Uniqueness and Approximated Computation of Optimal Incomplete Transportation Plans*

P. C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, C. Matrán.
Universidad de Cantabria and Universidad de Valladolid

October 14, 2008

Abstract

For a given trimming level $\alpha \in (0,1)$ an $\alpha-$trimmed version, P^*, of a probability P is a new probability obtained by re-weighting the probability of any Borel set, B, according a positive weight function, $f \leq \frac{1}{1-\alpha}$, in the way $P^*(B) = \int_B f(x)P(dx)$.

If P, Q are probability measures on an euclidean space, we consider the optimization problem of obtaining the best L_2-Wasserstein approximation between say a fixed probability and trimmed versions of the other, say trimmed versions of both probabilities. These best trimmed approximations naturally lead to new perspectives in the theory of Mass Transportation, 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 the uniqueness of the optimal solution. Notice that such solution involves an optimal map T transporting some trimmed version P^* of P to some other Q^* of Q, thus for any point x in the support of P the weight function associated to P^* allows to partially or completely avoid the consideration of x in the transport. Our results show that in fact only the non-trimmed points (verifying $f(x) = \frac{1}{1-\alpha}$) are transported, while the partially trimmed points (verifying $0 < f(x) < \frac{1}{1-\alpha}$) must remain untransported by T.

Keywords Approximated Computation, Incomplete Mass Transporting, Mass Transportation Problem, Multivariate Distributions, Optimal Transportation Plan, Similarity, Trimming, Uniqueness, Trimmed Probability.
A.M.S. classification: Primary 49Q20, 60A10. Secondary: 60B10, 28A50.

1 Introduction

This paper considers a modified version of the classical Mass Transportation Problem (MTP in the sequel). Broadly speaking, the MTP can be formulated as trying to relocate a certain amount of mass with a given initial distribution to another target distribution in such a way that the transportation cost is minimized. This seemingly simple problem has a long history which dates back to Monge. The initial formulation of the problem can be summarized in present-day language as follows. Let P_1, P_2 be two probability

*Research partially supported by the Spanish Ministerio de Educación y Ciencia and FEDER, grants MTM2008-06067-C02-01 and 02 and by the Consejería de Educación de la Junta de Castilla y León.

measures on the Euclidean space R^k with norm $\|\cdot\|$ and Borel $\sigma-$field β. Consider the set, $\mathcal{T}(P_1, P_2)$, of maps transporting P_1 to P_2, that is, the set of all measurable maps $T : R^k \to R^k$ such that, if the initial space is endowed with the probability P_1, then the distribution of the random variable T is P_2. Then Monge's problem consists of finding a transportation map, T_0, from P_1 to P_2 such that

$$T_0 := \arg\min_{T\in\mathcal{T}(P_1,P_2)} \int_{R^k} \|x - T(x)\| P_1(dx).$$

A later, fundamental generalization of this problem is the so-called Kantorovitch-Rubinstein-Wasserstein (KRW) formulation which consists in finding

$$\mathcal{W}_2^2(P_1, P_2) := \inf_{\pi\in\mathcal{M}(P_1,P_2)} \left\{ \int \|x - y\|^2 d\pi(x, y) \right\}, \tag{1}$$

where $\mathcal{M}(P_1, P_2)$ is the set of finite, positive measures on $\beta \times \beta$ with marginals P_1 and P_2.

Apart from the consideration of different cost functions, the main difference between the Monge and the KRW problem is that the later is not related to transportation maps. We mean that in the KRW formulation masses sharing the same initial position may end up in different locations. The KRW minimization allows also to consider the L_2-Wasserstein distance, $\mathcal{W}_2(P_1, P_2)$, between probability measures with finite moment of order two (see e.g. Bickel and Freedman [3] for details and properties of $\mathcal{W}_2$). Remarkably, the Monge and the KRW formulations turn out to be equivalent under some smoothness assumptions.

