ROBUST NON-PARAMETRIC REGRESSION VIA MEDIAN-OF-MEANS

ANNA BEN-HAMOU AND ARNAUD GUYADER

Abstract. In this paper, we apply the median-of-means principle to derive robust versions of local averaging rules in non-parametric regression. For various estimates, including nearest neighbors and kernel procedures, we obtain non-asymptotic exponential inequalities, with only a second moment assumption on the noise. We then show that these bounds cannot be significantly improved by establishing a corresponding lower bound on the tail probabilities.

Index Terms: Non-parametric regression, Median of Means, Sub-Gaussian estimators.
Mathematics Subject Classification (2020): 62G08, 62G15, 62G35.

Contents

1. Introduction

1.1. **Setting and main results.** Let (X, Y) be a pair of random variables, where X has distribution μ on $\mathbb{R}^d$, for $d \geq 1$, and Y is real-valued and satisfies $\mathbb{E}[Y^2] < \infty$. The regression function is defined for μ-almost every $x \in \mathbb{R}^d$ as

$$r(x) := \mathbb{E}[Y \mid X = x] .$$

The pair (X, Y) can be written as

$$Y = r(X) + \varepsilon ,$$

where the random variable ε, called the noise, satisfies $\mathbb{E}[\varepsilon \mid X] = 0$. Note that

$$\mathbb{E}\left[(Y - r(X))^2\right] = \inf_g \mathbb{E}\left[(Y - g(X))^2\right] ,$$

where the infimum is taken over all measurable functions $g : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}[g(X)^2] < \infty$. In other words, the regression function is an optimal approximation of Y by a square-integrable function of X, with respect to the L_2 risk.

When the distribution of the pair (X, Y) is unknown, one cannot predict Y using $r(X)$. However, assuming that one has access to an i.i.d. sample

$$\mathcal{D}_n := ((X_1, Y_1), \ldots, (X_n, Y_n))$$

with the same distribution as (X, Y), then one can use the data $\mathcal{D}_n$ in order to construct an estimate of the function r. In this respect, a local averaging estimate of the regression function is an estimate that can be written as

$$\forall x \in \mathbb{R}^d\,,\ \widehat{r}_n(x) := \widehat{r}_n(x, \mathcal{D}_n) = \sum_{i=1}^n W_i(x) Y_i\,,$$

where for all $i \in [\![1, n]\!]$, $W_i(x)$ is a Borel measurable function of x and $X_1, \ldots, X_n$ (but not of $Y_1, \ldots, Y_n$), with values in $[0, 1]$, and such that $\sum_{i=1}^n W_i(x) = 1$. This class includes nearest neighbors, kernel, and partitioning estimates.

The goal of this paper is to design robust versions of $\widehat{r}_n(x)$, which exhibit good concentration properties even if the noise ε does not have exponential moments. More in detail, we only assume that ε has a finite second moment. To do so, we use the *median-of-means* (MoM) technique: for $m \in [\![1, n]\!]$, we consider m disjoint subsets $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(m)}$ of $\mathcal{D}_n$, each of length $N = \lfloor n/m \rfloor$ (if n is not a multiple of m, we simply discard some observations). For each $j \in [\![1, m]\!]$ and all $x \in \mathbb{R}^d$, let

$$\widehat{r}^{(j)}(x) := \widehat{r}_N(x, \mathcal{D}^{(j)})\,,$$

for some estimate $\widehat{r}_N$, called the *base estimate*. Note that, for a given $x \in \mathbb{R}^d$, the variables $\widehat{r}^{(1)}(x), \ldots, \widehat{r}^{(m)}(x)$ are i.i.d., with the same distribution as $\widehat{r}_N(x) := \widehat{r}_N(x, \mathcal{D}_N)$. The median-of-means regression estimate is then defined as

$$\widehat{r}_n^{\mathsf{mom}}(x) := \mathsf{median}\left(\widehat{r}^{(1)}(x), \ldots, \widehat{r}^{(m)}(x)\right) ,$$

where $\mathsf{median}(r_1, \ldots, r_m) = r_{(\lceil m/2 \rceil)}$ corresponds to the smallest value $r \in \{r_1, \ldots, r_m\}$ such that

$$|\{j \in [\![1, m]\!]\,,\, r_j \leq r\}| \geq \frac{m}{2} \quad \text{and} \quad |\{j \in [\![1, m]\!]\,,\, r_j \geq r\}| \geq \frac{m}{2}\,.$$

Throughout the article, $\mathbb{R}^d$ is equipped with the Euclidean distance, and for $x \in \mathbb{R}^d$ and $\varepsilon > 0$, $\mathcal{B}(x, \varepsilon)$ denotes the Euclidean closed ball of center x and radius ε. In all what follows, we will be interested in the following model (see Section 1.2 for comments on this set of hypotheses).

Assumption 1. *The class $\mathcal{F} = \mathcal{F}_{\rho,\sigma}$, with $\rho, \sigma > 0$, is the class of distributions (X, Y) satisfying:*

(i) The support S of μ is bounded with diameter $D > 0$ and for all $x \in S$ and $\varepsilon \in (0, D]$, we have

$$\mu\left(\mathcal{B}(x, \varepsilon)\right) \geq \rho \varepsilon^d \,. \tag{1}$$

(ii) For all $x \in S$, we have $\mathrm{Var}(\varepsilon \,|\, X = x) \leq \sigma^2$.

(iii) The function r is Lipschitz with constant 1.

For a variety of base estimates, including nearest neighbors and kernel estimates, we show that, when $(X, Y) \in \mathcal{F}$, and when σ and ρ are known, the median-of-means estimate satisfies the following concentration inequality: for all $\delta \in [e^{-nc_{\mathcal{F}}+1}, 1[$, and for all $x \in S$, when the number of blocks m is chosen as $m = \lceil \ln(1/\delta) \rceil$, we have

$$\mathbb{P}\left(|\widehat{r}_n^{\,\mathsf{mom}}(x) - r(x)| \geq a \left(\frac{\sigma^2 \lceil \ln(1/\delta) \rceil}{\rho n} \right)^{\frac{1}{d+2}} \right) \leq \delta \,, \tag{2}$$

where $a > 0$ is an explicit numerical constant, possibly depending on d, and where $c_{\mathcal{F}} > 0$ is a constant that depends on ρ, σ and d only. Roughly speaking, this means that, in a large domain, the tail of $|\widehat{r}_n^{\,\mathsf{mom}}(x) - r(x)|$ is upper-bounded by that of $Z^{\frac{2}{d+2}}$, where $Z \sim \mathcal{N}(0, \frac{\sigma^2}{\rho n})$.

In fact, if for each $x \in S$, one has access to a local ρ_x such that (1) is satisfied, then (2) is fulfilled with ρ_x instead of ρ. Nevertheless, since (2) is valid for all $x \in S$, it implies that if $X \sim \mu$ independent of $\mathcal{D}_n$, then

$$\mathbb{P}\left(|\widehat{r}_n^{\,\mathsf{mom}}(X) - r(X)| \geq a \left(\frac{\sigma^2 \lceil \ln(1/\delta) \rceil}{\rho n} \right)^{\frac{1}{d+2}} \right) \leq \delta \,. \tag{3}$$

Furthermore, we show that this bound is optimal in the following sense: for all $d \geq 1$, there exists $\rho > 0$ such that for all $n \geq 1$ and all $\sigma^2 > 0$, for all $\delta \in]0, 2^{-(d+3)}]$, for any regression estimate $\widehat{r}_n$, there exists a distribution in $\mathcal{F} = \mathcal{F}_{\rho,\sigma}$ such that, when $X \sim \mu$ is independent of $\mathcal{D}_n$, we have

$$\mathbb{P}\left(|\widehat{r}_n(X) - r(X)| \geq b \left(\frac{\sigma^2 \ln\left(\frac{1}{2^{d+3}\delta}\right)}{\rho n} \right)^{\frac{1}{d+2}} \right) \geq \delta \,, \tag{4}$$

for some explicit numerical constant $b > 0$ depending only on the dimension d.

In addition to exhibiting concentration properties, the estimate $\widehat{r}_n^{\,\mathsf{mom}}$ also stands out through its strong robustness to outliers: as shown in Section 5, as soon as the number of outliers is less than $m/2$, the upper bound (2) still holds, modulo a slight modification of the constants a and $c_{\mathcal{F}}$.

1.2. **Related work.** It seems that the median-of-means principle was first introduced in works of Jerrum et al. [22], Alon et al. [1], in order to obtain sub-Gaussian estimators for the mean of a heavy-tailed random variable, or when outliers may contaminate the data (see also Catoni [9] for a different but related approach). Some variants that do not require any knowledge on the variance have also been proposed recently, see for example Lee and Valiant [28], Minsker and Ndaoud [38], or Gobet et al. [15].

One caveat of the MoM-estimator of the mean is its dependence on the confidence threshold δ. However, under stronger assumptions on the distribution, Minsker [37] showed that it is in fact adaptive to δ up to $\delta \approx e^{-\sqrt{n}}$. In the same vein, Devroye et al. [14] proposed a way to design δ-independent sub-Gaussian estimators up to $\delta \approx e^{-n}$.

The MoM principle was also generalized to multivariate settings by Minsker [36], Hsu and Sabato [19], Lerasle and Oliveira [30], Lugosi and Mendelson [35], and applied to a large variety of statistical problems, including linear regression (Audibert and Catoni [2]), empirical risk minimization (Brownlees et al. [7], Lecué and Lerasle [25], Lugosi and Mendelson [34]), classification (Lecué et al. [27]), bandits (Bubeck et al. [8]), least-squares density estimation (Lerasle and Oliveira [30]), and kernel density estimation (Humbert et al. [20]). We refer the reader to Lecué and Lerasle [26], Lerasle [29], or Lugosi and Mendelson [33] for more references on all of these subjects.

To our knowledge, there are only very few concentration results in non-parametric regression, at the exception of Jiang [23] for the k-nn estimate. However, in the latter, the noise is assumed sub-Gaussian. Hence, the first contribution of the present paper is to derive estimators satisfying (2), with only a second moment assumption on the noise. Our second contribution is to show that this bound on the tail probability cannot be significantly improved, by constructing scenarios where it is tight in the sense of (4).

Let us mention that, except for inequality (1), all points of Assumption 1 (*i.e.* bounded support, bounded variance and Lipschitz property) are standard to obtain L_2 rates of convergence in non-parametric regression estimation, see for example Chapters 4, 5, and 6 in Györfi et al. [17] for, respectively, partitioning, kernel, and nearest neighbors estimates. Concerning (1), the proof of Theorem 1 in Penrose and Yukich [39] (see in particular (13.1)) ensures that it is satisfied if S is a finite union of convex bounded sets and X has a density f that is bounded away from zero. Such an assumption is also made by Jiang [23]. Comparable assumptions are the so-called cone-condition in Korostelev and Tsybakov [24], Chapter 5, and the notion of standard support in Cuevas and Fraiman [13]. As will become clear in the remainder of the article, equation (1) allows us to obtain inequality (2) for all $x \in S$, and not only in average for X with law μ (see also Remark 1).

Finally, note that the variance factor in inequality (2), given by $\sigma^{\frac{4}{d+2}}(\rho n)^{-\frac{2}{d+2}}$, is the minimax rate of convergence for the L_2 risk $\mathbb{E}[(\widehat{r}_n(X) - r(X))^2]$ in this model. Usually, this minimax rate is written with respect to D^{-d} instead of ρ, where D is the diameter of S. Those two quantities are clearly related by the inequality $\rho \le D^{-d}$, which comes from (1) applied to $\varepsilon = D$. Still, one may check that lower bounds for the L_2 risk are actually obtained in situations where equation (1) is satisfied, namely with $X \sim \mu = \mathrm{Unif}([0,1]^d)$, see Stone [43] and Györfi et al. [17], Chapter 3. Thus, roughly speaking, we recover this minimax rate jointly with an exponential upper-bound for the tail of the error $|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)|$. This minimax rate is obtained by optimizing on the tuning parameter of the base estimate (*e.g.*, the number k of neighbors for k-nn estimates, the bandwidth h for kernel estimates), and this optimization step typically requires the knowledge of ρ and σ. In this respect, an open question would be to design procedures to choose this tuning parameter in an adaptive way. For the L_2 risk, this is possible for instance by splitting the sample or cross-validation, as explained for example in Györfi et al. [17], Chapters 7 and 8. Unfortunately, this issue seems to be more complicated in the present case of concentration bounds.

