Stone's theorem for distributional regression in Wasserstein distance

Clément Dombry, Thibault Modeste, Romain Pic

Stone's theorem for distributional regression in Wasserstein distance

Clément Dombry[1], Thibault Modeste[2], and Romain Pic[1]

[1]Université Franche-Comté, CNRS UMR 6623, Laboratoire de Mathématiques de Besançon, F-25000 Besançon, France
[2]Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, F-69622 Villeurbanne, France

Abstract

We extend the celebrated Stone's theorem to the framework of distributional regression. More precisely, we prove that weighted empirical distribution with local probability weights satisfying the conditions of Stone's theorem provide universally consistent estimates of the conditional distributions, where the error is measured by the Wasserstein distance of order $p \geq 1$. Furthermore, for $p = 1$, we determine the minimax rates of convergence on specific classes of distributions. We finally provide some applications of these results, including the estimation of conditional tail expectation or probability weighted moment.

Keywords: distributional regression, Wasserstein distance, non parametric regression, minimax rate of convergence.
MSC 2020 subject classification: 62G05.

Contents

1 Introduction

Forecast is a major task from statistics and often of crucial importance for decision making. In the simple case when the quantity of interest is univariate and quantitative, point forecast often takes the form of regression where one aims at estimating the conditional mean (or the conditional quantile) of the response variable Y given the available information encoded in a vector of covariates X. A point forecast is only a rough summary statistic and should at least be accompanied with an assessment of uncertainty (e.g. standard deviation or confidence interval). Alternatively, probabilistic forecasting and distributional regression (Gneiting and Katzfuss, 2014) suggest to estimate the full conditional distribution of Y given X, called the predictive distribution.

In the last decades, weather forecast has been a major motivation for the development of probabilistic forecast. Ensemble forecasts are based on a given number of deterministic models whose parameters vary slightly in order to take into account observation errors and incomplete physical representation of the atmosphere. This leads to an ensemble of different forecasts that overall also assess the uncertainty of the forecast. Ensemble forecasts suffer from bias and underdispersion (Hamill and Colucci, 1997) and need to be statistically postprocessed in order to be improved. Different postprocessing methods have been proposed, such as Ensemble Model Output Statistics (Gneiting et al., 2005), Quantile Regression Forests (Taillardat et al., 2019) or Neural Networks (Schulz and Lerch, 2021) among others. Distributional regression is now widely used beyond meteorology and recent methodological works include deep distribution regression by Li et al. (2021), distributional

random forest by Ćevid et al. (2022) or isotonic distributional regression by Henzi et al. (2021).

The purpose of the present paper is to provide an extension to the framework of distributional regression of the celebrated Stone's theorem (Stone, 1977) that states the consistency of local weight algorithm for the estimation of the regression function. The strength of Stone's theorem is that it is fully non-parametric and model-free, with very mild assumptions that covers many important cases such as kernel algorithms and nearest neighbor methods, see e.g. Györfi et al. (2002) for more details. We prove that Stone's theorem has a natural and elegant extension to distributional regression with error measured by the Wasserstein distance of order $p \geq 1$. Our result covers not only the case of a one-dimensional output $Y \in \mathbb{R}$ where the Wasserstein distance has a simple explicit form, but also the case of a multivariate output $Y \in \mathbb{R}^d$. The use of the Wasserstein distance is motivated by recent works revealing that it is a useful and powerful tool in statistics, see e.g. the review by Panaretos and Zemel (2020). Besides this main result, we characterize, in the case $d = 1$ and $p = 1$, the optimal minimax rate of convergence on suitable classes of distributions. We also discuss implications of our results to estimate various statistics of possible interest such as the expected shortfall or the probability weighted moment.

The structure of the paper is the following. In Section 2, we present the required background on Stone's theorem and Wasserstein spaces. Section 3 gathers our main results, including the extension of Stone's theorem to distributional regression (Theorem 2), the characterization of optimal minimax rates of convergence (Theorem 3) and some applications (Proposition 2 and the subsequent examples). All the technical proofs are postponed to Section 4.

2 Background

2.1 Stone's theorem

In a regression framework, we observe a sample (X_i, Y_i), $1 \leq i \leq n$, of independent copies of $(X, Y) \in \mathbb{R}^k \times \mathbb{R}^d$ with distribution P. Based on this sample and assuming Y integrable, the goal is to estimate the regression function

$$r(x) = \mathbb{E}[Y|X = x], \quad x \in \mathbb{R}^k.$$

Local average estimators take the form

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i \tag{1}$$

with $W_{n1}(x), \ldots, W_{nn}(x)$ the *local weights* at x. The local weights are assumed to be measurable functions of x and $X_1, \ldots, X_n$ but not to depend

on $Y_1, \dots, Y_n$, that is

$$W_{ni}(x) = W_{ni}(x; X_1, \dots, X_n), \quad 1 \leq i \leq n. \tag{2}$$

For the convenience of notation, the dependency on $X_1, \dots, X_n$ is implicit. In this paper, we focus only on the case of *probability weights* satisfying

$$W_{ni}(x) \geq 0, \ 1 \leq i \leq n, \quad \text{and} \quad \sum_{i=1}^{n} W_{ni}(x) = 1. \tag{3}$$

Stone's Theorem states the universal consistency of the regression estimate in L^p-norm.

Theorem 1 (Stone (1977))**.** *Assume the probability weights* (3) *satisfy the following three conditions:*

i) there is $C > 0$ *such that* $\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) g(X_i)\right] \leq C\mathbb{E}[g(X)]$ *for all* $n \geq 1$ *and measurable* $g : \mathbb{R}^k \to [0, +\infty)$ *such that* $\mathbb{E}[g(X)] < \infty$*;*

ii) for all $\varepsilon > 0$, $\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{\{\|X_i - X\| > \varepsilon\}} \to 0$ *in probability as* $n \to +\infty$*;*

iii) $\max_{1 \leq i \leq n} W_{ni}(X) \to 0$ *in probability as* $n \to +\infty$*.*

Then, for all $p \geq 1$ *and* $(X, Y) \sim P$ *such that* $\mathbb{E}[\|Y\|^p] < \infty$*,*

$$\mathbb{E}\left[\|\hat{r}_n(X) - r(X)\|^p\right] \longrightarrow 0 \quad \text{as } n \to +\infty. \tag{4}$$

Conversely, if Equation (4) *holds, then the probability weights must satisfy conditions* $i) - iii)$*.*

Remark 1. Stone's theorem is usually stated in dimension $d = 1$. Since the convergence of random vectors $\hat{r}_n(X) \to r(X)$ in L^p is equivalent to convergence in L^p of all the components, the extension to the dimension $d \geq 2$ is straightforward. Furthermore, more general weights than probability weights can be considered: condition (3) can be dropped and replaced by the weaker assumptions that

$$|W_{ni}(X)| \leq M \quad \text{a.s. for some } M > 0.$$

and

$$\sum_{i=1}^{n} W_{ni}(X) \to 1 \text{ in probability.}$$

Such general weights will not be considered in the present paper and we therefore stick to probability weights. The reader can refer to Biau and Devroye (2015) for a complete proof of Stone's theorem together with a discussion.

Example 1. The following two examples of kernel weights and nearest neighbor weights are the most important ones in the literature and we refer to Györfi et al. (2002) Chapter 5 and 6 respectively for more details.

- The kernel weights are defined by

$$W_{ni}(x) = \frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h_n}\right)}, \quad 1 \leq i \leq n \tag{5}$$

  if the denominator is nonzero, and $1/n$ otherwise. Here the bandwidth $h_n > 0$ depends only on the sample size n and the function $K : \mathbb{R}^k \to [0, +\infty)$ is called a kernel. In this case, the estimator (1) corresponds to the Nadaraya-Watson estimator of the regression function (Nadaraya, 1964; Watson, 1964). We say that K is a boxed kernel if there are constants $R_2 \geq R_1 > 0$ and $M_2 \geq M_1 > 0$ such that

$$M_1 \mathbb{1}_{\{\|x\| \leq R_1\}} \leq K(x) \leq M_2 \mathbb{1}_{\{\|x\| \leq R_2\}}, \quad x \in \mathbb{R}^k.$$

  Theorem 5.1 in Györfi et al. (2002) states that, for a boxed kernel, the kernel weights (5) satisfy conditions $i) - iii)$ of Theorem 1 if and only if $h_n \to 0$ and $nh_n^k \to +\infty$ as $n \to +\infty$.

