Sets of Probability Measures and Convex Combination Spaces

Miriam Alonso de la Fuente ALONSOFMIRIAM@UNIOVI.ES
Pedro Terán TERANPEDRO@UNIOVI.ES
Universidad de Oviedo, Spain

Abstract

The Wasserstein distances between probability distributions are an important tool in modern probability theory which has been generalized to sets of probability distributions. We will show that the (generalized) L^1-Wasserstein metric, with the operations of convolution and rescaling, fits in the abstract framework of convex combination spaces: nonlinear metric spaces preserving some of the nice properties of a normed space but accomodating other unusual behaviours. For instance, unlike in a linear space, a singleton $\{P\}$ is typically not convex (it is so only if P is degenerate). Also, some theorems for convex combination spaces are applied to this setting.

Keywords: compact sets of probabilities, convolution, credal set, law of large numbers, Wasserstein metric

1. Introduction

The Wasserstein L^p-metrics have become an essential tool for quantifying the disparity between probability measures [14, 9], due in part to their natural connection to optimal transport problems and techniques. They are an object of active interest at the theoretical level [3, 7], in the development of new statistical methods [2, 4] and in applications [5, 15].

Recently, Li and Lin [8] defined a generalized Wasserstein metric between sets of probabilities, rather than individual probability measures. Their immediate interest lies in connecting the convergence of sets of probabilities to that of functionals (sublinear expectations [10]) which can be written as suprema of integrals with respect to those probabilities.

But one can also envisage potential applications in which a set of probability measures represents a statistical model (parametric or nonparametric like, e.g., in robust statistics), a credal set, or is generated by other imprecise probability models (the core of a capacity, the selectionable distributions of a random set, and so on), or arises from decision-theoretical considerations (multiple priors, ambiguity).

Our aim is to show that both the L^1-Wasserstein metric w_1 and its generalization to sets of probability measures fit in the framework of the *convex combination spaces* defined by Terán and Molchanov [12]. These (non-linear) metric spaces provide a generalization of Banach spaces which is both amenable to probability theory and closed under uplifting to compact subsets. That means that, given a convex combination space $\mathbb{E}$, it is always true that the class of its non-empty compact subsets is also a convex combination space when endowed with the Hausdorff metric.

In our context, that will mean that Li and Lin's generalized Wasserstein metric immediately satisfies some known theorems as soon as those are proved for the usual Wasserstein metric, which itself just requires establishing that it defines a convex combination space and applying available results. Therefore it is an interesting path to showing that the generalized Wasserstein metric preserves some properties of the usual Wasserstein metric, without having to devise new proofs.

The basic operation in a convex combination space is the convex combination, whether an algebraic sum exists or not. A convex combination of points x_i with weights λ_i is directly connected to the expectation of a random element taking on values x_i with probabilities λ_i, and is sufficient to study averages and weighted averages of random elements. Moreover, note that elements need not admit an additive inverse since convex combinations only involve non-negative weights. That is essential in order to accommodate spaces of probability distributions since the neutral element is the degenerate distribution δ_0 at 0 but one cannot combine two non-degenerate probabilities to obtain δ_0 (this is only possible for δ_x and δ_{-x}).

The structure of the paper is as follows. Section 2 contains the necessary preliminaries. In Section 3, we prove that the Wasserstein space $W_1(\mathbb{R})$, Li and Lin's generalization, and an even larger space all satisfy the convex combination space axioms. In Section 4, a detailed discussion of the convexification axiom (CC5) below is carried out. Next section presents some consequences, namely versions of the strong law of large numbers, Jensen's inequality and the dominated convergence theorem, in the setting of random probability measures and random sets of probability measures.

2. Preliminaries

Let $(\mathbb{E}, d)$ be a metric space and let $A \subseteq \mathbb{E}$. Then the closure of A will be denoted by $\operatorname{cl} A$, its convex hull will be denoted by $\operatorname{co} A$ and its closed convex hull will be denoted by $\overline{\operatorname{co}}A$.

Denote by $\mathcal{K}(\mathbb{E})$ the space of non empty compact subsets of $\mathbb{E}$.

Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $v \in \mathbb{E}$ be an arbitrary point. A Borel measurable mapping $X : (\Omega, \mathcal{A}, P) \to (\mathbb{E}, d)$ is called *random element*. A random element $X : (\Omega, \mathcal{A}, P) \to \mathbb{E}$ is *integrable* if $d(X, v)$ is an integrable random variable.

Let X_n, X be random elements in $\mathbb{E}$. Then $\{X_n\}_n$ *converges weakly* to X if $E[f(X_n)] \to E[f(X)]$ for every continuous bounded function $f : \mathbb{E} \to \mathbb{R}$.

Let $W_1(\mathbb{R})$ be the set of probability measures in $\mathbb{R}$ with finite expectation. The L^1-norm of a real random variable with finite expectation is defined as

$$\|X\|_1 = E[|X|].$$

Endow $W_1(\mathbb{R})$ with the L^1-Wasserstein metric

$$w_1(P, Q) = \inf_{\mathcal{L}(X)=P, \mathcal{L}(Y)=Q} \|X - Y\|_1,$$

where $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ are the distributions of the random variables X and Y.

Assuming that Ω is a metric space, in [8, p. 2], a *sublinear expectation* is defined as

$$\mathbb{E}^{\mathcal{P}}[\varphi] = \sup_{\mu \in \mathcal{P}} E_\mu[\varphi], \forall \varphi \in C(\Omega),$$

where $\mathcal{P}$ is a set of Borel probability measures, $C(\Omega)$ is the space of continuous functions $\varphi : (\Omega, d) \to \mathbb{R}$ and d is a metric in Ω.