Existence, uniqueness or regularity of mappings $T \in \mathcal{T}(P_1, P_2)$ satisfying $\int_{R^k} \|x - T(x)\|^2 dP_1(x) = \mathcal{W}_2^2(P_1, P_2)$ are problems that have attracted the attention of mathematicians from very different points of view. Fluid Mechanics, Partial Differential Equations, Optimization, Probability Theory and Statistics are in the very broad range of applications of this and related MTP's justifying the interest and also the different technical approaches for their study. To avoid a formidable amount of references we refer to the books by Rachev and Rüschendorf [16] and by Villani [19] for an updated account of the interest and implications of the problem, as well as to recent works illustrating the permanent actuality of the topic, as Ambrosio [2], Caffarelli et al. [4], or Feldman and McCann [12].

Here we will analyze a variant of the KRW problem involving incomplete mass transportation. Let us introduce it through a motivating example. Gangbo and McCann consider in [13] the problem of identifying a leaf l_0 by comparing it with a catalog. They analyze the approach based on minimizing the transportation cost between the uniform distribution on the outline of l_0 and its couterparts in the catalog. To avoid technicalities, we assume that we are dealing with black and white pictures of l_0 and the leaves in the catalog, rather than their outlines. We identify the grey-levels with the density of a measure, compute the associated L_2-Wasserstein distances and identify l_0 with the closest leaf in the catalog.

Now, let us assume that, as it often happens, the picture of l_0 is corrupted at some spots. It seems reasonable to delete those spots before making the comparisons. However, it is

not always easy to tell a corrupted spot from a distinctive feature. A reasonable procedure would be to transport only a part of the initial mass, dismissing a small fraction, to minimize the transportation cost. If the leaves in the catalog are also corrupted by noise, then we should also allow some fraction of the target picture to remain unmatched.

In a natural way we end up in the problem of dismissing a (usually small) fraction of the masses represented by P_1 and P_2. This process resembles the trimming procedures employed in Statistics where, very often, outlying observations have to be deleted.

Trimming procedures are a common practice in Statistics. In the general setting considered here, the more suitable approach seems to be the one in Gordaliza [14]. If $\mathcal{P}(R^k)$ denotes the set of all Borel probability measures on R^k, the definition of a trimming of a probability which we employ here is the following:

Definition 1.1 *Given $0 \leq \alpha \leq 1$ and $P \in \mathcal{P}(R^k)$, we say that $P^* \in \mathcal{P}(R^k)$ is an α-trimming of P if P^* is absolutely continuous with respect to P, and $\frac{dP^*}{dP} \leq \frac{1}{1-\alpha}$. The set of all α-trimmings of P will be denoted by $\mathcal{R}_\alpha(P)$.*

This definition of trimming is more general than other usual alternatives in that it allows points to be partially trimmed. Cascos and López-Díaz [6] and [7] and our work [1] analyze some properties of these trimmings.

With this definition, the above outlined problem can be stated as follows. Consider $P_1, P_2 \in \mathcal{P}(R^k)$ with finite second order moment. Given $\alpha \in (0,1)$, we define the following measures of dissimilarity between P_1 and P_2 at level α:

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

Note that P_1 and P_2 do not play symmetric roles in $\mathcal{T}_1$ in applications. For instance, P_2 plays the role of l_0 and P_1 the role of one of the standards in the catalog in the case of a perfect, noise-free catalog. In the case of a noisy catalog $\mathcal{T}_2$ is the right choice.

An interesting feature of $\mathcal{T}_1$ and $\mathcal{T}_2$ is that the noisy spots do not have to be fixed in advance. In fact, the only requirement is to take α as an upper bound of the proportion of corrupted spots in the pictures; then, the procedure automatically determines the spots in the picture being noise by minimizing the transportation cost.

An analysis of similarity of distributions based on the comparison of trimmed versions of them has recently been developed in Álvarez-Esteban et al [1]. In fact, this paper can be considered as a generalization of [1]. The novelty of the approach in [1] consists in considering that two distributions are similar at level α whenever suitable chosen α-trimmed versions of such distributions coincide. The proposal focused on probability measures on the real line, and the same trimming pattern on both probabilities (see precise definitions in Remark 2.5). The measures $\mathcal{T}_1$ and $\mathcal{T}_2$ can also be employed in this setting. For instance, let us assume that we have a sample which we suspect that has been generated from a probability distribution P. If we compute the empirical distribution, P_n, associated to the sample at hand, then $\mathcal{T}_1(P,P_n)$ is appropriate to analyze if the sample

has been obtained from P plus a contamination. In problems where some observations have been censored, the right measure of closeness is $\mathcal{T}_1(P_n, P)$. If both facts are present, then the measure of choice is $\mathcal{T}_2(P_n, P)$. The statistical applications of these measures of similarity are the object of current research. In particular, the results in this paper are essential in the design and analysis of a similarity test to be reported soon.

