



HAL
open science

Mathematical Properties of Continuous Ranked Probability Score Forecasting

Romain Pic, Clément Dombry, Philippe Naveau, Maxime Taillardat

► **To cite this version:**

Romain Pic, Clément Dombry, Philippe Naveau, Maxime Taillardat. Mathematical Properties of Continuous Ranked Probability Score Forecasting. 2022. hal-03662994v1

HAL Id: hal-03662994

<https://hal.science/hal-03662994v1>

Preprint submitted on 9 May 2022 (v1), last revised 24 Oct 2023 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mathematical Properties of Continuous Ranked Probability Score Forecasting

Romain Pic^{a,*}, Clément Dombry^a, Philippe Naveau^b, Maxime Taillardat^{c,d}

^a*Univ. Bourgogne Franche-Comté, CNRS UMR 6623, Laboratoire de Mathématiques de Besançon, 25000 Besançon, France*

^b*Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, CEA/CNRS/UVSQ, EstimR, IPSL & U Paris-Saclay, Gif-sur-Yvette, France*

^c*CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France*

^d*Météo-France, Toulouse, France*

Abstract

The theoretical advances on 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, theoretical properties of such minimization of the CRPS have mostly considered the unconditional framework (i.e. without covariables) and infinite sample sizes. We circumvent these limitations and study the rate of convergence in terms of CRPS of distributional regression methods. We find the optimal minimax rate of convergence for a given class of distributions. Moreover, we show that the k nearest neighbor method and the kernel method for the distributional regression reach the optimal rate of convergence in dimension $d \geq 2$ and in any dimension, respectively.

Keywords: Distributional Regression, Probabilistic Forecast, CRPS, Minimax Rate of Convergence, Nearest Neighbor Method, Kernel Method.

1. Introduction

In meteorology, 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 & Colucci, 1997; Baran & Lerch, 2018) and need to be statistically postprocessed

*Corresponding author

Email address: `romain.pic@univ-fcomte.fr` (Romain Pic)

8 in order to be improved. Different postprocessing methods have been proposed,
 9 such as Ensemble Model Output Statistics (Gneiting et al., 2005), Quantile Re-
 10 gression Forests (Taillardat et al., 2019) or Neural Networks (Schulz & Lerch,
 11 2021). These references, among other, also discuss the stakes of weather forecast
 12 statistical postprocessing.

13
 14 Postprocessing methods rely on probabilistic forecast and distributional re-
 15 gression (Gneiting & Katzfuss, 2014) where the aim is to predict the conditional
 16 distribution of the quantity of interest (e.g. temperatures, wind-speed, or pre-
 17 cipitation) given a set of covariates (e.g. ensemble model output statistics).
 18 Algorithms are often based on the minimization of a proper scoring rule that
 19 compares actual observations with the predictive distribution. Scoring rules
 20 can be seen as an equivalent of a loss function in classical regression. A detailed
 21 review of scoring rules is given by Gneiting & Raftery (2007). The Continu-
 22 ous Ranked Probability Score (CRPS; Matheson & Winkler, 1976), defined in
 23 Equation (4), is one of the most popular score in meteorological forecasts. The
 24 CRPS is also minimized to infer parameters of statistical models used in post-
 25 processing (e.g. Gneiting et al., 2005; Naveau et al., 2016; Rasp & Lerch, 2018;
 26 Taillardat et al., 2019).

27
 28 To the best of our knowledge, most convergence statements about the CRPS
 29 are not only derived within an unconditional framework, i.e. without taking into
 30 account the covariates, but also these limiting results are based on infinite sam-
 31 ple sizes. In this work, our goal is to bypass these two limitations. To go further,
 32 we need to set the stage by including a few notations.

