Fast Computation of Wasserstein Barycenters

Marco Cuturi MCUTURI@I.KYOTO-U.AC.JP
Graduate School of Informatics, Kyoto University

Arnaud Doucet DOUCET@STAT.OXFORD.AC.UK
Department of Statistics, University of Oxford

Abstract

We present new algorithms to compute the mean of a set of empirical probability measures under the optimal transport metric. This mean, known as the Wasserstein barycenter, is the measure that minimizes the sum of its Wasserstein distances to each element in that set. We propose two original algorithms to compute Wasserstein barycenters that build upon the subgradient method. A direct implementation of these algorithms is, however, too costly because it would require the repeated resolution of large primal and dual optimal transport problems to compute subgradients. Extending the work of Cuturi (2013), we propose to smooth the Wasserstein distance used in the definition of Wasserstein barycenters with an entropic regularizer and recover in doing so a strictly convex objective whose gradients can be computed for a considerably cheaper computational cost using matrix scaling algorithms. We use these algorithms to visualize a large family of images and to solve a constrained clustering problem.

1. Introduction

Comparing, summarizing and reducing the dimensionality of empirical probability measures defined on a space Ω are fundamental tasks in statistics and machine learning. Such tasks are usually carried out using pairwise comparisons of measures. Classic information divergences (Amari and Nagaoka, 2001) are widely used to carry out such comparisons.

Unless Ω is finite, these divergences cannot be directly applied to empirical measures, because they are ill-defined for measures that do not have continuous densities. They also fail to incorporate prior knowledge on the geometry

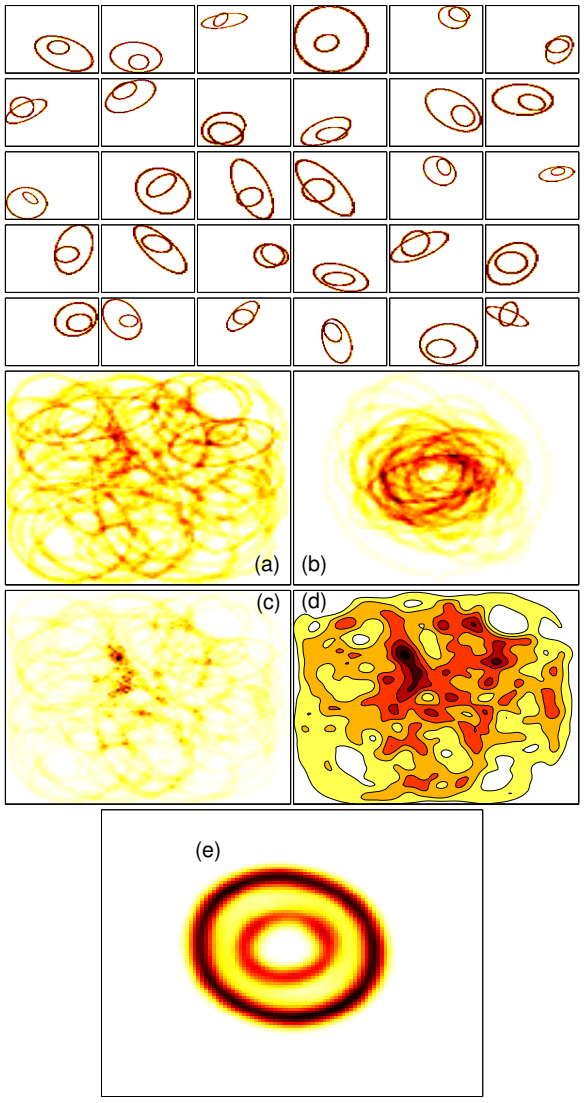

Figure 1. (Top) 30 images of two nested random ellipses. Mean measures using the (a) Euclidean distance (b) Euclidean after re-centering images (c) Jeffrey centroid (Nielsen, 2013) (d) RKHS distance (Gaussian kernel, $\sigma = 0.002$) (e) 2-Wasserstein distance.

(2016.01.21) Errata: corrected typos in Eq.(7) (added factor 2); def. of $\mathbf{d}_\lambda$ (added -1), sign for $\alpha^\star_\lambda$ in Prop. 2 and Alg.3.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

of Ω, which might be available if, for instance, Ω is also a Hilbert space. Both of these issues are usually solved using Parzen's approach (1962) to smooth empirical measures with smoothing kernels before computing divergences: the Euclidean (Gretton et al., 2007) and χ_2 distances (Harchaoui et al., 2008), the Kullback-Leibler and Pearson divergences (Kanamori et al., 2012a;b) can all be computed fairly efficiently by considering matrices of kernel evaluations.

The choice of a divergence defines implicitly the *mean* element, or barycenter, of a set of measures, as the particular measure that minimizes the sum of all its divergences to that set of target measures (Veldhuis, 2002; Banerjee et al., 2005; Teboulle, 2007; Nielsen, 2013). The goal of this paper is to compute efficiently barycenters (possibly in a constrained subset of all probability measures on Ω) defined by the *optimal transport distance* between measures (Villani, 2009, §6). We propose to minimize directly the sum of optimal transport distances from one measure (the variable) to a set of fixed measures by gradient descent. These gradients can be computed for a moderate cost by solving smoothed optimal transport problems as proposed by Cuturi (2013).

Wasserstein distances have many favorable properties, documented both in theory (Villani, 2009) and practice (Rubner et al., 1997; Pele and Werman, 2009). We argue that their versatility extends to the barycenters they define. We illustrate this intuition in Figure 1, where we consider 30 images of nested ellipses on a 100×100 grid. Each image is a discrete measure on $[0,1]^2$ with normalized intensities. Computing the Euclidean, Gaussian RKHS mean-maps or Jeffrey centroid of these images results in mean measures that hardly make any sense, whereas the 2-Wasserstein mean on that grid (defined in §3.1) produced by Algorithm 1 captures perfectly the structure of these images. Note that these results were recovered without any prior knowledge on these images other than that of defining a distance in $[0,1]^2$, here the Euclidean distance. Note also that the Gaussian kernel smoothing approach uses the same distance, in addition to a bandwidth parameter σ which needs to be tuned in practice.

