FAST COMPUTATION OF WASSERSTEIN BARYCENTERS

MARCO CUTURI AND ARNAUD DOUCET

Abstract. Wasserstein barycenters (Agueh and Carlier, 2011) define a new family of barycenters between N probability measures that builds upon optimal transport theory. We argue using a simple example that Wasserstein barycenters have interesting properties that differentiate them from other barycenters proposed recently, which all build either or both on kernel smoothing and Bregman divergences. We propose two algorithms to compute Wasserstein barycenters for finitely supported measures, one of which can be shown to be a generalization of Lloyd's algorithm. A naive implementation of these algorithms is intractable, because it would involve numerous resolutions of optimal transport problems, which are notoriously expensive to compute. We propose to follow recent work by Cuturi (2013) and smooth these transportation problems to recover faster optimization procedures. We apply these algorithms to the visualization of perturbed images and resampling in particle filters.

1. Introduction

Comparing, summarizing and reducing the dimensionality of empirical probability measures on a given probability 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.

These divergences are never directly applied to empirical measures and point clouds, because they are usually ill-defined for measures that do not have continuous densities. They also fail to incorporate any form of prior knowledge on the geometry of Ω, which might be available if, for instance, Ω is a metric or a Hilbert space in addition to being a probability space. Both of these issues are usually solved using Parzen's approach (1962) to smooth empirical measures with smoothing kernels. In addition to producing well-behaved densities from empirical measures, a smoothing kernel has two virtues: it encodes prior knowledge on the probability space Ω, and greatly facilitates computations if positive definite. For instance, 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.

A divergence defines implicitly the *mean* or the *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 and Nock, 2009; Nielsen, 2013). The goal of this paper is to compute efficiently barycenters defined by the *optimal transport distance* of measures (Villani, 2009, §6), a.k.a the *Wasserstein* distance. Although the juxtaposition of *efficient* and *optimal transport* may seem contradictory, since the latter is usually expensive

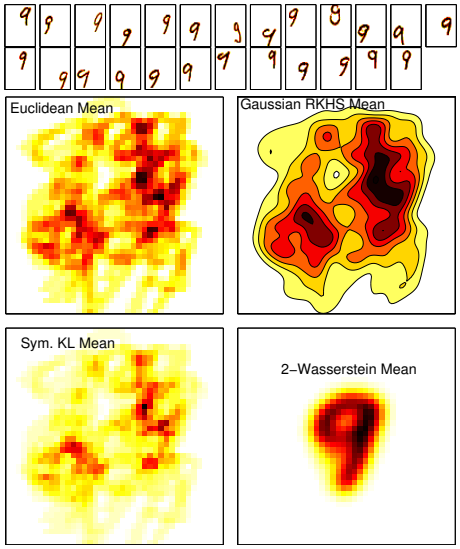

FIGURE 1. (Top) 25 images of "9" digits from the MNIST database scattered randomly in $[0,1]^2$. (Below, left-to-right, top-to-bottom) mean measures of these digits, using the Euclidean distance between histograms; the RKHS distance between smoothed densities with a spherical Gaussian kernel ($\sigma = 0.001$); the Symmetrized Kullback-Leibler divergence (Nielsen, 2013); the 2-Wasserstein distance.

to compute, we show in this work that we can obtain efficient algorithms using a computational approach introduced 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 25 handwritten "9" digits translated randomly on a 40×40 grid. Each image is a discrete measure on $[0,1]^2$ (stored as a histogram of size 1600 in memory) with normalized intensities. Computing the Euclidean, Gaussian RKHS mean-maps or Symmetrized Kullback-Leibler means of these images results in mean measures that hardly make any sense. The 2-Wasserstein mean (defined in §3.1) produced by Algorithm 1 and displayed in the lower-right corner captures perfectly the structure of these images. Note that the Wasserstein mean does not consider any prior on these images: not on their smoothness, nor their shape, nor their sparsity. The only ingredient that is used is that of defining a distance – the Euclidean distance – in $[0,1]^2$. Such a distance is also implicitly used in the definition of the Gaussian RKHS used to produce the blob in the upper-right corner.

This paper is structured as follows: we provide required background on optimal transport in §2, followed by the definition of Wasserstein barycenters with additional motivating examples in §3. Novel contributions are presented from §4: we start with two projected subgradient descent algorithms that can be used to compute Wasserstein barycenters. These algorithms are extremely costly, even

for measures of small support or histograms of small size. We show in §5 that smoothing a key element of these algorithms (the repeated computation of optimal transports) results in a projected gradient scheme and execution times that are orders of magnitudes faster. We conclude with two applications of our algorithms in §6.

2. Background on Optimal Transport

Let (Ω, D) be a metric Polish space. All measures considered in this paper are Borel measures, on the Borel σ-algebra induced by D. Let $P(\Omega)$ be the set of Borel probability measures on Ω, that is the set of non-negative measures with total unit mass. $\mathbb{R}_+$ is the half-line of nonnegative reals.