33
 In this article, we consider the regression framework $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ with
 distribution P . In forecast assessment, we make the distinction between the
 construction of the estimator relying of the training sample $D_n = \{(X_i, Y_i), 1 \leq$
 $i \leq n\}$ and its evaluation with respect to new data (X, Y) . Statistically, the goal
 of distributional regression is to estimate the conditional distribution of Y given
 $X = x$, noted $F_x^*(y) := P(Y \leq y | X = x)$. 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^* , $x \in \mathbb{R}^d$. In this context, it is crucial to assess if $\hat{F}_{n,x}$ is close to
 F_x^* over the range of possible values of $X = x$. We denote by P_X the marginal
 distribution of X . The main goal of this work is to study the following positive
 quantity :

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

34 where $\mathbb{E}_{X \sim P_X, D_n \sim P^n}$ denotes the expectation with respect to X and D_n follow-
 35 ing P_X and P^n respectively. This averaged L^2 -norm is the distance between the
 36 predicted distribution $\hat{F}_{n,x}$ and the true conditional distribution F_x^* . We focus
 37 on this specific distance because it corresponds to the excess of risk associated
 38 with the CRPS, i.e. it is the difference between the expected CRPS for the
 39 predicted distribution $\hat{F}_{n,x}$ and the expected CRPS for the ideal prediction F_x^* ,

40 see Section 2.1 for more details.

41
42 In order to study the rate of convergence of (1) as $n \rightarrow \infty$, we will adapt
43 the notion of *optimal minimax rate of convergence* that quantifies the best error
44 that an estimator can achieve uniformly on a given family of distributions \mathcal{D}
45 when the size of the training set D_n gets large. Stone (1982) provided minimax
46 rates of convergence within a point regression framework and the minimax theory
47 for nonparametric regression is well-developed, see e.g. Györfi et al. (2002)
48 or Tsybakov (2009). To the extent of our knowledge, this paper states the first
49 results for distributional regression.

50
51 Many predictors $\hat{F}_{n,x}$ can be studied and achieve the optimal minimax rate
52 of convergence. To go further, we focus on two cases : k -nearest neighbor and
53 kernel estimators.

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

$$\hat{F}_{n,x}(z) = \frac{1}{k_n} \sum_{i=1}^{k_n} \mathbb{1}_{Y_{i:n}(x) \leq z}, \quad (2)$$

where $1 \leq k_n \leq n$ and $Y_{i:n}(x)$ denotes the observation at the i -th nearest
neighbor of x . As usual, possible ties are broken at random to define nearest
neighbors.

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

$$\hat{F}_{n,x}(z) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) \mathbb{1}_{Y_i \leq z}}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}, \quad (3)$$

54 if the denominator is nonzero. When the denominator is zero, we use the con-
55 vention $\hat{F}_{n,x}(z) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{Y_i \leq z}$. Here the bandwidth $h_n > 0$ depends on the
56 sample size n , and the function $K : \mathbb{R}^d \rightarrow [0, \infty)$ is called the kernel.

57
58 Minimax rates of convergence of the k -NN and kernel models in point re-
59 gression are well-studied and it is known that, for suitable choices of number
60 of neighbors k_n and bandwidth h_n respectively, the methods are minimax rate
61 optimal on classes of distributions with Lipschitz or more generally Hölder con-
62 tinuous regression functions (see e.g. Theorem 14.5 in Biau & Devroye, 2015
63 and Theorem 5.2 in Györfi et al., 2002). For classes of distributions where the
64 conditional distribution F_x^* satisfies some regularity requirements with respect
65 to the covariates x (see Definition 1 of class $\mathcal{D}^{(h,C,M)}$), we are able to extend
66 these results to distributional regression. We obtain non asymptotic bounds for
67 the minimax rate of convergence for both the k -NN and kernel models.