The main results in this paper are given in Section 2. We show that, as in the classical optimal transportation problem, the MKW and Monge problems are equivalent under absolute continuity. We also prove the uniqueness of the optimal transportation plan. From the technical point of view, the most remarkable (and difficult) result concerns the uniqueness of the best pair of trimmed probabilities solving the corresponding minimization problems (Theorems 2.11 and 2.15). Remember that the trimmings that we employe allow to partially trim some points. In fact Theorem 2.14 shows that only the mass placed on non-trimmed points is transported, while the mass on partially trimmed points must remain fixed. Once this work was completed we learned about a recent paper by Caffarelli and McCann [5] where the problem of transporting a fraction of whole mass is also considered. Although their motivation and approach are very different, a main goal of that work is th analysis of the uniqueness of the optimal transportation plan. Our uniqueness result greatly improves the one obtained there, where disjoint supports for the initial measures is imposed.

Since there are not general explicit expressions for the solutions of the mutidimensional KRW problem, it is of primary interest to analize the possibilities of Monte-Carlo approximations and this will be also analyzed in Section 2 where we obtain a simple result that allows to represent the trimmings of any probability in terms of those of another (see Corollary 2.4). By handling these representations it is easy to prove the convergence of trimmings of convergent sequences (Lemma 2.7), which allows to obtain the consistency of the introduced dissimilarity measures (Theorem 2.17).

Finally Section 3 explores, through 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. The Lebesgue measure on the space (R^k, β) will be denoted by ℓ^k, while $\mathcal{F}_2(R^k)$ will denote the set of distributions in $\mathcal{P}(R^k)$ with finite second moment.

Given $P, Q \in \mathcal{P}(R^k)$, by $P \ll Q$ we will denote absolute continuity of P with respect to (w.r.t) Q, and by $\frac{dP}{dQ}$ the corresponding Radon-Nykodym derivative. By $\text{supp}(P)$ we will denote the support of P and by $P(\cdot|B)$ the conditional probability distribution given the set B. With a slight abuse of notation, given $P \in \mathcal{F}_2(R^k)$ and $\Theta, \Theta^* \subset \mathcal{F}_2(R^k)$, we will often denote

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

Unless otherwise stated, the random vectors will be assumed to be defined on the same probability space (Ω, σ, ν). Weak convergence of probabilities will be denoted by $\rightarrow_w$ and $\mathcal{L}(X)$ will denote the law of the random vector X.

2 Trimmings and Best Trimmed Approximations

We begin collecting some notation and properties of trimmed probabilities. A more detailed analysis can be found in [7]. From the definition of $\mathcal{R}_\alpha(P)$ it is obvious that $P^* \in \mathcal{R}_\alpha(P)$ if and only if $P^* \ll P$ and $\frac{dP^*}{dP} = \frac{1}{1-\alpha} f$ with $0 \leq f \leq 1$. Thus, the trimmings that we are handling allow to reduce the weight of some regions of the measurable space without completely removing them from the feasible set.

The following propositions contain some useful facts about trimmings that can be easily obtained (see also [7]).

Proposition 2.1 *For any probability measure, $P \in \mathcal{P}(R^k)$,*

(a) *If $\alpha < 1$ then $\mathcal{R}_\alpha(P)$ is compact for the topology of weak convergence.*

(b) *If $\alpha < 1$, $\{P_n\}_n$ in $\mathcal{P}(R^k)$ 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 $P^* \in \mathcal{R}_\alpha(P)$.*

Proposition 2.2 *Let $Q \in \mathcal{P}(R^k)$. 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:

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

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 $\frac{dP^*}{dP}(x) \leq \frac{1}{1-\alpha}$ a.s. (P) and $P = Q \circ T^{-1}$, also $\frac{dP^*}{dP}(T(y)) \leq \frac{1}{1-\alpha}$ a.s.(Q) hence $Q^* \in \mathcal{R}_\alpha(Q)$. •

Regarding the L_2-Wasserstein distance, it is well known (see e.g. [3]) that when $P, Q \in \mathcal{F}_2(R^k)$ the infimum in (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-optimal transport plan (L_2-o.t.p.) for (P, Q). (L_2-optimal coupling for (P, Q) is an alternative, sometimes used, terminology).