This paper is organized as follows: we provide background on optimal transport in §2, followed by the definition of Wasserstein barycenters with motivating examples in §3. Novel contributions are presented from §4: we present two subgradient methods to compute Wasserstein barycenters, one which applies when the support of the mean measure is known in advance and another when that support can be freely chosen in Ω. These algorithms are very costly even for measures of small support or histograms of small size. We show in §5 that the key ingredients of these approaches—the computation of primal and dual optimal transport solutions—can be bypassed by solving smoothed optimal transport problems. We conclude with two applications of our algorithms in §6.

2. Background on Optimal Transport

Let Ω be an arbitrary space, D a metric on that space and $P(\Omega)$ the set of Borel probability measures on Ω. For any point $x \in \Omega$, δ_x is the Dirac unit mass on x.

Definition 1 (Wasserstein Distances)**.** *For $p \in [1,\infty)$ and probability measures μ,ν in $P(\Omega)$, their p-Wasserstein distance (Villani, 2009, §6) is*

$$W_p(\mu,\nu) \stackrel{\text{def}}{=} \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega^2} D(x,y)^p d\pi(x,y) \right)^{1/p},$$

where $\Pi(\mu,\nu)$ is the set of all probability measures on Ω^2 that have marginals μ and ν.

2.1. Restriction to Empirical Measures

We will only consider empirical measures throughout this paper, that is measures of the form $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ where n is an integer, $X = (x_1,\ldots,x_n) \in \Omega^n$ and $(a_1,\ldots,a_n)$ lives in the probability simplex Σ_n,

$$\Sigma_n \stackrel{\text{def}}{=} \{u \in \mathbb{R}^n \mid \forall i \leq n, u_i \geq 0, \sum_{i=1}^n u_i = 1\}.$$

Let us introduce additional notations:

Measures on a Set X with Constrained Weights. Let Θ be a non-empty closed subset Θ of Σ_n. We write

$$P(X,\Theta) \stackrel{\text{def}}{=} \{\mu = \sum_{i=1}^n a_i \delta_{x_i},\ a \in \Theta\}.$$

Measures supported on up to k points. Given an integer k and a subset Θ of Σ_k, we consider the set $P_k(\Omega,\Theta)$ of measures of Ω that have discrete support of size up to k and weights in Θ,

$$P_k(\Omega,\Theta) \stackrel{\text{def}}{=} \bigcup_{X \in \Omega^k} P(X,\Theta).$$

When no constraints on the weights are considered, namely when the weights are free to be chosen anywhere on the probability simplex, we use the shorter notations $P(X) \stackrel{\text{def}}{=} P(X,\Sigma_n)$ and $P_k(\Omega) \stackrel{\text{def}}{=} P_k(\Omega,\Sigma_k)$.

2.2. Wasserstein & Discrete Optimal Transport

Consider two families $X = (x_1,\ldots,x_n)$ and $Y = (y_1,\ldots,y_m)$ of points in Ω. When $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{i=1}^n b_i \delta_{y_i}$, the Wasserstein distance $W_p(\mu,\nu)$ between μ and ν is the p^{th} root of the optimum of a network flow problem known as the *transportation problem* (Bertsimas and Tsitsiklis, 1997, §7.2). This problem builds upon two elements: the ***matrix M_{XY} of pairwise distances*** between elements of X and Y raised to the power p, which

acts as a cost parameter,

$$M_{XY} \stackrel{\text{def}}{=} [D(x_i, y_j)^p]_{ij} \in \mathbb{R}^{n \times m}, \quad (1)$$

and the ***transportation polytope*** $U(a, b)$ of $a \in \Sigma_n$ and $b \in \Sigma_m$, which acts as a feasible set, defined as the set of $n \times m$ nonnegative matrices such that their row and column marginals are equal to a and b respectively. Writing $\mathbb{1}_n$ for the n-dimensional vector of ones,

$$U(a, b) \stackrel{\text{def}}{=} \{T \in \mathbb{R}_+^{n \times m} \mid T\mathbb{1}_m = a,\, T^T\mathbb{1}_n = b\}. \quad (2)$$

Let $\langle A, B \rangle \stackrel{\text{def}}{=} \operatorname{tr}(A^T B)$ be the Frobenius dot-product of matrices. Combining Eq. (1) & (2), we have that $W_p^p(\mu, \nu)$—the distance $W_p(\mu, \nu)$ raised to the power p—can be written as the optimum of a parametric linear program $\mathbf{p}$ on $n \times m$ variables, parameterized by the marginals a, b and a (cost) matrix M_{XY}:

$$W_p^p(\mu, \nu) = \mathbf{p}(a, b, M_{XY}) \stackrel{\text{def}}{=} \min_{T \in U(a,b)} \langle T, M_{XY} \rangle. \quad (3)$$

3. Wasserstein Barycenters

We present in this section the Wasserstein barycenter problem, a variational problem involving all Wasserstein distances from one to many measures, and show how it encompasses known problems in clustering and approximation.

3.1. Definition and Special Cases

Definition 2 (Agueh and Carlier, 2011)**.** *A Wasserstein barycenter of N measures $\{\nu_1, \ldots, \nu_N\}$ in $\mathbb{P} \subset P(\Omega)$ is a minimizer of f over $\mathbb{P}$, where*

$$f(\mu) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} W_p^p(\mu, \nu_i). \quad (4)$$

Agueh and Carlier consider more generally a non-negative weight λ_i in front of each distance $W_p^p(\mu, \nu_i)$. The algorithms we propose extend trivially to that case but we use uniform weights in this work to keep notations simpler.

We highlight a few special cases where minimizing f over a set $\mathbb{P}$ is either trivial, relevant to data analysis and/or has been considered in the literature with different tools or under a different name. In what follows $X \in \Omega^n$ and $Y \in \Omega^m$ are arbitrary finite subsets of Ω.

• $N = 1, \mathbb{P} = P(X)$ When only one measure ν, supported on $Y \in \Omega^m$ is considered, its closest element μ in $P(X)$—if no constraints on weights a are given—can be computed by defining a weight vector a on the elements of X that results from assigning all of the mass b_i to the closest neighbor in metric D of y_i in X.

