Submitted to the Annals of Probability

SIMILARITY OF PROBABILITY MEASURES THROUGH TRIMMING. *

By Pedro C. Álvarez-Esteban, Eustasio del Barrio, Juan A. Cuesta-Albertos and Carlos Matrán.

Universidad de Cantabria and Universidad de Valladolid

Similarity of probability measures is considered in the multivariate setting through a distance between trimmed probabilities. The Wasserstein distance will be our choice, considering the best approximation between a fixed probability and trimmed versions of the other. We allow the possibility of equal or different trimming patterns on both probabilities. We show that trimmings of arbitrary probabilities can be parameterized in terms of the trimmings of a fixed probability, a fact we exploit through useful Skorohod representations.

Best trimmed approximations naturally lead to a Mass Transportation Problem where a part of the mass could be not necessarily transported. Since optimal transportation plans are not easily computable, we provide theoretical support for Monte-Carlo approximations, through a general consistency result. As a remarkable and unexpected additional result, with important implications for future work, we obtain uniqueness of the optimal solution.

1. Introduction.

An analysis of similarity of distributions based on the comparison of their trimmed versions has recently been introduced in Álvarez-Esteban et al [1]. The novelty of such approach consists in considering that two distributions are similar at level α whenever suitable chosen α-trimmed versions of such distributions coincide. Our proposal focused on probability measures on the real line, using the L_2-Wasserstein metric to measure distances between probabilities, and the same trimming pattern on both probabilities. In this work we will treat the problem in greater generality, considering probabilities on R^k and different kinds of trimming schemes.

Trimming probabilities is a frequent practice in Statistics, involving the sample or the population distributions in problems related to robust procedures. In the general setting considered here its explicit use (through trim-

[*] Research partially supported by the Spanish Ministerio de Educación y Ciencia and FEDER, grant BFM2005-04430-C02-01, and 02 and by the Consejería de Educación y Cultura de la Junta de Castilla y León, grant PAPIJCL VA102A06.

AMS 2000 subject classifications: Primary 60A10; secondary 60B10, 60B15

Keywords and phrases: Similarity, multivariate distributions, trimming, consistency, uniqueness, Mass Transportation Problem, optimal transportation plan, trimmed probability, Skorohod representation

ming functions) is already present in Gordaliza [11], although it seems that its analysis has not been considered until very recently. Cascos and López-Díaz [4] and [5] and our already cited work [1] contain some of this analysis. The definition of trimming in these papers is the following: Given $\alpha \in [0,1]$ and a probability measure, P, on the Euclidean space R^k, an α-trimming of P is any probability measure P^* such that $P^*(B) = \int_B w dP$, for some weight function, w, satisfying $0 \leq w \leq 1/(1-\alpha)$ and $\int w dP = 1$. This obviously generalizes the simplest version of trimming consisting in the conditional probability given a set (of probability greater or equal than $1-\alpha$). A more formal definition appears in Definition 2.1.

Assume that P_1 and P_2 are probabilities on R^k and let $\mathcal{R}_\alpha(P_1)$ and $\mathcal{R}_\alpha(P_2)$ be their corresponding sets of α-trimmed probabilities (see (2.1) below). If d is a metric over the set of probabilities, there are several problems of interest related to the trimmed versions of P_1 and P_2. By defining the measures of dissimilarity between P_1 and P_2 at level α:

$$\begin{aligned}
\mathcal{T}_1(P_1, P_2) &:= \min_{P_2^* \in \mathcal{R}_\alpha(P_2)} d(P_1, P_2^*), \\
\mathcal{T}_2(P_1, P_2) &:= \min_{P_1^* \in \mathcal{R}_\alpha(P_1), P_2^* \in \mathcal{R}_\alpha(P_2)} d(P_1^*, P_2^*),
\end{aligned}$$

we would be interested in the best way of trimming one or both probabilities to achieve the greatest similarity between the corresponding versions.

Note that P_1 and P_2 do not play symmetric roles in $\mathcal{T}_1$ in the applications. For instance, if P_n is the sample distribution and P is a hypothesized distribution, then $\mathcal{T}_1(P_n, P)$ measures the similarity between our sample and a trimmed version of the hypothesized model. On the other hand, $\mathcal{T}_1(P, P_n)$ measures the similarity between the main part of our sample and the model. Thus, $\mathcal{T}_1(P_n, P)$ can be appropriate to explore if our sample can be considered as coming from a model with some kind of censoring, while $\mathcal{T}_1(P, P_n)$ can be appropriate to analyze if our sample comes from some contaminated or distorted version of the model. The measure $\mathcal{T}_2$ is an appealing tool to analyze two-sample problems, when our samples are obtained from nearly similar populations, say because different small sub-populations merged into the main similar populations. Notice that the problem analyzed in [1] involves a problem related to $\mathcal{T}_2$, when it is assumed that the possible troubling sub-populations share similar proportional size and location in their corresponding populations (see Remark 2.9).

If d is the L_2-Wasserstein distance (see Definition 3.1), these problems lead to questions on the celebrated Mass Transportation Problem (MTP) concerning unexplored constraints in the Monge-Kantorovich formulation (see the books by Rachev and Rüschendorf [14] and by Villani [18] for an updated account of the interest and implications of the topic). Our problems involve the optimal transportations, with respect to the quadratic loss, between probabilities when a part could remain fixed. In fact, in the considered problems, minima are attained and the corresponding trimmed probabilities are solutions of MTP's which can be obtained or approximated. This gives an additional descriptive scope to the procedure, that might allow to discover underlying similar structures when they exist.

Since there are not general explicit expressions for the solutions of the mutidimensional MTP, it is of primary interest to analize the possibilities of Monte-Carlo approximations and this will be our main goal. However a keystone in this process concerns uniqueness of solutions, which we remark as other main result in the paper. The generality of this result was unexpected and opens greater perspectives for future work.

The program to be developed through the paper is the following: In Section 2 we will give the general definition, characterizations and properties of trimmed probabilities. A remarkable result concerns the representation of trimmings of any probability in terms of those of another (see Proposition 2.5 and Corollary 2.8). Our interest in MTP led to handle representations based on McCann's Theorem on MTP (see Theorem 2.6), generalizing those given in [1] in terms of the uniform distribution on the unit interval. This representation has the added value of providing a particular Skorohod's a.s. representation for weak convergence in R^k (see Theorem 2.7). Section 3 analyzes the different possibilities that naturally arise to measure the dissimilarity between two probabilities using trimmings at a given level, as well as for choosing trimming patterns. The main results in this section involve the use of the L_2-Wasserstein distance, a framework where the connection with MTP is patent. By handling suitable representations it is easy to prove the convergence of trimmings of convergent sequences (Lemma 3.7). This result generalizes the consistency obtained in [5], where only sample distributions are considered, and allows to obtain the consistency of the introduced dissimilarity measures (Theorem 3.14). Another remarkable result concerns the uniqueness of the best pair of trimmed probabilities solving the corresponding minimization problems (Theorems 3.6 and 3.12). Finally, Section 4 explores, with an example, the possibilities in descriptive analysis of probability measures that arise from this approach.

The notation to be employed in this paper is the following. $(\mathcal{X}, \beta)$ will denote a separable metric space endowed with its Borel σ-field although often, mostly in Section 3, will be the k-dimensional Euclidean space R^k. In this case the Lebesgue measure will be denoted by ℓ^k. The set of all probability measures defined on β will be denoted $\mathcal{P}(\mathcal{X}, \beta)$ and $\mathcal{F}_2(\mathcal{X})$ will represent the set of the distributions in $\mathcal{P}(\mathcal{X}, \beta)$ with finite second moment.

Given two probability distributions P, Q, by $P \ll Q$ we will denote absolute continuity of P with respect to (w.r.t) Q. We will say that the measurable map $T : \mathcal{X} \to \mathcal{X}$ transports P to Q if $Q = P \circ T^{-1}$. supp(P) will be the support of P and $P(\cdot|B)$ the conditional distribution given the set B.

Unless otherwise stated, the random vectors will be assumed to be defined on the same probability space (Ω, σ, ν). Convergence in distribution will be denoted by $\to_w$ and $\mathcal{L}(X)$ will denote the law of X.

2. Trimmings.

We begin with some general definitions and properties of trimmed probabilities defined on a separable metric space $(\mathcal{X}, \beta)$ with its Borel σ-field.

