Regularization of barycenters in the Wasserstein space

Elsa Cazelles, joint work with Jérémie Bigot & Nicolas Papadakis
Université de Bordeaux & CNRS

CMO Workshop - 30 April to 5 May 2017
Optimal Transport meets Probability, Statistics and Machine Learning

For $\Omega \subset \mathbb{R}^d$, $\mathcal{P}_2(\Omega)$ is the set of probability measures.

Definition

Let $\mathbb{P}_n^\nu = \frac{1}{n}\sum_{i=1}^n \delta_{\nu_i}$ where δ_{ν_i} is the dirac distribution at $\nu_i \in \mathcal{P}_2(\Omega)$. We define the regularized empirical barycenter of the discrete measure $\mathbb{P}_n^\nu$ as

$$\boldsymbol{\mu}_{\mathbb{P}_n^\nu}^\gamma = \underset{\mu \in \mathcal{P}_2(\Omega)}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^n W_2^2(\mu, \nu_i) + \gamma E(\mu)$$

where $\gamma > 0$ is a regularisation parameter and the penalty E is a proper, differentiable, lower semicontinuous and strictly convex function.

Case $\gamma = 0$: Wasserstein barycenter of [Agueh and Carlier].

As an example, take E the negative entropy defined as

$$E(\mu) = \begin{cases} \int_\Omega f(x) \log(f(x)) dx, & \text{if } \mu \text{ admits a pdf } f \\ +\infty & \text{otherwise.} \end{cases}$$

Advantage: It is possible to enforce the regularized barycenter to be absolutely continuous with respect to the Lebesgue measure on Ω.

We define the population Wasserstein barycenter defined as

$$\mu^0_{\mathbb{P}} \in \underset{\mu \in \mathcal{P}_2(\Omega)}{\operatorname{argmin}} \int W_2^2(\mu, \nu) d\mathbb{P}(\nu),$$

and its regularized version

$$\mu^{\gamma}_{\mathbb{P}} = \underset{\mu \in \mathcal{P}_2(\Omega)}{\operatorname{argmin}} \int W_2^2(\mu, \nu) d\mathbb{P}(\nu) + \gamma E(\mu).$$

where $\mathbb{P}$ is a probability measure on $\mathcal{P}_2(\Omega)$ and $\nu_1, \ldots, \nu_n$ iid of law $\mathbb{P}$. We recall that

$$\boldsymbol{\mu}^{\gamma}_{\mathbb{P}^{\nu}_n} = \underset{\mu \in \mathcal{P}_2(\Omega)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \nu_i) + \gamma E(\mu)$$

The *Bregman divergence* D_E associated to E is defined for two measures μ, ζ as

$$D_E(\mu, \zeta) := E(\mu) - E(\zeta) - \int_{\Omega} \nabla E(\zeta)(d\mu - d\zeta)$$

where ∇E denotes the gradient of E.
Thus the *symmetric Bregman divergence* d_E is given by

$$d_E(\mu, \zeta) := D_E(\mu, \zeta) + D_E(\zeta, \mu).$$

Theorem

For Ω compact in $\mathbb{R}^d$ and $\nabla E(\mu_{\mathbb{P}}^0)$ bounded,

$$\lim_{\gamma \to 0} D_E(\mu_{\mathbb{P}}^{\gamma}, \mu_{\mathbb{P}}^0) = 0,$$

which corresponds to showing that the squared bias term $d_E^2(\mu_{\mathbb{P}}^{\gamma}, \mu_{\mathbb{P}}^0)$ (as classically referred to in nonparametric statistics) converges to zero when $\gamma \to 0$.