1.3. **Organization of the paper.** Section 2 contains two key lemmas and provides a guideline for proving bounds like (2). In the next two sections, inequality (2) is established for various choices of local averaging procedures, namely nearest neighbor methods in Section 3, and kernel and partitioning methods in Section 4. Let us point out that, for partitioning estimates (Section 4.2), we are able to obtain a uniform control on $\sup_{x\in S} |\widehat{r}_n^{\mathsf{mom}}(x) - r(x)|$. In Section 5, we show that the estimator $\widehat{r}_n^{\mathsf{mom}}$ achieves robustness in a very generic contamination scheme. The lower bound (4) is established in Section 6. In Section 7, we mention how $\widehat{r}_n^{\mathsf{mom}}$ may be turned into a δ-independent estimator, using a strategy introduced by Devroye et al. [14]. Finally, Section 8 gathers the proofs of several technical results.

2. Preliminary results

This section exposes two generic results that will be of constant use throughout the paper. The first lemma relates deviation probabilities for the median-of-means estimate $\widehat{r}_n^{\mathsf{mom}}$ with deviation probabilities of the base estimate $\widehat{r}_N$. We point out that this result is valid for any base estimate $\widehat{r}_N$, not only for local averaging rules.

Lemma 1. *Let $\widehat{r}_n^{\mathsf{mom}}$ be the median-of-means estimate of r constructed on m blocks with base estimate $\widehat{r}_N$. For all $x \in \mathbb{R}^d$ and all $t \geq 0$, we have*

$$\mathbb{P}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \geq t\right) \leq 2^m p_t(x)^{m/2}\,,$$

where

$$p_t(x) := \mathbb{P}\left(|\widehat{r}_N(x) - r(x)| \geq t\right)\,.$$

Let us now turn to the specific context of this article, meaning that the base estimate takes the form

$$\forall x \in \mathbb{R}^d\,,\ \widehat{r}_N(x) = \widehat{r}_N(x, \mathcal{D}_N) = \sum_{i=1}^N W_i(x) Y_i\,, \tag{5}$$

where for all $i \in [\![1, N]\!]$, $W_i(x)$ is a Borel measurable function of x and $X_1, \ldots, X_N$ (but not of $Y_1, \ldots, Y_N$), with values in $[0, 1]$, and such that $\sum_{i=1}^N W_i(x) = 1$. Our second lemma gives a bias–variance decomposition for deviation probabilities of local averaging estimates.

Lemma 2. *Suppose that (ii) and (iii) in Assumption 1 are satisfied. Then, for all $x \in \mathbb{R}^d$, we have*

$$|\widehat{r}_N(x) - r(x)| \leq \left|\sum_{i=1}^N W_i(x)\varepsilon_i\right| + \sum_{i=1}^N W_i(x)\|X_i - x\|\,, \tag{6}$$

where $\varepsilon_i = Y_i - r(X_i)$. In addition, for all $s, t > 0$, we have

$$p_{t+s}(x) \leq \frac{\sigma^2 \mathbb{E}\left[\sum_{i=1}^N W_i(x)^2\right]}{t^2} + \mathbb{P}\left(\sum_{i=1}^N W_i(x)\|X_i - x\| \geq s\right)\,, \tag{7}$$

with $p_{t+s}(x)$ as defined in Lemma 1.

In the next two sections, we investigate several instances of local averaging procedures for the base estimate $\widehat{r}_N$. In each case, we first use Lemma 2 in order to determine t and s such that $p_{t+s}(x) \leq \frac{1}{4e^2}$. Lemma 1 then entails that

$$\mathbb{P}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \geq t + s\right) \leq e^{-m}\,.$$

The number of blocks m can then be chosen as $\lceil \ln(1/\delta) \rceil$, for some target confidence threshold $\delta \in [e^{-n}, 1[$, so that the probability above is less than δ. Next, provided σ and ρ are known, a tuning parameter of the base estimate (*e.g.*, the number k of neighbors for k-nn estimates, the bandwidth h for kernel estimates) can then be optimized to get a bound in the flavor of (2), imposing additional constraints on δ.

3. Nearest neighbors estimation

For $x \in \mathbb{R}^d$, let

$$\left(X_{(1)}(x), Y_{(1)}(x)\right), \ldots, \left(X_{(N)}(x), Y_{(N)}(x)\right)$$

a reordering of the data $\mathcal{D}_N$ according to increasing values of $\|X_i - x\|$, that is

$$\|X_{(1)}(x) - x\| \leq \cdots \leq \|X_{(N)}(x) - x\|\,,$$

where, if necessary, distance ties are broken by simulating auxiliary random variables $(U_1, \dots, U_N)$ i.i.d. with uniform law on $[0,1]$ and sorting them. The weighted nearest neighbors estimate is defined as

$$\widehat{r}_N(x) := \sum_{i=1}^{N} v_i Y_{(i)}(x)\,, \tag{8}$$

where $(v_1, \dots, v_N)$ is a deterministic vector in $[0,1]^N$ satisfying $\sum_{i=1}^{N} v_i = 1$. Note that this estimate is of the form (5), with $W_i(x) = v_{\sigma(i)}$, where σ is a random permutation (depending on x) such that $X_i = X_{(\sigma(i))}(x)$. We refer the interested reader to Biau and Devroye [4], Chapter 8, or Samworth [42] and references therein for more information on this topic.

In this context, the variance term of Lemma 2 reduces to

$$\mathbb{E}\left[\sum_{i=1}^{N} W_i(x)^2\right] = \sum_{i=1}^{N} v_{\sigma(i)}^2 = \sum_{i=1}^{N} v_i^2\,. \tag{9}$$

As for the bias term, letting $D_{(i)}(x) := \|X_{(i)}(x) - x\|$, we rewrite it as

$$\sum_{i=1}^{N} W_i(x)\|X_i - x\| = \sum_{i=1}^{N} v_i D_{(i)}(x)\,. \tag{10}$$

The following lemma, whose proof is housed in Section 8, allows us to control the expected nearest neighbor distances (see Remark 1 below for a comment on this result).

Lemma 3. *Under Assumption 1(i), for all $x \in S$ and $i \in [\![1, N]\!]$, one has*

$$\mathbb{E}\left[D_{(i)}(x)\right] \leq 2\left(\frac{i}{\rho(N+1)}\right)^{1/d}.$$

According to (10), we then deduce from Markov's inequality that, for all $x \in S$,

$$\mathbb{P}\left(\sum_{i=1}^{N} W_i(x)\|X_i - x\| \geq s\right) \leq \frac{2}{s}\sum_{i=1}^{N} v_i \left(\frac{i}{\rho(N+1)}\right)^{1/d}. \tag{11}$$

Combining (9) and (11), and applying Lemma 2 with

$$t = 2e\sigma\sqrt{2\sum_{i=1}^{N} v_i^2} \qquad \text{and} \qquad s = 16e^2 \sum_{i=1}^{N} v_i \left(\frac{i}{\rho(N+1)}\right)^{1/d}, \tag{12}$$

we see that, for all $x \in S$,

$$p_{t+s}(x) \leq \frac{1}{4e^2}\,,$$

which entails, by Lemma 1, that

$$\mathbb{P}\left(|\widehat{r}_n^{\,\mathsf{mom}}(x) - r(x)| \geq t + s\right) \leq e^{-m}\,. \tag{13}$$

We first propose to illustrate this result on two specific examples of nearest neighbors rules: the uniform k-nearest neighbors estimate (Section 3.1) and the bagged 1-nearest neighbor estimate (Section 3.2). As we will see, both satisfy the concentration inequality (2). Next, Section 3.3 details the mutual nearest neighbors estimate, which is not a weighted nearest neighbor rule, but still a local averaging procedure. The concentration inequality we obtain in this case is not exactly of the form (2) for the constant a depends on some extra parameter.

Remark 1. *The fact that the upper bound of Lemma 3 is valid for all $d \geq 1$ and, more importantly, for all $x \in S$, is due to inequality (1) in Assumption 1. If one only supposes that the support of X is bounded, then Biau et al. [5], Corollary 2.1, for $d = 1$ or $d \geq 3$ and a consequence of Liitiäinen et al. [31], Theorem 3.2, when $d = 2$, only ensure that there exists a constant c_d depending on the dimension d and the size of the support such that, for all $d \geq 2$,*

$$\mathbb{E}\left[D_{(i)}(X)^2\right] \leq c_d \left\lfloor \frac{N}{i} \right\rfloor^{-2/d},$$

whereas for $d = 1$ one has $\mathbb{E}\left[D_{(i)}(X)^2\right] \leq c_1 \frac{i}{N}$.

3.1. **The k Nearest Neighbors estimate.** Let us focus on the case of uniform k-nearest neighbors (k-nn). Namely, we now set

$$\widehat{r}_N(x) := \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}(x),$$

for some $k \in [\![1, N]\!]$. Then, for all $x \in S$, (12) becomes

$$t = 2e\sigma\sqrt{\frac{2}{k}} \qquad \text{and} \qquad s = \frac{16e^2}{k} \sum_{i=1}^{k} \left(\frac{i}{\rho(N+1)}\right)^{1/d} \leq 16e^2 \left(\frac{km}{\rho n}\right)^{1/d},$$

where we used that $N = \lfloor \frac{n}{m} \rfloor \geq \frac{n}{m} - 1$. We thus deduce from (13) that

$$\mathbb{P}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \geq 2e\sigma\sqrt{\frac{2}{k}} + 16e^2\left(\frac{km}{\rho n}\right)^{1/d}\right) \leq e^{-m}. \tag{14}$$

When σ and ρ are known, one may then choose k as the largest integer such that $\sigma\sqrt{\frac{2}{k}} \geq 8e\left(\frac{km}{\rho n}\right)^{1/d}$, i.e.

$$k^\star = \left\lfloor \left(\frac{\sigma^2}{32e^2}\right)^{\frac{d}{d+2}} \left(\frac{\rho n}{m}\right)^{\frac{2}{d+2}} \right\rfloor,$$

which belongs to $[\![1, N]\!] = [\![1, \lfloor \frac{n}{m} \rfloor]\!]$ provided

$$1 \leq \left(\frac{\sigma^2}{32e^2}\right)^{\frac{d}{d+2}} \left(\frac{\rho n}{m}\right)^{\frac{2}{d+2}} \leq \frac{n}{m},$$

i.e.

$$\frac{m}{n} \leq \rho\left(\frac{\sigma}{4e\sqrt{2}}\right)^d \wedge \frac{32e^2}{\sigma^2 \rho^{2/d}}.$$

In this case, using that $\lfloor u \rfloor \geq u/2$ for $u \geq 1$, we get

$$4e\sigma\sqrt{\frac{2}{k^\star}} \leq 8e\sigma\sqrt{\left(\frac{32e^2}{\sigma^2}\right)^{\frac{d}{d+2}} \left(\frac{m}{\rho n}\right)^{\frac{2}{d+2}}} \leq 32e^2\sqrt{2}\left(\frac{\sigma^2 m}{\rho n}\right)^{\frac{1}{d+2}}.$$

In view of (14), we have, for the optimal choice $k^\star$,

$$\mathbb{P}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \geq 32e^2\sqrt{2}\left(\frac{\sigma^2 m}{\rho n}\right)^{\frac{1}{d+2}}\right) \leq e^{-m}.$$

Finally, we arrive at the following result.