Let $\mathcal{P}$ and $\mathcal{Q}$ be sets of probability measures. The *generalized Wasserstein metric* (see [8, Definition 2.1]) between $\mathcal{P}$ and $\mathcal{Q}$ is

$$\mathcal{W}_1(\mathcal{P}, \mathcal{Q}) = \max\left\{\sup_{P \in \mathcal{P}} \inf_{Q \in \mathcal{Q}} w_1(P, Q), \sup_{Q \in \mathcal{Q}} \inf_{P \in \mathcal{P}} w_1(P, Q)\right\}.$$

Recall that a *weakly compact set* is compact with respect to the weak topology.

Definition 1 *We denote by $\mathcal{P}_1(\mathbb{R})$ the set of all sets $\mathcal{P}$ of probability measures in the real line such that*

(a) $\mathcal{P}$ is weakly compact

(b) For an arbitrary point $r \in \mathbb{R}$,

$$\lim_{K \to \infty} \mathbb{E}^{\mathcal{P}}[d(r, \cdot) I_{\{x \in \mathbb{R} : d(r,x) \geq K\}}] = 0.$$

In [8, Definition 2.4], Li and Lin imposed an additional condition, the convexity of the sets $\mathcal{P}$ with respect to the ordinary operations given by $(aP + bQ)(A) = aP(A) + bQ(A)$. Notice these are not the operations between probability measures that we will consider, which are based on convolution and rescaling. We will denote by $\mathcal{P}_1^c(\mathbb{R})$ the subset of $\mathcal{P}_1(\mathbb{R})$ in Li and Lin's definition (which they denote $\mathcal{P}_1(\mathbb{R})$).

Let K be a set of probability measures. Then K is *tight* if for every $\varepsilon > 0$ there exists a compact set L such that

$$\inf_{\mu \in L} \mu(L) > 1 - \varepsilon.$$

We will use the following characterization of the relatively compact subsets of $W_1(\mathbb{R}^d)$ [9, Proposition 2.2.3].

Lemma 2 *A tight subset $K \subseteq W_1(\mathbb{R}^d)$ has a compact closure in $W_1(\mathbb{R})$ if and only if*

$$\sup_{\mu \in K} \int_{\{x : \|x\| > R\}} \|x\| d\mu(x) \to 0 \qquad (1)$$

as $R \to \infty$.

Another important result related to tight subsets is Prokhorov's theorem, which is stated below (see [6]).

Lemma 3 (Prokhorov's theorem) *Let $(\mathbb{E}, d)$ be a complete and separable metric space and let $\mathcal{P}(\mathbb{E})$ be the set of probability measures defined in $\mathbb{E}$ with its Borel σ-algebra. Let $K \in \mathcal{P}(\mathbb{E})$. Then K is tight if and only if the closure of K in $\mathcal{P}(\mathbb{E})$ is compact.*

Remark 4 *Notice that a w_1-compact subset is always weakly compact and thus tight by Prokhorov's theorem (Lemma 3). Therefore, by Lemma 2 a set is w_1-compact if and only if it is tight, w_1-closed and satisfies (1).*

Definition 5 *[12] Let $(\mathbb{E}, d)$ be a metric space with a* convex combination operation *$[\cdot, \cdot]_{i=1}^n$ which for any $n \geq 2$ numbers $\lambda_1, \ldots \lambda_n > 0$ satisfying $\sum_{i=1}^n \lambda_i = 1$ and any $v_1, \ldots, v_n \in \mathbb{E}$ this operation produces an element of $\mathbb{E}$, denoted $[\lambda_i, v_i]_{i=1}^n$ or $[\lambda_1, v_1; \ldots; \lambda_n, v_n]$. We will say that $\mathbb{E}$ is a* convex combination space *if the following axioms are satisfied:*

(CC1) (Commutativity) For every permutation σ of $\{1, \ldots, n\}$,

$$[\lambda_i, v_i]_{i=1}^n = [\lambda_{\sigma(i)}, v_{\sigma(i)}]_{i=1}^n;$$

(CC2) (Associativity)

$$[\lambda_i, v_i]_{i=1}^{n+2} = [\lambda_1, v_1; \ldots; \lambda_n, v_n; \lambda_{n+1} + \lambda_{n+2}, [\frac{\lambda_{n+j}}{\lambda_{n+1} + \lambda_{n+2}}; v_{n+j}]_{j=1}^2];$$

(CC3) *(Continuity) If $u, v \in \mathbb{E}$ and $\lambda^{(k)} \to \lambda \in (0,1)$, then*

$$[\lambda^{(k)}, u; 1-\lambda^{(k)}, v] \to [\lambda, u; 1-\lambda, v];$$

(CC4) *(Negative curvature) For all $u_1, u_2, v_1, v_2 \in \mathbb{E}$ and $\lambda \in (0,1)$,*

$$\begin{aligned} &d([\lambda, u_1; 1-\lambda, u_2], [\lambda, v_1; 1-\lambda, v_2]) \\ &\quad \leq \lambda d(u_1, v_1) + (1-\lambda) d(u_2, v_2); \end{aligned}$$

(CC5) *(Convexification) For each $v \in \mathbb{E}$, there exists $\lim_{n\to\infty}[n^{-1}, v]_{i=1}^n$, which will be denoted by $\mathbf{K}_{\mathbb{E}}(v)$. The mapping $\mathbf{K}_{\mathbb{E}}$ is called the* convexification operator *of* $\mathbb{E}$.

The *barycenter* of a probability measure P in the real line (i.e., the expectation of a random variable whose distribution is P) will be denoted by $b(P)$.

3. Convex Combinations Based on Convolution

In this section, we will show that Li and Lin's space (with the convexity requirement or removing it) fits into the framework of convex combination spaces when the operations on probability measures are convolution and rescaling. The path to that result is as follows.