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

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

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

2.2. **Measures with Discrete and Finite Support.**

2.2.1. *Discrete Probability Simplex.* In what follows, we will mostly consider measures with discrete, finite support. In that context, the probability simplex Σ_n of n bins will play an important role throughout the paper:

$$\Sigma_n \stackrel{\text{def}}{=} \{u \in \mathbb{R}^n_+ \mid \sum_{i=1}^n u_i = 1\}.$$

2.2.2. *Measures supported on a finite set X.* For any point $x \in \Omega$, δ_x denotes the Dirac unit mass on point x. For any subset $X = \{x_1, \ldots, x_n\}$ of $n > 1$ points of Ω, we write $P(X)$ for the set of probability measures of Ω supported by X:

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

2.2.3. *Measures supported on up to k points.* Following the definition of $P(X)$, we consider the set $P_k(\Omega)$ of measures of Ω that have discrete support of size up to k, namely measures supported on any set $X \in \Omega^k$,

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

2.3. **Wasserstein and Optimal Transport.** Consider two sets $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ of points in Ω. When $\mu \in P(X)$ and $\nu \in P(Y)$, the Wasserstein distance $W_p(\mu, \nu)$ between μ and ν is the p^{th} root of the optimum of a Linear Program (LP) – a network flow problem to be precise – 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,

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

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

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

Combining these two definitions, the distance $W_p(\mu,\nu)$ raised to the power p, henceforth abbreviated as $W_p^p(\mu,\nu)$, is the optimum of a LP of $n \times m$ variables,

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

3. Wasserstein Barycenters

3.1. Definition and Previous Work.

Definition 1 (Agueh and Carlier (2011))**.** *A Wasserstein barycenter of N measures $\{\nu_1,\dots,\nu_N\}$ in any set $\mathbb{P} \subset P(\Omega)$ is any minimizer of f over $\mathbb{P}$, where*

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

The variational formulation[1] of barycenters used by Agueh and Carlier is similar to that used by Veldhuis (2002) or Banerjee et al. (2005). This idea can be traced back to the concept of Fréchet means (1948), as pointed out by Bigot and Klein (2012).

In their original paper, Agueh and Carlier (2011) consider conditions on the ν_i's for such a barycenter in $\mathbb{P} = P(\Omega)$ to be unique, describe its characteristics, relate it to the multi-marginal transportation problem and describe known solutions (Agueh and Carlier, 2011, §6) in the cases where (i) $\Omega = \mathbb{R}$; (ii) $N = 2$ using McCann's interpolant (1997); (iii) when all the measures ν_i are Gaussians in $\Omega = \mathbb{R}^d$ centered on 0.

Rabin et al. (2012) were – to our knowledge – the first to consider practical algorithmic approaches to compute Wasserstein barycenters between point clouds. Their method relies on *sliced* Wasserstein distances, a simplified proxy for the Wasserstein distance that uses 1-dimensional random projections of point clouds. The sliced approximation is very attractive for computational reasons: because W_p has a closed form for point clouds on the real line, which can be computed simply by sorting them, Rabin et al. can avoid solving general optimal transport problems. Negative results by Naor and Schechtman (2007) suggest however that the sliced Wasserstein distance is bound to have a large distortion factor with respect to the original Wasserstein distance.

We propose in this work computational answers to the problem raised by Agueh and Carlier: we describe algorithms to compute Wasserstein barycenters when (i) each of the N measures ν_i has a discrete and finite support (ii) the search for a barycenter is not considered on $P(\Omega)$ but restricted to either $P(X)$ (the set of measures supported on a predefined finite set X, §2.2) or $P_k(\Omega)$ (the set of measures supported on up to k atoms, §2.2). This setting is similar to that explored

[1] Agueh and Carlier consider more generally non-uniform weights on the distances from μ to the N target measures. The algorithms we propose extend trivially to that case. We use uniform weights to keep notations simpler.

by Rabin et al. who considered point clouds, except that we consider continuous weights. This setting is also known to encompass a few relevant problems:

3.2. **Related Problems and Relevance.**

3.2.1. $N = 1, \mathbb{P} = P(X)$, *X fixed.* When only one measure ν, supported on $Y \in \Omega^m$, $m > 1$ is considered, its closest element μ in $P(X)$ in the Wasserstein metric can be computing by simply defining a weight vector a on the atoms of X that results from assigning all of the mass b_i to the closest neighbor of y_i in X.

3.2.2. $N = 1, \mathbb{P} = P_k(\Omega)$ *and k-means.* When $N = 1$ and only one target measure ν of support Y of size m is considered, minimizing $f(\mu)$ over $P(\Omega)$ is trivially solved by setting $\mu = \nu$. When the feasible set is restricted to $P_k(\Omega)$, $k < m$, $\Omega = \mathbb{R}^d$ and $p = 2$, minimizing f over $P_k(\Omega)$ is known to be equivalent to the k-means problem (Pollard, 1982; Canas and Rosasco, 2012).

3.2.3. $N > 1, \mathbb{P} = P(X)$, *Clustering of Histograms.* When Ω can be reduced to the union of all supports of the ν_i with X, of total 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 most applications, histograms of features (bags-of-words, image features) are high-dimensional and sparse, as illustrated in Figure 1. Wasserstein barycenters, or EMD barycenters as they might be called in that case, can be used to produce mean elements that account for feature overlap and similarity. Suppose, for instance, that Ω is the set of all words in all languages. Suppose one is given a set of documents supported on a language $Y \subset \Omega$. Which bag-of-words, supported in a different language X, could summarize most efficiently such a set? Such problems appear naturally in cross-lingual document retrieval (Kraaij et al., 2003) and cross-lingual document categorization (Nastase and Strapparava, 2013).

