A few words on Wasserstein barycenters

Xavier Bacon

15/02/2024, SAMM's working group

Bibliography. — I based these notes on Agueh and Carlier seminal article [AC11] for the section on barycenters. The section about quadratic optimal transport and the one dimensional case comes from the two monographies of Villani [Vil21] and Santambrogio [San15]. To go further, especially concerning the use of Wasserstein barycenters in data science, we may look at the book [PC+19] of Cuturi and Peyré whose a part of it is devoted to this subject.

1 Definition and computation

Barycenter in normed vector spaces. — Let $N \in \mathbb{N}^*$ be an integer greater than 1. In a real normed vector space (say) $\mathcal{E}$, a **barycenter** (or weighted average) of a finite family of $\mathcal{E}$ $(y_i)_{i=1}^N$ is defined as the unique vector $x^* \in \mathcal{E}$ satisfying the equation

$$\sum_{i=1}^{N} \lambda_i (x^* - y_i) = 0$$

where $\lambda_1, \ldots, \lambda_N \geqslant 0$ are for some weights that sum to 1. It is clear that the equation above is equivalent to

$$x^* = \operatorname*{argmin}_{x \in \mathcal{E}} \sum_{i=1}^{N} \frac{\lambda_i}{2} \|x - y_i\|^2,$$

which is a more convenient equation once we want to extend this notion to non vector space such as $\mathcal{P}(X)$ the space of probability over a given set X.

Barycenters in Wasserstein spaces. — In what follows, let us fix a non-zero integer $d \in \mathbb{N}^*$. X will denote a subset of $\mathbb{R}^d$, typically $\mathbb{R}^d$ itself or sometimes for simplicity (especially when we will dealing with duality) a compact of $\mathbb{R}^d$. Finally notice that most of the following development can be extended to any metric space.

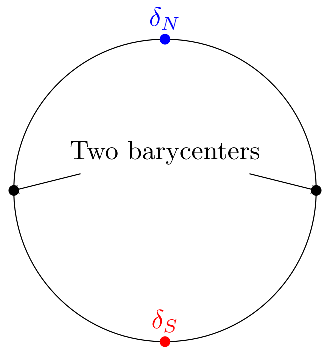
Let $\nu_1, \ldots, \nu_N \in \mathcal{P}(X)$ be N probability measures on X and $\lambda_1, \ldots, \lambda_N \in \mathbb{R}_+$ be N positive real numbers such that

$$\sum_{i=1}^{N} \lambda_i = 1.$$

A **Wasserstein barycenter** of the family $\nu = (\nu_i)_{i=1}^N$ associated to the weights $\lambda = (\lambda_i)_{i=1}^N$ and denoted by $\mathrm{bar}_\lambda(\nu)$ is defined as **a** solution of the following minimization problem

$$\mathrm{bar}_\lambda(\nu) \in \operatorname{argmin} \left\{ \sum_{i=1}^{N} \frac{\lambda_i}{2} W_2^2(\mu, \nu_i) : \mu \in \mathcal{P}(X) \right\}$$

Remark : non uniqueness. — Contrary to the normed case, the uniqueness of such a barycenter is not always satisfied (see the figure below).



A counterexample to uniqueness : $X = \mathbb{S}^1$ endowed with the angular distance.

However, uniqueness is satisfied once one of the ν_i admits a density with respect to the Lebesgue measure. For a proof of a more general result, see the seminal article of Wasserstein barycenters [AC11]. **From now on, we will assume that at least one ν_i admits a density with respect to the Lebesgue measure and so ensure the uniqueness of $\mathrm{bar}_\lambda(\nu)$.**

