Convergence of an estimator of the Wasserstein distance between two continuous probability distributions

Thierry Klein, Jean-Claude Fort, Philippe Berthet

Convergence of an estimator of the Wasserstein distance between two continuous probability distributions

Philippe Berthet [a], Jean-Claude Fort [b], Thierry Klein [a,c]

[a] *IMT- institut mathématiques de Toulouse, Université de Toulouse, France.*
[b] *MAP5, Université Paris Descartes, SPC, 45 rue des Saints Pères, 75006 Paris, France*
[c] *ENAC - Ecole Nationale de l'Aviation Civile , Université de Toulouse, France*

Abstract

This article is dedicated to the estimation of Wasserstein distances and Wasserstein costs between two distinct continuous distributions F and G on $\mathbb{R}$. The estimator is based on the order statistics of (possibly dependent) samples of F resp. G. We prove the consistency and the asymptotic normality of our estimators.

Nous étudions la convergence d'estimateurs des distances et des coûts de Wasserstein entre deux lois de probabilités continues F et G sur $\mathbb{R}$ distinctes. L'estimateur est construit à partir des statistiques d'ordres des échantillons (possiblement dépendants) de F et de G. Nous montrons leur consistance et un théorème de la limite centrale.

1. Introduction

The motivation of this work is to be found in the fast development of computer experiments. Nowadays the output of many computer codes is not only a real multidimensional variable but frequently a function computed on numerous sample points. In particular this function may be the density or the cumulative distribution function ($c.d.f$) of a real random variable ($r.v.$). To analyze such outputs one needs to choose a distance to compare various $c.d.f.$. Among the large possibilities offered by the literature the Wasserstein distances are now commonly used - for more details on general Wasserstein distances we refer to [10]. As many of the computer codes provide large samples of the underlying distributions, the statistical study of such distances is of primordial importance. For one dimensional probability distributions the p-Wasserstein distance is the L^p distance between simulated $r.v.$ from a universal simulator U uniform on $[0,1]$: $W_p^p(F,G) = \int_0^1 |F^-(u) - G^-(u)|^p du = \mathbb{E}|F^-(U) - G^-(U)|^p$, where F^- is the generalized inverse of F. The most relevant cases for applications are $p = 2$ and $p = 1$. It is then natural to estimate $W_p^p(F,G)$

Email addresses: philippe.berthet@math.univ-toulouse.fr (Philippe Berthet), fort43@gmail.com (Jean-Claude Fort), thierry.klein@math.univ-toulouse.fr or thierry01.klein@enac.fr (Thierry Klein).

by its empirical counterpart that is $W_p^p(\mathbb{F}_n, \mathbb{G}_n)$ where $\mathbb{F}_n$ and $\mathbb{G}_n$ are the empirical distribution functions of F and G built through *i.i.d.* samples of F and G. The two samples are not necessarily independent, for instance they may be issued from simultaneous experimentations.

Many authors were interested in the convergence of $W_p^p(\mathbb{F}_n, F)$, see the survey paper [3] and [8,6,7,1]. Up to our knowledge there are few works [9] trying to quantify the convergence of $W_p^p(\mathbb{F}_n, \mathbb{G}_n)$ or for more general Wasserstein costs.

This note is organized as follows. Section 2 is dedicated to the definition of the general Wasserstein costs we are considering and their estimators. In Section 3 we state the asymptotic properties of our estimators. In Section **??** we give a very brief sketch of the proof.

2. Wasserstein distances and costs for probability distributions on $\mathbb{R}$

2.1. *Wasserstein costs*

Let F and G two *c.d.f.* on $\mathbb{R}$. The p-Wasserstein distance between F and G is defined to be

$$W_p^p(F, G) = \min_{X \sim F, Y \sim G} \mathbb{E}|X - Y|^p, \tag{1}$$

where $X \sim F, Y \sim G$ means that X and Y are *r.v.*s with respective distribution F and G. The minimum in (3) has the following explicit expression,

$$W_p^p(F, G) = \int_0^1 |F^-(u) - G^-(u)|^p du. \tag{2}$$

The Wassertein distances can be generalized to Wasserstein costs. Given a real non negative function $c(x, y)$ of two real variables we consider

$$W_c(F, G) = \min_{X \sim F, Y \sim G} \mathbb{E}\ c(X, Y). \tag{3}$$

Among the costs c, those that define a negative measure on $\mathbb{R}^2$ are of special interest. They satisfy the "measure property" $\mathcal{P}$ defined by

$$\mathcal{P} : c(x', y') - c(x', y) - c(x, y') + c(x, y) \leq 0,\ x \leq x'\ , y \leq y'.$$

