Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Distributional regression and its evaluation with the CRPS: Bounds and convergence of the minimax risk

Romain Pic[a,*], Clément Dombry[a], Philippe Naveau[b], Maxime Taillardat[c,d]

[a] Laboratoire de Mathématiques de Besançon, CNRS UMR 6623, Univ. Bourgogne Franche-Comté, Besançon, France
[b] Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, CEA-CNRS-UVSQ, IPSL & U Paris-Saclay, Gif-sur-Yvette, France
[c] CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France
[d] Météo-France, Toulouse, France

ARTICLE INFO

Keywords:
Probabilistic forecasting
Distributional regression
CRPS
Minimax rate of convergence
Nearest neighbor method
Kernel method

ABSTRACT

The theoretical advances in the properties of scoring rules over the past decades have broadened the use of scoring rules in probabilistic forecasting. In meteorological forecasting, statistical postprocessing techniques are essential to improve the forecasts made by deterministic physical models. Numerous state-of-the-art statistical postprocessing techniques are based on distributional regression evaluated with the continuous ranked probability score (CRPS). However, the theoretical properties of such evaluations with the CRPS have solely considered the unconditional framework (i.e. without covariates) and infinite sample sizes. We extend these results and study the rate of convergence in terms of the CRPS of distributional regression methods. We find the optimal minimax rate of convergence for a given class of distributions and show that the k-nearest neighbor method and the kernel method reach this optimal minimax rate.

1. Introduction

In meteorology, ensemble forecasts are based on a given number of deterministic models whose parameters vary slightly in order to consider observation errors and incomplete physical representations of the atmosphere. This leads to an ensemble of different forecasts that also assess the overall uncertainty of the forecast. Ensemble forecasts suffer from bias and underdispersion (Hamill & Colucci, 1997; Baran & Lerch, 2018) 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., 2016), and neural networks (Schulz & Lerch, 2022). These references, among others, also discuss the stakes of weather forecast statistical postprocessing.

Postprocessing methods rely on distributional regression (Gneiting & Katzfuss, 2014) where the aim is to predict the conditional distribution of the quantity of interest (e.g. temperature, wind speed, or precipitation) given a set of covariates (e.g. raw outputs of a physical ensemble model). Algorithms are often based on the minimization of a proper scoring rule that compares actual observations with the predictive distribution. Scoring rules can be seen as the equivalent to loss functions in classical regression. A detailed review of scoring rules is given by Gneiting and Raftery (2007). The continuous ranked probability score (CRPS; Matheson & Winkler, 1976), defined in Eq. (2), is one of the most popular scores in meteorological forecasting. The CRPS is also minimized to infer the parameters of statistical models used in postprocessing (e.g. Gneiting et al., 2005; Naveau et al., 2016; Rasp & Lerch, 2018; Taillardat et al., 2019). Recently, under monotonicity assumptions, the isotonic distributional regression (Henzi et al., 2021) was shown to minimize the in-sample CRPS and to satisfy consistency in the sense of the Kolmogorov distance.

[*] Corresponding author.
E-mail address: romain.pic@univ-fcomte.fr (R. Pic).

https://doi.org/10.1016/j.ijforecast.2022.11.001
0169-2070/© 2022 The Authors. Published by Elsevier B.V. on behalf of International Institute of Forecasters. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

To the best of our knowledge, most convergence statements in distributional regression (e.g. Thorey et al., 2017 and Mösching & Dümbgen, 2020) are not only derived within an unconditional framework, i.e. without taking into account the covariates, but also these limiting results assume arbitrarily large sample sizes. In this work, our goal is to bypass these two limitations.

This paper is organized as follows. Section 2 introduces preliminary notions that are needed to state our main results in Section 3. Section 2.1 introduces our framework and notation for distributional regression. Section 2.2 provides the theoretical background on distributional regression and its evaluation using the CRPS, and Section 2.3 provides some elements on minimax risk theory. Section 2.4 briefly introduces the two models studied here: the k-nearest neighbor and kernel estimators. The main results on the minimax rate of convergence for distributional regression are stated in Section 3.1 where suitable classes of distributions $\mathcal{D}^{(h,C,M)}$ are defined. In Section 3.2, we study the k-NN estimators and derive a non-asymptotic upper bound for the excess risk of the CRPS uniformly on the class $\mathcal{D}^{(h,C,M)}$. Section 3.3 provides similar results for the kernel method. In Section 3.4, we find a lower minimax rate of convergence by reducing the problem to standard point regression solved by Györfi et al. (2002). We can deduce that the k-NN method for the distributional regression reaches the optimal rate of convergence in dimension $d \geq 2$, while the kernel method reaches the optimal rate of convergence in any dimension. All the proofs are postponed to and detailed in the Appendix. Section 4 provides a short conclusion and discussion.

2. Preliminaries

2.1. Distributional regression framework

We consider the regression framework $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ with distribution P. The goal of distributional regression is to estimate the conditional distribution of Y given $X = x$, noted

$$F_x^*(y) := P(Y \leq y | X = x), \quad x \in \mathbb{R}^d.$$

In forecast assessment, we make the distinction between the construction of the estimator relying on the training sample $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$ and its evaluation with respect to new data (X, Y). Given the training sample D_n, the forecaster constructs a predictor $\hat{F}_n : x \mapsto \hat{F}_{n,x}$ that estimates the conditional distribution F_x^*. In this context, it is crucial to assess whether $\hat{F}_{n,x}$ is close to F_x^* over the entire range of possible values of $X = x$. To this aim, we consider

$$\mathbb{E}_{X\sim P_X, D_n\sim P^n}\left[\int_{\mathbb{R}} |\hat{F}_{n,X}(z) - F_X^*(z)|^2 \mathrm{d}z\right] \quad (1)$$

where P_X denotes the marginal distribution of X, and $\mathbb{E}_{X\sim P_X, D_n\sim P^n}$ denotes the expectation with respect to X and D_n following P_X and P^n, respectively. The squared L^2-norm within the expectation is usually referred to as the squared second-order *Cramér's distance*. We focus on this specific distance because it corresponds to the excess risk associated with the CRPS, also called the *divergence* of the CRPS, as explained in the next section.

2.2. CRPS and evaluation of distributional regression

The continuous ranked probability score (CRPS; Matheson & Winkler, 1976) compares a predictive distribution F and a real-valued observation y by computing the following integral:

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}_{y\leq z})^2 \mathrm{d}z. \quad (2)$$

The expected CRPS of a predictive distribution F when observations Y are distributed according to G is defined as

$$\overline{\mathrm{CRPS}}(F, G) = \int_{\mathbb{R}} \mathrm{CRPS}(F, y) G(\mathrm{d}y), \quad F, G \in \mathcal{M}(\mathbb{R}), \quad (3)$$

where $\mathcal{M}(\mathbb{R})$ denotes the set of all distribution functions on $\mathbb{R}$. This quantity is finite when both F and G have a finite first moment. Then the difference between the expected CRPS of forecast F and the expected CRPS of the ideal forecast G can be written as

$$\overline{\mathrm{CRPS}}(F, G) - \overline{\mathrm{CRPS}}(G, G) = \int_{\mathbb{R}} |F(z) - G(z)|^2 \mathrm{d}z \geq 0. \quad (4)$$

This implies that the only optimal prediction, in the sense that it minimizes the expected CRPS, is the true distribution G. A score with this property is said to be *strictly proper*. This property is essential for distributional regression, as it justifies the minimization of the expected score in order to construct or evaluate a prediction.