(1) The Wasserstein space $W_1(\mathbb{R})$ is a convex combination space.

(2) $\mathcal{P}_1(\mathbb{R})$ is exactly $\mathcal{K}(W_1(\mathbb{R}))$.

(3) $\mathcal{P}_1(\mathbb{R})$ is a convex combination space.

(4) $\mathcal{P}_1^c(\mathbb{R}) \subseteq \mathcal{P}_1(\mathbb{R})$ is a convex combination space.

Notice that the intermediate steps have independent interest, as $W_1(\mathbb{R})$ is an important space in modern probability theory and $\mathcal{P}_1(\mathbb{R})$ removes the convexity assumption in the definition of $\mathcal{P}_1^c(\mathbb{R})$.

First, we have to show that the space of probability distributions with finite mean is a convex combination space. To that end, we define convex combinations using convolution and rescaling. The convolution of P and Q is the distribution of $X+Y$ where X, Y are independent and have distribution P, Q respectively. If P, Q are absolutely continuous then the density function of the convolution is the convolution of their density functions given by

$$f(x) = \int_{-\infty}^{\infty} f_X(x-y) f_Y(y) dy.$$

Rescaling P by a factor a means taking the distribution of aX for a random variable X with distribution P.

Theorem 6 *$(W_1(\mathbb{R}), w_1)$ is a convex combination space with the convex combination operation*

$$[\lambda_i, P_i]_{i=1}^n = \mathcal{L}(\sum_{i=1}^n \lambda_i X_i)$$

where X_i are independent random variables with distribution P_i, respectively. The convexification operator is $\mathbf{K}_{W_1(\mathbb{R})}(P) = \delta_{b(P)}$.

Proof Properties (CC1) and (CC2) are consequences of the commutativity and associativity of the sum and product in $\mathbb{R}$.

(CC3) Let $P, Q \in W_1(\mathbb{R})$ and let $\lambda^{(k)} \to \lambda \in (0,1)$. Then $\lambda^{(k)} X_1 + (1-\lambda^{(k)}) X_2 \to \lambda X_1 + (1-\lambda) X_2$.

(CC4) Let $P_1, P_2, Q_1, Q_2 \in W_1(\mathbb{R})$. Then

$$\begin{aligned} &w_1([\lambda, P_1; (1-\lambda), P_2], [\lambda, Q_1; (1-\lambda), Q_2]) \\ &= \inf_{\substack{\mathcal{L}(X)=[\lambda,P_1;(1-\lambda),P_2], \\ \mathcal{L}(Y)=[\lambda,Q_1;(1-\lambda),Q_2]}} \|X-Y\|_1 \\ &\leq \inf_{\substack{\mathcal{L}(X_1)=P_1,\mathcal{L}(X_2)=P_2, \\ \mathcal{L}(Y_1)=Q_1,\mathcal{L}(Y_2)=Q_2}} \left\| \begin{matrix} (\lambda X_1 + (1-\lambda)X_2) \\ -(\lambda Y_1 + (1-\lambda)Y_2) \end{matrix} \right\|_1 \\ &= \inf_{\substack{\mathcal{L}(X_1)=P_1,\mathcal{L}(X_2)=P_2, \\ \mathcal{L}(Y_1)=Q_1,\mathcal{L}(Y_2)=Q_2}} \left\| \begin{matrix} (\lambda X_1 - \lambda Y_1) \\ +((1-\lambda)X_2 - (1-\lambda)Y_2) \end{matrix} \right\|_1 \\ &= \inf_{\substack{\mathcal{L}(X_1)=P_1,\mathcal{L}(X_2)=P_2, \\ \mathcal{L}(Y_1)=Q_1,\mathcal{L}(Y_2)=Q_2}} \left\| \begin{matrix} \lambda(X_1 - Y_1) \\ +(1-\lambda)(X_2 - Y_2) \end{matrix} \right\|_1 \\ &\leq \inf_{\substack{\mathcal{L}(X_1)=P_1,\mathcal{L}(X_2)=P_2, \\ \mathcal{L}(Y_1)=Q_1,\mathcal{L}(Y_2)=Q_2}} \left(\begin{matrix} \lambda\|X_1 - Y_1\|_1 \\ +(1-\lambda)\|X_2 - Y_2\|_1 \end{matrix} \right) \\ &= \inf_{\mathcal{L}(X_1)=P_1,\mathcal{L}(Y_1)=Q_1} \lambda\|X_1 - Y_1\|_1 \\ &\quad + \inf_{\mathcal{L}(X_2)=P_2,\mathcal{L}(Y_2)=Q_2} (1-\lambda)\|X_2 - Y_2\|_1 \\ &= \lambda \inf_{\mathcal{L}(X_1)=P_1,\mathcal{L}(Y_1)=Q_1} \|X_1 - Y_1\|_1 \\ &\quad + (1-\lambda) \inf_{\mathcal{L}(X_2)=P_2,\mathcal{L}(Y_2)=Q_2} \|X_2 - Y_2\|_1 \\ &= \lambda w_1(P_1, Q_1) + (1-\lambda) w_1(P_2, Q_2). \end{aligned}$$

(CC5) Notice that convergence in the metric w_1 is equivalent to weak convergence and convergence of the first moment (see Theorem 7.12 in [13]). Let $\{X_n\}_n$ be an independent sequence of random variables with distribution P. Then $[n^{-1}, P]_{i=1}^n$ is the distribution of $n^{-1}\sum_{i=1}^n X_i$, which converges to $E(X_1)$ almost surely by the strong law of large numbers. In particular, convergence in distribution holds and thus

$$[n^{-1}, P]_{i=1}^n \to \delta_{E(X_1)} = \delta_{b(P)}$$

weakly. In its turn, convergence of the first moment is trivial since

$$E(n^{-1}\sum_{i=1}^{n} X_i) = E(X_1) = b(P) = b(\delta_{b(P)}).$$

■

We will characterize $\mathcal{P}_1(\mathbb{R})$ now as the set of all compact subsets of $(W_1(\mathbb{R}), w_1)$.