In Cuesta-Albertos and Matrán [8] (see also Rüschendorf and Rachev [17] and McCann [15]) 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 [15]). 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. [8], [9], [10], and Tuero [18]. A different approach, involving more analytical proofs, is summarized in [19].

Proposition 2.3 *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) defined on some (irrelevant) probability space (Ω, σ, ν). 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 T is an o.t.p. for (P, Q), then T is a.e. continuous on supp(P).*

(d) *Let $Q_n \in \mathcal{F}_2(R^k)$ such that $Q_n \to_w Q$, and $P \ll \ell^k$, and let T_n be o.t.p.'s for (P, Q_n). Then $T_n \to T$, P-a.s.*

Now let us return to the consideration of trimmed probabilities. From Proposition 2.2 and Proposition 2.3 (b) it obviously arises the following characterization.

Corollary 2.4 *If $P_0, Q \in \mathcal{F}_2(R^k)$, 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 2.5 Once we have chosen a particular probability measure $P_0 \in \mathcal{F}_2(R^k)$, $P_0 \ll \ell^k$, Corollary 2.4 allows to induce *trimmed versions similarly tailored according to the shape of* P_0: If $P_1, P_2 \in \mathcal{F}_2(R^k)$ 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 $P_0^* \in \mathcal{R}_\alpha(P_0)$ determines the pair (P_1^*, P_2^*) where $P_1^* = P_0^* \circ T_1^{-1}, P_2^* = P_0^* \circ T_2^{-1}$ and $P_1^* \in \mathcal{R}_\alpha(P_1)$, $P_2^* \in \mathcal{R}_\alpha(P_2)$ and we call them similarly tailored because they depend on the same trimming of P_0.

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

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

That was the kind of trimming adopted in [1] for probabilities on the real line with the U(0,1) law as distribution of reference.

From the definition of trimming, if $P \in \mathcal{F}_2(R^k)$ 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 2.1 (a) for the metric $\mathcal{W}_2$.

Proposition 2.6 *If $0 < \alpha < 1$ and $P \in \mathcal{F}_2(R^k)$, 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 order moments (Bickel and Freedman [3], Lemma 8.3). Since $\mathcal{R}_\alpha(P)$ is tight (Proposition 2.1 (a)), 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 Q be 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. •

Proposition 2.3 (d) and Corollary 2.4 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 2.7 *Let $0 < \alpha < 1$, $\{Q_n\}_n$ and Q be in $\mathcal{P}(R^k)$, and assume that $Q_n \to_w Q$. Then, if $Q^* \in \mathcal{R}_\alpha(Q)$, there exists a sequence $\{Q_n^*\}_n$ such that $Q_n^* \in \mathcal{R}_\alpha(Q_n)$, for all n, and $Q_n^* \to_w Q^*$.*

PROOF.- Let $P \in \mathcal{P}(R^k)$ 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, Proposition 2.3 (d) implies that $T_n \to T$, $P-$a.s.

By Corollary 2.4 $Q^* = P^* \circ T^{-1}$ 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 characterization in Corollary 2.4. Since $T_n \to T$, $P-$a.s., and $P^* \ll P$, also $T_n \to T$, P^*-a.s. Therefore $Q_n^* = P^* \circ T_n^{-1} \to_w P^* \circ T^{-1} = Q^*$. •

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 $\mathcal{W}_2^2(\gamma P_1 + (1-\gamma)P_2, Q) \leq \gamma\mathcal{W}_2^2(P_1, Q) + (1-\gamma)\mathcal{W}_2^2(P_2, Q)$, $\gamma \in (0,1)$, but when $Q \ll \ell^k$, property (a) in Proposition 2.3 leads to more:

Theorem 2.8 *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:

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

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

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

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 2.3 implies that a random map cannot be an o.t.p, thus T should be non-random, leading to

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

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 2.8, we obtain the following corollary, stating the strict convexity of $\mathcal{W}_2^2(P, \cdot)$, leading to a trivial proof for the convergence results in Theorem 2.12.

