FRÉCHET BARYCENTERS AND A LAW OF LARGE NUMBERS FOR MEASURES ON THE REAL LINE

ALEXEY KROSHNIN[1,2] AND ANDREI SOBOLEVSKI[1,3]

Abstract. Endow the space $\mathcal{P}(\mathbb{R})$ of probability measures on $\mathbb{R}$ with a transportation cost $J(\mu,\nu)$ generated by a translation-invariant convex cost function. For a probability distribution on $\mathcal{P}(\mathbb{R})$ we formulate a notion of average with respect to this transportation cost, called here the *Fréchet barycenter*, prove a version of the law of large numbers for Fréchet barycenters, and discuss the structure of $\mathcal{P}(\mathbb{R})$ related to the transportation cost J.

1. Introduction

The law of large numbers is probably the oldest result in statistics: already by Kepler's time, the sample arithmetic mean became universally accepted as estimator for a quantity whose measurements are subject to errors (see, e.g., [8]). About a century later, in Part 4 of his *Ars Conjectandi*, published posthumously in 1713 [2, 8], Jacob Bernoulli gave a rigorous proof that for two given outcomes, the probability of either outcome can be determined by averaging the number of its occurrences over a large sample.

Note that Bernoulli's argument is based on embedding the two distinct outcomes $\{0,1\}$ into the real line $\mathbb{R}$ whose affine structure is then used to perform averaging. To use a similar approach in non-affine spaces, such as collections of geometric shapes, Maurice Fréchet introduced in his memoir [5] a notion of averaging on a general metric space (M,d): for a Borel measure μ its *Fréchet mean* is the global minimum of $\int d^2(x,\cdot)\,\mathrm{d}\mu(x)$. This construction reduces to the conventional mean if the space (M,d) is Euclidean and the measure μ has finite second moment, but in more complex situations Fréchet means may fail to exist or to be unique.

In this paper we consider averaging in the space $\mathcal{P}(M)$ of measures over a metric space M, using a transport optimization procedure to define a suitable concept of a "typical element" [5, p. 224ff.], which extends the notion of Fréchet mean. For the first time a construction of this kind was introduced by M. Agueh and G. Carlier in [1]: a *Wasserstein barycenter* of a family of measures on the Euclidean space $\mathbb{R}^d$ is defined as the Fréchet mean using the 2-Wasserstein distance W_2 on $\mathcal{P}(\mathbb{R}^d)$, which is given by minimization of the mean-square displacement (a precise definition is recalled on p. 5). In [1], the authors establish existence, uniqueness, and regularity

[1] Institute for Information Transmission Problems of RAS (Kharkevich Institute)
[2] Moscow Institute of Physics and Technology
[3] National Research University "Higher School of Economics"
E-mail addresses: sobolevski@iitp.ru.
Date: December 2015.

results for the Wasserstein barycenter and, when $d = 1$, provide an explicit formula for the Wasserstein barycenter in terms of quantile functions of the measures involved.

Here we limit ourselves to the case $d = 1$ but take a general transportation cost

$$J(\mu, \nu) = \inf_{\substack{0 \le \gamma \in \mathcal{P}(\mathbb{R}^2):\\ \pi^x_{\#}\gamma = \mu, \pi^y_{\#}\gamma = \nu}} \int g(x-y)\,\mathrm{d}\gamma(x,y) \ge 0,$$

where $g(\cdot) \ge 0$ is a strictly convex function that satisfies $g(0) = 0$. Since $J(\mu, \nu) = 0$ iff $\mu = \nu$, this cost quanifiies separation between measures μ and ν in $\mathcal{P}(\mathbb{R})$ but does not necessarily satisfy the triangle inequality; in [5] such measures of separation (with an additional requirement of symmetry, $J(\mu, \nu) = J(\nu, \mu)$) are called "*écarts*" to distinguish them from *distances*, for which the triangle inequality is satisfied.

Let a measure ν be fixed and $\boldsymbol{\mu}$ be a random element of $\mathcal{P}(\mathbb{R})$ with distribution P_μ. We introduce a notion of Fréchet typical element of P_μ with respect to $J(\cdot,\cdot)$, which we propose to call the *Fréchet barycenter* of P_μ. It is defined as any measure ν for which the expected cost

$$\mathbb{E}\, J(\boldsymbol{\mu}, \nu) = \int_{\mathcal{P}(\mathbb{R})} J(\mu, \nu)\,\mathrm{d}P_\mu$$

attains its minimum over $\mathcal{P}(\mathbb{R})$. Rigorous definitions of such a distribution and an integral are formulated in Section 4.

Suppose $\mathbb{E}\, J(\boldsymbol{\mu}, \cdot)$ is not identically equal to $+\infty$ on $\mathcal{P}(\mathbb{R})$. Under this assumption the strict convexity of g ensures that a Fréchet barycenter ν^* is unique, and the one-dimensional setting allows to provide it with an explicit expression. Namely, the *quantile function* of the distribution ν^* is given by

$$\psi\colon (0,1) \ni x \mapsto \arg\min_{y\in\mathbb{R}} \mathbb{E}\, g(F^{-1}_{\boldsymbol{\mu}}(x) - y),$$

where $F^{-1}_{\boldsymbol{\mu}}$ is the quantile function (i.e., the inverse of the cumulative distribution function) of the random measure $\boldsymbol{\mu}$. In particular, when $g(x,y) = (x-y)^2$ the function ψ reduces to the usual arithmetic average of quantile functions, which was shown in [1] to give the Wasserstein barycenter on $\mathbb{R}$.

In other words, *every quantile of the Fréchet barycenter ν^*, defined with respect to a transportation cost on $\mathcal{P}(\mathbb{R})$ generated by the cost function $g(\cdot)$, is given by the typical value of the corresponding quantile, defined on $\mathbb{R}$ with respect to the function $g(\cdot)$*. We establish this result first for Fréchet barycenters of finite samples (Theorem 1) and then for an arbitrary probability distribution P_μ on $\mathcal{P}(\mathbb{R})$ such that $\mathbb{E}\, J(\boldsymbol{\mu}, \nu) < +\infty$ for at least one measure ν (Theorem 2).

These results are then used to obtain the following form of a law of large numbers: *the Fréchet barycenter of an independent sample of size n from the distribution P_μ weakly converges as $n \to \infty$ to the Fréchet barycenter of P_μ itself.*

To see this we first establish a similar statement for Fréchet typical elements in $\mathbb{R}$ with respect to the function g. Let the distribution P_X on $\mathbb{R}$ be such that the function $\xi(\cdot) = \mathbb{E}\, g(\boldsymbol{X} - \cdot)$ is finite for all $x \in \mathbb{R}$. Then the Fréchet typical element of the i.i.d. sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from the distribution P_X converges to $x^* := \arg\min_{x\in\mathbb{R}} \xi(x)$ as $n \to \infty$ (Theorem 3). This implies the law of large numbers for random measures on $\mathbb{R}$, which is again proved "quantile-wise" (Theorem 4).

This law of large numbers is then given another fornulation: *the sequence of "empirical" Fréchet barycenters* $\boldsymbol{\nu}_n$ *converges to* ν^* *in the sense that* $J(\boldsymbol{\nu}_n, \nu^*) \to 0$ *almost surely* (Theorem 5). This notion of convergence is shown to be somewhat stronger than the weak convergence of measures (Theorem 6). We conclude with a discussion of the topology on $\mathcal{P}(\mathbb{R})$ generated by this convergence, employing the Urysohn lemma to show that this topology is metrizable even if $J(\cdot,\cdot)$ is itself not a metric (Theorem 7).

The results reported here should be compared with recent results of T. Le Gouic and J.-M. Loubes [7] (see also the earlier series of works [4, 6, 3], where the law of large numbers based on Wasserstein barycenters in considered in the context of deformation models). The paper [7] contains a general consistency result for Wasserstein barycenters when a sequence of probability distributions over the 2-Wasserstein space $(\mathcal{P}(M), W_2)$ converges with respect to the "derivate" 2-Wasserstein distance (a 2-Wasserstein distence defined over $(\mathcal{P}(M), W_2)$ taken as a metric space itself). It should be stressed that in [7] the base space M is assumed to be an arbitrary locally compact geodesic space.