Wasserstein barycenters could also be used as intermediate centering steps in Lloyd type clustering algorithms (1982) when comparing databases of histograms.

3.2.4. *Best approximation of target measures by uniform measures.* Consider a single measure with support $Y \in \Omega^m$ and weights $b \in \Sigma_m$. One might be interested in the best approximation in Wasserstein sense of this weighted empirical measure by an empirical measure in $P(Y)$ restricted to have weights which are multiples of $1/m$. Such approximations are at the core of the key resampling step of particle filtering methods Doucet et al. (2001). An alternative consists of relaxing the fact that the approximation has to be in $P(Y)$ and considering the case where it lies in $P_k(\Omega)$ but is restricted to have identical weights $1/k$ (Reich, 2013).

4. Computing Wasserstein Barycenters

4.0.5. *Notations.* For each $i \leq N$, let the finite family of points $Y_i \in \Omega^{m_i}$ of size $m_i > 1$ describe the locations of the support of ν_i. Let b_i denote the probability weights of ν_i, which is by definition a vector in the simplex Σ_{m_i}.

To minimize f, we first review in §4.1 some important results on the convexity and differentiability of the linear program $S(a, b; M)$ defined in Equation (3). Since f is a sum of evaluations of S, as exhibited by Equations (3)&(4), the convexity and the differentiability of f follows from that of S.

4.1. **Convexity and Differentiability of $S(a,b;M)$.** A subscripted star $\cdot_\times$ appended to a vector (resp. a matrix) means that its last element (resp. line) has been removed.

4.1.1. *Redundancy in row & column-sum constraints.* The transportation polytope $U(a,b)$ introduced in Equation (2) is defined by n row-sum constraints and m column-sum constraints. However, one of these $n+m$ constraints is redundant because both a and b have equal sum. To ensure an independent set of constraints, we replace equality $T^T\mathbf{1}_{m-1} = b$ by $(T^T)_\times \mathbf{1}_{m-1} = b_\times$ in the definition of $U(a,b)$ and in the dual below.

4.1.2. *Dual transportation problem.* Given matrix $M \in \mathbb{R}^{n\times m}$, the optimum $S(a,b;M)$ admits the following dual LP form (Bertsimas and Tsitsiklis, 1997, §7.6,§7.8)

$$S(a,b;M) = \max_{(\alpha,\beta)\in C(M)} \alpha^T a + \beta^T b_\times \,, \tag{5}$$

where the polyhedron $C(M) \subset \mathbb{R}^{n+m-1}$ of dual variables is defined as

$$C(M) = \{(\alpha,\beta), \alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^{m-1} \,|\, \forall i \leq n,\, j \leq m-1,\; \alpha_i + \beta_j \leq m_{ij}\}.$$

4.1.3. *Relationship between primal and dual optima.* Any optimal solution $T^\star$ to the primal problem in Equation (3) obtained with the network simplex yields a dual optimal solution. For all the $n+m-1$ pairs (i,j) which are in the basis of $T^\star_{ij}$ (which correspond to the elements $T^\star_{ij} > 0$ if the solution is not degenerated) $\alpha_i + \beta_j = m_{ij}$ holds. Consequently $\alpha^\star$ and $\beta^\star$ can be recovered by inverting a trivial linear system, which is underdetermined when the primal solution is not unique. The optimal dual variable $\alpha^\star$ plays a key role:

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

Proof. These results follow from sensitivity analysis in LP's (Bertsimas and Tsitsiklis, 1997, §5.2). The dual expression of d_M in Equation (5) shows that $S(a,b;M)$ is bounded and can be computed as a maximum of linear functions indexed by the finite 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 S at a, and a subgradient otherwise. □

4.2. **Fixed Support: Minimizing f over $P(X)$.** Let $X \subset \Omega^n$. We propose in this section an algorithm to compute the weights $a \in \Sigma_n$ of a measure μ supported by X such that $f(\mu)$ is optimal in $P(X)$. When $N = 1$, the problem is trivially solved by setting $a_i = \sum_{j=1}^m b_j \Delta_{ij}$ where $\Delta_{ij} = 1$ iff y_j is in the i^{th} cell of the Voronoi partition seeded by X, namely the nearest neighbor of y_j in X is x_i (ties randomly attributed).

We assume $N > 1$. We overload function f in this section by defining it on any element $a \in \Sigma_n$ as

$$f(a) = \frac{1}{N}\sum_{i=1}^N W_p^p\left(\sum_{k=1}^n a_k \delta_{x_k}, \nu_i\right). \tag{6}$$

For $i \leq N$, following the notations of Equation (1), let

$$M_i \in \mathbb{R}^{n \times m_i},\ M_i \stackrel{\text{def}}{=} M_{XY_i}$$

Let $\alpha_i^\star$ be any optimal dual variable corresponding to the computation of $S(a, b_i; M_i)$ as detailed in Equation (5). f being a sum of terms $S(a, b_i, M_i)$, we get the following corollary of Proposition 1.