From Theorem 2 of [4], we optain an explicit formula of W_c for cost functions satisfying property $\mathcal{P}$, extending the formula (2). Namely if U is a *r.v.* uniformly distributed on $[0, 1]$ then

$$W_c(F, G) = \int_0^1 c(F^-(u), G^-(u))du = \mathbb{E}\ c(F^-(U), G^-(U)). \tag{4}$$

Remark 1 It is obvious that $c(x, y) = -xy$ satisfies the $\mathcal{P}$ property and if c satisfies $\mathcal{P}$ then any function of the form $a(x) + b(y) + c(x, y)$ also satisfies $\mathcal{P}$. In particular $(x - y)^2 = x^2 + y^2 - 2xy$ satisfies $\mathcal{P}$. More generally if ρ is a convex real function then $c(x, y) = \rho(x - y)$ satisfies $\mathcal{P}$. This is the case for $|x - y|^p$, $p \geqslant 1$.

2.2. *Empirical Wasserstein costs*

Let us assume that a *i.i.d.* sample $(X_i, Y_i)_{1 \leq i \leq n}$ of *r.v.* with marginal *c.d.f.* F and G is available. We write $\mathbb{F}_n$ and $\mathbb{G}_n$ the empirical *c.d.f* built on the two marginal samples and denote by $X_{(i)}$ the i^{th} order

statistic of the sample $(X_i)_{1\leq i\leq n}$, $X_{(1)} < \ldots < X_{(n)}$.
Let us study the following natural estimator of $W_c(F,G)$

$$W_c^n := W_c(\mathbb{F}_n, \mathbb{G}_n) = \frac{1}{n}\sum_{i=1}^{n} c(X_{(i)}, Y_{(i)}). \tag{5}$$

W_c^n is a consistent estimator of $W_c(F,G)$
Theorem 2.1 *Assume that the cost c(x,y) satisfies:* $c(x,y) \leq a(f(x)+f(y))$ *with* $f \geqslant 0$ *and* $\mathbb{E}[f(X)] < \infty, \mathbb{E}[f(Y)] < \infty$, *then* W_c^n *converges almost surely to* $W_c(F,G)$.

3. A Central Limit Theorem for W_c^n

The main result of this note is the weak convergence of

$$\sqrt{n}\left(W_c(\mathbb{F}_n, \mathbb{G}_n) - W_c(F,G)\right). \tag{6}$$

we need to fix the relative position of the two tails of F and G. Moreover these tails induce restrictions on the allowed costs. We assume that F and G are supported on whole $\mathbb{R}$ but we only deal with the right hand side of the real line in order to give a more synthetic presentation. Hence we restrict properties of the key functions to $\mathbb{R}_+$ and $(m,+\infty)$ for some $m > 0$ large enough.

3.1. *Assumptions on F and G.*

For $k \in \mathbb{N}_*$ denote $\mathcal{C}_k$ the set of functions that are k times continuously differentiable on $\mathbb{R}$, and $\mathcal{C}_0$ the set of continuous functions. We assume that F and G are $\mathcal{C}_2$ and that the densities $f = F'$ and $g = G'$ are positive and

$$(\mathbf{FG1}) \sup_{x>m} \frac{1-F(x)}{f(x)}\left(\frac{1}{x} + \frac{|f'(x)|}{f(x)}\right) < \infty \text{ and } \sup_{x>m} \frac{1-G(x)}{g(x)}\left(\frac{1}{x} + \frac{|g'(x)|}{g(x)}\right) < \infty \tag{7}$$

Moreover, the tails have to be strictly separated. Let us denote F^{-1} and G^{-1} the quantile functions. We assume that there exists $\tau_0 > 0$ such that for $u \geqslant F(m)$,

$$(\mathbf{FG2}) \quad F^{-1}(u) - G^{-1}(u) \geqslant \tau_0.$$

3.2. *Assumptions on the cost c.*

Among Wasserstein costs satisfying property $\mathcal{P}$ we consider the regular ones.
Let $\mathcal{M}_2(m)$ be the subset of functions $\varphi \in \mathcal{C}_2$ such that φ'' is monotone on $(m,+)$. For m large enough φ, φ' are also monotone on $(m,+\infty)$. Let $\mathcal{M}_0(m)$ be the set of continuous functions monotone on (m). Write $RV(\gamma)$ the set of regularly varying functions at $+\infty$ with index $\gamma \geqslant 0$. We introduce$RV_2^+(\gamma,m) = RV(\gamma) \cap \mathcal{M}_2(m), \quad \gamma > 0$.
When $L \in RV(0)$ we assume that

$$\frac{l_1}{x} \leqslant L'(x) = \frac{\varepsilon_1(x)L(x)}{x}, \quad \lim_{x\to+\infty} \varepsilon_1(x) = 0, \quad l_1 \geqslant 1. \tag{8}$$

Whence we define $RV_2^+(0,m) = \{L : L \in RV(0) \cap \mathcal{M}_2(m,+\infty)$ such that (8) holds$\}$.