• **Centroids of Histograms**: $N > 1, \Omega$ finite, $\mathbb{P} = P(\Omega)$. When Ω is a set of size d and a matrix $M \in \mathbb{R}_+^{d \times d}$ describes the pairwise distances between these d points (usually called in that case bins or features), the 1-Wasserstein distance is known as the Earth Mover's Distance (EMD) (Rubner et al., 1997). In that context, Wasserstein barycenters have also been called EMD prototypes by Zen and Ricci (2011).

• **Euclidean** Ω: $N = 1$, $D(x, y) = \|x - y\|_2, p = 2, \mathbb{P} = P_k(\Omega)$. Minimizing f on $P_k(\Omega)$ when (Ω, D) is a Euclidean metric space and $p = 2$ is equivalent to the k-means problem (Pollard, 1982; Canas and Rosasco, 2012).

• **Constrained k-Means**: $N = 1$, $\mathbb{P} = P_k(\Omega, \{\mathbb{1}_k/k\})$. Consider a measure ν with support $Y \in \Omega^m$ and weights $b \in \Sigma_m$. The problem of approximating this measure by a uniform measure with k atoms—a measure in $P_k(\Omega, \{\mathbb{1}_k/k\})$—in 2-Wasserstein sense was to our knowledge first considered by Ng (2000), who proposed a variant of Lloyd's algorithm (1982) for that purpose. More recently, Reich (2013) remarked that such an approximation can be used in the resampling step of particle filters and proposed in that context two ensemble methods inspired by optimal transport, one of which reduces to a single iteration of Ng's algorithm. Such approximations can also be obtained with kernel-based approaches, by minimizing an information divergence between the (smoothed) target measure ν and its (smoothed) uniform approximation as proposed recently by Chen et al. (2010) and Sugiyama et al. (2011).

3.2. Recent Work

Agueh and Carlier (2011) consider conditions on the ν_i's for a Wasserstein barycenter in $P(\Omega)$ to be unique using the multi-marginal transportation problem. They provide solutions in the cases where either (i) $\Omega = \mathbb{R}$; (ii) $N = 2$ using McCann's interpolant (1997); (iii) all the measures ν_i are Gaussians in $\Omega = \mathbb{R}^d$, in which case the barycenter is a Gaussian with the mean of all means and a variance matrix which is the unique positive definite root of a matrix equation (Agueh and Carlier, 2011, Eq.6.2).

Rabin et al. (2012) were to our knowledge the first to consider practical approaches to compute Wasserstein barycenters between point clouds in $\mathbb{R}^d$. To do so, Rabin et al. (2012) propose to approximate the Wasserstein distance between two point clouds by their *sliced* Wasserstein distance, the expectation of the Wasserstein distance between the projections of these point clouds on lines sampled randomly. Because the optimal transport between two point clouds on the real line can be solved with a simple sort, the sliced Wasserstein barycenter can be computed very efficiently, using gradient descent. Although their approach seems very effective in lower dimensions, it may not work for $d \geq 4$ and does not generalize to non-Euclidean metric spaces.

4. New Computational Approaches

We propose in this section new approaches to compute Wasserstein barycenters when (i) each of the N measures ν_i is an empirical measure, described by a list of atoms $Y_i \in \Omega^{m_i}$ of size $m_i \geq 1$, and a probability vector b_i in the simplex Σ_{m_i}; (ii) the search for a barycenter is not considered on the whole of $P(\Omega)$ but restricted to either $P(X, \Theta)$ (the set of measures supported on a predefined finite set X of size n with weights in a subset Θ of Σ_n) or $P_k(\Omega, \Theta)$ (the set of measures supported on up to k atoms with weights in a subset Θ of Σ_k).

Looking for a barycenter μ with atoms X and weights a is equivalent to minimizing f (see Eq. 3 for a definition of $\mathbf{p}$),

$$f(a, X) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \mathbf{p}(a, b_i, M_{XY_i}), \tag{5}$$

over relevant feasible sets for a and X. When X is *fixed*, we show in §4.1 that f is convex w.r.t a regardless of the properties of Ω. A subgradient for f w.r.t a can be recovered through the *dual optimal solutions* of all problems $\mathbf{p}(a, b_i, M_{XY_i})$, and f can be minimized using a projected subgradient method outlined in §4.2. If X is *free*, constrained to be of cardinal k, and Ω and its metric D are both *Euclidean*, we show in §4.4 that f is not convex w.r.t X but we can provide subgradients for f using the *primal optimal solutions* of all problems $\mathbf{p}(a, b_i, M_{XY_i})$. This in turn suggests an algorithm to reach a local minimum for f w.r.t. a and X in $P_k(\Omega, \Theta)$ by combining both approaches.

4.1. Differentiability of $\mathbf{p}(a, b, M_{XY})$ w.r.t a

Dual transportation problem. Given a matrix $M \in \mathbb{R}^{n \times m}$, the optimum $\mathbf{p}(a, b, M)$ admits the following dual Linear Program (LP) form (Bertsimas and Tsitsiklis, 1997, §7.6,§7.8), known as the dual optimal transport problem:

$$\mathbf{d}(a, b, M) = \max_{(\alpha,\beta) \in C_M} \alpha^T a + \beta^T b\,, \tag{6}$$

where the polyhedron C_M of dual variables is

$$C_M = \{(\alpha, \beta) \in \mathbb{R}^{n+m} \,|\, \alpha_i + \beta_j \leq m_{ij}\}.$$

By LP duality, $\mathbf{d}(a, b, M) = \mathbf{p}(a, b, M)$. The dual optimal solutions—which can be easily recovered from the primal optimal solution (Bertsimas and Tsitsiklis, 1997, Eq.7.10)—define a subgradient for $\mathbf{p}$ as a function of a:

Proposition 1. *Given $b \in \Sigma_m$ and $M \in \mathbb{R}^{n \times m}$, the map $a \mapsto \mathbf{p}(a, b, M)$ is a polyhedral convex function. Any optimal dual vector $\alpha^\star$ of $\mathbf{d}(a, b, M)$ is a subgradient of $\mathbf{p}(a, b, M)$ with respect to a.*