68

69 To summarize, this paper is organized as follows. Our main results are
70 presented in Section 2. Section 2.1 provides the theoretical background on
71 distributional regression and its evaluation using the CRPS. In Section 2.2, we
72 study the k -NN estimators (2) and derive a non asymptotic upper bound for the
73 excess of risk (1) uniformly on the class $\mathcal{D}^{(h,C,M)}$. Section 2.3 provides similar
74 results for the kernel method (3). In Section 2.4, we find a lower minimax rate
75 of convergence by reducing the problem to standard point regression solved by
76 Györfi et al. (2002). We can deduce that the k -NN method for the distributional
77 regression reaches the optimal rate of convergence in dimension $d \geq 2$, while the
78 kernel method reaches the optimal rate of convergence in any dimension. The
79 proofs of all the results presented in Section 2 are detailed in Appendix.

80 2. Main results

81 2.1. CRPS and 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

$$\text{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}_{y \leq z})^2 dz. \quad (4)$$

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

$$\overline{\text{CRPS}}(F, G) = \int_{\mathbb{R}} \text{CRPS}(F, y) G(dy), \quad F, G \in \mathcal{M}(\mathbb{R}) \quad (5)$$

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 moment of order 1. Then, the difference between the expected CRPS of the forecast F and the expected CRPS of the ideal forecast G can be written as

$$\overline{\text{CRPS}}(F, G) - \overline{\text{CRPS}}(G, G) = \int_{\mathbb{R}} |F(z) - G(z)|^2 dz \geq 0. \quad (6)$$

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

Example 1. (CRPS for Generalized Pareto distributions)

Explicit parametric formulas of the CRPS exist for most classical distribution families : e.g. Gaussian, logistic, censored logistic, Generalized Extreme Value, 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 GP distribution 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 (Friederichs & Thorarinsdottir, 2012)

$$\text{CRPS}(H_{\xi,\sigma}, y) = \left(y + \frac{\sigma}{\xi}\right) (2H_{\xi,\sigma}(y) - 1) - \frac{2\sigma}{\xi(\xi - 1)} \left(\frac{1}{\xi - 2} + (1 - H_{\xi,\sigma}(y)) \left(1 + \xi \frac{y}{\sigma}\right)\right). \quad (7)$$

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

$$\overline{\text{CRPS}}(H_{\xi,\sigma}, H_{\xi^*,\sigma^*}) = \frac{\sigma^*}{1 - \xi^*} + \frac{2\sigma}{1 - \xi} m_0 + \frac{2\xi}{1 - \xi} m_1 + 2\sigma^* \left(\frac{1}{1 - \xi} - \frac{1}{2(2 - \xi)}\right) \quad (8)$$

with

$$m_0 = \mathbb{E}_{Y \sim H_{\xi,\sigma}} \left[\left(1 + \frac{\xi Y}{\sigma}\right)^{-1/\xi} \right], \quad m_1 = \mathbb{E}_{Y \sim H_{\xi,\sigma}} \left[Y \left(1 + \frac{\xi Y}{\sigma}\right)^{-1/\xi} \right].$$

In particular,

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

In distributional regression, a predictor $\hat{F} : x \mapsto \hat{F}_x$ is evaluated thanks to its expected risk

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

This quantity is important as many distributional regression methods try to minimize it in order to improve predictions. When Y is integrable, Equation (6) implies

$$\begin{aligned} R_P(\hat{F}) - R_P(F^*) &= \mathbb{E}_{(X,Y) \sim P} \left[\text{CRPS}(\hat{F}_X, Y) - \text{CRPS}(F_X^*, Y) \right] \\ &= \mathbb{E}_{X \sim P_X} \left[\int_{\mathbb{R}} \left| \hat{F}_X(z) - F_X^*(z) \right|^2 dz \right] \geq 0. \quad (9) \end{aligned}$$

87 We recall that Bayes risk is the minimal theoretical risk over all possible pre-
88 dictors and that Bayes predictor is a predictor achieving Bayes risk. Thus,
89 Equation (9) implies that $R_P(F^*)$ is Bayes risk and that $R_P(\hat{F}) = R_P(F^*)$ if
90 and only if $\hat{F}_x = F_x^*$ P_X -a.e. An introduction to the notions of theoretical risk,
91 Bayes risk and excess of risk can be found in Section 2.4 of Hastie et al. (2001).

