A nearest neighbor estimate of the residual variance *

Luc Devroye † László Györfi ‡ Gábor Lugosi§
Harro Walk¶

March 27, 2018

Abstract

We study the problem of estimating the smallest achievable mean-squared error in regression function estimation. The problem is equivalent to estimating the second moment of the regression function of Y on $X \in \mathbb{R}^d$. We introduce a nearest-neighbor-based estimate and obtain a normal limit law for the estimate when X has an absolutely continuous distribution, without any condition on the density. We also compute the asymptotic variance explicitly and derive a non-asymptotic bound on the variance that does not depend on the dimension d. The asymptotic variance does not depend on the smoothness of the density of X or of the regression function. A non-asymptotic exponential concentration inequality is also proved. We illustrate the use of the new estimate through testing whether a component of the vector X carries information for predicting Y.

Key words: regression functional, nearest-neighbor-based estimate, asymptotic normality, concentration inequalities, dimension reduction.

*Luc Devroye was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada. László Györfi was supported by the National University of Public Service under the priority project KÖFOP-2.1.2-VEKOP-15-2016-00001 titled Public Service Development Establishing Good Governance in the Ludovika Workshop. Gábor Lugosi was supported by the Spanish Ministry of Economy and Competitiveness, Grant MTM2015-67304-P and FEDER, EU.

†McGill University, lucdevroye@gmail.com

‡Budapest University of Technology and Economics, gyorfi@cs.bme.hu

§ICREA, Pompeu Fabra University, and Barcelona Graduate School of Economics, gabor.lugosi@upf.edu

¶Universität Stuttgart, harro.walk@t-online.de

1 Introduction

In this paper we study the problem of estimating the smallest achievable mean-squared error in regression function estimation in multivariate problems. We introduce and analyze a nearest neighbor-based estimate of the second moment of the regression function. The second moment of the regression function is closely tied to the best possible achievable mean squared error. It is shown that the estimate is asymptotically normally distributed. It is remarkable that the asymptotic variance only depends on conditional moments of the regression function but not on its smoothness. Moreover, the non-asymptotic variance is bounded by a constant that is independent of the dimension. We also establish a non-asymptotic exponential concentration inequality. We illustrate these results studying variable selection. In particular, we construct and analyze a test for deciding whether a component of the observational vector has predictive power.

The formal setup is as follows. Let (X, Y) be a pair of random variables such that $X = (X^{(1)}, \ldots, X^{(d)})$ takes values in $\mathbb{R}^d$ and Y is a real-valued random variable with $\mathbb{E}[Y^2] < \infty$. We denote by μ the distribution of the observation vector X, that is, for all measurable sets $A \subset \mathbb{R}^d$, $\mu(A) = \mathbb{P}\{X \in A\}$. Then the *regression function*

$$m(x) = \mathbb{E}[Y \mid X = x] \tag{1.1}$$

is well defined for μ-almost all x. The center of our investigations is the functional

$$L^* = \mathbb{E}\left[(m(X) - Y)^2\right].$$

The importance of this functional stems from the fact that for each measurable function $g : \mathbb{R}^d \to \mathbb{R}$ one has

$$\mathbb{E}\left[(g(X) - Y)^2\right] = L^* + \mathbb{E}\left[(m(X) - g(X))^2\right]$$

and, in particular,

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

where the minimum is taken over all measurable functions $g : \mathbb{R}^d \to \mathbb{R}$. In other words, L^* is the minimal mean squared error of any "predictor" of Y based on observing X. L^* is often referred to as the *residual variance*.

In regression analysis the residual variance L^* is of obvious interest as it provides a lower bound for the performance of any regression function estimator. In this paper we study the problem of estimating L^* based on data consisting of independent, identically distributed (i.i.d.) copies of the pair (X, Y). It is convenient to assume that the number of samples is even and the $2n$ samples are split into two halves as

$$D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \qquad \text{and} \qquad D'_n = \{(X'_1, Y'_1), \ldots, (X'_n, Y'_n)\}$$

such that the $2n+1$ pairs $(X,Y),(X_1,Y_1),\ldots,(X_n,Y_n),(X'_1,Y'_1),\ldots,(X'_n,Y'_n)$ are independent and identically distributed.

