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

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

[1]IRIT, Université de Toulouse, CNRS
[2]Center for Mathematical Modeling, University of Chile

Journée GdR MIA, 22 Octobre 2021

Barycenters in the space of probability measures [Agueh, Carlier (2011)]

$\mathbb{L}_2$ barycenter

Wasserstein barycenter

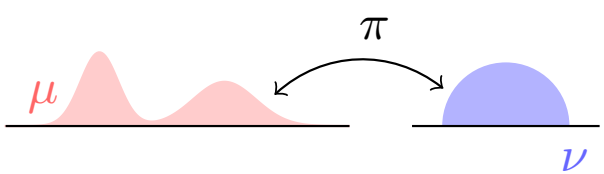
Wasserstein distance

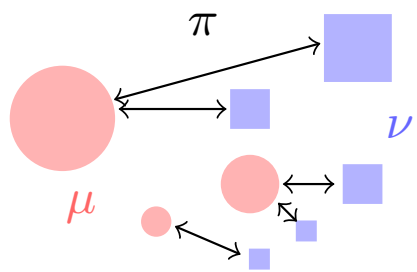
Let μ, ν be two measures supported on $\mathbb{R}^d$ with moment of order 2, i.e. $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$.

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 π is a transport plan that belongs to the space $\Pi(\mu, \nu)$ of the product measures with marginals μ and ν.





Wasserstein distance

Let μ, ν be two measures supported on $\mathbb{R}^d$ with moment of order 2, i.e. $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$.

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 π is a transport plan that belongs to the space $\Pi(\mu, \nu)$ of the product measures with marginals μ and ν.

If $\mu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ is absolutely continuous (a.c.) wrt Lebesgue measure, then

Monge's problem

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

where $\mathbb{T}(\mu, \nu)$ is the set of measurable functions $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $\nu = T\#\mu$, i.e. for any measurable set $B \subset \mathbb{R}^d, \nu(B) = \mu(T^{-1}(B))$.

Barycenter for probability measures

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

Wasserstein barycenter

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

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

Barycenter for probability measures

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

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

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

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

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* $\bar{\mu}$ verifies

$$\bar{\mu} = \left(\sum_{i=1}^{k} \lambda_i T_{\bar{\mu}}^{\nu_i} \right) \# \bar{\mu}, \rightarrow \textbf{ Iterative procedure}$$

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

How to get rid of the absolutely continuous assumptions on the measures?

For $\mu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$, there exists T^* and π^* such that $T^*\#\mu = \nu, \pi^* \in \Pi(\mu, \nu)$ and

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

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.

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

How to get rid of the absolutely continuous assumptions on the measures?

For $\mu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$, there exists T^* and π^* such that $T^*\#\mu = \nu, \pi^* \in \Pi(\mu, \nu)$ and

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

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

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

How to get rid of the absolutely continuous assumptions on the measures?

For $\mu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$, there exists T^* and π^* such that $T^*\#\mu = \nu, \pi^* \in \Pi(\mu,\nu)$ and

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

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

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

→ Which plan to choose for the construction?