Example 2. (GPD regression)

We illustrate the above statement 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

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

92 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
 93 GPD with shape parameter $\xi(x)$ and scale parameter $\sigma(x)$, then the risk $R_P(F)$
 94 is equal to Bayes risk if and only if $\xi(x) = \xi^*(x)$ and $\sigma(x) = \sigma^*(x)$ P_X -a.e.

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 the introduction, 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

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

with expectation taken both with respect to the training sample D_n and test observation (X, Y) . Once again, when Y is integrable, the expected risk has a unique minimum given by $R_P(F^*)$. The *excess of risk* becomes

$$\begin{aligned} & \mathbb{E}_{D_n \sim P^n} [R_P(\hat{F}_n)] - R_P(F^*) \\ &= \mathbb{E}_{D_n \sim P^n, X \sim P_X} \left[\int_{\mathbb{R}} |\hat{F}_{n, X}(z) - F_X^*(z)|^2 dz \right] \geq 0. \end{aligned} \quad (10)$$

95 For large sample sizes, one expects that the predictor correctly estimates
 96 the conditional distribution and that the excess of risk tends to zero. A genuine
 97 question is to investigate the rate of convergence of the excess of risk to zero as
 98 the sample size $n \rightarrow \infty$. The risk depends on the distribution of observations
 99 and we want the model to perform well on large classes of distributions. Hence,
 100 we consider the standard minimax approach, see for instance Györfi et al. (2002)
 101 for the standard cases of regression and classification.

102

103 We consider the following classes of distributions.

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

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

109 **Remark 1.** In condition *i*), $[0, 1]^d$ could be replaced by any compact set of \mathbb{R}^d .
 110 Condition *ii*) requires that $\overline{\text{CRPS}}(F_x^*, F_x^*)$ remains uniformly bounded by M ,
 111 which is a condition on the dispersion of the distribution F_X^* since it implies
 112 that the absolute mean error (MAE) remains uniformly bounded. Condition
 113 *iii*) is a regularity statement of the conditional distribution in the space $L^2(\mathbb{R})$.
 114 Conditions *i*) – *iii*) in Definition 1 are very similar to the conditions considered
 115 in the point regression framework, see Theorem 5.2 in Györfi et al. (2002).

116 **Example 3.** In the GPD regression framework, condition *ii*) is equivalent to
 117 $\sigma^*(x) \leq M(2 - \xi^*(x))(1 - \xi^*(x))$ when $0 < \xi^*(x) < 1$, for all $x \in [0, 1]^d$. The
 118 regularity condition *iii*) holds with constants C and h as soon as $x \mapsto \xi^*(x)$
 119 and $x \mapsto \sigma^*(x)$ are both h -Hölder. For example, the popular case were the
 120 shape parameter $\xi^*(x)$ and the scale parameter $\sigma^*(x)$ are assumed to be linearly
 121 dependent on x (i.e. $\xi^*(x) = \xi_0 + \alpha \cdot x$ and $\sigma^*(x) = \sigma_0 + \beta \cdot x$ with $\alpha, \beta \in \mathbb{R}^d$)
 122 is in a class of distributions of Definition 1.

123 In the following, we study the convergence rate of the excess of risk in order
 124 to obtain the optimal minimax convergence rate. The reasoning is divided into
 125 three steps:

- 126 1. We provide in Section 2.2 an explicit and nonasymptotic upper bound for
 127 the excess of risk of the k -nearest neighbor model uniformly on the class
 128 $\mathcal{D}^{(h,C,M)}$; the upper bound is then optimized with a suitable choice of
 129 $k = k(n)$.
- 130 2. In Section 2.3, we obtain similar results for the kernel model.
- 131 3. We show in Section 2.4 that $a_n = n^{-\frac{2h}{2h+d}}$ is a lower minimax rate of
 132 convergence; the main argument is that it is enough to consider a binary
 133 model when both the observation Y and prediction \hat{F}_X take values in
 134 $\{0, L\}$; we deduce that in this case, the CRPS coincides with the mean
 135 squared error so that we can appeal to standard results on lower minimax
 136 rate of convergence for regression.

