A novel notion of barycenter for probability distributions based on optimal weak mass transport

Elsa CAZELLES[1], Felipe TOBAR[2, 3], Joaquín FONTBONA[2]

[1]IRIT, Université de Toulouse, CNRS
118 Route de Narbonne, 31062 Toulouse, France

[2]CMM & [3]IDIA, Universidad de Chile
851 Avenida Beauchef, Santiago, Chile

elsa.cazelles@irit.fr, ftobar@uchile.cl, fontbona@dim.uchile.cl

Résumé – Afin de résumer l'information d'une famille finie de mesures de probabilité, il est naturel de considérer leur moyenne de Fréchet par rapport à la distance de Wasserstein-2, c'est-à-dire le barycentre de Wasserstein. Nous définissons de manière analogue un barycentre basé sur le transport optimal faible, qui correspond à un coût barycentrique quadratique de transport [8], [4]. Nous adressons l'interprétation de ces *barycentres faibles* au moyen de l'ordre convexe entre mesures de probabilité et nous montrons que, plutôt que de moyenner les géométries d'un ensemble de distributions (comme le fait le barycentre de Wasserstein classique), les barycentres faibles extraient une information géométrique commune à toutes les mesures, et celle-ci peut être encodée par une variable aléatoire latente sous-jacente.

Abstract – A natural method for averaging a finite family of probability measures is to compute their Fréchet mean with respect to the Wasserstein-2 distance, that is the Wasserstein barycenter. We define, in a similar way, a barycenter based on optimal weak transport, which corresponds to quadratic barycentric transport costs [8], [4]. We discuss the interpretation of these *weak barycenters* in the light of convex ordering between probability measures and we show that, rather than averaging the input distributions in a geometric way (as the Wasserstein barycenter based on classic optimal transport does), weak barycenters extract common geometric information shared by all the input distributions, encoded as a latent random variable that underlies all of them.

1 Introduction

This article [1] explores theoretical features and potential applications to machine learning of barycenters of probability measures analogously defined in terms of *optimal weak transport* (OWT, see [8]), or more precisely quadratic barycentric transport costs. In a nutshell, for a source measure μ and a target measure ν, the OWT problem aims to transport mass so that the conditional spatial mean of target support points y, given their source support points x, is close to x in average. This amounts to finding an intermediate measure η, possibly *more concentrated* than ν in the sense of convex ordering of probability measures, which is *close* to μ with respect to (wrt) the Wasserstein-2 distance.

Our main motivation is to investigate the effect and meaning of combining a family of probability measures using OWT instead of OT. To that end, we will define the *weak barycenter* of this family through an optimisation problem, and discuss some of its properties. Importantly, we will see that, rather than averaging the input distributions in a metric sense, solving a weak barycenter problem corresponds to finding probability measures that encode geometric or shape information *shared across* all of them. In fact, the weak barycenter problem will be interpreted as finding a latent random variable common to all the input distributions. Implications of this latent variable interpretation, in terms of robustness to outliers, will also be drawn in our work.

A second motivation is to develop and implement computational methods for weak barycenters, capitalising on the fact that the optimal weak coupling between *any* pair of distributions, with finite second moments, is always realised by a unique optimal *map*. This property is in sharp contrast to standard OT, where the absolute continuity (a.c.) wrt the Lebesgue measure of the source or target measure is typically needed to grant the existence and uniqueness of a map—the so-called Monge map—realising the optimal coupling between them. This map is often required in different ways to compute Wasserstein barycenters (see [2], [12] or [9]).