Proof. These results follow from sensitivity analysis in LP's (Bertsimas and Tsitsiklis, 1997, §5.2). $\mathbf{d}$ is bounded and is also the maximum of a finite set of linear functions, each indexed by the set of extreme points of C_M, evaluated at a and is therefore polyhedral convex. When the dual optimal vector is unique, $\alpha^\star$ is a gradient of $\mathbf{p}$ at a, and a subgradient otherwise. □

Because for any real value t the pair $(\alpha + t\mathbb{1}_n, \beta - t\mathbb{1}_m)$ is feasible if the pair (α, β) is feasible, and because their objective are identical, any dual optimum (α, β) is determined up to an additive constant. To remove this degree of freedom—which arises from the fact that one among all $n+m$ row/column sum constraints of $U(a, b)$ is redundant—we can either remove a dual variable or normalize any dual optimum $\alpha^\star$ so that it sums to zero, to enforce that it belongs to the tangent space of Σ_n. We follow the latter strategy in the rest of the paper.

4.2. Fixed Support: Minimizing f over $P(X)$

Let $X \subset \Omega^n$ be fixed and let Θ be a closed convex subset of Σ_n. The aim of this section is to compute weights $a \in \Theta$ such that $f(a, X)$ is minimal. Let $\alpha_i^\star$ be the optimal dual variable of $\mathbf{d}(a, b_i; M_{XY_i})$ normalized to sum to 0. f being a sum of terms $\mathbf{p}(a, b_i, M_{XY_i})$, we have that:

Corollary 1. *The function $a \mapsto f(a, X)$ is polyhedral convex, with subgradient*

$$\boldsymbol{\alpha} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \alpha_i^\star.$$

Assuming Θ is closed and convex, we can consider a naive projected subgradient minimization of f. Alternatively, if there exists a Bregman divergence $B(a, b) = \omega(b) - \omega(a) - \langle \nabla\omega(a), b - a \rangle$ for $a, b \in \Theta$ defined by a prox-function ω, we can define the proximal mapping $P_a(b) = \operatorname{argmin}_{c \in \Theta} (\langle b, c - a \rangle + B(a, c))$ and consider accelerated gradient approaches (Nesterov, 2005). We summarize this idea in Algorithm 1.

Algorithm 1 Wasserstein Barycenter in $P(X, \Theta)$

Inputs: $X \in \Omega^n, \Theta \subset \Sigma_n$. For $i \leq N : Y_i \in \Omega^{m_i}, b_i \in \Sigma_{m_i}, p \in [1, \infty), t_0 > 0$.
Form all $n \times m_i$ matrices $M_i = M_{XY_i}$, see Eq. (1).
Set $\hat{a} = \tilde{a} = \operatorname{argmin}_\Theta \omega$.
while not converged **do**
  $\beta = (t + 1)/2$, $a \leftarrow (1 - \beta^{-1})\hat{a} + \beta^{-1}\tilde{a}$.
  Form subgradient $\boldsymbol{\alpha} \leftarrow N^{-1} \sum_{i=1}^{N} \alpha_i^\star$ using all dual optima $\alpha_i^\star$ of $\mathbf{d}(a, b_i, M_i)$.
  $\tilde{a} \leftarrow P_a(t_0 \beta \boldsymbol{\alpha})$.
  $\hat{a} \leftarrow (1 - \beta^{-1})\hat{a} + \beta^{-1}\tilde{a}$, $t \leftarrow t + 1$.
end while

Notice that when $\Theta = \Sigma_n$ and B is the Kullback-Leibler divergence (Beck and Teboulle, 2003), we can initialize $\tilde{a}$

with $\mathbb{1}_n/n$ and use the multiplicative update to realize the proximal update: $\tilde{a} \leftarrow \tilde{a} \circ e^{-t_0\beta\boldsymbol{\alpha}}; \tilde{a} \leftarrow \tilde{a}/\tilde{a}^T\mathbb{1}_n$, where $\circ$ is Schur's product. Alternative sets Θ for which this projection can be easily carried out include, for instance, all (convex) level set of the entropy function H, namely $\Theta = \{a \in \Sigma_n | H(a) \geq \tau\}$ where $0 \leq \tau \leq \log n$.

4.3. Differentiability of $\mathbf{p}(a, b, M_{XY})$ w.r.t X

We consider now the case where $\Omega = \mathbb{R}^d$ with $d \geq 1$, D is the Euclidean distance and $p = 2$. When $\Omega = \mathbb{R}^d$, a family of n points X and a family of m points Y can be represented respectively as a matrix in $\mathbb{R}^{d\times n}$ and another in $\mathbb{R}^{d\times m}$. The pairwise squared-Euclidean distances between points in these sets can be recovered by writing $\mathbf{x} \stackrel{\text{def}}{=} \operatorname{diag}(X^TX)$ and $\mathbf{y} \stackrel{\text{def}}{=} \operatorname{diag}(Y^TY)$, and observing that

$$M_{XY} = \mathbf{x}\mathbb{1}_m^T + \mathbb{1}_n\mathbf{y}^T - 2X^TY \in \mathbb{R}^{n\times m}.$$

Transport Cost as a function of X. Due to the margin constraints that apply if a matrix T is in the polytope $U(a, b)$, we have:

$$\begin{aligned}\langle T, M_{XY}\rangle &= \langle T, \mathbf{x}\mathbb{1}_d^T + \mathbb{1}_d^T\mathbf{y} - 2X^TY\rangle \\ &= \operatorname{tr} T^T\mathbf{x}\mathbb{1}_d^T + \operatorname{tr} T^T\mathbb{1}_d^T\mathbf{y} - 2\langle T, X^TY\rangle \\ &= \mathbf{x}^Ta + \mathbf{y}^Tb - 2\langle T, X^TY\rangle.\end{aligned}$$

Discarding constant terms in $\mathbf{y}$ and b, we have that minimizing $\mathbf{p}(a, b, M_{XY})$ with respect to locations X is equivalent to solving

$$\min_{X\in\mathbb{R}^{d\times k}} \mathbf{x}^Ta + 2\mathbf{p}(a, b, -X^TY). \tag{7}$$

As a function of X, that objective is the sum of a convex quadratic function of X with a piecewise linear concave function, since

$$\mathbf{p}(a, b, -X^TY) = \min_{T\in U(a,b)} \langle X, -YT^T\rangle$$

is the minimum of linear functions indexed by the vertices of the polytope $U(a, b)$. As a consequence, $\mathbf{p}(a, b, M_{XY})$ is not convex with respect to X.