- The nearest neighbor (NN) weights are defined by

$$W_{ni}(x) = \begin{cases} \frac{1}{\kappa_n} & \text{if } X_i \text{ belongs to the } \kappa_n\text{-NN of } x \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

  where the number of neighbors $\kappa_n \in \{1, \ldots, n\}$ depends only on the sample size. Recall that the κ_n-NN of x within the sample $(X_i)_{1 \leq i \leq n}$ are obtained by sorting the distances $\|X_i - x\|$ in increasing order and keeping the κ_n points with the smallest distances – as discussed in Györfi et al. (2002) Chapter 6, several rules can be used to break ties such as lexicographic or random tie breaking. Theorem 6.1 in the same reference states that the nearest neighbor weights (6) satisfy conditions $i) - iii)$ of Theorem 1 if and only if $\kappa_n \to +\infty$ and $\kappa_n/n \to 0$ as $n \to +\infty$.

Example 2. Interestingly, some variants of the celebrated Breiman's Random Forest (Breiman, 2001) produce probability weights satisfying the assumptions of Stone's theorem. In Breiman's Random Forest, the splits involve both the covariates and the response variable so that the associated weighs $W_{ni}(x) = W_{ni}(x; (X_l, Y_l)_{1 \leq l \leq n})$ are not in the form (2). Scornet (2016) considers two simplified version of infinite random forest where the associated weights $W_{ni}(x)$ do not depend on the response values and satisfy the so call X-property, that is they are in the form (2). For totally

non adaptive forests, the trees are grown thanks to a binary splitting rule that does not use the training sample and is totally random; the author shows that the probability weights associated to the infinite forest satisfy the assumptions of Stone's theorem under the condition that the number of leaves grows to infinity at a rate smaller than n and the leaf volume tends to zero in probability (see Theorem 4.1 and its proof). For q-quantile forest, the binary splitting rules involves only the covariates and the author shows that the weights associated to the infinite forest satisfy the assumptions of Stone's theorem provided the subsampling number a_n satifies $a_n \to +\infty$ and $a_n/n \to 0$ (see Theorem 5.1 and its proof).

2.2 Wasserstein spaces

We recall the definition and some elementary facts on Wasserstein spaces on $\mathbb{R}^d$. More details and further results on optimal transport and Wasserstein spaces can be found in the monograph by Villani (2009), Chapter 6.

For $p \geq 1$, the Wasserstein space $\mathcal{W}_p(\mathbb{R}^d)$ is defined as the set Borel probability measures on $\mathbb{R}^d$ having a finite moment of order p, i.e. such that

$$M_p(\mu) = \Big(\int_{\mathbb{R}^d} \|y\|^p \, \mu(\mathrm{d}y) \Big)^{1/p} < \infty. \tag{7}$$

It is endowed with the distance defined, for $Q_1, Q_2 \in \mathcal{W}_p(\mathbb{R}^d)$, by

$$\mathcal{W}_p(Q_1, Q_2) = \inf_{\pi \in \Pi(Q_1, Q_2)} \left(\int \|y_1 - y_2\|^p \, \pi(\mathrm{d}y_1 \mathrm{d}y_2) \right)^{1/p}, \tag{8}$$

where $\Pi(Q_1, Q_2)$ denotes the set of measures on $\mathbb{R}^d \times \mathbb{R}^d$ with margins Q_1 and Q_2. A couple (Z_1, Z_2) of random variables with distributions Q_1 and Q_2 respectively is called a *coupling*. The Wasserstein distance is thus the minimal distance $\|Z_1 - Z_2\|_{\mathrm{L}^p} = \mathbb{E}[\|Z_1 - Z_2\|^p]^{1/p}$ over all possible couplings. Existence of optimal couplings is ensured since $\mathbb{R}^d$ is a complete and separable metric space so that the infimum is indeed a minimum.

Wasserstein distances are generally difficult to compute, but the case $d = 1$ is the exception. A simple optimal coupling is provided by the probability inverse transform: for $i = 1, 2$, let $Q_i \in \mathcal{W}_p(\mathbb{R})$, F_i denotes its cumulative distribution function and F_i^{-1} its generalized inverse (quantile function). Then, starting from an uniform random variable $U \sim \mathrm{Unif}(0, 1)$, an optimal coupling is given by $(Z_1, Z_2) = (F_1^{-1}(U), F_2^{-1}(U))$. Therefore, the Wasserstein distance is explicitly given by

$$\mathcal{W}_p(Q_1, Q_2) = \left(\int_0^1 |F_1^{-1}(u) - F_2^{-1}(u)|^p \mathrm{d}u \right)^{1/p}. \tag{9}$$

When $p = 1$, a simple change of variable yields

$$\mathcal{W}_1(Q_1, Q_2) = \int_{-\infty}^{+\infty} |F_1(u) - F_2(u)| \mathrm{d}u. \tag{10}$$

3 Main results

3.1 Stone's theorem for distributional regression

We now present the main result of the paper which is a natural extension of Stone's theorem to the framework of distributional regression. Given a distribution $(X, Y) \sim P$ on $\mathbb{R}^k \times \mathbb{R}^d$, we denote by F the marginal distribution of Y and by F_x its conditional distribution given $X = x$. This conditional distribution can be estimated on a sample $(X_i, Y_i)_{1 \leq i \leq n}$ of independent copies of (X, Y) by the weighted empirical distribution

$$\hat{F}_{n,x} = \sum_{i=1}^{n} W_{ni}(x) \delta_{Y_i} \tag{11}$$

where δ_y denotes the Dirac mass at point $y \in \mathbb{R}^d$. For probability weights satisfying (3), $\hat{F}_{n,x}$ is a probability measure and can be viewed as a random element in the complete and separable space $\mathcal{W}_p(\mathbb{R}^d)$. We recall that the weights $W_{ni}(x) = W_{ni}(x; X_1, \ldots, X_n)$ implicitly depend on $X_1, \ldots, X_n$ but not on $Y_1, \ldots, Y_n$.

Theorem 2. *Assume the probability weights satisfy conditions $i) - iii)$ from Theorem 1. Then, for all $p \geq 1$ and (X, Y) such that $\mathbb{E}[\|Y\|^p] < \infty$,*

$$\mathbb{E}\big[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)\big] \longrightarrow 0 \quad \text{as } n \to +\infty. \tag{12}$$

Conversely, if Equation (12) holds, then the probability weights must satisfy conditions $i) - iii)$.

It is worth noticing that

$$\mathbb{E}\left[\|\hat{r}_n(X) - r(X)\|^p\right] \leq \mathbb{E}\big[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)\big]$$

so that Theorem 2 implies Theorem 1 in a straightforward way. The proof of Theorem 2 is postponed to Section 4. It first considers the case $d = 1$ where the Wasserstein distance is explicitly given by formula (9). Then, the results is extended to higher dimension $d \geq 2$ thanks to the notion of max-sliced Wasserstein distance (Bayraktar and Guo, 2021) which allows to reduce the convergence of measures on $\mathbb{R}^d$ to the convergence of their uni-dimensional projections (a precise statement is given in Theorem 4 below).

3.2 Rates of convergence

We next consider rates of convergence in the minimax sense. Note that similar questions and results have been established in Pic et al. (2022), where the second order Cramér's distance was considered, i.e.

$$\|\hat{F}_{n,X} - F_X\|_{L_2}^2 = \int_{\mathbb{R}} |\hat{F}_{n,X}(y) - F_X(y)|^2 \, \mathrm{d}y.$$

We focus here on the Wasserstein distance $\mathcal{W}_p(\hat{F}_{n,X}, F_X)$ and consider only the case $d = 1$ and $p = 1$ which allows the explicit expression (10). The other cases seem harder to analyze and are beyond the scope of the present paper. Our first result considers the error in Wasserstein distance when $X = x$ is fixed.

Proposition 1. *Assume $d = 1$ and $(X, Y) \sim P$ such that $\mathbb{E}[|Y|] < \infty$. Then,*

$$\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,x}, F_x)\big] \leq \mathbb{E}\Big[\sum_{i=1}^n W_{ni}(x)\mathcal{W}_1(F_{X_i}, F_x)\Big] + M(x)\mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(x)\Big]^{1/2},$$

where $M(x) = \int_{\mathbb{R}} \sqrt{F_x(z)(1 - F_x(z))}\mathrm{d}z$.

The first term corresponds to an approximation error due to the fact that we use a biased sample to estimate F_x. The more regular the model is, the smaller the approximation error is. The second term is an estimation error due to the fact that we use an empirical mean to estimate F_x. This estimator error is smaller if the distribution error has a lower dispersion (as measured by $M(x)$) or if $\sum_{i=1}^n W_{ni}^2(x)$ is small. Note that in the case of nearest neighbor weights, $1/\sum_{i=1}^n W_{ni}^2(x)$ is exactly equal to κ so that this quantity is often referred to as the *effective sample size* and the estimation error is proportional to the square root of the expected reciprocal effective sample size.