An estimator $\widehat{L}_n$ of L^* is simply a function of the data D_n, D'_n. We are interested in "nonparametric" estimators of L^* that work under minimal assumptions on the underlying distribution. In particular, a desirable feature of any estimate is that it is strongly universally consistent, that is, $\widehat{L}_n \to L^*$ with probability one, for all possible distributions of (X,Y) with $\mathbb{E}Y^2 < \infty$. Such estimators may be constructed, for example, by constructing a strongly universally consistent regression function estimator m_n based on the data D_n (i.e., a function m_n is such that $\mathbb{E}[(m_n(X)-Y)^2|D_n] \to L^*$ with probability one for all distributions) and estimating its mean squared error by $(1/n)\sum_{i=1}^n (m_n(X'_i)-Y'_i)^2$. (For a detailed theory of universally consistent regression function estimation see [15].) However, the rate of convergence of such estimators is determined by the rate of convergence of the mean squared error of m_n which can be quite slow even under regularity assumptions on the underlying distribution. Estimating the entire regression function $m(x)$ is, intuitively, "harder" than estimating the value of L^*. Indeed, nearest-neighbor-based estimators of L^* have been constructed and analyzed by Devroye, Ferrario, Györfi, and Walk [6], Devroye, Schäfer, Györfi, and Walk [10], Evans and Jones [12], Liitiäinen, Corona, and Lendasse [17], [18], Liitiäinen, Verleysen, Corona, and Lendasse [19], and Ferrario and Walk [13]. These estimates have been shown to have a faster rate of convergence—under some natural assumptions–than estimates based on estimating the error of consistent regression function estimators. Moreover, the estimate in [6] is strongly universally consistent.

In this paper we introduce yet another universally consistent nearest-neighbor-based estimator of L^*. The advantage of this estimator, apart from sharing the fast rates of convergence of previously defined estimators, is that its random fluctuations may be bounded by dimension-, and distribution-independent quantities. In particular, we prove a central limit theorem and a distribution-free upper bound for the variance for the new estimator that show that it is concentrated around its expected value in an interval of width $O(1/\sqrt{n})$, independently of the dimension. The established concentration property is crucial in a variable-selection procedure that we discuss as an application. In particular, we design a test for deciding whether exclusion of a certain component of X increases L^* or not.

The paper is organized as follows. In Section 2 we introduce a novel estimate of L^* and establish some of its properties such as asymptotic normality and a non-asymptotic concentration inequality. The central limit theorem holds without any smoothness condition on the regression function, and the asymptotic variance depends only on the conditional moments of Y (Theorem 1). We prove a non-asymptotic bound on the variance that does not depend on the dimension of X (Theorem 2), and show an exponential concentration inequality for the centered estimate (Theorem 3). All these results are universal in the sense that we only

assume that X has a density and Y is bounded.

In Section 3 we briefly describe how the results method based on the results of Section 2 may be relevant for variable selection. Finally, the proofs are presented in Section 4.

2 A nearest-neighbor based estimate and its asymptotic normality

Denoting the second moment of the regression function by

$$S^* = \mathbb{E}\left[m(X)^2\right] ,$$

we have

$$L^* = \mathbb{E}\left[Y^2\right] - S^* ,$$

and therefore estimating L^* is essentially equivalent to estimating S^* (as the "easy" part $\mathbb{E}\left[Y^2\right]$ may be estimated by, e.g., $(1/n)\sum_{i=1}^n Y_i^2$ whose behavior is well understood).

Next we introduce a nearest neighbor-based estimator of S^*. Based on the data D_n, we start by constructing a nearest-neighbor (1-NN) regression function estimator as follows. Let $X_{1,n}(x)$ be the first nearest neighbor of x among $X_1, \ldots, X_n$ (with respect to the Euclidean distance in $\mathbb{R}^d$) and let $Y_{1,n}(x)$ be its label. (In order to rigorously define the nearest neighbor, we assume that ties are broken in order to favor points with smaller index. Since we assume the distribution of X to be absolutely continuous, this issue is immaterial since ties occur with probability zero.) The 1-NN estimator of the regression function m is defined as

$$m_n(x) = Y_{1,n}(x) .$$

The proposed estimate of S^* is

$$S_n = \frac{1}{n}\sum_{i=1}^n Y_i' m_n(X_i') .$$

By a straightforward adjustment of the arguments of Devroye, Ferrario, Györfi, and Walk [6], one may show that S_n is a strongly universal consistent estimate of S^*, that is,

$$\lim_n S_n = S^*$$

with probability one for any distribution of (X, Y) with $\mathbb{E}[Y^2] < \infty$. Note that the consistent functional estimate S_n is based on a non-consistent regression function estimate m_n.