Definition 2.1. *Given $0 \leq \alpha \leq 1$ and $P \in \mathcal{P}(\mathcal{X}, \beta)$, we say that $P^* \in \mathcal{P}(\mathcal{X}, \beta)$, is an α-trimming of P if $P^* \ll P$, and $\frac{dP^*}{dP} \leq \frac{1}{1-\alpha}$.*

We will denote the set of α-trimmings of P by $\mathcal{R}_\alpha(P)$, that is

$$\mathcal{R}_\alpha(P) = \left\{ P^* \in \mathcal{P}(\mathcal{X}, \beta) : P^* \ll P, \quad \text{and} \quad \frac{dP^*}{dP} \leq \frac{1}{1-\alpha} \quad P\text{-a.s.} \right\}, \tag{2.1}$$

Notice that $\mathcal{R}_1(P)$ is just the set of probability measures absolutely continuous with respect to P.

An equivalent characterization is that $P^* \in \mathcal{R}_\alpha(P)$ if and only if $P^* \ll P$ and $\frac{dP^*}{dP} = \frac{1}{1-\alpha} f$ with $0 \leq f \leq 1$. If f takes only the values 0 and 1 then it is the indicator of a set, say A, such that $P(A) \geq 1 - \alpha$ and the trimming corresponds to considering the probability measure $P(\cdot|A)$. Definition 2.1 allows to reduce the weight of some regions of the measurable space without completely removing them from the feasible set.

The following propositions collect some basic facts about trimmings.

Proposition 2.2. *For any probability measure, $P \in \mathcal{P}(\mathcal{X}, \beta)$,*

(a) *$\mathcal{R}_{\alpha_1}(P) \subset \mathcal{R}_{\alpha_2}(P)$ if $\alpha_1 \leq \alpha_2$.*
(b) *$\mathcal{R}_0(P) = \{P\}$.*
(c) *$\mathcal{R}_\alpha(P)$ is a convex set.*
(d) *For $\alpha < 1$, $P^* \in \mathcal{R}_\alpha(P)$ if and only if $P^*(A) \leq \frac{1}{1-\alpha} P(A)$ for all $A \in \beta$.*
(e) *If $\alpha < 1$ then $\mathcal{R}_\alpha(P)$ is closed for the topology of weak convergence in $\mathcal{P}(\mathcal{X}, \beta)$. If $\mathcal{X}$ is also complete then $\mathcal{R}_\alpha(P)$ is compact.*
(f) *If $\alpha < 1$ then $P^* \in \mathcal{R}_\alpha(P)$ if and only if for all continuous and bounded function $h \geq 0$: $\int h dP^* \leq \frac{1}{1-\alpha} \int h dP$.*

Proof. (a)-(c) are trivial and so is the *only if* part of (d). For the *if* part note that if $P^*(A) \le \frac{1}{1-\alpha}P(A)$ for all $A \in \beta$, then, $P^* \ll P$ and

$$P^*(A) = \int_A \frac{dP^*}{dP} dP \le \int_A \frac{1}{1-\alpha} dP,$$

for all $A \in \beta$ and, therefore, $\frac{dP^*}{dP} \le \frac{1}{1-\alpha}$ P-a.s. To prove the first part of (e) it suffices to show that if $\{P_n^*\}_n$ is a sequence such that $P_n^* \in \mathcal{R}_\alpha(P)$ and $P_n^* \to_w P^*$ then $P^* \in \mathcal{R}_\alpha(P)$. Thus, let us assume that $\{P_n^*\}_n$ is such a sequence. Then, by the portmanteau Theorem we have $P^*(A) \le \liminf_{n\to\infty} P_n^*(A) \le \frac{1}{1-\alpha}P(A)$, for every open set A. The result follows from (d) and regularity of probability measures on metric spaces. If $\mathcal{X}$ is complete then P is tight. It follows from (d) that $\mathcal{R}_\alpha(P)$ is tight and, by Prokhorov's Theorem (see, e.g., Theorem 1.2.10 in Araujo and Giné [2]), it is relatively compact, hence compact.

Only the *if* part of (f) needs some additional justification, so let us show that $P^*(A) \le \frac{1}{1-\alpha}P(A)$ if A is a closed set. Then, resorting to the metric d in the space, the functions $h_n(x) = \min(1, nd(x, A))$ are positive, continuous and bounded, and $h_n(x) \downarrow I_A(x)$, hence

$$P^*(A) = \lim_{n\to\infty} \int h_n dP^* \le \lim_{n\to\infty} \frac{1}{1-\alpha} \int h_n dP = \frac{1}{1-\alpha} P(A).$$

From the regularity of probability measures on separable metric spaces, the inequality extends to every Borel set. □

The next proposition guarantees that the weak limits of trimmed versions of weakly convergent sequences are trimmed versions of the limits. Note that the result would be false if our definition would have required exact trimming levels (thus, making property (a) in Proposition 2.2 false).

Proposition 2.3. *Let P, and $\{P_n\}_n$ be in $\mathcal{P}(\mathcal{X}, \beta)$. If $\alpha < 1$, $\{P_n\}_n$ is a tight sequence and $P_n^* \in \mathcal{R}_\alpha(P_n)$ for every n, then $\{P_n^*\}_n$ is tight.*

Moreover, if $P_n \to_w P$ and $P_n^* \to_w P^*$, then $P^* \in \mathcal{R}_\alpha(P)$.

Proof. The first part is trivial from (d) in Proposition 2.2, while the second easily follows from the portmanteau Theorem and (f) in the same proposition: For any $h \ge 0$ continuous and bounded,

$$\int h dP_n^* \le \frac{1}{1-\alpha} \int h dP_n \text{ for all } n, \tag{2.2}$$

thus in the limit we have

$$\int h dP^* = \lim_{n\to\infty} \int h dP_n^* \le \lim_{n\to\infty} \frac{1}{1-\alpha} \int h dP_n = \frac{1}{1-\alpha} \int h dP.$$

□

In [1] we gave an useful characterization of $\mathcal{R}_\alpha(P)$. It employs the set $\mathcal{C}_\alpha$, the class of absolutely continuous functions $h : [0,1] \to [0,1]$ such that, $h(0) = 0$, $h(1) = 1$, with derivative h' such that $0 \le h' \le \frac{1}{1-\alpha}$. We give such characterization in the following proposition as a matter of motivation and comparison with the result in the general setting, contained in Proposition 2.5. For this comparison remember that the quantile function of a distribution transports the uniform law, $U(0,1)$, on this distribution.

Proposition 2.4. *For any probability measure, P, on the real line*

$$\mathcal{R}_\alpha(P) = \{P^* \in \mathcal{P}(R,\beta) : P^*(-\infty,t] = h\left(P(-\infty,t]\right), \quad h \in \mathcal{C}_\alpha\}$$

Proposition 2.4 when applied to $U(0,1)$, states that the class $\mathcal{C}_\alpha$ is the class of all the distribution functions of α-trimmings of the uniform distribution. Then, this proposition characterizes the α-trimmings of every distribution in terms of the α-trimmings of the $U(0,1)$ distribution.

Proposition 2.5. *Let $Q \in \mathcal{P}(\mathcal{X},\beta)$. If T transports Q to P, then*

$$\mathcal{R}_\alpha(P) = \left\{P^* \in \mathcal{P}(\mathcal{X},\beta) : P^* = Q^* \circ T^{-1},\, Q^* \in \mathcal{R}_\alpha(Q)\right\}.$$

Proof. If $\alpha = 1$ and Q^* is any probability absolutely continuous with respect to Q, then $P^* := Q^* \circ T^{-1} \ll P$, because $P(B) = 0$ implies $Q(T^{-1}(B)) = 0$, thus $P^*(B) = Q^*(T^{-1}(B)) = 0$. On the other hand, if $P^* \ll P$, we can define $w(y) = \frac{dP^*}{dP}(T(y))$ and $Q^*(B) = \int_B w(y)Q(dy)$, hence, the change of variable formula shows for any set B in β:

$$\begin{aligned} Q^* \circ T^{-1}(B) &= \int_{T^{-1}(B)} \frac{dP^*}{dP}(T(y))Q(dy) \\ &= \int_B \frac{dP^*}{dP}(x)P(dx) = P^*(B). \end{aligned}$$

Let us assume that $\alpha < 1$. If $Q^* \in \mathcal{R}_\alpha(Q)$, then for any B in β:

$$\begin{aligned} Q^* \circ T^{-1}(B) &= \int_{T^{-1}(B)} \frac{dQ^*}{dQ}(x)Q(dx) \\ &\le \frac{1}{1-\alpha} Q\left(T^{-1}(B)\right) = \frac{1}{1-\alpha}P(B), \end{aligned}$$

thus $Q^* \circ T^{-1} \in \mathcal{R}_\alpha(P)$.

