ON DEVIATION PROBABILITIES IN NON-PARAMETRIC REGRESSION

ANNA BEN-HAMOU* AND ARNAUD GUYADER‡

Abstract. This paper is devoted to the problem of determining the concentration bounds that are achievable in non-parametric regression. We consider the setting where features are supported on a bounded subset of $\mathbb{R}^d$, the regression function is Lipschitz, and the noise is only assumed to have a finite second moment. We first specify the fundamental limits of the problem by establishing a general lower bound on deviation probabilities, and then construct explicit estimators that achieve this bound. These estimators are obtained by applying the median-of-means principle to classical local averaging rules in non-parametric regression, including nearest neighbors and kernel procedures.

Contents

1. Introduction

1.1. Setting and main results.

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

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

The pair (X, Y) can be written as

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

where the random variable ε, called the noise, satisfies $\mathbb{E}[\varepsilon \,|\, X] = 0$. 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 $\widehat{r}_n$ of the function r.

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. 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.

In this setting, it is known that the minimax rate of convergence for the L_2 risk $\mathbb{E}[(\widehat{r}_n(X) - r(X))^2]$ is given by

$$\left(\frac{\sigma^2}{\rho n}\right)^{\frac{2}{d+2}}\,,$$

up to some constants depending on d only, see [36] and [13], Chapter 3. In particular, this rate is achieved by nearest neighbors or kernel procedures. Let us note that usually, this minimax rate is written with D^{-d} instead of ρ, where D is the diameter of S. However, those two quantities are clearly related by the inequality $\rho \leq D^{-d}$, which comes from (1) applied to $\varepsilon = D$. Moreover, one may check that lower bounds for the L_2 risk are actually obtained in situations where equation (1) is satisfied, namely with X uniform on a d-dimensional hypercube.

The main goal of this paper is to obtain such minimax results for deviation probabilities rather than for the L_2 risk. More precisely, we start by establishing the following lower bound.

Theorem 1. *For any $\delta \in]0, 1/16]$, for any $\rho > 0$ and $\sigma > 0$, for any regression estimate $\widehat{r}_n$, there exists a distribution $(X, Y) \in \mathcal{F}_{\rho,\sigma}$ as defined in Assumption 1 such that, when $X \sim \mu$ independent of $\mathcal{D}_n$, we have*

$$\mathbb{P}\left(|\widehat{r}_n(X) - r(X)| \geq a_d \left(\frac{\sigma^2}{\rho n} \ln\left(\frac{1}{16\delta}\right)\right)^{\frac{1}{d+2}}\right) \leq \delta\,,$$

where $a_d > 0$ is an explicit constant depending on the dimension d only.

Next, we proceed by designing estimators that achieve such deviation bounds for any distribution in $\mathcal{F}_{\rho,\sigma}$. In particular, we are looking for exponential concentration with only a second moment assumption on the noise.

Let us recall that 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.

As will be shown, given a local averaging rule, an estimator that nearly achieves the bound of Theorem 1 may be constructed through 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

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

For a variety of base estimates, including nearest neighbors and kernel estimates, we show that, when $(X, Y) \in \mathcal{F}_{\rho,\sigma}$, 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, for all $x \in S$,

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

where $b_d > 0$ is an explicit constant possibly depending on d, $c_{\mathcal{F}} > 0$ is an explicit constant that depends on ρ, σ and d only, and both b_d and $c_{\mathcal{F}}$ depend on the base estimate at stake. Specifically, this is the purpose of Propositions 6 and 7 for nearest neighbors estimates, Proposition 8 for kernel estimates, and Proposition 9 for partitioning estimates. 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

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

which, in view of Theorem 1, is the optimal concentration bound.

1.2. **Related work.** To our knowledge, there are only very few results on deviation probabilities in non-parametric regression, at the exception of [18] for the k-nn estimate. However, in the latter, the noise is assumed sub-Gaussian. Therefore, the first contribution of the present paper is to delineate the fundamental limits of the problem by providing a general lower bound when only a second moment assumption on the noise is made (see Theorem 1). Our second contribution is to show that this bound can in fact be achieved by combining local averaging rules and the MoM principle.