In distributional regression, the quality of a predictor $\hat{F} : x \mapsto \hat{F}_x$ is assessed by its risk:

$$\begin{aligned} R_P(\hat{F}) &= \mathbb{E}_{(X,Y)\sim P}\left[\mathrm{CRPS}(\hat{F}_X, Y)\right] \\ &= \mathbb{E}_{X\sim P_X}\left[\overline{\mathrm{CRPS}}(\hat{F}_X, F_X^*)\right]. \end{aligned}$$

This quantity is important, as many distributional regression methods try to minimize it in order to improve predictions. When Y is integrable, Eq. (4) implies

$$\begin{aligned} R_P(\hat{F}) - R_P(F^*) &= \mathbb{E}_{(X,Y)\sim P}\left[\mathrm{CRPS}(\hat{F}_X, Y) - \mathrm{CRPS}(F_X^*, Y)\right] \\ &= \mathbb{E}_{X\sim P_X}\left[\int_{\mathbb{R}} \left|\hat{F}_X(z) - F_X^*(z)\right|^2 \mathrm{d}z\right] \geq 0. \quad (5) \end{aligned}$$

We recall that the Bayes risk is the minimal theoretical risk over all possible predictors and that a Bayes predictor is a predictor achieving the Bayes risk. Thus, Eq. (5) implies that $R_P(F^*)$ is the Bayes risk and that F^* is a Bayes predictor if and only if $\hat{F}_x = F_x^*$ P_X-a.e. An introduction to the notions of theoretical risk, Bayes risk, and excess risk can be found in Section 2.4 of Hastie et al. (2009).

Finally, we consider the case of a predictor $\hat{F}_n$ built on a training sample $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$, as presented in Section 2.1, to estimate the conditional distribution of Y given X. Then (X, Y) denotes a new independent observation used to evaluate the performances of $\hat{F}_n$. The predictor has the expected CRPS

$$\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] = \mathbb{E}_{D_n\sim P^n,(X,Y)\sim P}[\mathrm{CRPS}(\hat{F}_{n,X}, Y)],$$

with expectation taken both with respect to the training sample D_n and test observation (X, Y). Once again, when

Y is integrable, the theoretical risk has a unique minimum given by $R_P(F^*)$. The *excess risk* becomes

$$\mathbb{E}_{D_n\sim P^n}\left[R_P(\hat{F}_n)\right]-R_P(F^*)$$
$$=\mathbb{E}_{D_n\sim P^n,X\sim P_X}\left[\int_{\mathbb{R}}\left|\hat{F}_{n,X}(z)-F_X^*(z)\right|^2\mathrm{d}z\right]\geq 0. \qquad (6)$$

This justifies the choice of the squared Cramér's distance in Eq. (1).

For large sample sizes, one expects that the predictor correctly estimates the conditional distribution and that the excess risk (6) tends to zero. A genuine question is to investigate the rate of convergence of the excess risk to zero as the sample size $n\to\infty$. The risk depends on the distribution of observations, and we want the model to perform well on large classes of distributions. Hence, we consider the standard minimax approach, as described in the next section.

2.3. Optimal minimax rates of convergence

In order to study the rate of convergence, as $n\to\infty$, of the excess risk (6) to zero, we introduce the notion of the *optimal minimax rate of convergence*. The minimax risk corresponds to the best achievable risk in the worst-case scenario (whence the name *minimax*). More precisely, given a class of distributions $\mathcal{D}$, the optimal minimax rate of convergence quantifies the minimal error that an estimator $\hat{F}_n$ can achieve uniformly on a given class of distributions $\mathcal{D}$, when the size of the training set D_n gets large.

Stone (1982) provided minimax rates of convergence within a point regression framework and the minimax theory for nonparametric regression is well-developed, see e.g. Györfi et al. (2002) or Tsybakov (2009). To the extent of our knowledge, this paper states the first results for distributional regression.

The formal definition of the minimax rate of convergence for distributional regression is as follows:

Definition 1. A sequence of positive numbers (a_n) is called an optimal minimax rate of convergence on class $\mathcal{D}$ if

$$\liminf_{n\to\infty}\inf_{\hat{F}_n}\sup_{P\in\mathcal{D}}\frac{\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)]-R_P(F^*)}{a_n}>0 \qquad (7)$$

and

$$\limsup_{n\to\infty}\inf_{\hat{F}_n}\sup_{P\in\mathcal{D}}\frac{\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)]-R_P(F^*)}{a_n}<\infty, \qquad (8)$$

where the infimum is taken over all distributional regression models $\hat{F}_n$ trained on D_n. If the sequence (a_n) satisfies only the lower bound (7), it is called a lower minimax rate of convergence.

2.4. k-NN and kernel predictors in distributional regression

Many predictors $\hat{F}_n$ can be studied and possibly achieve the optimal minimax rate of convergence. In this paper, we focus on two simple cases: k-nearest neighbor and kernel estimators.

The k-nearest neighbor (k-NN) method is well known in the classical framework of regression and classification (see, e.g. Biau & Devroye, 2015). In distributional regression, the k-NN method can be suitably adapted to estimate the conditional distribution F_x^*, and the estimator is written as

$$\hat{F}_{n,x}(z)=\frac{1}{k_n}\sum_{i=1}^{k_n}\mathbb{1}_{Y_{i:n}(x)\leq z}, \qquad (9)$$

where $1\leq k_n\leq n$, and $Y_{i:n}(x)$ denotes the observation at the ith nearest neighbor of x. As usual, possible ties are broken at random to define nearest neighbors. Note that in weather forecast statistical postprocessing, the k-NN method corresponds to a type of analog ensemble method (see Delle Monache et al., 2013).

The kernel estimate in distributional regression (see, e.g., Chapter 5 of Györfi et al., 2002) can be expressed as

$$\hat{F}_{n,x}(z)=\frac{\sum_{i=1}^n K(\frac{x-X_i}{h_n})\mathbb{1}_{Y_i\leq z}}{\sum_{i=1}^n K(\frac{x-X_i}{h_n})}, \qquad (10)$$

where function $K:\mathbb{R}^d\to[0,\infty)$ is a density function, called a kernel; and $h_n>0$ is the so-called bandwidth, which depends on the sample size n. If the denominator in (10) vanishes, we use the convention $\hat{F}_{n,x}(z)=\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{Y_i\leq z}$.