Our results are comparable to the construction of [7] in the special situation when the two settings agree, i.e., when $M = \mathbb{R}$ and $g(x-y) = (x-y)^2$. Note that we do not consider general convergence of measures, but only convergence of an empirical measure of an i.i.d. sample to the distribution from which it is drawn. In this setting conditions of our Theorems 2 and 4 are somewhat easier to check then the condition of convergence under the "derivate" 2-Wasserstein metric. On the other hand, we use the one-dimensional geometry of the base space in an essential way, which is not readily extendable to a general manifold.

The paper is organized as follows. In Section 2 we introduce some standard definitions and notation and recall that when the cost function is defined on the real line and convex, the optimal transport plan is given by a monotone map defined solely by the marginal measures irrespective of the specific form of the cost function. Then we define in Section 3 the generalized barycenter of a finite set of measures in $\mathcal{P}(\mathbb{R})$ and provide it with an explicit representation. In Section 4 this construction is extended to barycenters of continuous families of measures in $\mathcal{P}(\mathbb{R})$. Then the central result of this paper is proved in Section 5: the convergence of empirical barycenters for an i.i.d. sequence of random measures to the "barycenter expected value" of the corresponding distribution. In order to prove it, we establish first a version of the law of large numbers for a suitable "nonlinear averaging" (Theorem 3) which has some independent interest. Section 6 establishes convergence with respect to the transportation cost $J(\cdot,\cdot)$, a property which turns out to be more informative than just the weak convergence, while Section 7 shows metrizability of this convergence. An Appendix contains proof of a technical measure-theoretic result used in Section 4.

2. Basic facts on mass transportation on the real line

For a measurable space X denote the space of probability measures on X by $\mathcal{P}(X)$. In particular, if the space X is topological, we assume it to be endowed with the standard Borel σ-algebra $\mathcal{B}(X)$. For two measurable spaces X, Y, a measurable map $T \colon X \to Y$ induces a map $T_\# \colon \mathcal{P}(X) \to \mathcal{P}(Y)$ given by $T_\#\mu(A) := \mu(T^{-1}(A))$

for any measurable $A \subset Y$. Recall that for any integrable function f

$$\int_Y f(y)\,\mathrm{d}(T_\# \mu) = \int_X f(T(x))\,\mathrm{d}\mu.$$

For two measures $\mu, \nu \in \mathcal{P}(X)$ define the set of *transport plans* taking μ to ν as

$$\Pi(\mu,\nu) := \Big\{\gamma \in \mathcal{P}(X \times X)\colon \pi^x_\# \gamma = \mu,\ \pi^y_\# \gamma = \nu\Big\},$$

where π^x and π^y are projections of $X \times X$ to the first and second factor respectively. Observe that $\Pi(\mu,\nu)$ is always nonempty because it contains the direct product measure $\mu \times \nu$.

Fix a measurable function $c\colon X \times X \to \mathbb{R}$ and call it the *cost function*. The *transportation cost* of a transport plan γ is defined as

$$K(\gamma) = \int_{\mathbb{R}\times\mathbb{R}} c(x,y)\,\mathrm{d}\gamma.$$

The *Monge–Kantorovich problem* for given $\mu, \nu \in \mathcal{P}(X)$ consists in minimizing the *transportation cost* $K(\gamma)$ over all $\gamma \in \Pi(\mu,\nu)$. A transport plan γ^* is called *optimal* if K attains its minimum over $\Pi(\mu,\nu)$ at γ^*. If moreover $\gamma^* = (\mathrm{id} \times T^*)_\# \mu$ for some measurable T^*, then T^* is called the *optimal transport map*. Observe that $T^*_\# \mu = \nu$ if T^* exists.

Recall that a measure μ on $\mathbb{R}$ is characterized by its (left-continuous) *cumulative distribution function* $F_\mu(x) := \mu((-\infty, x))$, and its (left-continuous) inverse $F_\mu^{-1}(y) := \inf\{x\colon F(x) \geq y\}$ for $0 < y < 1$ is called the *quantile function* of the measure μ. In this paper we consider $X = \mathbb{R}$ and take $c(x,y) = g(x-y)$, where the function g is strictly convex. We assume that g attains its minimum at $x = 0$ (unique because g is strictly convex) and, without loss or generality, that $g(0) = 0$. In this setting the Monge–Kantorovich problem has a well-known explicit solution, which we recall here.

Theorem 0. *Let the infimum of transportation cost $K(\gamma)$ on $\Pi(\mu,\nu)$ be finite. Then an optimal transport plan exists and has a uniquely defined form*

$$\gamma^* = (F_\mu^{-1} \times F_\nu^{-1})_\#\ \mathcal{L}|_{(0,1)}\,, \tag{1}$$

where $\mathcal{L}|_{(0,1)}$ is the standard Lebesque measure on $(0,1)$.

We first prove the following lemma, which is used repeatedly in the sequel.

Lemma 1. *For a convex function $g(\cdot)$, inequalities $x < x+\delta < y$ and $x < y-\delta < y$ imply that*

$$g(x) + g(y) \geq g(x+\delta) + g(y-\delta),$$

and when the function g is strictly convex, this inequality is strict.

Proof. Indeed, if $\lambda = \delta/(y-x)$, then clearly $0 < \lambda < 1$ and

$$x + \delta = (1-\lambda)x + \lambda y,$$
$$y - \delta = \lambda x + (1-\lambda) y.$$

Therefore convexity can be used to get

$$g(x+\delta) \leq (1-\lambda) g(x) + \lambda g(y),$$
$$g(y-\delta) \leq \lambda g(x) + (1-\lambda) g(y),$$

which implies the statement. The case of strict convexity is treated similarly. $\square$

Sketch of proof of Theorem 0. Consider points x_1, x_2, y_1, and y_2 in $\mathbb{R}$ such that $x_1 < x_2$ and $y_1 < y_2$. Then both $x_1 - y_1$ and $x_2 - y_2$ lie strictly between $x_1 - y_2$ and $x_2 - y_1$, which in view of the strict convexity of g and of Lemma 1 implies that

$$g(x_1 - y_1) + g(x_2 - y_2) < g(x_1 - y_2) + g(x_2 - y_1).$$

As $\inf K(\gamma)$ over $\Pi(\mu,\nu)$ is finite, there cannot exist sets A, B with $\gamma^*(A) > 0$, $\gamma^*(B) > 0$ such that $(x_1 - x_2)(y_1 - y_2) < 0$ for all $(x_1, y_1) \in A$ and $(x_2, y_2) \in B$. Thus, if an optimal map $T(\cdot)$ exists, it must be nondecreasing μ-a.e., and the support of γ^* must have the form $\{(x(t), y(t)) \colon t \in (a, b)\}$ for some interval (a, b). As $\gamma^* \in \Pi(\mu, \nu)$, such a transport plan is necessarily given by (1). $\square$

Remark 1. Observe that if the optimal map $T(\cdot)$ exists (in particular, if μ does not have atoms), then μ-a.e.

$$T(x) = F_\nu^{-1}(F_\mu(x)). \tag{2}$$

Hence the optimal map does not depend on the specific form of the cost function g (apart from the fact that it is strictly convex).

Remark 2. If $g(\cdot)$ is convex but not strictly so, then formulas (1) and (2) still give an optimal transport plan and an optimal map (provided the latter is well defined), but there may exist other optimal transport plans and maps.

3. Fréchet barycenters for convex cost functions on $\mathbb{R}$

Define a functional on $\mathcal{P}(\mathbb{R}) \times \mathcal{P}(\mathbb{R})$ by

$$J(\mu,\nu) := \inf\{K(\gamma) \colon \gamma \in \Pi(\mu,\nu)\} = \int_0^1 g(F_\mu^{-1}(x) - F_\nu^{-1}(x))\,\mathrm{d}x \tag{3}$$

(where we have used the above theorem) or, if an optimal map T exists, by

$$J(\mu,\nu) = \int_{\mathbb{R}} g(x - T(x))\,\mathrm{d}\mu.$$

One can show that if $c(\cdot,\cdot)$ is a distance function on $\mathbb{R}$, then J satisfies the triangle inequality on $\mathcal{P}(\mathbb{R})$, i.e., it is itself a distance. Another important particular case is when $c(x, y) = (x - y)^2$: in this case $\sqrt{J(\mu,\nu)}$ gives a distance on $\mathcal{P}(\mathbb{R})$, called the 2-*Wasserstein distance* and denoted $W_2(\mu,\nu)$.

