a set of point clouds looks like; see for instance the illustration given in Figure 10.

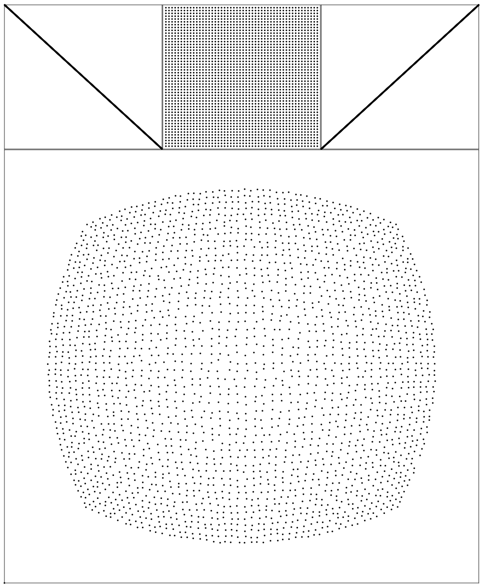

Figure 10: ISA approximation of the Wasserstein barycenter of the lower Fréchet bound M (top-left figure), the independence copula Π (top-middle) and the upper Fréchet bound W (top-right figure), as computed by the ISA. Each image consists of $k = 2500$ points.

References

Agueh, M. and G. Carlier (2011). Barycenters in the Wasserstein space. *SIAM J. Math. Anal. 43*(2), 904–924.

Álvarez-Esteban, P. C., E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán (2016). A fixed-point approach to barycenters in Wasserstein space. *J. Math. Anal. Appl. 441*(2), 744–762.

Anderes, E., S. Borgwardt, and J. Miller (2016). Discrete Wasserstein barycenters: optimal transport for discrete data. *Math. Methods Oper. Res. 84*(2), 389–409.

Benamou, J.-D., G. Carlier, M. Cuturi, L. Nenna, and G. Peyré (2015). Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput. 37*(2), A1111–A1138.

Birkhoff, G. (1946). Tres observaciones sobre el algebra lineal. *Rev. Univ. Nac. Tucumán (A) 5*, 147–151.

Bonneel, N., J. Rabin, G. Peyré, and H. Pfister (2015). Sliced and radon Wasserstein barycenters of measures. *J. Math. Imaging Vis. 51*, 22–45.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math. 44*(4), 375–417.

Burkard, R., M. Dell'Amico, and S. Martello (2009). *Assignment Problems.* SIAM, Philadelphia PA.

Carlier, G., A. Oberman, and E. Oudet (2015). Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM Math. Model. Numer. Anal. 49*(6), 1621–1642.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transportation. In *Proceedings of the Neural Information Processing Systems Conference*, 26, pp. 2292–2300.

Cuturi, M. and A. Doucet (2014). Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 685–693.

Dowson, D. C. and B. V. Landau (1982). The Fréchet distance between multivariate normal distributions. *J. Multivariate Anal. 12*(3), 450–455.

Gangbo, W. and A. Święch (1998). Optimal maps for the multidimensional Monge-Kantorovich problem. *Comm. Pure Appl. Math. 51*(1), 23–45.

Ho, N., X. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung (2017). Multilevel clustering via Wasserstein means. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1501–1509.

Irpino, A., R. Verde, and F. de A.T. De Carvalho (2014). Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. *Expert Syst. Appl. 41*(7), 3351 – 3366.

Kim, Y.-H. and B. Pass (2014). A general condition for Monge solutions in the multi-marginal optimal transport problem. *SIAM J. Math. Anal. 46*(2), 1538–1550.

Knott, M. and C. S. Smith (1994). On a generalization of cyclic monotonicity and distances among random vectors. *Linear Algebra and its Applications 199*, 363–371.

Le Gouic, T. and J.-M. Loubes (2017). Existence and consistency of Wasserstein barycenters. *Probab. Theory Related Fields 168*(3-4), 901–917.

Li, J. and J. Z. Wang (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell. 25*(9), 1075–1088.

Li, J. and J. Z. Wang (2008). Real-time computerized annotation of pictures. *IEEE Trans. Pattern Anal. Mach. Intell. 30*(6), 985–1002.

Oberman, A. M. and Y. Ruan (2015). An efficient linear programming method for optimal transportation. Available at https://arxiv.org/abs/1509.03668.

Peyré, G. and M. Cuturi (2019). Computational optimal transport. *Foundations and Tends in Machine Learning 11*(5–6), 355–607.

Puccetti, G. (2017). An algorithm to approximate the optimal expected inner product of two vectors with given marginals. *J. Math. Anal. Appl. 451*, 132–145.

Rabin, J., G. Peyré, J. Delon, and M. Bernot (2012). Wasserstein barycenter and its application to texture mixing. In A. M. Bruckstein, B. M. ter Haar Romeny, A. M. Bronstein, and M. M. Bronstein (Eds.), *Scale Space and Variational Methods in Computer Vision*, pp. 435–446. Springer, Berlin.

Rachev, S. T. and L. Rüschendorf (1998). *Mass Transportation Problems. Vol. I–II.* Springer-Verlag, New York.

Rüschendorf, L. and S. T. Rachev (1990). A characterization of random variables with minimum L^2-distance. *J. Multivariate Anal. 32*(1), 48–54.

Rüschendorf, L. and L. Uckelmann (1997). On optimal multivariate couplings. In V. Benes and I. Stepan (Eds.), *Distributions with given Marginals and Moment Problems*, pp. 261–273. Springer.

Rüschendorf, L. and L. Uckelmann (2002). On the n-coupling problem. *J. Multivariate Anal. 81*(2), 242–258.

Schmitzer, B. (2016). Stabilized sparse scaling algorithms for entropy regularized transport problems. Available at https://arxiv.org/abs/1610.06519.

Solomon, J., F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas (2015). Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (Proc. SIGGRAPH 2015) 34*(4), 66:1–66:11.

Ye, J. and J. Li (2014). Scaling up discrete distribution clustering using ADMM. In *IEEE Inter. Conf. Image Proces.*, pp. 5267–5271.

Ye, J., P. Wu, J. Z. Wang, and J. Li (2017). Fast discrete distribution clustering using Wasserstein barycenter with sparse support. *IEEE Trans. Signal Process. 65*(9), 2317–2332.