Application du Transport Optimal en Fair Learning

Paula Gordaliza [1], Eustasio del Barrio [2] & Jean-Michel Loubes [3]

[1] *IMUVA, Universidad de Valladolid & Institut de Mathématiques de Toulouse*
paula.gordaliza@math-univ.toulouse.fr
[2] *IMUVA, Universidad de Valladolid*
tasio@eio.uva.es
[3] *Institut de Mathématiques de Toulouse*
loubes@math-univ.toulouse.fr

Résumé. Nous fournissons un théorème de limite centrale pour la distance de Monge-Kantorovich entre deux distributions empiriques de tailles n et m, $\mathcal{W}_p(P_n, Q_m),\ p \geq 1$, pour les observations sur la droite réelle. Dans le cas $p > 1$, nos hypothèses sont précises en termes de moments et de finesse. Nous prouvons des résultats concernant le choix des constantes de centrage. Nous fournissons une estimateur consistant de la variance asymptotique qui permet de construire tests sur deux échantillons et des intervalles de confiance pour certifier la similarité entre deux distributions. Celles-ci sont ensuite utilisées pour évaluer un nouveau critère d'équité des ensembles de données dans la classification.

Mots-clés. Transport Optimal, Distance de Monge-Kantorovich, Théorème de Limite Centrale, Apprentissage Juste.

Abstract. We provide a Central Limit Theorem for the Monge-Kantorovich distance between two empirical distributions with sizes n and m, $\mathcal{W}_p(P_n, Q_m),\ p \geq 1$, for observations on the real line. In the case $p > 1$ our assumptions are sharp in terms of moments and smoothness. We prove results dealing with the choice of centering constants. We provide a consistent estimate of the asymptotic variance which enables to build two sample tests and confidence intervals to certify the similarity between two distributions. These are then used to assess a new criterion of data set fairness in classification.

Keywords. Optimal Transport, Monge-Kantorovich distance, Central Limit Theorem, Fair Learning.

1 Context, aim and scope of the paper

Applications of optimal transportation methods have witnessed a huge development in recent times, in a variety of fields, including machine learning and image processing, among others. The number of significant breakthroughs in the involved numerical procedures can help to understand some of the reasons for this interest. We refer to Chizat et al. (2018)

for a more detailed account. In the particular field of statistical inference, despite some early contributions (see, e.g., Munk and Czado (1998), del Barrio, Cuesta-Albertos et al. (1999), del Barrio et al. (2005) or Freitag et al. (2007)), progress has been more slow. Among the reasons for this different rythm we can quote the claim from Sommerfeld and Munk (2018) that transportation cost distance 'is an attractive tool for data analysis but statistical inference is hindered by the lack of distributional limits'. Let us try to give a more complete perspective on this claim.

With inferential goals in mind, the main object of interest is the transportation cost between two sets of random points or between an empirical and a reference measure. In the, by now classical, Kantorovich formulation, for probabilities P and Q on $\mathbb{R}^d$ a transportation plan is a joint probability, say π, on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals P and Q. The associated transportation cost is $I[\pi] = \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y) d\pi(x,y)$, where c is some cost function, and the optimal transportation cost is the minimal value of $I[\pi]$ among all choices of transportation plans, π, between P and Q. The problem admits a much more general formulation, but for our present purposes it is enough to know that for the choice $c(x,y) = c_p(x,y) = \|x-y\|^p$, $p \geq 1$, if we denote by $\mathcal{W}_p^p(P,Q)$ the corresponding optimal transportation cost, then $\mathcal{W}_p$ defines a metric in the set $\mathcal{F}_p(\mathbb{R}^d)$ of probabilities on $\mathbb{R}^d$ with finite p-th moment. We refer to Villani (2003) for general background on these facts.