Next we establish asymptotic normality of S_n under the condition that the response variable Y is bounded. In order to describe the asymptotic variance, we introduce the dimension-dependent constant $\alpha(d)$ as follows.

Let $B_{x,r}$ denote the closed ball of radius $r > 0$ centered at x in $\mathbb{R}^d$ and let λ denote the Lebesgue measure on $\mathbb{R}^d$. Let V be a random vector uniformly distributed in $B_{0,1}$. Define $\overline{1} = (1, 0, 0, \ldots, 0) \in \mathbb{R}^d$ and let $\overline{B} = B_{\overline{1},1} \bigcup B_{V,\|V\|}$. Introduce the random variable

$$W = \frac{\lambda(\overline{B})}{\lambda(B_{0,1})}$$

and define

$$\alpha(d) = \mathbb{E}\left[\frac{2}{W^2}\right]. \tag{2.1}$$

Theorem 1. *Assume that μ has a density and that there exists a constant $L > 0$ such that*

$$\mathbb{P}\{|Y| < L\} = 1\,. \tag{2.2}$$

Denote

$$M_2(X) = \mathbb{E}[Y^2 \mid X]$$

and define

$$\sigma_1^2 = \int M_2(x)^2 \mu(dx) - \left(\int m(x)^2 \mu(dx)\right)^2$$

and

$$\sigma_2^2 = \alpha(d)\left(\int M_2(x) m(x)^2 \mu(dx) - \int m(x)^4 \mu(dx)\right).$$

If $\sigma_1 > 0$, then

$$\sqrt{n}\,(S_n - \mathbb{E}\{S_n\})/\sigma \xrightarrow{\mathcal{D}} N(0,1)\,,$$

where

$$\sigma^2 = \sigma_1^2 + \sigma_2^2\,.$$

The dependence of the asymptotic variance on the dimension d is weak, merely via the constant $\alpha(d)$. Given $X_1, \ldots, X_n$, Devroye, Györfi, Lugosi, and Walk [8] considered the probability measures of the Voronoi cells. They proved that the asymptotic variance of n-times the probability measure of the Voronoi cell is equal to $\alpha(d) - 1$. Thus, this asymptotic variance is universal in the sense that it does not depend on the underlying density. A few values are $\alpha(1) = 1.5$, $\alpha(2) \approx 1.28$, $\alpha(3) \approx 1.18$. In general, $1 \le \alpha(d) \le 2$ and $\alpha(d) \to 1$ exponentially fast as $d \to \infty$. Thus, by (2.2) we have $\sigma^2 \le 3L^4$, and therefore Theorem 1 implies that

$$\limsup_{n\to\infty} n\mathrm{Var}(S_n) \le 3L^4.$$

The next theorem shows that, up to a constant factor, this bound holds non-asymptotically.

Theorem 2. *Assume that μ has a density and that $|Y| < L$. Then for all $n \geq 1$,*

$$\mathbb{V}ar(S_n) \leq \frac{9 \cdot L^4}{n}.$$

The next result is a non-asymptotic exponential inequality that extends Theorem 2. It implies that for all $t > 0$,

$$\mathbb{P}\left\{\sqrt{n}|S_n - \mathbb{E}S_n| > t\right\} \leq ce^{-\left(t/(cL^2)\right)^{2/3}}$$

for a universal constant $c > 0$. It is an interesting open question whether the right-hand side can be improved to $e^{-\left(t/(cL^2)\right)^2}$. This would give a non-asymptotic analog of the central limit theorem of Theorem 1.