2 Optimal –weak– transport and barycenters of distributions

2.1 Background on optimal transport

The optimal transport (OT) problem [11] aims to find the lowest cost to transfer the mass from one probability measure

1. The content of this paper is published at NeurIPS 2021 [5].

onto another, capitalising on their geometric information. In particular, the Wasserstein-2 distance W_2 metrises the space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures on $\mathbb{R}^d$ with finite 2-moment. Precisely, for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

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

where $\pi \in \Pi(\mu,\nu)$ is a *transport plan*, that is a probability measure on the product space $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν. For μ a.c., the unique optimal plan π^* is concentrated on the graph of an unique measurable map T^* and Eq. (1) is equivalent to Monge's problem

$$W_2(\mu,\nu) = \left(\min_{T \in \mathbb{T}(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-T(x)\|^2 \mathrm{d}\mu(x) \right)^{1/2}, \quad (2)$$

where $\mathbb{T}(\mu,\nu)$ is the set of measurable functions $T : \mathbb{R}^d \to \mathbb{R}^d$ such that [2] $\nu = T\#\mu$. In that case we have $\pi^* = (\mathrm{id}, T^*)\#\mu$. Interestingly, T^* can be written as the barycentric projection of π^*, that is $T^*(x) = \int_{\mathbb{R}^d} y \mathrm{d}\pi^*_x(y)$, where π^*_x is the disintegration of the transport plan $\pi^* \in \Pi(\mu,\nu)$ wrt the first marginal μ i.e.

$$\pi^*(\mathrm{d}x\mathrm{d}y) = \pi^*_x(\mathrm{d}y)\mu(\mathrm{d}x).$$

In fact, one can construct the barycentric projection from any transport plan $\pi \in \Pi(\mu,\nu)$:

$$S^\nu_\mu(x) := \int_{\mathbb{R}^d} y \mathrm{d}\pi_x(y) \overset{\text{also}}{=} \mathbb{E}(Y|X=x), \text{ with } (X,Y) \sim \pi.$$

Finally, the classical Wasserstein barycenter problem for a set of probability measures $\nu_1, \dots, \nu_n \in \mathcal{P}_2(\mathbb{R}^d)$ with weights $\lambda_1, \dots, \lambda_n$ in the simplex (i.e. $\lambda_i \geq 0$ and $\sum_{i=1}^n \lambda_i = 1$) is defined [1] by

$$\operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^n \lambda_i W_2^2(\mu,\nu_i). \quad (3)$$

2.2 Optimal weak transport

The OWT introduced in [8], and especially the case of barycentric transport costs, is defined for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ by

$$V(\mu|\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d} \|x - \int_{\mathbb{R}^d} y \mathrm{d}\pi_x(y)\|^2 \mathrm{d}\mu(x). \quad (4)$$

The following results from [3] (stated for our specific setting) lay the ground for our proposed weak barycenters. We denote by $\eta \leq_c \nu$ the *convex order of measures*, meaning that $\int \phi \mathrm{d}\eta \leq \int \phi \mathrm{d}\nu$ for any convex function ϕ that is nonnegative or integrable wrt $\eta + \nu$. The optimisation problem in Eq. (4) can then be reformulated thanks to the Brenier-Strassen theorem [7], [3], through the notion of convex ordering.

Theorem 1 ([3], Th. 1.2 & Th. 1.4)**.** *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_1(\mathbb{R}^d)$, then the OWT problem* (4) *admits a unique minimiser. Moreover, there exists a unique $\eta^* \leq_c \nu$ such that*

$$W_2^2(\mu,\eta^*) = \inf_{\eta \leq_c \nu} W_2^2(\mu,\eta) = V(\mu|\nu). \quad (5)$$

2. The *pushforward* operator $\#$ is defined such that for any measurable set $B \subset \mathbb{R}^d$, $\nu(B) = \mu(T^{-1}(B))$.

Additionally, there exists a convex function $\psi : \mathbb{R}^d \to \mathbb{R}$ of class C^1 with $\nabla\psi$ being 1-Lipschitz, such that $\nabla\psi\#\mu = \eta^*$. Finally, the optimal coupling $\pi^{\mu,\nu} \in \Pi(\mu,\nu)$ verifies $\int y \mathrm{d}\pi^{\mu,\nu}_x(y) = \nabla\psi(x)$ μ-a.s.