Minimax rates of convergence of the k-NN and kernel models in point regression are well studied, and it is known that, for suitable choices of the number of neighbors k_n and bandwidth h_n, respectively, the methods are minimax-rate optimal on classes of distributions with Lipschitz or, more generally, Hölder continuous regression functions (see e.g., Theorem 14.5 in Biau & Devroye, 2015 and Theorem 5.2 in Györfi et al. (2002)). For suitable classes of distributions defined hereafter, we are able to extend these results to distributional regression. Moreover, we obtain non-asymptotic bounds for the minimax rate of convergence for both the k-NN and kernel models (see Sections 3.2 and 3.3).

3. Main results

3.1. Optimal minimax rate of convergence

We consider the following classes of distributions:

Definition 2. For $h\in(0,1]$, $C>0$ and $M>0$, let $\mathcal{D}^{(h,C,M)}$ be the class of distributions P such that $F_x^*(y)=P(Y\leq y|X=x)$ satisfies

(i) $X\in[0,1]^d$ P_X-a.s.;
(ii) For all $x\in[0,1]^d$, $\int_{\mathbb{R}}F_x^*(z)(1-F_x^*(z))\mathrm{d}z\leq M$;
(iii) $\|F_{x'}^*-F_x^*\|_{L^2}\leq C\|x'-x\|^h$ for all $x,x'\in[0,1]^d$.

Conditions (i)–(iii) in Definition 2 are very similar to the conditions considered in the point regression framework; see Theorem 5.2 in Györfi et al. (2002). In condition (i), $[0,1]^d$ could be replaced by any compact set of $\mathbb{R}^d$. Condition (ii) requires that $\overline{\text{CRPS}}(F_x^*,F_x^*)$ remains uniformly bounded by M, which is a condition on the dispersion of the distribution F_X^*, since it implies that the mean absolute error (MAE) remains uniformly bounded. Condition (iii) is a regularity statement of the conditional distribution in the space $L^2(\mathbb{R})$. As an illustration, the different conditions are expressed for the generalized Pareto distribution model in Section 3.5 below.

Our main result is the following optimal minimax rate of convergence:

Theorem 1. *The sequence $a_n = n^{-\frac{2h}{2h+d}}$ is the optimal minimax rate of convergence on the class $\mathcal{D}^{(h,C,M)}$.*

It should be stressed that the rate of convergence $n^{-\frac{2h}{2h+d}}$ is the same as in point regression with square error, see Theorems 3.2 and 5.2 in Györfi et al. (2002) for the lower bound and upper bound, respectively.

Remark 1. As pointed out by a referee, conditions (i) and (iii) together with the integrability of Y imply condition (ii) for some $M > 0$. However, the dispersion, as measured by M, plays an important role throughout the proofs and, for this reason, we keep condition (ii) in order to obtain bounds as tight as possible.

The proof of Theorem 1 is divided into three steps:

1. We provide in Section 3.2 an explicit and non-asymptotic upper bound for the excess risk of the k-nearest neighbor model uniformly on class $\mathcal{D}^{(h,C,M)}$. The upper bound is then optimized with a suitable choice of $k = k_n$.
2. In Section 3.3, we obtain similar results for the kernel model.
3. We show in Section 3.4 that $a_n = n^{-\frac{2h}{2h+d}}$ is a lower minimax rate of convergence. The main argument is that it is enough to consider a binary model when both the observation Y and prediction $\hat{F}_X$ take values in $\{0, L\}$. We deduce that in this case, the CRPS coincides with the mean squared error so that we can appeal to standard results on a lower minimax rate of convergence for regression.

Combining these three steps, we finally obtain Theorem 1, providing the optimal minimax rate of convergence of the excess risk on the class $\mathcal{D}^{(h,C,M)}$. All the proofs are provided in the Appendix.

3.2. Upper bound for the k-nearest neighbor model

The k-NN method for distributional regression is defined in Eq. (9). Here we do not use only the mean of the nearest neighbor sample $(Y_{i:n}(x))_{1\leq i\leq k_n}$, but rather its entire empirical distribution. Interestingly, the tools developed to analyze the k-NN in point regression can be used in our distributional regression framework.

Proposition 1. *Assume $P \in \mathcal{D}^{(h,C,M)}$ and let $\hat{F}_n$ be the k-nearest neighbor model defined by Eq. (9). Then,*

$$\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq \begin{cases} 8^h C^2 \left(\frac{k_n}{n}\right)^h + \frac{M}{k_n} & \text{if } d = 1, \\ {c_d}^h C^2 \left(\frac{k_n}{n}\right)^{2h/d} + \frac{M}{k_n} & \text{if } d \geq 2, \end{cases}$$

where $c_d = \frac{2^{3+\frac{2}{d}}(1+\sqrt{d})^2}{V_d^{2/d}}$, and V_d is the volume of the unit ball in $\mathbb{R}^d$.

Let us stress that the upper bound is non-asymptotic and holds for all fixed n and k_n. Optimizing the upper bound in k_n yields the following corollary:

Corollary 1. *Assume $P \in \mathcal{D}^{(h,C,M)}$ and consider the k-NN model (9).*

- *For $d = 1$, the optimal choice $k_n = \left(\frac{M}{hC^2 8^h}\right)^{\frac{1}{h+1}} n^{\frac{h}{h+1}}$ yields*

  $$\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq Bn^{-\frac{h}{h+1}}$$

  with constant $B = C^{\frac{2}{h+1}} M^{\frac{h}{h+1}} 8^{\frac{h}{h+1}} \left(h^{-\frac{h}{h+1}} + h^{\frac{1}{h+1}}\right)$.
- *For $d \geq 2$, the optimal choice $k_n = \left(\frac{Md}{2hC^2 c_d^h}\right)^{\frac{d}{2h+d}} n^{\frac{2h}{2h+d}}$ yields*

  $$\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq Bn^{-\frac{2h}{2h+d}}$$

  with constant $B = (C^2 c_d^h)^{\frac{d}{2h+d}} M^{\frac{2h}{2h+d}} \left(\left(\frac{d}{2h}\right)^{\frac{2h}{2h+d}} + \left(\frac{2h}{d}\right)^{\frac{d}{2h+d}}\right)$.

3.3. Upper bound for the kernel model

Kernel methods adapted to distributional regression are defined in Eq. (10). For convenience and simplicity of notation, we develop our result for the simple uniform kernel $K(x) = \mathbb{1}_{\{\|x\|\leq 1\}}$. However, it should be stressed that all the results can be extended to boxed kernels (Györfi et al., 2002, Figure 5.7, p. 73) to the price of some extra multiplicative constants. For the uniform kernel, the estimator writes

$$\hat{F}_{n,x}(z) = \frac{\sum_{i=1}^n \mathbb{1}_{\{\|X_i - x\|\leq h_n\}} \mathbb{1}_{\{Y_i\leq z\}}}{\sum_{i=1}^n \mathbb{1}_{\{\|X_i - x\|\leq h_n\}}}, \tag{11}$$

when the denominator is non-zero, and $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Y_i\leq z\}}$ otherwise.