In view of Proposition 1, we introduce the following classes of functions.

Definition 1. *Let $\mathcal{D}(H, L, M)$ be the class of distributions $(X, Y) \sim P$ on $\mathbb{R}^k \times \mathbb{R}$ satisfying:*

a) $X \in [0, 1]^k$ a.s. and $\mathbb{E}|Y| < \infty$,

b) for all $x, x' \in [0, 1]^k$, $\mathcal{W}_1(F_x, F_{x'}) \leq L\|x - x'\|^H$,

c) for all $x \in [0, 1]^k$, $\int_{\mathbb{R}} \sqrt{F_x(z)(1 - F_x(z))}\,\mathrm{d}z \leq M$.

The definition of the class together with Proposition 1 entails that the expected error is uniformly bounded on the class $\mathcal{D}(H, L, M)$ by

$$\begin{aligned}&\mathbb{E}\Big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\Big]\\ &\leq L\mathbb{E}\Big[\sum_{i=1}^n W_{ni}(X)\|X_i - X\|^H\Big] + M\mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(X)\Big]^{1/2}.\end{aligned} \tag{13}$$

As a consequence, Proposition 1 allows to derive explicit bounds uniformly on $\mathcal{D}(H, L, M)$ for the kernel and nearest neighbor methods from Example 1. For the sake of simplicity, we consider the uniform kernel only.

Corollary 1. *Let $\hat{F}_{n,X}$ be given by the kernel method with uniform kernel $K(x) = \mathbb{1}_{\{\|x\|\leq 1\}}$ and weights given by Equation (5). If $P \in \mathcal{D}(H, L, M)$, then*

$$\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\big] \leq Lh_n^H + M\sqrt{(2+1/n)c_k}(nh_n^k)^{-1/2} + Lk^{H/2}c_k(nh_n^k)^{-1}$$

with $c_k = k^{k/2}$.

Corollary 2. *Let $\hat{F}_{n,X}$ be given by the nearest neighbor method with weights given by Equation (6) and assume $P \in \mathcal{D}(H, L, M)$. Then,*

$$\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\big] \leq \begin{cases} L8^{H/2}(\kappa_n/n)^{H/2} + M\kappa_n^{-1/2} & \text{if } k = 1, \\ L\tilde{c}_k^{H/2}(\kappa_n/n)^{H/k} + M\kappa_n^{-1/2} & \text{if } k \geq 2, \end{cases}$$

where $\tilde{c}_k$ depends only on the dimension k and is defined in Biau and Devroye (2015, Theorem 2.4).

One can see that consistency holds — i.e. the expected error tends to 0 as $n \to +\infty$ — as soon as $h_n \to 0$ and $nh_n^k \to +\infty$ for the kernel method and $\kappa_n/n \to 0$ and $\kappa_n \to +\infty$ for the nearest neighbor method.

The next theorem provides the optimal minimax rate of convergence on the class $\mathcal{D}(H, L, M)$. We say that two sequences of positive numbers (a_n) and (b_n) have the same rate of convergence, noted $a_n \asymp b_n$, if the ratios a_n/b_n and b_n/a_n remain bounded as $n \to +\infty$.

Theorem 3. *The optimal minimax rate of convergence on the class $\mathcal{D}(H, L, M)$ is given by*

$$\inf_{\hat{F}_n} \sup_{P\in\mathcal{D}(H,L,M)} \mathbb{E}[\mathcal{W}_1(\hat{F}_{n,X}, F_X)] \asymp n^{-H/(2H+k)}.$$

Theorem 3 is the counterpart of Pic et al. (2022, Theorem 1) where the minimax rate of convergence for the second order Cramér's distance has been considered. The strategy of proof is similar: i) we prove a lower bound by considering a suitable class of binary distributions where the error in Wasserstein distance corresponds to an absolute error in point regression for which the minimax lower rate of convergence is known; ii) we check that the upper bound for the kernel and/or nearest neighbor algorithm has the same rate of convergence as the lower bound, which proves that the optimal minimax rate of convergence has been identified. In particular, our proof shows that the kernel method defined in Equation (5) reaches the minimax rate of convergence in any dimension $k \geq 1$ with the choice of bandwidth $h_n \asymp n^{-1/(2H+k)}$; the nearest neighbor method defined in Equation (6) reaches the minimax rate of convergence in any dimension $k \geq 2$ with the number of neighbors $\kappa_n \asymp n^{H/(H+k/2)}$.

Remark 2. Our estimate of the minimax rate of convergence holds only for $d = p = 1$ and we briefly discuss what can be expected in other cases.

When $p = 1$ and $d \geq 2$, one may hope to use the strong equivalence between the max-sliced Wasserstein distance and the Wasserstein distance (Bayraktar and Guo, 2021, Theorem 2.3.ii). This requires to estimate the expectation of a supremum over the sphere and this line of research is left for further work.

When $p > 1$, even in dimension $d = 1$, it seems difficult to obtain bounds for the Wasserstein distance of order p without very strong assumptions. Bobkov and Ledoux (2019) consider the rate of convergence of the empirical distribution $\hat{F}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ for an i.i.d. sample $Y_1, \dots, Y_n$ with distribution F on $\mathbb{R}$. A first consistency result (Theorem 2.14) states that $\mathbb{E}[\mathcal{W}_p^p(\hat{F}_n, F)] \to 0$ as soon as F has a finite moment of order $p \geq 1$. Regarding rates of convergence, they show (Corollary 3.9) that for $p = 1$ the standard rate of convergence holds, i.e. $\mathbb{E}[\mathcal{W}_1(\hat{F}_n, F)] = O(1/\sqrt{n})$, if and only if

$$J_1(F) = \int_{\mathbb{R}} \sqrt{F(z)(1 - F(z))}\mathrm{d}z < \infty.$$

On the other hand, rate of convergences for higher order $p > 1$ require the condition

$$J_p(F) = \int_{\mathbb{R}} \frac{[F(z)(1 - F(z))]^{p/2}}{f(z)^{p-1}}\mathrm{d}z < \infty,$$

where f is the density of the absolutely continuous component of F. They show (Corollary 5.5) that the standard rate holds, i.e. $\mathbb{E}[\mathcal{W}_p^p(\hat{F}_n, F)] = O(n^{-p/2})$, if and only if $J_p(F) < \infty$. However, this condition is very strong: it does not hold for the Gaussian distribution or for distributions with disconnected support.

3.3 Applications

We briefly illustrate Theorem 2 with some applications and examples. In statistics, we commonly face the following generic situation: we are interested in a summary statistic S with real values, e.g. quantiles or tail expectation, and we want to assess the effect of X on Y through S, that is we want to assess $S_{Y|X=x}$. Assuming that S is well-defined for distributions on $\mathbb{R}^d$ with a finite moment of order $p \geq 1$, it can be seen as a map $S : \mathcal{W}_p(\mathbb{R}^d) \to \mathbb{R}$ and then $S_{Y|X=x} = S(F_x)$ with F_x the conditional distribution of Y given $X = x$. A natural plug-in estimate of $S_{Y|X=x}$ is

$$\hat{S}_{n,x} = S(\hat{F}_{n,x}) \quad \text{with } \hat{F}_{n,x} \text{ defined by (11).}$$

In this generic situation, our extension of Stone's theorem directly implies the following proposition. Recall that $M_p(\mu)$ is defined in Equation (7).

Proposition 2. *Assume* $\mathbb{E}[\|Y\|^p] < \infty$ *and* $\mathbb{P}(F_X \in \mathcal{C}) = 1$ *where* $\mathcal{C} \subset \mathcal{W}_p(\mathbb{R}^d)$ *denotes the continuity set of the statistic* $S : \mathcal{W}_p(\mathbb{R}^d) \to \mathbb{R}$. *Then weak consistency holds, i.e.*

$$\hat{S}_{n,X} \longrightarrow S_{Y|X} \quad \textit{in probability as } n \to +\infty.$$

If furthermore the statistic S *admits a bound of the form*

$$|S(\mu)| \leq aM_p^q(\mu) + b, \quad \textit{with } a, b \geq 0 \textit{ and } 0 < q \leq p, \tag{14}$$

then consistency holds in $\mathrm{L}^{p/q}$, *i.e.*

$$\mathbb{E}\big[|\hat{S}_{n,X} - S_{Y|X}|^{p/q}\big] \longrightarrow 0 \quad \textit{as } n \to +\infty$$

Example 3. (quantile). For a distribution G on $\mathbb{R}$, we define the associated quantile function

$$G^{-1}(\alpha) = \inf\{z \in \mathbb{R} : G(z) \geq \alpha\}, \quad 0 < \alpha < 1.$$

It is well-known that the weak convergence $G_n \overset{d}{\to} G$ implies the quantile convergence $G_n^{-1}(\alpha) \to G^{-1}(\alpha)$ at each continuity point α of G^{-1}. Equivalently, considering $\mathcal{P}(\mathbb{R})$ endowed with the weak convergence topology, the α-quantile statistic $S_\alpha(G) = G^{-1}(\alpha)$ is continuous at G as soon as G^{-1} is continuous at α.