Definition 1. Consider a finite set $\mu_1, \mu_2, \ldots, \mu_n$ of measures in $\mathcal{P}(\mathbb{R})$ and the *weights* $\lambda_1 > 0, \lambda_2 > 0, \ldots, \lambda_n > 0$. The *Fréchet barycenter* $\operatorname{bar}(\mu_i, \lambda_i)_{1\le i\le n} \in \mathcal{P}(\mathbb{R})$ with respect to the cost function $c(x, y) = g(x-y)$ is a measure that minimizes

$$\sum_{1\le i\le n} \lambda_i J(\mu_i, \nu)$$

over $\nu \in \mathcal{P}(\mathbb{R})$.

Without loss of generality we can assume the weights to be normalized so that $\sum_i \lambda_i = 1$. In particular, the definition of Wasserstein barycenter given in [1] is recovered when $J(\mu,\nu) = W_2^2(\mu,\nu)$.

Theorem 1. *Suppose $\sum_i \lambda_i J(\mu_i, \nu) < +\infty$ for some $\nu \in \mathcal{P}(\mathbb{R})$. Then the Fréchet barycenter exists and is uniquely defined as $\nu^* = \psi_\# \mathcal{L}|_{(0,1)}$, where for $0 < x < 1$*

$$\psi(x) := \arg\min_{y\in\mathbb{R}} \sum_{1\le i\le n} \lambda_i g(F_{\mu_i}^{-1}(x) - y). \tag{4}$$

Proof. Observe first that $J \geq 0$ and therefore the condition of the theorem ensures that $\inf \sum_i \lambda_i J(\mu_i, \nu)$ over $\nu \in \mathcal{P}(\mathbb{R})$ is finite.

Now fix some n-tuple $a = (a_1, a_2, \dots, a_n) \in \mathbb{R}^n$ and consider the function

$$f_a(y) := \sum_{1\le i\le n} \lambda_i g(a_i - y).$$

The assumptions on g imply that for all a this function is also strictly convex and attains a unique minimum on $\mathbb{R}$, so the function ψ introduced in (4) is well-defined.

Let us show that ψ is nondecreasing. Indeed, assume on the contrary that $\psi(x_1) > \psi(x_2)$ for some $0 < x_1 < x_2 < 1$. Then convexity of g, monotonicity of $F_{\mu_i}^{-1}(\cdot)$, and Lemma 1 imply that

$$g(F_{\mu_i}^{-1}(x_1) - \psi(x_2)) + g(F_{\mu_i}^{-1}(x_2) - \psi(x_1)) \\ \le g(F_{\mu_i}^{-1}(x_1) - \psi(x_1)) + g(F_{\mu_i}^{-1}(x_2) - \psi(x_2)),$$

for all $1 \le i \le n$, where $y_1 = \psi(x_2)$ and $y_2 = \psi(x_1)$ (the inequality will be nonstrict if $F_{\mu_i}^{-1}(x_1) = F_{\mu_i}^{-1}(x_2)$). Multiplying each of these inequalities by λ_i and summing over i, we get

$$\sum_{1\le i\le n} \lambda_i \big[g(F_{\mu_i}^{-1}(x_1) - \psi(x_2)) + g(F_{\mu_i}^{-1}(x_2) - \psi(x_1))\big] \\ \le \sum_{1\le i\le n} \lambda_i \big[g(F_{\mu_i}^{-1}(x_1) - \psi(x_1)) + g(F_{\mu_i}^{-1}(x_2) - \psi(x_2))\big]. \tag{5}$$

On the other hand, $\psi(x_1)$ and $\psi(x_2)$ are the unique minima of the right-hand side of (4) for the respective values x_1 and x_2. Thus

$$\sum_{1\le i\le n} \lambda_i g(F_{\mu_i}^{-1}(x_1) - \psi(x_2)) > \sum_{1\le i\le n} \lambda_i g(F_{\mu_i}^{-1}(x_1) - \psi(x_1))$$

and

$$\sum_{1\le i\le n} \lambda_i g(F_{\mu_i}^{-1}(x_2) - \psi(x_1)) > \sum_{1\le i\le n} \lambda_i g(F_{\mu_i}^{-1}(x_2) - \psi(x_2)).$$

Adding the latter two inequalitites term by term, we obtain a contradiction with (5), which proves that ψ is nondecreasing.

Define now $F(x) := \inf\{y \in (0,1) \colon \psi(y) \ge x\}$ whenever the set in the r.h.s. is non-empty, and $F(x) = 1$ if $\psi(y) < x$ for all $0 < y < 1$. Then $F(-\infty) = 0$, $F(+\infty) = 1$ and F is left-continuous, i.e., there exists a measure $\nu^* \in \mathcal{P}(\mathbb{R})$ such that $\nu^*((-\infty, x)) = F(x)$ for all $x \in \mathbb{R}$. Note that $\psi(\cdot)$ as defined in (4) is left-continuous because all $F_{\mu_i}^{-1}(\cdot)$ are, so $F^{-1}(x) = \psi(x)$ on $(0,1)$.

We now check that ν^* is a Fréchet barycenter. For any $\nu \in \mathcal{P}(\mathbb{R})$,

$$\sum_{1\le i\le n} \lambda_i J(\mu_i,\nu) = \sum_{1\le i\le n} \lambda_i \int_0^1 g(F_{\mu_i}^{-1} - F_\nu^{-1})\,\mathrm{d}x$$
$$= \int_0^1 \sum_{1\le i\le n} \lambda_i g(F_{\mu_i}^{-1} - F_\nu^{-1})\,\mathrm{d}x \ge \int_0^1 \sum_{1\le i\le n} \lambda_i g(F_{\mu_i}^{-1} - \psi)\,\mathrm{d}x$$
$$= \sum_{1\le i\le n} \lambda_i \int_0^1 g(F_{\mu_i}^{-1} - F_{\nu^*}^{-1})\,\mathrm{d}x = \sum_{1\le i\le n} \lambda_i J(\mu_i,\nu^*).$$

The latter quantity is finite, and the strict convexity of g ensures that the inequality of the middle line is strict unless $\nu = \nu^*$. Thus $\nu^* = \mathrm{bar}(\mu_i,\lambda_i)_{1\le i\le n}$. □

Remark 3. If $g(\cdot)$ is smooth, then the function ψ is determined from the equation

$$\sum_i \lambda_i g'(F_{\mu_i}^{-1} - \psi) = 0.$$

In particular when $g(x) = x^2$ we recover the formula $\psi(x) = \sum_i \lambda_i F_{\mu_i}^{-1}(x)$ for the Wasserstein barycenter [1].

Remark 4. If g is convex but not strictly so, then $\arg\min$ in (4) may be attained on an interval rather than at a single point. Define $\psi(x)$ to be the left endpoint of this interval; we will then still obtain a Fréchet barycenter as the measure ν^* for which $\nu^*((-\infty,x)) = \psi^{-1}(x)$, but this barycenter will not necessarily be unique.

4. The Fréchet barycenter of a continuous distribution

Now we extend the notion of Fréchet barycenter to continuous families of measures, which allows to define an "expected value" for a probability distribution over $\mathcal{P}(\mathbb{R})$. Endow $\mathcal{P}(\mathbb{R})$ with topology of weak convergence, and let $\mathcal{B}(\mathcal{P}(\mathbb{R}))$ be the corresponding Borel σ-algebra. We need a technical measurability lemma whose proof is postponed to Appendix.

Lemma 2. *The function $K(\mu,x) := F_\mu^{-1}(x)$, where $\mu \in \mathcal{P}(\mathbb{R})$ and $0 < x < 1$, is measurable with respect to the product σ-algebra $\mathcal{B}(\mathcal{P}(\mathbb{R})) \otimes \mathcal{B}((0,1))$.*

Let now $\boldsymbol{\mu}$ be a random element of $\mathcal{P}(\mathbb{R})$ distributed according to a law P_μ. Recall from the last section that $g(\cdot)$ is assumed to be a strictly convex function attaining on $\mathbb{R}$ a minimal value $g(0) = 0$. For the functional $J(\mu,\nu)$ defined in (3)

$$\mathbb{E}\,J(\boldsymbol{\mu},\nu) = \int_{\mathcal{P}(\mathbb{R})} J(\mu,\nu)\,\mathrm{d}P_\mu \in [0,+\infty].$$

Definition 2. Consider the problem of minimizing $\mathbb{E}\,J(\boldsymbol{\mu},\nu)$ over $\mathcal{P}(\mathbb{R})$ and denote its solution by $\mathrm{bar}(P_\mu)$ or ν^* for short. We call the measure $\mathrm{bar}(P_\mu)$ the *Fréchet barycenter* of the distribution P_μ.