Theorem

If Ω is a compact of $\mathbb{R}^d$, then one has that

$$\mathbb{E}(d_E^2(\boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n}, \mu^{\gamma}_{\mathbb{P}})) \leq \frac{C\ I(1,\mathcal{H})\|H\|_{\mathbb{L}_2(\mathbb{P})}}{\gamma^2 n}$$

where C is a positive constant,

$$\mathcal{H} = \{h_\mu : \nu \in \mathcal{P}_2(\Omega) \mapsto W_2^2(\mu,\nu) \in \mathbb{R}; \mu \in \mathcal{P}_2(\Omega)\}$$

is a class of functions defined on $\mathcal{P}_2(\Omega)$ with envelope H, and

$$I(1,\mathcal{H}) = \sup_Q \int_0^1 (1 + \underbrace{\log N(\varepsilon\|H\|_{\mathbb{L}_2(Q)}, \mathcal{H}, \|\cdot\|_{\mathbb{L}_2(Q)})}_{\text{Metric entropy}})^{\frac{1}{2}} d\varepsilon$$

The metric entropy is of order $\frac{1}{\varepsilon^d}$.

Remarks on the metric entropy.

$$\text{Metric entropy of } \mathcal{H}$$
$$\downarrow$$
$$\text{Metric entropy of } \mathcal{P}_2(\Omega)$$
$$\downarrow$$
$$\text{Metric entropy of } \Omega$$
$$\downarrow$$
$$\text{Compacity of } \Omega$$

Theorem (1-D)

When $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ are iid random measures with support included in a compact interval Ω,

$$\mathbb{E}\left(d_E^2 \left(\boldsymbol{\mu}_{\mathbb{P}_n}^{\gamma}, \mu_{\mathbb{P}}^{\gamma} \right) \right) \leq \frac{C}{\gamma^2 n}.$$

where $C > 0$ does not depend on n and γ.

Remark: Metric entropy of the space of quantiles.

It follows that if $\gamma = \gamma_n$ is such that $\lim_{n\to\infty} \gamma_n^2 n = +\infty$ then

$$\lim_{n\to\infty} \mathbb{E}(d_E^2 \left(\boldsymbol{\mu}_{\mathbb{P}_n^{\boldsymbol{\nu}}}^{\gamma}, \mu_{\mathbb{P}}^{0}) \right) = 0.$$

When $d > 1$, the class of functions $\mathcal{H}$ is too large, so we have to add regularity.
For Ω smooth and uniformly convex, and $(\nu_i)_{i=1,\dots,n}$ of law $\mathbb{P}$. We specify the penalty function E:

$$E(\mu) = \begin{cases} \int_{\mathbb{R}^d} f(x)\log(f(x))dx + \|f\|_{H^k(\Omega)}, & \text{if } f = \frac{d\mu}{d\lambda} \text{ and } f > \alpha \\ +\infty & \text{otherwise.} \end{cases}$$

where $\|\cdot\|_{H^k(\Omega)}$ designates the Sobolev norm associated to the $\mathbb{L}^2(\Omega)$ space and $\alpha > 0$ is arbitrarily small.

- Sobolev embedding theorem: $H^k(\Omega)$ is included in the Hölder space $C^{m,\beta}$ for $\beta = k - m - d/2$.
- Regularity on optimal maps is obtained from regularity on probability measures (e.g. [De Philippis and Figalli])

Hence we can bound the metric entropy [Van der Vaart] by

$$K\left(\frac{1}{\varepsilon}\right)^a \text{ for any } a \geq d/(m+1).$$

Hence, as soon as $a/2 < 1$, for which $k > d-1$ is necessary, we get a rate of convergence.

Stability of the minimizer for the symmetric Bregman distance d_E.

Theorem

Let $\nu_1, \ldots, \nu_n$ and $\eta_1, \ldots, \eta_n$ be two sequences of probability measures in $\mathcal{P}_2(\Omega)$. Let $\mu^\gamma_{\mathbb{P}^\nu_n}$ and $\mu^\gamma_{\mathbb{P}^\eta_n}$ be the regularized empirical barycenters associated to the discrete measures $\mathbb{P}^\nu_n$ and $\mathbb{P}^\eta_n$, then

$$d_E\left(\mu^\gamma_{\mathbb{P}^\nu_n}, \mu^\gamma_{\mathbb{P}^\eta_n}\right) \leq \frac{2}{\gamma n} \inf_{\sigma \in \mathcal{S}_n} \sum_{i=1}^{n} W_2(\nu_i, \eta_{\sigma(i)}),$$

where $\mathcal{S}_n$ denotes the permutation group of the set of indices $\{1, \ldots, n\}$.