Corollary 1. *f is a polyhedral convex function on $P(X)$ and*

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

is a subgradient of f at a.

This result suggests a projected subgradient descent algorithm, outlined in Algorithm 1, to find the barycenter of measures $\{\nu_i\}$ when restricting such a search to $P(X)$. We write P_{Σ_d} for a projector onto the probability simplex in a suitable norm. Note that this projector may include non-trivial constraints, such as sparsity constraints, either in explicit (Becker et al., 2013) or regularized (Pilanci et al., 2012) form.

Algorithm 1 p-Wasserstein Barycenter in $P(X)$

Inputs: $X \in \Omega^n$, $Y_i \in \Omega^{m_i}$, $b_i \in \Sigma_{m_i}$, $i \leq N$,
$p \in [1, \infty)$, $t_0 > 0$.
if $N = 1$ **then**
  a obtained using allocation of each b_i to nearest neighbor of Y in X. **stop**
end if
Initialize $a = a_0$, $t = 1$
Form all $n \times m_i$ matrices $M_i = M_{XY_i}$, see Eq. (1).
while not converged **do**
  for $i \in \{1, \ldots, N\}$ **do**
    Compute $\alpha_i^\star$, the dual optimal variable of $S(a, b_i, M_i)$, see Eq. (5)
  end for
  Subgradient: $\boldsymbol{\alpha} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \alpha_i^\star$
  $a \leftarrow P_{\Sigma_d}\left(a - \frac{t_0 \boldsymbol{\alpha}}{\sqrt{t}}\right)$; $t \leftarrow t + 1$
end while

4.3. **Free Support: Minimizing f over $P_k(\mathbb{R}^d)$.** We now consider the more general case of minimizing f over any probability measure μ supported by *at most k atoms*. We require now that $\Omega = \mathbb{R}^d$ for $d \geq 1$, D is the Euclidean distance and $p = 2$. We overload again function f to consider for $X \in (\mathbb{R}^d)^k, a \in \Sigma_k$,

$$f(X, a) = \frac{1}{N} \sum_{i=1}^{N} W_p^p \left(\sum_{k=1}^{n} a_k \delta_{x_k}, \nu_i \right). \tag{7}$$

4.3.1. *Ground Metrics M_{XY_i} in the Euclidean Case.* When $\Omega = \mathbb{R}^d$, the sets X and Y_i, $i \leq N$ can be respectively represented by a matrix in $\mathbb{R}^{d \times n}$ and N matrices

in $\mathbb{R}^{d\times m_i}$. The pairwise squared-Euclidean distances of points in these sets can be recovered by writing $\mathbf{x} \stackrel{\text{def}}{=} \operatorname{diag}(X^T X)$ and $\mathbf{y}_i \stackrel{\text{def}}{=} \operatorname{diag}(Y_i^T Y_i)$, and observing that

$$M_{XY_i} = \mathbf{x}\mathbf{1}_m^T + \mathbf{1}_n \mathbf{y}_i^T - 2X^T Y_i \in \mathbb{R}^{n\times m_i}.$$

Due to the margin constraints that apply if a matrix T is in the polytope $U(a, b_i)$, we have:

$$\begin{aligned}\langle T, M_{XY_i}\rangle &= \langle T, \mathbf{x}\mathbf{1}_d^T + \mathbf{1}_d^T \mathbf{y}_i - 2X^T Y_i\rangle \\ &= \operatorname{tr} T^T \mathbf{x}\mathbf{1}_d^T + \operatorname{tr} T^T \mathbf{1}_d^T \mathbf{y}_i - 2\langle T, X^T Y_i\rangle \\ &= \mathbf{x}^T a + \mathbf{y}^T b_i - 2\langle T, X^T Y_i\rangle.\end{aligned}$$

4.3.2. *Properties of $f(X, a)$.* Discarding constant terms in $\mathbf{y}_i$ and b_i, we have that computing $\min_{\mu\in P_k(\Omega)} f(\mu)$ is equivalent to solving

$$\min_{\substack{X\in\mathbb{R}^{d\times k}\\ a\in\Sigma_k}} \mathbf{x}^T a + \frac{2}{N}\sum_{i=1}^N S(a, b_i, -X^T Y_i) \tag{8}$$

The objective function on the righthand side of Equation (8) is convex in a. As a function of X, that objective is the sum of a convex quadratic function of X with a piecewise linear concave function, since, as noted in (Cuturi and Avis, 2011), each term

$$S(a, b_i, -X^T Y_i) = \min_{T\in U(a,b_i)} \langle X, -Y_i T^T\rangle$$

can be seen as the minimum of linear functions indexed by the vertices of the polytope $U(a, b_i)$. As a consequence $f(X, a)$ is only convex with respect to a but not with respect to X.