Proposition 7

$$\mathcal{P}_1(\mathbb{R}) = \mathcal{K}(W_1(\mathbb{R}))$$

Proof First, let us show that $\mathcal{P}_1(\mathbb{R}) \subseteq \mathcal{K}(W_1(\mathbb{R}))$, that is, every set of probabilities $\mathcal{P}$ verifying (a) and (b) is a w_1-compact set of integrable probabilities. Let $\mathcal{P}$ be a singleton $\{P\}$. Then, by condition (b), $\{P\}$ is uniformly integrable, hence integrable.

Moreover, let $\mathcal{P} \in \mathcal{P}_1(\mathbb{R})$, which is weakly compact by hypothesis. Since $(W_1(\mathbb{R}), w_1)$ is complete and separable (see Proposition 2.2.8 and Theorem 2.2.7 in [9]), by Prokhorov's theorem (Lemma 3) $\mathcal{P}$ is tight. Since condition (b) in Definition 1 is satisfied, by Lemma 2, $\mathcal{P}$ has compact closure in the Wasserstein metric.

There remains to show that $\mathcal{P}$ is actually closed in w_1. Let $\{P_n\}_n \subseteq \mathcal{P}$ be a convergent sequence to some P in w_1. Since the weak topology is weaker than the topology induced by the Wasserstein metric, the sequence $\{P_n\}_n$ converges to P in the weak topology. Then, since $\mathcal{P}$ is compact in the weak topology, it is closed, hence $P \in \mathcal{P}$. In conclusion, $\mathcal{P}$ is closed in the Wasserstein metric, so it is compact.

Next, we have to show $\mathcal{K}(W_1(\mathbb{R})) \subseteq \mathcal{P}_1$, that is, every compact subset of integrable probabilities satisfies conditions (a) and (b). Condition (a) is inmediate, since every w_1-compact set is weakly compact. For (b), Lemma 2 ensures that condition (b) is satisfied, since the equivalence between tightness and compact closure in $W_1(\mathbb{R})$ is true only if the sequences are uniformly integrable. ■

In [12, Theorem 6.2], it was proven that the property of $(\mathbb{E}, d)$ being a convex combination space is inherited by $\mathcal{K}(\mathbb{E})$ endowed with the Hausdorff metric

$$d_H(K, L) = \max\{\sup_{x\in K}\inf_{y\in L} d(x, y), \sup_{y\in L}\inf_{x\in K} d(x, y)\}.$$

As a consequence, we obtain the following result.

Theorem 8 *$(\mathcal{P}_1(\mathbb{R}), \mathcal{W}_1)$ is a convex combination space with the convex combination operation*

$$[\lambda_i, \mathcal{P}_i]_{i=1}^{n} = \{\mathcal{L}(\sum_{i=1}^{n} \lambda_i X_i) \mid \mathcal{L}(X_i) \in \mathcal{P}_i,$$
$$X_i \textit{ independent}, i \in \{1, \ldots, n\}\}$$

and the convexification operator $\mathbf{K}_{\mathcal{K}(W_1(\mathbb{R}))} = \overline{\text{co}} \circ \mathbf{K}_{W_1(\mathbb{R})}$.

Proof By Theorem 6.2 in [12], that convex combination operation is well defined and $\mathcal{P}_1(\mathbb{R})$ becomes a convex combination operation when endowed the Hausdorff metric. But that is just the generalized Wasserstein metric $\mathcal{W}_1$ defined by Lin and Li. By the same result, the convexification operator is the composition of the closed convex hull and the convexification operator in the underlying space. ■

Finally, we will complete step (4) above.

Theorem 9 *$\mathcal{P}_1^c(\mathbb{R})$ is a convex combination space with the operations and $\mathcal{W}_1$-metric inherited from $\mathcal{P}_1(\mathbb{R})$.*

Proof Since $\mathcal{P}_1^c(\mathbb{R})$ is a subset of $\mathcal{P}_1(\mathbb{R})$, properties (CC1) through (CC5) of the latter will ensure that $\mathcal{P}_1^c(\mathbb{R})$ is a convex combination space, as soon as we prove that the convex combination of elements in $\mathcal{P}_1^c(\mathbb{R})$ is in $\mathcal{P}_1^c(\mathbb{R})$. Moreover, due to the associativity property (CC2), it suffices to prove it for the convex combination of two elements, as any larger convex combination can iteratively be reduced to convex combinations of two elements.

Let $\lambda \in (0, 1)$ and $\mathcal{P}, \mathcal{Q} \in \mathcal{P}_1^c(\mathbb{R})$. We need to show $[\lambda, \mathcal{P}; 1-\lambda, \mathcal{Q}] \in \mathcal{P}_1^c(\mathbb{R})$, i.e., $[\lambda, \mathcal{P}; 1-\lambda, \mathcal{Q}]$ is convex in the sense used in [8]. Specifically, we will show that, whenever $p \in (0, 1)$, $P, P' \in \mathcal{P}$ and $Q, Q' \in \mathcal{Q}$, the probability measure $p \cdot [\lambda, P; 1-\lambda, Q] + (1-p) \cdot [\lambda, P'; 1-\lambda, Q']$ is in $[\lambda, \mathcal{P}; 1-\lambda, \mathcal{Q}]$. As before, the general case follows by iterating convex combinations of two elements.