Lemma 3. *Let $\boldsymbol{X}$ be a random element of $\mathbb{R}$ with distribution P_X. For any $x \in \mathbb{R}$ consider the function*

$$\xi(x) := \mathbb{E}\,g(\boldsymbol{X} - x) \tag{6}$$

taking values in $[0,+\infty]$. This function attains a unique minimum provided $\xi(x) < +\infty$ for some x.

Proof. Indeed consider the set

$$I := \{x \in \mathbb{R} \colon \xi(x) < +\infty\}.$$

Clearly $\xi(\cdot)$ is convex because so is $g(\cdot)$, which implies that I is a segment.

We now show that $\xi(\cdot)$ is lower semicontinuous. Take an arbitrary $x_0 \in \mathbb{R}$.

Case 1: $\xi(x_0) = +\infty$. Fix an arbitrary $C > 0$. Then there exists a segment $[-A, A]$ such that $\int_{[-A,A]} g(t - x_0)\,\mathrm{d}P_X(t) > C$. The function g is continuous and therefore uniformly continuous on $[-A, A]$, so for a suitable $\delta > 0$ the condition $|x - x_0| < \delta$ implies that

$$\xi(x) = \mathbb{E}\, g(\boldsymbol{X} - x) \geq \int_{[-A,A]} g(t - x)\,\mathrm{d}P_X > C.$$

Therefore $\lim_{x \to x_0} \xi(x) = +\infty = \xi(x_0)$.

Case 2: $\xi(x_0) < +\infty$. Fix $\epsilon > 0$. Then there exists A such that $\int_{-A}^{A} g(t - x_0)\,\mathrm{d}P_X(t) > \xi(x_0) - \epsilon$. The continuity of g implies that there exists $\delta > 0$ such that from $|x - x_0| < \delta$ it follows that

$$\xi(x) = \mathbb{E}\, g(\boldsymbol{X} - x) \geq \int_{-A}^{A} g(t - x)\,\mathrm{d}P_X(t) > \xi(x_0) - \epsilon.$$

Thus $\lim_{x \to x_0} \xi(x) \geq \xi(x_0)$.

Note also that $\lim_{x \to \pm\infty} g(x) = +\infty$ implies that $\lim_{x \to \pm\infty} \xi(x) = +\infty$.

Together the convexity and lower semicontinuity of ξ mean that $\xi(\cdot)$ is continuous on I. Moreover, if the interval I is open from either end, then $\xi(x)$ grows indefinitely as x approaches that endpoint. Thus a minimum of $\xi(\cdot)$ on I exists and is unique due to the strict convexity of ξ. □

Theorem 2. *Suppose $\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R})$ is a measure-valued random variable with distribution P_μ and $\mathbb{E}\, J(\boldsymbol{\mu}, \nu) < +\infty$ for some $\nu \in \mathcal{P}(\mathbb{R})$. Then there exists a unique solution $\nu^* = \psi_{\#}\, \mathcal{L}|_{(0,1)}$, where for a.e. $x \in (0, 1)$ (cf (4))*

$$\psi(x) := \arg\min_{y \in \mathbb{R}} \mathbb{E}\, g(F_{\boldsymbol{\mu}}^{-1}(x) - y).$$

Proof. Examination of equation (3) shows that for Lebesgue-a.e. $0 < x < 1$ there exists a value y_x such that $\mathbb{E}\, g(F_{\boldsymbol{\mu}}^{-1}(x) - y_x) < +\infty$. Therefore the function

$$\psi(x) := \arg\min_{y \in \mathbb{R}} \mathbb{E}\, g(F_{\boldsymbol{\mu}}^{-1}(x) - y)$$

is defined according to Lemma 2 for Lebesgue-a.e. $0 < x < 1$. Using the same argument as in Theorem 1 (though with integrals over P_μ instead of sums over λ_i), one can show that the function ψ is nondecreasing. Thus it can be extended over the whole interval $(0, 1)$, e.g., by left continuity. Then ψ generates a measure $\nu^* = \psi_{\#}\, \mathcal{L}|_{(0,1)} \in \mathcal{P}(\mathbb{R})$, where $F_{\nu^*}^{-1}(\cdot) = \psi(\cdot)$ a.e.

We are left with a task of checking that ν^* is a Fréchet barycenter. This can again be done in the same way as in Theorem 1, exchanging the order of integration over $0 < x < 1$ and integration over $\mathcal{P}(\mathbb{R})$ (which replaces summation over i) by Fubini's theorem. In the process we employ the measurability lemma formulated at the beginning of this section. Uniqueness of the Fréchet barycenter follows because ψ (more precisely, its left-continuous version) is defined uniquely. □

5. Weak convergence of the empirical Fréchet barycenter

First we prove a version of law of large numbers for a nonlinear version of averaging performed in terms of the function g over ordinary scalar random variables. We then employ this result to prove convergence of quantiles for a sequence of empirical Fréchet barycenters, which implies the weak convergence of the barycenters themselves.

Theorem 3. *Let $\{\boldsymbol{X}_n\}_{n\in\mathbb{N}}$ be a sequence of i.i.d. real random variables with distribution P_X such that the function $x \mapsto \xi(x) = \mathbb{E}\, g(\boldsymbol{X} - x)$ defined in (6) is finite for all $x \in \mathbb{R}$. Let $x^* := \arg\min_{x\in\mathbb{R}} \xi(x)$ and define the random variables*

$$\overline{\boldsymbol{X}}_n := \arg\min_{x\in\mathbb{R}} \sum_{1\le i\le n} g(\boldsymbol{X}_i - x), \qquad n = 1, 2, \ldots \tag{7}$$

Then $\overline{\boldsymbol{X}}_n$ converges to x^* a.s.

Proof. It follows from Lemma 3 that x^* is well defined. Suppose that $\overline{\boldsymbol{X}}_n$ fails to converge to x^*, and assume specifically that for some $\delta > 0$ there exists a subsequence $n_k \to \infty$ such that $\overline{\boldsymbol{X}}_{n_k} > x^* + 2\delta$ (the remaining case $\overline{\boldsymbol{X}}_{n_k} < x^* - 2\delta$ can be treated similarly).

We will show that $\xi(x^*) - \xi(x^* + \delta) \ge \epsilon$ for a suitable $\epsilon > 0$, which gives the desired contradiction because $\xi(\cdot)$ is supposed to attain its minimum at x^*. It is enough to show for all sufficiently large n_k that

$$\frac{1}{n_k} \sum_{1\le i\le n_k} g(\boldsymbol{X}_i - x^*) - \frac{1}{n_k} \sum_{1\le i\le n_k} g(\boldsymbol{X}_i - x^* - \delta) > \epsilon \tag{8}$$

and to invoke the law of large numbers for the random variables $g(\boldsymbol{X}_i - x)$.

Define

$$\epsilon(\boldsymbol{X}_i) := g(\boldsymbol{X}_i - x^*) + g(\boldsymbol{X}_i - x^* - 2\delta) - 2g(\boldsymbol{X}_i - x^* - \delta) > 0;$$

this quantity is positive for any i because of the strict convexity of $g(\cdot)$. Observe that the points $\boldsymbol{X}_i - x^* - 2\delta$ and $\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{n_k} + \delta$ lie between $\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{n_k}$ and $\boldsymbol{X}_i - x^* - \delta$, and therefore the strict convexity of g and Lemma 1 imply that

$$g(\boldsymbol{X}_i - x^* - \delta) - g(\boldsymbol{X}_i - x^* - 2\delta) > g(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{n_k} + \delta) - g(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{n_k}).$$

Combining this with the definition of $\epsilon(\boldsymbol{X}_i)$, we get

$$\begin{aligned} g(\boldsymbol{X}_i - x^*) - g(\boldsymbol{X}_i - x^* - \delta) &= g(\boldsymbol{X}_i - x^* - \delta) - g(\boldsymbol{X}_i - x^* - 2\delta) + \epsilon(\boldsymbol{X}_i) \\ &> g(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{n_k} + \delta) - g(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{n_k}) + \epsilon(\boldsymbol{X}_i). \end{aligned}$$

Averaging these inequalities over $1 \le i \le n_k$, we obtain

$$\begin{aligned} &\frac{1}{n_k} \sum_{1\le i\le n_k} g(\boldsymbol{X}_i - x^*) - \frac{1}{n_k} \sum_{1\le i\le n_k} g(\boldsymbol{X}_i - x^* - \delta) \\ &> \frac{1}{n_k} \sum_{1\le i\le n_k} g(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{n_k} + \delta) - \frac{1}{n_k} \sum_{1\le i\le n_k} g(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{n_k}) + \frac{1}{n_k} \sum_{1\le i\le n_k} \epsilon(\boldsymbol{X}_i) \\ &\ge \frac{1}{n_k} \sum_{1\le i\le n_k} \epsilon(\boldsymbol{X}_i), \end{aligned}$$

where the last inequality follows from the definition (7) of $\overline{\boldsymbol{X}}_{n_k}$.

