

Wasserstein Barycenters Computation

Tarek FRAHI, Romain PERQUY, Giang TRAN

MASEF, Paris Dauphine

May 7th, 2018

Introduction

Recent progress in optimal transport yields to new algorithms that allow computing the mean of empirical probability measures under the Wasserstein distance. The aim of this presentation is therefore to introduce the framework of the Wasserstein barycenters detailed by Agueh and Carlier (2010) and highlight the algorithms developed by Cuturi and Doucet (2014).

- Define **the Wasserstein Barycenter** between N probability measures.
- Present **two original algorithms** to compute Wasserstein barycenters:
 - Using Subgradient Method;
 - Updating Algorithm using Local Quadratic Approximation;
- **Smooth the Wasserstein barycenters** with **an entropic regularization**, and present an updated algorithm with cheaper computational cost.

- 1 Background on Optimal Transport
 - Wasserstein Distances
 - Measures with discrete and finite support
 - Optimal Transport Formulation
 - Wasserstein Barycenters
- 2 Computation of the Barycenters
 - Convexity and Differentiability
 - Fixed Support: Minimizing f over $P(X)$
 - Free Support
- 3 Fast Computation
 - Smoothed Transportation Problems
 - Barycenters Computation
- 4 Applications

Transport Problem

Let (Ω, D) be a metric Polish space and $P(\Omega)$ a set of Borel probability measures on Ω . Let X and Y be two compact metric subspaces of Ω .

Monge-Kantorovich Optimal Transport Problem

Let $\mu, \nu \in P(\Omega)$ and c a continuous cost function $c : X \times Y \longrightarrow \mathbb{R}$. The goal is to find a cost minimizing transport between μ and ν i.e.

$$\inf_{\pi \in \Pi(\mu, \nu)} \left(\int_{X \times Y} c(x, y) d\pi(x, y) \right)$$

where $\Pi(\mu, \nu)$ denotes the set of transport plans between μ and ν i.e. the set of probability measures on $\Omega \times \Omega$ having μ and ν as marginals.

Wasserstein Distance

Definition

Given $p \in [1, +\infty)$, the Wasserstein distance between two probability measures μ and ν in $P(\Omega)$ is defined by

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

We can also write it as

$$W_p(\mu, \nu) = \inf_{(X_1, X_2)} E(|X_1 - X_2|^p)^{1/p}$$

where (X_1, X_2) are all couples of random variables with law μ and ν respectively.

Measures with discrete and finite support

Discrete Probability Simplex

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

Measure supported on a finite set X

Given $X = \{x_1, \dots, x_n\}$ of $n > 1$ points of Ω , we define

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

Measures supported on up to k points

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

Optimal Transport Formulation

Let $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_m\}$ sets of points in Ω .

Define the matrix M_{XY} of **pairwise distances** between elements of X and Y raised to the power p as

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

and the **transport polytope** $U(a, b)$ of $a \in \Sigma_n$ and $b \in \Sigma_m$ as

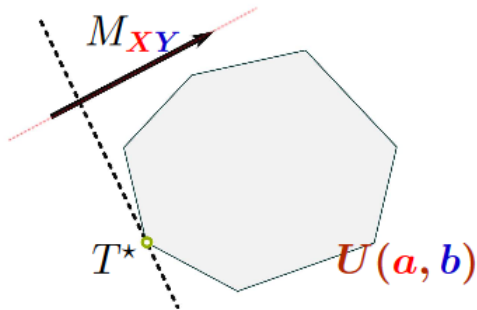
$$U(a, b) \stackrel{\text{def}}{=} \{T \in \mathbb{R}_+^{n \times m} \mid T\mathbf{1}_m = a, T^t\mathbf{1}_n = b\}$$

.