Let X, X', Y, Y' be independent random variables whose distributions, respectively, are P, P', Q, Q'. While they might possibly be defined on different sample spaces $\Omega_X, \Omega_{X'}, \Omega_Y, \Omega_{Y'}$, without loss of generality we may assume that this is not the case. Indeed, otherwise we define $\Omega = \Omega_X \times \Omega_{X'} \times \Omega_Y \times \Omega_{Y'}$ and replace X, X', Y, Y' by random variables with the same distribution given by $\hat{X} : (\omega_1, \omega_2, \omega_3, \omega_4) \in \Omega \mapsto X(\omega_1)$ and so on.

Consider a random variable $\xi : \Omega \times [0, 1] \to \mathbb{R}$ defined in the product probability space $\Omega \times [0, 1]$ with the probability measure $\mathbb{P}$ being the product of P and the uniform distribution in $[0, 1]$,

$$\xi(\omega, t) = \begin{cases} \lambda X(\omega) + (1-\lambda)Y(\omega), & t \in [0, p] \\ \lambda X'(\omega) + (1-\lambda)Y'(\omega), & t \in (p, 1] \end{cases}$$

Since, for any Borel set $A \subseteq R$,

$$\mathbb{P}(\xi \in A) = p \cdot P(\lambda X + (1-\lambda)Y \in A)$$
$$+(1-p) \cdot P(\lambda X + (1-\lambda)Y \in A)$$

and

$$P_{\lambda X+(1-\lambda)Y} = [\lambda, P_X; 1-\lambda, P_Y] = [\lambda, P; 1-\lambda, Q],$$

$$P_{\lambda X'+(1-\lambda)Y'} = [\lambda, P_{X'}; 1-\lambda, P_{Y'}] = [\lambda, P'; 1-\lambda, Q'],$$

we have

$$\mathcal{L}(\xi) = p \cdot [\lambda, P; 1-\lambda, Q] + (1-p) \cdot [\lambda, P'; 1-\lambda, Q'].$$

There remains to prove $\mathcal{L}(\xi) \in [\lambda, \mathcal{P}; 1-\lambda, \mathcal{Q}]$. For that purpose, notice $\xi = \lambda\xi_1 + (1-\lambda)\xi_2$ where

$$\xi_1(\omega, t) = \begin{cases} X(\omega), & t \in [0,p] \\ X'(\omega), & t \in (p,1] \end{cases}$$

and

$$\xi_2(\omega, t) = \begin{cases} Y(\omega), & t \in [0,p] \\ Y'(\omega), & t \in (p,1]. \end{cases}$$

The distributions of ξ_1 and ξ_2 are the mixtures $p \cdot P + (1-p) \cdot P'$ and $p \cdot Q + (1-p) \cdot Q'$. Since $P, P' \in \mathcal{P} \in \mathcal{P}_1^c(\mathbb{R})$, we have $\mathcal{L}(\xi_1) \in \mathcal{P}$. Similarly, $\mathcal{L}(\xi_2) \in \mathcal{Q}$. Therefore

$$\begin{aligned} \mathcal{L}(\xi) &= \mathbb{P}_{\lambda\xi_1 + (1-\lambda)\xi_2} \\ &= [\lambda, \mathcal{L}(\xi_1); 1-\lambda, \mathcal{L}(\xi_2)] \in [\lambda, \mathcal{P}; 1-\lambda, \mathcal{Q}] \end{aligned}$$

as wished. ■

4. On Property (CC5)

Convexification property (CC5) is trivially satisfied in a linear space, while convolution is not a group and moreover rescaling a distribution by the factor −1 does not provide its additive inverse. It is therefore interesting to study this property more specifically in the spaces $W_1(\mathbb{R})$ and $\mathcal{P}_1(\mathbb{R})$.

Proposition 10 *Let $P \in W_1(\mathbb{R})$. Then $\{P\}$ is convex if and only if P is a degenerate distribution.*

Proof By [12, Proposition 3.2], $\{P\}$ is convex if and only if $P = \mathbf{K}_{W_1(\mathbb{R})}(Q)$ for some Q, which by (CC5) implies $P = \delta_{b(Q)}$.

For the converse, notice $P = \delta_x$ for some $x \in \mathbb{R}$ implies every convex combination $[\lambda_i, P]_{i=1}^n$ is the distribution of the random variable $\sum_{i=1}^n \lambda_i x = x$, i.e., $[\lambda_i, P]_{i=1}^n = P$. ■

Remark 11 *If P has a finite variance σ^2, an alternative argument goes by noticing that $[1/2, P; 1/2, P]$ has variance $\sigma^2/2$, which is impossible unless P is degenerate.*

Let us present an explicit expression of the convexification operator in $\mathcal{P}_1(\mathbb{R})$. To that end, given $\mathcal{P} \in \mathcal{P}_1(\mathbb{R})$ we define

$$\underline{b}(\mathcal{P}) = \inf_{P\in\mathcal{P}} b(P) \quad \textit{(lower barycenter)},$$

$$\overline{b}(\mathcal{P}) = \sup_{P\in\mathcal{P}} b(P) \quad \textit{(upper barycenter)}.$$

Notice these are just the lower and upper expectation defined by all random variables whose distribution is in $\mathcal{P}$.