4.3.3. *Local Quadratic Approximation of $f(X, a)$.* Suppose that, for all $i \le N$, $T_i^\star$ is optimal for problem $S(a, b_i, -X^T Y_i)$. Updating Equation (8), we get

$$\begin{aligned}&\langle XX^T, \operatorname{diag}(a)\rangle - \frac{2}{N}\sum_{i=1}^N \langle T_i^\star, X^T Y_i\rangle = \\ &\quad \|X\operatorname{diag}(a^{1/2}) - \frac{1}{N}\sum_{i=1}^N Y_i T_i^{\star T}\operatorname{diag}(a^{-1/2})\|^2 - \|\frac{1}{N}\sum_{i=1}^N Y_i T_i^{\star T}\operatorname{diag}(a^{-1/2})\|^2.\end{aligned}$$

Minimizing a local quadratic approximation of $f(\cdot, a)$ around X yields $X^\star = \left(\frac{1}{N}\sum_{i=1}^N Y_i T_i^{\star T}\right)\operatorname{diag}(a^{-1})$. A natural step to update X is thus to consider a Newton step, which we assume to be of fixed length $\theta \in [0, 1]$ in what follows. A simple interpretation of this update is as follows: each matrix $T_i^{\star T}\operatorname{diag}(a^{-1})$ has n column-vectors in the simplex Σ_{m_i}. The suggested update for X is the mean of all $Y_i T_i^{\star T}\operatorname{diag}(a^{-1})$. Each of these terms is itself a mixture of n vectors in Y_i, namely n barycenters of points enumerated in Y_i with weights defined by the optimal transport $T_i^\star$.

4.3.4. *Alternating Optimization.* As hinted in Section 3.2, since our problem supersedes the k-means problem, minimizing $f(X, a)$ cannot to be a convex problem in the general case. To obtain an approximate solution of $f(X, a)$ we propose in Algorithm 2 to update alternatively X (with the local quadratic approximation above) and a (with Algorithm 1).

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

Input: $Y_i \in \Omega^{m_i}$ for $i \leq N$. $\theta \in]0,1]$.
initialize X and a
while X and a have not converged **do**
  $a \leftarrow a^\star$ found with Algorithm 1.
  for $i \in \{1, \dots, N\}$ **do**
    $T_i^\star \leftarrow$ optimal solution of $S(a, b_i; -X^T Y)$
  end for
  $X \leftarrow (1-\theta)X + \theta \left(\frac{1}{N} \sum_{i=1}^N Y_i T_i^{\star T} \right) \mathrm{diag}(a^{-1})$
end while

4.3.5. *Particular Cases of Algorithm 2.* As argued in §3.2, minimizing f over $P_k(\mathbb{R}^d)$ in the case where only one $N = 1$ measure is considered is equivalent to the k-means problem. One can extend this analogy to Algorithm 2 and show that it is strictly equivalent to Lloyd's algorithm when $N = 1$: computing $a^\star$ reduces to the computation of the Voronoi diagram of X and the subsequent allocations of the weights of ν lying in these cells, while the integration of the Voronoi cells is equivalent to the computation of barycenters of Y using an optimal $T^\star$. As a result of this analogy, our algorithm also suffers from a dependance on the initial values of X (Arthur and Vassilvitskii, 2007), which needs to be investigated in the case where $N > 1$.

Reich has recently proposed an ensemble transform method grounded on optimal transport (2013), whose ensemble location updates (2013, Equation 3.8, where the notation x^f corresponds to Y in this paper, and x^a to updated X) can be interpreted as a single iteration of the loop of Algorithm 2, when $N = 1$; the weights a are fixed to be equal to $\mathbf{1}_k/k$; $\theta = 1$. Reich's derivation of this approach follows a different route, and is grounded on an asymptotic result of McCann (1995). We consider Reich's approach in §6.2, and compare it with our approach which simply suggests instead to continue (and not stop after one iteration) applying optimal transport maps to Y to recover better locations.

4.3.6. *In Summary.* We have proposed two algorithms, one convex (Algorithms 1) and another that is not (Algorithm 2) to compute Wasserstein barycenters of measures. 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 builds upon N transportation problems at each subgradient step. Namely, N network flow problems each of size $n \times m_i$ with $n + m_i - 1$ constraints. These become quickly expensive when n and m_i are above a few hundreds. Algorithm 2 incurs an even higher cost, since it involves running Algorithm 1 at each iteration. We propose alternative computational approaches in §5.

5. Fast & Smooth Optimization

To circumvent the major computational roadblock posed by the repeated computation of optimal transports, we propose to use Cuturi's approach (2013) and smooth all of the transportation problems encountered in Algorithms 1 and 2.

5.1. **Smoothed Transportation Problems S_λ.** Consider the entropy $h(T)$ of a $n \times m$ transport variable T, which is by definition in $\Sigma_{n\times m}$:

$$h(T) \stackrel{\text{def}}{=} -\sum_{i,j=1}^{n,m} t_{ij} \log(t_{ij}).$$

Cuturi (2013) has recently proposed to consider, for $\lambda > 0$, a regularized transportation problem S_λ as

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

5.1.1. *Dual Optimal Variables of S_λ.* Cuturi uses an entropic regularization to define a novel family of distances – Sinkhorn distances – which, as argued by the author, have favorable properties over classic transportation distances. In that sense, Cuturi focuses exclusively on the value of the optimum of S_λ, and its interest as a novel distance by itself. Our focus in this work is different, since we are not interested in the objective, but instead on the optimal transport as well as the optimal dual variables of S_λ. As shown below, the optimal dual variables can be used to recover a gradient for weights a in Algorithm 1, while an approximate optimal transport can be used to update locations X in Algorithm 2.