This strongly differs from the classical OT setting, for which the uniqueness of an optimal transport plan is not guaranteed for arbitrary measures. Note that V is no longer a distance : it is not symmetric, the homogeneity property is not verified i.e. $V(\mu|\nu)$ does not imply $\mu = \nu$, but it is still positive and finite.

Fig. 1 illustrates the differences between the distributions $S^{OT}\#\mu$ and $S^{OWT}\#\mu$ constructed respectively from an OT plan in Eq. (1) and the optimal weak plan in Eq. (4) between two measures μ and ν. The measure $S^{OT}\#\mu$ reasonably fits the target distribution ν, and $S^{OWT}\#\mu \leq_c \nu$.

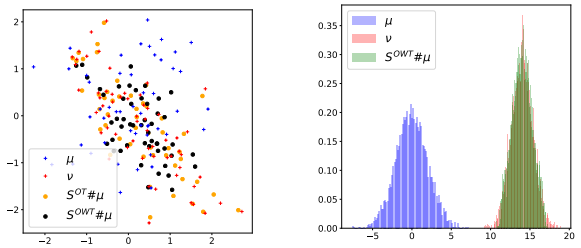

FIGURE 1 – Examples of pushforward measures constructed from barycentric projections for two measures μ and ν in two dimensions **(left)** and one dimension **(right)**.

Computation of OWT. For two discrete measures $\mu = \sum_{i=1}^r a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$, the OWT problem boils down to solving the following quadratic programming problem with linear constraints :

$$\min_{\pi \in \mathbb{R}^{r \times m}} \left\{ \sum_{i=1}^r a_i \left| x_i - \left(\frac{\pi \mathbf{y}}{\mathbf{a}} \right)_i \right|^2, \pi_{ij} \geq 0,\ \pi \mathbb{1} = a,\ \pi^T \mathbb{1} = b \right\},$$

which can be solved using a solver such as **cvxpy**. We also propose to solve the OWT problem with a proximal algorithm. The optimal barycentric projection is then constructed as $\frac{\pi \mathbf{y}}{\mathbf{a}}$.

2.3 Weak barycenters

As in the OT framework, and based on OWT in Eq. (4), we define weak barycenters as follows.

Definition 1. *The set of weak barycenters of a finite family of measures $\{\nu_i\}_{i=1,\dots,n} \in \mathcal{P}_2(\mathbb{R}^d)$ with weights $\{\lambda_i\}_{i=1,\dots,n}$ in the simplex is defined as*

$$\operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^n \lambda_i V(\mu|\nu_i). \quad (6)$$

A weak barycenter averages, wrt to the Wasserstein metric, an optimally chosen set of probability measures $\{\eta_1, \dots, \eta_n\}$ which are *more concentrated* than the corresponding ν_i, in the sense that $\eta_i \leq_c \nu_i$ for each $1 \leq i \leq n$. The existence of a solution is established in the next proposition (see proof in [5]).

Proposition 1. *The weak barycenter problem in Eq.* (6) *admits a minimiser $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.*

In the following, we denote by X and Y_i random variables with respective laws μ and ν_i, and δ_a the Dirac measure supported on $a \in \mathbb{R}^d$. First intuitions are presented in the following.

Lemma 1. *If μ is a weak barycenter of $\{\nu_i\}_{i=1\dots,n}$ and $\mu' \leq_c \mu$, then μ' also is a weak barycenter. In particular, the Dirac measure supported on $\mathbb{E}_\mu(X)$ is always a weak barycenter. Moreover, a Dirac distribution $\delta_{\bar{\omega}}$ is a weak barycenter if and only if $\bar{\omega} = \sum_{i=1}^n \lambda_i \mathbb{E}_{\nu_i}(Y_i)$.*