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

140 2.2. Upper bound for the k -nearest neighbor model

141 The k -NN method for distributional regression is defined in Equation (2).
 142 Here, we do not use only the mean of the nearest neighbor sample $(Y_{i:n}(x))_{1 \leq i \leq k_n}$
 143 but its entire empirical distribution. Interestingly, the tools developed to ana-
 144 lyze the k -NN in point regression can be used in our distributional regression
 145 framework.

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

model defined by Equation (2). Then,

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

146 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 .

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

149 **Corollary 1.** Assume $P \in \mathcal{D}^{(h,C,M)}$ and consider the k -NN model (2).

- 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

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

150 with constant $B = C^{\frac{2}{h+1}} M^{\frac{h}{h+1}} \left(8^{\frac{h}{h+1}} h^{-\frac{h}{h+1}} + (8^h h C)^{\frac{1}{h+1}}\right)$.

- 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

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

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

153 2.3. Upper bound for the kernel model

Kernel methods adapted to distributional regression are defined in Equation (3). For convenience and simplicity of notations, 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 p73) to the price of some extra multiplicative constants. For the uniform kernel, the estimator writes

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

154 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 Equation (11). Then,

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

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

156 Once again, the upper bound is non-asymptotic and holds for all fixed n and
157 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

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

yields

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

with

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

158 2.4. Optimal minimax rates of convergence

159 We finally compare the rates of convergence obtained in Corollaries 1 and 2
160 with a lower minimax rate of convergence in order to see whether the opti-
161 mal rate of convergence are achieved. We first recall these different notions of
162 minimax rates of convergence.

Definition 2. A sequence of positive numbers (a_n) is called an optimal minimax rate of convergence on the class \mathcal{D} if

$$\liminf_{n \rightarrow \infty} \inf_{\hat{F}_n} \sup_{P \in \mathcal{D}} \frac{\mathbb{E}_{D_n \sim P^n} [R_P(\hat{F}_n)] - R_P(F^*)}{a_n} > 0 \quad (12)$$

and

$$\limsup_{n \rightarrow \infty} \inf_{\hat{F}_n} \sup_{P \in \mathcal{D}} \frac{\mathbb{E}_{D_n \sim P^n} [R_P(\hat{F}_n)] - R_P(F^*)}{a_n} < \infty, \quad (13)$$

163 where the infimum is taken over all distributional regression models \hat{F}_n trained
164 on D_n . If the sequence (a_n) satisfies only the lower bound (12), it is called a
165 lower minimax rate of convergence.

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 $\int_{\mathbb{R}} F_x^*(z)(1 - F_x^*(z)) dz \leq L/4$, condition ii) in Definition 1 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,*

$$\liminf_{n \rightarrow \infty} \inf_{\hat{F}_n} \sup_{P \in \mathcal{B}^{(h,C,L)}} \frac{\mathbb{E}_{D_n \sim P^n} [R_P(\hat{F}_n)] - R_P(F^*)}{C^{\frac{2d}{2h+d}} n^{-\frac{2h}{2h+d}}} \geq C_1 \quad (14)$$

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 minimax lower 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 minimax lower rate of convergence in any dimension d . This shows that the lower rate of convergence is in fact the optimal rate of convergence and proves the following theorem.

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.

3. 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

194 evaluated with the CRPS converges at the same rate as point regression even
195 though the distributional estimate carries more information on the prediction
196 of the underlying process.