Proposition 12 *Let $\mathcal{P} \in \mathcal{P}_1(\mathbb{R})$. Then*

$$\mathbf{K}_{\mathcal{P}_1(\mathbb{R})}(\mathcal{P}) = \{\delta_x \mid x \in [\underline{b}(\mathcal{P}), \overline{b}(\mathcal{P})]\}.$$

Proof By Proposition 7, $\mathcal{P}_1(\mathbb{R}) = \mathcal{K}(W_1(\mathbb{R}))$. Therefore, using Theorems 8 and 6,

$$\mathbf{K}_{\mathcal{P}_1(\mathbb{R})}(\mathcal{P}) = \overline{\operatorname{co}}\,\mathbf{K}_{W_1(\mathbb{R})}(\mathcal{P}) = \overline{\operatorname{co}}\,\{\delta_{b(P)} \mid P \in \mathcal{P}\}.$$

The convex hull of $\{\delta_{b(P)} \mid P \in \mathcal{P}\}$ is formed by that set together with all convex combinations $[\lambda_i, \delta_{b(P_i)}]_{i=1}^n = \delta_{\sum_{i=1}^n \lambda_i b(P_i)}$ with $P_i \in \mathcal{P}$. It is therefore the set $\{\delta_x \mid x \in \operatorname{co}\{b(P) \mid P \in \mathcal{P}\}\}$.

From the fact that a sequence of degenerate distributions cannot w_1-converge to a non-degenerate distribution, it is not hard to show that for any convex set $A \subseteq \mathbb{R}$, the w_1-closure of the set $\{\delta_x \mid x \in A\}$ is $\{\delta_x \mid x \in \operatorname{cl} A\}$. Accordingly,

$$\mathbf{K}_{\mathcal{P}_1(\mathbb{R})}(\mathcal{P}) = \{\delta_x \mid x \in \overline{\operatorname{co}}\,\{b(P) \mid P \in \mathcal{P}\}\}.$$

The set $\{b(P) \mid P \in \mathcal{P}\}$ is bounded due to the weak compactness of $\mathcal{P}$. Therefore its closed convex hull is a compact interval, and it follows from the definition that its endpoints are $\underline{b}(\mathcal{P})$ and $\overline{b}(\mathcal{P})$. ■

The set of all independent sequences of random variables X_n whose distributions are in a set $\mathcal{P}$ will be denoted by $\mathcal{X}(\mathcal{P})$. Also, whenever $P \in W_1(\mathbb{R})$ and $\mathcal{Q} \subseteq W_1(\mathbb{R})$, we will denote by $d(P, \mathcal{Q})$ the w_1-distance from P to $\mathcal{Q}$,

$$d(P, \mathcal{Q}) = \inf_{Q\in\mathcal{Q}} w_1(P, Q).$$

From the fact that $\mathcal{P}_1(\mathbb{R})$ is a convex combination space we obtain the following consequence of (CC5).

Theorem 13 *Let $\mathcal{P} \in \mathcal{P}_1(\mathbb{R})$. Then*

$$d(\mathcal{L}(n^{-1}\sum_{i=1}^n X_i), \{\delta_x \mid x \in [\underline{b}(\mathcal{P}), \overline{b}(\mathcal{P})]\}) \to 0$$

uniformly over all independent sequences $\{X_n\}_n$ such that $\mathcal{L}(X_n) \in \mathcal{P}$ for all $n \in \mathbb{N}$.

Proof Combining the preceding results, $[n^{-1}, \mathcal{P}]_{i=1}^n$ converges in the generalized Wasserstein metric $\mathcal{W}_1$ to the set $\{\delta_x \mid x \in [\underline{b}(\mathcal{P}), \overline{b}(\mathcal{P})]\}$. Recall $\mathcal{W}_1$ is the Hausdorff metric between elements of $\mathcal{K}(W_1(\mathbb{R})) = \mathcal{P}_1(\mathbb{R})$. In particular,

$$\sup_{P\in[n^{-1},\mathcal{P}]_{i=1}^n} \inf_{Q\in\{\delta_x \mid x\in[\underline{b}(\mathcal{P}),\overline{b}(\mathcal{P})]\}} w_1(P, Q) \to 0,$$

equivalently

$$\sup_{P_1,\ldots,P_n\in\mathcal{P}} d([n^{-1}, P_i]_{i=1}^n, \{\delta_x \mid x \in [\underline{b}(\mathcal{P}), \overline{b}(\mathcal{P})]\})$$

$$= \sup_{P_1,\dots,P_n\in\mathcal{P}} \inf_{x\in[\underline{b}(\mathcal{P}),\overline{b}(\mathcal{P})]} w_1([n^{-1},P_i]_{i=1}^n,\delta_x) \to 0.$$

Since every $\{X_n\}_n \in \mathcal{X}(\mathcal{P})$ satisfies $\mathcal{L}(n^{-1}\sum_{i=1}^n X_i) = [n^{-1},\mathcal{L}(X_i)]_{i=1}^n$ and $\mathcal{L}(X_i)\in\mathcal{P}$, we deduce

$$\sup_{\{X_n\}_n\in\mathcal{X}(\mathcal{P})} d(\mathcal{L}(n^{-1}\sum_{i=1}^n X_i), \{\delta_x \mid x\in[\underline{b}(\mathcal{P}),\overline{b}(\mathcal{P})]\}) \to 0.$$

That means the uniform convergence on the statement holds. ■

When $\mathcal{P}$ is a singleton, that reduces to $w_1(\mathcal{L}(n^{-1}\sum_{i=1}^n X_i),\delta_{b(P)}) \to 0$, which is equivalent to the weak law of large numbers. In that sense, Proposition 13 can be interpreted as a generalization of the law of large numbers to sets of probability measures. It shows that the distributions of sample averages of independent random variables (non-identically) distributed according to $\mathcal{P}$ eventually tend to be concentrated arbitrarily close to the set of means of $\mathcal{P}$. Moreover, the Wasserstein metric provides a quantification of the disparity for which that convergence is uniform across all possible distribution choices for the sequence.