If we assume that $P^* \in \mathcal{R}_\alpha(P)$, by defining Q^* as above: $Q^*(B) = \int_B \frac{dP^*}{dP}(T(y))Q(dy)$, we have $Q^* \ll Q$, and, $Q^* \circ T^{-1} = P^*$. Moreover, since $\frac{dP^*}{dP}(x) \le \frac{1}{1-\alpha}$ a.s. (P) and $P = Q \circ T^{-1}$, also $\frac{dP^*}{dP}(T(y)) \le \frac{1}{1-\alpha}$ a.s.(Q) hence $Q^* \in \mathcal{R}_\alpha(Q)$. □

Obviously if P is atomic, every transported probability from P will also be atomic. Thus, in every $\mathcal{X}$ there are probabilities that cannot be tranported to any other probability on $\mathcal{P}(\mathcal{X}, \beta)$. On the other hand, it is well known (see e.g. Theorem 11.7.5 in Dudley [9]) that any law in a complete separable metric space can be obtained as the law of a random variable defined on the space $[0, 1]$ endowed with the Lebesgue measure. Moreover, when a weakly convergent sequence of laws is involved, it is also possible to obtain a sequence of random variables in that space with the given laws, converging almost surely (see e.g. Theorem 11.7.2 in [9]). This is known as a Skorohod or Skorohod-Dudley-Wichura a.s. representation, and our results concerning convergences could be based on such a representation in a general framework. However, the possibility of handling as representations only suitable transports between probabilities on the same space, can be of independent interest and we have opted by an alternative construction.

In our transport context, McCann [13] proved the existence of special maps transporting any probability satisfying certain regularity conditions to any other probability if $\mathcal{X} = R^k$. Such maps are cyclically monotone, characterized by the fact that they are sub-gradients of convex functions, and the regularity condition on the probability can be described as *giving zero probability to small sets* (see Villani [18]). However to avoid technical details we will assume the more familiar (and restrictive) absolute continuity, so that we state the following version of McCann's result.

Theorem 2.6 (McCann)**.** *If $P, Q \in \mathcal{P}(R^k, \beta)$, and $P \ll \ell^k$, then there exists an (essentially) unique cyclically monotone map transporting P to Q.*

Regarding asymptotic properties, the following theorem is a particular case of a theorem of Heinich and Lootgieter [12], extending results in Cuesta et al. [7] and Tuero [17]. It is included for completeness, and shows that the particular representation obtained through Theorem 2.6 provides a particular Skorohod's a.s. representation for weak convergence.

Theorem 2.7. *Let $Q_n, Q, P \in \mathcal{P}(R^k, \beta)$ such that $Q_n \rightarrow_w Q$, and $P \ll \ell^k$, and let T_n (resp. T) be cyclically monotone maps transporting P to Q_n (resp. to Q). Then $T_n \rightarrow T$, P-a.s.*

The maps T involved in Theorem 2.6 are intrinsically related to the optimal transportation from a probability to another when the cost function is quadratic. Under the additional hypothesis of finite second order moment of the involved probabilities and the general case of separable Hilbert spaces, the existence and uniqueness of a map T transporting P to Q with minimum integrated cost was obtained in Cuesta-Albertos and Matrán [6]. Therefore, any such function T, given by Theorem 2.6, transporting P to Q

will be called an *optimal transportation plan* (o.t.p.) between P and Q or, simply, an o.t.p. (in fact such function is an o.t.p. between any other P' and $P' \circ T^{-1}$). A general overview of the MTP in the Probability context can be found in the book by Rachev and Rüschendorf [14]. A more analytical approach can be found in the book by Villani [18].

From Proposition 2.5 and Theorem 2.6 we obtain the following corollary on the existence of universal representations of the sets of trimming of the probabilities, on R^k, based on the set of trimmings of a given probability.

Corollary 2.8. *If $P_0, Q \in \mathcal{P}(R^k, \beta)$, and $P_0 \ll \ell^k$, then $\mathcal{R}_\alpha(Q)$ coincides with the set of all probabilities which can be written as $P_0^* \circ T^{-1}$ where $P_0^* \in \mathcal{R}_\alpha(P_0)$ and T is the (essentially) unique o.t.p. between P_0 and Q.*

Remark 2.9. Once we have chosen a particular probability measure P_0 on R^k, $P_0 \ll \ell^k$, Corollary 2.8 allows to induce trimmed versions "similarly tailored" according to the shape of P_0: If P_1, P_2 are probabilities on R^k and T_1, T_2 are the respective o.t.p. between P_0 and P_1 and between P_0 and P_2, any $P_0^* \in \mathcal{R}_\alpha(P_0)$ determine through the o.t.p.'s the trimmed probabilities $P_1^* \in \mathcal{R}_\alpha(P_1)$, $P_2^* \in \mathcal{R}_\alpha(P_2)$, by the relations $P_1^* = P_0^* \circ T_1^{-1}, P_2^* = P_0^* \circ T_2^{-1}$.

This representation of the trimmed versions of two probabilities through those of another permits the consideration of a new measure of similarity between P_1 and P_2 according to the shape of P_0 through the relation

$$\mathcal{T}_3(P_1, P_2) = \min_{P_0^* \in \mathcal{R}_\alpha(P_0)} d(P_0^* \circ T_1^{-1}, P_0^* \circ T_2^{-1}).$$

Note the role of P_0 as a common pattern to measuring dissimilarities. Trimming of two probabilities according to the same P_0 will be called *similarly tailored*. That was the kind of trimming adopted in [1] for probabilities on the real line and the U(0,1) law as distribution of reference.

3. Best trimmed approximations.

Given a metric, d, on $\mathcal{P}(\mathcal{X}, \beta)$ (or a convenient subset of it) and two probability measures, P and Q in $\mathcal{P}(\mathcal{X}, \beta)$, we can consider the problem of finding the α-trimmed version of P which is closest to Q in d metric, namely,

$$R_\alpha = R_{\alpha,d,P,Q} = \operatorname*{argmin}_{P^* \in \mathcal{R}_\alpha(P)} d(P^*, Q).$$

Compactness of $\mathcal{R}_\alpha(P)$ in the topology of weak convergence guarantees the existence of such *best trimmed approximation* if d metrizes weak convergence (e.g., if d is the bounded Lipschitz or the Prokhorov metric). This best trimmed approximation needs not be unique, but convexity of the set of trimmings ensures that the set of best approximations is a convex, compact set if d is a convex metric (meaning that $d(\gamma P + (1-\gamma)Q, R) \leq \gamma d(P,R) + (1-\gamma)d(Q,R)$ for all $\gamma \in [0,1]$). This holds, for instance, for the bounded Lipschitz metric. Nevertheless, the bounded Lipschitz or the Prokhorov metric are not easily computed in general and may not be the most interesting choice for applications.

In this section we will assume that $\mathcal{X}$ is a Banach space with norm $\|\cdot\|$ (usually $\mathcal{X} = R^k$) and we focus on the Wasserstein metric $\mathcal{W}_2$, defined by

$$\mathcal{W}_2^2(P,Q) = \inf_{\pi \in \mathcal{M}(P,Q)} \left\{ \int \|x-y\|^2 d\pi(x,y) \right\}, \tag{3.1}$$

where $\mathcal{M}(P,Q)$ is the set of Borel probability measures on $\beta \times \beta$ with marginals P and Q. It can be shown that $\mathcal{W}_2$ is a metric on the set $\mathcal{F}_2(\mathcal{X})$ provided $\mathcal{X}$ is a separable Banach space (see Bickel and Freedman [3]). With an slight abuse of notation, given $P \in \mathcal{F}_2(\mathcal{X})$ and $\Theta, \Phi \subset \mathcal{F}_2(\mathcal{X})$, we will often denote

$$\mathcal{W}_2(P,\Theta) = \inf_{Q\in\Theta} \mathcal{W}_2(P,Q) \text{ and } \mathcal{W}_2(\Theta,\Phi) = \inf_{(P,Q)\in\Theta\times\Phi} \mathcal{W}_2(P,Q).$$

The infimum in (3.1) is attained, so that to find $\mathcal{W}_2^2(P,Q)$ it is enough to obtain a pair (X,Y) of random vectors with distributions laws $\mathcal{L}(X) = P$ and $\mathcal{L}(Y) = Q$ and satisfying

$$\int \|X-Y\|^2 \, d\nu = \inf \left\{ \int \|U-V\|^2 \, d\nu, \quad \mathcal{L}(U) = P, \quad \mathcal{L}(V) = Q \right\}$$

Such a pair (X,Y) is called an L_2-o.t.p. for (P,Q). (L_2-optimal coupling for (P,Q) is an alternative, sometimes used, terminology).