Proposition 4. *Under Assumption 1, let*

$$c = \rho \left(\frac{\sigma}{4e\sqrt{2}} \right)^d \wedge \frac{32e^2}{\sigma^2 \rho^{2/d}} \wedge 1 \,, \quad \delta \in [e^{-cn+1}, 1[\,, \quad \textit{and} \quad m = \lceil \ln(1/\delta) \rceil \,,$$

ensuring that $1 \le m \le cn$. Then the estimator $\widehat{r}_n^{\mathsf{mom}}$ constructed on m blocks with $k^\star$-nn base estimators, where

$$k^\star = \left\lfloor \left(\frac{\sigma^2}{32e^2} \right)^{\frac{d}{d+2}} \left(\frac{\rho n}{m} \right)^{\frac{2}{d+2}} \right\rfloor \,,$$

satisfies

$$\mathbb{P}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \ge 32e^2\sqrt{2} \left(\frac{\sigma^2 \lceil \ln(1/\delta) \rceil}{\rho n} \right)^{\frac{1}{d+2}} \right) \le \delta \,.$$

Hence, when $k^\star$-nn is chosen as base estimate, inequality (2) is satisfied with $a = 32e^2\sqrt{2}$ and $c_{\mathcal{F}} = c$.

Remark 2. *One may notice that the optimal value for k has the same dependency with respect to σ^2 and n, that is $k^\star = O(\sigma^{\frac{2d}{d+2}} n^{\frac{2}{d+2}})$, as the one that balances bias and variance when minimizing the L_2 risk, see Györfi et al. [17] Theorem 6.2. In a different setting, the conclusion is the same in the work of Jiang, see Remark 1 in Jiang [23].*

3.2. **Bagging and Nearest Neighbors.** We now turn to the bagged 1-nn estimate with replacement. Bagging (for **b**ootstrap **agg**regat**ing**) is a simple way to combine estimates in order to improve their performance. This method, suggested by Breiman [6], proceeds by resampling from the original data set, constructing a predictor from each subsample, and decide by combining. By bagging an N-sample, the crude nearest neighbor regression estimate is turned into a consistent weighted nearest neighbor regression estimate, which is amenable to statistical analysis. In particular, one may find experimental performances, consistency results, rates of convergence, and minimax properties in Hall and Samworth [18], Biau and Devroye [3], Biau et al. [5], to cite just a few references.

Without going into details, it turns out that bagged 1-nn estimates, with or without replacement, can easily be reformulated as weighted nearest neighbor rules (see for example Biau and Devroye [3]). For sampling without replacement, the weights in (8) are, for some $k \in [\![1, N]\!]$,

$$v_i := \frac{\binom{N-i}{k-i}}{\binom{N}{k}} \mathbf{1}_{i \in [\![1, N-k+1]\!]} \,,$$

while for sampling with replacement, we get

$$v_i := \left(1 - \frac{i-1}{N} \right)^k - \left(1 - \frac{i}{N} \right)^k .$$

From now on, let us focus on sampling with replacement. Concerning the variance term in (12), Proposition 2.2 in Biau et al. [5] and the fact that $k \le N$ yield

$$\sum_{i=1}^N v_i^2 \le \frac{2k}{N} \left(1 + \frac{1}{N} \right)^{2k} \le \frac{2e^2 k}{N} \le \frac{4e^2 km}{n} \,,$$

where for the last inequality, we used that $N = \left\lfloor \frac{n}{m} \right\rfloor \ge \frac{n}{2m}$. For the bias term in (12), we have to upper bound the quantity

$$\sum_{i=1}^N v_i \left(\frac{i}{N+1} \right)^{1/d} .$$

This is the purpose of the upcoming result, whose proof is detailed in Section 8.4.

Lemma 5. *For all $d \geq 1$, one has*

$$\sum_{i=1}^{N} v_i \left(\frac{i}{N+1}\right)^{1/d} \leq 2ek^{-1/d}\,.$$

In view of (12), we are led to

$$t \leq 4e^2\sigma\sqrt{\frac{2km}{n}} \qquad \text{and} \qquad s \leq \frac{32e^3}{(\rho k)^{1/d}}\,,$$

and, by Lemma 1, for all $x \in S$,

$$\mathbb{P}\left(|\widehat{r}_n^{\,\mathsf{mom}}(x) - r(x)| \geq 4e^2\sigma\sqrt{\frac{2km}{n}} + \frac{32e^3}{(\rho k)^{1/d}}\right) \leq e^{-m}\,.$$

As for the k-nn case, when σ and ρ are known, the integer k may be chosen as the largest integer such that $\frac{8e}{(\rho k)^{1/d}} \geq \sigma\sqrt{\frac{2km}{n}}$, that is

$$k^\star = \left\lfloor \left(\frac{32e^2 n}{\rho^{2/d}\sigma^2 m}\right)^{\frac{d}{d+2}} \right\rfloor,$$

which belongs to $[\![1, N]\!] = [\![1, \lfloor \frac{n}{m} \rfloor]\!]$ if

$$\frac{m}{n} \leq \rho\left(\frac{\sigma}{4e\sqrt{2}}\right)^d \wedge \frac{32e^2}{\sigma^2\rho^{2/d}}\,.$$

In this case, using that $\lfloor u \rfloor \geq u/2$ for $u \geq 1$, we get after some simplification

$$\mathbb{P}\left(|\widehat{r}_n^{\,\mathsf{mom}}(x) - r(x)| \geq 128e^3\left(\frac{\sigma^2 m}{\rho n}\right)^{\frac{1}{d+2}}\right) \leq e^{-m}\,.$$

Putting all pieces together gives the following statement.

Proposition 6. *Under Assumption 1, let*

$$c = \rho\left(\frac{\sigma}{4e\sqrt{2}}\right)^d \wedge \frac{32e^2}{\sigma^2\rho^{2/d}} \wedge 1\,, \quad \delta \in [e^{-cn+1}, 1[\,, \quad \textit{and} \quad m = \lceil \ln(1/\delta) \rceil\,,$$

ensuring that $1 \leq m \leq cn$. Then the estimator $\widehat{r}_n^{\,\mathsf{mom}}$ constructed on m blocks with $k^\star$-bagged 1-nn base estimators, where

$$k^\star = \left\lfloor \left(\frac{32e^2 n}{\rho^{2/d}\sigma^2 m}\right)^{\frac{d}{d+2}} \right\rfloor,$$

satisfies

$$\mathbb{P}\left(|\widehat{r}_n^{\,\mathsf{mom}}(x) - r(x)| \geq 128e^3\left(\frac{\sigma^2 \lceil \ln(1/\delta) \rceil}{\rho n}\right)^{\frac{1}{d+2}}\right) \leq \delta\,.$$

In summary, the result for the bagged 1-nn estimate is the same as for the k-nn estimate as established in Proposition 4, that is inequality (2) with this time $a = 128e^3$ and exactly the same expression for $c_{\mathcal{F}} = c$.

3.3. **Mutual Nearest Neighbors.** The term of mutual nearest neighbors (mnn) seems to date back to Chidananda Gowda and Krishna [11, 12] in the context of clustering. Since then, it has raised interest in image analysis for object retrieval (see, *e.g.*, Jégou et al. [21] and Qin et al. [41]) as well as for classification purposes (see Liu et al. [32]). Some theoretical results (consistency, rates of convergence) can be found in Guyader and Hengartner [16].

Denote $\mathcal{N}_k(x)$ the set of the k nearest neighbors of x in $\mathcal{D}_N$, $\mathcal{N}'_k(X_i)$ the set of the k nearest neighbors of X_i in $(\mathcal{D}_N \setminus \{X_i\}) \cup \{x\}$, and

$$\mathcal{M}_k(x) := \{X_i \in \mathcal{N}_k(x) : \ x \in \mathcal{N}'_k(X_i)\}$$

the set of the mutual nearest neighbors of x. Its cardinal $M_k(x) := |\mathcal{M}_k(x)|$ is a random variable with values between 0 and k. The mutual nearest neighbors regression estimate is defined as

$$\widehat{r}_N(x) := \frac{1}{M_k(x)} \sum_{i:X_i \in \mathcal{M}_k(x)} Y_i \,,$$

with the understanding that $0/0 = 0$. Setting $W_i(x) := \frac{1}{M_k(x)} \mathbf{1}_{X_i \in \mathcal{M}_k(x)}$ if $M_k(x) > 0$, and 0 otherwise, we can also write

$$\widehat{r}_N(x) := \sum_{i=1}^{N} W_i(x) Y_i \,.$$

However, notice that it does not admit formulation (8) since the weights are not deterministically linked to the order statistics $X_{(1)}(x), \ldots, X_{(N)}(x)$. Hence, *stricto sensu*, the mnn estimator does not enter into the framework of weighted nearest neighbors rules. Nonetheless, it is locally averaging and equation (7) combined with Markov's inequality gives

$$p_{t+s}(x) \leq \frac{\sigma^2 \mathbb{E}\left[\sum_{i=1}^N W_i(x)^2\right]}{t^2} + \frac{\mathbb{E}\left[\sum_{i=1}^N W_i(x) \|X_i - x\|\right]}{s} \,. \tag{15}$$

In the latter, we can reformulate the first term as

$$\mathbb{E}\left[\sum_{i=1}^{N} W_i(x)^2\right] = \mathbb{E}\left[\frac{\mathbf{1}_{M_k(x)>0}}{M_k(x)}\right] \,. \tag{16}$$

Now, according to equation (2.1) in Guyader and Hengartner [16], we have

$$M_k(x) \geq \left| \left\{ i \in [\![1, N]\!] \,, \ \|X_i - x\| < \frac{\|X_{(k+1)}(x) - x\|}{2} \right\} \right| \,.$$

Then, conditionally on $D_{(k+1)}(x) = \|X_{(k+1)}(x) - x\|$, it can be shown (see for instance Lemma A.1 in Cérou and Guyader [10]) that $M_k(x)$ is stochastically larger than a Binomial random variable with parameters k and

$$\frac{\mu\left(\mathring{\mathcal{B}}(x, D_{(k+1)}(x)/2)\right)}{\mu\left(\mathcal{B}(x, D_{(k+1)}(x))\right)} \,,$$

where $\mathring{\mathcal{B}}(x, r)$ stands for the open ball with center x and radius r. In light of this, a convenient way to control $\mathbb{E}\left[M_k(x)^{-1}\right]$ in (16) is to add the assumption that the measure μ is "doubling" in the following sense: there exists a constant $\alpha > 0$ such that for all $x \in S$ and all $\varepsilon > 0$,

$$\frac{\mu\left(\mathring{\mathcal{B}}(x, \varepsilon/2)\right)}{\mu\left(\mathcal{B}(x, \varepsilon)\right)} \geq \alpha \cdot \tag{17}$$

Let us mention that, in the standard definition of a doubling measure, the denominator in (17) is $\mu(\mathring{\mathcal{B}}(x, \varepsilon))$. However, under this additional assumption, $M_k(x)$ is stochastically larger than a Binomial random variable with parameters k and α, and by Györfi et al. [17], Lemma 4.1, we can finally upper bound the variance term (16) by

$$\mathbb{E}\left[\sum_{i=1}^{N} W_i(x)^2\right] = \mathbb{E}\left[\frac{\mathbf{1}_{M_k(x)>0}}{M_k(x)}\right] \leq \frac{2\alpha}{k+1}\,.$$

Concerning the bias term in (15), simply notice that any mutual nearest neighbors of x belongs to the k nearest neighbors of x. By Lemma 3, this implies

$$\mathbb{E}\left[\sum_{i=1}^{N} W_i(x)\|X_i - x\|\right] \leq \mathbb{E}\left[D_{(k)}(x)\right] \leq 2\left(\frac{km}{\rho n}\right)^{1/d}\,.$$

By the same reasoning as before, if we set

$$t = 4e\sigma\sqrt{\frac{\alpha}{k}} \qquad \text{and} \qquad s = 16e^2\left(\frac{km}{\rho n}\right)^{1/d}$$

then Lemma 1 yields $\mathbb{P}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \geq t + s\right) \leq e^{-m}$. Therefore, one may carry out the optimization exactly as in the k-nn case but with σ replaced by $\sigma\sqrt{2\alpha}$, and obtain a proposition akin to Proposition 4, with σ modified accordingly.

