Regularized Barycenters in the Wasserstein Space

Elsa Cazelles, Jérémie Bigot, Nicolas Papadakis

Regularized Barycenters in the Wasserstein Space

Elsa Cazelles, Jérémie Bigot, and Nicolas Papadakis

Université de Bordeaux, CNRS, Institut de Mathématiques de Bordeaux, UMR 5251,
elsa.cazelles@u-bordeaux.fr

Abstract. This paper is an overview of results that have been obtain in [2] on the convex regularization of Wasserstein barycenters for random measures supported on $\mathbb{R}^d$. We discuss the existence and uniqueness of such barycenters for a large class of regularizing functions. A stability result of regularized barycenters in terms of Bregman distance associated to the convex regularization term is also given. Additionally we discuss the convergence of the regularized empirical barycenter of a set of n iid random probability measures towards its population counterpart in the real line case, and we discuss its rate of convergence. This approach is shown to be appropriate for the statistical analysis of discrete or absolutely continuous random measures. In this setting, we propose an efficient minimization algorithm based on accelerated gradient descent for the computation of regularized Wasserstein barycenters.

Keywords: Wasserstein space, Fréchet mean, Barycenter of probability measures, Convex regularization, Bregman divergence

1 Introduction

This paper is concerned by the statistical analysis of data sets whose elements may be modeled as random probability measures supported on $\mathbb{R}^d$. It is an overview of results that have been obtain in [2]. In the special case of one dimension ($d = 1$), we are able to provide refined results on the study of a sequence of discrete measures or probability density functions (e.g. histograms) that can be viewed as random probability measures. Such data sets appear in various research fields. Examples can be found in neuroscience [10], biodemographic and genomics studies [11], economics [7], as well as in biomedical imaging [9]. In this paper, we focus on first-order statistics methods for the purpose of estimating, from such data, a population mean measure or density function.

The notion of averaging depends on the metric that is chosen to compare elements in a given data set. In this work, we consider the Wasserstein distance W_2 associated to the quadratic cost for the comparison of probability measures. Let Ω be a subset of $\mathbb{R}^d$ and $\mathcal{P}_2(\Omega)$ be the set of probability measures supported on Ω with finite order second moment.

Definition 1. *As introduced in [1], an empirical Wasserstein barycenter $\bar{\nu}_n$ of a set of n probability measures $\nu_1, \ldots, \nu_n$ (not necessarily random) in $\mathcal{P}_2(\Omega)$ is*

defined as a minimizer of

$$\mu \mapsto \frac{1}{n}\sum_{i=1}^{n} W_2^2(\mu,\nu_i), \ \ over\ \mu \in \mathcal{P}_2(\Omega). \tag{1}$$

The Wasserstein barycenter corresponds to the notion of empirical Fréchet mean [6] that is an extension of the usual Euclidean barycenter to nonlinear metric spaces.

However, depending on the data at hand, such a barycenter may be irregular. As an example let us consider a real data set of neural spike trains which is publicly available from the MBI website[1]. During a squared-path task, the spiking activity of a movement-encoded neuron of a monkey has been recorded during 5 seconds over $n = 60$ repeated trials. Each spike train is then smoothed using a Gaussian kernel (further details on the data collection can be found in [10]). For each trial $1 \leq i \leq n$, we let ν_i be the measure with probability density function (pdf) proportional to the sum of these Gaussian kernels centered at the times of spikes. The resulting data are displayed in Fig. 1(a). For probability measures supported on the real line, computing a Wasserstein barycenter simply amounts to averaging the quantile functions of the ν_i's (see e.g. Section 6.1 in [1]). The pdf of the Wasserstein barycenter $\bar{\nu}_n$ is displayed in Fig. 1(b). This approach clearly leads to the estimation of a very irregular mean template density of spiking activity.

In this paper, we thus introduce a convex regularization of the optimization problem (1) for the purpose of obtaining a regularized Wasserstein barycenter. In this way, by choosing an appropriate regularizing function (e.g. the negative entropy in Subsection 2.1), it is of possible to enforce this barycenter to be absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$.