Multi-marginal transportation problem. — In order to calculate $\mathrm{bar}_\lambda(\nu)$, we introduce an auxiliary problem. For this purpose, for every $x = (x_1, \dots, x_N) \in \mathbb{R}^N$, we denote by $B(x)$ the euclidean barycenter of $x_1, \dots, x_N$ that is

$$B(x) = \sum_{i=1}^{N} \lambda_i x_i$$

and introduce the auxiliary problem

$$\inf \left\{ \int_{X^d} \sum_{i=1}^{N} \frac{\lambda_i}{2} |x_i - B(x)|^2 \, \mathrm{d}\gamma(x_1, \dots, x_N) : \gamma \in \Pi(\nu_1, \dots, \nu_N) \right\}, \tag{$\mathcal{Q}$}$$

where $\Pi(\nu_1, \dots, \nu_N)$ denotes the set of probability measures on X^N having ν_i as marginals. The problem above is called **the multi-marginal transportation problem** and its interest lies in the next crucial result:

Link between $\mathrm{bar}_\lambda(\nu)$ and ($\mathcal{Q}$). — Let γ be a solution a solution of ($\mathcal{Q}$), then

$$\mathrm{bar}_\lambda(\nu) = B\#\gamma.$$

Proof. See Proposition 4.2. in [AC11]. The essential tool in this proof is the concept of measure's disintegration.

2 A case study for $N = 2$

2.1 A few words on quadratic optimal transport

From now on, we fix the number of marginals N equals to 2, the dimension d to 1 and for calculus simplicity, we assume moreover that $\lambda_1 = \lambda_2 = 1$. Then, $(\mathcal{Q})$ becomes

$$\inf \left\{ \iint_{\mathbb{R}\times\mathbb{R}} \frac{1}{2}|x_1 - B(x)|^2 + \frac{1}{2}|x_2 - B(x)|^2 \,\mathrm{d}\gamma(x) : \gamma \in \Pi(\nu_1, \nu_2) \right\}, \tag{1}$$

where B is given for every $x = (x_1, x_2) \in \mathbb{R}^2$ by

$$B(x) = \frac{x_1 + x_2}{2}.$$

Notice that this minimization problem is nothing less than the classical Kantorovitch optimal transport problem for the particular cost

$$c(x_1, x_2) = \frac{1}{2}|x_1 - M(x)|^2 + \frac{1}{2}|x_2 - M(x)|^2.$$

Primal and dual. — Developing the squares in (1) leads us easily to the equivalent maximization problem[1]

$$\sup \left\{ \iint_{\mathbb{R}\times\mathbb{R}} x_1 x_2 \,\mathrm{d}\gamma(x) : \gamma \in \Pi(\nu_1, \nu_2) \right\}. \tag{$\mathcal{P}$}$$

In order to study this new problem, the key tool is the so-called dual problem of $(\mathcal{Q})$, defined as the minimization problem

$$\inf \left\{ \int_{\mathbb{R}} \varphi_1(x_1) \,\mathrm{d}\nu_1(x_1) + \int_{\mathbb{R}} \varphi_(x_2) \,\mathrm{d}\nu_2(x_2) : \varphi_1(x_1) + \varphi_2(x_2) \geqslant x_1 x_2, \forall (x_1, x_2) \in \mathbb{R}^2 \right\}. \tag{$\mathcal{P}^*$}$$

where the infimum is taken over continuous functions φ_1, φ_2 vanishing at infinity[2].

Reminder : why duality? — The main interest of introducing the dual problem $(\mathcal{P}^*)$ is to obtain an equation relying any solution (say) γ of $(\mathcal{P})$ to any solution (say) (φ_1, φ_2) of $(\mathcal{P}^*)$. To establish such an equation, recall that according to the *no-duality gap* theorem, we have the equality

$$(\mathcal{P}) = (\mathcal{P}^*)$$

rewritten as, admitting the existence of a maximizer and a minimizer,

$$\iint_{\mathbb{R}\times\mathbb{R}} x_1 x_2 \,\mathrm{d}\gamma(x) = \int_{\mathbb{R}} \varphi_1(x_1) \,\mathrm{d}\nu_1(x_1) + \int_{\mathbb{R}} \varphi_(x_2) \,\mathrm{d}\nu_2(x_2)$$

where

$$\gamma \in \operatorname{argmax}(\mathcal{P}) \text{ and } (\varphi_1, \varphi_2) \in \operatorname{argmin}(\mathcal{P}^*).$$

[1] In the sense that any minimizer of the first one is a maximizer of the second and *vice versa*.