Proposition 7. *Suppose that Assumption 1 and (17) are satisfied. Let*

$$c = \rho\left(\frac{\sigma\alpha}{4e}\right)^d \wedge \frac{16e^2}{\alpha\sigma^2\rho^{2/d}} \wedge 1\,, \quad \delta \in [e^{-cn+1}, 1[\,, \quad \text{and} \quad m = \lceil \ln(1/\delta)\rceil\,,$$

ensuring that $1 \leq m \leq cn$. *Then the estimator* $\widehat{r}_n^{\mathsf{mom}}$ *constructed on* m *blocks with* $k^\star$*-*mnn *base estimators, where*

$$k^\star = \left\lfloor \left(\frac{\alpha\sigma^2}{16e^2}\right)^{\frac{d}{d+2}} \left(\frac{\rho n}{m}\right)^{\frac{2}{d+2}} \right\rfloor\,,$$

satisfies

$$\mathbb{P}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \geq 64e^2\alpha^{1/3}\left(\frac{\sigma^2\lceil\ln(1/\delta)\rceil}{\rho n}\right)^{\frac{1}{d+2}}\right) \leq \delta\,.$$

In summary, the mnn estimate does not exactly satisfy inequality (2), since $a = 64e^2\alpha^{1/3}$ depends on the extra parameter α.

4. Kernel and partitioning estimation

4.1. Kernel estimates.

Let $0 < h \leq D$ and consider the kernel estimator

$$\widehat{r}_N(x) := \frac{1}{N_h(x)}\sum_{i=1}^{N} Y_i \mathbf{1}_{\|X_i - x\| \leq h}\,,$$

where

$$N_h(x) := \sum_{i=1}^{N} \mathbf{1}_{\|X_i - x\| \leq h}\,,$$

with the convention $0/0 = 0$. Observe that this is again of the form (5) with

$$W_i(x) = N_h(x)^{-1}\mathbf{1}_{\|X_i - x\| \leq h}\,.$$

By inequality (6), we get

$$|\widehat{r}_N(x) - r(x)| \leq \left|\sum_{i=1}^{N} W_i(x)\varepsilon_i\right| + \sum_{i=1}^{N} W_i(x)\|X_i - x\|\,.$$

In this case, we have, deterministically,

$$\sum_{i=1}^{N} W_i(x)\|X_i - x\| = \frac{1}{N_h(x)} \sum_{i=1}^{N} \|X_i - x\| \mathbf{1}_{\|X_i - x\| \le h} \le h\,.$$

Hence, for all $t > 0$ and all $x \in S$, Markov's inequality yields

$$p_{t+h}(x) = \mathbb{P}(|\widehat{r}_N(x) - r(x)| \ge t + h) \le \mathbb{P}\left(\left|\sum_{i=1}^{N} W_i(x)\varepsilon_i\right| \ge t\right) \le \frac{\sigma^2 \mathbb{E}\left[\sum_{i=1}^{N} W_i(x)^2\right]}{t^2}\,,$$

so that

$$p_{t+h}(x) \le \frac{\sigma^2 \mathbb{E}\left[\frac{\mathbf{1}_{N_h(x)>0}}{N_h(x)}\right]}{t^2}\,.$$

Since $N_h(x)$ is distributed as a Binomial random variable with parameters N and $\mu\left(\mathcal{B}(x,h)\right)$, we have, by Györfi et al. [17], Lemma 4.1, and Assumption 1,

$$\mathbb{E}\left[\frac{\mathbf{1}_{N_h(x)>0}}{N_h(x)}\right] \le \frac{2}{(N+1)\mu\left(\mathcal{B}(x,h)\right)} \le \frac{2m}{\rho n h^d}\,.$$

Hence, we obtain

$$p_{t+h}(x) \le \frac{2\sigma^2 m}{\rho n h^d t^2}\,,$$

Since the right-hand side equals $1/4e^2$ for $t = 2e\sqrt{\frac{2\sigma^2 m}{\rho n h^d}}$, Lemma 1 implies

$$\mathbb{P}\left(|\widehat{r}_n^{\,\mathsf{mom}}(x) - r(x)| \ge 2e\sqrt{\frac{2\sigma^2 m}{\rho n h^d}} + h\right) \le e^{-m}\,.$$

When σ and ρ are known, the bandwidth h can then be optimized: taking h such that $2e\sqrt{\frac{2\sigma^2 m}{\rho n h^d}} = h$, i.e.

$$h^\star = \left(\frac{8e^2\sigma^2 m}{\rho n}\right)^{\frac{1}{d+2}}\,,$$

we see that if $m \le \frac{\rho D^{d+2} n}{8e^2\sigma^2}$, ensuring $h^\star \le D$, we have

$$\mathbb{P}\left(|\widehat{r}_n^{\,\mathsf{mom}}(x) - r(x)| \ge 2\left(\frac{8e^2\sigma^2 m}{\rho n}\right)^{\frac{1}{d+2}}\right) \le e^{-m}\,.$$

All in all, we arrive at the following proposition.

Proposition 8. *Under Assumption 1, let*

$$c = \frac{\rho D^{d+2}}{8e^2\sigma^2} \wedge 1\,, \qquad \delta \in [e^{-cn+1}, 1[\,, \qquad \textit{and} \qquad m = \lceil \ln(1/\delta) \rceil\,,$$

ensuring that $1 \le m \le cn$. Then the estimator $\widehat{r}_n^{\,\mathsf{mom}}$ constructed on m blocks with $h^\star$-kernel base estimators, where

$$h^\star = \left(\frac{8e^2\sigma^2 m}{\rho n}\right)^{\frac{1}{d+2}}\,,$$

satisfies

$$\mathbb{P}\left(|\widehat{r}_n^{\,\mathsf{mom}}(x) - r(x)| \ge 4e^{2/3}\left(\frac{\sigma^2\lceil \ln(1/\delta)\rceil}{\rho n}\right)^{\frac{1}{d+2}}\right) \le \delta\,.$$

In other words, inequality (2) is fulfilled with $a = 4e^{2/3}$ and $c_{\mathcal{F}} = c$.

4.2. **Partitioning estimates.** To simplify the presentation, let us here assume that $S = [0,1]^d$. For some integer $K \geq 1$, let $\mathcal{P} = \{A_1, A_2, \ldots, A_{K^d}\}$ be a cubic partition of $[0,1]^d$ by K^d cubes with side length $1/K$. For $k \in [\![1, K^d]\!]$, if $x \in A_k$, the partitioning estimate of the regression function takes the form

$$\widehat{r}_N(x) = \sum_{i=1}^{N} W_i(x) Y_i := \frac{1}{N_k} \sum_{i=1}^{N} Y_i \mathbf{1}_{X_i \in A_k} \,,$$

where

$$N_k := \sum_{i=1}^{N} \mathbf{1}_{X_i \in A_k} \,,$$

with the usual convention $0/0 = 0$. The reasoning is *mutatis mutandis* the same as for kernel estimates. Here again, the bias term can indeed be deterministically bounded:

$$\sum_{i=1}^{N} W_i(x) \|X_i - x\| = \frac{1}{N_k} \sum_{i=1}^{N} \mathbf{1}_{X_i \in A_k} \|X_i - x\| \leq \sqrt{d} K^{-1} \,. \tag{18}$$

Hence, for all $t > 0$ and $x \in A_k$, we are led to

$$p_{t+\sqrt{d}K^{-1}}(x) \leq \frac{\sigma^2 \mathbb{E}\left[\frac{\mathbf{1}_{N_k > 0}}{N_k}\right]}{t^2} \leq \frac{2\sigma^2}{(N+1)\mu(A_k) t^2} \leq \frac{2^{d+1} K^d \sigma^2 m}{\rho n t^2} \,, \tag{19}$$

where we used that, if a_k denotes the center of A_k, then by assumption (1)

$$\mu(A_k) \geq \mu\left(\mathcal{B}(a_k, (2K)^{-1})\right) \geq \rho (2K)^{-d} \,.$$

Again, by Lemma 1, we get

$$\mathbb{P}\left(|\widehat{r}_n^{\,\mathsf{mom}}(x) - r(x)| \geq 2e \sqrt{\frac{2^{d+1} K^d \sigma^2 m}{\rho n}} + \sqrt{d} K^{-1} \right) \leq e^{-m} \,.$$

One may then choose K as the largest integer such that $\sqrt{d} K^{-1} \geq 2e \sqrt{\frac{2^{d+1} K^d \sigma^2 m}{\rho n}}$, i.e.

$$K_\star = \left\lfloor \left(\frac{\rho d n}{2^{d+3} e^2 \sigma^2 m} \right)^{\frac{1}{d+2}} \right\rfloor \,,$$

which belongs to $\mathbb{N} \setminus \{0\}$ as soon as

$$\frac{m}{n} \leq \frac{\rho d}{2^{d+3} e^2 \sigma^2} \,.$$

Once again, using that $\lfloor u \rfloor \geq u/2$ for $u \geq 1$, we obtain

$$2\sqrt{d} K_\star^{-1} \leq 4\sqrt{d} \left(\frac{2^{d+3} e^2 \sigma^2 m}{\rho d n} \right)^{\frac{1}{d+2}} \leq 16 e^{2/3} \sqrt{d} \left(\frac{\sigma^2 m}{\rho n} \right)^{\frac{1}{d+2}} \,,$$

which yields

$$\mathbb{P}\left(|\widehat{r}_n^{\,\mathsf{mom}}(x) - r(x)| \geq 16 e \sqrt{d} \left(\frac{\sigma^2 m}{\rho n} \right)^{\frac{1}{d+2}} \right) \leq e^{-m} \,.$$

Putting all things together gives the upcoming result, which shows that inequality (2) is fulfilled with $a = 16 e \sqrt{d}$ and $c_{\mathcal{F}} = c$.