In [6] (see also Rüschendorf and Rachev [16] and McCann [13]) it was proved that, under continuity assumptions on the probability P, the L_2-o.t.p. (X,Y) for (P,Q) can be represented as $(X, T(X))$ for some suitable "optimal map" T. This map coincides with the (essentially unique) cyclically monotone map transporting P to Q (see [13]). Since in this section we only use the $\mathcal{W}_2$ distance, in the sequel we will use the term o.t.p. for the pair (X,Y) which will also apply to the map T. For posterior use we summarize some properties in the following statement. The interested reader can find the proofs in Cuesta-Albertos et al. [6], [8], and Tuero [17]. A different approach, involving more analytical proofs, is summarized in [18]

Proposition 3.1. *Assume that $P,Q \in \mathcal{F}_2(R^k)$, and that $P \ll \ell^k$, and let (X,Y) be an o.t.p. for (P,Q). Then we have:*

(a) *The cardinal of the support of a regular conditional distribution of Y given $X = x$ is one, P-a.s.*

(b) *There exists a P-probability one set D and a Borel measurable cyclically monotone map $T : D \to R^k$ such that $Y = T(X)$, $\nu-a.s.$*

(c) *If (X, Y_1) and (X, Y_2) are o.t.p.'s for (P, Q), then $Y_1 = Y_2$ $\nu-a.s.$*

(d) *If T is an o.t.p. for (P, Q), then T is a.e. continuous on supp(P).*

Regarding the convexity of the $\mathcal{W}_2-$metric we have a nice property. It is easy to check that the Wasserstein metric always satisfies the inequality $\mathcal{W}_2^2(\gamma P+(1-\gamma)Q, R) \leq \gamma\mathcal{W}_2^2(P, R)+(1-\gamma)\mathcal{W}_2^2(Q, R)$, $\gamma \in (0, 1)$, but when $R \ll \ell^k$, property (a) in Proposition 3.1 leads to more:

Theorem 3.2. *Let $P_i, Q_i, i = 1, 2$, be probability measures in $\mathcal{F}_2(R^k)$ such that $P_i \ll \ell^k, i = 1, 2$. If $Q_1 \neq Q_2$ and there is not a common o.t.p. T such that $Q_1 = P_1 \circ T^{-1}$ and $Q_2 = P_2 \circ T^{-1}$, then, for every γ in $(0, 1)$,*

$$\mathcal{W}_2^2(\gamma P_1+(1-\gamma)P_2, \gamma Q_1+(1-\gamma)Q_2) < \gamma\mathcal{W}_2^2(P_1, Q_1)+(1-\gamma)\mathcal{W}_2^2(P_2, Q_2).$$

Proof. Assume that f_i is the density function of P_i, and let $(X_i, T_i(X_i))$, $i = 1, 2$ be o.t.p.'s for $(P_i, Q_i), i = 1, 2$. If $P_\gamma := \gamma P_1 + (1 - \gamma)P_2$ and $Q_\gamma := \gamma Q_1 + (1 - \gamma)Q_2$, then $f_\gamma := \gamma f_1 + (1 - \gamma)f_2$ is a density function for P_γ. Let us define on the support of P_γ the following random function:

$$T(x) = \begin{cases} T_1(x) & \text{with probability } \gamma f_1(x)/(\gamma f_1(x) + (1-\gamma)f_2(x)) \\ T_2(x) & \text{with probability } (1-\gamma)f_2(x)/(\gamma f_1(x) + (1-\gamma)f_2(x)) \end{cases}$$

If X_γ is any r.v. with probability law $\mathcal{L}(X_\gamma) = P_\gamma$, we have:

$$\begin{aligned} \nu[T(X_\gamma) \in A] &= \int \mu[T(X_\gamma) \in A | X_\gamma = x] P_\gamma(dx) \\ &= \int I_A[T_1(x)] \frac{\gamma f_1(x)}{\gamma f_1(x) + (1-\gamma)f_2(x)} P_\gamma(dx) \\ &\quad + \int I_A[T_2(x)] \frac{(1-\gamma)f_2(x)}{\gamma f_1(x) + (1-\gamma)f_2(x)} P_\gamma(dx) \\ &= \gamma \int I_A[T_1(x)] f_1(x) dx + (1-\gamma) \int I_A[T_2(x)] f_2(x) dx \\ &= \gamma\nu[T_1(X_1) \in A] + (1-\gamma)\nu[T_2(X_2) \in A] \\ &= \gamma Q_1(A) + (1-\gamma)Q_2(A) = Q_\gamma(A). \end{aligned}$$

Since $\mathcal{L}(T(X_\gamma)) = Q_\gamma$, by the same argument, we have:

$$\begin{aligned} \mathcal{W}_2^2(P_\gamma, Q_\gamma) &\leq \int \|X_\gamma - T(X_\gamma)\|^2 d\nu \\ &= \gamma \int \|X_1 - T_1(X_1)\|^2 d\nu + (1-\gamma) \int \|X_2 - T_2(X_2)\|^2 d\nu \\ &= \gamma\mathcal{W}_2^2(P_1, Q_1) + (1-\gamma)\mathcal{W}_2^2(P_2, Q_2). \end{aligned}$$

This shows that $\mathcal{W}_2^2(P_\gamma, Q_\gamma) < \gamma \mathcal{W}_2^2(P_1, Q_1) + (1-\gamma)\mathcal{W}_2^2(P_2, Q_2)$ unless T is an o.t.p. for (P_γ, Q_γ). But a) in Proposition 3.1 implies that a random map cannot be an o.t.p, thus T should be non-random, leading to

$$T(x) = \begin{cases} T_1(x) & \text{if } x \in \text{Supp}(P_1) - \text{Supp}(P_2) \\ T_1(x) & (= T_2(x)) \text{ if } x \in \text{Supp}(P_1) \cap \text{Supp}(P_2) \\ T_2(x) & \text{if } x \in \text{Supp}(P_2) - \text{Supp}(P_1) \end{cases}$$

This fact would contradict our hypothesis because it implies that T would be an o.t.p. common for (P_1, Q_1) and (P_2, Q_2). □

Taking $P_1 = P_2$ in Theorem 3.2, we obtain the following corollary, stating the strict convexity of $\mathcal{W}_2^2(P, \cdot)$.

Corollary 3.3. *Let P, Q_1, Q_2, be probability measures in $\mathcal{F}_2(R^k)$ and assume that $P \ll \ell^k$. If $Q_1 \neq Q_2$, then, for every γ in $(0,1)$,*

$$\mathcal{W}_2^2(P, \gamma Q_1 + (1-\gamma)Q_2) < \gamma \mathcal{W}_2^2(P, Q_1) + (1-\gamma)\mathcal{W}_2^2(P, Q_2).$$

Now let us return to the consideration of trimmed probabilities. If P has finite second moment and $P^* \in \mathcal{R}_\alpha(P)$ then

$$\int \|x\|^2 dP^*(x) \leq \frac{1}{1-\alpha} \int \|x\|^2 dP(x).$$

This shows that $\mathcal{R}_\alpha(P) \subset \mathcal{F}_2(R^k)$ if $P \in \mathcal{F}_2(R^k)$. Our next result is a version of Proposition 2.2 (e) for the metric $\mathcal{W}_2$.

Proposition 3.4. *If $P \in \mathcal{F}_2(\mathcal{X})$, where $\mathcal{X}$ is a separable Banach space, then $\mathcal{R}_\alpha(P)$ is compact in the $\mathcal{W}_2$ topology.*

Proof. Convergence in $\mathcal{W}_2$ is equivalent to weak convergence plus convergence of second moments (Bickel and Freedman [3], Lemma 8.3). We saw in the proof of Proposition 2.2 that $\mathcal{R}_\alpha(P)$ is tight. Now, given an infinite set $\mathcal{R} \subset \mathcal{R}_\alpha(P)$ we can extract a sequence $\{Q_n\}_n \subset \mathcal{R}$ that converges weakly. Let us call Q its weak limit. Then $\mathcal{W}_2(Q_n, Q) \to 0$ iff $\|x\|^2$ is uniformly Q_n-integrable. Fix $t > 0$. Then

$$\int_{\|x\|>t} \|x\|^2 dQ_n(x) = \int_{\|x\|>t} \|x\|^2 \frac{dQ_n}{dP}(x) dP(x) \leq \frac{1}{1-\alpha} \int_{\|x\|>t} \|x\|^2 dP(x),$$

from which the uniform integrability of $\|x\|^2$ is immediate. □

The last proposition implies that there always exists a best trimmed approximation in Wasserstein metric and the set of best trimmed approximants is compact. From the convexity of the metric the set of best approximations is convex. The following example shows that the best trimmed approximation is not always unique.

Example 3.5. Set $P = \frac{1}{2}\delta_{\{-1\}} + \frac{1}{2}\delta_{\{1\}}$ and $Q = \delta_{\{0\}}$. Obviously, every $P^* \in \mathcal{R}_\alpha(P)$ satisfies that $\mathcal{W}_2(P^*, Q) = 1$, and, then, the set of best trimmed approximations is $\mathcal{R}_\alpha(P)$.

Of course, under the absolutely continuity hypothesis, the strict convexity property in Corollary 3.3 ensures the uniqueness of the best trimmed approximation.