In view of this, we let $\mathcal{C} = \{G \in \mathcal{P}(\mathbb{R})\colon G^{-1} \text{ continuous on } (0,1)\}$ and assume that the conditional distribution satisfies $\mathbb{P}(F_X \in \mathcal{C}) = 1$. Then weak convergence holds for the conditional quantiles, i.e.

$$\hat{F}_{n,X}^{-1}(\alpha) \to F_X^{-1}(\alpha) \quad \text{in probability.}$$

Note that no integrability condition is needed here because we can apply Proposition 2 on the transformed data $(X_i, \tilde{Y}_i)_{1 \leq i \leq n}$, where $\tilde{Y}_i = \tan^{-1}(Y_i)$ is bounded so that convergence in Wasserstein distance is equivalent to weak convergence. If furthermore Y is p-integrable, then the bound

$$\begin{aligned} |S_\alpha(G)|^p &\leq \frac{1}{\alpha}\int_0^\alpha |G^{-1}(u)|^p \mathrm{d}u + \frac{1}{1-\alpha}\int_\alpha^1 |G^{-1}(u)|^p \mathrm{d}u \\ &\leq \Big(\frac{1}{\alpha} + \frac{1}{1-\alpha}\Big) M_p^p(G) \end{aligned}$$

implies the strengthened convergence

$$\hat{F}_{n,X}^{-1}(\alpha) \to F_X^{-1}(\alpha) \quad \text{in } \mathrm{L}^p.$$

Example 4. (tail expectation) The tail expectation above level $\alpha \in (0,1)$ is the risk measure defined for $G \in \mathcal{W}_1(\mathbb{R})$ by

$$S_\alpha(G) = \frac{1}{1-\alpha}\int_\alpha^1 G^{-1}(u)\,\mathrm{d}u.$$

The name comes from the equivalent definition

$$S_\alpha(G) = \mathbb{E}[Y \mid Y > G^{-1}(\alpha)], \quad Y \sim G,$$

which holds when G^{-1} is continuous at α. One can see that

$$\begin{aligned} |S_\alpha(G_1) - S_\alpha(G_2)| &\leq \frac{1}{1-\alpha} \int_\alpha^1 |G_1^{-1}(u) - G_1^{-1}(u)| \, \mathrm{d}u \\ &\leq \frac{1}{1-\alpha} \int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)| \, \mathrm{d}u \\ &= \frac{1}{1-\alpha} \mathcal{W}_1(G_1, G_2). \end{aligned}$$

so that S_α is Lipschitz continuous with respect to the Wasserstein distance $\mathcal{W}_1$. As a consequence, the conditional tail expectation $S_\alpha(F_x)$ can be estimated in a consistent way by the plug-in estimator $S_\alpha(\hat{F}_{n,x})$ since

$$\mathbb{E}[|S_\alpha(\hat{F}_{n,X}) - S_\alpha(F_X)|] \leq \frac{1}{1-\alpha} \mathbb{E}[\mathcal{W}_1(\hat{F}_{n,X}, F_X)] \longrightarrow 0.$$

Example 5. (probability weighted moment) A similar result holds for the probability weighted moment of order $p, q > 0$ defined by

$$S_{p,q}(G) = \int_0^1 G^{-1}(u) u^p (1-u)^q \, \mathrm{d}u, \quad G \in \mathcal{W}_1(\mathbb{R}).$$

(Greenwood et al. (1979)). The name comes from the equivalent definition

$$S(G) = \mathbb{E}[Y G(Y)^p (1 - G(Y))^q], \quad Y \sim G,$$

which holds when G^{-1} is continuous on $(0, 1)$. One can again check that the statistic $S_{p,q}$ is Lipschitz continuous with respect to the Wasserstein distance $\mathcal{W}_1$ since

$$\begin{aligned} |S_{p,q}(G_1) - S_{p,q}(G_2)| &\leq \int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)| u^p (1-u)^q \, \mathrm{d}u \\ &\leq \max_{0 \leq u \leq 1} u^p (1-u)^q \times \int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)| \, \mathrm{d}u \\ &= \Big(\frac{p}{p+q}\Big)^p \Big(\frac{q}{p+q}\Big)^q \mathcal{W}_1(G_1, G_2). \end{aligned}$$

Example 6. (covariance) We conclude with a simple example in dimension $d = 2$ where the statistic of interest is the covariance between the two components of $Y = (Y_1, Y_2)$ given $X = x$. Here, we consider

$$S(G) = \int_{\mathbb{R}^2} y_1 y_2 \, \mathrm{d}G - \int_{\mathbb{R}^2} y_1 \, \mathrm{d}G \int_{\mathbb{R}^2} y_2 \, \mathrm{d}G, \quad G \in \mathcal{W}_2(\mathbb{R}^2).$$

Considering square integrable random vectors $Y = (Y_1, Y_2)$ and $Z = (Z_1, Z_2)$ with distribution G and H respectively, we compute

$$\begin{aligned}
&|S(G) - S(H)| \\
&= \big|\mathrm{Cov}(Y_1, Y_2) - \mathrm{Cov}(Z_1, Z_2)\big| \\
&= \big|\mathrm{Cov}(Y_1, Y_2 - Z_2) - \mathrm{Cov}(Z_1 - Y_1, Z_2)\big| \\
&\leq \mathrm{Var}(Y_1)^{1/2}\mathrm{Var}(Y_2 - Z_2)^{1/2} + \mathrm{Var}(Z_2)^{1/2}\mathrm{Var}(Z_1 - Y_1)^{1/2}
\end{aligned}$$

were the last line is a consequence of Cauchy-Schwartz inequality. We have the upper bounds

$$\mathrm{Var}(Y_1)^{1/2} \leq M_2(G), \quad \mathrm{Var}(Z_2)^{1/2} \leq M_2(H)$$

and, choosing an optimal coupling (Y, Z) between G and H,

$$\mathrm{Var}(Z_1 - Y_1)^{1/2} \leq \|Y - Z\|_{L^2} = \mathcal{W}_2(G, H), \quad \mathrm{Var}(Y_2 - Z_2)^{1/2} \leq \mathcal{W}_2(G, H).$$

Altogether, we obtain,

$$|S(G) - S(H)| \leq \big(M_2(G) + M_2(H)\big)\mathcal{W}_2(G, H).$$

This proves that S is locally Lipschitz and hence continuous with respect to the distance $\mathcal{W}_2$. Taking $H = \delta_0$, we obtain

$$|S(G)| \leq M_2(G)^2$$

and the bound (14) holds with $q = 2$. Thus Proposition 2 implies that the plug-in estimator

$$S(\hat{F}_{n,x}) = \sum_{i=1}^{n} W_{ni}(x) Y_{1i} Y_{2i} - \sum_{i=1}^{n} W_{ni}(x) Y_{1i} \sum_{i=1}^{n} W_{ni}(x) Y_{2i}$$

is consistent in absolute mean for the conditional covariance

$$S(F_x) = \mathbb{E}(Y_1 Y_2 \mid X = x) - \mathbb{E}(Y_1 \mid X = x)\mathbb{E}(Y_2 \mid X = x),$$

i.e. $\mathbb{E}[|S(\hat{F}_{n,X}) - S(F_X)|] \longrightarrow 0$ as $n \to +\infty$.

4 Proofs

4.1 Proof of Theorem 2

Proof of Theorem 2 - case $d = 1$. We first consider the case when Y is uniformly bounded and takes its values in $[-M, M]$ for some $M > 0$. Then, it holds

$$F_x(z) = \begin{cases} 0 & \text{if } z < -M \\ 1 & \text{if } z \geq M \end{cases} \quad \text{and} \quad \hat{F}_{n,x}(z) = \begin{cases} 0 & \text{if } z < -M \\ 1 & \text{if } z \geq M \end{cases}.$$

and the generalized inverse functions (quantile functions) are bounded in absolute value by M. As a consequence,

$$\begin{aligned}\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)\right] &= \mathbb{E}\left[\int_0^1 |\hat{F}_{n,X}^{-1}(u) - F_X^{-1}(u)|^p \mathrm{d}z\right] \\ &\leq (2M)^{p-1}\mathbb{E}\left[\int_0^1 |\hat{F}_{n,X}^{-1}(u) - F_X^{-1}(u)|\mathrm{d}u\right] \\ &= (2M)^{p-1}\int_{-M}^{M} \mathbb{E}\left[|\hat{F}_{n,X}(z) - F_X(z)|\right] \mathrm{d}z. \qquad (15)\end{aligned}$$

In this lines, we have used Equations (9) and (10) together with Fubini's theorem.