Proposition 9. *Under Assumption 1 with $S = [0,1]^d$, let*

$$c = \frac{\rho d}{2^{d+3} e^2 \sigma^2} \wedge 1 \,, \quad \delta \in [e^{-cn+1}, 1[\,, \quad \text{and} \quad m = \lceil \ln(1/\delta) \rceil \,,$$

ensuring that $1 \le m \le cn$. Then the estimator $\widehat{r}_n^{\mathsf{mom}}$ on $[0,1]^d$ constructed on m blocks with partitioning base estimators on $K_\star^d$ hypercubes, where

$$K_\star = \left\lfloor \left(\frac{\rho d n}{2^{d+3} e^2 \sigma^2 m} \right)^{\frac{1}{d+2}} \right\rfloor \,,$$

satisfies

$$\mathbb{P}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \ge 16 e \sqrt{d} \left(\frac{\sigma^2 \lceil \ln(1/\delta) \rceil}{\rho n} \right)^{\frac{1}{d+2}} \right) \le \delta \,.$$

One enjoyable feature of partitioning estimates is that uniform bounds for $x \in S$ can easily be obtained. Indeed, for all $t > 0$, we have

$$\mathbb{P}\left(\sup_{x \in S} |\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| > t + \sqrt{d} K^{-1} \right) \le \mathbb{P}\left(\sup_{x \in S} \sum_{j=1}^m \mathbf{1}_{\{|\widehat{r}^{(j)}(x) - r(x)| > t + \sqrt{d} K^{-1}\}} \ge \frac{m}{2} \right) .$$

Then, by inequality (6) of Lemma 2 and the deterministic bound (18) on the bias term, it comes

$$\mathbb{P}\left(\sup_{x \in S} |\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| > t + \sqrt{d} K^{-1} \right) \le \mathbb{P}\left(\sup_{x \in S} \sum_{j=1}^m \mathbf{1}_{\left\{\left|\sum_{i=1}^N W_i^{(j)}(x) \varepsilon_i^{(j)}\right| > t\right\}} \ge \frac{m}{2} \right) ,$$

where $\varepsilon_1^{(j)}, \dots, \varepsilon_N^{(j)}$ stand for the noise variables in block j, and where

$$W_i^{(j)}(x) := \sum_{k=1}^{K^d} \mathbf{1}_{x \in A_k} \frac{1}{N_k^{(j)}} \mathbf{1}_{X_i^{(j)} \in A_k} \,,$$

with $X_1^{(j)}, \dots, X_N^{(j)}$ the features in block j, and $N_k^{(j)}$ the number of features in block j falling into A_k. Noticing that $W_i^{(j)}(x)$ does not depend on the exact position of x but only on the cube A_k in which it lies, we obtain, with the notation $B_{i,k}^{(j)} := \mathbf{1}_{\left\{X_i^{(j)} \in A_k\right\}}$,

$$\begin{aligned}
\mathbb{P}\left(\sup_{x \in S} |\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| > t + \sqrt{d} K^{-1} \right) &\le \mathbb{P}\left(\sup_{1 \le k \le K^d} \sum_{j=1}^m \mathbf{1}_{\left\{\left|\frac{1}{N_k^{(j)}} \sum_{i=1}^N B_{i,k}^{(j)} \varepsilon_i^{(j)}\right| > t\right\}} \ge \frac{m}{2} \right) \\
&\le \sum_{k=1}^{K^d} \mathbb{P}\left(\sum_{j=1}^m \mathbf{1}_{\left\{\left|\frac{1}{N_k^{(j)}} \sum_{i=1}^N B_{i,k}^{(j)} \varepsilon_i^{(j)}\right| > t\right\}} \ge \frac{m}{2} \right) \\
&= \sum_{k=1}^{K^d} \mathbb{P}\left(\sum_{j=1}^m B_k^{(j)} \ge \frac{m}{2} \right) ,
\end{aligned}$$

thanks to the union bound and with the notation

$$B_k^{(j)} := \mathbf{1}_{\left\{\left|\frac{1}{N_k^{(j)}} \sum_{i=1}^N B_{i,k}^{(j)} \varepsilon_i^{(j)}\right| > t\right\}} .$$

Clearly, for each $k \in [\![1, K]\!]$, the Bernoulli random variables $(B_k^{(j)})_{1\le j\le m}$ are i.i.d. with parameter

$$p_k := \mathbb{P}\left(\left|\frac{1}{N_k^{(j)}}\sum_{i=1}^{N} B_{i,k}^{(j)}\varepsilon_i^{(j)}\right| > t\right) \le \frac{2^{d+1}K^d\sigma^2 m}{\rho n t^2},$$

where the upper bound, which does not depend on k, comes from (19). We may now apply inequality (25) in the proof of Lemma 1 to deduce that

$$\mathbb{P}\left(\sup_{x\in S}|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| > t + \sqrt{d}K^{-1}\right) \le K^d \cdot 2^m \left(\frac{2^{d+1}K^d\sigma^2 m}{\rho n t^2}\right)^{m/2}.$$

Choosing t appropriately, we get

$$\mathbb{P}\left(\sup_{x\in S}|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| > e\sqrt{\frac{2^{d+3}\sigma^2K^{d+\frac{2d}{m}}m}{\rho n}} + \sqrt{d}K^{-1}\right) \le e^{-m}.$$

In particular, we see that if $K \to +\infty$, and if there exists $\varepsilon > 0$ such that $n^{-1}K^{d+\varepsilon} \to 0$, then, choosing m such that $n^{-1}mK^{d+\frac{2d}{m}} \to 0$ and e^{-m} is summable (*e.g.*, $m = (\log n)^2$), Borel–Cantelli Lemma entails that

$$\sup_{x\in S}|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \underset{n\to+\infty}{\longrightarrow} 0 \quad \text{almost surely.}$$

5. Robustness to outliers

In this section, we consider the contamination scheme introduced by Lecué and Lerasle [25]. We assume that the index set $\{1, \dots, n\}$ is divided into two disjoint subsets: the subset $\mathcal{I}$ of inliers, and the subset $\mathcal{O}$ of outliers. The sequence $(X_i, Y_i)_{i\in\mathcal{I}}$ is i.i.d. with the same law as $(X, Y) \in \mathcal{F}$. No assumption is made on the variables $(X_i, Y_i)_{i\in\mathcal{O}}$. We denote by $\mathbb{P}_{\mathcal{O}\cup\mathcal{I}}$ the distribution corresponding to such a contaminated sample.

Let $\widehat{r}_n^{\mathsf{mom}}$ be the median-of-means estimate of r constructed on m blocks, with base estimate $\widehat{r}_N$. Define ε as the proportion of outliers in the original sample and assume that $m > 2|\mathcal{O}| = 2\varepsilon n$, where $|\mathcal{O}|$ is the number of outliers. Then, letting $\mathcal{B}$ be the set of blocks that do not intersect $\mathcal{O}$, we have, for all $s, t > 0$,

$$\begin{aligned}\mathbb{P}_{\mathcal{O}\cup\mathcal{I}}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \ge t + s\right) &\le \mathbb{P}_{\mathcal{O}\cup\mathcal{I}}\left(\sum_{j=1}^{m}\mathbf{1}_{\{|\widehat{r}^{(j)}(x)-r(x)|\ge t+s\}} \ge \frac{m}{2}\right)\\ &\le \mathbb{P}_{\mathcal{O}\cup\mathcal{I}}\left(|\mathcal{B}^c| + \sum_{j\in\mathcal{B}}\mathbf{1}_{\{|\widehat{r}^{(j)}(x)-r(x)|\ge t+s\}} \ge \frac{m}{2}\right).\end{aligned}$$

Observing that $|\mathcal{B}^c| \le |\mathcal{O}| = \varepsilon n < m/2$, we get, with a slight abuse of notation,

$$\mathbb{P}_{\mathcal{O}\cup\mathcal{I}}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \ge t + s\right) \le \mathbb{P}\left(\sum_{j=1}^{m}\mathbf{1}_{\{|\widehat{r}^{(j)}(x)-r(x)|\ge t+s\}} \ge \frac{m}{2} - \varepsilon n\right).$$

Now, the proof of Lemma 1 reveals that

$$\mathbb{P}\left(\sum_{j=1}^{m}\mathbf{1}_{\{|\widehat{r}^{(j)}(x)-r(x)|\ge t+s\}} \ge \frac{m}{2} - \varepsilon n\right) \le 2^m p_{t+s}(x)^{\frac{m}{2}-\varepsilon n},$$

which is less than e^{-m} for $p_{t+s}(x) \leq (2e)^{-\frac{2m}{m-2\varepsilon n}}$. Hence, provided $m = \lceil \ln(1/\delta) \rceil > 2|\mathcal{O}| = 2\varepsilon n$, the same strategy as the one described just after the statement of Lemma 2 can be applied to the contaminated setting.

For example, in the case of uniform nearest neighbors detailed in Section 3.1, some computation shows that the estimator $\widehat{r}_n^{\mathsf{mom}}$ constructed on m blocks with $k^\star$-nn base estimators, where

$$k^\star = \left\lfloor \left(\frac{1}{8} (2e)^{-\frac{2m}{m-2\varepsilon n}} \sigma^2 \left(\frac{\rho n}{m} \right)^{\frac{2}{d}} \right)^{\frac{d}{d+2}} \right\rfloor ,$$

which belongs to $[\![1, N]\!] = [\![1, \lfloor \frac{n}{m} \rfloor]\!]$ provided

$$1 \leq \left(\frac{1}{8} (2e)^{-\frac{2m}{m-2\varepsilon n}} \sigma^2 \left(\frac{\rho n}{m} \right)^{\frac{2}{d}} \right)^{\frac{d}{d+2}} \leq \frac{n}{m} ,$$

satisfies

$$\mathbb{P}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \geq 8\sqrt{2}(2e)^{\frac{2m}{m-2\varepsilon n}} \left(\frac{\sigma^2 \lceil \ln(1/\delta) \rceil}{\rho n} \right)^{\frac{1}{d+2}} \right) \leq \delta .$$

6. Lower bound

For some $h > 0$ to be specified later, let

$$S = \left[0\,,\, h \left\lceil h^{-1} \right\rceil\right]^d .$$

We consider the model

$$Y = r(X) + \varepsilon ,$$

where $X \sim \mu = \mathrm{Unif}(S)$, the noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is independent of X, and r is a 1-Lipschitz function on S. Notice that this model satisfies Assumption 1, meaning that $(X, Y) \in \mathcal{F}$. Indeed, the support S is bounded with diameter $D = h \left\lceil h^{-1} \right\rceil \sqrt{d}$ and, for all $x \in S$ and $\varepsilon \in (0, D]$, one has

$$\mu\left(\mathcal{B}(x, \varepsilon)\right) \geq \mu\left(\mathcal{B}(x, \varepsilon/\sqrt{d})\right) \geq \mu\left(\mathcal{B}(0, \varepsilon/\sqrt{d})\right) = \frac{\pi^{d/2}}{2^d d^{d/2} \Gamma(1 + d/2)} \varepsilon^d =: \rho \varepsilon^d .$$

Let us stress that ρ depends on d but not on h. Our purpose is to show that the concentration properties we have obtained so far are optimal in the sense of inequality (4).

Proposition 10. *For all $\delta \in]0, 2^{-(d+3)}]$, for any regression estimate $\widehat{r}_n$, there exists a 1-Lipschitz mapping $r = r_n : S \to \mathbb{R}$ such that, when $X \sim \mu$ independent of $\mathcal{D}_n$, we have*

$$\mathbb{P}\left(|\widehat{r}_n(X) - r(X)| \geq \frac{1}{4} \left(\frac{\sigma^2 \ln\left(\frac{1}{2^{d+3}\delta}\right)}{n} \right)^{\frac{1}{d+2}} \right) \geq \delta . \tag{20}$$

Proof. This lower bound is established using the same idea as in the proof of Theorem 3.2 in Györfi et al. [17] (see also Section 2 in Stone [43]). Namely, let $\mathcal{C} := \left[-\frac{1}{2}, \frac{1}{2}\right]^d$ and $\partial\mathcal{C}$ its frontier, then define

$$g(x) := \mathrm{dist}(x, \partial\mathcal{C})\; \mathbf{1}_{x \in \mathcal{C}} = \inf\{\|x - y\|, y \in \partial\mathcal{C}\}\; \mathbf{1}_{x \in \mathcal{C}} .$$

Clearly, g is 1-Lipschitz and one can check that

$$\int g(x)^2 \mathrm{d}x = \frac{1}{2(d+1)(d+2)} . \tag{21}$$

Consider a partition of S by $K := \lceil h^{-1} \rceil^d$ hypercubes A_j of sidelength h and with centers a_j, and let the functions $g_1, \ldots, g_K$ be defined by

$$\forall j \in [\![1, K]\!] \,,\; g_j(x) := hg\left(h^{-1}(x - a_j)\right) \,.$$

Hence the support of g_j is $A_j = \left[a_j - \frac{h}{2}; a_j + \frac{h}{2}\right]^d$ and

$$\int g_j(x)^2 \mathrm{d}x = \frac{h^{d+2}}{2(d+1)(d+2)} \,.$$

The set of regression functions we consider in what follows is

$$\mathcal{R} := \left\{ r^{(c)} = \sum_{j=1}^{K} c_j g_j \,,\, c \in \{-1, 1\}^K \right\} \,.$$