Then, the Wasserstein distance raised to the power p , is the optimum of a Linear Program of $n \times m$ variables

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

Optimal Transport Representation



$$\begin{aligned} W_p^p(\mu, \nu) &= \langle T^*, M_{XY} \rangle \\ &= \min_{T \in U(a, b)} \langle T, M_{XY} \rangle \end{aligned}$$

Definition

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

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

Some special cases and recent work

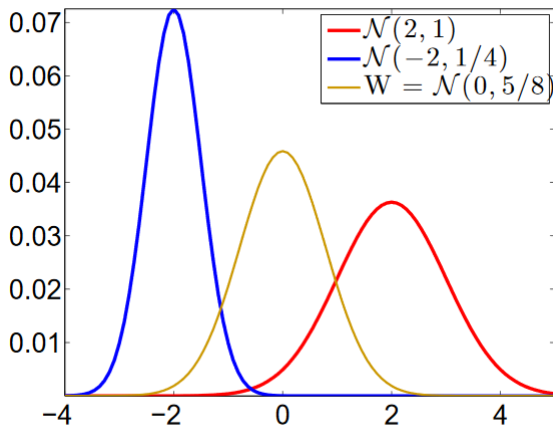
Some special cases:

- $N = 1, \mathbb{P} = P(X)$, a weight vector a can be defined by setting $a_i = \sum_{j=1}^m b_j \Delta_{ij}$ where $\Delta_{ij} = 1$ iff nearest neighbor of y_j in X is x_i .
- Euclidean Ω : $N = 1, D(x, y) = \|x - y\|_2, p = 2; \mathbb{P} = P_k(\Omega)$.
- Constrained k -Means: $N = 1, P = P_k(\Omega, \{\mathbf{1}_k/k\})$.

Recent work:

- Agueh and Carlier (2011) consider conditions on the ν_i for a Wasserstein barycenter in $P(\Omega)$ to be unique in cases:
 - (i) $\Omega = \mathbb{R}$;
 - (ii) $N = 2$ using McCann's interpolant;
 - (iii) all the measures ν_i are Gaussians in $\Omega = \mathbb{R}^d$.
- Rabin et al. (2012) considered practical approaches to compute Wasserstein barycenters between point clouds in \mathbb{R}^d .

Averaging two gaussians



Constraints redundancy

By the definition of the polytope $U(a, b)$, we have $n + m$ constraints, and one of these is redundant because a and b have the same sum equal to 1. We therefore introduce $b_x = T^T \mathbf{1}_{m-1}$ to truncate the constraints.

Dual transportation problem

Given $M \in \mathbb{R}^{n \times m}$, the problem admits the dual form :

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

where the polyhedron

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

Dual optimum and subgradient

Proposition

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

Dual optimum and subgradient

We recall that the epigraph of a function f is the subspace

$$\{(x, e) : x \in \mathbb{R}^n, e \in \mathbb{R}, e \geq f(x)\}$$

Proof

(i) We have that epigraph of $a \rightarrow \alpha^T a + \beta^T b_x$ is a closed halfspace. The epigraph of $S(a, b; M) = \max \alpha^T a + \beta^T b_x$ is then a finite intersection of closed halfspaces, thus $a \rightarrow S(a, b, M)$ is a polyhedral function.

(ii) Suppose that α^* is an optimal solution to the dual. The strong duality implies that $\alpha^{*T} a^* = S'(a^*, M)$ (where S' is the dual map without constants). Consider arbitrary a , then by weak duality we have $\alpha^{*T} a \leq S'(a, M)$. Then $(\alpha^{*T} a - \alpha^{*T} a^*) \leq (S'(a, M) - S'(a^*, M))$. Therefore α^* is a subgradient of S' then S at a^* .

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

Let $X \subset \Omega^n$ be fixed.