Consider the regression model $(X, \mathbb{1}_{\{Y\leq z\}}) \in \mathbb{R}^d \times \mathbb{R}$ where $z \in [-M, M]$ is fixed. The corresponding regression function is

$$x \mapsto \mathbb{E}[\mathbb{1}_{\{Y\leq z\}}|X = x] = F_x(z)$$

and the local weight estimator associated with the sample $(X_i, \mathbb{1}_{\{Y_i\leq z\}})$, $1 \leq i \leq n$ is

$$x \mapsto \sum_{i=1}^{n} W_{ni}(x)\mathbb{1}_{\{Y_i\leq z\}} = \hat{F}_{n,x}(z).$$

An application of Stone's theorem with $p = 1$ yields

$$\mathbb{E}\left[|\hat{F}_{n,X}(z) - F_X(z)|\right] \longrightarrow 0, \quad \text{as } n \to +\infty,$$

whence we deduce, by the dominated convergence theorem,

$$\int_{-M}^{M} \mathbb{E}\left[|\hat{F}_{n,X}(z) - F_X(z)|\right] \mathrm{d}z \longrightarrow 0.$$

The upper bound (15) finally implies

$$\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)\right] \longrightarrow 0.$$

We next consider the general case when Y is not necessarily bounded. For $M > 0$, we define the truncation Y^M of Y by

$$Y^M = \begin{cases} -M & \text{if } Y < -M \\ Y & \text{if } -M \leq Y < M \\ M & \text{if } Y \geq M \end{cases}.$$

We define similarly $Y_1^M, \ldots, Y_n^M$ the truncations of $Y_1, \ldots, Y_n$ respectively. The conditional distribution associated with Y^M is

$$F_x^M(z) = \mathbb{P}(Y^M \leq z|X = x) = \begin{cases} 0 & \text{if } z < -M \\ F_x(z) & \text{if } -M \leq Y < M \\ 1 & \text{if } z \geq M \end{cases}.$$

The local weight estimation built on the truncated sample is

$$\hat{F}_{n,x}^M(z) = \sum_{i=1}^n W_{ni}(x) \mathbb{1}_{\{Y_i^M \leq z\}}.$$

By the triangle inequality,

$$\mathcal{W}_p(\hat{F}_{n,x}, F_x) \leq \mathcal{W}_p(\hat{F}_{n,x}, \hat{F}_{n,x}^M) + \mathcal{W}_p(\hat{F}_{n,x}^M, F_x^M) + \mathcal{W}_p(F_x^M, F_x),$$

whence we deduce

$$\begin{aligned}
&\mathbb{E}[\mathcal{W}_p^p(\hat{F}_{n,x}, F_x)] \\
&\leq 3^{p-1} \left(\mathbb{E}[\mathcal{W}_p^p(\hat{F}_{n,X}, \hat{F}_{n,X}^M)] + \mathbb{E}[\mathcal{W}_p(\hat{F}_{n,X}^M, F_X^M)] + \mathbb{E}[\mathcal{W}_p^p(F_X^M, F_X)] \right).
\end{aligned}$$

By the preceding result in the bounded case, for any fixed M, the second term converge to 0 as $n \to +\infty$. We next focus on the first and third term.

For fixed $X = x$, there is a natural coupling between the distribution $\hat{F}_{n,x}$ and $\hat{F}_{n,x}^M$ given by (Z_1, Z_2) such that

$$(Z_1, Z_2) = (Y_i, Y_i^M) \quad \text{with probability } W_{ni}(x).$$

Clearly $Z_1 \sim \hat{F}_{n,x}$ and $Z_2 \sim \hat{F}_{n,x}^M$ and this coupling provides the upper bound

$$\mathcal{W}_p^p(\hat{F}_{n,x}, \hat{F}_{n,x}^M) \leq \|Z_1 - Z_2\|_{\mathrm{L}^p}^p = \sum_{i=1}^n W_{ni}(x) |Y_i - Y_i^M|^p. \tag{16}$$

Let us introduce the function $g_M(x)$ defined by

$$g_M(x) = \mathbb{E}\left[|Y - Y^M|^p \mid X = x \right].$$

Using the fact that, conditionally on $X_1, \ldots, X_n$, the random variables $Y_1, \ldots, Y_n$ are independent with distribution $F_{X_1}, \ldots, F_{X_n}$, we deduce

$$\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,x}, \hat{F}_{n,x}^M) \right] \leq \mathbb{E}\left[\sum_{i=1}^n W_{ni}(x) g_M(X_i) \right].$$

The condition i) on the weights in Stone's Theorem then implies

$$\mathbb{E}\left[\sum_{i=1}^n W_{ni}(X) g_M(X_i) \right] \leq C \mathbb{E}[g_M(X)].$$

Because $|Y - Y^M|^p$ converges almost surely to 0 as $M \to +\infty$ and is bounded by $2^p |Y|^p$ which is integrable, Lebesgue's convergence theorem implies

$$\mathbb{E}[g_M(X)] = \mathbb{E}\left[|Y - Y^M|^p \right] \longrightarrow 0 \quad \text{as } M \to +\infty.$$

We deduce that the first term satisfies

$$\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,X}, \hat{F}_{n,X}^M)\right] \leq C\mathbb{E}[g_M(X)] \longrightarrow 0, \quad \text{as } M \to +\infty$$

where the convergence is uniform in n.

We now consider the third term. Since Y^M is obtained from Y by truncation, the distribution functions and quantile functions of Y and Y^M are related by

$$F_x^M(z) = \begin{cases} 0 & \text{if } z < -M \\ F_x(z) & \text{if } -M \leq z < M \\ 1 & \text{if } z \geq M \end{cases}$$

and

$$(F_x^M)^{-1}(u) = \begin{cases} -M & \text{if } F_x^{-1}(u) < -M \\ (F_x)^{-1}(u) & \text{if } -M \leq F_x^{-1}(u) < M \\ M & \text{if } F_x^{-1}(u) \geq M \end{cases}.$$

As a consequence

$$\begin{aligned} \mathcal{W}_p^p(F_x^M, F_x) &= \int_0^1 |(F_x^M)^{-1}(u) - F_x^{-1}(u)|^p \mathrm{d}u \\ &= \mathbb{E}\left[|Y^M - Y|^p \mid X = x\right] = g_M(x). \end{aligned}$$

We deduce

$$\mathbb{E}\left[\mathcal{W}_p^p(F_X^M, F_X)\right] = \mathbb{E}[g_M(X)] \longrightarrow 0, \quad \text{as } M \to +\infty$$

where the convergence is uniform in n.

We finally combine the three terms. The sum can be made smaller than any $\varepsilon > 0$ by first choosing M large enough so that the first and third terms are smaller than $\varepsilon/3$ and then choosing n large enough so that the second term is smaller than $\varepsilon/3$. This proves Equation (12) and concludes the proof. □

In order to extend the proof from $d = 1$ to $d \geq 2$, we need the notion of *sliced Wasserstein distance*, see Bayraktar and Guo (2021) for instance. Let $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\| = 1\}$ be the unit sphere in $\mathbb{R}^d$ and, for $u \in \mathbb{R}^d$, let $u_* : \mathbb{R}^d \to \mathbb{R}$ be the linear form defined by $u_*(x) = u \cdot x$. The projection in direction u of a measure μ on $\mathbb{R}^d$ is defined as the pushforward $\mu \circ u_*^{-1}$ which is a measure on $\mathbb{R}$. The inequality $|u \cdot x| \leq \|x\|$ implies that $\mu \circ u_*^{-1} \in \mathcal{W}_p(\mathbb{R})$ for all $\mu \in \mathcal{W}_p(\mathbb{R}^d)$ and $u \in \mathbb{S}^{d-1}$. The sliced and max-sliced Wasserstein distances between $\mu, \nu \in \mathcal{W}_p(\mathbb{R}^d)$ are then defined respectively by

$$S\mathcal{W}_p(\mu, \nu) = \left(\int_{\mathbb{S}^{d-1}} \mathcal{W}_p^p(\mu \circ u_*^{-1}, \nu \circ u_*^{-1})\, \sigma(\mathrm{d}u)\right)^{1/p},$$

where σ denotes the uniform measure on $\mathbb{S}^{d-1}$ and

$$\overline{SW}_p(\mu,\nu) = \max_{u\in\mathbb{S}^{d-1}} \mathcal{W}_p(\mu\circ u_*^{-1}, \nu\circ u_*^{-1}).$$

In plain words, the sliced and max-sliced Wasserstein distance are respectively the average and the maximum over all the 1-dimensional Wasserstein distances between the projections of μ and ν. The following result is crucial in our proof.