We now observe that due the strict convexity of $g(\cdot)$, the quantity $2\epsilon := \mathbb{E}\,\epsilon(\boldsymbol{X})$ is positive. We can then invoke the strong law of large numbers for the random variables $g(\boldsymbol{X} - x)$ (and therefore for $\epsilon(\boldsymbol{X})$) to see that a.s.

$$\frac{1}{n}\sum_{1\le i\le n} \epsilon(\boldsymbol{X}_i) \ge \epsilon > 0$$

for all sufficiently large n. This gives (8) and completes the proof. □

Theorem 4. *Take a sequence $\{\boldsymbol{\mu}_n\}$ of random measures in $\mathcal{P}(\mathbb{R})$, independent and identically distributed with the law P_μ such that conditions of Theorem 2 are satisfied. Then empirical Fréchet barycenters $\boldsymbol{\nu}_n = \mathrm{bar}(\boldsymbol{\mu}_i, 1/n)_{1\le i\le n}$ weakly converge to $\nu^* := \mathrm{bar}(P_\mu)$ almost surely.*

Proof. Define a function $\psi(\cdot)$ as in Theorem 2:

$$\psi(x) := \arg\min_{y\in\mathbb{R}} \mathbb{E}\, g(F^{-1}_{\boldsymbol{\mu}}(x) - y)$$

for a.e. $0 < x < 1$. Theorem 3 implies that the quantile $F^{-1}_{\boldsymbol{\nu}_n}(x)$ converges to $\psi(x)$ almost surely for a.e. $0 < x < 1$. As $\psi(\cdot) = F^{-1}_{\nu^*}(\cdot)$ a.e., there exists a dense set $M \subset (0,1)$ such that almost surely $F^{-1}_{\boldsymbol{\nu}_n}(x) \to F^{-1}_{\nu^*}(x)$ for all $x \in M$. But this implies that $F_{\boldsymbol{\nu}_n} \to F_{\nu^*}$ at all points where F_{ν^*} is continuous.

Indeed, let this convergence fail at some x_0 where F_{ν^*} is continuous. Without loss of generality we can assume that there exists a subsequence n_k such that $F_{\boldsymbol{\nu}_{n_k}}(x_0) > F_{\nu^*}(x_0) + \epsilon$ for a suitable $\epsilon > 0$. Take $y \in M$ such that $F_{\nu^*}(x_0) < y < F_{\nu^*}(x_0) + \epsilon$. Then continuity of F_{ν^*} at x_0 implies that there exists $\gamma > 0$ such that $F_{\nu^*}(x) < y$ whenever $x < x_0 + \gamma$. Then $F^{-1}_{\nu^*}(y) \ge x_0 + \gamma$ whereas $F^{-1}_{\nu_{n_k}}(y) \le x_0$, which in the limit $n_k \to \infty$ gives a contradiction. □

6. Convergence of the empirical Fréchet barycenter with respect to the transportation cost J

In this and the next section we will additonally assume that the cost function $g(\cdot) \ge 0$ satisfies the following condition: there exist nonnegative constants A, B such that

$$g(x-y) \le A + B\big(g(x) + g(y)\big) \tag{9}$$

for all $x, y \in \mathbb{R}$ (note that a similar condition with $A = 0$ appears in Fréchet's memoir [5, p. 228]). This condition is not exceedingly restrictive. In particular, it is satisfied for cost functions of algebraic growth:

Lemma 4. *For $p > 0$ suppose that constants $0 < C_1 < C_2$, $x_0 > 0$ are such that $C_1|x|^p < g(x) < C_2|x|^p$ for any $|x| > x_0$. Then (9) holds.*

Proof. Since $g(\cdot)$ is continuous, there exists a finite positive $C_0 := \max_{|x|\le x_0} g(x)$. Then for any x, y on $\mathbb{R}$

$$\begin{aligned} g(x-y) &\le \max\{C_0, C_2|x-y|^p\} \le C_0 + C_2|x-y|^p \\ |x-y|^p &\le 2^p \max\{|x|^p, |y|^p\} \le 2^p(|x|^p + |y|^p) \\ |x|^p &\le \max\{x_0^p, \frac{g(x)}{C_1}\} \le x_0^p + \frac{g(x)}{C_1} \end{aligned}$$

Gathering these results, we get the desired inequality:

$$g(x-y) \le (C_0 + 2^{p+1}C_2 x_0^p) + 2^p \frac{C_2}{C_1}\big(g(x)+g(y)\big). \qquad \square$$

Under condition (9), convergence of Fréchet barycenters can be characterized in terms of the transportation cost J defined in (3).

Theorem 5. *Let $\{\boldsymbol{\mu}_n\}_{n\ge1} \subset C(\nu_0) \subset \mathcal{P}(\mathbb{R})$ be a sequence of i.i.d. random elements with distribution P_μ such that $P_\mu(C(\nu_0)) = 1$. Then, under the hypotheses of Theorem 2, the sequence $\overline{\boldsymbol{\mu}}_n := \operatorname{bar}(\boldsymbol{\mu}_i, 1/n)_{1\le i\le n}$ of empirical Fréchet barycenters satisfies*

$$\lim_{n\to\infty} J(\overline{\boldsymbol{\mu}}_n, \nu^*) = \lim_{n\to\infty} J(\nu^*, \overline{\boldsymbol{\mu}}_n) = J(\nu^*,\nu^*) = 0 \quad a.s.,$$

where $\nu^* := \operatorname{bar}(P_\mu)$ is the Fréchet barycenter of P_μ, and the functional J.

Proof. Consider the function

$$\psi(x) := \arg\min_{y\in\mathbb{R}} \mathbb{E}\, g(F_{\boldsymbol{\mu}}^{-1}(x) - y).$$

According to Theorem 2, $F_{\nu^*}^{-1}(x) = \psi(x)$ a.e. and

$$\mathbb{E}\, J(\boldsymbol{\mu}, \nu^*) = \int_0^1 \mathbb{E}\, g(F_{\boldsymbol{\mu}}^{-1} - \psi)\,\mathrm{d}x < +\infty.$$

It follows that

$$\lim_{k\to\infty} \int_0^{\frac1k} \mathbb{E}\, g(F_{\boldsymbol{\mu}}^{-1} - \psi)\,\mathrm{d}x = \lim_{k\to\infty} \int_{1-\frac1k}^1 \mathbb{E}\, g(F_{\boldsymbol{\mu}}^{-1} - \psi)\,\mathrm{d}x = 0.$$

For the random variables $\int_0^{\frac1k} g(F_{\boldsymbol{\mu}_i}^{-1} - \psi)\,\mathrm{d}x$, $k \ge 1$, the strong law of large numbers implies that

$$\frac1n \sum_{i=1}^n \int_0^{\frac1k} g(F_{\boldsymbol{\mu}_i}^{-1} - \psi)\,\mathrm{d}x \xrightarrow[n\to\infty]{} \mathbb{E} \int_0^{\frac1k} g(F_{\boldsymbol{\mu}}^{-1} - \psi)\,\mathrm{d}x < +\infty \text{ a.s.}$$

Using the bound (9) on $g(\cdot)$ and the construction of the empirical Fréchet barycenter (Theorem 1), we get

$$\begin{aligned}
(10)\qquad \int_0^{\frac1k} g(F_{\overline{\boldsymbol{\mu}}_n}^{-1} - \psi)\,\mathrm{d}x &\le \frac{A}{k} + B\int_0^{\frac1k} \frac1n \sum_{i=1}^n \big(g(F_{\boldsymbol{\mu}_i}^{-1} - F_{\overline{\boldsymbol{\mu}}_n}^{-1}) + g(F_{\boldsymbol{\mu}_i}^{-1} - \psi)\big)\,\mathrm{d}x \le \\
&\le \frac{A}{k} + \frac{2B}{n}\sum_{i=1}^n \int_0^{\frac1k} g(F_{\boldsymbol{\mu}_i}^{-1} - \psi)\,\mathrm{d}x \\
&\xrightarrow[n\to\infty]{} \frac{A}{k} + 2B\,\mathbb{E}\int_0^{\frac1k} g(F_{\boldsymbol{\mu}}^{-1} - \psi)\,\mathrm{d}x \xrightarrow[k\to\infty]{} 0.
\end{aligned}$$

