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

Elsa Cazelles[1], Felipe Tobar[2,3] and Joaquin Fontbona[2]

[1]IRIT, Université de Toulouse, CNRS
[2]Center for Mathematical Modeling, Universidad de Chile
[3]Initiative for Data & Artificial Intelligence, Universidad de Chile

NeurIPS 2021

Optimal transport problem: Wasserstein distance

Let μ, ν be two measures supported on $\mathbb{R}^d$ with finite moment of order 2,

Kantorovich's problem

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

where $\Pi(\mu, \nu) =$\{product measures with marginals μ and ν\}.

Optimal transport problem: Wasserstein distance

Let μ, ν be two measures supported on $\mathbb{R}^d$ with finite moment of order 2,

Kantorovich's problem

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

where $\Pi(\mu,\nu) =$\{product measures with marginals μ and ν\}.

Monge's problem

If μ is absolutely continuous,

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

with $\mathbb{T}(\mu,\nu) =$\{measurable functions $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $\nu = T\#\mu$\}.

Barycentric projection

For μ absolutely continuous,

$$W_2^2(\mu,\nu) = \int \|x - T^*(x)\|^2 \mathrm{d}\mu(x) = \iint \|x-y\|^2 \mathrm{d}\pi^*(x,y), \text{ with } \pi^* = (\mathrm{id}, T^*)\#\mu.$$

And $T^*(x) = \int_{\mathbb{R}^d} y \mathrm{d}\pi_x^*(y)$, where π_x^* is the disintegration of the transport plan $\pi^* \in \Pi(\mu,\nu)$ with respect to the first marginal μ i.e.

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

Barycentric projection

For μ absolutely continuous,

$$W_2^2(\mu,\nu) = \int \|x - T^*(x)\|^2 \mathrm{d}\mu(x) = \iint \|x-y\|^2 \mathrm{d}\pi^*(x,y), \text{ with } \pi^* = (\mathrm{id}, T^*)\#\mu.$$

And $T^*(x) = \int_{\mathbb{R}^d} y \mathrm{d}\pi_x^*(y)$, where π_x^* is the disintegration of the transport plan $\pi^* \in \Pi(\mu,\nu)$ with respect to the first marginal μ i.e.

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

Barycentric projection

$$S_\mu^\nu(x) := \int_{\mathbb{R}^d} y \mathrm{d}\pi_x^{\mu,\nu}(y)$$

$\rightarrow$ Which plan to choose for the construction?

Optimal weak transport problem

Optimal weak transport [Gozlan, Roberto, Samson, Tetali (2017)]

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

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

Optimal weak transport problem

Optimal weak transport [Gozlan, Roberto, Samson, Tetali (2017)]

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

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

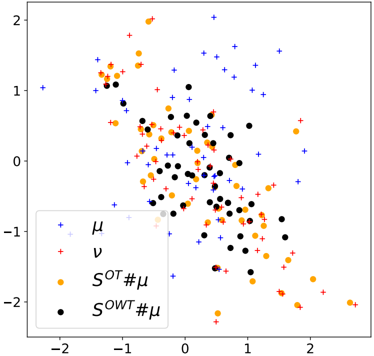
Main advantages:

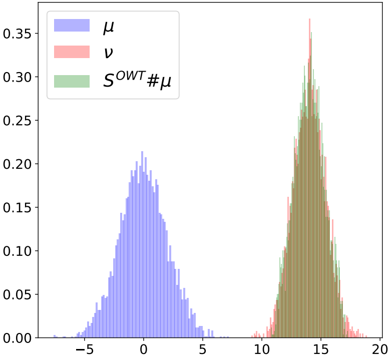
- The optimal plan π is unique for any distribution.
- Characterization via convex ordering [Gozlan and Juillet (2020)] and [Backhoff-Veraguas, Beiglböck, Pammer (2019)]:

$$V(\mu|\nu) = \inf_{\eta \leqslant_c \nu} W_2^2(\mu, \eta) = W_2^2(\mu, S_\mu^\nu \# \mu),$$

where $\eta \leqslant_c \nu$ stands for the *convex ordering of measures*: for any ϕ convex function, $\int \phi \, \mathrm{d}\eta \leqslant \int \phi \, \mathrm{d}\nu$.