Proposition 2. *A measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ is a weak barycenter of $\{\nu_i\}_{i=1\dots,n}$ if and only if its mean satisfies $\mathbb{E}_\mu(X) = \sum_{i=1}^n \lambda_i \mathbb{E}_{\nu_i}(Y_i)$ and $\hat{\mu} \leq_c \hat{\nu}_i$ holds for all $1 \leq i \leq n$, where $\hat{\nu}$ denotes the centered version of a law ν.*

For instance, in the case of one dimensional Gaussian distributions $\nu_i = \mathcal{N}(m, \sigma_i^2)$, the set of weak barycenters includes $\{\mu = \mathcal{N}(m, \sigma^2) \mid 0 \leq \sigma^2 \leq \sigma_i^2 \ \forall i\}$. As for the Wasserstein barycenter, it is given by $\mathcal{N}(m, (\sum \lambda_i \sigma_i)^2)$.

2.4 Weak barycenters as latent variables

A weak barycenter encodes common geometric information present in all the input measures, and this can be intuitively and rigorously interpreted as being the distribution of a latent variable underlying the realisations of random variables of laws ν_i for all $i = 1, \dots, n$.

Theorem 2. *Let μ be a weak barycenter of $\{\nu_i\}_{i=1\dots,n}$. Then, for each $1 \leq i \leq n$, a random variable $Y_i \sim \nu_i$ can be realised as*

$$Y_i = X + (\mathbb{E}Y_i - \mathbb{E}X) + \bar{Y}_i,$$

where $X \sim \mu$ and $\bar{Y}_i = Y_i - \mathbb{E}(Y_i|X)$ is centered conditionally on X. Moreover, one has $S_\mu^\nu(X) = X + (\mathbb{E}Y_i - \mathbb{E}X)$ for all $i = 1, \dots, n$. Finally, we have $\mathbb{E}(Y_i - \mathbb{E}Y_i|X - \mathbb{E}X) = X - \mathbb{E}X$ or, equivalently, $\hat{\mu} \leq_c \hat{\nu}_i$, with $\hat{\mu}$ and $\hat{\nu}_i$ the laws of $X - \mathbb{E}X$ and $Y_i - \mathbb{E}Y_i$ respectively.

That is to say, each $Y_i \sim \nu_i$ can be realised by sampling a random variable X common to all $i = 1, \dots, n$ and distributed according to a weak barycenter μ, translating that value by $\mathbb{E}Y_i - \mathbb{E}X$ and adding a *cluster-specific* component $\bar{Y}_i$ or idiosyncratic noise, centered conditionally on X.

Robustness to outliers. The observations of each class can be naturally interpreted as perturbations wrt the (translated) law of the weak barycenter, or outliers wrt the associated latent random variable. We illustrate this in Fig. 2. We consider two sets of 50 observations sampled from 2D Gaussian measures, where each observation may be corrupted by random translations (Bernoulli $p = 0.1$) thus producing outliers. We show the resulting barycenters (right) for Wasserstein barycenter (red) and weak barycenter (black), the latest shows robustness to outliers.

3 Algorithms and numerical experiments

Computation of weak barycenters. Akin to the fixed-point methodology in the classical Wasserstein scenario, we define an

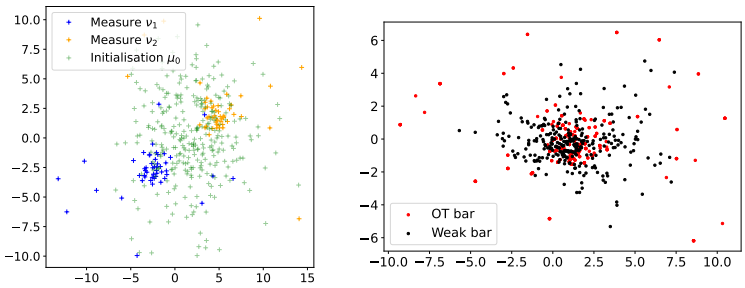

FIGURE 2 – **(Left)** Empirical Gaussian distributions with corrupted observations, and initialisation μ_0 of the iterative procedure presented in Section 3. **(Right)** The associated OWT (black) and OT (red) barycenters.