Theorem 4 (Bayraktar and Guo (2021)). *For all $p \geq 1$, $S\mathcal{W}_p$ and $\overline{SW}_p$ are distances on $\mathcal{W}_p(\mathbb{R}^d)$ which are equivalent to $\mathcal{W}_p$, i.e. for all sequence $\mu, \mu_1, \mu_2, \ldots \in \mathcal{W}_p(\mathbb{R}^d)$*

$$S\mathcal{W}_p(\mu_n,\mu) \to 0 \iff \overline{SW}_p(\mu_n,\mu) \to 0 \iff \mathcal{W}_p(\mu_n,\mu) \to 0.$$

Proof of Theorem 2 - case $d \geq 2$. For the sake of clarity, we divide the proof into three steps:

1) we prove that the result holds in max-sliced Wasserstein distance, i.e. $\mathbb{E}[\overline{SW}_p^p(\hat{F}_{n,X}, F_X)] \to 0$;

2) we deduce that $\mathcal{W}_p(\hat{F}_{n,X}, F_X) \to 0$ in probability;

3) we show that the sequence $\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)$ is uniformly integrable.

Points 2) and 3) together imply $\mathbb{E}[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)] \to 0$ as required.

Step 1). For all $u \in \mathbb{S}^{d-1}$, the projection $\hat{F}_{n,X}\circ u_*^{-1}$ is the weighted empirical distribution

$$\hat{F}_{n,X}\circ u_*^{-1} = \sum_{i=1}^n W_{ni}(X)\delta_{Y_i\cdot u}.$$

An application of Theorem 2 to the 1-dimensional sample $(Y_i\cdot u)_{i\geq 1}$ yields

$$\mathbb{E}[\mathcal{W}_p^p(\hat{F}_{n,X}\circ u_*^{-1}, F_X\circ u_*^{-1})] \longrightarrow 0. \tag{17}$$

Note indeed that $\mathbb{E}[|Y|^p] < \infty$ implies $\mathbb{E}[|Y\cdot u|^p] < \infty$ and that the conditional laws of $Y\cdot u$ are the pushforward of those of Y, i.e. $\mathcal{L}(Y\cdot u \mid X) = F_X\circ u_*^{-1}$.

We next consider the max-sliced Wasserstein distance. Regularity in the direction $u \in \mathbb{S}^{d-1}$ will be useful and we recall that the Wasserstein distance between projections depends on the direction in a Lipschitz way. More precisely, according to Bayraktar and Guo (2021, Proposition 2.2),

$$|\mathcal{W}_p(\mu\circ u_*^{-1}, \nu\circ u_*^{-1}) - \mathcal{W}_p(\mu\circ v_*^{-1}, \nu\circ v_*^{-1})| \leq (M_p(\mu)+M_p(\nu))\|u-v\|,$$

for all $\mu, \nu \in \mathcal{W}_p(\mathbb{R}^d)$ and $u, v \in \mathbb{S}^{d-1}$ (recall Equation (7) for the definition of $M_p(\mu)$, $M_p(\nu)$).

The sphere $\mathbb{S}^{d-1}$ being compact, for all $\varepsilon > 0$, one can find $K \geq 1$ and $u_1, \ldots, u_K \in \mathbb{S}^{d-1}$ such that the balls $B(u_i, \varepsilon)$ with centers u_i and radius ε cover the sphere. Then, due to the Lipschitz property, the max-sliced Wasserstein distance is controlled by

$$\begin{aligned}
&\overline{SW}_p(\hat{F}_{n,X}, F_X) \\
&= \max_{u \in \mathbb{S}^{d-1}} \mathcal{W}_p^p(\hat{F}_{n,X} \circ u_*^{-1}, F_X \circ u^{-1}) \\
&\leq \max_{1 \leq k \leq K} \mathcal{W}_p(\hat{F}_{n,X} \circ u_{k*}^{-1}, F_X \circ u_{k*}^{-1}) + \varepsilon(M_p(\hat{F}_{n,X}) + M_p(F_X)).
\end{aligned}$$

Elevating to the p-th power and taking the expectation, we deduce

$$\begin{aligned}
&\mathbb{E}\big[\overline{SW}_p^p(\hat{F}_{n,X}, F_X)\big] \\
&\leq 3^{p-1}\mathbb{E}\big[\max_{1 \leq k \leq K} \mathcal{W}_p^p(\hat{F}_{n,X} \circ u_{k*}^{-1}, F_X \circ u_{k*}^{-1})\big] + 3^{p-1}\varepsilon^p(\mathbb{E}\big[M_p^p(\hat{F}_{n,X})\big] + \mathbb{E}\big[M_p^p(F_X)\big]).
\end{aligned}$$

The first term converges to 0 thanks to Eq. (17), i.e.

$$\mathbb{E}[\max_{1 \leq i \leq K} \mathcal{W}_p^p(\hat{F}_{n,X} \circ u_{i*}^{-1}, F_X \circ u_{i*}^{-1})] \longrightarrow 0.$$

The second term is controlled by a constant times ε^p since

$$\mathbb{E}[M_p^p(\hat{F}_{n,X})] = \mathbb{E}\big[\sum_{i=1}^n W_{ni}(X)\|Y_i\|^p\big] \leq C\mathbb{E}[\|Y\|^p]$$

(by property i) of the weights) and

$$\mathbb{E}[M_p^p(F_X)] = \mathbb{E}\big[\mathbb{E}[\|Y\|^p \mid X]\big] = \mathbb{E}[\|Y\|^p]$$

(by the tower property of conditional expectation). Letting $\varepsilon \to 0$, the second term can be made arbitrarily small. We deduce $\mathbb{E}[\overline{SW}_p^p(\hat{F}_{n,X}, F_X)] \to 0$.

Step 2). As a consequence of step 1), $\overline{SW}_p(\hat{F}_{n,X}, F_X) \to 0$ in probability, or equivalently $\hat{F}_{n,X} \to F_X$ in probability in the metric space $(\mathcal{W}_p(\mathbb{R}^d), \overline{SW}_p)$. Theorem 4 implies that the identity mapping is continuous from $(\mathcal{W}_p(\mathbb{R}^d), \overline{SW}_p)$ into $(\mathcal{W}_p(\mathbb{R}^d), \mathcal{W}_p)$. The continuous mapping theorem implies that $\hat{F}_{n,X} \to F_X$ in probability in the metric space $(\mathcal{W}_p(\mathbb{R}^d), \mathcal{W}_p)$. Equivalently, $\mathcal{W}_p(\hat{F}_{n,X}, F_X) \to 0$ in probability.

Step 3). By the triangle inequality,

$$\mathcal{W}_p(\hat{F}_{n,X}, F_X) \leq \mathcal{W}_p(\hat{F}_{n,X}, \delta_0) + \mathcal{W}_p(\delta_0, F_X)$$

with δ_0 the Dirac mass at 0. Furthermore, for any $\mu \in \mathcal{W}_p(\mathbb{R}^d)$,

$$\mathcal{W}_p(\mu, \delta_0) = \left(\int_{\mathbb{R}^d} \|x\|^p \, \mu(\mathrm{d}x)\right)^{1/p} = M_p(\mu).$$

We deduce

$$\mathcal{W}_p^p(\hat{F}_{n,X}, F_X) \leq 2^{p-1} M_p^p(\hat{F}_{n,X}) + 2^{p-1} M_p^p(F_X).$$

In order to prove the uniform integrability of the left hand side, it is enough to prove that

$$M_p^p(F_X) \text{ is integrable and } M_p^p(\hat{F}_{n,X}),\ n \geq 1, \text{ is uniformly integrable.} \quad (18)$$

We have

$$M_p^p(F_X) = \mathbb{E}[\|Y\|^p \mid X]$$

which is integrable because $\mathbb{E}[\|Y\|^p] < \infty$. Furthermore,

$$M_p^p(\hat{F}_{n,X}) = \sum_{i=1}^n W_{ni}(X)\|Y_i\|^p$$

and Stone's Theorem ensures that

$$\sum_{i=1}^n W_{ni}(X)\|Y_i\|^p \longrightarrow \mathbb{E}[\|Y\|^p \mid X] \quad \text{in } L^1.$$

Since the sequence $M_p^p(\hat{F}_{n,X})$ converges in L^1, it is uniformly integrable and the claim follows. □