Similarly one can prove that

$$(11)\qquad \int_{1-\frac1k}^1 g(F_{\overline{\boldsymbol{\mu}}_n}^{-1} - \psi)\,\mathrm{d}x \xrightarrow[n\to\infty]{} \frac{A}{k} + 2B\,\mathbb{E}\int_{1-\frac1k}^1 g(F_{\boldsymbol{\mu}}^{-1} - \psi)\,\mathrm{d}x \xrightarrow[k\to\infty]{} 0 \text{ a.s.}$$

But by Theorem 2 $\overline{\boldsymbol{\mu}}_n$ weakly converges to ν^* a.s., hence for all $k \ge 1$

$$(12)\qquad \int_{\frac1k}^{1-\frac1k} g(F_{\overline{\boldsymbol{\mu}}_n}^{-1} - \psi)\,\mathrm{d}x \xrightarrow[n\to\infty]{} \int_{\frac1k}^{1-\frac1k} g(F_{\nu^*}^{-1} - \psi)\,\mathrm{d}x = 0 \text{ a.s.}$$

Now it follows from formulas (10), (11) and (12) that

$$J(\overline{\boldsymbol{\mu}}_n, \nu^*) = \int_0^1 g(F_{\overline{\boldsymbol{\mu}}_n}^{-1} - \psi)\,\mathrm{d}x \xrightarrow[n\to\infty]{} J(\nu^*, \nu^*) = 0 \text{ a.s.}$$

In the same way, one can obtain that a.s.

$$J(\nu^*, \boldsymbol{\nu}_n) \xrightarrow[n\to\infty]{} J(\nu^*, \nu^*) = 0. \qquad \square$$

It remains to clarify the relation between the convergence defined in terms of the transportation cost ($\mu_n \to \mu$ iff $J(\mu_n, \mu) \to 0$), which is used in the preceding theorem, and the weak convergence of measures $\mu_n \rightharpoonup \mu$.

Lemma 5. *Let the sequence $\{\mu_n\}_{n\geq 1}$ be such that $\lim_{n\to\infty} J(\mu_n, \mu^*) = 0$ for some $\mu^* \in \mathcal{P}(\mathbb{R})$; then $\mu_n \rightharpoonup \mu^*$.*

Proof. Suppose on the contrary that the weak convergence of μ_n to μ^* does not hold. Then there exists a point $0 < x_0 < 1$ where $F_{\mu^*}^{-1}$ is continuous but the sequence $F_{\mu_n}^{-1}(x_0)$ does not converge to $F_{\mu^*}^{-1}(x_0)$. Assume specifically that there exists $\epsilon > 0$ such that $F_{\mu_n}^{-1}(x_0) \geq F_{\mu^*}^{-1}(x_0) + 2\epsilon$ for a suitable subsequence, which we still denote μ_n. Monotonicity of $F_{\mu_n}^{-1}$ and continuity of $F_{\mu^*}^{-1}$ at x_0 imply that $F_{\mu_n}^{-1}(x) \geq F_{\mu^*}^{-1}(x) + \epsilon$ for $x_0 < x < x_0 + \delta$ with some $\delta > 0$. Then

$$J(\mu_n, \mu^*) \geq \int_{x_0}^{x_0+\delta} g(F_{\mu_n}^{-1} - F_{\mu^*}^{-1})\,\mathrm{d}x \geq \int_{x_0}^{x_0+\delta} g(\epsilon)\,\mathrm{d}x > 0,$$

which contradicts the assumption $J(\mu_n, \mu^*) \to 0$. $\square$

Theorem 6. *For a sequence of measures $\{\mu_n\}_{n\geq 1}$ and a measure μ^* the following conditions are equivalent:*

(1) *$\mu_n \rightharpoonup \mu^*$ and $J(\mu_n, \nu) \to J(\mu^*, \nu)$ for all $\nu \in \mathcal{P}(\mathbb{R})$;*
(2) *$\mu_n \rightharpoonup \mu^*$ and there exists a measure ν_0 such that $J(\mu_n, \nu_0) \to J(\mu^*, \nu_0)$;*
(3) *$J(\mu_n, \mu^*) \to 0$.*

Proof. Obviously (1) implies (2). To show that (2) implies (3), observe that for any $\epsilon > 0$ we have

$$\int_\epsilon^{1-\epsilon} g(F_{\mu_n}^{-1} - F_\nu^{-1})\,\mathrm{d}x \to \int_\epsilon^{1-\epsilon} g(F_{\mu^*}^{-1} - F_\nu^{-1})\,\mathrm{d}x$$

due to the weak convergence of measures. Since

$$J(\mu_n, \nu) = \Big(\int_0^\epsilon + \int_\epsilon^{1-\epsilon} + \int_{1-\epsilon}^1\Big) g(F_{\mu_n}^{-1} - F_\nu^{-1})\,\mathrm{d}x$$
$$\xrightarrow[n\to\infty]{} J(\mu^*, \nu) = \Big(\int_0^\epsilon + \int_\epsilon^{1-\epsilon} + \int_{1-\epsilon}^1\Big) g(F_{\mu^*}^{-1} - F_\nu^{-1})\,\mathrm{d}x,$$

we have

$$\Big(\int_0^\epsilon + \int_{1-\epsilon}^1\Big) g(F_{\mu_n}^{-1} - g_\nu^{-1})\,\mathrm{d}x \xrightarrow[n\to\infty]{} \Big(\int_0^\epsilon + \int_{1-\epsilon}^1\Big) g(F_{\mu^*}^{-1} - g_\nu^{-1})\,\mathrm{d}x,$$

where the right-hand side vanishes as $\epsilon \to 0$ because the integral $\int_\epsilon^{1-\epsilon} g(F_{\mu^*}^{-1} - F_\nu^{-1})\,\mathrm{d}x$ converges to $J(\mu^*,\nu)$. Thus

$$\begin{aligned}\Big(\int_0^\epsilon + \int_{1-\epsilon}^1\Big) g(F_{\mu_n}^{-1} - F_{\mu^*}^{-1})\,\mathrm{d}x \\ \le \Big(\int_0^\epsilon + \int_{1-\epsilon}^1\Big) (A + Bg(F_{\mu_n}^{-1} - F_\nu^{-1}) + Bg(F_{\mu^*}^{-1} - F_\nu^{-1}))\,\mathrm{d}x \\ \xrightarrow[n\to\infty]{} 2A\epsilon + 2B\Big(\int_0^\epsilon + \int_{1-\epsilon}^1\Big) g(F_{\mu^*}^{-1} - F_\nu^{-1})\,\mathrm{d}x \xrightarrow[\epsilon\to 0]{} 0.\end{aligned}$$

Using the weak convergence again we observe that

$$\int_\epsilon^{1-\epsilon} g(F_{\mu_n}^{-1} - F_{\mu^*}^{-1})\,\mathrm{d}x \to 0$$

for any $\epsilon > 0$. Thus

$$J(\mu_n,\mu) = \int_0^1 g(F_{\mu_n}^{-1} - F_{\mu^*}^{-1})\,\mathrm{d}x \to 0.$$

It remains to prove that (3) imples (1). Fix a measure $\nu \in \mathcal{P}(\mathbb{R})$. By Lemma 5 convergence $J(\mu_n,\mu^*) \to 0$ implies weak convergence $\mu_n \rightharpoonup \mu^*$. Consider

$$\begin{aligned}\int_0^\epsilon g(F_{\mu_n}^{-1} - F_\nu^{-1})\,\mathrm{d}x \le \int_0^\epsilon (A + Bg(F_{\mu_n}^{-1} - F_{\mu^*}^{-1}) + Bg(F_\nu^{-1} - F_{\mu^*}^{-1})]\,\mathrm{d}x \\ \le \int_0^\epsilon (A + Bg(F_{\mu_n}^{-1} - F_{\mu^*}^{-1}) + AB + B^2 g(F_{\mu^*}^{-1} - F_\nu^{-1}))\,\mathrm{d}x \\ \xrightarrow[n\to\infty]{} A(1+B)\epsilon + B^2 \int_0^\epsilon g(F_{\mu^*}^{-1} - F_\nu^{-1})\,\mathrm{d}x \xrightarrow[\epsilon\to 0]{} 0.\end{aligned}$$