[2] If we are reduced to a compact subset of $\mathbb{R}$, the *vanishing at infinity assumption* disappears.

The following equation, satisfied for almost every (x_1, x_2) with respect to γ, follows

$$\varphi_1(x_1) + \varphi_2(x_2) = x_1 x_2. \qquad \text{(Optimality equation)}$$

If we assume moreover that $\varphi_1(x_1)$ and $\varphi_2(x_2)$ are differentiable, we obtain

$$\varphi_1^{'}(x_1) = x_2$$
$$\varphi_2^{'}(x_2) = x_1$$

for every (x_1, x_2) belonging to the support of γ. Looking at the first equation shows that x_2 is a function of x_1, so the support of γ is included in a graph of a function $T = T(x_1)$.

2.2 One dimensional case

From now on, we will recall basic tools used in the optimal transport theory in the one dimensional case in order to achieve our goal :

How do we compute γ the solution of $(\mathcal{P})$?

From now on, we assume that both ν_1 and ν_2 admits a density with respect to the Lebesgue measure. We begin by a fundamental criterion.

Optimality criterion in one dimension. — If the support of γ is included onto a graph of a non-decreasing function $T = T(x)$, then γ is optimal in $(\mathcal{P})$. The converse is true as well.

Proof. See [Vil21], Proposition 2.24 and the commentary below the Open Problem 2.25.

As a consequence, it is sufficient to find a non-decreasing function T such that $T\#\nu_1 = \nu_2$ and set $\gamma = (\mathrm{Id}, T)\#\nu_1$. For this purpose, we must recall some basics definitions.

Cumulative distribution function. — Given a probability measure $\mu \in \mathcal{P}(X)$, its cumulative distribution function F_μ is defined for $x \in \mathbb{R}$ as

$$F_\mu(x) = \mu((-\infty, x])$$

and is non-decreasing, right-continuous everywhere and continuous at any non-atomic point of μ.

We wish to inverse such a function but unfortunately F_μ is not strictly non-decreasing (it is the case when the support of μ is $\mathbb{R}$). It admits nonetheless a kind of a pseudo-inverse denoted by $F_\mu^{[-1]}$. Moreover a formula is given by

$$F_\mu^{[-1]}(x) = \inf\{t \in \mathbb{R} : F(t) \geqslant x\}$$

Now we set

$$T = F_{\nu_2}^{[-1]} \circ F_{\nu_1}$$

Notice that such a map is monotone non-decreasing by composition. It is sufficient to prove that

$$T\#\nu_1 = \nu_2$$

For that purpose we need two lemmas:

Lemma 1. — If μ is atomless, then

$$(F_\mu)\,\#\mu = \mathrm{Leb}_{[0,1]}$$

Lemma 2. — Without any restriction on μ,

$$\left(F_\mu^{[-1]}\right)\#\,\mathrm{Leb}_{[0,1]} = \mu$$

Proof's idea of Lemma 1. If the support of μ is assume to be equals to $\mathbb{R}$, then the infimum in the formula of $F_\mu^{[-1]}$ is attained and moreover $F_\mu^{[-1]}$ is in fact the inverse of F_μ. Then it is easy to check that for every $a \in (0,1)$,

$$\mu([0,a)) = a$$

and conclude by characterization of the Lebesgue measure.

Proof's idea of Lemma 2. For every $a \in (0,1)$,

$$\begin{aligned}\mathrm{Leb}_{[0,1]}\left(\left\{x \in [0,1] : F_\mu^{[-1]}(x) \leqslant a\right\}\right) &= \mathrm{Leb}_{[0,1]}\left(\{x \in [0,1] : x \leqslant F_\mu(a)\}\right) \\ &= F_\mu(a)\end{aligned}$$

which is sufficient to conclude.

References

[AC11] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[PC+19] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[San15] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[Vil21] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.