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 [13] for, respectively, partitioning, kernel, and nearest neighbors estimates. Concerning (1), the proof of Theorem 1 in [33] (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 [18]. Comparable assumptions are the so-called cone-condition in [19], Chapter 5, and the notion of standard support in [10]. 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 2).

Concerning the MoM principle, it seems that it was first introduced in works of [17], [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 [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 [23], [32], or [12]. One caveat of the MoM-estimator of the mean is its dependence on the confidence threshold δ. However, under stronger assumptions on the distribution, [31] showed that it is in fact adaptive to δ up to $\delta \approx e^{-\sqrt{n}}$. In the same vein, [11] 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 [30], [15], [25], [29], and applied to a large variety of statistical problems, including linear regression ([2]), empirical risk minimization ([7], [20], [28]), classification ([22]), bandits ([8]), least-squares density estimation ([25]), and kernel density estimation ([16]). We refer the reader to [21], [24], or [27] for more references on all of these subjects.

Finally, let us note that the minimax deviation inequalities (2) are 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 [13], 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 is devoted to the proof of Theorem 1. Then Section 3 provides a guideline for proving (2). It also contains two general observations about our estimators: the fact that they are robust to the presence of outliers in the sample, and the fact that they can be turned into δ-independent estimators, using a strategy introduced by [11]. In the next two sections, Equation (2) is established for various choices of local averaging procedures, namely nearest neighbor methods in Section 4, and kernel and partitioning methods in Section 5. Let us point out that, for partitioning estimates (Section 5.2), we are able to obtain a uniform control on $\sup_{x\in S} |\widehat{r}_n^{\mathsf{mom}}(x) - r(x)|$. Finally, Section 6 gathers the proofs of several technical results.

2. Lower bound

In order to establish Theorem 1, we consider the following setting. For some $A > 0$, let

$$S = [0\,,\,A]^d\,,$$

and consider the model

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

where $X \sim \mu = \mathrm{Unif}(S)$, where the noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is independent of X, and where 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 $D = A\sqrt{d}$ and, for all $x \in S$ and $\varepsilon \in (0, D]$, one has

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

and any value of $\rho > 0$ can be achieved by adjusting the sidelength A.

The starting point of the proof uses the same idea as in the proof of Theorem 3.2 in [13] (see also Section 2 in [36]). Namely, let $\mathcal{C} := \left[-\frac{1}{2}, \frac{1}{2}\right]^d$ and $\partial\mathcal{C}$ its frontier, then define the 1-Lipschitz function

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

Let $M \geq 1$ be some integer to be specified later and consider a partition of S by $K := M^d$ hypercubes A_j of sidelength $h := A/M$ and with centers a_j, and let the functions $g_1, \ldots, g_K$ be defined by

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

We then look for a "bad" regression function among the functions

$$r^{(c)} = \sum_{j=1}^{K} c_j g_j\,,\ c \in \{-1, 1\}^K\,,$$

which are all 1-Lipschitz by [13]. In this setting, we are led to the following result, which is proved in Section 6.1.

Proposition 2. *For any $\delta \in]0, 1/16]$, for any regression estimate $\widehat{r}_n$, there exists $c \in \{-1,1\}^K$ such that, when $X \sim \mu$ independent of $\mathcal{D}_n$, we have*

$$\mathbb{P}\left(\left|\widehat{r}_n(X) - r^{(c)}(X)\right| \geq \frac{1-2^{-\frac{1}{d}}}{2}\left(\frac{\pi(d+1)^2\sigma^2A^d}{n}\ln\left(\frac{1}{16\delta}\right)\right)^{\frac{1}{d+2}}\right) \geq \delta\,. \tag{3}$$

Since $A^d = \frac{\pi^{d/2}}{2^d d^{d/2}\Gamma(1+d/2)}\rho^{-1}$, this establishes Theorem 1 with

$$a_d = \frac{1-2^{-\frac{1}{d}}}{2}\left(\frac{\pi^{1+d/2}(d+1)^2}{2^d d^{d/2}\Gamma(1+d/2)}\right)^{\frac{1}{d+2}}\,.$$

3. Median-of-means versions of local averaging procedures

3.1. **Two key lemmas.**

This section exposes two generic results that will be of constant use for proving upper bounds. 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 3. *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