Optimal weak transport problem

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

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

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

Optimal weak transport problem

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

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

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

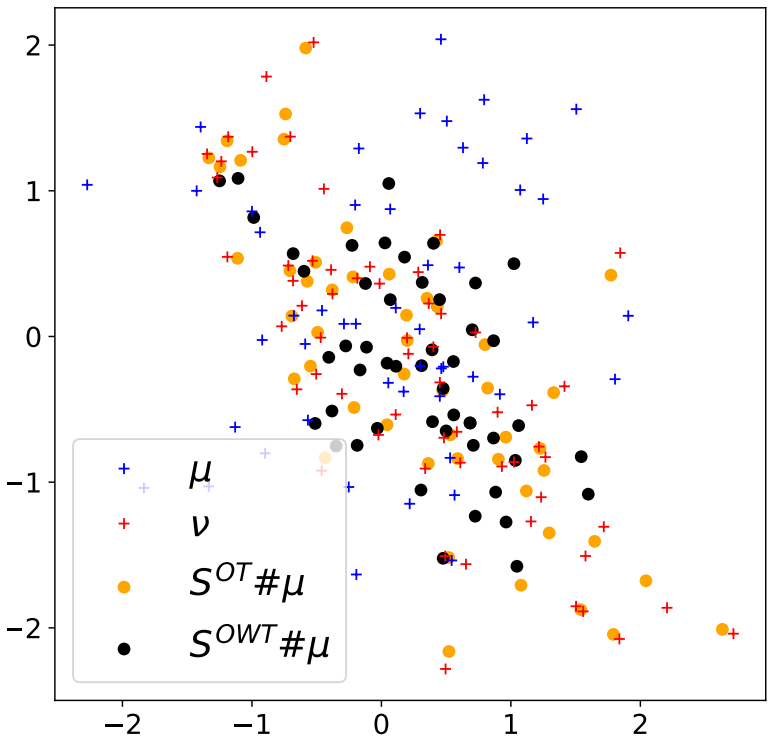
Main advantages:

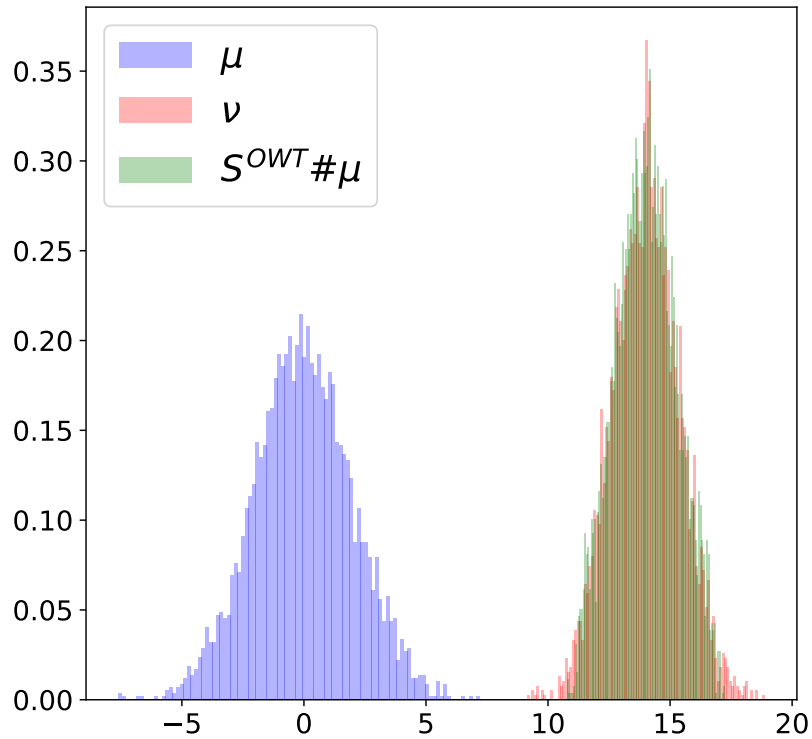
- The optimal plan π is unique for any distributions.
- Characterization via convex ordering [Gozlan and Juillet (2020)] and [Backhoff-Veraguas, Beiglböck, Pammer (2019)]:

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

  where $\eta \leqslant_c \nu$ stands for the convex ordering of measures: for any ϕ convex function, $\int \phi \, \mathrm{d}\eta \leqslant \int \phi \, \mathrm{d}\nu$.
- Two random variables with the same mean can be compared by how spread out their distributions are, this is captured by the convex ordering.

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.

Barycenter for optimal weak transport

Weak barycenter

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

Theorem

The weak barycenter problem admits solutions.

Barycenter for optimal weak transport

Weak barycenter

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

Theorem

The weak barycenter problem admits solutions.