Theorem 3.6. *Assume that P and Q, belong to $\mathcal{F}_2(R^k)$ and that $P \ll \ell^k$. Then there exists an unique $Q_\alpha \in \mathcal{R}_\alpha(Q)$, verifying:*

$$\mathcal{W}_2(P, Q_\alpha) = \mathcal{W}_2(P, \mathcal{R}_\alpha(Q)).$$

This uniqueness result shows that in the measure of dissimilarity $\mathcal{T}_1(P, Q) = \mathcal{W}_2(P, \mathcal{R}_\alpha(Q))$, considered in the introduction, the minimum is attained by just a trimmed probability if P is absolutely continuous.

Theorem 2.7 and Corollary 2.8 allow also to show that any trimmed version of a probability in $\mathcal{F}_2(R^k)$, which is the limit of probabilities in $\mathcal{F}_2(R^k)$, can be obtained as the limit of trimmed versions of these probabilities.

Lemma 3.7. *Let $\{Q_n\}_n$ and Q be in $\mathcal{P}(R^k, \beta)$, and assume that $Q_n \rightarrow_w Q$. Then, if $Q^* \in \mathcal{R}_\alpha(Q)$, there exist a sequence $\{Q_n^*\}_n$ such that $Q_n^* \in \mathcal{R}_\alpha(Q_n)$, for all n, and $Q_n^* \rightarrow_w Q^*$.*

Proof. Let P be any probability measure on R^k such that $P \ll \ell^k$, and consider the sequence $\{T_n\}_n$ of o.t.p.'s between P and P_n. If T is the o.t.p. between P and Q, Theorem 2.7 implies that $T_n \to T$, $P-$a.s.

By Corollary 2.8 $Q^* = P^* \circ T^{-1}$ for some $Q^* \in \mathcal{R}_\alpha(Q)$. Define now $Q_n^* = P^* \circ T_n^{-1}$, that belongs to $\mathcal{R}_\alpha(Q_n)$ by the already used characterization in Corollary 2.8. Since $T_n \to T$, $P-$a.s., and $P^* \ll P$, also $T_n \to T$, P^*-a.s. Therefore $Q_n^* = P^* \circ T_n^{-1} \rightarrow_w P^* \circ T^{-1} = Q^*$. □

Remark 3.8. As it was already noticed, this lemma (in conjunction with Proposition 2.3) generalizes the main result in [5]. There the proof involved a smart, ad hoc, construction of trimmed versions of the sample distributions. A generalization of that proof, for arbitrary convergent sequences of probability measures, seems to be impossible.

Theorem 3.9. *Let $\{P_n\}_n$, P and Q be in $\mathcal{F}_2(R^k)$, such that $\mathcal{W}_2(P_n, P) \to 0$.*

a) If $Q \ll \ell^k$ *and* $P_{n,\alpha} := \underset{P_n^* \in \mathcal{R}_\alpha(P_n)}{\text{arg min}}\ \mathcal{W}_2(P_n^*, Q)$*, then*

$$\mathcal{W}_2(P_{n,\alpha}, P_\alpha) \to 0, \ \textit{where } P_\alpha := \underset{P^* \in \mathcal{R}_\alpha(P)}{\text{arg min}}\ \mathcal{W}_2(P^*, Q).$$

b) If $P \ll \ell^k$ *and* $Q_{n,\alpha} \in \mathcal{R}_\alpha(Q)$ *satisfies that* $\mathcal{W}_2(P_n, Q_{n,\alpha}) = \mathcal{W}_2(P_n, \mathcal{R}_\alpha(Q))$*, then*

$$\mathcal{W}_2(Q_{n,\alpha}, Q_\alpha) \to 0, \ \textit{where } Q_\alpha := \underset{Q^* \in \mathcal{R}_\alpha(Q)}{\text{arg min}}\ \mathcal{W}_2(P, Q^*).$$

Proof. Both statements have similar proofs, so let us consider only statement a). By Proposition 2.3 the sequence $\{P_{n,\alpha}\}_n$ is tight and by the same argument that in the proof of Proposition 3.4, the function $\|x\|^2$ is uniformly integrable for $\{P_n\}_n$ thus also for $\{P_{n,\alpha}\}_n$. Therefore to show $\mathcal{W}_2(P_{n,\alpha}, P_\alpha) \to 0$ it suffices to guarantee that if $\{P_{r_n,\alpha}\}_n$ is any weakly convergent subsequence then $P_{r_n,\alpha} \to_w P_\alpha$.

By Proposition 2.3, if $P_{r_n,\alpha} \to_w P^*$, then $P^* \in \mathcal{R}_\alpha(P)$ and, therefore

$$\mathcal{W}_2(P_\alpha, Q) \le \mathcal{W}_2(P^*, Q) = \lim \mathcal{W}_2(P_{r_n,\alpha}, Q) \le \liminf \mathcal{W}_2(P^*_{r_n,\alpha}, Q), \tag{3.2}$$

for any choice $P^*_{r_n,\alpha} \in \mathcal{R}_\alpha(P_{r_n})$. Lemma 3.7 and the uniform integrability argument allow to choose this last sequence verifying $\mathcal{W}_2(P^*_{r_n,\alpha}, P_\alpha) \to 0$, hence $\mathcal{W}_2(P^*_{r_n,\alpha}, Q) \to \mathcal{W}_2(P_\alpha, Q)$, which joined with (3.2) and with the uniqueness of the best trimmed approximation P_α given by Theorem 3.6 shows that $P^* = P_\alpha$. □

3.1. *Trimming in both probabilities.* To state the uniqueness of the best trimmed approximations we will use some additional notation and basic results. Given $v_0 \in R^k$ with $\|v_0\| = 1$, we will consider H_0 an hyperplane orthogonal to v_0. The orthogonal projection on H_0 will be denoted by π_0 and for every $y \in R^k$, we will denote $r_y = \langle y - \pi_0(y), v_0 \rangle$. Given a measurable set $B \subset R^k$, and $z \in H_0$, we will also denote

$$B_z := \{y \in B : \pi_0(y) = z\}, \text{ and } z_{v_0} := \{r_y : y \in B_z\},$$

Given the probability distribution P, we will denote with P° the marginal distribution of P on H_0 and with P_z a regular conditional distribution given z, where $z \in H_0$. This conditional probability induces in an obvious way a probability on the real line through the isometry $\mathcal{I}_z$ between $(R^k)_z$ and R, given by $y \to r_y$. This probability will be denoted λ_z and its distribution (resp. quantile) function will be denoted $F(x|z)$ (resp. $q_z(t)$. We stress on the joint measurability of these functions in the following lemma, that we include for future reference.

Lemma 3.10. *The maps $(x,z) \to F(x|z)$ and $(t,z) \to q_z(t)$ are jointly measurable in their arguments.*

Proof. Note that if $F(x,y)$ is a joint distribution function on $R \times R^{k-1}$ and $G(z)$ is the marginal on R^{k-1}, then they are measurable (for probabilities supported on finite sets it is obvious and the generalization carries over through standard arguments). On the other hand, let us consider the measures η_x and μ respectively associated to the increasing functions $F(x,\cdot)$ and $G(\cdot)$. As a consequence of the Differentiation Theorem for Radon Measures (see e.g. Sections 1.6.2 and 1.7.1 in Evans and Gariepy [10]), if we consider for any $z = (z_1, ..., z_{k-1}) \in R^{k-1}$, the sequence of rectangles $A_n(z) := \{(y_1, ..., y_{k-1}) : z_i - \frac{1}{n} < y_i \leq z_i + \frac{1}{n},\ i = 1, ..., k-1\}$, we have the following a.s. convergence, leading to the measurability:

$$F(x|z) = \lim_{n\to\infty} \frac{\eta_x(A_n(z))}{\mu(A_n(z))}.$$

The measurability of $q_z(t)$ follows from the key property $x \leq q_z(t)$ if and only if $F(x|z) \leq t$. □

Theorem 3.11 gives a nice property of the best trimmed approximations of two probabilities when trimming is allowed in both probabilities. According to this result, the best trimming functions involved in this problem are basically indicator functions of appropriate sets with, may be, the exception of points that remain fixed in the transport. In particular, partial trimming is impossible on supp$(P)-$supp(Q).

Theorem 3.11. *Let $\alpha > 0$, and let $P, Q \in \mathcal{P}(R^k, \beta)$. Assume that $P \ll \ell^k$ has density f w.r.t. ℓ_k. If $P_1 \in \mathcal{R}_\alpha(P)$ and $Q_1 \in \mathcal{R}_\alpha(Q)$ verify that*

$$\mathcal{W}_2^2(P_1, Q_1) = \mathcal{W}_2^2[\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)] > 0,$$

and T is an o.t.p. for (P_1, Q_1), then $T(x) = x$ P-a.s. on the set $\mathcal{A} := \{x \in R^k : a_1(x) \in (0,1)\}$, where $a_1 := (1-\alpha)f_1$ and f_1 is the density function of P_1 with respect to P.