Proposition 2. *Assume $P \in \mathcal{D}^{(h,C,M)}$ and let $\hat{F}_n$ be the kernel model defined by Eq. (11). Then,*

$$\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq \tilde{c}_d \frac{2M + C^2 d^h + \frac{M}{n}}{nh_n^d} + C^2 h_n^{2h}$$

where $\tilde{c}_d$ only depends on d.

Once again, the upper bound is non-asymptotic and holds for all fixed n and h_n. Optimizing the upper bound in h_n yields the following corollary:

Corollary 2. *Assume $P \in \mathcal{D}^{(h,C,M)}$ and consider the kernel model (11). For any d, the optimal choice*

$$h_n = \left(\frac{\tilde{c}_d d(2M + C^2 d^h + \frac{M}{n})}{2hC^2}\right)^{\frac{1}{2h+d}} n^{-\frac{1}{2h+d}}$$

yields

$$\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq Bn^{-\frac{2h}{2h+d}}$$

with

$$B = C^{\frac{2d}{2h+d}} \left(\tilde{c}_d(2M + C^2 d^h + \frac{M}{n}) \right)^{\frac{2h}{2h+d}} \times \left(\left(\frac{d}{2h}\right)^{-\frac{d}{2h+d}} + \left(\frac{d}{2h}\right)^{\frac{2h}{2h+d}} \right).$$

3.4. Lower minimax rate of convergence

We finally compare the rates of convergence obtained in Corollaries 1 and 2 with a lower minimax rate of convergence in order to see whether the optimal rate of convergence is achieved.

To prove a lower bound on a class $\mathcal{D}$, it is always possible to consider a smaller class $\mathcal{B}$. Indeed, if $\mathcal{B} \subset \mathcal{D}$, we clearly have

$$\inf_{\hat{F}_n} \sup_{P\in\mathcal{B}} \left\{ \mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \right\} \leq \inf_{\hat{F}_n} \sup_{P\in\mathcal{D}} \left\{ \mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \right\}$$

so that any lower minimax rate of convergence on $\mathcal{B}$ is also a lower minimax rate of convergence on $\mathcal{D}$.

To establish the lower minimax rate of convergence, we focus on the following classes of binary responses:

Definition 3. Let $\mathcal{B}^{(h,C,L)}$ be the class of distributions of (X, Y), such that

(i) $Y \in \{0, L\}$ and X is uniformly distributed on $[0, 1]^d$;
(ii) $\|F^*_{x'} - F^*_x\|_{L^2} \leq C\|x' - x\|^h$ for all $x, x' \in [0, 1]^d$.

Since a binary outcome $Y \in \{0, L\}$ satisfies $\int_{\mathbb{R}} F^*_x(z)(1 - F^*_x(z))\mathrm{d}z \leq L/4$, condition (ii) in Definition 2 holds with $M \geq L/4$. Then $\mathcal{B}^{(h,C,L)} \subset \mathcal{D}^{(h,C,M)}$, and the following lower bound established on the smaller class also holds on the larger class.

Proposition 3. *The sequence* $a_n = n^{-\frac{2h}{2h+d}}$ *is a lower minimax rate of convergence on the class* $\mathcal{B}^{(h,C,L)}$. *More precisely,*

$$\liminf_{n\to\infty} \inf_{\hat{F}_n} \sup_{P\in\mathcal{B}^{(h,C,L)}} \frac{\mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*)}{C^{\frac{2d}{2h+d}} n^{-\frac{2h}{2h+d}}} \geq C_1 \qquad (12)$$

for some constant $C_1 > 0$ *independent of* C.

Combining Corollaries 1 and 2 and Proposition 3, we can deduce that for $d \geq 2$, the k-NN model reaches the lower minimax rate of convergence $a_n = n^{-\frac{2h}{2h+d}}$ for the class $\mathcal{D}^{(h,C,M)}$ and that the kernel model reaches the lower minimax rate of convergence a_n in any dimension d. This shows that this lower rate of convergence is in fact the optimal rate of convergence and proves Theorem 1.

3.5. Generalized pareto distributions

Explicit parametric formulas of the CRPS exist for most classical distribution families: e.g. Gaussian, logistic, censored logistic, generalized extreme value, and generalized Pareto (see Gneiting et al., 2005; Taillardat et al., 2016; Friederichs & Thorarinsdottir, 2012). We focus here on the generalized Pareto distribution (GPD) family and we denote by $H_{\xi,\sigma}$ the GPD with shape parameter $\xi \in \mathbb{R}$ and scale parameter $\sigma > 0$. Recall that it is defined, when $\xi \neq 0$, by

$$H_{\xi,\sigma}(z) = 1 - \left(1 + \frac{\xi z}{\sigma}\right)_+^{-1/\xi}, \quad z > 0,$$

with the notation $(\cdot)_+ = \max(0, \cdot)$. When $\xi = 0$, the standard limit by continuity is used. For $\xi < 1$, the GPD has a finite first moment, and the associated CRPS is given by the following (Friederichs & Thorarinsdottir, 2012):

$$\mathrm{CRPS}\left(H_{\xi,\sigma}, y\right) = \left(y + \frac{\sigma}{\xi}\right)\left(2H_{\xi,\sigma}(y) - 1\right) - \frac{2\sigma}{\xi(\xi-1)} \times \left(\frac{1}{\xi-2} + (1 - H_{\xi,\sigma}(y))\left(1 + \xi\frac{y}{\sigma}\right)\right). \qquad (13)$$

When $Y \sim H_{\xi^*,\sigma^*}$, the expected CRPS is as follows (Taillardat et al., 2022):

$$\overline{\mathrm{CRPS}}\left(H_{\xi,\sigma}, H_{\xi^*,\sigma^*}\right) = \frac{\sigma^*}{1-\xi^*} + \frac{2\sigma}{1-\xi}m_0 + \frac{2\xi}{1-\xi}m_1 + 2\sigma\left(\frac{1}{1-\xi} - \frac{1}{2(2-\xi)}\right) \qquad (14)$$

with

$$m_0 = \mathbb{E}_{Y\sim H_{\xi^*,\sigma^*}}\left[\left(1 + \frac{\xi}{\sigma}Y\right)^{-1/\xi}\right],$$

$$m_1 = \mathbb{E}_{Y\sim H_{\xi^*,\sigma^*}}\left[Y\left(1 + \frac{\xi}{\sigma}Y\right)^{-1/\xi}\right].$$