5.1.2. *Smoothing of non-smooth objective.* Smoothing S, the non-smooth support function of the polytope $U(a, b)$, by a strongly convex term, minus the entropy, can also be interpreted as a form of Nesterov smoothing (2005). Compared to other regularizers however, such as the quadratic norm $\langle T, T \rangle$ of T, far more is gained computationally by using its entropy $h(T)$: the unique solution of S_λ has a factorized form (Wilson, 1969), a property derived from the maximum entropy principle (Darroch and Ratcliff, 1972) and described in detail by Erlander and Stewart (1990, §3,4):

Lemma 1 (Wilson (1969))**.** *Given $M \in \mathbb{R}^{n\times m}, a \in \Sigma_n, b \in \Sigma_m$, the unique optimal solution $T^\sharp \in U(a, b)$ to $S_\lambda(a, b, M)$ admits the factorization*

$$T^\sharp = \operatorname{diag}(u) e^{-\lambda M} \operatorname{diag}(v),$$

where $u \in \mathbb{R}^n_+, v \in \mathbb{R}^m_+$.

This fact can be obtained by the method of Lagrangian multipliers, as shown in the proof of Proposition 2 below. These factors u and v can be recovered using Sinkhorn's algorithm:

Lemma 2 (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 and v, 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*

$$g_{Aab}(u, v) \in \mathbb{R}^n_+ \times \mathbb{R}^m_+ \mapsto (Av^{-1}./b, A^T u^{-1}./a).$$

Sinkhorn's algorithm consists in applying the Sinkhorn map g_{Aab} iteratively to any pair of arbitrary initial vectors until convergence. The convergence of the algorithm is linear when using Hilbert's projective metric between these scaling factors (Franklin and Lorenz, 1989, §3). Although we highlight this algorithm because of its simplicity, as can be seen in the outline of Algorithm 3, other algorithms exist (Knight and Ruiz, 2012) which are known to be more reliable numerically – yet not necessarily faster – when the regularization term λ is large (Knight, 2008).

5.2. **Convexity and Differentiability of S_λ.** This section echoes the earlier claims made for S in §4.1

Proposition 2. *For fixed b and M, $S_\lambda(a, b; M)$ is a strongly convex function of a with parameter $1/\lambda$. Let (u, v) be* any *vectors of $\mathbb{R}_+^n \times \mathbb{R}_+^m$ such that $\operatorname{diag}(u)e^{-\lambda M}\operatorname{diag}(v) \in U(a, b)$. The gradient of $S_\lambda(a, b; M)$ with respect to a is equal to*

$$\nabla S_\lambda = \frac{1}{\lambda}\left(\log(v_m u) + \mathbf{1}_n\right).$$

Proof. Minus the entropy is strongly convex with parameter 1 on $U(a, b)$. As a result the map $T \to \langle T, M\rangle - \frac{1}{\lambda}h(T)$ is strongly convex with parameter $1/\lambda$. The strong convexity of $S_\lambda(a, b; M)$ with parameter $1/\lambda$ with respect to a follows from (Fiacco and Kyparisis, 1986, Prop.2.11). Slater's weak conditions (Boyd and Vandenberghe, 2004, §5.2.3) hold since T is only constrained by equality and affine (non-negativity) constraints. As a result strong duality applies. Let $\mathcal{L}$ be the Lagrangian of S_λ with respect to (T, α, β), where $\alpha \in \mathbb{R}^n$ is the vector of dual variables for row-sums while $\beta \in \mathbb{R}^{m-1}$ is the vector of $m-1$ dual variables for column-sums.

$$\mathcal{L} = \sum_{ij} t_{ij}m_{ij} + \frac{1}{\lambda}t_{ij}\log t_{ij} + \alpha^T(T\mathbf{1}_n - a) + \beta^T((T^T)_\times \mathbf{1}_{m-1} - b_\times).$$

First order conditions imply that, for optimal primal $T^\star$ and dual variables $\alpha^\star, \beta^\star$,

$$\forall i \le n, j \le m,\ t^\star_{ij} = e^{-\frac{1}{2}-\lambda\alpha^\star_i}e^{-\lambda m_{ij}}e^{-\frac{1}{2}-\lambda\beta^\star_j},$$

where we have set $\beta^\star_m \stackrel{\text{def}}{=} 0$. Let (u, v) be *any* fixed point of the Sinkhorn map $g_{e^{-\lambda M}ab}$. The dual variables $(\alpha^\star, \beta^\star)$ can be recovered from any solution (u, v) by noting that (u, v) must be rescaled to $(\rho u, v/\rho)$ so that $v_m/\rho = e^{-1/2}$. Consequently we obtain that

$$\alpha^\star = -\frac{1}{\lambda}\left(\log(v_m u) + \mathbf{1}_n\right).$$

Using results from local sensitivity analysis (Boyd and Vandenberghe, 2004, §5.6.3) we recover that the gradient of S_λ with respect to a is $-\alpha^\star$. □

The computation of the gradient of S_λ is summarized in Algorithm 3, with a one-line iteration of Sinkhorn's fixed point map introduced in (Cuturi, 2013), to which we also refer to for a quantification of the speedups that result from computing S_λ instead of solving S exactly.