The proof in Györfi et al. [17] ensures that each $r^{(c)}$ is 1-Lipschitz. Our goal is to show that for all $\delta \in]0, 2^{-(d+3)}]$, and for all regression estimate $\widehat{r}_n$ taking as input i.i.d. couples $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ with the same law as (X, Y), we have

$$\sup_{c \in \{-1,1\}^K} \mathbb{P}\left(\left| \widehat{r}_n(X) - r^{(c)}(X) \right| \geq \frac{h}{4} \right) \geq \delta \,, \tag{22}$$

with

$$h := \left(\frac{\pi \sigma^2 (d+1)(d+2) \ln\left(\frac{1}{2^{d+3}\delta}\right)}{n} \right)^{\frac{1}{d+2}} \,. \tag{23}$$

For this, let $\delta \in]0, 2^{-(d+3)}]$, h defined by (23), and $\widehat{r}_n$ denote a regression estimate. For $j \in [\![1, K]\!]$, and for $x \in A_j$, we start by defining

$$\widetilde{r}_n(x) := \mathrm{sign}\left(\widehat{r}_n(x)\right) g_j(x) \,.$$

Note that for all $x \in S$, and $c \in \{-1, 1\}^K$, we have

$$\left| \widehat{r}_n(x) - r^{(c)}(x) \right| \geq \frac{1}{2} \left| \widetilde{r}_n(x) - r^{(c)}(x) \right| \,.$$

We proceed by designing estimated signs $\widetilde{c}_j$ as follows: for all $j \in [\![1, K]\!]$, define the hypercube

$$A'_j := \left[a_j - \tfrac{h}{4}; a_j + \tfrac{h}{4}\right]^d \subset A_j$$

and set

$$\widetilde{c}_j := \begin{cases} +1 & \text{if } \mathbb{P}\left(\widetilde{r}_n(X) = g_j(X) \,\middle|\, X \in A'_j \right) \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

In other words, for each hypercube A_j, we take a majority vote, but only on A'_j. Observe that if X falls in the subset A'_j of a bad hypercube A_j, *i.e.* such that $\widetilde{c}_j \neq c_j$, then $\left| \widetilde{r}_n(X) - r^{(c)}(X) \right| \geq h/2$ with probability at least $1/2$. Combining those observations and the fact that $\mathbb{P}(X \in A'_j) = (2^d K)^{-1}$, we have

$$\mathbb{P}\left(\left| \widehat{r}_n(X) - r^{(c)}(X) \right| \geq \frac{h}{4} \right) \geq \mathbb{P}\left(\left| \widetilde{r}_n(X) - r^{(c)}(X) \right| \geq \frac{h}{2} \right) \geq \frac{1}{2^{d+1}K} \sum_{j=1}^{K} \mathbb{P}(\widetilde{c}_j \neq c_j) \,.$$

We are now left to show that

$$\sup_{c \in \{-1,1\}^K} \frac{1}{2^{d+1}K} \sum_{j=1}^{K} \mathbb{P}(\widetilde{c}_j \neq c_j) \geq \delta \,.$$

To do so, consider a uniform random vector $(C_1, \dots, C_K) \in \{-1, 1\}^K$ (that is, i.i.d. Rademacher random variables with parameter $1/2$). Clearly,

$$\sup_c \frac{1}{2^{d+1}K} \sum_{j=1}^{K} \mathbb{P}(\widetilde{c}_j \neq c_j) \geq \frac{1}{2^{d+1}K} \sum_{j=1}^{K} \mathbb{P}(\widetilde{c}_j \neq C_j)\,.$$

Now, for each $j \in [\![1, K]\!]$, the estimated sign $\widetilde{c}_j$ might be seen as a decision rule on C_j, based on the data $\mathcal{D}_n$. The minimal error probability is attained by the Bayes decision rule:

$$C_j^\star := \mathbf{1}_{\mathbb{P}(C_j=1|\mathcal{D}_n)\geq 1/2} - \mathbf{1}_{\mathbb{P}(C_j=1|\mathcal{D}_n)<1/2}\,.$$

Hence,

$$\mathbb{P}(\widetilde{c}_j \neq C_j) \geq \mathbb{P}(C_j^\star \neq C_j) = \mathbb{E}\left[\mathbb{P}(C_j^\star \neq C_j \,\middle|\, X_1, \dots, X_n)\right]\,.$$

Let $X_{i_1}, \dots, X_{i_\ell}$ be the variables X_i that fall in the hypercube A_j. Conditionally on $X_1, \dots, X_n$, the Bayesian rule for C_j based on $Y_1, \dots, Y_n$ only depends on $Y_{i_1}, \dots, Y_{i_\ell}$, and the problem comes down to the Bayesian estimation of $C \sim \mathrm{Rad}(1/2)$ in the model $Y = Cu + W$, where u is a fixed vector of $\mathbb{R}^\ell$ and W is a centered Gaussian vector with covariance matrix $\sigma^2 I_\ell$, independent of C. In this situation, Györfi et al. [17], Lemma 3.2, ensures that

$$\mathbb{P}(C_j^\star \neq C_j \,|\, X_1, \dots, X_n) = \Phi\left(-\frac{\sqrt{\sum_{s=1}^{\ell} g_j(X_{i_s})^2}}{\sigma}\right) = \Phi\left(-\frac{\sqrt{\sum_{i=1}^{n} g_j(X_i)^2}}{\sigma}\right)\,.$$

By Jensen's Inequality, we have

$$\mathbb{P}(C_j^\star \neq C_j) \geq \Phi\left(-\frac{\sqrt{\sum_{i=1}^{n} \mathbb{E}\left[g_j(X_i)^2\right]}}{\sigma}\right) = \Phi\left(-\frac{\sqrt{n\mathbb{E}\left[g_j(X)^2\right]}}{\sigma}\right)\,.$$

Since, by (21),

$$\mathbb{E}\left[g_j(X)^2\right] = \int g_j(x)^2 \mathrm{d}x = \frac{h^{d+2}}{2(d+1)(d+2)}\,,$$

we are led to

$$\begin{aligned}\sup_c \frac{1}{2^{d+1}K} \sum_{j=1}^{K} \mathbb{P}(\widetilde{c}_j \neq c_j) &\geq \frac{1}{2^{d+1}} \Phi\left(-\frac{1}{\sigma}\sqrt{\frac{nh^{d+2}}{2(d+1)(d+2)}}\right) \\ &\geq \frac{1}{2^{d+1}} \Phi\left(-\sqrt{\frac{\pi}{2}\ln\left(\frac{1}{2^{d+3}\delta}\right)}\right)\,,\end{aligned}$$

by definition of h in (23). We now use the following lower bound on the Gaussian tail (see for instance Pólya [40], equation (1.5)):

$$\forall x \geq 0,\ \ \Phi(-x) \geq \frac{1}{2}\left(1 - \sqrt{1 - e^{-\frac{2}{\pi}x^2}}\right) \geq \frac{1}{4} e^{-\frac{2}{\pi}x^2}\,,$$

where the second inequality is by $\sqrt{1-u} \leq 1 - u/2$ for $u \in [0, 1]$. Hence,

$$\sup_c \frac{1}{2^{d+1}K} \sum_{j=1}^{K} \mathbb{P}(\widetilde{c}_j \neq c_j) \geq \frac{1}{2^{d+3}} e^{-\ln\left(\frac{1}{2^{d+3}\delta}\right)} = \delta\,,$$

as desired. Returning to (22), there exists $r = r^{(c)}$ such that

$$\mathbb{P}\left(|\widehat{r}_n(X) - r(X)| \geq \frac{1}{4}\left(\frac{\pi\sigma^2(d+1)(d+2)\ln\left(\frac{1}{2^{d+3}}\delta\right)}{n}\right)^{\frac{1}{d+2}}\right) \geq \delta$$

with, for all $d \geq 1$,

$$(\pi(d+1)(d+2))^{\frac{1}{d+2}} \geq 1\,,$$

which concludes the proof. ■

7. From δ-dependent to δ-independent estimators

One feature of $\widehat{r}_n^{\mathsf{mom}}(x)$ is that it depends on m, the number of blocks, and thus on the pre-chosen confidence threshold $\delta = e^{-m}$. In this section, we give an argument due to Devroye et al. [14], allowing to turn $\widehat{r}_n^{\mathsf{mom}}(x)$ into an estimator satisfying (2) simultaneously for an infinity of δ. Let us first recall that, when σ and ρ are known, then for all integer m between 1 and $\lfloor cn \rfloor$ (for some constant c depending only on ρ, σ and d), one is able to construct a confidence interval $\widehat{I}_m$ for $r(x)$ with level $1 - e^{-m}$. Now let

$$\widehat{m} := \min \left\{ 1 \leq m \leq \lfloor cn \rfloor\,, \bigcap_{j=m}^{\lfloor cn \rfloor} \widehat{I}_j \neq \emptyset \right\}\,,$$

and define the estimator $\widetilde{r}_n(x)$ as the midpoint of the interval $\bigcap_{j=\widehat{m}}^{\lfloor cn \rfloor} \widehat{I}_j$. Let $\delta \in \left[\frac{e^{-\lfloor cn \rfloor}}{1-e^{-1}}, 1\right[$ and let m_δ be the smallest integer $m \in [\![1, \lfloor cn \rfloor]\!]$ such that $\delta \geq \frac{e^{-m}}{1-e^{-1}}$. Then $\widetilde{r}_n(x)$ satisfies

$$\mathbb{P}\left(|\widetilde{r}_n(x) - r(x)| > |\widehat{I}_{m_\delta}|\right) \leq \delta\,. \tag{24}$$

Indeed, by a union bound, we have

$$\mathbb{P}\left(\bigcap_{j=m_\delta}^{\lfloor cn \rfloor} \left\{ r(x) \in \widehat{I}_j \right\}\right) \geq 1 - \sum_{j=m_\delta}^{\lfloor cn \rfloor} e^{-j} \geq 1 - \frac{e^{-m_\delta}}{1-e^{-1}} \geq 1 - \delta\,.$$

Now, on the event $\bigcap_{j=m_\delta}^{\lfloor cn \rfloor} \left\{ r(x) \in \widehat{I}_j \right\}$, one has $\bigcap_{j=m_\delta}^{\lfloor cn \rfloor} \widehat{I}_j \neq \emptyset$, hence $\widehat{m} \leq m_\delta$. But if $\widehat{m} \leq m_\delta$, then $\widetilde{r}_n(x)$ also belongs to $\bigcap_{j=m_\delta}^{\lfloor cn \rfloor} \widehat{I}_j$, and in particular

$$|\widetilde{r}_n(x) - r(x)| \leq |\widehat{I}_{m_\delta}|\,,$$

which establishes (24).