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

Assume now that the base estimate takes the form

$$\forall x \in \mathbb{R}^d \,, \; \widehat{r}_N(x) = \widehat{r}_N(x, \mathcal{D}_N) = \sum_{i=1}^{N} W_i(x) Y_i \,, \tag{4}$$

where for all $i \in [\![1, N]\!]$, $W_i(x)$ is a Borel measurable function of x and $X_1, \dots, X_N$ (but not of $Y_1, \dots, Y_N$), 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 4. *Suppose that (ii) and (iii) in Assumption 1 are satisfied. Then, for all $x \in \mathbb{R}^d$, we have*

$$|\widehat{r}_N(x) - r(x)| \leq \left|\sum_{i=1}^{N} W_i(x)\varepsilon_i\right| + \sum_{i=1}^{N} W_i(x)\|X_i - x\| \,, \tag{5}$$

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

$$p_{t+s}(x) \leq \frac{\sigma^2 \mathbb{E}\left[\sum_{i=1}^N W_i(x)^2\right]}{t^2} + \mathbb{P}\left(\sum_{i=1}^{N} W_i(x)\|X_i - x\| \geq s\right) \,, \tag{6}$$

with $p_{t+s}(x)$ as defined in Lemma 3.

In Sections 4 and 5, we will investigate several instances of local averaging procedures for the base estimate $\widehat{r}_N$. In each case, we first use Lemma 4 in order to determine t and s such that $p_{t+s}(x) \leq \frac{1}{4e^2}$. Lemma 3 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.

3.2. **Robustness.**

In addition to exhibiting concentration properties, the estimate $\widehat{r}_n^{\mathsf{mom}}$ also stands out through its strong robustness to outliers. In this section, we consider the contamination scheme introduced by [20]. 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 $\widehat{r}_N$. Assume that the number $|\mathcal{O}|$ of outliers in the original sample is stricly less than $m/2$. For simplicity, we here further assume that $|\mathcal{O}| \leq m/4$. Then, letting $\mathcal{B}$ be the set of blocks that do not intersect $\mathcal{O}$, we have, for all $s, t > 0$,

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

Observing that $|\mathcal{B}^c| \leq |\mathcal{O}| \leq m/4$, we get, with a slight abuse of notation,

$$\mathbb{P}_{\mathcal{O}\cup\mathcal{I}}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \geq t+s\right) \leq \mathbb{P}\left(\sum_{j=1}^{m} \mathbf{1}_{\left\{\left|\widehat{r}^{(j)}(x)-r(x)\right|\geq t+s\right\}} \geq \frac{m}{4}\right).$$

Now, the proof of Lemma 3 reveals that

$$\mathbb{P}\left(\sum_{j=1}^{m} \mathbf{1}_{\left\{\left|\widehat{r}^{(j)}(x)-r(x)\right|\geq t+s\right\}} \geq \frac{m}{4}\right) \leq \left(\frac{4}{3^{3/4}}\right)^m p_{t+s}(x)^{m/4},$$

which is less than e^{-m} for $p_{t+s}(x) \leq \frac{27}{(4e)^4}$. Hence, provided $m \geq 4|\mathcal{O}|$ and modulo some minor changes in the constants, the same strategy as the one described just after the statement of Lemma 4 can be applied to the contaminated setting.

Remark 1. *More generally, define ε as the proportion of outliers in the original sample and assume that $m > 2|\mathcal{O}| = 2\varepsilon n$, then a straightforward adaptation of the previous reasoning shows that*

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

which is less than e^{-m} for $p_{t+s}(x) \leq (2e)^{-\frac{2m}{m-2\varepsilon n}}$.

3.3. 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 [11], 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 $\delta \geq \frac{e^{-m}}{1-e^{-1}}$, namely $m_\delta = \left\lceil \ln\left(\frac{1}{(1-e^{-1})\delta}\right)\right\rceil$. Then $\widetilde{r}_n(x)$ satisfies

$$\mathbb{P}\left(|\widetilde{r}_n(x) - r(x)| > |\widehat{I}_{m_\delta}|\right) \leq \delta. \tag{7}$$

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 (7).