Quadratic Approximation. Suppose that $T^\star$ is optimal for problem $\mathbf{p}(a, b, M_{XY})$. Updating Eq. (7),

$$\begin{aligned}\mathbf{x}^Ta - 2\langle T^\star, X^TY\rangle = \|X\operatorname{diag}(a^{1/2}) - YT^{\star T}\operatorname{diag}(a^{-1/2})\|^2 \\ - \|YT^{\star T}\operatorname{diag}(a^{-1/2})\|^2.\end{aligned}$$

Minimizing a local quadratic approximation of $\mathbf{p}$ at X yields thus the Newton update

$$X \leftarrow YT^{\star T}\operatorname{diag}(a^{-1}). \tag{8}$$

A simple interpretation of this update is as follows: the matrix $T^{\star T}\operatorname{diag}(a^{-1})$ has n column-vectors in the simplex Σ_m. The suggested update for X is to replace it by n barycenters of points enumerated in Y with weights defined by the optimal transport $T^\star$. Note that, because the minimization problem we consider in X is not convex to start with, one could be fairly creative when it comes to choosing D and p among other distances and exponents. This substitution would only involve more complicated gradients of M_{XY} w.r.t. X that would appear in Eq. (7).

4.4. Free Support: Minimizing f over $P_k(\mathbb{R}^d, \Theta)$

We now consider, as a natural extension of §4.2 when $\Omega = \mathbb{R}^d$, the problem of minimizing f over a probability measure μ that is (i) supported by *at most k atoms* described in X, a matrix of size $d \times k$, (ii) with weights in $a \in \Theta \subset \Sigma_k$.

Alternating Optimization. To obtain an approximate minimizer of $f(a, X)$ we propose in Algorithm 2 to update alternatively locations X (with the Newton step defined in Eq. 8) and weights a (with Algorithm 1).

Algorithm 2 2-Wasserstein Barycenter in $P_k(\mathbb{R}^d, \Theta)$

Input: $Y_i \in \mathbb{R}^{d\times m_i}, b_i \in \Sigma_{m_i}$ for $i \leq N$.
initialize $X \in \mathbb{R}^{d\times k}$ and $a \in \Theta$
while X and a have not converged **do**
  $a \leftarrow a^\star$ using Algorithm 1.
  for $i \in (1, \dots, N)$ **do**
    $T_i^\star \leftarrow$ optimal solution of $\mathbf{p}(a, b_i; M_{XY_i})$
  end for
  $X \leftarrow (1-\theta)X + \theta\left(\frac{1}{N}\sum_{i=1}^N Y_iT_i^{\star T}\right)\operatorname{diag}(a^{-1})$, setting $\theta \in [0, 1]$ with line-search or a preset value.
end while

Algorithm 2 and Lloyd/Ng Algorithms. As mentioned in §2, minimizing f defined in Eq. (5) over $P_k(\mathbb{R}^d)$, with $N = 1$, $p = 2$ and no constraints on the weights ($\Theta = \Sigma_k$), is equivalent to solving the k-means problem applied to the set of points enumerated in ν_1. In that particular case, Algorithm 2 is also equivalent to Lloyd's algorithm. Indeed, the assignment of the weight of each point to its closest centroid in Lloyd's algorithm (the maximization step) is equivalent to the computation of $a^\star$ in ours, whereas the re-centering step (the expectation step) is equivalent to our update for X using the optimal transport, which is in that case the trivial transport that assigns the weight (divided by N) of each atom in Y_i to its closest neighbor in X. When the weight vector a is constrained to be uniform ($\Theta = \{\mathbb{1}_k/k\}$), Ng (2000) proposed a heuristic to obtain uniform k-means that is also equivalent to Algorithm 2, and which also relies on the repeated computation of optimal transports. For more general sets Θ, Algorithm 1 ensures that the weights a remain in Θ at each iteration of Algorithm 2, which cannot be guaranteed by neither Lloyd's nor Ng's approach.

Algorithm 2 and Reich's (2013) Transform. Reich (2013) has recently suggested to approximate a weighted measure ν by a uniform measure supported on as many atoms. This approximation is motivated by optimal transport theory, notably asymptotic results by McCann (1995), but does not attempt to minimize, as we do in Algorithm 2, any Wasserstein distance between that approximation and the original measure. This approach results in one application of the Newton update defined in Eq. (8), when X is first initialized to Y and $a = \mathbb{1}_m/m$ to compute the optimal transport $T^\star$.

Summary We have proposed two original algorithms to compute Wasserstein barycenters of probability measures: one which applies when the support of the barycenter is fixed and its weights are constrained to lie in a convex subset Θ of the simplex, another which can be used when the support can be chosen freely. These algorithms are relatively simple, yet—to the best of our knowledge—novel. We suspect these approaches were not considered before because of their prohibitive computational cost: Algorithm 1 computes at each iteration the dual optima of N transportation problems to form a subgradient, each with $n + m_i$ variables and $n \times m_i$ inequality constraints. Algorithm 2 incurs an even higher cost, since it involves running Algorithm 1 at each iteration, in addition to solving N primal optimal transport problems to form a subgradient to update X. Since both objectives rely on subgradient descent schemes, they are also likely to suffer from a very slow convergence. We propose to solve these issues by following Cuturi's approach (2013) to smooth the objective f and obtain strictly convex objectives whose gradients can be computed more efficiently.

5. Smoothed Dual and Primal Problems

To circumvent the major computational roadblock posed by the repeated computation of primal and dual optimal transports, we extend Cuturi's approach (2013) to obtain smooth and strictly convex approximations of both primal and dual problems $\mathbf{p}$ and $\mathbf{d}$. The matrix scaling approach advocated by Cuturi was motivated by the fact that it provided a fast approximation $\mathbf{p}_\lambda$ to $\mathbf{p}$. We show here that the same approach can be used to smooth the objective f and recover for a cheap computational price its gradients w.r.t. a and X.