A similar result holds for $\int_{1-\epsilon}^1 g(F_{\mu_n}^{-1} - F_\nu^{-1})\,\mathrm{d}x$. Now the weak convergence $\mu_n \rightharpoonup \mu^*$ implies that

$$\int_\epsilon^{1-\epsilon} g(F_{\mu_n}^{-1} - F_\nu^{-1})\,\mathrm{d}x \xrightarrow[n\to\infty]{} \int_\epsilon^{1-\epsilon} g(F_{\mu^*}^{-1} - F_\nu^{-1})\,\mathrm{d}x \xrightarrow[\epsilon\to 0]{} J(\mu^*,\nu).$$

Gathering the results for $\int_0^\epsilon$, $\int_\epsilon^{1-\epsilon}$, and $\int_{1-\epsilon}^1$, we obtain $J(\mu_n,\nu) \to J(\mu^*,\nu)$. □

Remark 5. The arguments of $J(\cdot,\cdot)$ can be simultaneously swapped in each of the conditions (1)–(3) without violating the theorem. In particular $J(\mu_n,\mu^*) \to 0$ is equivalent to $J(\mu^*,\mu_n) \to 0$.

Remark 6. In conditions (1) and (2), it is sufficient to require that

$$\limsup_{\nu\to\infty} J(\mu_n,\nu) \le J(\mu^*,\nu);$$

this is equivalent to convergence $J(\mu_n,\nu) \to J(\mu^*,\nu)$ thanks to features of the weak convergence of measures.

7. Topology of the space $\mathcal{P}(\mathbb{R})$

Here we observe that $\mathcal{P}(\mathbb{R})$ can be endowed with a bundle structure where individual fibers are composed of measures connected by transport maps of finite cost. Recall the cost function g is assumed to satisfy the additional condition (9).

Lemma 6. *Let the relation $\mu \sim \nu$ be defined as $J(\mu,\nu) < +\infty$; then it is an equivalence on $\mathcal{P}(\mathbb{R})$.*

Proof. The relation $\sim$ is obviously reflective. It is symmetric because $g(-x) \leq A + Bg(x)$ according to (9) and therefore

$$J(\nu,\mu) = \int_0^1 g(F_\nu^{-1} - F_\mu^{-1})\,\mathrm{d}x \leq A + BJ(\mu,\nu) \leq +\infty.$$

Finally to prove that it is transitive we observe that for any measures λ, μ, ν

$$J(\mu,\nu) = \int_0^1 g(F_\mu^{-1} - F_\nu^{-1})\,\mathrm{d}x \leq \int_0^1 \big(A + Bg(F_\mu^{-1} - F_\lambda^{-1}) \\ + Bg(F_\nu^{-1} - F_\lambda^{-1})\big)\,\mathrm{d}x = A + BJ(\mu,\lambda) + BJ(\nu,\lambda) < +\infty. \quad \square$$

Denote the equivalence class of ν with respect to the relation $\sim$ by $C(\nu) := \{\mu \in \mathcal{P}(\mathbb{R})\colon \mu \sim \nu\}$. Observe that for any $\nu \in \mathcal{P}(\mathbb{R})$ the function $J(\cdot,\nu)$ is measurable with respect to the Borel σ-algebra generated by the topology of weak convergence in $\mathcal{P}(\mathbb{R})$, and therefore $C(\nu)$ is measurable for any ν. Observe also that for $g(x) = |x|^p$, $p \geq 1$, the corresponding Wasserstein space W_p coincides with $C(\delta_0)$, the equivalence class of a Dirac unit mass at the origin.

Remark 7. Equivalence classes $C(\nu)$ form a continuum. Indeed, take a measure ν_1 for which $\int_0^1 |F_{\nu_1}^{-1}|\,\mathrm{d}x = +\infty$ and define measures ν_α for $\alpha \geq 0$ by rescaling: $F_{\nu_\alpha}^{-1} = \alpha F_{\nu_1}^{-1}$. For $\alpha \neq \beta$

$$\int_0^1 |F_{\nu_\alpha}^{-1} - F_{\nu_\beta}^{-1}|\,\mathrm{d}x = |\alpha - \beta| \int_0^1 |F_\nu^{-1}|\,\mathrm{d}x = +\infty.$$

As the function $g(\cdot)$ is convex and attains a minimum at 0, its growth at infinity is at least linear. Hence there exist constants $C_1 > 0$ and C_2 such that $f(x) \geq C_1|x| + C_2$ for all x. Therefore $J(\nu_\alpha,\nu_\beta) = +\infty$, i.e., $C(\nu_\alpha) \neq C(\nu_\beta)$ for all nonnegative reals $\alpha \neq \beta$.

Lemma 7. (1) *The barycenter* $\operatorname{bar}(\mu_i,\lambda_i)_{1\leq i\leq n}$ *is defined iff the measures $\mu_1,\ldots,\mu_n$ all lie in the same equivalence class.*

(2) *If the barycenter* $\operatorname{bar}(P_\mu)$ *exists, then* $\operatorname{supp} P_\mu$ *belongs to a single equivalence class, i.e., there exists $\nu_0 \in \mathcal{P}(\mathbb{R})$ such that $P_\mu(C(\nu_0)) = 1$.*

Proof. (1) If $\mu_1,\ldots,\mu_n$ belong to different equivalence classes, then the quantity $\sum_{1\leq i\leq n} \lambda_i J(\mu_i,\nu)$ is infinite for any $\nu \in \mathcal{P}(\mathbb{R})$, and no barycenter exists. Conversely, if all measures are equivalent, then $\sum_{1\leq i\leq n} \lambda_i J(\mu_i,\nu)$ is finite e.g. for $\nu = \mu_1$, and a unique barycenter exists by Theorem 1.

(2) Let $P_\mu(C(\nu)) < 1$ for all $\nu \in \mathcal{P}(\mathbb{R})$. Then $P_\mu(\mathcal{P}(\mathbb{R})\setminus C(\nu)) = P_\mu(\boldsymbol{\mu} \notin C(\nu)) = P_\mu(J(\boldsymbol{\mu},\nu) = +\infty) > 0$ and $\mathbb{E}\,J(\boldsymbol{\mu},\cdot) = +\infty$ on $\mathcal{P}(\mathbb{R})$, which again implies there is no barycenter. $\square$

Now we fix some $\nu_0 \in \mathcal{P}(\mathbb{R})$. By Theorem 6, the convergence $J(\mu_n, \mu^*) \to 0$ implies that $J(\mu_n, \nu) \to J(\mu^*, \nu)$ for all $\nu \in C(\nu_0)$. This enables us to define on $C(\nu_0)$ the following topology, which is at least as strong as the topology of weak convergence.

Definition 3. The balls $B_r(\nu) := \{\mu \in C(\nu_0) \colon J(\mu, \nu) < r\}$ form a basis of a *topology* τ_g on $C(\nu_0)$.

Lemma 8. *The space* $(C(\nu_0), \tau_g)$ *is separable.*

Proof. Consider the set S of measures $\mu \in C(\nu_0)$ given by the following construction: fix some $n \geq 1$, choose the points $0 < x_0 < x_1 < \cdots < x_n < 1$ and $q_1 < q_2 < \cdots < q_n$ such that all x_i, $0 \leq i \leq n$, and q_j, $1 \leq j \leq n$, are rational, and define the quantile function of the measure $\mu \in S$ by

$$F_\mu^{-1}(x) = \begin{cases} F_{\nu_0}^{-1}(x), & x \leq x_0, \\ q_i, & x_{i-1} < x \leq x_i,\ 1 \leq i \leq n, \\ F_{\nu_0}^{-1}(x), & x > x_n. \end{cases}$$

Clearly the set S is countable, and it remains to show that it is dense in $C(\nu_0)$.