Proof. Assume, on the contrary, that $P(\mathcal{A} \cap \{x \in R^k : \|T(x) - x\| > 0\}) > 0$ and let us denote by $\hat{P}$ the conditional distribution of P given this set.

From (e) in Proposition 3.1 we have that T is a.e. continuous. Let x_0 be a point in the support of $\hat{P}$ in which T is continuous. Then, for every $\epsilon > 0$ there exists $\delta > 0$ such that $T(B(x_0, \delta)) \subset B(T(x_0), \epsilon)$. Let us denote $A = B(x_0, \delta) \cap \mathcal{A}$.

Let $v_0 = (T(x_0) - x_0)/\|T(x_0) - x_0\|$ and H_0 be the hyperplane orthogonal to v_0 which contains x_0. With the notation at the beginning of this subsection, taking ϵ small enough, we can assume that $m := \inf_{y \in B(T(x_0),\epsilon)} r_y$ is greater than $M := \sup_{y\in B(x_0,\delta)} r_y$. Therefore,

$$\|T(y)-\pi_0[T(y)]\| > r_y, \text{ for every } y\in A. \tag{3.3}$$

On the other hand, we have

$$P[A] = \int_{H_0} P_z(A_z)P^\circ(dz) = \int_{H_0} \lambda_z(z_{v_0})P^\circ(dz). \tag{3.4}$$

Since x_0 belongs to the support of $\hat{P}$, then $P[A]>0$, thus

$$P^\circ\{z\in H_0 : \lambda_z(z_{v_0})>0\} > 0. \tag{3.5}$$

Let $z\in H_0$ such that $\lambda_z(z_{v_0})>0$. If $y_1, y_2 \in A_z$ satisfy that $r_{y_1} < r_{y_2}$, the orthogonality between $(\pi_0(y)-x_0)$ and $(y-\pi_0(y))$ for every $y\in R^k$ and (3.3) lead to

$$\begin{aligned}
\|y_1 - T(y_1)\|^2 &= \|T(y_1)-\pi_0[T(y_1)]+\pi_0(y_1)-y_1+\pi_0(T(y_1))-\pi_0(y_1)\|^2\\
&= \left(r_{T(y_1)}-r_{y_1}\right)^2 + \|\pi_0[T(y_1)]-z\|^2\\
&> \left(r_{T(y_1)}-r_{y_2}\right)^2 + \|\pi_0[T(y_1)]-\pi_0(y_2)\|^2 \qquad (3.6)\\
&= \|y_2 - T(y_1)\|^2.
\end{aligned}$$

Now, we consider the partition of the set $A = A^- \cup A^+$ given by

$$\begin{aligned}
A^- &:= \{y\in A : F\left(r_y|\pi_0(y)\right) \le 1/2\}, \text{ and}\\
A^+ &:= \{y\in A : F\left(r_y|\pi_0(y)\right) > 1/2\}.
\end{aligned}$$

From Lemma 3.10 we have that these sets are measurable. For almost every $z\in H_0$ satisfying $\lambda_z(z_{v_0})>0$ they define a value R_z, such that the sets

$$\begin{aligned}
A_z^- &:= \{y\in A_z : r_y < R_z\}, & A_z^+ &:= \{y\in A_z : r_y > R_z\},\\
z_{v_0}^- &:= \{r_y : y\in A_z^-\}, & z_{v_0}^+ &:= \{r_y : y\in A_z^+\}
\end{aligned}$$

verify $\lambda_z[z_{v_0}^-] = \lambda_z[z_{v_0}^+] > 0$. Let λ_z^- and λ_z^+ be the probability λ_z conditioned to the sets $z_{v_0}^-$ and $z_{v_0}^+$ respectively, and let their corresponding distribution (resp. quantile) functions be $F^-(x|z)$ and $F^+(x|z)$ (resp. $q_z^-(t)$ and $q_z^+(t)$). Then, recalling the isometry $\mathcal{I}_z$ and the way to obtain o.t.p.'s in the real line, the map $\Gamma : A^- \to A^+$ defined by

$$\Gamma(y) = \mathcal{I}_{\pi_0(y)}^{-1}\left[q_{\pi_0(y)}^+\left[F^-\left(r_y\ |\pi_0(y)\right)\right]\right]$$

is an o.t.p. between P_z^- and P_z^+ for almost every $z \in H_0$ satisfying $P_z(z_{v_0}) > 0$. To end the construction, let us consider the function $a^* : R^k \to R$ defined as follows:

$$a^*(y) = \begin{cases} a_1(y) & \text{if } y \notin A \\ a_1(y) - \min\{1 - a_1[\Gamma(y)], a_1(y)\} & \text{if } y \in A^- \\ a_1(y) + \min\{1 - a_1(y), a_1[\Gamma^{-1}(y)]\} & \text{if } y \in A^+. \end{cases}$$

From this point, the proof involves three steps:

Step 1. *$f^* := a^*/(1-\alpha)$ is a density with respect to P that defines a probability $P^* \in \mathcal{R}_\alpha(P)$.*

Obviously $a^*(R^k) \subset [0,1]$. On the other hand

$$\begin{aligned} \int_{R^k} a^*(y)P(dy) \quad = \quad & \int_{R^k} a_1(y)P(dy) \\ & - \int_{A^-} \min\{1 - a_1[\Gamma(y)], a_1(y)\} P(dy) \\ & + \int_{A^+} \min\left\{1 - a_1(y), a_1[\Gamma^{-1}(y)]\right\} P(dy). \end{aligned} \tag{3.7}$$

For almost every $z \in H_0$ satisfying $P_z(A_z) > 0$, by construction, the law of a_1 under P_z^+, $P_z^+ \circ a_1^{-1}$, coincides with the law $P_z^- \circ (a_1(\Gamma))^{-1}$, while $P_z^+ \circ (a_1(\Gamma^{-1}))^{-1} = P_z^- \circ a_1^{-1}$. Therefore the last term verifies

$$\begin{aligned} & \int_{A^+} \min\left\{1 - a_1(y), a_1[\Gamma^{-1}(y)]\right\} P(dy) \\ & \qquad = \int_{H_0} \left(\int_{A_z^+} \min\left\{1 - a_1(y), a_1[\Gamma^{-1}(y)]\right\} P_z(dy) \right) P^\circ(dz) \\ & \qquad = \int_{H_0} \left(\int_{A_z^-} \min\{1 - a_1(\Gamma(y)), a_1(y)\} P_z(dy) \right) P^\circ(dz) \\ & \qquad = \int_{A^-} \min\{1 - a_1[\Gamma(y)], a_1(y)\} P(dy), \end{aligned} \tag{3.8}$$

what, joined to (3.7) leads to $\int_{R^k} a^*(y)P(dy) = \int_{R^k} a_1(y)P(dy) = 1 - \alpha$, which proves this step.

Step 2. *There exists a random map, T^*, transporting P^* to Q_1.*

Let us consider the random map T^* defined by $T^*(y) = T(y)$ on the complementary of A^+ and, for $y \in A^+$, taking the values $T(y)$ or $T[\Gamma(y)]$ with probabilities $f_1(y)/f^*(y)$ $(= a_1(y)/a^*(y))$ and $[f^*(y) - f_1(y)]/f^*(y)$

$(= [a^*(y) - a_1(y)]/a^*(y))$ respectively. These values are positive because, by construction, $a^*(y) > a_1(y)$ on A^+.

The argument to show that T^* transports P^* to Q_1 is analogous to that developed in Theorem 3.2, taking into account that $P_z^+ \circ a_1^{-1} = P_z^- \circ (a_1(\Gamma))^{-1}$.

Step 3. $\mathcal{W}_2^2(P_1, Q_1) > \mathcal{W}_2^2(P^*, Q_1)$.