Corollary 2.9 *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 $(0,1)$,*

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

Proposition 2.6 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 2.10 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 2.9 ensures the uniqueness of the best trimmed approximation.

Theorem 2.11 *Assume that P and Q, belong to $\mathcal{F}_2(R^k)$ and that $P \ll \ell^k$. Then, for every $0 < \alpha < 1$, 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 2.12 *Let $\{P_n\}_n$, P and Q be in $\mathcal{F}_2(R^k)$, such that $\mathcal{W}_2(P_n, P) \to 0$. Let $0 < \alpha < 1$, then*

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

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

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

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

PROOF.- Both statements have similar proofs, so let us consider only statement a). By Proposition 2.1 (a) the sequence $\{P_{n,\alpha}\}_n$ is tight and by the same argument that in the proof of Proposition 2.6, 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 2.1 (b), if $P_{r_n,\alpha} \to_w P^*$, then $P^* \in \mathcal{R}_\alpha(P)$ and, therefore

$$\mathcal{W}_2(P_\alpha, Q) \leq \mathcal{W}_2(P^*, Q) = \lim \mathcal{W}_2(P_{r_n,\alpha}, Q) \leq \liminf \mathcal{W}_2(P^*_{r_n,\alpha}, Q), \tag{2}$$

for any choice $P^*_{r_n,\alpha} \in \mathcal{R}_\alpha(P_{r_n})$. Lemma 2.7 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 (2) and with the uniqueness of the best trimmed approximation P_α given by Theorem 2.11 shows that $P^* = P_\alpha$. •

2.1 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 2.13 *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 [11]), 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 2.14 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 $\text{supp}(P) - \text{supp}(Q)$.

Theorem 2.14 *Let $\alpha > 0$, and let $P, Q \in \mathcal{P}(R^k)$. 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 (c) in Proposition 2.3 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,

$$\|T(y) - \pi_0[T(y)]\| > r_y, \text{ for every } y \in A. \tag{3}$$

On the other hand, we have

$$P[A] = \int_{H_0} P_z(A_z) P^\circ(dz) = \int_{H_0} \lambda_z(z_{v_0}) P^\circ(dz). \tag{4}$$

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

$$P^\circ\{z \in H_0 : \lambda_z(z_{v_0}) > 0\} > 0. \tag{5}$$

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) lead to

$$\begin{aligned} \|y_1 - T(y_1)\|^2 &= \|T(y_1) - \pi_0[T(y_1)] + \pi_0(y_1) - y_1 + \pi_0(T(y_1)) - \pi_0(y_1)\|^2 \\ &= \left(r_{T(y_1)} - r_{y_1}\right)^2 + \|\pi_0[T(y_1)] - z\|^2 \\ &> \left(r_{T(y_1)} - r_{y_2}\right)^2 + \|\pi_0[T(y_1)] - \pi_0(y_2)\|^2 \\ &= \|y_2 - T(y_1)\|^2. \end{aligned} \tag{6}$$

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

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

From Lemma 2.13 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

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

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

$$\begin{aligned} \int_{R^k} a^*(y)P(dy) &= \int_{R^k} a_1(y)P(dy) \\ &\quad - \int_{A^-} \min\{1 - a_1[\Gamma(y)], a_1(y)\} P(dy) \\ &\quad + \int_{A^+} \min\left\{1 - a_1(y), a_1[\Gamma^{-1}(y)]\right\} P(dy). \end{aligned} \tag{7}$$

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

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

$$= \int_{H_0} \left(\int_{A_z^-} \min\left\{1 - a_1(\Gamma(y)), a_1(y)\right\} P_z(dy) \right) P^\circ(dz)$$
$$= \int_{A^-} \min\left\{1 - a_1[\Gamma(y)], a_1(y)\right\} P(dy), \tag{8}$$

what, joined to (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 2.8, 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 (6), we have

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

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 2.15 (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*

$$\mathcal{W}_2^2(P_1, Q_1) = \mathcal{W}_2^2[\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)]. \tag{9}$$

PROOF.- Assume that (P_1, Q_1) and (P_2, Q_2) are two different pairs fulfilling (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 2.8, 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 2.14 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 2.15, the generalization of Theorem 2.12 to this framework of double trimming is straightforward.