If we observe $X_1, \ldots, X_n$ i.i.d. P, $Y_1, \ldots, Y_m$ i.i.d. Q and write P_n and Q_m for the associated empirical measures, then, assuming that P and Q have finite p-th moment it is well-known that $\mathcal{W}_p^p(P_n, Q) \to \mathcal{W}_p^p(P,Q)$ and $\mathcal{W}_p^p(P_n, Q_m) \to \mathcal{W}_p^p(P,Q)$ almost surely. Enhancing this result with a distributional limit theorem would yield a useful inferential tool in different problems. Early work focused on the case $P = Q$. From an inferential point of view this corresponds to goodness-of-fit problems, with a distributional limit result providing approximate distributions under the null model $P = Q$. In this line we must cite Atjai et al. (1984) and Talagrand and Yukich (1993) dealing with the case when $P = Q$ is the uniform distribution on the unit hypercube, with later contributions (see Dobrić and Yukich (1995), Fournier and Guillin (2015)) covering an increasingly wider setup. These references dealt with general dimension d, but were not satisfactory for inferential goals, since they only dealt with rates of convergence. Until very recently, distributional limits were only available in the one-dimensional case ($d = 1$). In this case, if $p = 1$ then, under some integrability assumptions $\mathcal{W}_1(P_n, P) = O_P(n^{-1/2})$, with $\sqrt{n}\mathcal{W}_1(P_n, P)$ converging weakly to a non Gaussian limit, see del Barrio, Giné et al. (1999). If $p > 1$ then it is still possible to get a limiting distribution for $\sqrt{n}\mathcal{W}_p(P_n, P)$, but now integrability assumptions are not enough and the available results require some smoothness conditions on P (and on its density), see del Barrio, Cuesta-Albertos et al. (1999) and del Barrio et al. (2005) for the case $p = 2$. Some degree of smoothness (absolute continity of P with positive density on an interval) is, in fact, necessary for boundedness of the sequence $\sqrt{n}E(\mathcal{W}_p(P_n, P))$ if $p > 1$, see Bobkov and Ledoux (2014).

In some statistical applications (in bioequivalence testing, but also in the application to fair learning that we present later) the goal is to provide some statistical certification

that the data are not too far from a model, say homogeneity, $P = Q$. Not rejecting the null $H_0 : P = Q$ would be a mere sanity check, but would not provide statistical evidence that the null holds (even approximately). However, this kind of evidence would be granted from rejection of the null $H_0 : \rho(P, Q) \geq \Delta_0$ for some distance ρ. Computation of approximate p-values in this setup would be possible through distributional limit theory for the case $P \neq Q$. Hence, in the case of transportation cost metrics it would be useful to prove a CLT for

$$r_n\big(\mathcal{W}_p^p(P_n, Q) - a_n\big) \tag{1.1}$$

for some centering a_n and scaling $r_n > 0$ (and similarly for the two-sample case) in the case $P \neq Q$. It would be also useful to guarantee that we can take $a_n = \mathcal{W}_p^p(P, Q)$ as centering constants.

For the metric $\mathcal{W}_2$ (or a trimmed version of it) some limiting results for (1.1) were given in Munk and Czado (1998) for one-dimensional data. More recently, Sommerfeld and Munk (2018) handle d-dimensional data and general p, but it is constrained to the case when P and Q are finitely supported (extensions to probabilities with countable support are given in Tameling et al. (2017)). The picture is less complete in the case of continuous distributions. Back to the case $p = 2$, a CLT in general dimension has been provided in del Barrio and Loubes (2017): if Q has a positive density in the interior of its convex support and P and Q have finite moments of order $4 + \delta$ for some $\delta > 0$ then

$$\sqrt{n}\big(\mathcal{W}_2^2(P_n, Q) - E(\mathcal{W}_2^2(P_n, Q))\big) \to_w N(0, \sigma^2(P, Q)) \tag{1.2}$$

for some $\sigma^2(P, Q)$ which is not null if and only if $P \neq Q$. A two-sample version of such results are also given in this work. Note that $\to_w$ denotes weak convergence in probabilities.

In this paper we provide extensions of (1.2) to general distances $\mathcal{W}_p$, $p \geq 1$. We cover only the case of one-dimensional data. In turn, from a probabilistic point of view the main contributions of this paper are that *i)* we prove the analogue of (1.2) for general $p > 1$ under sharp moment and smoothness assumptions:

Theorem 1.1 (Central Limit Theorem for $\mathcal{W}_p$ with $p > 1$) *Assume that $F, G \in \mathcal{F}_{2p}$ and G^{-1} is continuous on $(0, 1)$ and $p > 1$. Then*

(i) If $X_1, \ldots, X_n$ are i.i.d. F and F_n is the empirical d.f. based on the X_i's

$$\sqrt{n}(\mathcal{W}_p^p(F_n, G) - E\mathcal{W}_p^p(F_n, G)) \to_w N(0, \sigma_p^2(F, G)).$$