4.2 Proof of Proposition 1, Corollaries 1-2 and Theorem 3

Proof of Proposition 1. The proof of the upper bound relies on a coupling argument. Without loss of generality, we can assume that the Y_i's are generated from uniform random variables U_i's by the inversion method – i.e. we assume that U_i, $1 \leq i \leq n$, are independent identically distributed random variables with uniform distribution on $(0,1)$ that are furthermore independent from the covariates X_i, $1 \leq i \leq n$ and we set $Y_i = F_{X_i}^{-1}(U_i)$. Then the sample (X_i, Y_i) is i.i.d. with distribution P. In order to compare $\hat{F}_{n,x}$ and F_x, we introduce the random variables $\tilde{Y}_i = F_x^{-1}(U_i)$ and we define

$$\tilde{F}_{n,x}(z) = \sum_{i=1}^n W_{ni}(x) \mathbb{1}_{\{\tilde{Y}_i \leq z\}}.$$

By the triangle inequality,

$$\mathcal{W}_1(\hat{F}_{n,x}, F_x) \leq \mathcal{W}_1(\hat{F}_{n,x}, \tilde{F}_{n,x}) + \mathcal{W}_1(\tilde{F}_{n,x}, F_x).$$

In the right hand side, the first term is interpreted as an *approximation error* comparing the weighted sample $(Y_i, W_{ni}(x))$ to $(\tilde{Y}_i, W_{ni}(x))$ where the $\tilde{Y}_i$ have the target distribution F_x. The second term is an *estimation error*

where we use the weighted sample $(\tilde{Y}_i, W_{ni}(x))$ with the correct distribution to estimate F_x.

We first consider the approximation error. A similar argument as for the proof of Equation (16) implies

$$\mathcal{W}_1(\hat{F}_{n,x}, \tilde{F}_{n,x}) \leq \sum_{i=1}^{n} W_{ni}(x)|Y_i - \tilde{Y}_i|.$$

Introducing the uniform random variables U_i's, we get

$$\begin{aligned}
\mathbb{E}[\mathcal{W}_1(\hat{F}_{n,x}, \tilde{F}_{n,x})] &\leq \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(x)|F_{X_i}^{-1}(U_i) - F_x^{-1}(U_i)|\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(x) \int_0^1 |F_{X_i}^{-1}(u) - F_x^{-1}(u)| \, \mathrm{d}u\Big] \quad \text{by independence} \\
&= \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(x)\mathcal{W}_1(F_{X_i}, F_x)\Big],
\end{aligned}$$

where the equality relies on Equation (9). Note that this control of the approximation error is very general and could be extended to the Wasserstein distance of order $p > 1$.

We next consider the estimation error and our approach works for $p = 1$ only. By Equation (10),

$$\mathbb{E}[\mathcal{W}_1(\tilde{F}_{n,x}, F_x)] = \mathbb{E}\Big[\int_{\mathbb{R}} \Big|\sum_{i=1}^{n} W_{ni}(x)\big(\mathbb{1}_{\{\tilde{Y}_i \leq z\}} - F_x(z)\big)\Big| \mathrm{d}z\Big].$$

Applying Fubini's theorem and using the upper bound

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\sum_{i=1}^{n} W_{ni}(x)\big(\mathbb{1}_{\{\tilde{Y}_i \leq z\}} - F_x(z)\big)\Big|\Big] \\
&\leq \mathbb{E}\Big[\Big|\sum_{i=1}^{n} W_{ni}(x)\big(\mathbb{1}_{\{\tilde{Y}_i \leq z\}} - F_x(z)\big)\Big|^2\Big]^{1/2} \\
&= \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}^2(x)\Big]^{1/2} \sqrt{F_x(z)(1 - F_x(z))},
\end{aligned}$$

we deduce

$$\mathbb{E}[\mathcal{W}_1(\tilde{F}_{n,x}, F_x)] \leq \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}^2(x)\Big]^{1/2} \int_{\mathbb{R}} \sqrt{F_x(z)(1 - F_x(z))} \mathrm{d}z.$$

Collecting the two terms yields Proposition 1. □

Proof of Corollary 1. For the kernel algorithm with uniform kernel and weights (5), we denote by

$$N_n(X) = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in B(X,h_n)\}}$$

the number of points in the ball $B(X, h_n)$ with center X and radius h_n. If $N_n \geq 1$, only the points in $B(X, h_n)$ have a nonzero weight which is equal to $1/N_n$. If $N_n = 0$, then by convention all the weights are equal to $1/n$. Thus we deduce

$$\mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}^2(X)\Big] = \mathbb{E}\Big[\frac{1}{N_n(X)}\mathbb{1}_{\{N_n(X)\geq 1\}}\Big] + \frac{1}{n}\mathbb{P}(N_n(X) = 0)$$

and

$$\mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(X)\|X_i - X\|^H\Big] \leq h_n^H \mathbb{P}(N_n(X) \geq 1) + k^{H/2}\mathbb{P}(N_n(X) = 0)$$

because the distance to X for the points with non zero weight can be bounded from above by h_n if $N_n(X) \geq 1$ and by $\sqrt{k}$ otherwise (note that $\sqrt{k}$ is the diameter of $[0,1]^k$).

Next, we use the fact that, conditionally on $X = x$, $N_n(x)$ has a binomial distribution with parameters n and $p_n(x) = \mathbb{P}(X_1 \in B(x, h_n))$. This implies

$$\mathbb{E}\Big[\frac{1}{N_n(X)}\mathbb{1}_{\{N_n(X)\geq 1\}}\Big] \leq \mathbb{E}\Big[\frac{2}{np_n(X)}\Big] \leq \frac{2c_k}{nh_n^k}$$

where the first inequality follows from Györfi et al. (2002, Lemma 4.1) and the second one from Györfi et al. (2002, Equation 5.1) where the constant $c_k = k^{k/2}$ can be taken. Similarly,

$$\begin{aligned}\mathbb{P}(N_n(X) = 0) = \mathbb{E}[(1 - p_n(X))^n] &\leq \mathbb{E}[e^{-np_n(X)}] \\ &\leq \big(\max_{u>0} ue^{-u}\big) \times \mathbb{E}\Big[\frac{1}{np_n(X)}\Big] \\ &\leq \frac{c_k}{nh_n^k}.\end{aligned}$$

In view of these different estimates, Equation (13) entails

$$\begin{aligned}\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\big] &\leq L\Big(h_n^H + k^{H/2}\frac{c_k}{nh_n^k}\Big) + M\left(\frac{(2+1/n)c_k}{nh_n^k}\right)^{1/2} \\ &\leq Lh_n^H + M\sqrt{(2+1/n)c_k}(nh_n^k)^{-1/2} + Lk^{H/2}c_k(nh_n^k)^{-1}.\end{aligned}$$

□

Proof of Corollary 2. For the nearest neighbor weights (6), there are exactly κ_n non-vanishing weights with value $1/\kappa_n$ whence

$$\sum_{i=1}^{n} W_{ni}^2(X) = \frac{1}{\kappa_n}.$$

Furthermore, the κ_n nearest neighbors of X satisfy

$$\|X_{i:n}(X) - X\| \leq \|X_{\kappa_n:n}(X) - X\|, \quad i = 1, \ldots, \kappa_n.$$

In view of this, Equation (13) entails

$$\begin{aligned}\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\big] &\leq L\mathbb{E}\big[\|X_{\kappa_n:n}(X) - X\|^H\big] + M\kappa_n^{-1/2} \\ &\leq L\mathbb{E}\big[\|X_{\kappa_n:n}(X) - X\|^2\big]^{H/2} + M\kappa_n^{-1/2}\end{aligned}$$

where the last line relies on Jensen's inequality. We conclude thanks to Biau and Devroye (2015, Theorem 2.4) stating that

$$\mathbb{E}\left[\|X_{\kappa_n:n}(X) - X\|^2\right] \leq \begin{cases} 8(\kappa_n/n) & \text{if } k = 1, \\ \tilde{c}_k(\kappa_n/n)^{2/k} & \text{if } k \geq 2. \end{cases}$$

□

Proof of Theorem 3 (lower bound). The proof of a lower bound for the minimax risk in Wasserstein distance is adapted from the proof of Proposition 3 in Pic et al. (2022, Appendix C) and we give only the main lines.