197 We have also shown that the k -NN method and the kernel method reach
198 this optimal rate of convergence, respectively in dimension $d \geq 2$ and in any
199 dimension. However, these methods are not used in practice because of the
200 limitations of their predictive power in moderate or high dimension $d \geq 3$ due
201 to the curse of dimension. An extension of this work could be to study if state-of-
202 the-art techniques reach the optimal rate of convergence obtained in this article.
203 Random Forests (Breiman, 2001) methods, such as Quantile Regression Forests
204 (Meinshausen, 2006) and Distributional Random Forests (Cévid et al., 2020),
205 appear to be natural candidates as they are based on a generalized notion of
206 neighborhood and have been subject to recent development in weather forecast
207 statistical postprocessing (see, e.g., Taillardat et al., 2016).

The results of this article 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

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

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

212 Moreover, an interesting development would be to obtain similar results for
213 rate of convergence with respect to different strictly proper scoring rules or met-
214 rics, for instance energy scores or Wasserstein distances.

215
216 *Acknowledgments:* The authors acknowledge the support of the French Agence
217 Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-
218 REX project) and of the Energy oriented Centre of Excellence-II (EoCoE-II),
219 Grant Agreement 824158, funded within the Horizon2020 framework of the Eu-
220 ropean Union. Part of this work was also supported by the ExtremesLearning
221 grant from 80 PRIME CNRS-INSU and the ANR project Melody (ANR-19-
222 CE46-0011).

223 References

- 224 Baran, S., & Lerch, S. (2018). Combining predictive distributions for the sta-
225 tistical post-processing of ensemble forecasts. *International Journal of Fore-*
226 *casting*, *34*. doi:10.1016/j.ijforecast.2018.01.005.
- 227 Biau, G., & Devroye, L. (2015). *Lectures on the Nearest Neighbor Method*.
228 Springer Series in the Data Sciences. Springer.
- 229 Breiman, L. (2001). Random forests. *Machine Learning*, *45*. doi:10.1023/a:
230 1010933404324.

- 231 Brier, G. W. (1950). Verification of forecasts expressed in terms of probabil-
 232 ity. *Monthly Weather Review*, 78. doi:10.1175/1520-0493(1950)078<0001:
 233 VOFEIT>2.0.CO;2.
- 234 Čevič, D., Michel, L., Näf, J., Meinshausen, N., & Bühlmann, P. (2020). Distri-
 235 butional random forests: Heterogeneity adjustment and multivariate distri-
 236 butional regression. <https://arxiv.org/abs/2005.14458>. doi:10.48550/
 237 ARXIV.2005.14458.
- 238 Friederichs, P., & Thorarinsdottir, T. L. (2012). Forecast verification for extreme
 239 value distributions with an application to probabilistic peak wind prediction.
 240 *Environmetrics*, 23, 579–594. doi:10.1002/env.2176.
- 241 Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *An-
 242 nual Review of Statistics and its Applications*, . doi:10.1146/
 243 annurev-statistics-062713-085831.
- 244 Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction,
 245 and estimation. *Journal of the American Statistical Association*, 102. doi:10.
 246 1198/016214506000001437.
- 247 Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. (2005). Cal-
 248 ibrated Probabilistic Forecasting Using Ensemble Model Output Statistics
 249 and Minimum CRPS Estimation. *Monthly Weather Review*, 133, 1098 –
 250 1118. doi:10.1175/MWR2904.1.
- 251 Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold-
 252 and quantile-weighted scoring rules. *Journal of Business and Economic Statis-
 253 tics*, 29. doi:10.1198/jbes.2010.08110.
- 254 Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2002). *A Distribution-Free
 255 Theory of Nonparametric Regression*. Springer Series in Statistics. Springer.
- 256 Hamill, T. M., & Colucci, S. J. (1997). Verification of eta-rsm short-
 257 range ensemble forecasts. *Monthly Weather Review*, 125. doi:10.1175/
 258 1520-0493(1997)125<1312:VOERSR>2.0.CO;2.
- 259 Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical
 260 Learning*. Springer Series in Statistics. Springer New York Inc.
- 261 Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous proba-
 262 bility distributions. *Management Science*, 22. doi:10.2307/2629907.
- 263 Meinshausen, N. (2006). Quantile regression forests. *The Journal of Machine
 264 Learning Research*, 7, 983–999.
- 265 Naveau, P., Huser, R., Ribereau, P., & Hannart, A. (2016). Modeling jointly
 266 low, moderate, and heavy rainfall intensities without a threshold selection.
 267 *Water Resources Research*, 52. doi:10.1002/2015wr018552.