In particular,

$$\overline{\mathrm{CRPS}}\left(H_{\xi^*,\sigma^*}, H_{\xi^*,\sigma^*}\right) = \frac{\sigma^*}{(2-\xi^*)(1-\xi^*)}.$$

We now consider the distributional regression framework, and we illustrate the statement of Section 2.2 on Bayes risk in the case of a generalized Pareto regression model where Y given $X = x$ follows a GPD with shape parameter $\xi^*(x)$ and scale parameter $\sigma^*(x)$. Then it is possible to show that Bayes risk is equal to

$$R_P(F^*) = \int_{\mathbb{R}^d} \frac{\sigma^*(x)}{(2-\xi^*(x))(1-\xi^*(x))}P_X(\mathrm{d}x)$$

when $0 < \xi^*(x) < 1$ for all $x \in \mathbb{R}^d$. For a forecast in the GPD class, i.e. F_x is a GPD with shape parameter $\xi(x)$ and scale parameter $\sigma(x)$, then the risk $R_P(F)$ is equal to Bayes risk if and only if $\xi(x) = \xi^*(x)$ and $\sigma(x) = \sigma^*(x)$ P_X-a.e.

In the GPD regression framework, the conditions of the classes of distributions $\mathcal{D}^{(h,C,M)}$ can be interpreted as conditions on the parameters $\xi^*(x)$ and $\sigma^*(x)$. Condition (ii) is equivalent to $\sigma^*(x) \leq M(2-\xi^*(x))(1-\xi^*(x))$ when $0 < \xi^*(x) < 1$, for all $x \in [0, 1]^d$. The regularity condition (iii) holds with constants C and h as soon as $x \mapsto \xi^*(x)$ and $x \mapsto \sigma^*(x)$ are both h-Hölder.

For example, the popular case where the shape parameter $\xi^*(x)$ and the scale parameter $\sigma^*(x)$ are assumed to be linearly dependent on x (i.e. $\xi^*(x) = \xi_0 + \xi_1 \cdot x$ and $\sigma^*(x) = \sigma_0 + \sigma_1 \cdot x$ with $\xi_1, \sigma_1 \in \mathbb{R}^d$) is in a class of distributions of Definition 2.

4. Conclusion and discussion

We found that the optimal rate of convergence for distributional regression on $\mathcal{D}^{(h,C,M)}$ is of the same order as the optimal rate of convergence for point regression. Thus, with regard to the sample size n, distributional regression evaluated with the CRPS converges at the same rate as point regression even though the distributional estimate carries more information regarding the prediction of the underlying process.

We also showed that the k-NN method and the kernel method reach this optimal rate of convergence, respectively in dimension $d \geq 2$ and in any dimension. However, these methods are not widely used in practice because of the limitations of their predictive power in moderate or high dimension $d \geq 3$ due to the curse of dimensionality. An extension of this work could be to study whether state-of-the-art techniques reach the optimal rate of convergence obtained in our study. Random forests (Breiman, 2001) methods, such as quantile regression forests (Meinshausen, 2006) and distributional random forests (Ćevid et al., 2020), appear to be natural candidates, as they are based on a generalized notion of neighborhood and have been subject to recent development in weather forecast statistical postprocessing (see, e.g., Taillardat et al., 2016).

Our results were obtained for the CRPS, which is widely used in practice, but can easily be extended to the weighted CRPS in its standard uses. The weighted CRPS is defined as

$$\mathrm{wCRPS}(F,y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}_{y \leq z})^2 w(z)\mathrm{d}z$$

where w denotes the chosen weight. The weighted CRPS is used to put the focus of the score on specific regions of the outcome space (Gneiting & Ranjan, 2011). It is used in the study of extreme events by giving more weight to the extreme behavior of the distribution.

Moreover, an interesting development would be to obtain similar results for the rate of convergence with respect to different strictly proper scoring rules or metrics, for instance energy scores or Wasserstein distances.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project) and the Energy-oriented Centre of Excellence II (EoCoE-II), Grant Agreement 824158, funded within the Horizon2020 framework of the European Union. Part of this work was also supported by the ExtremesLearning grant from 80 PRIME CNRS-INSU and the ANR project Melody (ANR-19-CE46-0011).

Appendix A. Proof of Proposition 1

For simplicity of notation, we simply write $\mathbb{E}$ for the expectation with respect to $(X, Y) \sim P$ and $D_n \sim P^n$. The context makes it sufficiently clear so as to avoid confusion.

Proof. Recall that for the CRPS, the excess risk is equal to

$$\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) = \mathbb{E}\left[\int_{\mathbb{R}} |\hat{F}_{n,X}(z) - F^*_X(z)|^2 \mathrm{d}z\right]. \qquad \text{(A.1)}$$

We first estimate $\mathbb{E}[|\hat{F}_{n,x}(z) - F^*_x(z)|^2]$ for fixed $x \in [0,1]^d$ and $z \in \mathbb{R}$. Denote by $X_{1:n}(x), \ldots, X_{k_n:n}(x)$ the nearest neighbors of x, and by $Y_{1:n}(x), \ldots, Y_{k_n:n}(x)$ the associated values of the response variable. Conditionally on $X_{i:n}(x) = x_i$, $1 \leq i \leq k_n$, the random variables $Y_{i:n}(x)$, $1 \leq i \leq k_n$ are independent and with distribution $F^*_{x_i}$, $1 \leq i \leq k_n$. This implies that, conditionally, $\hat{F}_{n,x}(z)$ is the average of the k_n independent random variables $\mathbb{1}_{\{Y_{i:n}(x) \leq z\}}$ that have a Bernoulli distribution with parameter $F^*_{x_i}(z)$. Therefore, the conditional bias and variance are given by

$$\mathbb{E}[\hat{F}_{n,x}(z) - F^*_x(z) \mid X_i(x) = x_i, 1 \leq i \leq k_n]$$
$$= \frac{1}{k_n}\sum_{i=1}^{k_n} \left(F^*_{x_i}(z) - F^*_x(z)\right)$$
$$\mathrm{Var}[\hat{F}_{n,x}(z) \mid X_i(x) = x_i, 1 \leq i \leq k_n]$$
$$= \frac{1}{k_n^2}\sum_{i=1}^{k_n} F^*_{x_i}(z)(1 - F^*_{x_i}(z)).$$

Adding up the squared conditional bias and variance, and integrating with respect to $X_{i:n}(x)$, $1 \leq i \leq k_n$, we obtain the mean squared error:

$$\mathbb{E}\Big[|\hat{F}_{n,x}(z) - F^*_x(z)|^2\Big]$$
$$= \mathbb{E}\Big[\Big(\frac{1}{k_n}\sum_{i=1}^{k_n}\big(F^*_{X_{i:n}(x)}(z) - F^*_x(z)\big)\Big)^2\Big]$$
$$+ \frac{1}{k_n^2}\sum_{i=1}^{k_n} \mathbb{E}\Big[F^*_{X_{i:n}(x)}(z)(1 - F^*_{X_{i:n}(x)}(z))\Big].$$