For instance, for k-nn base estimates, one may combine Proposition 4 and this method to construct an estimator $\widetilde{r}_n(x)$ which is such that, for all $\delta \in \left[\frac{e^{-cn+1}}{1-e^{-1}}, 1\right[$,

$$\mathbb{P}\left(|\widetilde{r}_n(x) - r(x)| \geq 64e^2\sqrt{2}\left(\frac{\sigma^2\left\lceil \ln\left(\frac{1}{(1-e^{-1})\delta}\right)\right\rceil}{\rho n}\right)^{\frac{1}{d+2}}\right) \leq \delta\,,$$

with

$$c = \rho\left(\frac{\sigma}{4e\sqrt{2}}\right)^d \wedge \frac{32e^2}{\sigma^2\rho^{2/d}} \wedge 1\,.$$

8. Proofs of technical results

8.1. Proof of Lemma 1.

By definition of the median, we have

$$\mathbb{P}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \geq t\right) \leq \mathbb{P}\left(\sum_{j=1}^m \mathbf{1}_{\left\{|\widehat{r}^{(j)}(x) - r(x)| \geq t\right\}} \geq \frac{m}{2}\right)\,.$$

The variables $\mathbf{1}_{\{|\widehat{r}^{(j)}(x)-r(x)|\geq t\}}$ are i.i.d. Bernoulli variables with parameter $p_t(x)$. Now, if $B_1,\ldots,B_m$ are i.i.d. Bernoulli random variables with parameter $p\in[0,1]$, then for any real number $\ell\in[0,m]$ we may write

$$\mathbb{P}\left(\sum_{j=1}^m B_j \geq \ell\right) = \sum_{k=\lceil \ell\rceil}^m \binom{m}{k} p^k(1-p)^{m-k} \leq p^{\lceil \ell\rceil}\sum_{k=\lceil \ell\rceil}^m \binom{m}{k} \leq p^\ell \sum_{k=0}^m \binom{m}{k} = 2^m p^\ell\,. \quad (25)$$

In particular, taking $\ell = \frac{m}{2}$ gives the desired result in Lemma 1.

8.2. **Proof of Lemma 2.** For a given $x\in\mathbb{R}^d$, the difference $\widehat{r}_N(x)-r(x)$ can be decomposed as

$$\widehat{r}_N(x)-r(x) = \sum_{i=1}^N W_i(x)\varepsilon_i + \sum_{i=1}^N W_i(x)\left(r(X_i)-r(x)\right)\,,$$

where $\varepsilon_i = Y_i - r(X_i)$. By the triangle inequality and the fact that r is 1-Lipschitz, we have

$$\left|\sum_{i=1}^N W_i(x)r(X_i) - r(x)\right| \leq \sum_{i=1}^N W_i(x)\|X_i - x\|\,,$$

which establishes inequality (6). Next, for $t,s>0$, a union bound gives

$$p_{t+s}(x) \leq \mathbb{P}\left(\left|\sum_{i=1}^N W_i(x)\varepsilon_i\right| \geq t\right) + \mathbb{P}\left(\sum_{i=1}^N W_i(x)\|X_i - x\| \geq s\right)\,,$$

By Markov's inequality, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^N W_i(x)\varepsilon_i\right| \geq t\right) \leq \frac{\mathbb{E}\left[\left(\sum_{i=1}^N W_i(x)\varepsilon_i\right)^2\right]}{t^2}\,,$$

and the assumption on the conditional variance of ε implies that

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^N W_i(x)\varepsilon_i\right)^2\right] &= \sum_{i,j=1}^N \mathbb{E}\left[W_i(x)W_j(x)\mathbb{E}\left[\varepsilon_i\varepsilon_j \,\middle|\, X_1,\ldots,X_n\right]\right] \\ &= \sum_{i=1}^N \mathbb{E}\left[W_i(x)^2\mathbb{E}\left[\varepsilon_i^2 \,\middle|\, X_i\right]\right] \\ &\leq \sigma^2\mathbb{E}\left[\sum_{i=1}^N W_i(x)^2\right]\,,\end{aligned}$$

which concludes the proof of (7).

8.3. **Proof of Lemma 3.** We have

$$\mathbb{E}\left[D_{(i)}(x)\right] = \int_0^D \mathbb{P}\left(D_{(i)}(x) > \varepsilon\right)\mathrm{d}\varepsilon \leq a + \int_a^D \mathbb{P}\left(D_{(i)}(x) > \varepsilon\right)\mathrm{d}\varepsilon\,,$$

for some $a\geq 0$ to be specified later. Observe that $D_{(i)}(x)>\varepsilon$ if and only if there are strictly less than i observations in $\mathcal{B}(x,\varepsilon)$. Since the number of observations in $\mathcal{B}(x,\varepsilon)$ is distributed as a Binomial random variable with parameters N and $\mu\left(\mathcal{B}(x,\varepsilon)\right)\geq\rho\varepsilon^d$, we have

$$\mathbb{P}\left(D_{(i)}(x) > \varepsilon\right) \leq \sum_{j=0}^{i-1}\binom{N}{j}(\rho\varepsilon^d)^j(1-\rho\varepsilon^d)^{N-j}\,. \quad (26)$$

Applying Biau et al. [5], Lemma 3.1, gives, for all $p \in [0,1]$,

$$\sum_{j=0}^{i-1} \binom{N}{j} p^j (1-p)^{N-j} \leq \frac{i}{p(N+1)} \,.$$

Hence,

$$\mathbb{E}\left[D_{(i)}(x)\right] \leq a + \frac{i}{N+1} \int_a^D \frac{1}{\rho \varepsilon^d} \mathrm{d}\varepsilon \,.$$

For $d \geq 2$, we obtain

$$\mathbb{E}\left[D_{(i)}(x)\right] \leq a + \frac{i}{\rho(N+1)} \cdot \frac{a^{1-d}}{d-1} \leq a \left(1 + \frac{i a^{-d}}{\rho(N+1)}\right) \,.$$

Taking $a = \left(\frac{i}{\rho(N+1)}\right)^{1/d}$, we get

$$\mathbb{E}\left[D_{(i)}(x)\right] \leq 2 \left(\frac{i}{\rho(N+1)}\right)^{1/d} \,.$$

For $d = 1$, we set $a = 0$ and use (26) to deduce that

$$\begin{aligned}
\mathbb{E}\left[D_{(i)}(x)\right] &\leq \int_0^D \sum_{j=0}^{i-1} \binom{N}{j} (\rho\varepsilon)^j (1-\rho\varepsilon)^{N-j} \mathrm{d}\varepsilon \\
&= \frac{1}{\rho} \sum_{j=0}^{i-1} \binom{N}{j} \int_0^{\rho D} u^j (1-u)^{N-j} \mathrm{d}u \\
&\leq \frac{1}{\rho} \sum_{j=0}^{i-1} \binom{N}{j} \int_0^{1} u^j (1-u)^{N-j} \mathrm{d}u \,.
\end{aligned}$$

Recognizing the Beta function

$$\int_0^1 u^j (1-u)^{N-j} \mathrm{d}u = B(j+1, N-j+1) = \frac{j!(N-j)!}{(N+1)!} \,,$$

we obtain

$$\mathbb{E}\left[D_{(i)}(x)\right] \leq \frac{1}{\rho} \sum_{j=0}^{i-1} \binom{N}{j} \frac{j!(N-j)!}{(N+1)!} = \frac{i}{\rho(N+1)} \,.$$

8.4. **Proof of Lemma 5.** Our objective here is to prove that, for all $d \geq 1$, and for all $k \in [\![1, N]\!]$,

$$\sum_{i=1}^{N} v_i \left(\frac{i}{N+1}\right)^{1/d} \leq 2 e k^{-1/d} \,,$$

where

$$v_i = \left(1 - \frac{i-1}{N}\right)^k - \left(1 - \frac{i}{N}\right)^k \,.$$

If $d = 1$, we can simply apply Proposition 2.3 in Biau et al. [5] to deduce

$$\sum_{i=1}^{N} v_i \left(\frac{i}{N+1}\right) \leq \frac{2}{k} \left(1 + \frac{1}{N}\right)^k \leq \frac{2e}{k} \,.$$

If $d \geq 2$, since

$$v_i = \sum_{j=1}^{k} \binom{k}{j} \frac{1}{N^j} \left(1 - \frac{i}{N}\right)^{k-j},$$

we may write

$$\begin{aligned}
\sum_{i=1}^{N} v_i \left(\frac{i}{N+1}\right)^{1/d} &\leq \sum_{i=1}^{N} v_i \left(\frac{i}{N}\right)^{1/d} \\
&= \sum_{i=1}^{N} \left(\sum_{j=1}^{k} \binom{k}{j} \frac{1}{N^j} \left(1 - \frac{i}{N}\right)^{k-j} \right) \left(\frac{i}{N}\right)^{1/d} \\
&= \sum_{j=1}^{k} \binom{k}{j} \frac{1}{N^{j-1}} \left(\frac{1}{N} \sum_{i=1}^{N} \left(\frac{i}{N}\right)^{1/d} \left(1 - \frac{i}{N}\right)^{k-j} \right).
\end{aligned}$$

In the latter, we can compare the Riemann sum to the associated integral thanks to the next inequality (see Section 8.5 for its justification).

Lemma 11. *For all $d \geq 2$ and $1 \leq j \leq k \leq N$, we have*

$$\frac{1}{N} \sum_{i=1}^{N} \left(\frac{i}{N}\right)^{1/d} \left(1 - \frac{i}{N}\right)^{k-j} \leq 2 \int_0^1 x^{1/d} (1-x)^{k-j} \mathrm{d}x = 2B(1/d + 1, k - j + 1).$$

This yields

$$\begin{aligned}
\sum_{i=1}^{N} v_i \left(\frac{i}{N+1}\right)^{1/d} &\leq 2 \sum_{j=1}^{k} \binom{k}{j} \frac{1}{N^{j-1}} B(1/d + 1, k - j + 1) \\
&= 2 \sum_{j=1}^{k} \binom{k}{j} \frac{1}{N^{j-1}} \frac{\Gamma(1/d + 1)(k-j)!}{\Gamma(1/d + k - j + 2)}.
\end{aligned}$$

Next, notice that

$$\Gamma(1/d + k - j + 2) = \Gamma(1/d + 1)(k - j + 1)! \prod_{\ell=1}^{k-j+1} \left(1 + \frac{1}{d\ell}\right),$$

so that

$$\begin{aligned}
\sum_{i=1}^{N} v_i \left(\frac{i}{N+1}\right)^{1/d} &\leq 2 \sum_{j=1}^{k} \binom{k}{j} \frac{1}{N^{j-1}(k-j+1) \prod_{\ell=1}^{k-j+1} \left(1 + \frac{1}{d\ell}\right)} \\
&= 2 \sum_{j=1}^{k} \binom{k}{j-1} \frac{1}{N^{j-1} j \prod_{\ell=1}^{k-j+1} \left(1 + \frac{1}{d\ell}\right)} \\
&= 2 \sum_{j=0}^{k-1} \binom{k}{j} \frac{1}{N^{j}(j+1) \prod_{\ell=1}^{k-j} \left(1 + \frac{1}{d\ell}\right)}.
\end{aligned}$$

To go further, we need the following technical result (see Section 8.6 for the proof).

Lemma 12. *For all $d \geq 2$, $k \geq 1$, and $j \in [\![0, k-1]\!]$, we have*

$$\frac{1}{(j+1) \prod_{\ell=1}^{k-j} \left(1 + \frac{1}{d\ell}\right)} \leq k^{-1/d}.$$

Accordingly, we get

$$\sum_{i=1}^{N} v_i \left(\frac{i}{N+1}\right)^{1/d} \leq 2k^{-1/d} \sum_{j=0}^{k-1} \binom{k}{j} \frac{1}{N^j}\,,$$

so that, finally, for all $d \geq 2$, we are led to

$$\sum_{i=1}^{N} v_i \left(\frac{i}{N+1}\right)^{1/d} \leq 2\left(1+\frac{1}{N}\right)^{k} k^{-1/d} \leq 2ek^{-1/d}\,,$$

and the proof of Lemma 5 is complete.