It is important to note that the distributions $\mathcal{L}(n^{-1}\sum_{i=1}^n X_i)$ may not converge to a point in $[\underline{b}(\mathcal{P}),\overline{b}(\mathcal{P})]$. Indeed, if X_1 is distributed as P, the next 10 variables are distributed as Q, the next 100 as P, the next 1000 as Q, and so on, the distributions of $n^{-1}\sum_{i=1}^n X_i$ will oscillate between $b(P)$ and $b(Q)$ without converging (provided $b(P)\neq b(Q)$). Thus the claim in Theorem 13 that $\mathcal{L}(n^{-1}\sum_{i=1}^n X_i)$ will eventually be confined very close to that 'limit set' of degenerate distributions is the best that can be said.

It can also be reasoned, from the other part of the $\mathcal{W}_1$-convergence, that the limit set $[\underline{b}(\mathcal{P}),\overline{b}(\mathcal{P})]$ is optimal in that each of its points is approached by the sample averages of some appropriate sequence $\{X_n\}_n$, and that happens for all of them uniformly in w_1.

In the next section, among other results we apply the law of large numbers for convex combination spaces which adds another layer of complexity by replacing, in the condition $\mathcal{L}(X_i)\in\mathcal{P}$, the fixed $\mathcal{P}$ by randomly chosen $\mathcal{P}_i$.

5. Some Consequences

Known theorems for convex combination spaces apply in particular to $\mathcal{P}_1(\mathbb{R}), \mathcal{P}_1^c(\mathbb{R})$ and $W_1(\mathbb{R})$. This section illustrates some of the possibilities. For space reasons we focus on $\mathcal{P}_1(\mathbb{R})$ which is the most general of them, leaving the particularizations to the reader.

Since $W_1(\mathbb{R})$ is a complete separable metric space by [9, Theorem 2.2.7 and Proposition 2.2.8], and $\mathcal{W}_1$ is the Hausdorff metric in $\mathcal{P}_1(\mathbb{R}) = \mathcal{K}(W_1(\mathbb{R}))$, the latter is complete and separable as well. Thus the integration theory for convex combination spaces in [12] applies, allowing one to define a notion of expectation in $\mathcal{P}_1(\mathbb{R})$.

In order to do so, one considers as *random elements* of $\mathcal{P}_1(\mathbb{R})$ the Borel measurable mappings from a probability space $(\Omega,\mathcal{A},\mathbb{P})$ to $\mathcal{P}_1(\mathbb{R})$ (i.e., the preimage of each ρ-open set is measurable). These are random (w_1-compact) sets of probability measures.

The expectation of a simple random element $\Gamma = \sum_{i=1}^n I_{\Omega_i}\cdot \mathcal{Q}_i$ (where $\{\Omega_i\}_{i=1}^n$ is a partition of Ω) is defined to be $E(\Gamma) = [\mathbb{P}(\Omega_i), \mathbf{K}_{\mathcal{P}_1(\mathbb{R})}(\mathcal{Q}_i)]_{i=1}^n$. The rationale of applying the convexification operator is to achieve invariance under partition refinements (an unnecessary step if every point in the space were convex). Taking into account Proposition 12 above,

$$E(\Gamma) = [\mathbb{P}(\Omega_i), \{\delta_x \mid x\in[\underline{b}(\mathcal{Q}_i),\overline{b}(\mathcal{Q}_i)]\}]$$

$$= \{\delta_x \mid x\in[\sum_{i=1}^n \mathbb{P}(\Omega_i)\underline{b}(\mathcal{Q}_i), \sum_{i=1}^n \mathbb{P}(\Omega_i)\overline{b}(\mathcal{Q}_i)]\}.$$

The expectation in the general case is well defined using approximation by simple functions. Integrable random elements are those whose expected distance from an arbitrary element is finite, i.e.,

$$E(\mathcal{W}_1(\Gamma,\{\delta_0\})) = E(\sup_{P\in\Gamma} w_1(P,\delta_0))$$

$$= E(\sup_{P\in\Gamma}\inf_{\mathcal{L}(X)=P} E|X|) < \infty.$$

The strong law of large numbers is then a generalization for random elements of property (CC5) of non-random elements.

Theorem 14 *Let Γ be an integrable random element of $\mathcal{P}_1(\mathbb{R})$. Let $\{\Gamma_n\}_n$ be pairwise independent random elements of $\mathcal{P}_1(\mathbb{R})$ identically distributed as Γ. Then*

$$\mathcal{W}_1([n^{-1};\Gamma_i]_{i=1}^n, E(\Gamma)) \to 0$$

almost surely.

Accordingly, for almost every $\omega\in\Omega$ the set of distributions

$$\{\mathcal{L}(n^{-1}\sum_{i=1}^n X_i) \mid X_i \textit{ has a distribution}$$
$$\textit{in } \Gamma_i(\omega), X_i \textit{ independent}\}$$

$\mathcal{W}_1$-converges to the set of degenerate distributions $E(\Gamma)$ which is independent of ω.

Proof Apply [12, Theorem 5.1]. ■

In the case of singletons, which can be identified with elements of $W_1(\mathbb{R})$, the distribution $\mathcal{L}(n^{-1}\sum_{i=1}^n X_i)$ is shown to converge to a deterministic limit independent of ω while the distributions of the X_n are chosen randomly in a way that depends on ω. That is, $\Gamma_i(\omega)$ is a set of distributions from which a probability measure P_i is taken and then X_i is a random variable, on a different probability space, whose distribution is P_i.

Another classical result that extends to this setting is Jensen's inequality.