Theorem 3. *Assume that μ has a density and that $|Y| < L$. Write*

$$S_n - \mathbb{E}[S_n] = U_n + V_n$$

with

$$U_n = S_n - \mathbb{E}[S_n \mid D_n] \quad \textit{and} \quad V_n := \mathbb{E}[S_n \mid D_n] - \mathbb{E}[S_n] \ .$$

Then for every $n \geq 1$ and $\epsilon > 0$, we have

$$\mathbb{P}\{|U_n| > \epsilon\} \leq 2e^{-n\epsilon^2/(2L^4)}$$

and

$$\mathbb{P}\{|V_n| \geq \epsilon\} \leq 2e^{-n^{1/3}\epsilon^{2/3}/(42eL^4)^{1/3}+1} \ . \tag{2.3}$$

The proofs of Theorems 1, 2 and 3 are presented in Section 4.

3 Illustration: testing for dimension reduction

In standard nonparametric regression design, one considers a finite number of real-valued features $X^{(i)}$, $i \in I \subset \{1,\ldots,d\}$ for predicting the value of a response variable Y. A first question one may try to answer is whether these features suffice to explain Y. In case they do, an estimation method can be applied on the basis of the features already under consideration. Otherwise more or different features

need to be considered. The quality of a subvector $\{X^{(i)},\, i \in I\}$ of X is measured by the minimum mean squared error

$$L^*(I) := \mathbb{E}\Big[Y - \mathbb{E}[Y \mid X^{(i)} : i \in I]\Big]^2$$

that can be achieved using the features as explanatory variables. $L^*(I)$ depends upon the unknown distribution of $(Y, X^{(i)} : i \in I)$.

Thus, even before a regression function estimate is chosen, one may be interested in estimating L^*. For possible dimensionality reduction, one needs, in general, to test the hypothesis

$$L^* = L^*(I) \tag{3.1}$$

for a particular (proper) subset I of $\{1,\dots,d\}$. A natural way of approaching this testing problem is by estimating both L^* and $L^*(I)$, and accept the hypothesis if the two estimates are close to each other (De Brabanter, Ferrario and Györfi [5]).

Introduce the notation

$$S^*(I) := \mathbb{E}\Big[\mathbb{E}[Y \mid X^{(i)}, i \in I]^2\Big]\,.$$

Then the hypothesis (3.1) is equivalent to

$$S^* = S^*(I)\,.$$

Without loss of generality, consider the case $I = \{1,\dots,d-1\}$, that is, the case when one tests whether the last component $X^{(d)}$ of the observation vector $(X^{(1)},\dots,X^{(d)})$ is ineffective. Let the transformation T be defined by

$$T((x^{(1)},\dots,x^{(d)})) = (x^{(1)},\dots,x^{(d-1)})\,.$$

Thus, dropping the component $X^{(d)}$ from the observation vector $X = (X^{(1)},\dots,X^{(d)})$ leads to the observation vector

$$\widehat{X} = T(X) = (X^{(1)},\dots,X^{(d-1)})$$

of dimension $d-1$.

Using the notation

$$m(X) = \mathbb{E}[Y \mid X] \text{ and } \widetilde{m}(T(X)) = \mathbb{E}[Y \mid T(X)]$$

and

$$S^* = \mathbb{E}[m(X)^2] \text{ and } \widehat{S^*} = \mathbb{E}[\widetilde{m}(T(X))^2]\,,$$

the null-hypothesis $\widehat{S^*} = S^*$ is equivalent to

$$m(X) = \widetilde{m}(T(X)) \quad \text{with probability one.} \tag{3.2}$$

We propose to approach this testing problem by considering the nearest-neighbor estimate defined in Section 2. Let S_n be the estimate of S^* using the sample

$$\mathcal{D}_{2n} = \{(X_1, Y_1), \ldots, (X_{2n}, Y_{2n})\} .$$

Assume that an independent sample of size $2n$ is available:

$$\overline{\mathcal{D}}_{2n} = \{(\overline{X}_1, \overline{Y}_1), \ldots, (\overline{X}_{2n}, \overline{Y}_{2n})\} .$$

We use $\overline{\mathcal{D}}_{2n}$ to construct an estimate $\widetilde{S}_n$ of $\widehat{S^*}$. $\widetilde{S}_n$ is defined as the nearest-neighbor estimate computed from the sample

$$\{(T(\overline{X}_1), \overline{Y}_1), \ldots, (T(\overline{X}_{2n}), \overline{Y}_{2n})\} .$$