2 Regularization of barycenters

We choose to add a penalty directly into the computation of the Wasserstein barycenter in order to smooth the Fréchet mean and to remove the influence of noise in the data.

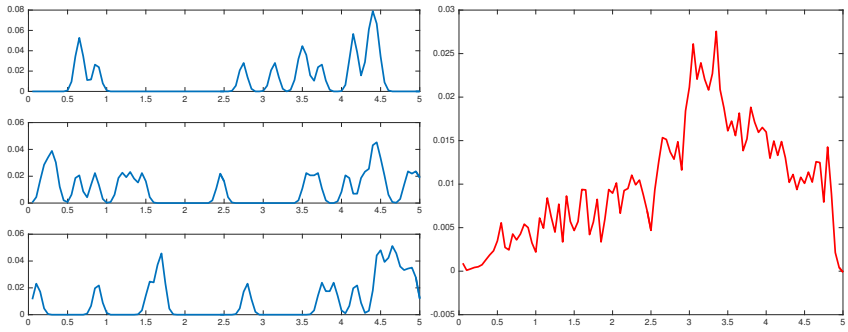
Definition 2. *Let $\mathbb{P}_n^\nu = \frac{1}{n}\sum_{i=1}^n \delta_{\nu_i}$ where δ_{ν_i} is the dirac distribution at ν_i. We define a regularized empirical barycenter $\mu^\gamma_{\mathbb{P}_n^\nu}$ of the discrete measure $\mathbb{P}_n^\nu$ as a minimizer of*

$$\mu \mapsto \frac{1}{n}\sum_{i=1}^{n} W_2^2(\mu,\nu_i) + \gamma E(\mu) \ over\ \mu \in \mathcal{P}_2(\Omega), \tag{2}$$

where $\mathcal{P}_2(\Omega)$ is the space of probability measures on Ω with finite second order moment, $E : \mathcal{P}_2(\Omega) \to \mathbb{R}_+$ is a smooth convex penalty function, and $\gamma > 0$ is a regularization parameter.

In what follows, we present the main properties on the regularized empirical Wasserstein barycenter $\mu^\gamma_{\mathbb{P}_n^\nu}$.

[1] http://mbi.osu.edu/2012/stwdescription.html



(a) (b)

Fig. 1. (a) A subset of 3 smoothed neural spike trains out of $n = 60$. Each row represents one trial and the pdf obtained by smoothing each spike train with a Gaussian kernel of width 50 milliseconds. (b) Probability density function of the empirical Wasserstein barycenter $\bar{\nu}_n$ for this data set.

2.1 Existence and uniqueness

We consider the wider problem of

$$\min_{\mu \in \mathcal{P}_2(\Omega)} J^{\gamma}_{\mathbb{P}}(\mu) = \int W_2^2(\mu, \nu) d\mathbb{P}(\nu) + \gamma E(\mu). \tag{3}$$

Hence, (2) corresponds to the minimization problem (3) where $\mathbb{P}$ is discrete ie $\mathbb{P} = \mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\nu_i}$.

Theorem 1 (Theorem 3.2 in [2]) *Let $E : \mathcal{P}_2(\Omega) \to \mathbb{R}_+$ be a proper, lower semicontinuous and differentiable function that is strictly convex on its domain $\mathcal{D}(E) = \{\mu \in \mathcal{P}_2(\Omega)$ such that $E(\mu) < +\infty\}$. Then, the functional $J^{\gamma}_{\mathbb{P}}$ define by (3) admits a unique minimizer.*

Such assumptions on E are supposed to be always satisfied throughout the paper. A typical example of regularization function satisfying such assumptions is the negative entropy defined as

$$E(\mu) = \begin{cases} \int_{\mathbb{R}^d} f(x)\log(f(x))dx, & \text{if } \mu \text{ admits a density } f \text{ with respect to} \\ & \text{the Lebesgue measure } dx \text{ on } \Omega, \\ +\infty & \text{otherwise.} \end{cases}$$