Definition

Let $N > 1$. For $a \in \Sigma_n$, the Wasserstein barycenter $f(a)$ is defined by

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

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

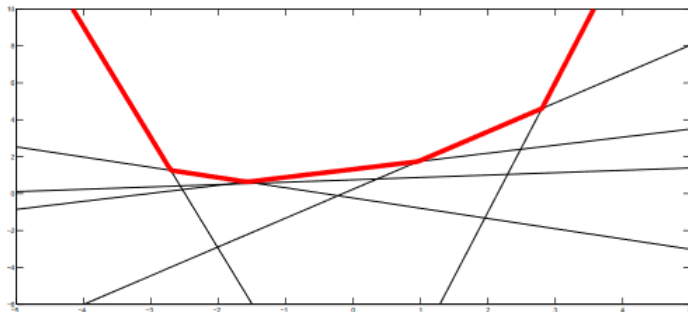
Let α_i^* be the optimal dual variable of $S(a, b_i, M_{XY_i})$.

Corollary

The function f is polyhedral convex on $P(X)$, with subgradient

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

Subgradient iteration



Algorithm 1. p -Wasserstein Barycenters in $P(X)$

Inputs: $X \in \Omega^n$; For $i \leq N$: $Y_i \in \Omega^{m_i}$, $b_i \in \Sigma_{m_i}$, $p \in [1, \infty)$, $t_0 > 0$.

Initialize $a = a_0$, $t = 1$

Form all $n \times m_i$ matrices $M_i = M_{XY_i}$

while not converged **do**

for $i \in \{1, \dots, N\}$ **do**

 Compute α_i^* the dual optimal variable of $S(a, b_i, M_i)$.

end for Subgradient: $\alpha \leftarrow \frac{1}{N} \sum_{i=1}^N \alpha_i^*$

$$a \leftarrow P_{\Sigma_d} \left(a - \frac{t_0 \alpha}{\sqrt{t}} \right); t \leftarrow t + 1$$

end while

Definition

Looking for a barycenter μ with atom X and weight a is equivalent to minimizing the functional f

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

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

Consider $\Omega = \mathbb{R}^d$ with $d \geq 1$, D is the Euclidean distance; $p = 2$.

Euclidean Wasserstein Distance

$$W_2(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 d\pi(x, y) \right)^{1/2}$$

Let $\mathbf{x} \stackrel{\text{def}}{=} \text{diag}(X^T X)$ and $\mathbf{y}_i \stackrel{\text{def}}{=} \text{diag}(Y_i^T Y_i)$, then

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

Transport Cost as a function of X

Let $T \in U(a, b)$, we have

$$\langle T, M_{XY_i} \rangle = \mathbf{x}^T a + \mathbf{y}_i^T b - 2\langle T, X^T Y_i \rangle$$

Discarding constant terms in b and y , minimizing $S(a, b, M_{XY_i})$ w.r.t X is equivalent to

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

Local Quadratic Approximation

Suppose T_i^* is optimal for problem $S(a, b_i, M_{XY_i})$. Updating Eq. (1),

$$\begin{aligned} \langle X^T X, \text{diag}(a) \rangle - \frac{2}{N} \sum_{i=1}^N \langle T_i^*, X^T Y_i \rangle = \\ \|X \text{diag}(a^{1/2}) - \frac{1}{N} \sum_{i=1}^N Y_i T_i^{*T} \text{diag}(a^{-1/2})\|^2 - \left\| \frac{1}{N} \sum_{i=1}^N Y_i T_i^{*T} \text{diag}(a^{-1/2}) \right\|^2 \end{aligned}$$

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

Iterative update

Minimize the local quadratic approximation of S at X yields thus

$$X^* \leftarrow \left(\frac{1}{N} \sum_{i=1}^N Y_i T_i^{*T} \right) \text{diag}(a^{-1})$$

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

Input: $Y_i \in \mathbb{R}^{d \times m_i}$, $b_i \in \Sigma_{m_i}$ for $i \leq N$, $\theta \in [0, 1]$

Initialize X and a

while X and a have not converged **do**

$a \leftarrow a^*$ using Algorithm 1.

for $i \in \{1, \dots, N\}$ **do**

$T_i^* \leftarrow$ optima solution of $S(a, b_i, -X^T Y_i)$

end for

$$X \leftarrow (1 - \theta)X + \theta \left(\frac{1}{N} \sum_{i=1}^N Y_i T_i^{*T} \right) \text{diag}(a^{-1})$$

end while

Smoothed formulation

For a $n \times m$ transport T the entropy $h(T)$ is defined by

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

The primal problem can then be regularized using a constant $\lambda > 0$

Regularized Primal

$$P_{\lambda}(a, b; M) = \min_{T \in U(a,b)} \langle T, M_{XY} \rangle - \frac{1}{\lambda} h(T)$$

Smoothed formulation

The dual problem is a smoothed version of the original dual transportation problem, where the positivity constraints of each term are replaced by an exponential penalty.