Application: Let $\nu_1, \ldots, \nu_n$ be n absolutely continuous probability measures and $\boldsymbol{X} = (\boldsymbol{X}_{i,j})_{1 \leq i \leq n;\ 1 \leq j \leq p_i}$ a dataset of random variables such that $\boldsymbol{X}_{i,j} \sim \nu_i$. Then

$$\mathbb{E}\left(d_E^2\left(\mu^\gamma_{\mathbb{P}_n^\nu}, \boldsymbol{\mu}^\gamma_{\boldsymbol{X}}\right)\right) \leq \frac{4}{\gamma^2 n} \sum_{i=1}^{n} \mathbb{E}\left(W_2^2(\nu_i, \boldsymbol{\nu}_{p_i})\right),$$

where $\boldsymbol{\mu}^\gamma_{\boldsymbol{X}}$ is the random measure satisfying

$$\boldsymbol{\mu}^\gamma_{\boldsymbol{X}} = \underset{\mu \in \mathcal{P}_2(\Omega)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} W_2^2\left(\mu, \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{i,j}}\right) + \gamma E(\mu).$$

We consider $1 \leq i \leq n = 100$ random Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$ where

- $\boldsymbol{\mu}_i$ random uniform variable on $[-2, 2]$
- $\boldsymbol{\sigma}_i^2$ random uniform variable on $[0, 1]$.

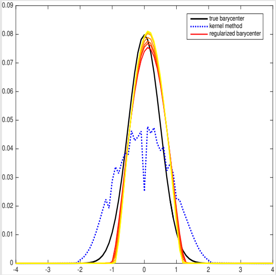
Then we generate $(\boldsymbol{X}_{ij})_{1 \leq i \leq n; 1 \leq j \leq p_i}$, $5 \leq p_i \leq 10$, random variables such that

$$\boldsymbol{X}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2) \text{ for each } 1 \leq i \leq j.$$

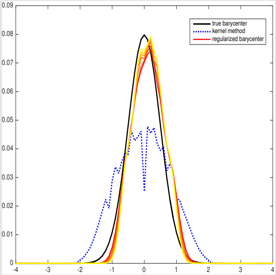
Finally, let

$$\boldsymbol{\nu}_i = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{ij}} \text{ for each } i$$

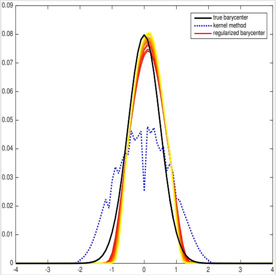
.



Dirichlet regularization



Entropy regularization



Dirichlet + Entropy regularization

- Dashed and black curve = density of the population Wasserstein barycenter.
- Blue and dotted curve = the smoothed Wasserstein barycenter obtained by a preliminary kernel smoothing step of the discrete measures $\boldsymbol{\nu}_i$ that is followed by quantile averaging.
- Warn color = regularized Wasserstein barycenters for several *gamma* and regularization.