For any distributions $\nu_1, \dots, \nu_k \in \mathcal{P}_2(\mathbb{R}^d)$

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

with $S_\mu^{\nu_i}$ optimal in the weak problem.

Interpretation as a latent variable model

Theorem

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

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

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

Robustness to outlier

For μ a weak barycenter which is not a Dirac measure, and $X \sim \mu$:

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

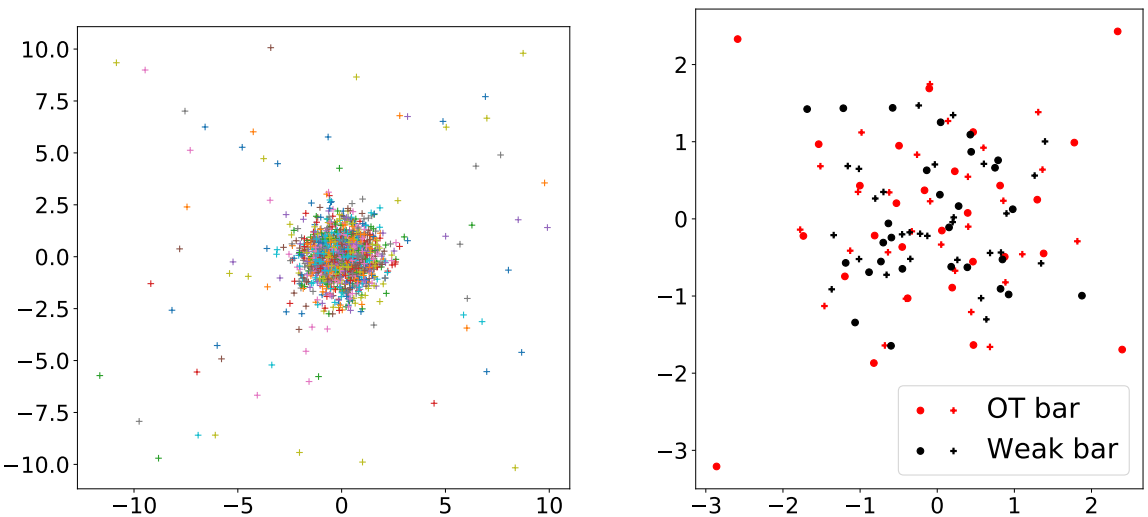


Figure: Empirical Gaussian distributions and their OWT (black) and OT (red) barycenters for Gaussian observations (crosses) and corrupted observations (dots).

Fixed-point approach for the weak barycenter problem

Iterative procedure

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

with for each $i = 1, \dots, k$, $S_\mu^{\nu_i} = \int y \mathrm{d}\pi_x^{\mu,\nu_i}(y)$, with $\pi^{\mu,\nu_i} \in \Pi(\mu, \nu_i)$ achieving the minimum for the optimal weak problem.

Fixed-point approach for the weak barycenter problem

Iterative procedure

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

with for each $i = 1, \dots, k$, $S_\mu^{\nu_i} = \int y \mathrm{d}\pi_x^{\mu,\nu_i}(y)$, with $\pi^{\mu,\nu_i} \in \Pi(\mu, \nu_i)$ achieving the minimum for the optimal weak problem.

Proposition

If μ is a weak-barycenter then $G(\mu) = \mu$.

Proposition

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

Fixed-point approach for the weak barycenter problem

Iterative procedure

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

with for each $i = 1, \dots, k$, $S_\mu^{\nu_i} = \int y \mathrm{d}\pi_x^{\mu,\nu_i}(y)$, with $\pi^{\mu,\nu_i} \in \Pi(\mu, \nu_i)$ achieving the minimum for the optimal weak problem.

Remark: Assuming that μ is a fixed-point doesn't imply that μ is a weak barycenter.