We impose $(wlog)$ that $c(x,x)=0$ and assume that

$$(\mathbf{C1}) \quad c(x,y) \geqslant 0, \quad c \in \mathcal{C}_2$$

$$(\mathbf{C2}) \; c(x,y) = \exp(l(|x-y|)), \quad \text{for } (x,y) \in (m,+\infty)^2 \quad \text{and} \quad l \in RV_2^+(\gamma,0), \gamma \geqslant 0.$$

Thus c is asymptotically regular and symetric. Finally we need the following contraction of $c\,(x,y)$ along the diagonal $x=y$. We assume that there exists $d(m,\tau) \to 0$ as $\tau \to 0$ such that

$$(\mathbf{C3}) \quad |c\,(x',y') - c\,(x,y)| \leqslant d(m,\tau)\,(|x'-x| + |y'-y|) \quad \text{for } (x,y)\,,(x',y') \in D_m(\tau),$$

where $D_m(\tau) = \{(x,y) : \max(|x|\,,|y|) \leqslant m, |x-y| \leqslant \tau\}$.

3.3. *Cross assumptions between c, F and G.*

Consider the tail functions $\psi_X(x) = -\log \mathbb{P}(X > x)$ and $\psi_Y(x) = -\log \mathbb{P}(Y > x)$.
We require that for some $\theta > 2$ it holds, for $x \in (l(m), +\infty)$,

$$(\mathbf{CFG}) \quad (\psi_X \circ l^{-1})'(x) \geqslant 2 + \frac{2\theta}{x}.$$

We point out that (**CFG**) implies $\psi_X(x) \geqslant 2l(x) + 2\theta \log l(x)$ which in turn guaranties that the heaviest tail satisfies

$$\int_m^{+\infty} \sqrt{\mathbb{P}\left(\exp(l(X)) > x\right)} dx < +\infty.$$

The later condition is close to necessity for the finiteness of the limiting variance in Theorem 3.1 below. This is the same kind of condition (3.4) in [3] that ensures the convergence of $W_1(\mathbb{F}_n, F)$ at rate $\sqrt{n}$.

3.4. *The main theorem*

We say that conditions (**C**), (**FG**) and (**CFG**) hold if they are satisfied by the right and the left tails of the *c.d.f* F and G(possibly exchanging F and G). We denote $h_X = f \circ F^{-1}$, $h_Y = g \circ G^{-1}$ the so-called density quantile functions. Now define

$$\Pi(u,v) = \mathbb{P}\left(X \leqslant F^{-1}(u), Y \leqslant G^{-1}(v)\right),$$

then the covariance matrix

$$\Sigma(u,v) = \begin{pmatrix} \dfrac{\min(u,v) - uv}{h_X(u)h_X(v)} & \dfrac{\Pi(u,v) - uv}{h_X(u)h_Y(v)} \\ \dfrac{\Pi(v,u) - uv}{h_X(v)h_Y(u)} & \dfrac{\min(u,v) - uv}{h_Y(v)h_Y(u)} \end{pmatrix}, \tag{9}$$

and the gradient

$$\nabla(u) = \left(\frac{\partial}{\partial x} c(F^{-1}(u), G^{-1}(u)), \frac{\partial}{\partial y} c(F^{-1}(u), G^{-1}(u))\right). \tag{10}$$

Theorem 3.1 *If* (**C**), (**FG**) *and* (**CFG**) *hold then*

$$\sqrt{n}\,(W_c(\mathbb{F}_n, \mathbb{G}_n) - W_c(F,G)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(\Pi, c))$$

with

$$\sigma^2(\Pi, c) = \int_0^1 \int_0^1 \nabla(u)\Sigma(u,v)\nabla(v) du dv < +\infty. \tag{11}$$