Using Jensen's inequality and integrating with respect to $P_X(\mathrm{d}x)\mathrm{d}z$, we deduce that the excess risk (A.1) satisfies

$$\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*)$$
$$\leq \frac{1}{k_n}\sum_{i=1}^{k_n} \mathbb{E}\left[\int_{\mathbb{R}} (F^*_{X_{i:n}(X)}(z) - F^*_X(z))^2 \mathrm{d}z\right]$$
$$+ \frac{1}{k_n^2}\sum_{i=1}^{k_n} \mathbb{E}\left[\int_{\mathbb{R}} F^*_{X_{i:n}(X)}(z)(1 - F^*_{X_{i:n}(X)})\mathrm{d}z\right].$$

Using conditions (ii) and (iii) in the definition of class $\mathcal{D}^{(h,C,M)}$ to bound from above the first and second terms, respectively, we get

$$\begin{aligned}\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) &\le \frac{C^2}{k_n}\sum_{i=1}^{k_n}\mathbb{E}\Big[\|X_{i:n}(X)-X\|^{2h}\Big] + \frac{M}{k_n}\\ &\le C^2\mathbb{E}\Big[\|X_{k_n:n}(X)-X\|^{2h}\Big] + \frac{M}{k_n},\end{aligned}$$

where the last inequality uses the fact that, by definition of nearest neighbors, the distances $\|X_{i:n}(X)-X\|$, $1 \le i \le k_n$, are non-increasing.

The last step of the proof is to use Theorem 2.4 from Biau and Devroye (2015) stating that

$$\mathbb{E}[\|X_{k_n:n}(X)-X\|^2] \le \begin{cases} 8\frac{k_n}{n} & \text{if } d=1,\\ c_d\left(\frac{k_n}{n}\right)^{2/d} & \text{if } d \ge 2.\end{cases}$$

Together with the concavity inequality (as $h \in (0,1]$)

$$\mathbb{E}[\|X_{k_n:n}(X)-X\|^{2h}] \le \mathbb{E}[\|X_{k_n:n}(X)-X\|^2]^h,$$

we deduce

$$\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) \le \begin{cases} C^2 8^h\left(\frac{k_n}{n}\right)^h + \frac{M}{k_n} & \text{if } d=1,\\ C^2{c_d}^h\left(\frac{k_n}{n}\right)^{2h/d} + \frac{M}{k_n} & \text{if } d \ge 2,\end{cases}$$

concluding the proof of Proposition 1. □

Appendix B. Proof of Proposition 2

Proof. Eq. (11) can be rewritten as

$$\hat{F}_{n,x}(z) = \frac{\sum_{i=1}^n \mathbb{1}_{\{X_i\in S_{x,h_n}\}}\mathbb{1}_{\{Y_i\le z\}}}{nP_n(S_{x,h_n})},$$

with $S_{x,\epsilon}$ as the closed ball centered at x of radius $\epsilon > 0$ and

$$P_n(\cdot) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i\in\cdot\}}$$

as the empirical measure corresponding to $X_1,\dots,X_n$. Recall that we use the estimator $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Y_i\le z\}}$ when $nP_n(S_{x,h_n}) = 0$.

Similarly to the proof of Proposition 1, a bias/variance decomposition of the squared error yields

$$\begin{aligned}&\mathbb{E}\Big[|\hat{F}_{n,x}(z) - F_x^*(z)|^2\Big]\\ &= \mathbb{E}\left[\left(\frac{\sum_{i=1}^n\big(F^*_{X_i(x)}(z)-F_x^*(z)\big)\mathbb{1}_{\{X_i\in S_{x,h_n}\}}}{nP_n(S_{x,h_n})}\right)^2\mathbb{1}_{\{nP_n(S_{x,h_n})>0\}}\right]\\ &\quad+\mathbb{E}\left[\frac{\sum_{i=1}^n F^*_{X_i}(z)(1-F^*_{X_i}(z))\mathbb{1}_{\{X_i\in S_{x,h_n}\}}}{(nP_n(S_{x,h_n}))^2}\mathbb{1}_{\{nP_n(S_{x,h_n})>0\}}\right]\\ &\quad+\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{\{Y_i\le z\}} - F_x^*(z)\right)^2\mathbb{1}_{\{nP_n(S_{x,h_n})=0\}}\right]\\ &:= A_1(z)+A_2(z)+A_3(z).\end{aligned}$$

The excess risk at $X = x$ is thus decomposed into three terms

$$\begin{aligned}&\mathbb{E}\left[\int_{\mathbb{R}}|\hat{F}_{n,x}(z)-F_x^*(z)|^2\mathrm{d}z\right]\\ &\quad=\int_{\mathbb{R}}A_1(z)\mathrm{d}z+\int_{\mathbb{R}}A_2(z)\mathrm{d}z+\int_{\mathbb{R}}A_3(z)\mathrm{d}z\end{aligned}$$

that we analyze successively.

The first term (bias) is bounded from above using Jensen's inequality and property (iii) of $\mathcal{D}^{(h,C,M)}$:

$$\begin{aligned}\int_{\mathbb{R}}A_1(z)\mathrm{d}z &\le \mathbb{E}\left[\frac{\sum_{i=1}^n\int_{\mathbb{R}}\big(F^*_{X_i(x)}(z)-F_x^*(z)\big)^2\mathrm{d}z\mathbb{1}_{\{X_i\in S_{x,h_n}\}}}{nP_n(S_{x,h_n})}\right.\\ &\qquad\left.\times\mathbb{1}_{\{nP_n(S_{x,h_n})>0\}}\right]\\ &\le \mathbb{E}\left[\frac{\sum_{i=1}^n C^2\|X_i-x\|^{2h}\mathbb{1}_{\{X_i\in S_{x,h_n}\}}}{nP_n(S_{x,h_n})}\right.\\ &\qquad\left.\times\mathbb{1}_{\{nP_n(S_{x,h_n})>0\}}\right]\\ &\le C^2{h_n}^{2h}.\end{aligned}$$

The second term (variance) is bounded using property (ii) of $\mathcal{D}^{(h,C,M)}$ and an elementary result for the binomial distribution:

$$\begin{aligned}\int_{\mathbb{R}}A_2(z)\mathrm{d}z &= \mathbb{E}\left[\frac{\sum_{i=1}^n\int_{\mathbb{R}}F^*_{X_i}(z)(1-F^*_{X_i}(z))\mathrm{d}z\mathbb{1}_{\{X_i\in S_{x,h_n}\}}}{(nP_n(S_{x,h_n}))^2}\right.\\ &\qquad\left.\times\mathbb{1}_{\{nP_n(S_{x,h_n})>0\}}\right]\\ &\le M\mathbb{E}\left[\frac{\mathbb{1}_{\{nP_n(S_{x,h_n})>0\}}}{nP_n(S_{x,h_n})}\right]\\ &\le \frac{2M}{nP_X(S_{x,h_n})}.\end{aligned}$$

In the last line, we use that $Z = nP_n(S_{x,h_n})$ follows a binomial distribution with parameters n and $p = P_X(S_{x,h_n})$ so that $\mathbb{E}\left[\frac{1}{Z}\mathbb{1}_{\{Z>0\}}\right] \le \frac{2}{(n+1)p}$; see Lemma 4.1 in Györfi et al. (2002).