(ii) If, furthermore, F^{-1} is continuous, $Y_1, \ldots, Y_m$ are i.i.d. G, independent of the X_i's, G_m is the empirical d.f. based on the Y_j's and $\frac{n}{n+m} \to \lambda \in (0, 1)$ then

$$\sqrt{\tfrac{nm}{n+m}}(\mathcal{W}_p^p(F_n, G_m) - E\mathcal{W}_p^p(F_n, G_m)) \to_w N(0, (1-\lambda)\sigma_p^2(F, G) + \lambda\sigma_p^2(G, F)).$$

ii) we show that in the case $p = 1$, when strict convexity of the cost function is lost, non-normal limits can occur, even in the case $P \neq Q$. For the statistical applications that we present, the centering constants in the former CLT's are of crucial importance. We provide general conditions under which $E(\mathcal{W}_p^p(P_n, Q))$ can be replaced by $\mathcal{W}_p^p(P, Q)$ as centering constant in (1.2). Combined with a consistent estimator of the asymptotic variance in the CLT's, this enables us to define a consistent test

$$H_0 : \mathcal{W}_p(P, Q) \geq \Delta_0 \quad \text{vs} \quad H_a : \mathcal{W}_p(P, Q) < \Delta_0, \tag{1.3}$$

that is, a consistent method for gathering statistical evidence to ensure if $\mathcal{W}_p(P, Q) < \Delta_0$.

We would like to note at this point that our approach to prove Theorem 1.1 uses the fact that if P and Q are probabilities on the real line with distribution functions (d.f.'s) F and G, respectively, then $\mathcal{W}_p^p(P, Q)$ is simply the L_p-distance between quantile functions

$$\mathcal{W}_p^p(P, Q) = \int_0^1 |F^{-1} - G^{-1}|^p. \tag{1.4}$$

(see, Remark 2.19 in Villani (2003)). For this reason, with some abuse of notation, we will write $\mathcal{W}_p(F, G)$ instead of $\mathcal{W}_p(P, Q)$ in the sequel. We remark, however, that we do not rely on strong approximations for the quantile process (as in Munk and Czado (1998) or del Barrio, Cuesta-Albertos (1999)). This kind of approach would require much stronger smoothness assumptions on F. Our technique, in contrast, is much closer to that in del Barrio and Loubes (2017) and (1.4) is only used to prove some sharp variance bounds.

Currently, the increasingly frequent use of machine learning techniques affects many aspects of our lives. This has yielded to a growing scientific attention to the framework of fair learning. We refer for instance to Romei and Ruggieri (2014), Pedreschi et al. (2012), Chouldechova (2017) or Friedler et al. (2018). In this setting, decisions are made by algorithmic procedures and the main concern is to detect whether decision rules, learnt from variables X, are biased with respect to a subcategory of the population. Formally, the problem consists in forecasting a binary variable $Y \in \{0, 1\}$ using observed covariates $X \in \mathbb{R}^d$, $d \geq 1$, and assuming that the population is divided into two categories that represent a bias, modeled by a protected variable $S \in \{0, 1\}$. A decision rule would be unfair for S when it favours individuals in the main protected group, usually $S = 1$, in the sense that the outcome of the algorithm is not just driven by the values of the covariates X but also by the values of S, leading to treating differently individuals from both groups while they have similar covariates. This discrimination may come from the algorithm or from a biased situation that would have been learnt from the training sample.

In the first situation, many criteria have been given in the recent literature on fair learning to detect whether an algorithm is committing discrimination (see Berk et al. (2017) or Besse et al. (2018) for a review). A majority of these definitions consider that the decision should be independent from the protected attribute S. In Berk et al. (2017), a classifier $g : \mathbb{R}^d \to \{0, 1\}$ is said to achieve Statistical Parity, with respect to (X, S), if

$$\mathbb{P}(g(X) = 1 \mid S = 0) = \mathbb{P}(g(X) = 1 \mid S = 1). \tag{1.5}$$

Therefore, if $\mathcal{L}$ denotes the distribution of a random variable, then (1.5) is reached by a classifier g when $\mathcal{L}(g(X) \mid S = 0) = \mathcal{L}(g(X) \mid S = 1)$ and $g(X)$ and S are independent.