3.4.1. *Sketch of proof*

Let $c_n(u) = \sqrt{n}\left(c(\mathbb{F}_n^{-1}(u), \mathbb{G}_n^{-1}(u)) - c(F^{-1}(u), G^{-1}(u))\right)$. We write

$$\begin{aligned}\int_{\frac{1}{2}}^{1} c_n(u)du &= \int_{\frac{1}{2}}^{u_0} c_n(u)du + \int_{u_0}^{1-\xi_n} c_n(u)du + \int_{1-\xi_n}^{1-\varepsilon_n} c_n(u)du + \int_{1-\varepsilon_n}^{1} c_n(u)du \\ &:= I_1(n) + I_2(n) + I_3(n) + I_4(n).\end{aligned}$$

Under (**CFG**) and (**C2**) we can find $\varepsilon_n \to 0$ such that $I_4(n) \to 0$ in probability.

The sequence $\sqrt{n}\varepsilon_n \to 0$ is an explicit function of n involving the functions $l(x)$, $h_X(x)$ and $h_Y(x)$. Then our condition (**FG1**) implies the minimal conditions we indeed need on $h_X(x)$, $h_Y(x)$ and their companion functions $(1-u)/F_X(x)h_X(x)$ and $(1-u)/F_Y(x)h_Y(x)$ to apply the strong gaussian approximation of [5]. When combined with (**CFG**) and (**C2**), we are able to show that for any $\xi_n \to 0$ such that $\sqrt{n}\xi_n \to +\infty$ we have $I_3(n) \to 0$ *a.s.*.

If ξ_n is slow enough we establish a distribution free a joint Brownian strong approximation of the quantile processes $(\mathbb{F}_n^{-1}(u), \mathbb{G}_n^{-1}(u))$ that we derive from [2].

Then, thanks to the regularity (**C3**) of $l(x)$ and the stochastic ordering (**FG2**), when $u_0 \to 1$ the term $I_2(n)$ is small in probability as $n \to \infty$.

Finally, the latter strong approximation allows under (**C1**) to show that the main term $I_1(n)$

is tight and has a limit law that is arbitrarily close to the desired Gaussian limit law when $u_0 \to 1$. Working with the strongly approximated versions entails the weak convergence claimed at Theorem 3.1. Moreover (**CFG**) implies that the variance of the limiting Gaussian process is integrable and $\sigma^2(\Pi, c) < +\infty$.

4. Conclusion

In this work we established consistency and asymptotic normality of the natural estimators of a large class of Wasserstein costs between two smooth distributions F and G having separated tails.

The case W_1 is not included in the previous theorem since it does not satisfy (**C3**), but the theorem still holds thanks to a specific proof. Theorem 3.1 applies to all the classical distributions with regularly decreasing tail when choosing an adapted cost. For instance we may take two distributions with polynomial (Pareto) tail of same order $\beta > 2$ but shifted and apply our result when choosing a Wasserstein distance W_p^p with $p < \beta/2$. For Gaussian tailed distributions one may consider exponential costs of type $e^{|x-y|^\gamma} - 1, \gamma < 2$.

Clearly this result allows to build a test for equality of two smooth distributions.

References

[1] P. C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. Uniqueness and approximate computation of optimal incomplete transportation plans. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(2):358–375, 2011.

[2] P. Berthet and D. M. Mason. Revisiting two strong approximation results of Dudley and Philipp. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 155–172. Inst. Math. Statist., Beachwood, OH, 2006.

[3] M. Bobkov, S. G.and Ledoux. One-dimensional empirical measures, order statistics and kantorovich transport distances. *To appear in: Memoirs of the AMS*, Preprint 2016.

[4] Stamatis Cambanis, Gordon Simons, and William Stout. Inequalities for $Ek(X, Y)$ when the marginals are fixed. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 36(4):285–294, 1976.

[5] M. Csörgö and P. Révész. Strong approximations of the quantile process. *Ann. Statist.*, 6(4):882–894, 1978.

[6] E. del Barrio, E. Giné, and C. Matrán. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.*, 27(2):1009–1071, 1999.

[7] E. del Barrio, E. Giné, and C. Matrán. Correction: "Central limit theorems for the Wasserstein distance between the empirical and the true distributions" [Ann. Probab. **27** (1999), no. 2, 1009–1071; MR1698999 (2000g:60034)]. *Ann. Probab.*, 31(2):1142–1143, 2003.

[8] E. del Barrio, E. Giné, and F. Utzet. Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11(1):131–189, 2005.

[9] M. Sommerfeld and A. Munk. Inference for Empirical Wasserstein Distances on Finite Spaces. *ArXiv e-prints*, October 2016.

[10] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.