4. 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 $(U_1,\ldots,U_N)$ i.i.d. with uniform law on $[0,1]$ and sorting them. The weighted nearest neighbors estimate is defined as

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

where $(v_1,\ldots,v_N)$ is a deterministic vector in $[0,1]^N$ satisfying $\sum_{i=1}^N v_i=1$. Note that this estimate is of the form (4), 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 [4], Chapter 8, or [35] and references therein for more information on this topic.

In this context, the variance term of Lemma 4 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

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

The following lemma, whose proof is housed in Section 6, allows us to control the expected nearest neighbor distances (see Remark 2 below for a comment on this result).

Lemma 5. *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 4 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 3, 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 (Subsection 4.1) and the bagged 1-nearest neighbor estimate (Section 4.2). As we will see, both satisfy the concentration inequality (2).

Remark 2. *The fact that the upper bound of Lemma 5 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 [5], Corollary 2.1, for $d = 1$ or $d \geq 3$ and a consequence of [26], 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 $\mathbb{E}\left[D_{(i)}(X)^2\right] \leq c_1 \frac{i}{N}$.

4.1. 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 $k \in [\![1, N]\!]$. We then have the following proposition, whose proof can be found in Section 6.3.

Proposition 6. *Under Assumption 1, let*

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

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

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

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

Remark 3. *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 [13] Theorem 6.2. In a different setting, the conclusion is the same in the work of Jiang, see Remark 1 in [18].*

4.2. **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 [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 [14], [3], [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 [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

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

Here we focus on sampling with replacement and get the following bound, whose proof can be found in Section 6.3.

Proposition 7. *Under Assumption 1, let*

$$c = \rho \left(\frac{\sigma}{2e\sqrt{2}}\right)^d \wedge \frac{8e^2}{\sigma^2 \rho^{2/d}} , \qquad \delta \in [e^{-\lfloor cn \rfloor}, 1[, \qquad \text{and} \qquad m = \lceil \ln(1/\delta) \rceil ,$$

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

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

satisfies

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

Hence, for $k^\star$-bagged 1-nn base estimate, inequality (2) still holds, with this time $b_d = 64e^3$ and $c_{\mathcal{F}} = c$.

5. Kernel and partitioning estimation

5.1. **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 (4) with

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

Proposition 8. *Under Assumption 1, let*

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

ensuring that $1 \le m \le cn$. Then the estimator $\widehat{r}_n^{\mathsf{mom}}$ constructed on m blocks with $h^\star$-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 $b_d = 4e^{2/3}$ and $c_{\mathcal{F}} = c$. The proof can be found in Section 6.4.

5.2. **Partitioning estimates.** To simplify the presentation, let us here assume that $S = [0,1]^d$. For some integer $K \ge 1$, 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 upcoming result shows that inequality (2) is then fulfilled with $b_d = 16e\sqrt{d}$ and $c_{\mathcal{F}} = c$.

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

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

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

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

One enjoyable feature of partitioning estimates is that uniform bounds for $x \in S$ can easily be obtained.

Proposition 10. *Under Assumption 1 with $S = [0,1]^d$, let*

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

ensuring that $1 \leq m \leq cn$. 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

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

satisfies

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

In particular, we see that if $m = m_n$ satisfies $m = o(n)$ and e^{-m} is summable (*e.g.*, $m = (\log n)^2$), then Borel–Cantelli Lemma entails that

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

6. Proofs

6.1. **Proofs from Section 2.**

Proof of Proposition 2. Recall that for $\mathcal{C} := \left[-\frac{1}{2}, \frac{1}{2}\right]^d$ and $\partial\mathcal{C}$ its frontier, the function g is defined by

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

This function is 1-Lipschitz and one can check that

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

Next, for an integer $M \geq 1$ to be specified later, consider a partition of S by $K := M^d$ hypercubes A_j of sidelength $h := A/M$ and with centers a_j, and let the functions $g_1, \dots, g_K$ 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

$$\int g_j(x)^2 \mathrm{d}x = \frac{h^{d+2}}{2(d+1)(d+2)} . \tag{14}$$

We then restrict to regression functions of the form

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

Now let $\widehat{r}_n$ denote any 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

$$\left|\widehat{r}_n(x) - r^{(c)}(x)\right| \geq \frac{1}{2}\left|\widetilde{r}_n(x) - r^{(c)}(x)\right| . \tag{15}$$