By construction of T^* and inequality (3.6), we have

$$\begin{aligned}
\mathcal{W}_2^2(P^*, Q_1) &\leq \int_{R^k} \|y - T^*(y)\|^2 P^*(dy) \\
&= \int_{(A^+)^c} \|y - T(y)\|^2 P^*(dy) \\
&\quad + \int_{A^+} \left(\|y - T(y)\|^2 \frac{f_1(y)}{f^*(y)} + \|y - T[\Gamma^{-1}(y)]\|^2 \frac{f^*(y) - f_1(y)}{f^*(y)} \right) f^*(y) P(dy) \\
&< \int_{(A^- \cup A^+)^c} \|y - T(y)\|^2 f_1(y) P(dy) + \int_{A^-} \|y - T(y)\|^2 f^*(y) P(dy) \\
&\quad + \int_{A^+} \left(\|y - T(y)\|^2 f_1(y) + \|\Gamma^{-1}(y) - T[\Gamma^{-1}(y)]\|^2 (f^*(y) - f_1(y)) \right) P(dy).
\end{aligned}$$

Moreover, by construction of the map Γ, recalling the relation $P_z^+ \circ (a_1(\Gamma^{-1}))^{-1} = P_z^- \circ (a_1)^{-1}$, we obtain that

$$\begin{aligned}
&\int_{A^+} \|\Gamma^{-1}(y) - T[\Gamma^{-1}(y)]\|^2 (f^*(y) - f_1(y)) P(dy) \\
&\quad = -\int_{A^-} \|y - T(y)\|^2 (f^*(y) - f_1(y)) P(dy),
\end{aligned}$$

what, by construction of f^*, gives

$$\mathcal{W}_2^2(P^*, Q_1) < \mathcal{W}_2^2(P_1, Q_1),$$

contradicting the optimality of the pair (P_1, Q_1). □

Theorem 3.12 (Uniqueness). *Let $\alpha > 0$ and let $P, Q \in \mathcal{P}(R^k, \beta)$, with $P \ll \ell^k$. If $\mathcal{W}_2^2[\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)] > 0$, then there exists a unique pair of probability distributions $P_1 \in \mathcal{R}_\alpha(P)$ and $Q_1 \in \mathcal{R}_\alpha(Q)$ such that*

$$\mathcal{W}_2^2(P_1, Q_1) = \mathcal{W}_2^2[\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)]. \tag{3.9}$$

Proof. Assume that (P_1, Q_1) and (P_2, Q_2) are two different pairs fulfilling (3.9), and let $a_i := (1-\alpha) f_i$, $i = 1, 2$, where f_i is the density function of P_i with respect to P. By using convex combinations $P_{\delta_i} = \delta_i P_1 + (1-\delta_i) P_2$ and

$Q_{\delta_i} = \delta_i Q_1 + (1-\delta_i)Q_2,\ i = 1,2$, with $\delta_1 \neq \delta_2$, from Theorem 3.2, we can assume that P_1 and P_2 have common support, and that T is the common o.t.p. for both solutions. That is, $Q_i = P_i \circ T^{-1}$, for $i = 1,2$. Moreover, in the set $\{a_1 \neq a_2\}$ it is satisfied that $0 < a_1(y) < 1$, so that Theorem 3.11 implies that $T(x) = x$ on this set. But then it is easy to show that there exist sets $A \subset \{a_1 = a_2\}$ and $B \subset \{a_1 < a_2\}$ such that, defining

$$a^*(x) = \begin{cases} 0 & \text{if } x \in A \\ a_2(x) & \text{if } x \in B \\ a_1(x) & \text{if } x \notin A \cup B, \end{cases}$$

thus, $f^* := a^*/(1-\alpha)$ is the density function of a probability, say P^*, in $\mathcal{R}_\alpha(P)$, $Q^* := P^* \circ T^{-1}$ belongs to $\mathcal{R}_\alpha(Q)$ and:

$$\begin{aligned} \mathcal{W}_2^2(P^*, Q^*) &= \int_{R^k} \|x - T(x)\|^2 f^*(x) P(dx) \\ &= \int_{\{a_1 = a_2\} - A} \|x - T(x)\|^2 f_1(y) P(dx) \\ &< \int_{\{a_1 = a_2\}} \|x - T(x)\|^2 f_1(x) P(dx) = \mathcal{W}_2^2(P_1, Q_1). \end{aligned}$$

□

Once we have the uniqueness result given in Theorem 3.12, the generalization of Theorem 3.9 to this framework of double trimming is straightforward.

Theorem 3.13. *Let $\{P_n\}_n, \{Q_n\}_n$, P and Q be in $\mathcal{F}_2(R^k)$, satisfying*

$$\mathcal{W}_2(P_n, P) \to 0,\ \mathcal{W}_2(Q_n, Q) \to 0, \quad \textit{and } P \ll \ell^k.$$

If $P_n^* \in \mathcal{R}_\alpha(P_n)$ and $Q_n^* \in \mathcal{R}_\alpha(Q_n)$ satisfy

$$\mathcal{W}_2(P_n^*, Q_n^*) = \mathcal{W}_2(\mathcal{R}_\alpha(P_n), \mathcal{R}_\alpha(Q_n)),$$

then $\mathcal{W}_2(P_n^*, P^*) \to 0$ and $\mathcal{W}_2(Q_n^*, Q^*) \to 0$, where $P^* \in \mathcal{R}_\alpha(P)$, $Q^* \in \mathcal{R}_\alpha(Q)$ and $W_2(P^*, Q^*) = W_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q))$.

The Strong Law of Large Numbers and the Glivenko-Cantelli Theorem assure (through the uniform integrability argument) that when $\{P_n^\omega\}_n$ is the sequence of empirical probability distributions based on a sequence $\{X_n\}_n$ of independent identically distributed (i.i.d.) random vectors, with law $P \in \mathcal{F}_2(R^k)$, then $\mathcal{W}_2(P_n^\omega, P) \to 0$ for a.s. ω. Therefore the following theorem on the consistency of the trimmed approximations is immediate. This result allows the use of Monte-Carlo simulations to approximate any of the dissimilarity measures $\mathcal{T}_1$ and $\mathcal{T}_2$ between probabilities.

Theorem 3.14 (Consistency). *Let $\{X_n\}_n$, $\{Y_n\}_n$ be two sequences of i.i.d. random vectors with $\mathcal{L}(X_n) = P$, $\mathcal{L}(Y_n) = Q$, $P, Q \in \mathcal{F}_2(R^k)$, and let P_n^ω, Q_n^ω be the empirical distributions based on the samples $\{X_1(\omega), ...X_n(\omega)\}$ and $\{Y_1(\omega), ...Y_n(\omega)\}$.*

(a) *If $Q \ll \ell^k$ and $P_{n,\alpha}^\omega := \underset{P^* \in \mathcal{R}_\alpha(P_n^\omega)}{\text{arg min}} \mathcal{W}_2(P^*, Q)$, then*

$$\mathcal{W}_2(P_{n,\alpha}^\omega, P_\alpha) \to 0 \ \nu\text{-a.s., where } P_\alpha := \underset{P^* \in \mathcal{R}_\alpha(P)}{\text{arg min}} \mathcal{W}_2(P^*, Q).$$

(b) *If $P \ll \ell^k$ and $Q_{n,\alpha}^\omega \in \mathcal{R}_\alpha(Q)$ verifies $\mathcal{W}_2(P_n^\omega, Q_{n,\alpha}^\omega) = \mathcal{W}_2(P_n^\omega, \mathcal{R}_\alpha(Q))$, then*

$$\mathcal{W}_2(Q_{n,\alpha}^\omega, Q_\alpha) \to 0 \ \nu\text{-a.s., where } Q_\alpha := \underset{Q^* \in \mathcal{R}_\alpha(Q)}{\text{arg min}} \mathcal{W}_2(P, Q^*).$$

(c) *If P or $Q \ll \ell^k$ and $P_{n,\alpha}^\omega \in \mathcal{R}_\alpha(P_n^\omega)$ and $Q_{n,\alpha}^\omega \in \mathcal{R}_\alpha(Q)$ satisfy*

$$\mathcal{W}_2(P_{n,\alpha}^\omega, Q_{n,\alpha}^\omega) = W_2(\mathcal{R}_\alpha(P_n^\omega), \mathcal{R}_\alpha(Q)),$$

then $\mathcal{W}_2(P_{n,\alpha}^\omega, P_\alpha) \to 0$ and $\mathcal{W}_2(Q_{n,\alpha}^\omega, Q_\alpha) \to 0$ ν-a.s., where

$$(P_\alpha, Q_\alpha) := \text{arg min}\{\mathcal{W}_2(P^*, Q^*) : \ P^* \in \mathcal{R}_\alpha(P), Q^* \in \mathcal{R}_\alpha(Q)\}.$$

(d) *If P or $Q \ll \ell^k$ and $P_{n,\alpha}^\omega \in \mathcal{R}_\alpha(P_n^\omega)$ and $Q_{n,\alpha}^\omega \in \mathcal{R}_\alpha(Q_n^\omega)$ satisfy*

$$\mathcal{W}_2(P_{n,\alpha}^\omega, Q_{n,\alpha}^\omega) = \mathcal{W}_2(\mathcal{R}_\alpha(P_n^\omega), \mathcal{R}_\alpha(Q_n^\omega)),$$

then $\mathcal{W}_2(P_{n,\alpha}^\omega, P_\alpha) \to 0$ and $\mathcal{W}_2(Q_{n,\alpha}^\omega, Q_\alpha) \to 0$ ν-a.s., where

$$(P_\alpha, Q_\alpha) := \text{arg min}\{\mathcal{W}_2(P^*, Q^*) : \ P^* \in \mathcal{R}_\alpha(P), Q^* \in \mathcal{R}_\alpha(Q)\}.$$

4. Example.

To end the paper, we present in Figure 4 a display showing different levels of similarity between a standard normal distribution, P, and a mixture of normal distributions with variance 1 and means 0 and 4, and respective weights 0.8 and 0.2, $Q = 0.8N(0, 1) + 0.2N(4, 1)$.