The proposed test is based of the test statistic

$$T_n = S_n - \widetilde{S}_n$$

and accepts the null hypothesis (3.2) if and only if

$$T_n \le a_n := \omega_n \left(n^{-1/2} + n^{-2/d} \right)$$

where ω_n is an increasing unbounded sequence such that $a_n \to 0$. Under the alternative hypothesis, according to the consistency result of Devroye, Ferrario, Györfi, and Walk [6], for bounded Y,

$$T_n \to S^* - \widehat{S^*} > 0 \quad \text{with probability one,} \tag{3.3}$$

and this convergence is universal, that is, it holds without any conditions. Thus, since $a_n \to 0$, if $\widehat{S^*} \neq S^*$, then, with probability one, the test does not make any mistake for a sufficiently large n.

Theorem 1 implies that

$$\sqrt{n}\,(S_n - \mathbb{E}S_n)/\sigma \xrightarrow{\mathcal{D}} N(0,1)$$

and

$$\sqrt{n}\left(\widetilde{S}_n - \mathbb{E}\widetilde{S}_n\right)/\widetilde{\sigma} \xrightarrow{\mathcal{D}} N(0,1)$$

with $\sigma^2, \widetilde{\sigma}^2 < 3L^4$. Since S_n and $\widetilde{S}_n$ are independent, we have

$$\sqrt{n}(T_n - \mathbb{E}T_n)/(\sqrt{\sigma^2 + \widetilde{\sigma}^2}) \xrightarrow{\mathcal{D}} N(0,1) . \tag{3.4}$$

In order to understand the behavior of the test, one needs to study the difference of the biases of the estimates

$$\mathbb{E}T_n = \mathbb{E}S_n - \mathbb{E}\widetilde{S}_n$$

under the null hypothesis (3.2). In this case we have

$$\mathbb{E}S_n - \mathbb{E}\widetilde{S}_n = (\mathbb{E}S_n - \mathbb{E}\{m(X)^2\}) - (\mathbb{E}\widetilde{S}_n - \mathbb{E}\{\widetilde{m}(T(X))^2\}) .$$

If $\widetilde{m}$ and f are Lipschitz continuous and f is bounded away from 0, then, by Devroye, Ferrario, Györfi, and Walk [6],

$$n^{2/d}(\mathbb{E}S_n - \mathbb{E}\{m(X)^2\}) = O(1)$$

when $d \geq 2$ and

$$n^{2/(d-1)}(\mathbb{E}\widetilde{S}_n - \mathbb{E}\{\widetilde{m}(T(X))^2\}) = O(1)$$

when $d \geq 3$.

Thus, under the null hypothesis (3.2),

$$\mathbb{E}T_n = O(n^{-2/d}) , \tag{3.5}$$

for $d \geq 2$. Note that for $d \leq 4$, the bias is at most of the order of the random fluctuations of the test statistic. However, for $d > 4$ the bias may dominate. Such a dependence on the dimension is inevitable under fully nonparametric conditions like the ones assumed here.

Under the null hypothesis, (3.4) and (3.5) imply that the probability of error may be bounded as

$$\mathbb{P}\{T_n > a_n\} \leq \mathbb{P}\{T_n - \mathbb{E}T_n > \omega_n \cdot n^{-1/2}\} + \mathbb{1}_{\{\mathbb{E}T_n > \omega_n \cdot n^{-2/d}\}} \to 0 .$$

Thus, the test is consistent.

The condition that the density f is bounded away from zero may be avoided at the price of a worse rate of convergence. In particular, if m is C-Lipschitz and X is bounded, then

$$\begin{aligned}
& n^{1/d}|\mathbb{E}S_n - \mathbb{E}[m(X)^2]| \\
&= n^{1/d}|\mathbb{E}[m(X)m_n(X)] - \mathbb{E}[m(X)^2]| \\
&= n^{1/d}|\mathbb{E}[m(X)m(X_{1,n}(X))] - \mathbb{E}[m(X)^2]| \\
&\leq n^{1/d} LC\mathbb{E}\|X_{1,n}(X) - X\| \\
&= O(1) \quad \text{(by a packing argument of Liitiäinen et al. [18, Theorem 3.2] and by Biau and Devroye [1, Theorem 2.1]).}
\end{aligned}$$

In this case the threshold should be larger:

$$a_n := \omega_n\left(n^{-1/2} + n^{-1/d}\right)$$

One may prove that the test is not only consistent in the sense that $\mathbb{P}\{T_n > a_n\} \to 0$ under the null hypothesis but also in the sense that $\limsup_{n\to\infty} \mathbb{1}_{\{T_n > a_n\}} =$

0 with probability one. For a discussion and references on the notion of strong consistency we refer the reader to Devroye and Lugosi [9], Biau and Györfi [2], Gretton and Györfi [14].