The last term is a remainder term and is bounded by

$$\begin{aligned}\int_{\mathbb{R}}A_3(z)\mathrm{d}z &\le \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\int_{\mathbb{R}}\big(F^*_{X_i}(z)-F^*_x(z)\big)^2\mathrm{d}z\mathbb{1}_{\{nP_n(S_{x,h_n})=0\}}\right]\\ &\quad+\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n\int_{\mathbb{R}}F^*_{X_i}(z)(1-F^*_{X_i}(z))\mathrm{d}z\mathbb{1}_{\{nP_n(S_{x,h_n})=0\}}\right].\end{aligned}$$

Properties (ii) and (iii) of $\mathcal{D}^{(h,C,M)}$ and the fact that $\|X_i - x\| \le \sqrt{d}$ imply

$$\begin{aligned}\int_{\mathbb{R}}A_3(z)\mathrm{d}z &\le \left(C^2d^h+\frac{M}{n}\right)\mathbb{E}\left[\mathbb{1}_{\{nP_n(S_{x,h_n})=0\}}\right]\\ &\le \left(C^2d^h+\frac{M}{n}\right)e^{-nP_X(S_{x,h_n})}.\end{aligned}$$

For the second inequality, we use that $\mathbb{P}(Z=0) = (1-p)^n \le e^{-np}$, where $Z = nP_n(S_{x,h_n})$ follows a binomial distribution with parameters n and $p = P_X(S_{x,h_n})$.

Collecting the three terms, we obtain the following upper bound for the excess risk at $X = x$:

$$\mathbb{E}\left[\int_{\mathbb{R}} |\hat{F}_{n,x}(z) - F_x^*(z)|^2 \mathrm{d}z\right] \\ \leq C^2 {h_n}^{2h} + \frac{2M}{nP_X(S_{x,h_n})} + \left(C^2 d^h + \frac{M}{n}\right) e^{-nP_X(S_{x,h_n})}.$$

We finally integrate this bound with respect to $P_X(\mathrm{d}x)$. According to Equation (5.1) in Györfi et al. (2002), there exists a constant $\tilde{c}_d$ depending only on d such that

$$\int_{[0,1]^d} \frac{1}{nP_X(S_{x,h_n})} P_X(\mathrm{d}x) \leq \frac{\tilde{c}_d}{nh_n^d}.$$

Note that $\tilde{c}_d$ can be chosen as $\tilde{c}_d = d^{d/2}$. We also have

$$\int_{[0,1]^d} e^{-nP_X(S_{x,h_n})} P_X(\mathrm{d}x) \leq \max_{u\geq 0} ue^{-u} \int_{[0,1]^d} \frac{1}{nP_X(S_{x,h_n})} \\ \times P_X(\mathrm{d}x) \\ \leq \frac{\tilde{c}_d}{nh_n^d}.$$

We thus obtain

$$\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) = \mathbb{E}\left[\int_{\mathbb{R}} |\hat{F}_{n,x}(z) - F_x^*(z)|^2 \mathrm{d}z\right] \\ \leq C^2 {h_n}^{2h} + \tilde{c}_d \frac{2M + C^2 d^h + \frac{M}{n}}{n{h_n}^d}. \quad \square$$

Appendix C. Proof of Proposition 3

The proof of Proposition 3 relies on the next two elementary lemmas. The first one states that for a binary outcome $Y \in \{0, L\}$, forecasters should focus only on a binary forecast $F \in \mathcal{M}(\{0, L\})$, which is very natural. More precisely, any predictive distribution $F \in \mathcal{M}(\mathbb{R})$ can be associated with $F \in \mathcal{M}(\{0, L\})$ with a better expected CRPS.

Lemma 1. *Let $G \in \mathcal{M}(\{0, L\})$. For $F \in \mathcal{M}(\mathbb{R})$, the distribution*

$$\tilde{F}(z) = (1 - m)\mathbb{1}_{0\leq z} + m\mathbb{1}_{L\leq z} \text{ with } m = \frac{1}{L}\int_0^L (1 - F(z))\mathrm{d}z$$

satisfies

$$\overline{\mathrm{CRPS}}(\tilde{F}, G) \leq \overline{\mathrm{CRPS}}(F, G).$$

Proof. Let $F \in \mathcal{M}(\mathbb{R})$ and $G \in \mathcal{M}(\{0, L\})$. We have

$$\overline{\mathrm{CRPS}}(F, G) = \int_{\mathbb{R}}\int_{\mathbb{R}} (F(z) - \mathbb{1}_{y\leq z})^2 \mathrm{d}z G(\mathrm{d}y) \\ \geq \int_{\mathbb{R}}\int_0^L (F(z) - \mathbb{1}_{y\leq z})^2 \mathrm{d}z G(\mathrm{d}y)$$

Because $1 - m$ is the mean value of F on $[0, L]$, we have for $y \in \{0, L\}$

$$\int_0^L (F(z) - \mathbb{1}_{y\leq z})^2 \mathrm{d}z \geq \int_0^L ((1 - m) - \mathbb{1}_{y\leq z})^2 \mathrm{d}z.$$

Integrating with respect to $G(\mathrm{d}y)$, we deduce

$$\overline{\mathrm{CRPS}}(F, G) \geq \int_{\mathbb{R}}\int_0^L ((1 - m) - \mathbb{1}_{y\leq z})^2 \mathrm{d}z G(\mathrm{d}y).$$

The right-hand side equals $\overline{\mathrm{CRPS}}(\tilde{F}, G)$ and we conclude $\overline{\mathrm{CRPS}}(F, G) \geq \overline{\mathrm{CRPS}}(\tilde{F}, G)$. □

Lemma 2 shows that for binary outcomes and predictions, the CRPS reduces to a quantity proportional to the Brier score: Brier, 1950

$$\mathrm{Brier}(p, y) = (y - p)^2, \quad y \in \{0, 1\}, p \in [0, 1],$$

which is closely related to the mean squared error used in regression.