In the classical Wassertein barycenter framework, for a.c. measures, with at least one measure with bounded density, this result is not straightforward either: μ barycenter implies μ fixed-point

- if $x = \sum_{i=1}^{k} \lambda_i T_\mu^{\nu_i}(x)$ for every $x \in \mathbb{R}^d$, not only μ-almost everywhere [Agueh, Carlier, 2011].
- by invoking more smoothness on the distributions $\nu_1, \dots, \nu_k$ [Zemel, Panaretos, 2019]. Additionally, they only conjecture that the fixed-point is unique under these smoothness conditions.

An algorithm for a stream of data

Let $\mathbb{Q}$ be a probability distribution supported on a set of measures living in $\mathcal{P}_2(\mathbb{R}^d)$, then

Weak population barycenter

$$\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min} \int_{\mathcal{P}_2(\mathbb{R}^d)} V(\mu|\nu) \mathrm{d}\mathbb{Q}(\nu)$$

An algorithm for a stream of data

Let $\mathbb{Q}$ be a probability distribution supported on a set of measures living in $\mathcal{P}_2(\mathbb{R}^d)$, then

Weak population barycenter

$$\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min} \int_{\mathcal{P}_2(\mathbb{R}^d)} V(\mu|\nu) \mathrm{d}\mathbb{Q}(\nu)$$

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 $\sum_{k=1}^{\infty} \gamma_k = \infty$

Theorem

The sequence $(\mu_k)_k$ is a.s. relatively compact in W_q for all $q < 2$ (in particular it is tight). Moreover, a limit point μ verifies $x = \int S_\mu^\nu(x) \mathrm{d}\mathbb{Q}(\nu), \mu(x)$-a.s.

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

Computation of S_μ^ν

Let $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$

Quadratic programming

$$\min_{\pi \in \mathbb{R}^{n \times m}} \left\{ \left\| x - \left(\frac{\pi \mathbf{y}^T}{\mathbf{a}} \right) \right\|_\mu^2, \ \pi \geqslant 0, \ \pi \mathbb{1} = \mu, \ \pi^T \mathbb{1} = \nu \right\}$$

Proximal algorithm

$$\min_{\pi \in \mathbb{R}^{n \times m}} \underbrace{\sum_{i=1}^n a_i \| x_i - \overbrace{\left(\frac{\pi \mathbf{y}^T}{\mathbf{a}} \right)}^{S_\mu^\nu}_i \|^2}_{f(\pi)} + \underbrace{1_{\Pi(\mu,\nu)}(\pi)}_{g(\pi)},$$

where is the indicator function of the set C is $1_C(\pi) = \begin{cases} \pi & \text{if } \pi \in C \\ \infty & \text{otherwise.} \end{cases}$

Then

$$\pi^{k+1} = \mathrm{prox}_{\gamma^k g}(\pi^k - \gamma^k \nabla f(\pi^k)),$$

and prox_g is the projection operator onto $\Pi(\mu, \nu)$.
$\rightarrow$ Accelerated version FISTA [Beck, Teboulle (2009)].

Comparison setting

Recall the online algorithm

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

where ${S_{\mu_k}^{\nu^k}} = \frac{\pi \mathbf{y}^T}{a}$ can be constructed from any transport plan π.

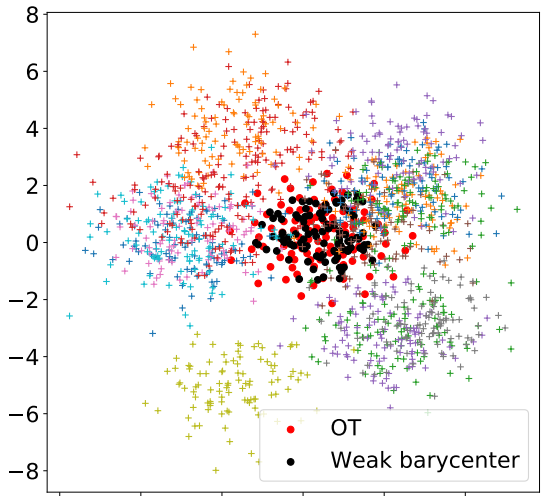
We consider the same iterative algorithm for π computed as:

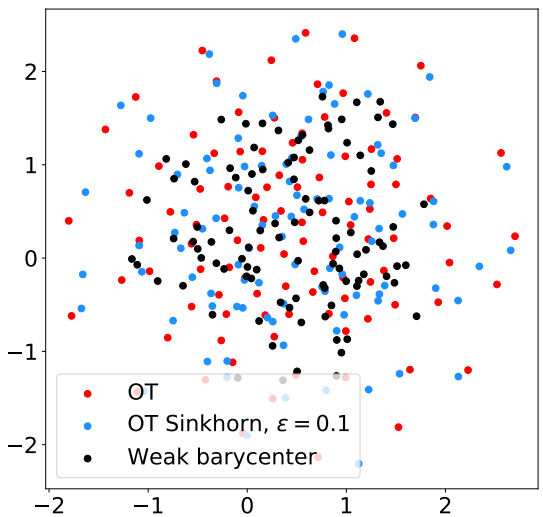
- an optimal transport plan of $W_2 \rightarrow$ "OT barycenter"
- an optimal transport plan of the entropy regularized OT problem (or Sinkhorn problem):

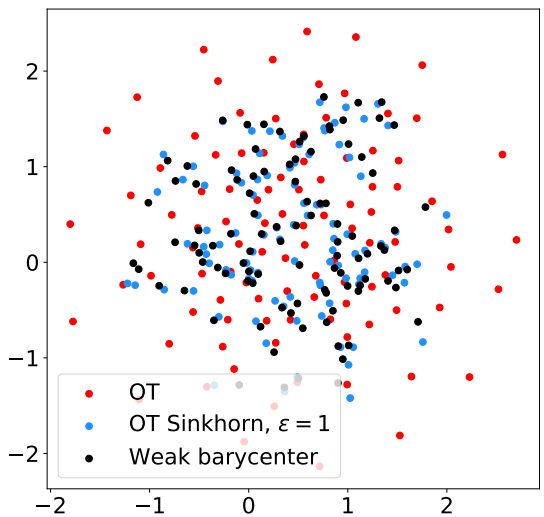
$$\underset{\pi \in \Pi(\mu,\nu)}{\arg\min} \int \|x-y\|^2 \mathrm{d}\pi(x,y) + \varepsilon KL(\pi|\mu \otimes \nu)$$

$\rightarrow$ "Sinkhorn barycenter"

Gaussian distributions







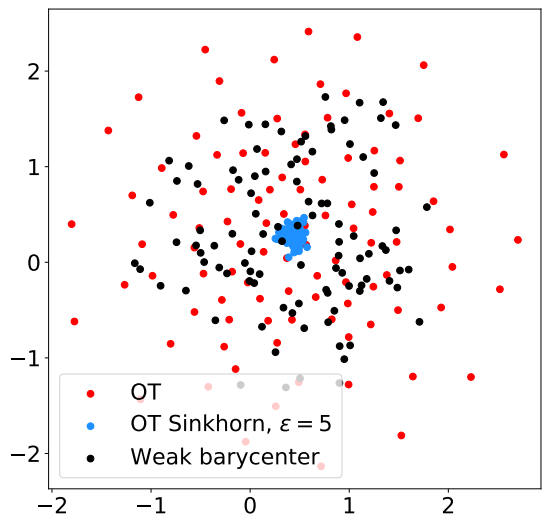


Figure: First row : $K = 15$ points clouds of $n = 100$ observations from Gaussian distributions each. Weak barycenter (black) and OT barycenter (red) computed from the streaming algorithm. Second row : illustration of the weak (black), OT (red) and OT Sinkhorn (blue) barycenters for different values of $\varepsilon = 0.1, 1$ and 5.