Yet, in most real problems the independence described in (1.5) is difficult to achieve and, in addition, it refers to a given rule when in fact very different classifiers could be trained from the same learning sample. Furthermore, algorithms are usually inaccessible, in the sense that explaining how the classifier is chosen may be seen too intrusive by most companies or it may be simply not possible for many of them to change the way their models are built. To beat these shortchomings, another solution originally proposed in Feldman et al. (2015) and further developed in del Barrio et al. (2018), tries to look for a condition on the learning sample that ensures that every classifier trained from it is fair. This condition must guarantee that (1.5) holds for every classifier $g : \mathbb{R}^d \to \{0, 1\}$. If we denote $\mu_s := \mathcal{L}(X|S = s)$, $s \in \{0, 1\}$, then this means that μ_0 and μ_1 are equal. But certifying this equality is equivalent to the homogeneity testing problem and, as pointed out before, a goodness-of-fit test does not allow such certification. The most we can aspire to is providing statistical evidence that μ_0 and μ_1 are close. In this work we argue in favour of the Wasserstein metrics to measure the distances between the distributions.

As noted above, the CLT's provided in this paper enable to construct a new test to assess the degree of dissimilarity of different distributions, P and Q, using our procedure for testing (1.3). In the setup of fair learning, rejecting the null with this test we will be able to statistically certify that the distributions μ_0 and μ_1 are not too different. This will guarantee that the data set is fair, in the sense described above. Additionally, we provide a new way of assessing fairness in machine learning by considering confidence intervals for the degree of dissimilarity between these distributions (with respect to the Wasserstein distance). Finally, we also outline how our fairness assessment procedure can be tuned in order to use it with high-dimensional data.

References

Ajtai, M., Komlós, J. and Tusnády, G. (1984). On optimal matchings, *Combinatorica*, 4, 259–264

Berk, R., Heidari, H., Jabbari, S., Kearns, M. and Roth, A. (2017). Fairness in criminal justice risk assessments: the state of the art, *arXiv preprint arXiv:1703.09207*.

Besse, P., del Barrio, E., Gordaliza, P. and Loubes, J.-M. (2018). Confidence intervals for testing disparate impact in fair learning, *arXiv preprint arXiv:1807.06362*.

Bobkov, S. and Ledoux, M. (2014). One-dimensional empirical measures, order statistics and kantorovich transport distances, *preprint*.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.X. (2018). Scaling algorithms for unbalanced optimal transport problems, *Math. Comp.*

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in

recidivism prediction instruments, *Big data*, 5, 153–163.
Del Barrio, E., Cuesta-Albertos, J.A., Matrán, C. and Rodríguez-Rodríguez, J.M. (1999). Tests of goodness-of-fit based on the l_2-Wasserstein distance, *Ann. Statist.*, 1230-1239.
Del Barrio, E., Gamboa, F.,Gordaliza, P. and Loubes, J.-M. (2018). Obtaining fairness using optimal transport theory, *arXiv preprint arXiv:1806.03195*.
Del Barrio, E., Giné, E. and Matrán, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.*, 1009–1071.
Del Barrio, E., Giné, E., Utzet, F. et al. (2005). Asymptotics for l_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances, *Bernoulli*, 11, 131–189.
Del Barrio, E. and Loubes, J.-M. (2017). Central limit theorem for empirical transportation cost in general dimension. *arXiv preprint arXiv:1705.01299*.
Dobrić, V. and Yukich, J. E. (1995). Asymptotics for transportation cost in high dimensions, *Journal of Theoretical Probability*, 8, 97–118.
Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C. and Venkatasubramanian, S. (2015). Certifying and removing disparate impact, *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.
Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Rel.*, 162, 707–738.
Freitag, G., Czado, C. and Munk, A. (2007). A nonparametric test for similarity of marginals with applications to the assessment of population bioequivalence. *J. Stat. Plan. Infer.*, 137, 697–711.
Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P. and Roth, D. (2018). A comparative study of fairness-enhancing interventions in machine learning. *ArXiv e-prints*.
Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60, 223–241.
Pedreschi, D., Ruggieri, S. and Turini, F. (2017). A study of top-k measures for discrimination discovery, *In Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 126–131.
Romei, A. and Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis, *Knowl. Eng. Rev.*, 29, 582–638.
Sommerfeld, M. and Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 80, 219–238.
Talagrand, M. and Yukich, J.E. (1993). The integrability of the square exponential transportation cost, *Ann. App. Probab.*, 1100–1111.
Tameling, C., Sommerfeld, M. and Munk, A. (2017). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications, *arXiv preprint arXiv:1707.00973*.
Villani, C. (2003). Topics in Optimal Transportation, *Graduate studies in mathematics. American Mathematical Soc.*