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

$$A'_j := \left[a_j - \frac{h}{2^{1+\frac{1}{d}}}; a_j + \frac{h}{2^{1+\frac{1}{d}}}\right]^d \subset A_j$$

and set

$$\widetilde{c}_j := \begin{cases} +1 & \text{if } \lambda\left(\{x \in A'_j,\, \mathrm{sign}(\widehat{r}_n(x)) = +1\}\right) \geq \frac{|A'_j|}{2} \\ -1 & \text{otherwise} \end{cases}$$

where λ is the Lebesgue measure. In other words, for each hypercube A_j, we take a majority vote, but only on A'_j. In particular, if X is uniform on S and falls in the subset A'_j of a bad hypercube A_j, *i.e.* such that $\widetilde{c}_j \neq c_j$, then

$$\mathbb{P}\left(\left|\widetilde{r}_n(X) - r^{(c)}(X)\right| \geq h\left(1 - 2^{-\frac{1}{d}}\right) \middle| X \in A'_j, \widetilde{c}_j \neq c_j\right) \geq \frac{1}{2}\,.$$

Next, (15) gives

$$\mathbb{P}\left(\left|\widehat{r}_n(X) - r^{(c)}(X)\right| \geq \frac{h}{2}\left(1 - 2^{-\frac{1}{d}}\right)\right) \geq \mathbb{P}\left(\left|\widetilde{r}_n(X) - r^{(c)}(X)\right| \geq h\left(1 - \frac{1}{2^{\frac{1}{d}}}\right)\right)\,.$$

For any event B, the independency between $\widetilde{c}_j$ and X, and the fact that $\mathbb{P}(X \in A'_j) = (2K)^{-1}$ imply

$$\mathbb{P}(B) \geq \sum_{j=1}^{K} \mathbb{P}\left(B \,\middle|\, X \in A'_j, \widetilde{c}_j \neq c_j\right)\mathbb{P}(X \in A'_j)\mathbb{P}(\widetilde{c}_j \neq c_j) = \frac{1}{2K}\sum_{j=1}^{K} \mathbb{P}\left(B \,\middle|\, X \in A'_j, \widetilde{c}_j \neq c_j\right)\mathbb{P}(\widetilde{c}_j \neq c_j)\,.$$

Combining those observations yields

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

We are now left to show that, for some well chosen value of M, we have

$$\sup_{c \in \{-1,1\}^K} \frac{1}{4K}\sum_{j=1}^{K}\mathbb{P}(\widetilde{c}_j \neq c_j) \geq \delta\,.$$

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,

$$\sup_{c \in \{-1,1\}^K} \frac{1}{4K}\sum_{j=1}^{K}\mathbb{P}(\widetilde{c}_j \neq c_j) \geq \frac{1}{4K}\sum_{j=1}^{K}\mathbb{P}(\widetilde{c}_j \neq C_j)\,.$$

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:

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

Hence,

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

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, [13], Lemma 3.2, ensures that

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

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 (14),

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

we are led to

$$\sup_{c\in\{-1,1\}^K} \frac{1}{4K}\sum_{j=1}^K \mathbb{P}(\widetilde{c}_j \neq c_j) \geq \frac{1}{4}\Phi\left(-\frac{1}{\sigma}\sqrt{\frac{nh^{d+2}}{2(d+1)(d+2)A^d}}\right) .$$

We now use the following lower bound on the Gaussian tail (see for instance [34], 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,

$$\sup_{c\in\{-1,1\}^K} \frac{1}{4K}\sum_{j=1}^K \mathbb{P}(\widetilde{c}_j \neq c_j) \geq \frac{1}{16}\exp\left(-\frac{nh^{d+2}}{\pi(d+1)(d+2)\sigma^2A^d}\right) \geq \frac{1}{16}\exp\left(-\frac{nh^{d+2}}{\pi(d+1)^2\sigma^2A^d}\right) .$$