2.2 Stability

We study the stability of the minimizer of (3) with respect to the discrete distribution $\mathbb{P}_n^{\nu} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\nu_i}$ on $\mathcal{P}_2(\Omega)$. This result is obtained for the symmetric Bregman distance $d_E(\mu, \zeta)$ between two measures μ and ζ. Bregman distances associated to a convex penalty E are known to be appropriate error measures for

various regularization methods in inverse problems (see e.g. [4]). This Bregman distance between two probability measures μ and ζ is defined as

$$d_E(\mu,\zeta) := \langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta \rangle = \int_\Omega (\nabla E(\mu)(x) - \nabla E(\zeta)(x))(d\mu - d\zeta)(x),$$

where $\nabla E : \Omega \to \mathbb{R}$ denotes the gradient of E. In the setting where E is the negative entropy and $\mu = \mu_f$ (resp. $\zeta = \zeta_g$) admits a density f (resp. g) with respect to the Lebesgue measure, then d_E is the symmetrised Kullback-Leibler divergence

$$d_E(\mu_f, \zeta_g) = \int (f(x) - g(x)) \log\left(\frac{f(x)}{g(x)}\right) dx.$$

The stability result of the regularized empirical barycenter can then be stated as follows.

Theorem 2 (Theorem 3.3 in [2]) *Let $\nu_1, \ldots, \nu_n$ and $\eta_1, \ldots, \eta_n$ be two sequences of probability measures in $\mathcal{P}_2(\Omega)$. If we denote by $\mu^\gamma_{\mathbb{P}^\nu_n}$ and $\mu^\gamma_{\mathbb{P}^\eta_n}$ the regularized empirical barycenter associated to the discrete measures $\mathbb{P}^\nu_n$ and $\mathbb{P}^\eta_n$, then the symmetric Bregman distance (associated to E) between these two barycenters is bounded as follows*

$$d_E\left(\mu^\gamma_{\mathbb{P}^\nu_n}, \mu^\gamma_{\mathbb{P}^\eta_n}\right) \leq \frac{2}{\gamma n} \inf_{\sigma \in \mathcal{S}_n} \sum_{i=1}^n W_2(\nu_i, \eta_{\sigma(i)}), \tag{4}$$

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

In particular, inequality (4) allows to compare the case of data made of n absolutely continuous probability measures $\nu_1, \ldots, \nu_n$, with the more realistic setting where we have only access to a dataset of random variables $\boldsymbol{X} = (\boldsymbol{X}_{i,j})_{1 \leq i \leq n;\, 1 \leq j \leq p_i}$ organized in the form of n experimental units, such that $\boldsymbol{X}_{i,1}, \ldots, \boldsymbol{X}_{i,p_i}$ are iid observations in $\mathbb{R}^d$ sampled from the measure ν_i for each $1 \leq i \leq n$. If we denote by $\boldsymbol{\nu}_{p_i} = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{i,j}}$ the usual empirical measure associated to ν_i, it follows from inequality (4) that

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

where $\boldsymbol{\mu}^\gamma_{\boldsymbol{X}}$ is given by $\boldsymbol{\mu}^\gamma_{\boldsymbol{X}} = \operatorname*{argmin}_{\mu \in \mathcal{P}_2(\Omega)} \frac{1}{n} \sum_{i=1}^n W_2^2(\mu, \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{i,j}}) + \gamma E(\mu)$. This result allows to discuss the rate of convergence (for the symmetric squared Bregman distance) of $\boldsymbol{\mu}^\gamma_{\boldsymbol{X}}$ to $\mu^\gamma_{\mathbb{P}^\nu_n}$ as a function of the rate of convergence (for the squared Wasserstein distance) of the empirical measure $\boldsymbol{\nu}_{p_i}$ to ν_i for each $1 \leq i \leq n$ (in the asymptotic setting where $p = \min_{1 \leq i \leq n} p_i$ is let going to infinity). As an illustrative example, in the one-dimensional case (that is $d = 1$), one may use the work in [3] on a detailed study of the variety of rates of convergence of an empirical measure on the real line toward its population counterpart for the

expected squared Wasserstein distance. For example, by Theorem 5.1 in [3], it follows that