About the barycentric projection





- $S^{OT}(x) = \int y \mathrm{d}\pi_x^{OT}(y)$, with π^{OT} optimal in the OT sense.
- $S^{OWT}(x) = \int y \mathrm{d}\pi_x^{OWT}(y)$, with π^{OWT} optimal in the OWT sense.

Wasserstein barycenters

Let $\nu_1, \dots, \nu_k \in \mathcal{P}_2(\mathbb{R}^d)$ and $\lambda_1, \dots, \lambda_k$ weights in the simplex.

Wasserstein barycenter [Agueh and Carlier(2011)]

$$\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min} \sum_{i=1}^{k} \lambda_i W_2^2(\mu, \nu_i)$$

For distributions $\nu_1, \dots, \nu_k$ absolutely continuous such that ν_1 has a bounded density

$$\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min} \sum_{i=1}^{k} \lambda_i W_2^2(\mu, \nu_i) = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min} \sum_{i=1}^{k} \lambda_i \int_{\mathbb{R}^d} \|x - T_\mu^{\nu_i}(x)\|^2 \mathrm{d}\mu(x),$$

where $T_\mu^{\nu_i}$ is optimal in the Monge problem and in particular $T_\mu^{\nu_i} \# \mu = \nu_i$, and the unique barycenter* $\tilde{\mu}$ verifies

$$\tilde{\mu} = \left(\sum_{i=1}^{k} \lambda_i T_{\tilde{\mu}}^{\nu_i} \right) \# \tilde{\mu},$$

*Fixed point characterisation [Álvarez-Esteban, del Barrio, Cuesta-Albertos and Matrán (2016)] and [Zemel and Panaretos (2019)]

Our contribution : weak barycenters for probability measures

Weak barycenter

$$\arg\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^k \lambda_i V(\mu|\nu_i)$$

For any distribution $\nu_1, \dots, \nu_k$,

$$\arg\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^k \lambda_i V(\mu|\nu_i) = \arg\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^k \lambda_i \int_{\mathbb{R}^d} \|x - S_\mu^{\nu_i}(x)\|^2 \mathrm{d}\mu(x),$$

where $S_\mu^{\nu_i}$ is optimal in the weak problem, and and a weak barycenter $\bar{\mu}$ verifies

$$\bar{\mu} = \left(\sum_{i=1}^k \lambda_i S_{\bar{\mu}}^{\nu_i}\right) \# \bar{\mu},$$

Interpretation as a latent variable model

Theorem

Assume that μ is a weak barycenter of $\{\nu_i\}_{i=1,\ldots,k}$, which is not a Dirac measure. Then, for each $i = 1, \ldots, k$, the random variable $Y_i \sim \nu_i$ can be realised as

$$Y_i = X \quad + \quad \underbrace{(\mathbb{E}Y_i + \mathbb{E}X)}_{\text{translation}} \quad + \quad \underbrace{\bar{Y}_i}_{\substack{\text{idiosyncratic or cluster} \\ \text{specific component}}}$$

where $X \sim \mu$ and $\bar{Y}_i = Y_i - \mathbb{E}(Y_i|X)$.

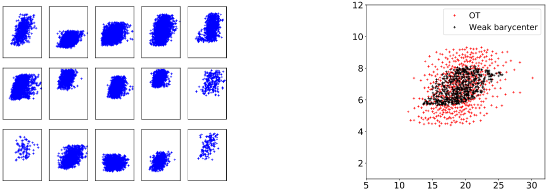


Figure: **Left**: Cytometry dataset for $n = 15$ patients and FSC vs. SSC cell's marker. **Right** : The weak barycenter (black) and the OT barycenter (red). The data are represented with the same axis as the figure of barycenters.

Robustness to outliers

$$Y_i = X \quad + \quad \underbrace{(\mathbb{E}Y_i + \mathbb{E}X)}_{\text{translation}} \quad + \quad \underbrace{\bar{Y}_i}_{\text{cluster specific}}$$

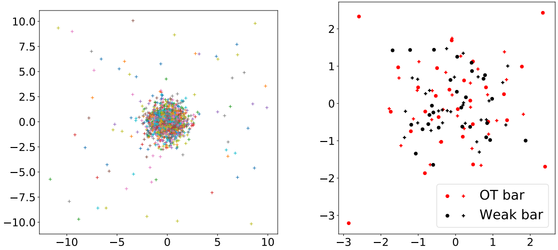


Figure: Empirical Gaussian distributions and their OWT (black) and OT (red) barycenters for Gaussian observations (crosses) and corrupted observations (dots).