5.1. Regularized Primal and Smoothed Dual

A $n \times m$ transport T, which is by definition in the nm-simplex, has entropy $h(T) \stackrel{\text{def}}{=} -\sum_{i,j=1}^{n,m} t_{ij} \log(t_{ij})$. Cuturi (2013) has recently proposed to consider, for $\lambda > 0$, a regularized primal transport problem $\mathbf{p}_\lambda$ as

$$\mathbf{p}_\lambda(a, b; M) = \min_{T \in U(a,b)} \langle X, M \rangle - \frac{1}{\lambda} h(T).$$

We introduce in this work its dual problem, which is a smoothed version of the original dual transportation problem, where the positivity constraints of each term $m_{ij} - \alpha_i - \beta_j$ have been replaced by penalties $\frac{1}{\lambda} e^{-\lambda(m_{ij} - \alpha_i - \beta_j)}$:

$$\mathbf{d}_\lambda(a, b; M) = \max_{(\alpha,\beta) \in \mathbb{R}^{n+m}} \alpha^T a + \beta^T b - \sum_{i \leq n, j \leq m} \frac{e^{\lambda(\alpha_i + \beta_j - m_{ij}) - 1}}{\lambda}.$$

These two problems are related below in the sense that their respective optimal solutions are linked by a unique positive vector $u \in \mathbb{R}^n_+$:

Proposition 2. *Let K be the elementwise exponential of $-\lambda M_{XY}$, $K \stackrel{\text{def}}{=} e^{-\lambda M_{XY}}$. Then there exists a pair of vectors $(u, v) \in \mathbb{R}^n_+ \times \mathbb{R}^m_+$ such that the optimal solutions of $\mathbf{p}_\lambda$ and $\mathbf{d}_\lambda$ are respectively given by*

$$T^\star_\lambda = \operatorname{diag}(u) K \operatorname{diag}(v), \alpha^\star_\lambda = \frac{\log(u)}{\lambda} - \frac{\log(u)^T \mathbb{1}_n}{\lambda n} \mathbb{1}_n.$$

Proof. The result follows from the Lagrange method of multipliers for the primal as shown by Cuturi (2013, Lemma 2), and a direct application of first-order conditions for the dual, which is an unconstrained convex problem. The term $\frac{\log(u)^T \mathbb{1}_n}{\lambda n} \mathbb{1}_n$ in the definition of $\alpha^\star_\lambda$ is used to normalize $\alpha^\star_\lambda$ so that it sums to zero as discussed in the end of §4.1. □

5.2. Matrix Scaling Computation of (u, v)

The positive vectors (u, v) mentioned in Proposition 2 can be computed through Sinkhorn's matrix scaling algorithm applied to K, as outlined in Algorithm 3:

Lemma 1 (Sinkhorn, 1967)**.** *For any positive matrix A in $\mathbb{R}^{n \times m}_+$ and positive probability vectors $a \in \Sigma_n$ and $b \in \Sigma_m$, there exist positive vectors $u \in \mathbb{R}^n_+$ and $v \in \mathbb{R}^m_+$, unique up to scalar multiplication, such that $\operatorname{diag}(u) A \operatorname{diag}(v) \in U(a, b)$. Such a pair (u, v) can be recovered as a fixed point of the Sinkhorn map*

$$(u, v) \mapsto (Av^{-1}./b, A^T u^{-1}./a).$$

The convergence of the algorithm is linear when using Hilbert's projective metric between the scaling factors (Franklin and Lorenz, 1989, §3). Although we use this algorithm in our experiments because of its simplicity, other algorithms exist (Knight and Ruiz, 2012) which are known to be more reliable numerically when λ is large.

Summary: Given a smoothing parameter $\lambda > 0$, using Sinkhorn's algorithm on matrix K, defined as the elementwise exponential of $-\lambda M$ (the pairwise Gaussian kernel matrix between the supports X and Y when $p = 2$, using bandwidth $\sigma = 1/\sqrt{2\lambda}$) we can recover smoothed optima $\alpha^\star_\lambda$ and $T^\star_\lambda$ for *both* smoothed primal $\mathbf{p}_\lambda$ and dual $\mathbf{d}_\lambda$ transport problems. To take advantage of this, we simply propose to substitute the smoothed optima $\alpha^\star_\lambda$ and $T^\star_\lambda$ to the original optima $\alpha^\star$ and $T^\star$ that appear in Algorithms 1 and 2.

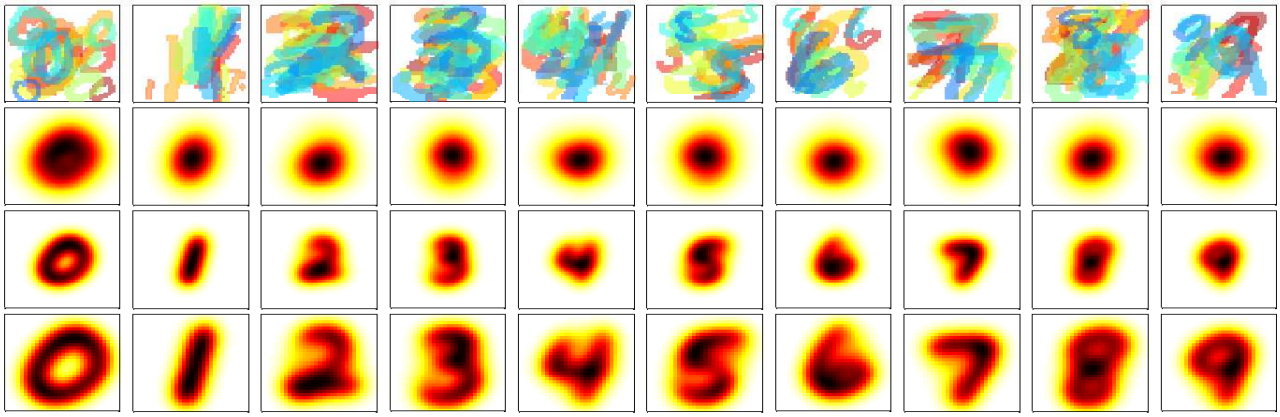

Figure 2. (top) For each digit, 15 out of the ≈ 5.000 scaled and translated images considered for each barycenter. (bottom) Barycenters after $t = 1, 10, 60$ gradient steps. For $t = 60$, images are cropped to show the 30×30 central pixels.