The right-hand side is larger than $\delta \in]0,1/16]$ as soon as

$$h \leq \left(\frac{\pi(d+1)^2\sigma^2A^d}{n}\ln\left(\frac{1}{16\delta}\right)\right)^{\frac{1}{d+2}} .$$

This concludes the proof of Proposition 2. ■

6.2. Proofs from Section 3.

Proof of Lemma 3. By definition of the median, we have

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

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

$$\mathbb{P}\left(\sum_{j=1}^m B_j \geq \ell\right) = \sum_{k=\lceil\ell\rceil}^m \binom{m}{k}p^k(1-p)^{m-k} \leq p^{\lceil\ell\rceil}\sum_{k=\lceil\ell\rceil}^m \binom{m}{k} \leq p^\ell\sum_{k=0}^m \binom{m}{k} = 2^m p^\ell . \quad (16)$$

In particular, taking $\ell = \frac{m}{2}$ gives the desired result in Lemma 3. ■

Proof of Lemma 4. 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 (5). 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

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

which concludes the proof of (6). ■

6.3. Proofs from Section 4.

Proof of Lemma 5. 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

$$\mathbb{P}\left(D_{(i)}(x) > \varepsilon\right) \leq \sum_{j=0}^{i-1} \binom{N}{j} (\rho\varepsilon^d)^j (1 - \rho\varepsilon^d)^{N-j}. \tag{17}$$

Applying [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

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

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 (17) to deduce that

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

Recognizing the Beta function $\int_0^1 u^j(1-u)^{N-j}\mathrm{d}u = \frac{j!(N-j)!}{(N+1)!}$, we obtain

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

■

Proof of Proposition 6. In the k-nn case, (12) becomes

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

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

$$\mathbb{P}\left(|\widehat{r}_n^{\,\mathsf{mom}}(x) - r(x)| \ge 2e\sigma\sqrt{\frac{2}{k}} + 16e^2 \left(\frac{km}{\rho n}\right)^{1/d} \right) \le e^{-m} \,. \tag{18}$$

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

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

i.e.

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

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

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

In view of (18), we have, for the optimal choice $k^\star$,

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

Taking $\delta \in [e^{-\lfloor cn \rfloor}, 1[$ and $m = \lceil \ln(1/\delta) \rceil$, we arrive at the desired result. ■

Proof of Proposition 7. Concerning the variance term in (12), Proposition 2.2 in [5] and the fact that $k \le N$ yield

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

where for the last inequality, we used that $N = \lfloor \frac{n}{m} \rfloor \ge \frac{n}{2m}$. 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 just below.

Lemma 11. *For all $d \ge 1$, one has*

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

In view of (12), we are led to

$$t \le 4e^2\sigma\sqrt{\frac{2km}{n}} \qquad \text{and} \qquad s \le \frac{16e^3}{(\rho k)^{1/d}}\,,$$

and, by Lemma 3, for all $x \in S$,

$$\mathbb{P}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \ge 4e^2\sigma\sqrt{\frac{2km}{n}} + \frac{16e^3}{(\rho k)^{1/d}}\right) \le e^{-m}\,.$$

As for the k-nn case, when σ and ρ are known, the integer k may be chosen as the largest integer such that $\frac{4e}{(\rho k)^{1/d}} \ge \sigma\sqrt{\frac{2km}{n}}$, that is

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

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

$$\frac{m}{n} \le \rho\left(\frac{\sigma}{2e\sqrt{2}}\right)^d \wedge \frac{8e^2}{\sigma^2\rho^{2/d}} = c\,.$$

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

$$\mathbb{P}\left(|\widehat{r}_n^{\mathsf{mom}}(x) - r(x)| \ge 64e^3\left(\frac{\sigma^2 m}{\rho n}\right)^{\frac{1}{d+2}}\right) \le e^{-m}\,.$$

Taking $\delta \in [e^{-\lfloor cn \rfloor}, 1[$ and $m = \lceil \ln(1/\delta) \rceil$ yields the desired result. ■

Proof of Lemma 11. Recall that

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

Since $\sum_{i=1}^N v_i = 1$ and since $x \mapsto x^{1/d}$ is concave on $\mathbb{R}_+$, Jensen's inequality gives

$$\sum_{i=1}^{N} v_i \left(\frac{i}{N+1}\right)^{1/d} \le \left(\sum_{i=1}^{N} v_i \frac{i}{N+1}\right)^{1/d}. \tag{19}$$