Algorithms for computing the weak barycenter problem

Iterative procedure

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

where $S_\mu^{\nu_i} = \displaystyle\int y \mathrm{d}\pi_x^{\mu,\nu_i}(y)$, with $\pi^{\mu,\nu_i} \in \Pi(\mu, \nu_i)$ achieving the minimum for the optimal weak problem.

Algorithms for computing the weak barycenter problem

Iterative procedure

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

where $S_\mu^{\nu_i} = \int y \mathrm{d}\pi_x^{\mu,\nu_i}(y)$, with $\pi^{\mu,\nu_i} \in \Pi(\mu, \nu_i)$ achieving the minimum for the optimal weak problem.

For a stream of data

Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d), \nu_k \overset{i.i.d.}{\sim} \mathbb{Q}$ and $\gamma_k > 0$. We define the following iterative procedure for $k \geqslant 0$:

$$\mu_{k+1} = \left[(1-\gamma_k)\mathrm{id} + \gamma_k S_{\mu_k}^{\nu^k}\right] \# \mu_k,$$

with $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$ and $\sum_{k=1}^{\infty} \gamma_k = \infty$.

Stochastic gradient descent in the classical Wasserstein setting: [Backhoff-Veraguas, Fontbona, Rios, Tobar (2018)] and [Chewi, Maunu, Rigollet, Stromme (2020)].

Spiral distributions

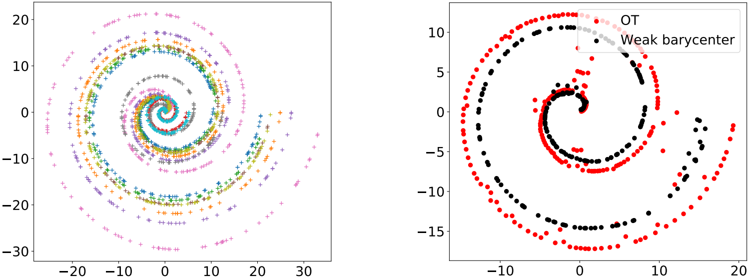


Figure: **Left**: $k = 10$ distributions supported on spiral, each distribution consists of p random points, with p randomly chosen in $(200, 225)$. **Right**: Weak (black) and OT (red) barycenters.

MNIST dataset

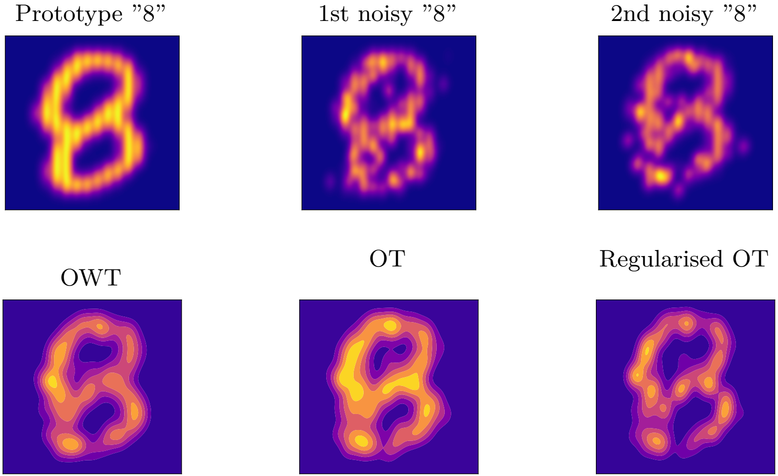


Figure: Digit "8" from MNIST dataset. **Top**: (left) Prototype "8". (middle & right) Noisy versions of the prototype by randomly (Bernoulli $p = 0.1$) moving pixels. **Bottom**: Comparison of three barycenters : OWT plan (left), OT plan (middle) and entropy regularised OT plan for $\varepsilon = 1$ (right).

Closing remarks

Conclusions

- Definition of a weak barycenter, that compiles the common geometric information of the input distributions.
- Interpretation as a latent variable.
- Two algorithms for i) a fixed set of data and ii) streaming data.

Future work:

→ 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 $d \geqslant 2$, among all the solutions of the weak barycenter problem. When $d = 1$, a maximal barycenter exists thanks to the complete lattice property of the set of probability measures wrt the convex ordering.