5.3. **Implementation of the Gradient of S_λ.** The computation of the gradient of $f(X, a)$ (defined in Equation (7)) with respect to a can be naively carried out by computing the N gradients for $S_\lambda(a, b_i, M_{XY_i})$. This computation can also be parallelized if one considers a single matrix M_{XY} where $Y = \cup_i Y_i$ and the optimization is carried out with sparse histograms. When the size of $\cap_i Y_i$ is comparable to that of $\cup_i Y_i$ this can result in large speedups.

5.3.1. *In summary:* we have shown that the optimal solution to S can be approximated by the solution to a smoothed problem S_λ. These solutions can be used as such, to replace optimal transports by smooth optimal transports as in Algorithm 2, or to recover smoothed optimal dual variables to be used in Algorithm 1. Algorithm 3 can be warm-started with the scaling factor z computed at a previous iteration.

Algorithm 3 Computation of S_λ's optimal solution $T^\sharp$ and gradient ∇ w.r.t a

Input M, λ, a, b
$K = \exp(-\lambda M)$;
$\widetilde{K} = \mathrm{diag}(a^{-1})K$ % use `bsxfun(@rdivide,`K`,`a`)`
Set $z =$`ones(`n`,1)/`n;
while z changes **do**
  $z = \widetilde{K}$`*(`b`.*(`$K^T z^{-1}$`)^-1)`
end while
$\rho = K$`(:,m)`$^T z^{-1}/a_n$;
$\nabla = \frac{1}{\lambda}\left(\mathbf{1}_N - \log(\rho z)\right)$;
$T^\sharp = \mathrm{diag}(l)K\,\mathrm{diag}(m)$
% use `bsxfun(@times,`m`,(bsxfun(@times,`K, l`))'`);

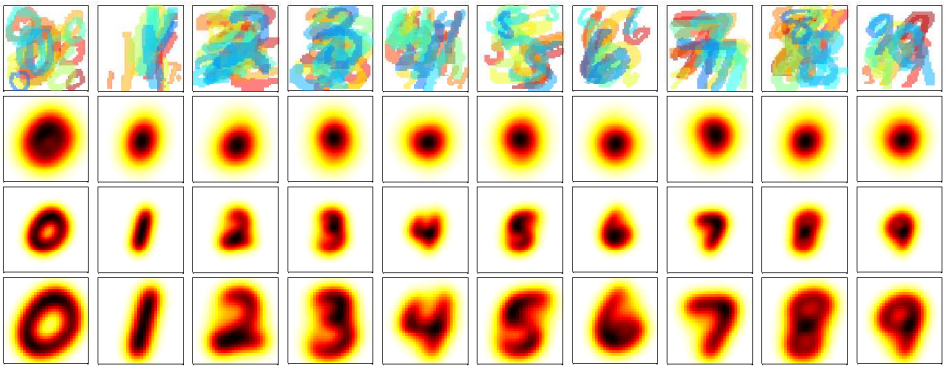

FIGURE 2. (top-row) For each digit, 15 out of the ≈ 5.000 scaled and translated images considered for each barycenter. (bottom rows) Barycenters after $t = 1, 10, 60$ gradient steps. For $t = 60$, images are cropped to show only the 30×30 central pixels.

6. Applications

We review two applications of Algorithms 1 and 2 which can only only work in these settings due to the speed-ups we gain from Algorithm 3.

6.1. **Mean of Histograms.** Wasserstein means for 10 digits are reported in Figure 2. We use the 50.000 first images of the MNIST database, which provide approximately 5.000 images for each digit. Each image (20×20 pixels) is scaled randomly, uniformly between half-size and double-size, and translated randomly within the $50 \times 50 = 2.500$ grid, with a bias towards corners. We display intermediate barycenter solutions, for $t = 1, 10, 60$ gradient iterations. The algorithm is initialized with the uniform measure in Σ_{2500} and we use a naive gradient descent setup with exponentiated gradient updates (Beck and Teboulle, 2003) and a step size $t_0 = 10$. λ is set to $60/\mathrm{median}(M)$. Using a Quadro K5000 GPU with close to 1500 cores, the computation of a single barycenter takes about 2 hours to reach 100 iterations. Since we always use warm starts to run Algorithm (3), the first iterations are typically more computationally intensive than those carried out near the end.

	$N=30$	$N=50$	$N=70$	$N=100$	$N=150$	$N=250$	$N=500$
Multinomial	64.4 ± 144	31 ± 60.5	21.2 ± 26.9	15.6 ± 16.9	11.1 ± 12.6	7.33 ± 8.57	3.76 ± 4.27
Reich	40.1 ± 56.6	29.9 ± 21.6	26.3 ± 16.2	24.2 ± 13.4	22.2 ± 10.2	20.8 ± 7.53	20 ± 5.27
Systematic	35.2 ± 66.4	22.3 ± 31.5	17 ± 19.4	12.3 ± 14.4	8.42 ± 9.31	5.48 ± 6.63	2.78 ± 3.26
Algorithm 2	31.7 ± 53.8	20.1 ± 21.1	16 ± 16.9	11.9 ± 13.1	8.42 ± 9.27	5.4 ± 6.14	2.88 ± 3.34