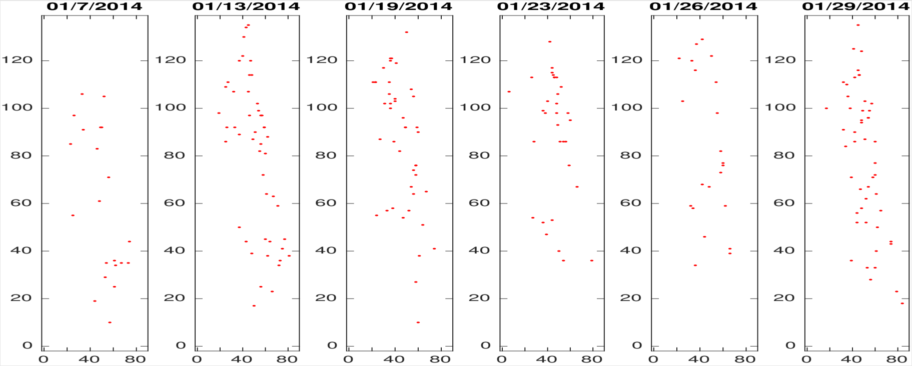
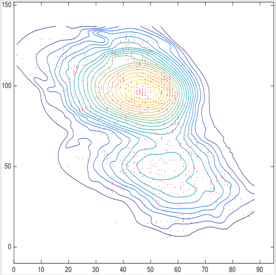
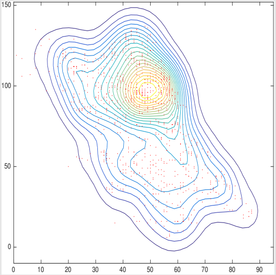


Figure: All crimes registered in the city of Chicago (i.e. an image 137×88) for 6 days of January 2014.

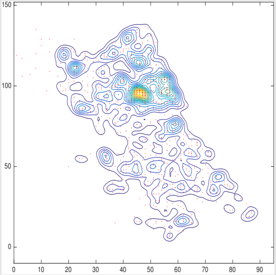
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data from the Chicago Police Department's CLEAR.



Regularized Wasserstein barycenter (Dirichlet)



Kernel density estimator gkde2

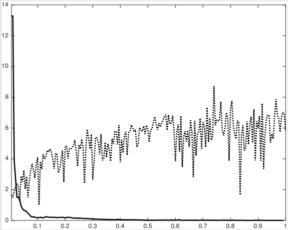


Kernel density estimator kde2d

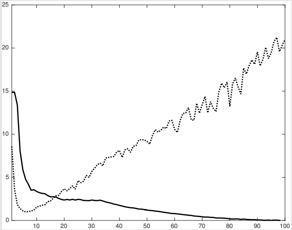
Figure: Location of Crimes in the city of Chicago during the month of January 2014

[1] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904-924, 2011.
[2] Martin Burger, Marzena Franek and Carola-Bibiane Schönlieb. Regularized regression and density estimation based on optimal transport. *Applied Mathematics Research eXpress*, 2012(2):209-253, 2012.
[3] Thibault Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein barycenters. *arXiv preprint arXiv:1506.04153*, 2015.
[4] Amir Beck and Marc Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1): 183-202, 2009.
[5] Guido De Philippis and Alessio Figalli. The Monge–Ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51(4), 527-580, 2014.
[6] Frank Bauer and Axel Munk. Optimal regularization for ill-posed problems in metric spaces. *Journal of Inverse and Ill-posed Problems jiip*, 15(2), 137-148, 2007

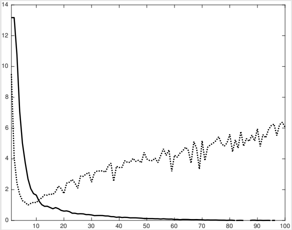
Automatic selection of the parameter γ through an adaptation of Lepskii balancing principal [Bauer and Munk].



Dirichlet



Entropy



Dirichlet + Entropy

Figure: Balancing functional (times 10^{-6}) in solid line and $d_E(\mu_{\mathbb{P}}, \mu_{\mathbb{P}_n}^{1/\lambda}) / \min_\lambda d_E(\mu_{\mathbb{P}}, \mu_{\mathbb{P}_n}^{1/\lambda})$ (dotted line) are plotted as functions of λ for 3 different regularizations.

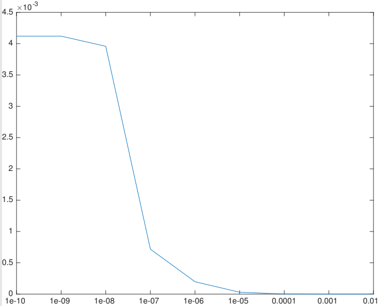


Figure: Smooth balancing functional associated to regularized barycenters for different value of λ.