From left to right, the columns in the display correspond to respective trimming levels 0, 0.1, 0.15 and 0.2. In descending order, the rows show the results for the best trimming according to $\mathcal{T}_2(P, Q)$, $\mathcal{T}_1(P, Q)$, $\mathcal{T}_1(Q, P)$ and $\mathcal{T}_3(P, Q)$, that is respectively when trimming is allowed in both probabilities, only in Q, only in P, and in both probabilities but with the similarly tailored trimming of Remark 2.9.

A few comments on the similarity shown in these figures are in order. Taking into account that Q can be considered the result of adding a 20% of

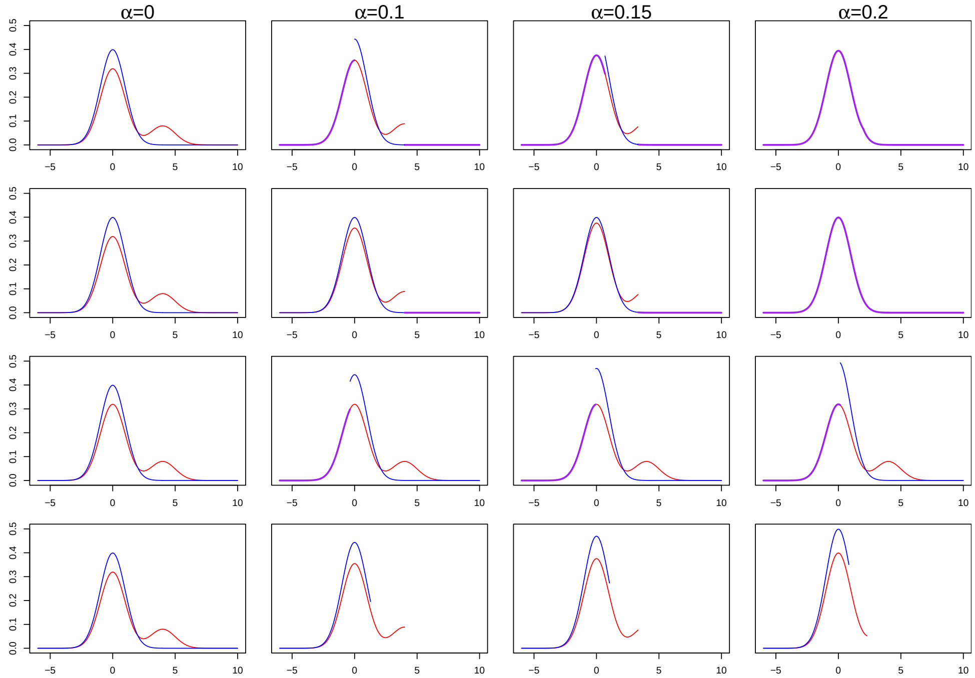

FIG 1. *Densities arising from the minimization of the measures of dissimilarity* $\mathcal{T}_2(P,Q)$, $\mathcal{T}_1(P,Q)$, $\mathcal{T}_1(Q,P)$ *and* $\mathcal{T}_3(P,Q)$ *(top to bottom) with different trimming levels* ($\alpha = 0, 0.1, 0.15$ *and* 0.2*, left to right).* P *is a N(0,1) distribution and* Q *is the mixture 0.8 N(0,1) + 0.2 N(4,1). The figures show the densities of the probabilities obtained as best trimmed approximations of* P *(blue) and* Q *(red).*

contamination to P, it is obvious that $\mathcal{T}_1(P,Q)$ and $\mathcal{T}_2(P,Q)$ should be 0 for every $\alpha \geq 0.2$. This is what happens in the first two rows. In fact, it can be checked that $\mathcal{T}_1(P,Q) > 0$ for every $\alpha < 0.2$. However, $\mathcal{T}_2$ allows to move P a bit closer to Q and then, $\mathcal{T}_2(P,Q) = 0$ even at level 0.1909.

On the other hand, it is impossible to obtain Q by simply trimming P. Thus, $\mathcal{T}_1(Q,P) > 0$ for every trimming level α. The same happens with $\mathcal{T}_3(P,Q)$ because the differences between P and Q can not be eliminated through a similarly tailored trimming.

It is also worth to pay some attention to the differences in the o.t.p.'s associated to the considered trimmings. The small bump in the density of Q is responsible for most of the dissimilarity between P and Q. Optimal trimming tries to decrease the Q density on the right tail whenever it is possible, as it is the case in the first two rows. In such cases there is true trimming ($a_1 < 1$ in Theorem 3.11) and there is no mass transportation in this range. A secondary source of dissimilarity comes from the different scale between the P density and the main bump in the density of Q. When trimming is allowed in P, the P density is decreased on the left tail and there is no mass transportation on the left, as in the first and third rows in the display. Note that in the first row, true mass transportation happens only in the central region. On the opposite, in the fourth row the trimming function is always zero or one and there is true mass transportation on the non-trimmed range (the o.t.p. is a.s. different from the identity).

This example stresses, in a descriptive way, the differences between the measures of dissimilarity considered through the paper. In particular intuition mostly agrees with the use of $\mathcal{T}_2$, while the right use of $\mathcal{T}_3$ should involve some extra caution in practice.

References.

[1] Álvarez-Esteban P. C.; del Barrio, E.; Cuesta-Albertos, J. A. and Matrán, C. (2008). Trimmed comparison of distributions. To appear in *J. Amer. Statist. Assoc.*

[2] Araujo, A. and Giné, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables.* John Wiley & Sons.

[3] Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.,* 9, 1196-1217.

[4] Cascos, I. and López-Díaz, M. (2005). Integral trimmed regions. *J. Multivariate Anal.* 96, 404-424.

[5] Cascos, I. and López-Díaz, M. (2008). Consistency of the α-trimming of a probability. Applications to central regions. *Bernoulli,* 14(2), 580-592.

[6] Cuesta-Albertos, J. A. and Matrán, C. (1989). Notes on the Wasserstein metric in Hilbert spaces. *Ann. Probab.,* 17, 1264-1276.

[7] Cuesta-Albertos, J. A.; Matrán, C. and Tuero, A. (1997). Optimal Transportation Plans and Convergence in Distribution. *J. Multivariate Anal.,* 60, 72-83.

[8] Cuesta-Albertos, J. A.; Matrán, C. and Tuero, A. (1997). On the monotonicity of optimal transportation plans. *J. Math. Anal. Appl.,* 215, 86-94.

[9] Dudley, R.M. (1989). *Real Analysis and Probability.* Wadsworth and Brooks-Cole, CA.

[10] Evans, L. C. and Gariepy, R. F. (1992). *Measure Theory and Fine Properties of Functions.* Studies in Advanced Mathematics. CRC Press. Boca Raton.

[11] Gordaliza, A. (1991). Best approximations to random variables based on trimming procedures. *J. Approx. Theory,* Vol. 64, No. 2, 162-180.

[12] Heinich, H. and Lootgieter, J-C. (1996). Convergence des fonctions monotones. *C. R. Acad. Sci. Paris,* t. 322, Série I, 869-874.

[13] McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.* 80, 309-323.

[14] Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems. (2 vol.)* Springer Series in Statistics. Probability and its Applications. Springer. New York.

[15] Rockafellar, R. T. (1970). *Convex Analysis.* Princeton University Press.

[16] Rüschendorf, L. and Rachev, S. T. (1990). A characterization of random variables with minimum L^2-distance. *J. Multivariate Anal.,* 32, 48-54.

[17] Tuero, A. (1993). On the stochastic convergence of representations based on Wasserstein metrics. *Ann. Probab.,* 21, 72-85.

[18] Villani, C. (2003). *Topics in Optimal Transportation.* Graduate Studies in Mathematics Vol. 58, Amer. Math. Soc.

Dept. Matemáticas, Estadística y Computación
Universidad de Cantabria
Avda. los Castros s.n.
39005 SANTANDER
Spain
e-mail: cuestaj@unican.es

Dept. de Estadística e Investigación Operativa
Universidad de Valladolid
Prado de la Magdalena s.n.
47005 VALLADOLID
Spain
e-mail: pedroc@eio.uva.es
tasio@eio.es
matran@eio.uva.es