Consider the subclass of $\mathcal{D}(H, L, M)$ where Y is a binary variable with possible values 0 and B. Note that condition c) of Definition 1 is automatically satisfied if $B \leq 4M$. The conditional distribution of Y given $X = x$ is characterized by

$$p(x) = \mathbb{P}(Y = B \mid X = x)$$

and the Wasserstein distance by

$$\mathcal{W}_1(F_x, F_{x'}) = B|p(x) - p(x')|,$$

so that property b) of Definition 1 is equivalent to

$$B|p(x) - p(x')| \leq L\|x - x'\|^H. \tag{19}$$

Similarly as in Pic et al. (2022, Lemma 1), one can show that a general prediction with values in $\mathbb{R}$ can always be improved (in terms of Wasserstein error) into a binary prediction with values in $\{0, B\}$. Indeed, for a given prediction $\hat{F}_{n,x}$, the binary prediction

$$\tilde{F}_{n,x} = (1 - \tilde{p}_n(x))\delta_0 + \tilde{p}_n(x)\delta_B$$

with

$$\tilde{p}_n(x) = \frac{1}{B}\int_0^B \big(1 - \hat{F}_{n,x}(z)\big)\mathrm{d}z$$

always satisfies

$$\mathbb{E}[\mathcal{W}_1(\tilde{F}_{n,X}, F_X)] \leq \mathbb{E}[\mathcal{W}_1(\hat{F}_{n,X}, F_X)].$$

This simple remark implies that, when considering the minimax risk on the restriction of the class $\mathcal{D}(H, L, M)$ to binary distributions, we can focus on binary predictions. But for binary predictions,

$$\mathbb{E}[\mathcal{W}_1(\tilde{F}_{n,X}, F_X)] = B|\tilde{p}_n(X) - p(X)|,$$

showing that the minimax rate of convergence for distributional regression in Wasserstein distance is equal to the minimax rate of convergence for estimating the regression function $\mathbb{E}[Y|X = x] = Bp(x)$ in absolute error under the regularity assumption (19) . According to Stone (1980, 1982), a lower bound for the minimax risk in L^1-norm is $n^{-H/(2H+k)}$ (in the first paper, we consider the Bernoulli regression model referred to as Model 1 Example 5 and the L^q distance with $q = 1$). □

Proof of Theorem 3 (upper bound). For the kernel method, Corollary 1 states that the expected Wasserstein error is upper bounded by

$$Lh_n^H + M\sqrt{(2 + 1/n)c_k}(nh_n^k)^{-1/2} + Lk^{H/2}c_k(nh_n^k)^{-1}.$$

Minimizing the sum of the first two terms in the right-hand side with respect to h_n leads to $h_n \propto n^{1/(2H+1)}$ and implies that right-hand side is of order $n^{-H/(2H+k)}$ (the last term is negligible). This matches the minimax lower rate of convergence previously stated previously and proves that the optimal minimax risk is of order $n^{-H/(2H+k)}$.

For the nearest neighbor method, minimizing the upper bound for the expected Wasserstein error from Corollary 2 leads to

$$\kappa_n \propto \begin{cases} n^{H/(H+1)} & \text{if } k = 1 \\ n^{H/(H+k/2)} & \text{if } k \geq 2 \end{cases},$$

with a corresponding risk of order

$$\begin{cases} n^{-H/(2H+2)} & \text{if } k = 1 \\ n^{-H/(2H+k)} & \text{if } k \geq 2 \end{cases},$$

whence the nearest neighbor method reaches the optimal rate when $k \geq 2$. □

4.3 Proof of Proposition 2

Proof of Proposition 2. The first point follows from the fact that composition by a continuous application respects convergence in probability. Indeed, as the estimator $\hat{F}_{n,X}$ converges to F_X in probability for the Wasserstein distance $\mathcal{W}_p$, $S(\hat{F}_{n,X})$ converges to $S(F_X)$ in probability.

In order to prove the consistency in $\mathrm{L}^{p/q}$, it is enough to prove furthermore the uniform integrability of $|S(\hat{F}_{n,X}) - S(F_X)|^{p/q}$, $n \geq 1$. With the convexity inequality of power functions as $p/q \geq 1$, Equation (14) entails

$$\begin{aligned}
|S(\hat{F}_{n,X}) - S(F_X)|^{p/q} &\leq 2^{p/q-1}\big(|S(\hat{F}_{n,X})|^{p/q} + |S(F_X)|^{p/q}\big) \\
&\leq 2^{p/q-1}\Big((aM_p^q(\hat{F}_{n,X}) + b)^{p/q} + (aM_p^q(F_X) + b)^{p/q}\Big) \\
&\leq 2^{2(p/q-1)}\Big(a^{p/q}M_p^p(\hat{F}_{n,X}) + a^{p/q}M_p^p(F_X) + 2b^{p/q}\Big).
\end{aligned}$$

This upper bound together with Equation (18) implies the uniform integrability of $|S(\hat{F}_{n,X}) - S(F_X)|^{p/q}$, $n \geq 1$, which concludes the proof. □

Acknowledgements

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project). They are also grateful to Mehdi Dagdoug for suggesting the example of random forest weights (Example 2).

References

Erhan Bayraktar and Gaoyue Guo. Strong equivalence between metrics of Wasserstein type. *Electronic Communications in Probability*, 26(none):1 – 13, 2021. doi: 10.1214/21-ECP383. URL https://doi.org/10.1214/21-ECP383.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015.

Sergey Bobkov and Michel Ledoux. One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Mem. Amer. Math. Soc.*, 261(1259):v+126, 2019. ISSN 0065-9266. doi: 10.1090/memo/1259. URL https://doi.org/10.1090/memo/1259.

Leo Breiman. Random forests. *Machine Learning*, 45, 2001. doi: 10.1023/a:1010933404324.

Tilmann Gneiting and Matthiass Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and its Applications*, 2014. doi: 10.1146/annurev-statistics-062713-085831.

Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, 133 (5):1098 – 1118, 2005. doi: 10.1175/MWR2904.1.

J. Arthur Greenwood, J. Maciunas Landwehr, N. C. Matalas, and J. R. Wallis. Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, 15(5):1049–1054, 1979. doi: https://doi.org/10.1029/WR015i005p01049. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR015i005p01049.

Lászlò Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, 2002.

Thomas M. Hamill and Stephen J. Colucci. Verification of eta–rsm short-range ensemble forecasts. *Monthly Weather Review*, 125, jun 1997. doi: 10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

Alexander Henzi, Johanna F. Ziegel, and Tilmann Gneiting. Isotonic distributional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 83(5): 963–993, 2021. ISSN 1369-7412.

Rui Li, Brian J. Reich, and Howard D. Bondell. Deep distribution regression. *Computational Statistics & Data Analysis*, 159:107203, 2021. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2021.107203.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964. doi: 10.1137/1109020. URL https://doi.org/10.1137/1109020.

Victor M. Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. SpringerBriefs in Probability and Mathematical Statistics. Springer, Cham, 2020. ISBN 978-3-030-38437-1; 978-3-030-38438-8. doi: 10.1007/978-3-030-38438-8. URL https://doi.org/10.1007/978-3-030-38438-8.

Romain Pic, Clément Dombry, Philippe Naveau, and Maxime Taillardat. Distributional regression and its evaluation with the crps: Bounds and convergence of the minimax risk. *International Journal of Forecasting*, 2022. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2022.11.001. URL https://www.sciencedirect.com/science/article/pii/S0169207022001443.

Benedikt Schulz and Sebastian Lerch. Machine learning methods for post-processing ensemble forecasts of wind gusts: A systematic comparison, 2021. arXiv:2106.09512.

Erwan Scornet. On the asymptotics of random forests. *J. Multivariate Anal.*, 146:72–83, 2016. ISSN 0047-259X. doi: 10.1016/j.jmva.2015.06.009. URL https://doi.org/10.1016/j.jmva.2015.06.009.

Charles J. Stone. Consistent nonparametric regression. *Ann. Statist.*, 5(4):595–645, 1977. ISSN 0090-5364. URL http://links.jstor.org/sici?sici=0090-5364(197707)5:4<595:CNR>2.0.CO;2-O&origin=MSI With discussion and a reply by the author.

Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6):1348–1360, 1980. ISSN 0090-5364. URL http://links.jstor.org/sici?sici=0090-5364(198011)8:6<1348:OROCFN>2.0.CO;2-Q&origin

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982. ISSN 0090-5364. URL http://links.jstor.org/sici?sici=0090-5364(198212)10:4<1040:OGROCF>2.0.CO;2-2&orig

Maxime Taillardat, Anne-Laure Fougères, Philippe Naveau, and Olivier Mestre. Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, 34, jun 2019. doi: 10.1175/WAF-D-18-0149.1.

Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9. URL https://doi.org/10.1007/978-3-540-71050-9. Old and new.

Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372, 1964. ISSN 0581572X. URL http://www.jstor.org/stable/25049340.

Domagoj Ćevid, Loris Michel, Jeffrey Näf, Nicolai Meinshausen, and Peter Bühlmann. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022. URL http://jmlr.org/papers/v23/21-0585.html.