TABLE 1. Mean square errors and standard deviations for a varying number of particles over 10.000 simulations

6.2. **Filtering.** Consider the non-linear dynamic model

$$X_{1,t} = \frac{X_{1,t-1}}{2} + \frac{25X_{1,t-1}}{1+X_{1,t-1}^2} + 8\cos(1.2t) + V_{1,t},$$
$$X_{2,t} = 0.99X_{2,t-1} + V_{2,t},\ Y_t = X_{1,t}^2/20 + X_{2,t} + W_t$$

where $X_{1,1} \sim \mathcal{N}(0,1)$, $X_{2,1} \sim \mathcal{N}(0,1)$, $V_{1,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $V_{2,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,2)$. We use particle filtering so as to estimate the analytically intractable multimodal posterior distributions $p(x_t|\, y^t)$ where $x_t := (x_{1,t}, x_{2,t})^T$, $y^t := (y_1, ..., y_t)$. At the core of these methods is the resampling step: given an approximation $\widehat{p}(x_t|\, y^t) = \sum_{i=1}^n w_t^i \delta_{x_t^i}$ of $p(x_t|\, y^t)$, resampling provides a new approximation $\widetilde{p}(x_t|\, y^t) = n^{-1}\sum_{i=1}^n n_t^i \delta_{x_t^i}$ where $n_t^i \in \mathbb{N}$, $\sum_{i=1}^n n_t^i = n$. Standard resampling are such that $\{n_t^i\}$ depends of $\{w_t^i\}$ but not of $\{x_t^i\}$. This is clearly an undesirable feature. Consider a bimodal distribution with few samples in a mode and numerous samples in the other. Given two particles with similar weights, one in the "small" mode and one the "big" mode, we would like our resampling scheme to distinguish those particles and preserve the particle in the small mode.

We compare the performance of multinomial and systematic resampling against that of Reich's algorithm and Algorithm 2, with a fixed to $\mathbf{1}_N/N$. For both approaches we used smoothed transports as outlined in Algorithm 3, and not the true transport, whose cost would be prohibitive in these settings. We measure performance in terms of the mean and standard deviation – over 10.000 runs of the particle filter – of the mean square error

$$\sum_{t=1}^{10} \|\widehat{\mathbb{E}}_n\left(X_t|\, y^t\right) - \mathbb{E}\left(X_t|\, y^t\right)\|^2$$

where y^{10} is a fixed realization of the observations. Here $\widehat{\mathbb{E}}_n(X_t|\, y^t)$ is the particle filter estimate with n samples and $\mathbb{E}(X_t|\, y^t)$ the "ground-truth" estimated using a very large number of particles. We set λ adaptively to the inverse of the second decile of all elements in M and threshold entries such that $\lambda M > 200$. We set $\theta = 0.5$. These results agree with Reich's claim that optimal transport resampling is specially useful when the number of particles is small, both in terms of mean error and especially of variance. They also suggest that the repeated updates of locations in Algorithm 2, compared to the single step advocated by Reich, seem to yield better results.

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.

D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proc. of SODA*, pages 1027–1035. SIAM, 2007.

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.

S. Becker, V. Cevher, C. Koch, and A. Kyrillidis. Sparse projections onto the simplex. *Proc. of the Internation Conference on Machine Learning*, 2013.

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

J. Bigot and T. Klein. Consistent estimation of a population barycenter in the Wasserstein space. *arXiv preprint arXiv:1212.2562*, 2012.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013. to appear.

M. Cuturi and D. Avis. Ground metric learning. *Arxiv preprint arXiv:1110.2306*, 2011.

J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

S. Erlander and N. Stewart. *The gravity model in transportation analysis: theory and extensions*. VSP, 1990.

A. V. Fiacco and J. Kyparisis. Convexity and concavity properties of the optimal value function in parametric nonlinear programming. *Journal of Optimization Theory and Applications*, 48(1):95–126, 1986.

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

M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'Institut Henri Poincaré*, volume 10, pages 215–310. Presses Universitaires de France, 1948.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schlkopf, 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. The sinkhorn-knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

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

W. Kraaij, J.-Y. Nie, and M. Simard. Embedding web-based statistical translation models in cross-language information retrieval. *Comput. Linguistics*, 29(3):381–419, 2003.

S. Lloyd. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 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.

A. Naor and G. Schechtman. Planar earthmover is not in l_1. *SIAM J. Comput.*, 37(3):804–826, 2007.

V. Nastase and C. Strapparava. Bridging languages through etymology: The case of cross language text categorization. In *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, 2013.

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

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.

F. Nielsen and R. Nock. Sided and symmetrized bregman centroids. *IEEE Trans. on Information Theory*, 55(6):2882–2904, 2009.

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.

M. Pilanci, L. El Ghaoui, and V. Chandrasekaran. Recovery of sparse probability measures via convex programming. In *Adv. in Neural Infor. Processing Systems 25*, 2012.

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.

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.

A. G. Wilson. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy*, pages 108–126, 1969.

E-mail address: mcuturi@i.kyoto-u.ac.jp

Graduate School of Informatics, Kyoto University

E-mail address: doucet@stat.oxford.ac.uk

Department of Statistics, University of Oxford