$$\mathbb{E}\left(W_2^2(\nu_i, \boldsymbol{\nu}_{p_i})\right) \leq \frac{2}{p_i+1} J_2(\nu_i), \text{ with } J_2(\nu_i) = \int_\Omega \frac{F_i(x)(1-F_i(x))}{f_i(x)} dx,$$

where f_i is the pdf of ν_i, and F_i denotes its cumulative distribution function. Therefore, provided that $J_2(\nu_i)$ is finite for each $1 \leq i \leq n$, one obtains the following rate of convergence of $\boldsymbol{\mu}_{\boldsymbol{X}}^\gamma$ to $\mu_{\mathbb{P}_n^\nu}^\gamma$ (in the case of measures ν_i supported on an interval Ω of $\mathbb{R}$)

$$\mathbb{E}\left(d_E^2\left(\mu_{\mathbb{P}_n^\nu}^\gamma, \boldsymbol{\mu}_{\boldsymbol{X}}^\gamma\right)\right) \leq \frac{8}{\gamma^2 n} \sum_{i=1}^n \frac{J_2(\nu_i)}{p_i+1} \leq \frac{8}{\gamma^2}\left(\frac{1}{n}\sum_{i=1}^n J_2(\nu_i)\right) p^{-1}. \tag{5}$$

Note that by the results in Appendix A in [3], a necessary condition for $J_2(\nu_i)$ to be finite is to assume that f_i is almost everywhere positive on the interval Ω.

2.3 Convergence to a population Wasserstein barycenter

Introducing this symmetric Bregman distance also allows to analyze the consistency of the regularized barycenter $\mu_{\mathbb{P}_n}^\gamma$ as the number of observations n tends to infinity and the parameter γ is let going to zero. When $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ are supposed to be independent and identically distributed (iid) random measures in $\mathcal{P}_2(\Omega)$ sampled from a distribution $\mathbb{P}$, we analyze the convergence of $\boldsymbol{\mu}_{\mathbb{P}_n^{\boldsymbol{\nu}}}^\gamma$ with respect to the population Wasserstein barycenter defined as

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

and its regularized version

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

In the case where Ω is a compact of $\mathbb{R}^d$ and $\nabla E(\mu_{\mathbb{P}}^0)$ is bounded, we prove that $\mu_{\mathbb{P}}^\gamma$ converges to $\mu_{\mathbb{P}}^0$ as $\gamma \to 0$ for the Bregman divergence associated to E. This result corresponds to showing that the bias term (as classically referred to in nonparametric statistics) converges to zero when $\gamma \to 0$. We also analyze the rate of convergence of the variance term when Ω is a compact of $\mathbb{R}$:

Theorem 3 (Theorem 4.5 in [2]) *For Ω compact included in $\mathbb{R}$, there exists a constant $C > 0$ (not depending on n and γ) such that*

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

Therefore, when $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ are iid random measures with support included in a compact interval Ω, it follows that if $\gamma = \gamma_n$ is such that $\lim_{n\to\infty} \gamma_n^2 n = +\infty$ then

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

3 Numerical experiments

We consider a simulated example where the measures $\boldsymbol{\nu}_i$ are discrete and supported on a small number p_i of data points ($5 \leq p_i \leq 10$). To this end, for each $i = 1, \ldots, n$, we simulate a sequence $(\boldsymbol{X}_{ij})_{1 \leq j \leq p_i}$ of iid random variables sampled from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$, and the $\boldsymbol{\mu}_i$'s (resp. $\boldsymbol{\sigma}_i$) are iid random variables such that $-2 \leq \boldsymbol{\mu}_i \leq 2$ and $0 \leq \boldsymbol{\sigma}_i \leq 1$ with $\mathbb{E}(\boldsymbol{\mu}_i) = 0$ and $\mathbb{E}(\boldsymbol{\sigma}_i) = 1/2$. The target measure that we wish to estimate in these simulations is the population (or true) Wasserstein barycenter of the random distribution $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$ which is $\mathcal{N}(0, 1/4)$ thanks to the assumptions $\mathbb{E}(\boldsymbol{\mu}_1) = 0$ and $\mathbb{E}(\boldsymbol{\sigma}_1) = 1/2$. Then, let $\boldsymbol{\nu}_i = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{ij}}$, where δ_x is the Dirac measure at x.