iterative rule valid for arbitrary probability measures $\nu_1, \dots, \nu_n \in \mathcal{P}_2(\mathbb{R}^d)$ based on the barycentric projection :

$$\mu_{k+1} = G(\mu_k), \text{ with } G(\mu) = \left(\sum_{i=1}^n \lambda_i S_\mu^{\nu_i}\right) \# \mu, \qquad (7)$$

where for each $i = 1, \dots, n$, the optimal barycentric projection is given by $S_\mu^{\nu_i} : x \mapsto \int y \mathrm{d}\pi_x^{\mu,\nu_i}(y)$, for $\pi^{\mu,\nu_i} \in \Pi(\mu, \nu_i)$ achieving the minimum in the OWT problem in Eq. (4).

A fundamental difference with the Wasserstein barycenter [2] is that the optimal Monge map T_μ^ν in the OT problem verifies $T_\mu^\nu \# \mu = \nu$, whereas the pushforward measure $S_\mu^\nu \# \mu$ in the OWT setting still depends on μ. We will then prove that the iterative algorithm in Eq. (7), based on the maps $S_\mu^{\nu_i}$, admits converging subsequences. For that purpose, we prove the following fundamental result.

Theorem 3. *The function $\mu \mapsto G(\mu)$ defined in Eq. (7) is W_2-continuous from $\mathcal{P}_2(\mathbb{R}^d)$ to $\mathcal{P}_2(\mathbb{R}^d)$.*

Using an approach similar to [2] for the Wasserstein barycenter, the proposed fixed-point procedure is built on the next proposition.

Proposition 3. *If μ is a weak-barycenter, that is, a solution of problem (6), then $G(\mu) = \mu$, i.e., $x = \sum_{i=1}^n \lambda_i S_\mu^{\nu_i}(x), \mu(x)$-a.s.*

The inverse implication of Proposition 3 is not necessarily true, that is, some fixed points may not be weak barycenters. However, a Dirac delta $\delta_\omega, \omega \in \mathbb{R}^d$, that meets the fixed-point condition $\delta_\omega = G(\delta_\omega)$, is a weak barycenter (see Lemma 1).

Proposition 4. *Let $(\mu_k)_k$ be the sequence defined by the iterative procedure $\mu_{k+1} = G(\mu_k)$ and starting from $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then $(\mu_k)_k$ is tight and every converging subsequence must converge to a fixed point of G.*

These results also hold for the Wasserstein barycenter of *a.c.* measures $\{\nu_i\}_{i=1\dots,n}$ such that at least one of them has a bounded density. Moreover, the inverse implication, namely if μ is a fixed-point then it is a barycenter, is not straightforward even in the Wasserstein barycenter case. For that case one considers the fixed-point equation given by $\mu = (\sum_{i=1}^n \lambda_i T_\mu^{\nu_i}) \# \mu$, with $T_\mu^{\nu_i}$

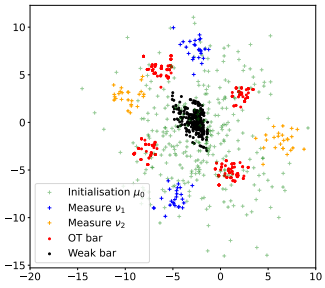

FIGURE 3 – Two Gaussian mixtures point clouds ν_1, ν_2 and their OWT (black) and OT (red) barycenters.