Theorem 15 *Let $\varphi : \mathcal{P}_1(\mathbb{R}) \to \mathbb{R}$ be a lower semicontinuous function, i.e.,*

$$\mathcal{W}_1(\mathcal{Q}_n, \mathcal{Q}) \to 0 \Rightarrow \liminf_n \varphi(\mathcal{Q}_n) \geq \varphi(\mathcal{Q}),$$

and midpoint convex, i.e., such that

$$\varphi([1/2, \mathcal{P}; 1/2, \mathcal{Q}]) \leq \frac{\varphi(\mathcal{P}) + \varphi(\mathcal{Q})}{2}$$

for all $\mathcal{P}, \mathcal{Q} \in \mathcal{P}_1(\mathbb{R})$. Let Γ be an integrable random element of $\mathcal{P}_1(\mathbb{R})$ such that $E(\varphi(\Gamma)) < \infty$. Then

$$\varphi(E(\Gamma)) \leq E(\varphi(\Gamma)).$$

Proof This is an application of [11, Theorem 3.1]. ∎

A similar result holds also for $W_1(\mathbb{R})$ (since it is a complete separable convex combination space) but it would not follow trivially from Theorem 15 due to the necessity to extend φ from $W_1(\mathbb{R})$ to $\mathcal{P}_1(\mathbb{R})$.

Finally, we present a dominated convergence theorem under weak convergence. For similar results obtaining convergence in more general quasimetrics, the reader is referred to [1].

Theorem 16 *Let Γ_n, Γ be random elements of $\mathcal{P}_1(\mathbb{R})$ such that*

$$\mathcal{W}_1(\Gamma_n, \{\delta_0\}) \leq g$$

for some $g \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. If $\Gamma_n \to \Gamma$ weakly then

$$\mathcal{W}_1(E(\Gamma_n), E(\Gamma)) \to 0.$$

Proof It follows from [1, Corollary 5.2]. ∎

Notice weak convergence (a generalization to metric spaces of convergence in distribution) is weaker than convergence in probability and almost sure convergence, the usual assumptions in the dominated convergence theorem.

Due to the definition of the convex combination via convolution and rescaling, these extensions of classical results can be restated as properties of averages of independent random variables whose distributions belong to randomly chosen sets of probability measures.

6. Concluding Remarks

First, in Section 3, we have shown that the spaces $W_1(\mathbb{R})$, $\mathcal{P}_1(\mathbb{R})$ and $\mathcal{P}_1^c(\mathbb{R})$ are convex combination spaces when endowed with the Wasserstein metric and its generalization to sets of probability measures. In the same section, we have shown that the space $\mathcal{P}_1(\mathbb{R})$ is exactly the space of compact sets of probability measures with finite expectation. In addition, we have identified the convexification operator on these spaces, which has a simple expression (see Sections 3 and 4). Furthermore, it has been seen that $W_1(\mathbb{R})$ and $\mathcal{P}_1(\mathbb{R})$ have the same properties due to the fact that they both fit within the framework of convex combination spaces. This shows that there are properties of $\mathcal{P}_1(\mathbb{R})$ that need not be constructed as if $\mathcal{P}_1(\mathbb{R})$ were superior to $W_1(\mathbb{R})$, which is what has been done in [8]. Finally, in Section 5, we have extended some classical results to this setting using already existing theorems for convex combination spaces.

As a possible future line of research, the relationships between the strong laws of large numbers for random sets and the one we have shown here (Theorem 14) could be studied.

Acknowledgments

Research in this paper was supported by grants from Spain (PID2019-104486GB-I00) and the Principality of Asturias (SV-PA-21-AYUD/2021/50897, PA-21-PF-BP20-112).

Author Contributions

Both authors contributed equally to the manuscript. Section 4 was prepared and written by P. Terán.

References

[1] Miriam Alonso de la Fuente and Pedro Terán. Convergence theorems for random elements in convex combination spaces. *Fuzzy Sets and Systems*, 458: 69–93, 2023.

[2] Pedro César Álvarez-Esteban, Eustasio del Barrio, Juan Antonio Cuesta-Albertos, and Carlos Matrán. Trimmed comparison of distributions. *J. Amer. Statist. Assoc.*, 103:697–704, 2008.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savare. *Gradient flows in metric spaces and in the Wasserstein space of probability measures*. Birkhaüser, 2005.

[4] Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *J. Amer. Statist. Assoc., to appear*.

[5] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Math. Programming*, 171: 115–166, 2018.

[6] Olav Kallenberg. *Foundations of Modern Probability*. Springer, second edition, 2002.

[7] Young-Heon Kim and Brendan Pass. Wasserstein barycenters over Riemannian manifolds. *Adv. Math.*, 307:640–683, 2017.

[8] Xinpeng Li and Yiqing Lin. Generalized Wasserstein distance and weak convergence of sublinear expectations. *J. Theoret. Probab.*, 30:581–593, 2017.

[9] Victor M. Panaretos and Yoav Zemel. *An invitation to Statistics in Wasserstein space*. Springer, 2020.

[10] Shige Peng. *Nonlinear expectations and stochastic calculus under uncertainty*. Springer, 2019.

[11] Pedro Terán. Jensen's inequality for random elements in metric spaces and some applications. *J. Math. Anal. Appl.*, 414:756–766, 2014.

[12] Pedro Terán and Ilya Molchanov. The law of large numbers in a metric space with a convex combination operation. *J. Theoret. Probab.*, 9:875–898, 2006.

[13] Cédric Villani. *Topics in Optimal Transport*. American Mathematical Society, 2003.

[14] Cédric Villani. *Optimal Transport: Old and New*. Springer, 2008.

[15] Yuan Zhao and Xinchang Zhang. Calculating spatial configurational entropy of a landscape mosaic based on the Wasserstein metric. *Landscape Ecol.*, 34:1849–1858, 2019.