Now

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

and by comparing the latter to the associated integral we have

$$\sum_{i=1}^{N}\left(\frac{i}{N}\right)^k \le N\int_{1/N}^{1+1/N} u^k \mathrm{d}u \le \frac{N}{k+1}\left(1+\frac{1}{N}\right)^{k+1} = \frac{N+1}{k+1}\left(1+\frac{1}{N}\right)^{k} \le \frac{N+1}{k} e\,.$$

Coming back to (19), we get

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

■

6.4. **Proofs from Section 5.**

Proof of Proposition 8. By inequality (5), we get

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

In the kernel 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 [13], 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 3 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 $m \leq \frac{\rho D^{d+2} n}{8e^2\sigma^2}$, ensuring $h^\star \leq D$, we have

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

Taking $\delta \in [e^{-\lfloor cn \rfloor}, 1[$ and $m = \lceil \ln(1/\delta) \rceil$ yields the desired result. ■

Proof of Proposition 9. Mutatis mutandis, the reasoning is the same as for kernel estimates. Here again, the bias term can indeed be deterministically bounded:

$$\sum_{i=1}^{N} W_i(x)\|X_i - x\| = \frac{1}{N_k}\sum_{i=1}^{N} \mathbf{1}_{X_i \in A_k}\|X_i - x\| \leq \sqrt{d}K^{-1}\,. \tag{20}$$

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

$$p_{t+\sqrt{d}K^{-1}}(x) \leq \frac{\sigma^2 \mathbb{E}\left[\frac{\mathbf{1}_{N_k>0}}{N_k}\right]}{t^2} \leq \frac{2\sigma^2}{(N+1)\mu(A_k)t^2} \leq \frac{2^{d+1}K^d\sigma^2 m}{\rho n t^2}\,, \tag{21}$$

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 3, 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

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

which yields

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

■

Proof of Proposition 10. For all $t > 0$, we have

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

Then, by inequality (5) of Lemma 4 and the deterministic bound (20) on the bias term, it comes

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

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 $B_{i,k}^{(j)} := \mathbf{1}_{\left\{X_i^{(j)}\in A_k\right\}}$,

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

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 $k \in \llbracket 1, K \rrbracket$, 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 (21). We may now apply inequality (16) in the proof of Lemma 3 to deduce that

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

Choosing t appropriately, we get

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

The end of the proof is then the same as for Proposition 9. Indeed, one may choose K as the largest integer such that $\sqrt{d}K^{-1} \geq e\sqrt{\frac{2^{d+3}\sigma^2 K^{d+\frac{2d}{m}} m}{\rho n}}$, i.e.

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

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

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

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

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

which yields

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

■

Acknowledgments. The authors would like to thank Julien Reygner for fruitful discussions 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. 4

[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. 10

[4] G. Biau and L. Devroye. *Lectures on the nearest neighbor method.* Springer Series in the Data Sciences. Springer, 2015. 8

[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. 9, 10, 15, 17

[6] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. 10

[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. 4

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

[11] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016. 4, 7

[12] 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. 4

[13] 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. 2, 4, 5, 9, 13, 18

[14] 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. 10

[15] 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. 4

[16] 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

[17] 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. 4

[18] 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, 9

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

[20] 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, 6

[21] 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

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

[23] 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. 4

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

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

[26] E. Liitiäinen, F. Corona, and A. Lendasse. Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis*, 101(4):811–823, 2010. 9

[27] 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

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

[29] G. Lugosi and S. Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019. 4

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

[31] S. Minsker. Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics*, 13(2):5213–5252, 2019. 4

[32] 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. 4

[33] 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

[34] 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. 14

[35] R. J. Samworth. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012. 8

[36] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. 2, 5

($\star$) LPSM, Sorbonne University, Paris
Email address, $\star$: anna.ben_hamou@sorbonne-universite.fr

(‡) LPSM, Sorbonne University, Paris
Email address, ‡: arnaud.guyader@sorbonne-universite.fr