Algorithm 3 Smoothed Primal $T^\star_\lambda$ and Dual $\alpha^\star_\lambda$ Optima

Input M, λ, a, b
$K = \exp(-\lambda M)$;
$\widetilde{K} = \mathrm{diag}(a^{-1})K$ % use `bsxfun(@rdivide,K,a)`
Set $u =$`ones(n,1)/n`;
while u changes **do**
  $u = 1./(\widetilde{K}(b./(K^T u)))$.
end while
$v = b./(K^T u)$.
$\alpha^\star_\lambda = \frac{1}{\lambda}\log(u) - \frac{\log(u)^T \mathbb{1}_n}{\lambda n}\mathbb{1}_n$.
$T^\star_\lambda = \mathrm{diag}(u)K\,\mathrm{diag}(v)$.
% use `bsxfun(@times,` v, `(bsxfun(@times,` K, u`))'`);

6. Applications

We present two applications, one of Algorithm 1 and one of Algorithm 2, that both rely on the smooth approximations presented in §5. The settings we consider involve computing respectively tens of thousands or tens of high-dimensional optimal transport problems—2.500×2.500 for the first application, 57.647×48 for the second—which cannot be realistically carried out using network flow solvers. Using network flow solvers, the resolution of a single transport problem of these dimensions could take between several minutes to several hours. We also take advantage in the first application of the fact that Algorithm 3 can be run efficiently on GPGPUs using vectorized code (Cuturi, 2013, Alg.1).

6.1. Visualization of Perturbed Images

We use 50.000 images of the MNIST database, with approximately 5.000 images for each digit from 0 to 9. Each image (originally 20×20 pixels) is scaled randomly, uniformly between half-size and double-size, and translated randomly within a 50×50 grid, with a bias towards corners. We display intermediate barycenter solutions for each of these 10 datasets of images for $t = 1, 10, 60$ gradient iterations. λ is set to $60/\mathrm{median}(M)$, where M is the squared-Euclidean distance matrix between all 2,500 pixels in the grid. Using a Quadro K5000 GPU with close to 1500 cores, the computation of a single barycenter takes about 2 hours to reach 100 iterations. Because we use warm starts to initialize u in Algorithm 3 at each iteration of Algorithm 1, the first iterations are typically more computationally intensive than those carried out near the end.

6.2. Clustering with Uniform Centroids

In practice, the k-means cost function applied to a given empirical measure could be minimized with a set of centroids X and weight vector a such that the entropy of a is very small. This can occur when most of the original points in the dataset are attributed to a very small subset of the k centroids, and could be undesirable in applications of k-means where a more regular attribution is sought. For instance, in sensor deployment, when each centroid (sensor) is limited in the number of data points (users) it can serve, we would like to ensure that the attributions agree with those limits.

Whereas the original k-means cannot take into account such limits, we can ensure them using Algorithm 2. We illustrate the difference between looking for optimal centroids with "free" assignments ($\Theta = \Sigma_k$), and looking for optimal "uniform" centroids with constrained assignments ($\Theta = \{\mathbb{1}_k/k\}$) using US census data for income and population repartitions across 57.647 spatial locations in the 48 contiguous states. These weighted points can be interpreted as two empirical measures on $\mathbb{R}^2$ with weights directly proportional to these respective quantities. We initialize both "free" and "uniform" clustering with the actual 48 state capitals. Results displayed in Figure 3 show that by forcing our approximation to be uniform, we recover centroids that induce a

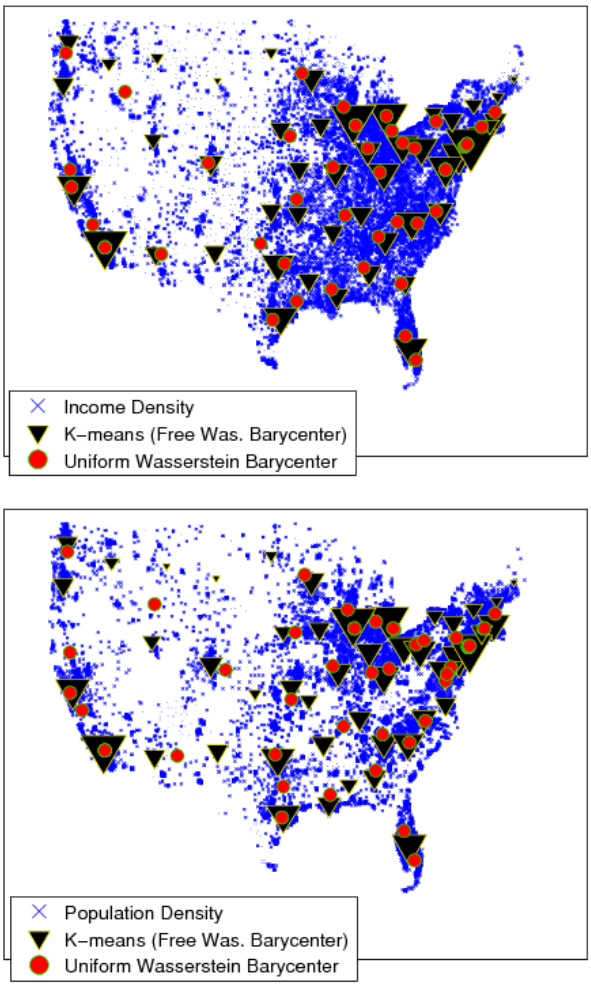

Figure 3. Comparison of two Wasserstein barycenters on spatial repartitions of income and population in the 48 contiguous states, using as many ($k = 48$) centroids. The size of each of the 57.647 blue crosses is proportional to the local average of the relevant variable (income above and population below) at that location, normalized to sum to 1. Each downward triangle is a centroid of the k-means clustering (equivalent to a Wasserstein barycenter with $\Theta = \Sigma_k$) whose size is proportional to the portion of mass captured by that centroid. Red dots indicate centroids obtained with a uniform constraint on the weights, $\Theta = \{\mathbb{1}_k/k\}$. Since such centroids are constrained to carry a fixed portion of the total weight, one can observe that they provide a more balanced clustering than the k-means solution.