8.5. **Proof of Lemma 11.** We have to show that for all $d \geq 2$ and $1 \leq j \leq k \leq N$,we have

$$\frac{1}{N}\sum_{i=1}^{N} \left(\frac{i}{N}\right)^{1/d} \left(1-\frac{i}{N}\right)^{k-j} \leq 2\int_0^1 x^{1/d}(1-x)^{k-j}\mathrm{d}x\,.$$

To lighten the notation, set $m := k-j$ so that $0 \leq m \leq N-1$, and we want to establish that

$$\frac{1}{N}\sum_{i=1}^{N} \left(\frac{i}{N}\right)^{1/d} \left(1-\frac{i}{N}\right)^{m} \leq 2\int_0^1 x^{1/d}(1-x)^{m}\mathrm{d}x = 2B(1+1/d, 1+m)\,.$$

If $m = 0$, since the mapping $x \mapsto x^{1/d}$ is increasing, we get

$$\frac{1}{N}\sum_{i=1}^{N} \left(\frac{i}{N}\right)^{1/d} \leq \frac{1}{N} + \int_0^1 x^{1/d}\mathrm{d}x = \frac{1}{N} + \frac{d}{d+1} \leq 2\frac{d}{d+1} = 2\int_0^1 x^{1/d}\mathrm{d}x\,.$$

Therefore, from now on we can safely assume that $d \geq 2$ and $1 \leq m \leq N-1$. Let us denote $\varphi(x) = x^{1/d}(1-x)^m$ for $0 \leq x \leq 1$, and notice that if we set $x^\star := (1+md)^{-1}$, the function φ is increasing on $[0, x^\star]$ and decreasing on $[x^\star, 1]$. With this in mind, it is readily seen that

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \varphi(i/N) - \int_0^1 \varphi(x)\mathrm{d}x &= \sum_{i=1}^{N}\int_{\frac{i-1}{N}}^{\frac{i}{N}} (\varphi(i/N)-\varphi(x))\mathrm{d}x = \sum_{i=1}^{N}\int_{\frac{i-1}{N}}^{\frac{i}{N}} \left(\int_x^{\frac{i}{N}} \varphi'(y)\mathrm{d}y\right)\mathrm{d}x \\
&= \sum_{i=1}^{N}\int_{\frac{i-1}{N}}^{\frac{i}{N}} \left(\int_{\frac{i-1}{N}}^{y} \varphi'(y)\mathrm{d}x\right)\mathrm{d}y = \sum_{i=1}^{N}\int_{\frac{i-1}{N}}^{\frac{i}{N}} \left(y-\frac{i-1}{N}\right)\varphi'(y)\mathrm{d}y \\
&\leq \sum_{i=1}^{N}\int_{\frac{i-1}{N}}^{\frac{i}{N}} \left(y-\frac{i-1}{N}\right)\min(0,\varphi'(y))\mathrm{d}y \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\int_{\frac{i-1}{N}}^{\frac{i}{N}} \min(0,\varphi'(y))dy = \frac{\varphi(x^\star)}{N}\,.
\end{aligned}$$

Consequently, taking into account that $N \geq m+1$, it suffices to show that

$$\frac{\varphi(x^\star)}{m+1} \leq B(1+1/d, 1+m)\,.$$

For this, we write

$$B(1+1/d,1+m) = \frac{\Gamma(1+1/d)\Gamma(m+1)}{\Gamma(m+1/d+2)} = \frac{\Gamma(1+1/d)\Gamma(m+1)}{(m+1/d+1)(m+1/d)\Gamma(m+1/d)}\,.$$

Since $d \geq 2$, Gautschi's inequality then implies that

$$\frac{\Gamma(m+1/d)}{\Gamma(m+1)} \leq m^{1/d-1}\,.$$

Since

$$\varphi(x^\star) = \frac{(1/d)^{1/d} m^m}{(m+1/d)^{m+1/d}},$$

we deduce that

$$\frac{\varphi(x^\star)}{(m+1)B(1+1/d, 1+m)} \leq \frac{(1/d)^{1/d}(1+1/m+1/(md))}{\Gamma(1+1/d)(1+1/m)(1+1/(md))^{m+1/d-1}}.$$

To conclude the analysis, we distinguish between two cases. First, if $m \geq 2$, then

$$(1+1/(md))^{m+1/d-1} \geq 1+1/(md),$$

and remarking that

$$(1+1/m)(1+1/(md))^{m+1/d-1} \geq (1+1/m)(1+1/(md)) \geq 1+1/m+1/(md),$$

we get

$$\frac{\varphi(x^\star)}{(m+1)B(1+1/d, 1+m)} \leq \frac{(1/d)^{1/d}}{\Gamma(1+1/d)} =: \psi(1/d).$$

and it is easy to check numerically that, for all $u \in (0, 1/2]$, one has $\psi(u) \leq 1$, hence the desired inequality when $m \geq 2$. Finally, when $m = 1$, we have

$$\frac{\varphi(x^\star)}{(m+1)B(1+1/d, 1+m)} = \frac{\varphi(x^\star)}{2B(1+1/d, 2)} \leq \frac{(1/d)^{1/d}(2+1/d)}{2\Gamma(1+1/d)(1+1/d)^{1/d}} =: \phi(1/d),$$

and one can again numerically verify that, for all $u \in (0, 1/2]$, one has $\phi(u) \leq 1$. This concludes the proof.

8.6. **Proof of Lemma 12.** If we set $p = j$ and $q = k - j$, this amounts to show that, for all $p \geq 0$ and $q \geq 1$,

$$\frac{1}{(p+1)\prod_{\ell=1}^{q}\left(1+\frac{1}{d\ell}\right)} \leq (p+q)^{-1/d},$$

or, equivalently,

$$\varphi_q(p) := \frac{1}{d}\ln(p+q) \leq \ln(p+1) + \sum_{\ell=1}^{q} \ln\left(1+\frac{1}{d\ell}\right) =: \psi_q(p).$$

In this aim, note that

$$0 \leq \varphi_q'(p) = \frac{1}{d(p+q)} \leq \frac{1}{p+1} = \psi_q'(p),$$

and $\varphi_q(0) := \frac{1}{d}\ln(q)$ while

$$\psi_q(0) = \sum_{\ell=1}^{q} \ln\left(1+\frac{1}{d\ell}\right).$$

Thus, the lemma holds true if and only if, for all $d \geq 2$ and $q \geq 1$,

$$\frac{1}{d}\ln(q) \leq \sum_{\ell=1}^{q} \ln\left(1+\frac{1}{d\ell}\right).$$

Taking into account that $\ln(1+x) \geq x - x^2/2$ for all $x \geq 0$, we may write

$$\sum_{\ell=1}^{q} \ln\left(1+\frac{1}{d\ell}\right) \geq \frac{1}{d}\sum_{\ell=1}^{q}\frac{1}{\ell} - \frac{1}{2d^2}\sum_{\ell=1}^{q}\frac{1}{\ell^2}.$$

Now, if γ stands for Euler's constant, we know that

$$\sum_{\ell=1}^{q} \frac{1}{\ell} \geq \ln(q) + \gamma\,,$$

whereas

$$\sum_{\ell=1}^{q} \frac{1}{\ell^2} \leq \frac{\pi^2}{6}\,.$$

Putting all pieces together, we have shown that

$$\sum_{\ell=1}^{q} \ln\left(1 + \frac{1}{d\ell}\right) - \frac{1}{d}\ln(q) \geq \frac{1}{d}\left(\gamma - \frac{\pi^2}{12d}\right).$$

Since the right-hand side is positive for all $d \geq 2$, the proof is complete.

Acknowledgments. The authors would like to thank Julien Reygner for his help with the proof of Lemma 11, as well as Matthieu Lerasle for pointing out several references and for relevant comments on a first version of the article.

References

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29, 1996. 3

[2] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011. 4

[3] G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518, 2010. 8

[4] G. Biau and L. Devroye. *Lectures on the nearest neighbor method*. Springer Series in the Data Sciences. Springer, 2015. 6

[5] G. Biau, F. Cérou, and A. Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research (JMLR)*, 11:687–712, 2010. 7, 8, 21

[6] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. 8

[7] C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015. 4

[8] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013. 4

[9] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'IHP Probabilités et statistiques*, 48(4):1148–1185, 2012. 3

[10] F. Cérou and A. Guyader. Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics*, 10:340–355, 2006. 10

[11] K. Chidananda Gowda and G. Krishna. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern recognition*, 10(2):105–112, 1978. 10

[12] K. Chidananda Gowda and G. Krishna. The condensed nearest neighbor rule using the concept of mutual nearest neighborhood (corresp.). *IEEE Transactions on Information Theory*, 25(4):488–490, 1979. 10

[13] A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, 25(6):2300–2312, 1997. 4

[14] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016. 3, 4, 19

[15] E. Gobet, M. Lerasle, and D. Métivier. Mean estimation for Randomized Quasi Monte Carlo method. working paper or preprint, 2022. URL `https://hal.science/hal-03631879`. 3

[16] A. Guyader and N. Hengartner. On the mutual nearest neighbors estimate in regression. *Journal of Machine Learning Research (JMLR)*, 14:2361–2376, 2013. 10

[17] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002. 4, 8, 11, 12, 16, 17, 18

[18] P. Hall and R. J. Samworth. Properties of bagged nearest neighbour classifiers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):363–379, 2005. 8

[19] D. Hsu and S. Sabato. Heavy-tailed regression with a generalized median-of-means. In *International Conference on Machine Learning*, pages 37–45. PMLR, 2014. 3

[20] P. Humbert, B. Le Bars, and L. Minvielle. Robust kernel density estimation with median-of-means principle. In *International Conference on Machine Learning*, pages 9444–9465. PMLR, 2022. 4

[21] H. Jégou, C. Schmid, H. Harzallah, and J. Verbeek. Accurate image search using the contextual dissimilarity measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):2–11, 2008. 10

[22] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical computer science*, 43:169–188, 1986. 3

[23] H. Jiang. Non-asymptotic uniform rates of consistency for k-nn regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3999–4006, 2019. 4, 8

[24] A. P. Korostelev and A. B. Tsybakov. *Minimax theory of image reconstruction*, volume 82. Springer Science & Business Media, 2012. 4

[25] G. Lecué and M. Lerasle. Learning from MOM's principles: Le Cam's approach. *Stochastic Processes and their Applications*, 129(11):4385–4410, 2019. 4, 15

[26] G. Lecué and M. Lerasle. Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics*, 48(2):906–931, 2020. 4

[27] G. Lecué, M. Lerasle, and T. Mathieu. Robust classification via MOM minimization. *Machine Learning*, 109(8):1635–1665, 2020. 4

[28] J. C. Lee and P. Valiant. Optimal sub-Gaussian mean estimation in $\mathbb{R}$. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 672–683. IEEE, 2022. 3

[29] M. Lerasle. Lecture notes: Selected topics on robust statistical learning theory. *arXiv preprint arXiv:1908.10761*, 2019. 4

[30] M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011. 3, 4

[31] E. Liitiäinen, F. Corona, and A. Lendasse. Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis*, 101(4):811–823, 2010. 7

[32] H. Liu, S. Zhang, J. Zhao, X. Zhao, and Y. Mo. A new classification algorithm using mutual nearest neighbors. In *2010 Ninth International Conference on Grid and Cloud Computing*, pages 52–57. IEEE, 2010. 10

[33] G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019. 4

[34] G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3):925–965, 2019. 4

[35] G. Lugosi and S. Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019. 3

[36] S. Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21 (4):2308–2335, 2015. 3

[37] S. Minsker. Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics*, 13(2):5213–5252, 2019. 3

[38] S. Minsker and M. Ndaoud. Robust and efficient mean estimation: an approach based on the properties of self-normalized sums. *Electronic Journal of Statistics*, 15(2):6036–6070, 2021. 3

[39] M. D. Penrose and J. Yukich. Laws of large numbers and nearest neighbor distances. In *Advances in directional and linear statistics*, pages 189–199. Physica-Verlag/Springer, Heidelberg, 2011. 4

[40] G. Pólya. Remarks on computing the probability integral in one and two dimensions. In *Proceedings of the 1st Berkeley symposium on mathematical statistics and probability*, pages 63–78, 1945. 18

[41] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. Van Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *CVPR 2011*, pages 777–784. IEEE, 2011. 10

[42] R. J. Samworth. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012. 6

[43] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. 4, 16