Stream of spiral distributions

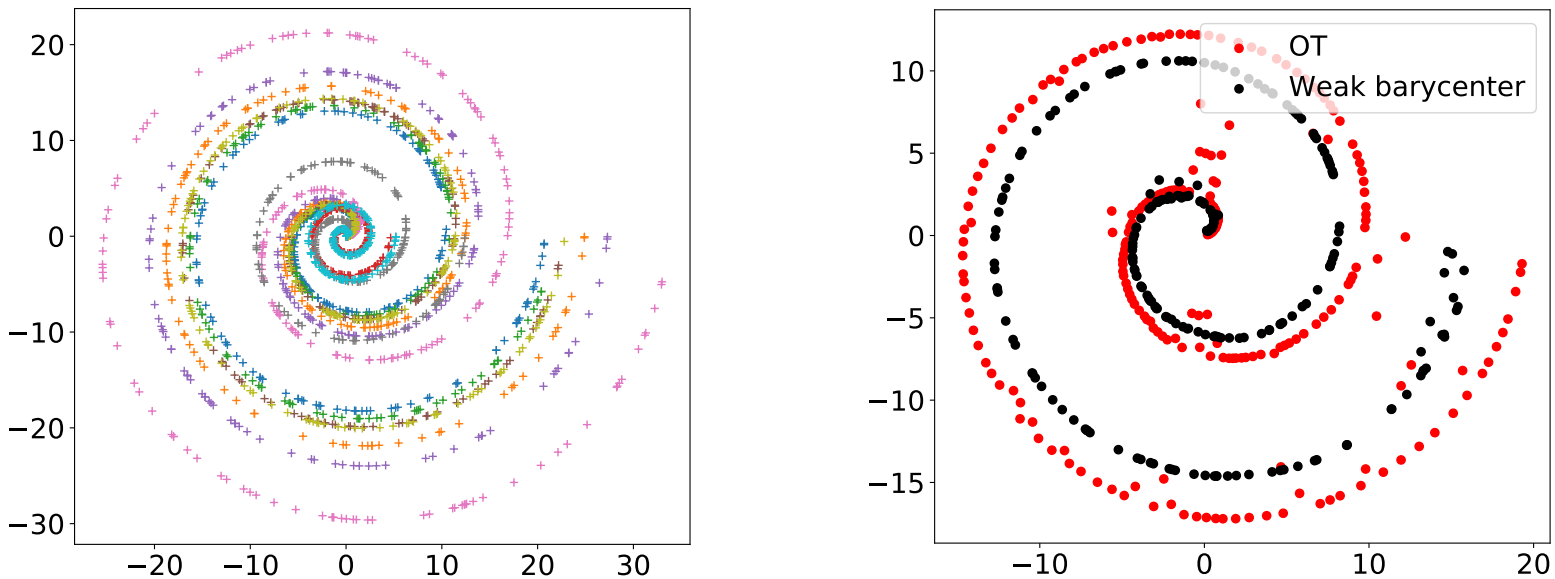


Figure: (left) $k = 10$ distributions supported on spiral, each distribution consists of p random points, with p randomly chosen in $(200, 225)$. (right) Weak (black) and OT (red) barycenters.