In order to compute the regularized barycenter, we solve (3) with an efficient minimization algorithm based on accelerated gradient descent (see [5]) for the computation of regularized barycenters in 1-D (see Appendix C in [2]).

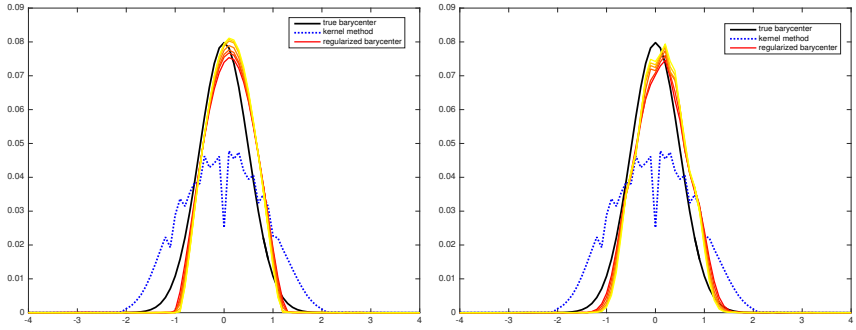
To illustrate the benefits of regularizing the Wasserstein barycenter of the $\boldsymbol{\nu}_i$'s, we compare our estimator with the one obtained by the following procedure which we refer to as the kernel method. In a preliminary step, each measure $\boldsymbol{\nu}_i$ is smoothed using a standard kernel density estimator to obtain

$$\hat{\boldsymbol{f}}_{i,h_i}(x) = \frac{1}{p_i h_i} \sum_{j=1}^{p_i} K\left(\frac{x - \boldsymbol{X}_{ij}}{h_i}\right), \; x \in \Omega,$$

where K is a Gaussian kernel. The bandwidth h_i is chosen by cross-validation. An alternative estimator is then defined as the Wasserstein barycenter of the smoothed measures with density $\hat{\boldsymbol{f}}_{1,h_1}, \ldots, \hat{\boldsymbol{f}}_{n,h_n}$. Thanks, to the well-know quantile averaging formula, the quantile function $\bar{F}_n^{-1}$ of this smoothed Wasserstein barycenter is given by $\bar{F}_n^{-1} = \frac{1}{n} \sum_{i=1}^{n} F^{-1}_{\hat{\boldsymbol{f}}_{i,h_i}}$ where F_g^{-1} denotes the quantile function of a given pdf g. The estimator $\bar{F}_n^{-1}$ corresponds to the notion of smoothed Wasserstein barycenter of multiple point processes as considered in [8]. The density of $\bar{F}_n^{-1}$ is denoted by $\hat{\boldsymbol{f}}_n$, and it is displayed in Fig. 2. Hence, it seems that a preliminary smoothing of the $\boldsymbol{\nu}_i$ followed quantile averaging is not sufficient to recover a satisfactory Gaussian shape when the number p_i of observations per unit is small.

Alternatively, we have applied our algorithm directly on the (non-smoothed) discrete measures $\boldsymbol{\nu}_i$ to obtain the regularized barycenter $\boldsymbol{f}^{\gamma}_{\mathbb{P}_n}$ defined as the minimizer of (2). For the penalty function E, we took either the negative entropy or a Dirichlet regularization. The densities of the penalized Wasserstein barycenters associated to these two choices for E and for different values of γ are displayed as solid curves in warm colors in Fig. 2. For both penalty functions and despite a small number of observations per experimental units, the shape of these densities better reflects the fact that the population Wasserstein barycenter is a Gaussian distribution.