more balanced clustering. Indeed, each cell of the Voronoi diagram built with these centroids is now constrained to hold the same aggregate wealth or population. These centroids could form the new state capitals of equally rich or equally populated states. On an algorithmic note, we notice in Figure 4 that Algorithm 2 converges to its (local) optimum at a speed which is directly comparable to that of the k-means in terms of iterations, with a relatively modest computational overhead. Unsurprisingly, the Wasserstein distance between the clusters and the original measure is higher when adding uniform constraints on the weights.

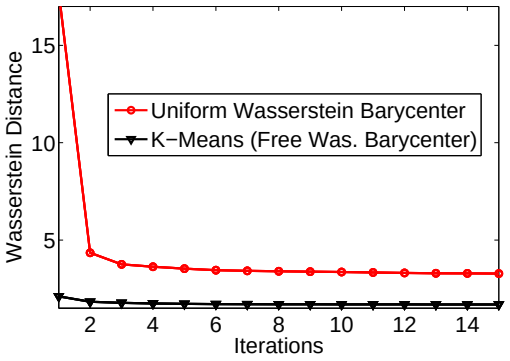

Figure 4. Wasserstein distance of the uniform Wasserstein barycenter (weights constrained to be in $\Theta = \{\mathbb{1}_k/k\}$) and its unconstrained equivalent (k-means) to the income empirical measure. Note that, because of the constraints on weights, the Wasserstein distance of the uniform Wasserstein barycenter is necessarily larger. On a single CPU core, these computations require 12.5 seconds for the constrained case, using Sinkhorn's approximation, and 1.55 seconds for the regular k-means algorithm. Using a regular transportation solver, computing the optimal transport from the 57.647 points to the 48 centroids would require about 1 hour for a single iteration

Conclusion We have proposed in this paper two original algorithms to compute Wasserstein barycenters of empirical measures. Using these algorithms in practice for measures of large support is a daunting task for two reasons: they are inherently slow because they rely on the subgradient method; the computation of these subgradients involves solving optimal and dual optimal transport problems. Both issues can be substantially alleviated by smoothing the primal optimal transport problem with an entropic penalty and considering its dual. Both smoothed problems admit gradients which can be computed efficiently using only matrix vector products. Our aim in proposing such algorithms is to demonstrate that Wasserstein barycenters can be used for visualization, constrained clustering, and hopefully as a core component within more complex data analysis techniques in future applications. We also believe that our smoothing approach can be directly applied to more complex variational problems that involve multiple Wasserstein distances, such as Wasserstein propagation (Solomon et al., 2014).

Acknowledgements We thank reviewers for their comments and Gabriel Peyré for fruitful discussions. MC was supported by grant 26700002 from JSPS. AD was partially supported by EPSRC.

References

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

S.-I. Amari and H. Nagaoka. *Methods of Information Geometry*. AMS vol. 191, 2001.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

G. Canas and L. Rosasco. Learning probability measures with respect to optimal transport metrics. In *Adv. in Neural Infor. Proc. Systems 25*, pages 2501–2509. 2012.

Y. Chen, M. Welling, and A. J. Smola. Supersamples from kernel-herding. In *Uncertainty in Artificial Intelligence (UAI)*, 2010.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schoelkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007.

Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel fisher discriminant analysis. In *Advances in Neural Information Processing Systems 20*, pages 609–616. MIT Press, 2008.

T. Kanamori, T. Suzuki, and M. Sugiyama. f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Trans. on Information Theory*, 58 (2):708–720, 2012a.

T. Kanamori, T. Suzuki, and M. Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012b.

P. A. Knight and D. Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 2012.

S. Lloyd. Least squares quantization in pcm. *IEEE Trans. on Information Theory*, 28(2):129–137, 1982.

R. J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309–324, 1995.

R. J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, May 2005.

M. K. Ng. A note on constrained k-means algorithms. *Pattern Recognition*, 33(3):515–519, 2000.

F. Nielsen. Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms. *IEEE Signal Processing Letters (SPL)*, 2013.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

O. Pele and M. Werman. Fast and robust earth mover's distances. In *ICCV'09*, 2009.

D. Pollard. Quantization and the method of k-means. *IEEE Trans. on Information Theory*, 28(2):199–205, 1982.

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, volume 6667 of *Lecture Notes in Computer Science*, pages 435–446. Springer, 2012.

S. Reich. A non-parametric ensemble transform method for bayesian inference. *SIAM Journal of Scientific Computing*, 35 (4):A2013–A2024, 2013.

Y. Rubner, L. Guibas, and C. Tomasi. The earth movers distance, multi-dimensional scaling, and color-based image retrieval. In *Proceedings of the ARPA Image Understanding Workshop*, pages 661–668, 1997.

R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4): 402–405, 1967.

J. Solomon, R. Rustamov, G. Leonidas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *Proceedings of The 31st International Conference on Machine Learning*, pages 306–314, 2014.

M. Sugiyama, M. Yamada, M. Kimura, and H. Hachiya. On information-maximization clustering: Tuning parameter selection and analytic solution. In *Proceedings of the 28th International Conference on Machine Learning*, pages 65–72, 2011.

M. Teboulle. A unified continuous optimization framework for center-based clustering methods. *The Journal of Machine Learning Research*, 8:65–102, 2007.

R. Veldhuis. The centroid of the symmetrical kullback-leibler distance. *Signal Processing Letters, IEEE*, 9(3):96–99, 2002.

C. Villani. *Optimal transport: old and new*, volume 338. Springer Verlag, 2009.

G. Zen and E. Ricci. Earth mover's prototypes: A convex learning approach for discovering activity patterns in dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3225–3232. IEEE, 2011.