- 268 Rasp, S., & Lerch, S. (2018). Neural networks for post-processing en-
 269 semble weather forecasts. *Monthly Weather Review*, *146*. doi:10.1175/
 270 MWR-D-18-0187.1.
- 271 Schulz, B., & Lerch, S. (2021). Machine learning methods for postprocessing en-
 272 semble forecasts of wind gusts: A systematic comparison. [arXiv:2106.09512](https://arxiv.org/abs/2106.09512)
 273 [arXiv:2106.09512](https://arxiv.org/abs/2106.09512).
- 274 Stone, C. J. (1982). Optimal global rates of convergence for nonparametric
 275 regression. *The Annals of Statistics vol. 10 iss. 4, 10*. doi:10.1214/aos/
 276 1176345969.
- 277 Taillardat, M., Fougères, A.-L., Naveau, P., & de Fondeville, R. (2022).
 278 Extreme events evaluation using CRPS distributions. [https://hal.](https://hal.archives-ouvertes.fr/hal-02121796v3)
 279 [archives-ouvertes.fr/hal-02121796v3](https://hal.archives-ouvertes.fr/hal-02121796v3).
- 280 Taillardat, M., Fougères, A.-L., Naveau, P., & Mestre, O. (2019). Forest-based
 281 and semiparametric methods for the postprocessing of rainfall ensemble fore-
 282 casting. *Weather and Forecasting*, *34*. doi:10.1175/WAF-D-18-0149.1.
- 283 Taillardat, M., Mestre, O., Zamo, M., & Naveau, P. (2016). Calibrated en-
 284 semble forecasts using quantile regression forests and ensemble model output
 285 statistics. *Monthly Weather Review*, *144*. doi:10.1175/MWR-D-15-0260.1.
- 286 Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer
 287 Series in Statistics. Springer, New York. doi:10.1007/b13794.

288 Appendix A. Proof of Proposition 1

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

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

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

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 $X_{1:n}(x), \dots, X_{k_n:n}(x)$ the nearest neighbors of x and by $Y_{1:n}(x), \dots, Y_{k_n:n}(x)$ 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

$$\begin{aligned}\mathbb{E}[\hat{F}_{n,x}(z) - F_x^*(z) \mid X_i(x) = x_i, 1 \leq i \leq k_n] &= \frac{1}{k_n} \sum_{i=1}^{k_n} (F_{x_i}^*(z) - F_x^*(z)) \\ \text{Var}[\hat{F}_{n,x}(z) \mid X_i(x) = x_i, 1 \leq i \leq k_n] &= \frac{1}{k_n^2} \sum_{i=1}^{k_n} F_{x_i}^*(z)(1 - F_{x_i}^*(z)).\end{aligned}$$

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

$$\begin{aligned}&\mathbb{E}[|\hat{F}_{n,x}(z) - F_x^*(z)|^2] \\ &= \mathbb{E}\left[\left(\frac{1}{k_n} \sum_{i=1}^{k_n} (F_{X_{i:n}(x)}^*(z) - F_x^*(z))\right)^2\right] + \frac{1}{k_n^2} \sum_{i=1}^{k_n} \mathbb{E}\left[F_{X_{i:n}(x)}^*(z)(1 - F_{X_{i:n}(x)}^*(z))\right].\end{aligned}$$