The proof of strong consistency under the alternative hypothesis follows simply from (3.3). Under the null hypothesis it follows from Theorem 3. Indeed, Theorem 3 implies that

$$\mathbb{P}\{|T_n - \mathbb{E}T_n| > \epsilon\} \le 2e^{-n\epsilon^2/(2L^4)} + 2e^{-n^{1/3}\epsilon^{2/3}/(42eL^4)^{1/3}+1} .$$

For $\delta > 3/2$, choose

$$a_n := (\ln n)^\delta n^{-1/2} + \omega_n \cdot n^{-2/d}$$

with increasing unbounded $\omega_n = o(n^{2/d})$. Then, under the null hypothesis

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}\{T_n > a_n\} &\le \sum_{n=1}^{\infty} \Big(\mathbb{P}\{T_n - \mathbb{E}T_n > (\ln n)^\delta n^{-1/2}\} + \mathbb{1}_{\{\mathbb{E}T_n > \omega_n \cdot n^{-2/d}\}}\Big) \\ &\le \sum_{n=1}^{\infty} \Big(2e^{-(\ln n)^{2\delta}/(2L^4)} + 2e^{-(\ln n)^{2\delta/3}/(42eL^4)^{1/3}+1} + \mathbb{1}_{\{\mathbb{E}T_n > \omega_n \cdot n^{-2/d}\}}\Big) \\ &< \infty ,\end{aligned}$$

and so the Borel-Cantelli Lemma implies that the test makes error only finitely many times almost surely.

Remark. In applications, one would like to test not only if a given component of X carries predictive information but rather test the same for each of the d variables. In such cases, one faces a *multiple testing* problem with d dependent tests. In order to analyze such multiple testing procedures, say, by the Bonferroni approach, one needs a uniform control over the fluctuations of the test statistic. In such cases a non-asymptotic concentration inequality of Theorem 3 is particularly useful.

4 Proofs

In the proofs below we use two lemmas on the measure of Voronoi cells. Let

$$A_n(X_j) = \{x \in \mathbb{R}^d : X_j \text{ is the nearest neighbor of } x \text{ among } X_1, \dots, X_n\}$$

($j = 1, \dots, n$), be the cells of the Voronoi partition of $\mathbb{R}^d$.

Lemma 1. *If μ has a density, then*

$$n^k \mathbb{E}\Big[\mu(A_n(X_1))^k\Big] \le k! ,$$

$k = 1, 2, \dots$

Proof. Devroye, Györfi, Lugosi, and Walk [8] proved that there exists a positive constant c_k such that

$$n^k \mathbb{E}\Big[\mu(A_n(X_1))^k\Big] \le c_k \ ,$$

and $n\mu(A_n(X_1))$ converges in distribution to a random variable Z such that

$$\mathbb{E}\Big[Z^k\Big] \le k! \ ,$$

$k = 1, 2, \ldots$ This lemma is on the same non-asymptotic bound. We show that

$$\begin{aligned} &\mathbb{E}\Big\{\mu(A_n(X_1))^k\Big\} \\ &\le \mathbb{P}\{X_{n+1}, \ldots, X_{n+k} \text{ are the nearest neighbors of } X_1 \text{ among } X_2, \ldots, X_{n+k}\} \ , \end{aligned} \tag{4.1}$$

which implies that

$$\mathbb{E}\Big\{(n\mu(A_n(X_1)))^k\Big] \le \frac{n^k}{\binom{n+k-1}{k}} \le k! \ .$$