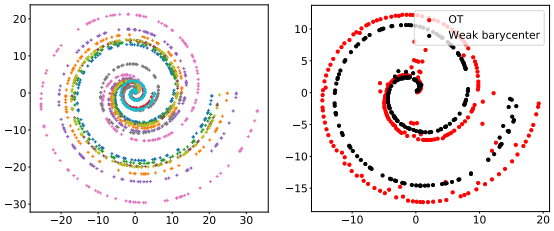

FIGURE 4 – **(Left)** 10 distributions supported on spiral. **(Right)** OWT (black) and OT (red) barycenters.

the Monge map verifying $\nu_i = T_\mu^{\nu_i} \# \mu$ in Eq. (2). Indeed, [1] and [12, Theorem 2] provide additional conditions for this to be true by essentially invoking more smoothness on the distributions $\{\nu_i\}_{i=1\dots,n}$.

Case of a stream of distributions. We also propose a stochastic iterative algorithm for computing a weak barycenter for distributions obtained sequentially, that can be found in [5].

Numerical experiment. We compare weak and OT barycenters by running the iterative procedure in Eq. (7) for the barycentric projections associated respectively to an optimal weak plan in 4 and an OT plan in (1) as in [6, 10].

We present three experiments. In Fig. 3, we consider two points clouds sampled from Gaussian mixtures (orange and blue). The weak and OT barycenters clearly behaves differently : the OWT barycenter shows the convex order property. In Fig. 4, we have 10 empirical distributions supported on spirals (left), with random ratio in $(0, 3)$. The weak barycenter (right) seems to better preserve the shape of the spiral than the OT barycenter, which is in line with the latent variable interpretation of Theorem 2. In Fig. 5, we consider two noisy points clouds sampled from distributions supported on ellipses. The OWT barycenter (black) seems to be more robust to the noise coming from ν_1 and ν_2.

Open questions

We identify two main theoretical aspects for further research :

- General conditions on the family of input measures for the existence of weak barycenters that are not Dirac masses.

- Conditions on input measures for a maximal weak barycenter (in terms of convex ordering) to exist when $d \geq 2$.

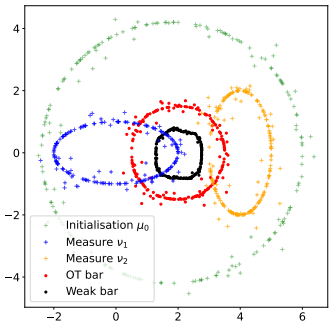

FIGURE 5 – Two distributions supported on ellipses and their OWT (black) and OT (red) barycenters.

Références

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

[2] P.C. Álvarez-Esteban, E. Del Barrio, J.A. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2) :744–762, 2016.

[3] J. Backhoff-Veraguas, M. Beiglböck, and G. Pammer. Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations*, 58(6) :203, 2019.

[4] J. Backhoff-Veraguas, M. Beiglböck, and G. Pammer. Weak monotone rearrangement on the line. *Electronic Communications in Probability*, 25, 2020.

[5] E. Cazelles, F. Tobar, and J. Fontbona. A novel notion of barycenter for probability distributions based on optimal weak mass transport. *Advances in Neural Information Processing Systems*, 34, 2021.

[6] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. *Proceedings of the 31st International Conference on Machine Learning*, 32(2) :685–693, 2014.

[7] N. Gozlan and N. Juillet. On a mixture of Brenier and Strassen theorems. *Proceedings of the London Mathematical Society*, 120(3) :434–463, 2020.

[8] N. Gozlan, C. Roberto, P.-M. Samson, and P. Tetali. Kantorovich duality for general transport costs and applications. *Journal of Functional Analysis*, 273(11) :3327–3405, 2017.

[9] G. Rios, J. Backhoff-Veraguas, J. Fontbona, and F. Tobar. Bayesian learning with Wasserstein barycenters. *arXiv preprint arXiv :1805.10833v4*, 2018.

[10] V. Seguy, B.B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large-scale optimal transport and mapping estimation. In *International Conference on Learning Representations (ICLR)*, 2018.

[11] C. Villani. *Optimal Transport : Old and New*. Springer Berlin Heidelberg, 2008.

[12] Y. Zemel and V. Panaretos. Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, 25(2) :932–976, 2019.