Using Jensen's inequality and integrating with respect to $P_X(dx)dz$, we deduce that the excess of risk (A.1) satisfies

$$\begin{aligned}\mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) &\leq \frac{1}{k_n} \sum_{i=1}^{k_n} \mathbb{E}\left[\int_{\mathbb{R}} (F_{X_{i:n}(X)}^*(z) - F_X^*(z))^2 dz\right] \\ &\quad + \frac{1}{k_n^2} \sum_{i=1}^{k_n} \mathbb{E}\left[\int_{\mathbb{R}} F_{X_{i:n}(X)}^*(z)(1 - F_{X_{i:n}(X)}^*(z)) dz\right].\end{aligned}$$

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

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

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

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

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

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

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

we deduce

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

294 concluding the proof of Proposition 1. □

295 Appendix B. Proof of Proposition 2

Proof. Equation (11) can be rewritten as

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

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

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

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

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

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

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

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

298 that we analyze successively.

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

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

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

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

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

The last term is a remainder term and is bounded by

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

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

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

302 For the second inequality, we use that $\mathbb{P}(Z = 0) = (1 - p)^n \leq e^{-np}$ where
 303 $Z = nP_n(S_{x,h_n})$ follows a binomial distribution with parameters n and $p =$
 304 $P_X(S_{x,h_n})$.

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

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

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

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

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

$$\begin{aligned} \int_{[0,1]^d} e^{-nP_X(S_{x,h_n})} P_X(dx) &\leq \max_{u \geq 0} u e^{-u} \int_{[0,1]^d} \frac{1}{nP_X(S_{x,h_n})} P_X(dx) \\ &\leq \frac{\tilde{c}_d}{nh_n^d}. \end{aligned}$$

We obtain thus

$$\begin{aligned} \mathbb{E}[R_P(\hat{F}_n)] - R_P(F^*) &= \mathbb{E} \left[\int_{\mathbb{R}} |\hat{F}_{n,x}(z) - F_x^*(z)|^2 dz \right] \\ &\leq C^2 h_n^{2h} + \tilde{c}_d \frac{2M + Cd^{h/2} + \frac{M}{n}}{nh_n^d}. \end{aligned}$$

305

□

306 Appendix C. Proof of Proposition 3

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

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

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

satisfies

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

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

$$\begin{aligned} \overline{\text{CRPS}}(F, G) &= \int_{\mathbb{R}} \int_{\mathbb{R}} (F(z) - \mathbb{1}_{y \leq z})^2 dz G(dy) \\ &\geq \int_{\mathbb{R}} \int_0^L (F(z) - \mathbb{1}_{y \leq z})^2 dz G(dy) \end{aligned}$$

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

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

Integrating with respect to $G(dy)$, we deduce

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

The right hand side equals $\overline{\text{CRPS}}(\tilde{F}, G)$ and we conclude

$$\overline{\text{CRPS}}(F, G) \geq \overline{\text{CRPS}}(\tilde{F}, G).$$

312

□

Lemma 2 shows that for binary outcome and predictions, the CRPS reduces to a quantity proportional to the Brier score (Brier, 1950)

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

313 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

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

Proof. We compute

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

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

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

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

315 where the infimum are taken over models \hat{F}_n and \tilde{F}_n trained on the first ob-
316 servations $(X_i, Y_i)_{1 \leq i \leq n}$ and with values in $\mathcal{M}(\mathbb{R})$ and $\mathcal{M}(\{0, L\})$, respectively.

317 Indeed, the left hand side is a priori smaller since the family \hat{F}_n is larger but
318 Lemma 1 ensures that each model \hat{F}_n can be associated with a model \tilde{F}_n with
319 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 $m(x) = \frac{1}{L} \int_0^L (1 - F_x^*(z)) dz$ and $m_n(x) = \frac{1}{L} \int_0^L (1 - \hat{F}_{n,x}(z)) dz$.
Then Lemma 2 implies

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

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

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

320 which is the standard regularity assumption on the regression function m . Using
321 the result of the Problem 3.3 in Györfi et al. (2002) dealing with binary models,
322 we finally obtain that the sequence $a_n = n^{-\frac{2h}{2h+d}}$ is a lower minimax rate of
323 convergence for this class of distributions and more precisely that Equation (14)
324 holds. \square