Recall that $B_{x,r}$ denotes the closed ball of radius $r > 0$ centered at x and note that

$$\begin{aligned} \mathbb{E}\Big\{\mu(A_n(X_1))^k\Big\} &= \mathbb{P}\{X_{n+1}, \ldots, X_{n+k} \in A_n(X_1)\} \\ &= \mathbb{E}\Big[(1 - \mu(B_{X_{n+1}, \|X_{n+1} - X_1\|} \cup \cdots \cup B_{X_{n+k}, \|X_{n+k} - X_1\|}))^{n-1}\Big] \\ &\le \mathbb{E}\Big[(1 - \max\{\mu(B_{X_{n+1}, \|X_{n+1} - X_1\|}), \ldots, \mu(B_{X_{n+k}, \|X_{n+k} - X_1\|})\})^{n-1}\Big] \ , \end{aligned}$$

and

$$\begin{aligned} &\mathbb{P}\{X_{n+1} \ldots, X_{n+k} \text{ are the nearest neighbors of } X_1 \text{ among } X_2, \ldots, X_{n+k}\} \\ &= \mathbb{E}\Big[(1 - \max\{\mu(B_{X_1, \|X_{n+1} - X_1\|}), \ldots, \mu(B_{X_1, \|X_{n+k} - X_1\|})\})^{n-1}\Big] \ . \end{aligned}$$

(4.1) follows from comparing the right-hand sides of the two equations above. On the one hand,

$$\begin{aligned} &\mathbb{P}\Big\{\max\{\mu(B_{X_1, \|X_{n+1} - X_1\|}), \ldots, \mu(B_{X_1, \|X_{n+k} - X_1\|})\} \le z\Big\} \\ &= \mathbb{P}\Big\{\mu(B_{X_1, \|X_{n+1} - X_1\|}) \le z, \ldots, \mu(B_{X_1, \|X_{n+k} - X_1\|}) \le z\Big\} \\ &= \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X_1, \|X_{n+1} - X_1\|}) \le z, \ldots, \mu(B_{X_1, \|X_{n+k} - X_1\|}) \le z \mid X_1\Big\}\Big] \\ &= \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X_1, \|X_{n+1} - X_1\|}) \le z \mid X_1\Big\} \cdot \ldots \cdot \mathbb{P}\Big\{\mu(B_{X_1, \|X_{n+k} - X_1\|}) \le z \mid X_1\Big\}\Big] \\ &= \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X_1, \|X_{n+1} - X_1\|}) \le z \mid X_1\Big\}^k\Big] \\ &= z^k \ , \end{aligned}$$

while on the other hand,

$$
\begin{aligned}
&\mathbb{P}\Big\{\max\{\mu(B_{X_{n+1},\|X_{n+1}-X_1\|}),\ldots,\mu(B_{X_{n+k},\|X_{n+k}-X_1\|})\} \le z\Big\} \\
&= \mathbb{P}\Big\{\mu(B_{X_{n+1},\|X_{n+1}-X_1\|}) \le z,\ldots,\mu(B_{X_{n+k},\|X_{n+k}-X_1\|}) \le z\Big\} \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X_{n+1},\|X_{n+1}-X_1\|}) \le z,\ldots,\mu(B_{X_{n+k},\|X_{n+k}-X_1\|}) \le z \mid X_1\Big\}\Big] \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X_{n+1},\|X_{n+1}-X_1\|}) \le z \mid X_1\Big\}\cdot\ldots\cdot\mathbb{P}\Big\{\mu(B_{X_{n+k},\|X_{n+k}-X_1\|}) \le z \mid X_1\Big\}\Big] \\
&= \mathbb{E}\left[\mathbb{P}\Big\{\mu(B_{X_{n+1},\|X_{n+1}-X_1\|}) \le z \mid X_1\Big\}^k\right] \\
&\ge \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X_{n+1},\|X_{n+1}-X_1\|}) \le z \mid X_1\Big\}\Big]^k \\
&= \mathbb{P}\Big\{\mu(B_{X_{n+1},\|X_{n+1}-X_1\|}) \le z\Big\}^k \\
&= z^k\,.
\end{aligned}
$$

□

Lemma 2. *(Devroye, Györfi, Lugosi, and Walk [8]) Assume that μ has a density. Then*

$$n^2\mathbb{E}\Big[\mu(A_n(X_1))^2 \mid X_1 = x\Big] \to \alpha(d)$$

for μ-almost all x, where α_d is defined in (2.1).

Proof of Theorem 2

We prove the variance bound of Theorem 2 first. The proof relies of the following version of the Efron-Stein inequality, see, for example, [4, Theorem 3.1].