Take an arbitrary $\epsilon > 0$ and a measure $\nu \in C(\nu_0)$. As $J(\nu_0, \nu) < +\infty$, the integrals

$$\int_0^\delta g(F_{\nu_0}^{-1} - F_\nu^{-1})\,\mathrm{d}x, \quad \int_{1-\delta}^1 g(F_{\nu_0}^{-1} - F_\nu^{-1})\,\mathrm{d}x$$

can be made smaller than $\epsilon/3$ by choosing a sufficiently small $\delta > 0$. Fix δ at some suitable rational value and choose a sufficiently rich set of rational points $\delta = x_0 < \cdots < x_n = 1 - \delta$ and rational values $q_1 < \cdots < q_n$ such that the corresponding measure μ satisfies $\int_\delta^{1-\delta} g(q_i - F_\nu^{-1})\,\mathrm{d}x < \epsilon/3$. If $q_1 < F_{\nu_0}^{-1}(x_0)$, we replace x_0 with a smaller rational value x_0' such that $q_1 \geq F_0^{-1}(x_0')$ and

$$\int_{x_0'}^{x_0} g(q_1 - F_\nu^{-1})\,\mathrm{d}x \leq \int_{x_0'}^{x_0} g(F_{\nu_0}^{-1} - F_\nu^{-1})\,\mathrm{d}x.$$

The same operation van be performed if necessary in the neighborhood of $x = 1$. We have thus constructed a measure $\mu \in S$ such that $J(\mu, \nu) < \epsilon$. □

Theorem 7. *The space* $(C(\nu_0), \tau_g)$ *is metrizable.*

Proof. We will use Tychonoff's formulation of the Urysohn lemma, which states that a regular topological space with a countable basis is metrizable.

Clearly the T_1 axiom holds for $(C(\nu_0), \tau_g)$, i.e., all one-point sets in $(C(\nu_0), \tau_g)$ are closed. Moreover the closure of a ball $B_r(\nu)$ is given by $\overline{B}_r(\nu) = \{\mu \in C(\nu_0) \colon J(\mu, \nu) \leq r\}$. Thus for any measure $\nu \in C(\nu_0)$ and a ball $B_r(\nu)$ we have $\nu \in B_{r/2}(\nu) \subset \overline{B}_{r/2}(\nu) \subset B_r(\nu)$, which implies that $(C(\nu_0), \tau_g)$ is regular.

Next we construct a countable basis for $(C(\nu_0), \tau_g)$. Take some $\mu^* \in C(\nu_0)$ and consider sequences $\{\mu_n\}_{n\geq 1}$, $\{\nu_n\}_{n\geq 1}$ such that all $\mu_n \in S$, where S was constructed in the proof of the preceding lemma and both $J(\mu^*, \mu_n) \to 0$ and $J(\nu_n, \mu_n) \to 0$.

Repeating the proof of Lemma 5 one can show that $\nu_n \rightharpoonup \mu^*$. Moreover, similarly to the proof of Theorem 6 one can obtain that $J(\nu_n, \mu^*) \to 0$, due to condition (9).

Let us now consider a ball $B_r(\mu^*)$. For a sufficiently small $\epsilon > 0$ and for any measure $\mu \in S$ such that $J(\mu^*, \mu) < \varepsilon$, the argument just given implies that $J(\nu, \mu^*) < r$ as soon as $J(\nu, \mu) < \varepsilon$. Thus for a rational $\epsilon > 0$ we constructed a ball $B_\epsilon(\mu)$ such that $\mu^* \in B_\epsilon(\mu) \subset B_r(\mu^*)$. Therefore the space $(C(\nu_0), \tau_g)$ has a countable basis $\{B_r(\mu)\colon \mu \in S, r \in \mathbb{Q}_+\}$. It now follows from the Urysohn lemma that $(C(\nu_0), \tau_g)$ is metrizable. □

Appendix: Proof of the measurability lemma

Proof of Lemma 2. We have to show that for any $a \in \mathbb{R}$ the set $U_a := K^{-1}((a, +\infty))$ is $\mathcal{B}(\mathcal{P}(\mathbb{R})) \otimes \mathcal{B}((0, 1))$-measurable. Fix some $x \in \mathbb{R}$ and consider the function

$$k_x(\mu) := K(\mu, x) = F_\mu^{-1}(x)$$

on $\mathcal{P}(\mathbb{R})$; we will show first that $k_x^{-1}((a, +\infty)) \subset \mathcal{P}(\mathbb{R})$ is open.

Take an arbitrary measure $\bar\mu \in U_a$ and denote $y := k_x(\bar\mu) = F_{\bar\mu}^{-1}(x) > a$. It suffices to show that there exists an open neighbourhood V of $\bar\mu$ in the weak topology of $\mathcal{P}(\mathbb{R})$ such that $V \subset k_x^{-1}((a, +\infty))$.

The left continuity of $F_{\bar\mu}$ ensures that $F_{\bar\mu}(\frac{a+y}{2}) < x - \delta$ for some $\delta > 0$. Take a continuous function $v\colon \mathbb{R} \to [0, 1]$ such that $v(t) = 1$ for $t \le a$ and $v(t) = 0$ for $t \ge (a + y)/2$. It follows that

$$\int_{\mathbb{R}} v \, \mathrm{d}\bar\mu \le \int_{-\infty}^{\frac{a+y}{2}} \mathrm{d}\bar\mu = F_{\bar\mu}\Big(\frac{a + y}{2}\Big) < x - \delta.$$

On the other hand, for any $\mu \in \mathcal{P}(\mathbb{R})$ such that $F_\mu(a + 0) \ge x$ we have

$$\int_{\mathbb{R}} v \, \mathrm{d}\mu \ge \int_{-\infty}^{a} \mathrm{d}\mu = F_\mu(a + 0) \ge x,$$

which implies $\int_{\mathbb{R}} v \, \mathrm{d}\mu - \int_{\mathbb{R}} v \, \mathrm{d}\bar\mu > \delta$. Therefore for all measures ν in the weak neighborhood

$$V_\delta(\bar\mu) = \Big\{\nu \in \mathcal{P}(\mathbb{R})\colon \Big|\int_{\mathbb{R}} v \, \mathrm{d}\nu - \int_{\mathbb{R}} v \, \mathrm{d}\bar\mu\Big| < \delta\Big\}.$$

it follows that $F_\nu(a + 0) < x$, or $k_x(\nu) > a$. Thus the set $k_x^{-1}((a, +\infty))$ is open in the weak topology of $\mathcal{P}(\mathbb{R})$.

From monotonicity of the inverse cumulative distribution function we obtain that $A \times [x, 1) \subset U_a$ whenever $A \times \{x\} \in U_a$. The left continuity of F_μ implies that for any $(\mu, x) \in U_a$ there exists a rational $s \in (0, x]$ such that $F_\mu^{-1}(s) > a$, whence $(\mu, s) \in U_a$. Therefore

$$U_a = \cup_{x\in(0,1)} k_x^{-1}((a, +\infty)) \times \{x\} = \cup_{s\in(0,1)\cap\mathbb{Q}} k_s^{-1}((a, +\infty)) \times [s, 1).$$

Thus U_a is measurable because it is a countable union of measurable sets, and $K(\cdot, \cdot)$ is measurable. □

References

[1] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011. URL: http://dx.doi.org/10.1137/100805741.

[2] J. Bernoulli. *Ars conjectandi.* Impensis Thurnisiorum, Fratrum, Basileae, 1713. URL: https://books.google.ru/books?id=kD4PAAAAQAAJ.

[3] J. Bigot and T. Klein. Characterization of barycenters in the wasserstein space by averaging optimal transport maps. 2015. URL: http://arXiv.org/abs/1212.2562v5.

[4] E. Boissard, T. L. Gouic, and J.-M. Loubes. Distribution's template estimate with Wasserstein metrics. *Bernouilli*, 21(2):740–759, 2015. URL: https://projecteuclid.org/euclid.bj/1429624959.

[5] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Inst. Henri Poincaré*, 10(4):215–310, 1948. URL: http://archive.numdam.org/ARCHIVE/AIHP/AIHP_1948__10_4/AIHP_1948__10_4_215_0/AIHP_1948__10_4_215_0.pdf.

[6] T. Le Gouic. *Localisation de masse et espaces de Wasserstein*. PhD thesis, Université de Toulouse III-Paul Sabatier, 2013. URL: http://thesesups.ups-tlse.fr/2163/1/2013TOU30236.pdf.

[7] T. Le Gouic and J.-M. Loubes. Existence and consistency of wasserstein barycenters. 2015. URL: http://arXiv.org/abs/1506.04153.

[8] O. Sheynin. *Jacob Bernoulli. On the law of large numbers*, chapter Foreword. NG Verlag, Berlin, 2005. URL: http://www.sheynin.de/download/bernoulli.pdf.