Finally, we provide Monte-Carlo simulations to illustrate the influence of the number $n = 100$ of observed measures on the convergence of these estimators. For a given $10 \leq n_0 \leq n$, we randomly draw n_0 measures $\boldsymbol{\nu}_i$ from the whole

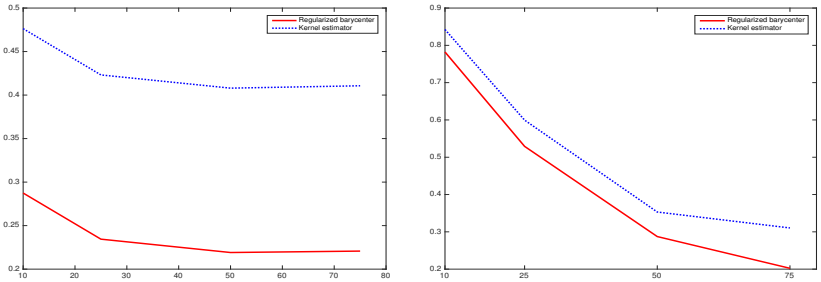


(a) Dirichlet regularization (b) Entropy regularization

Fig. 2. Simulated data from Gaussian distributions with random means and variances. In all the figures, the black curve is the density of the true Wasserstein barycenter. The blue and dotted curve represents the pdf of the smoothed Wasserstein barycenter obtained by a preliminary kernel smoothing step. Pdf of the regularized Wasserstein barycenter $\boldsymbol{\mu}_{\mathbb{P}_n}^{\gamma}$ (a) for $20 \leq \gamma \leq 50$ with $E(f) = ||f'||^2$ (Dirichlet), and (b) for $0.08 \leq \gamma \leq 14$ with $E(f) = \int f \log(f)$ (negative entropy)

.

sample, and we compute a smoothed barycenter via the kernel method and a regularized barycenter for a chosen γ. For given value of n_0, this procedure is repeated 200 times, which allows to obtain an approximation of the expected error $\mathbb{E}\left(d(\hat{\mu}, \mu_{\mathbb{P}})\right)$ of each estimator $\hat{\mu}$, where d is either d_E or W_2. The penalty used is a linear combinaison of Dirichlet and negative entropy functions. The results are displayed in Figure 3. It can be observed that our approach yields better results than the kernel method for both types of error (using either the Bregman or Wasserstein distance).



(a) Error with Bregman distance (b) Error with W_2 distance

Fig. 3. Errors in terms of expected Bregman and Wasserstein distances between the population barycenter and the estimated barycenters (kernel method in dashed blue, regularized barycenter in red) for a sample of size $n_0 = 10, 25, 50$ and 75.

4 Conclusion

In this paper, we have summarize some of the results of [2]. We provide a study on regularized barycenters in the Wasserstein space, which is of interest when the data are irregular or for noisy probability measures. Future works will concern the numerical computation and the study of the convergence to a population Wasserstein barycenter for $\Omega \subset \mathbb{R}^d$.

Acknowledgment

This work has been carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of the GOTMI project (ANR-16-CE33-0010-01).

References

1. M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.
2. J. Bigot, E. Cazelles, and N. Papadakis. Penalized barycenters in the Wasserstein space. *Submitted. Available at https://128.84.21.199/abs/1606.01025.*
3. S. Bobkov and M. Ledoux. *One-dimensional empirical measures, order statistics and Kantorovich transport distances.* 2014. Book in preparation. Available at http://perso.math.univ-toulouse.fr/ledoux/files/2013/11/Order.statistics.10.pdf.
4. M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse problems*, 20(5):1411, 2004.
5. Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.
6. M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H.Poincaré, Sect. B, Prob. et Stat.*, 10:235–310, 1948.
7. A. Kneip and K. J. Utikal. Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc.*, 96(454):519–542, 2001. With comments and a rejoinder by the authors.
8. V. M. Panaretos and Y. Zemel. Amplitude and phase variation of point processes. *Annals of Statistics*, 44(2):771–812, 2016.
9. K. Petersen and H.-G. Müller. Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics*, To be published, 2015.
10. W. Wu and A. Srivastava. An information-geometric framework for statistical inferences in the neural spike train space. *Journal of Computational Neuroscience*, 31(3):725–748, November 2011.
11. Z. Zhang and H.-G. Müller. Functional density synchronization. *Computational Statistics & Data Analysis*, 55(7):2234–2249, 2011.