Smoothed Dual

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

Proposition (Wilson, 1969 - Cuturi, 2013)

Let $K = e^{-\lambda M_{XY}}$. There exists a pair $(u, v) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$ such that the solutions for the primal and dual are respectively

$$T_\lambda^* = \text{diag}(u) K \text{diag}(v)$$

and

$$\alpha_\lambda^* = -\frac{\log(u)}{\lambda} + \frac{\log(u)^T \mathbf{1}_n}{\lambda n} \mathbf{1}_n$$

Lemma (Sinkhorn, 1967)

For any positive matrix $A \in \mathbb{R}_+^{nm}$ 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 $\text{diag}(u)A\text{diag}(v) \in U(a, b)$.

Such a pair (u, v) can be recovered as a fixed point of the Sinkhorn map

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

Algorithm 3. Fast Computation of Wasserstein Barycenters

Input: M, λ, a, b

$K = \exp(-\lambda M);$

$\tilde{K} = \text{diag}(a^{-1})K;$

$u = \text{ones}(n, 1)/n;$

while u change **do**

$u = 1./(\tilde{K}(b./(K^T u)))$

end while

$v = b./(K^T u);$

$\alpha_{\lambda}^* = -\frac{\log(u)}{\lambda} + \frac{\log(u)^T \mathbf{1}_n}{\lambda n} \mathbf{1}_n;$

$T_{\lambda}^* = \text{diag}(u)K\text{diag}(v);$

Clustering and k-means

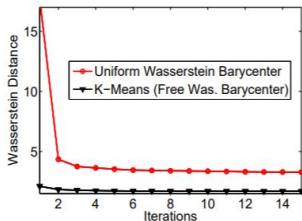
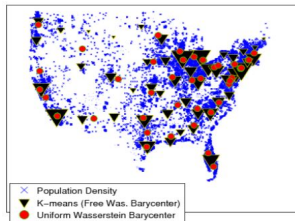
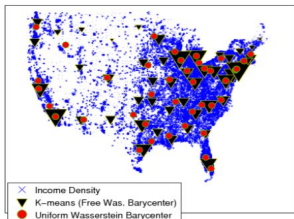
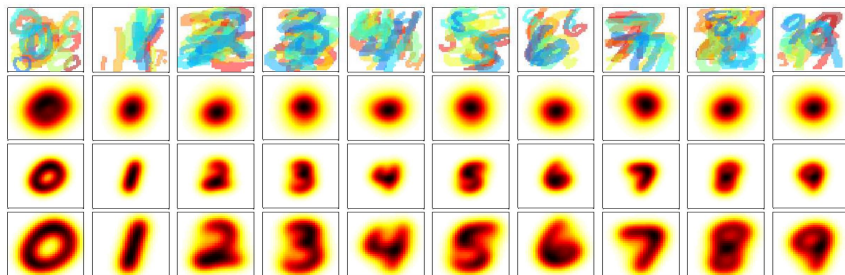





Image processing




References I

 M. Cuturi and A. Doucet
Fast computation of Wasserstein barycenters
2014

 M. Agueh and G. Carlier
Barycenters in the Wasserstein space
2010

 M. Cuturi
Light speed computation of optimal transport distances
2013

 J. Benamou, G. Carlier, M. Cuturi, L. Nenna, G. Peyre
Iterative Bregman Projections for Regularized Transportation Problems
2014

References II



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



C. Villani
Optimal transport: old and new
Springer, 2009