Lemma 2. *For all $y \in \{0, L\}$ and $F(z) = (1 - p)\mathbb{1}_{0\leq z} + p\mathbb{1}_{L\leq z} \in \mathcal{M}(\{0, L\})$ with $p \in [0, 1]$, it holds that*

$$\mathrm{CRPS}(F, y) = L\mathrm{Brier}(p, \frac{y}{L}) = L(\frac{y}{L} - p)^2.$$

Proof. We compute

$$\mathrm{CRPS}(F, y) = \int_0^L (1 - p - \mathbb{1}_{y\leq z})^2 \mathrm{d}z \\ = \begin{cases} Lp^2 & \text{if y=0} \\ L(1 - p)^2 & \text{if y=L} \end{cases}.$$

In both cases, this equals $L(\frac{y}{L} - p)^2 = L\mathrm{Brier}(p, \frac{y}{L})$. □

Proof of Proposition 3. Since only binary outcomes are considered in the class $\mathcal{B}^{(h,C,L)}$, Lemma 1 implies that

$$\inf_{\hat{F}_n} \sup_{P\in\mathcal{B}^{(h,C,L)}} \left\{\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*)\right\} \\ = \inf_{\tilde{F}_n} \sup_{P\in\mathcal{B}^{(h,C,L)}} \left\{\mathbb{E}[R_P(\tilde{F}_n)] - R_P(F^*)\right\}$$

where the infima are taken over models $\hat{F}_n$ and $\tilde{F}_n$ trained on the first observations $(X_i, Y_i)_{1\leq i\leq n}$ and with values in $\mathcal{M}(\mathbb{R})$ and $\mathcal{M}(\{0, L\})$, respectively. Indeed, the left-hand side is a priori smaller, since the family $\hat{F}_n$ is larger, but Lemma 1 ensures that each model $\hat{F}_n$ can be associated with a model $\tilde{F}_n$ with equal or lower expected score.

We then apply Lemma 2. For a binary outcome, the conditional distribution of Y given $X = x$ writes

$$F_x^*(z) = (1 - m(x))\mathbb{1}_{0\leq z} + m(x)\mathbb{1}_{L\leq z},$$

and the model $\tilde{F}_n$ with values in $\mathcal{M}(\{0, L\})$ takes the form

$$\tilde{F}_{n,x}(z) = (1 - m_n(x))\mathbb{1}_{0\leq z} + m_n(x)\mathbb{1}_{L\leq z},$$

with $m(x) = \frac{1}{L}\int_0^L (1 - F_x^*(z))\mathrm{d}z$ and $m_n(x) = \frac{1}{L}\int_0^L (1 - \hat{F}_{n,x}(z))\mathrm{d}z$.

Then Lemma 2 implies

$$\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) \\ = \mathbb{E}\left[\mathrm{CRPS}(\hat{F}_{n,X}, Y) - \mathrm{CRPS}(F_X^*, Y)\right] \\ = L\mathbb{E}\left[(Y/L - m_n(X))^2 - (Y/L - m(X))^2\right] \\ = L\mathbb{E}\left[(m_n(X) - m(X))^2\right],$$

which corresponds to the excess risk in regression with squared error loss. Property (iii) of $\mathcal{B}^{(h,C,L)}$ is equivalent to

$$|m(x) - m(x')|^h \leq C\|x - x'\|^h, \quad x \in [0, 1]^d,$$

which is the standard regularity assumption on the regression function m. Using the result of Problem 3.3 in Györfi et al. (2002) dealing with binary models, we finally obtain that the sequence $a_n = n^{-\frac{2h}{2h+d}}$ is a lower minimax rate of convergence for this class of distributions, and more precisely that Eq. (12) holds. □

References

Baran, S., & Lerch, S. (2018). Combining predictive distributions for the statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, *34*(3), 477–496. http://dx.doi.org/10.1016/j.ijforecast.2018.01.005.

Biau, G., & Devroye, L. (2015). *Springer Series in the Data Sciences, Lectures on the nearest neighbor method*. Springer Cham.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. http://dx.doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

Ćevid, D., Michel, L., Näf, J., Meinshausen, N., & Bühlmann, P. (2020). Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. http://dx.doi.org/10.48550/ARXIV.2005.14458, arXiv https://arxiv.org/abs/2005.14458.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., & Searight, K. (2013). Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, *141*(10), 3498–3516. http://dx.doi.org/10.1175/MWR-D-12-00281.1.

Friederichs, P., & Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, *23*(7), 579–594. http://dx.doi.org/10.1002/env.2176.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*(1), 125–151. http://dx.doi.org/10.1146/annurev-statistics-062713-085831.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. http://dx.doi.org/10.1198/016214506000001437.

Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*(5), 1098–1118. http://dx.doi.org/10.1175/mwr2904.1.

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, *29*, http://dx.doi.org/10.1198/jbes.2010.08110.

Györfi, L., Kohler, M., Krzyżak, A., & Walk, H. (2002). *Springer Series in Statistics, A Distribution-Free Theory of Nonparametric Regression*. Springer New York, http://dx.doi.org/10.1007/b97848.

Hamill, T. M., & Colucci, S. J. (1997). Verification of eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, *125*(6), 1312–1327. http://dx.doi.org/10.1175/1520-0493(1997)125<1312:voersr>2.0.co;2.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Springer Series in Statistics, The Elements of Statistical Learning*. Springer New York, http://dx.doi.org/10.1007/978-0-387-84858-7.

Henzi, A., Ziegel, J. F., & Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *83*(5), 963–993. http://dx.doi.org/10.1111/rssb.12450.

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*(10), 1087–1096. http://dx.doi.org/10.1287/mnsc.22.10.1087.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*(35), 983–999.

Mösching, A., & Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, *14*(1), 24–49. http://dx.doi.org/10.1214/19-ejs1659.

Naveau, P., Huser, R., Ribereau, P., & Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, *52*(4), 2753–2769. http://dx.doi.org/10.1002/2015wr018552.

Rasp, S., & Lerch, S. (2018). Neural networks for post-processing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. http://dx.doi.org/10.1175/mwr-d-18-0187.1.

Schulz, B., & Lerch, S. (2022). Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review*, *150*(1), 235–257. http://dx.doi.org/10.1175/MWR-D-21-0150.1.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, *10*(4), 1040–1053. http://dx.doi.org/10.1214/aos/1176345969.

Taillardat, M., Fougères, A.-L., Naveau, P., & de Fondeville, R. (2022). Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2022.07.003.

Taillardat, M., Fougères, A.-L., Naveau, P., & Mestre, O. (2019). Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Journal Weather and Forecasting*, *34*(3), 617–634. http://dx.doi.org/10.1175/waf-d-18-0149.1.

Taillardat, M., Mestre, O., Zamo, M., & Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, *144*(6), 2375–2393. http://dx.doi.org/10.1175/mwr-d-15-0260.1.

Thorey, J., Mallet, V., & Baudin, P. (2017). Online learning with the continuous ranked probability score for ensemble forecasting. *Quarterly Journal of the Royal Meteorological Society*, *143*(702), 521–529. http://dx.doi.org/10.1002/qj.2940.

Tsybakov, A. B. (2009). *Springer Series in Statistics, Introduction to Nonparametric Estimation*. Springer New York, NY, http://dx.doi.org/10.1007/b13794.