MNIST dataset

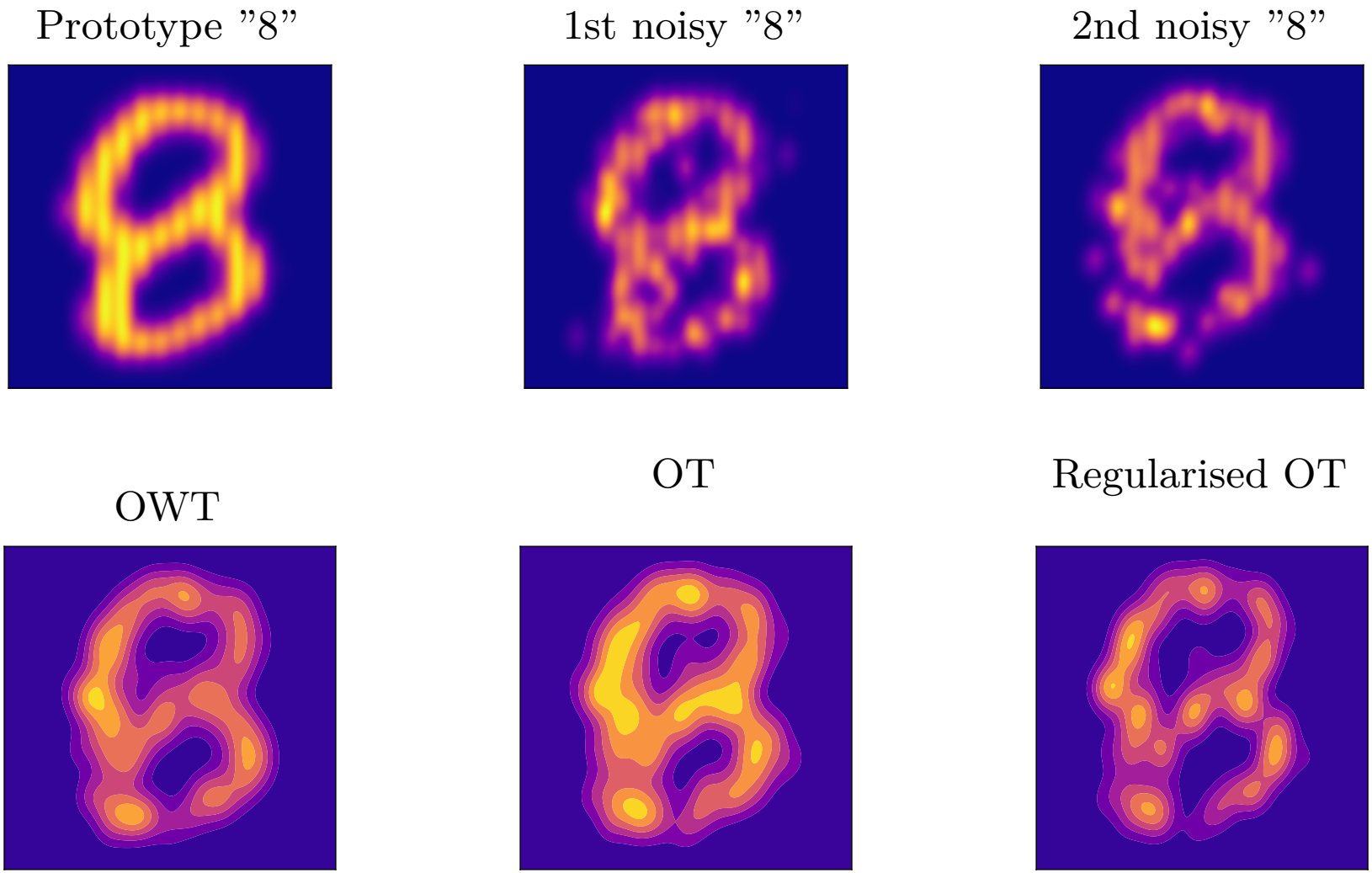


Figure: Digit "8" from MNIST dataset. First row. (left) Prototype "8". (middle & right) Noisy versions of the prototype by randomly (Bernoulli $p = 0.1$) moving pixels. Second row. Comparison of three barycenters : OWT plan (left), OT plan (middle) and entropy regularised OT plan for $\varepsilon = 1$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

Further 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.

Bibliography

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

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.

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.

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.

Elsa Cazelles, Felipe Tobar, Joaquin Fontbona. Streaming computation of optimal weak transport barycenters. https://arxiv.org/abs/2102.13380.