Theorem 2.16 *Let $\{P_n\}_n, \{Q_n\}_n$, P and Q be in $\mathcal{F}_2(R^k)$, satisfying*

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

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 2.17 (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 $P_{n,\alpha}^\omega := arg\ min_{P^* \in \mathcal{R}_\alpha(P_n^\omega)} \mathcal{W}_2(P^*, Q)$, then*

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

(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*

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

(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

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

(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

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

3 Example

To end the paper, we present in Figure 3 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 2.5 using the $U(0,1)$ distribution as reference.

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 contamination to P, it is

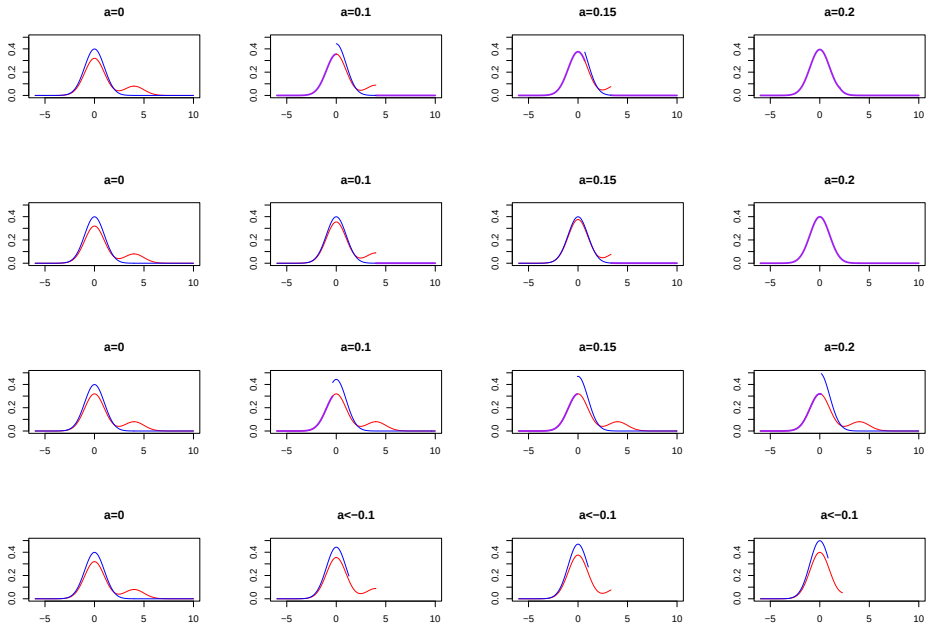

Figure 1: Densities arising from the minimization of the measures of dissimilarity $\mathcal{T}_2(P,Q)$, $\mathcal{T}_1(P,Q)$, $\mathcal{T}_1(Q,P)$ and $\mathcal{T}_3(P,Q)$ (top to bottom) with different trimming levels ($\alpha = 0, 0.1, 0.15$ and 0.2, left to right). P is a N(0,1) distribution and 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).

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 2.14) 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. *J. Amer. Statist. Assoc. (Theory and Methods)* 103, No. 482, 697-704.

[2] Ambrosio, L. (2003). *Lecture Notes on Optimal Transport Problems, Mathematical Aspects of Evolving Interfaces*, Lecture Notes in Math. vol. 1812, pp. 12. Springer-Verlag, Berlin/New York.

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

[4] Caffarelli, L.A., Feldman, M. and McCann, R.J.(2002). Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs. *J. Amer. Math. Soc.,* 15, 1-26.

[5] Caffarelli, L.A. and McCann, R.J.(2008). Free boundaries in optimal transport and Monge-Ampére obstacle problems. *Ann. of Math.*, to appear.

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

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

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

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

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

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

[12] Feldman, M and McCann, R.J. (2002). Uniqueness and transport density in Monge's mass transportation problem. *Calc. Var.,* 15, 81-113.

[13] Gangbo, W. and McCann, R.J. (2000). Shape recognition via Wasserstein distance. *Quart. Appl. Math.,* 58, 705-737.

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

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

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

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

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

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