Lemma 3. *(Efron-Stein inequality) Let $Z = (Z_1,\ldots,Z_n)$ be a collection of independent random variables taking values in some measurable set A and denote by $Z^{(i)} = (Z_1,\ldots,Z_{i-1},Z_{i+1},\ldots,Z_n)$ the collection with the i-th random variable dropped. Let $f : A^n \to \mathbb{R}$ and $g : A^{n-1} \to \mathbb{R}$ be measurable real-valued functions. Then*

$$\mathrm{Var}(f(Z)) \le \mathbb{E}\left[\sum_{i=1}^{n}\Big(f(Z) - g(Z^{(i)})\Big)^2\right].$$

By the decomposition

$$S_n = S_n - \mathbb{E}[S_n \mid D_n] + \mathbb{E}[S_n \mid D_n],$$

we have that

$$\mathbb{V}ar(S_n) = \mathbb{E}\Big[(S_n - \mathbb{E}[S_n \mid D_n])^2\Big] + \mathbb{V}ar(\mathbb{E}[S_n \mid D_n])\,.$$

Conditionally on D_n, S_n is an average of independent, identically distributed (i.i.d.) random variables bounded by L^2, and therefore

$$\mathbb{E}\Big[(S_n - \mathbb{E}[S_n \mid D_n])^2\Big] \leq \frac{L^4}{n} .$$

Notice that we may write

$$m_n(x) = \sum_{j=1}^{n} Y_j \mathbb{1}_{\{x \in A_n(X_j)\}} .$$

Then

$$\mathbb{E}[S_n \mid D_n] = \int m(x) m_n(x) \mu(dx) = \sum_{j=1}^{n} Y_j \int_{A_n(X_j)} m(x) \mu(dx) .$$

Putting $L_n = \mathbb{E}[S_n \mid D_n]$, this implies

$$L_n = \sum_{i=1}^{n} Y_i \mathbb{E}\{\mathbb{1}_{X \in A_n(X_i)} m(X) \mid D_n\} .$$

Considering L_n as a function of the n i.i.d. pairs $(X_i, Y_i)_{i=1}^n$, we may use the Efron-Stein inequality to bound the variance of L_n. Define $L_n^{(j)}$ as L_n when (X_j, Y_j) is omitted from the sample. By Lemma 3,

$$\mathbb{V}ar(L_n) \leq \mathbb{E}\left[\sum_{j=1}^{n} \left(L_n - L_n^{(j)}\right)^2\right] = n\mathbb{E}\left[\left(L_n - L_n^{(1)}\right)^2\right] .$$

Let $\{A'_n(X_2), \ldots, A'_n(X_n)\}$ be the Voronoi partition, when X_1 is omitted from the sample. Then

$$\begin{aligned} |L_n - L_n^{(1)}| &= \left| Y_1 \int_{A_n(X_1)} m(x)\mu(dx) - \sum_{i=2}^{n} Y_i \int_{A'_n(X_i) \setminus A_n(X_i)} m(x)\mu(dx) \right| \\ &\leq L^2 \left(\mu(A_n(X_1)) + \sum_{i=2}^{n} \mu(A'_n(X_i) \setminus A_n(X_i)) \right) \\ &= 2L^2 \mu(A_n(X_1)) . \end{aligned}$$

Thus, Lemma 1 implies

$$\mathbb{V}ar(L_n) \leq 4nL^4 \mathbb{E}\Big[\mu(A_n(X_1))^2\Big] \leq 8L^4/n$$

leading to

$$\mathbb{V}ar\left(\mathbb{E}[S_n \mid D_n]\right) \leq \frac{8L^4}{n} ,$$

and therefore to the desired bound

$$\mathbb{V}ar(S_n) \leq \frac{9L^4}{n} .$$

Proof of Theorem 1

Introduce the notation

$$\sqrt{n}(S_n - \mathbb{E}S_n) = U_n + V_n + W_n \ ,$$

where

$$U_n = \sqrt{n}(S_n - \mathbb{E}[S_n \mid D_n])$$

and

$$V_n = \sqrt{n}(\mathbb{E}[S_n \mid D_n] - \mathbb{E}[S_n \mid X_1, \ldots, X_n])$$

and

$$W_n = \sqrt{n}(\mathbb{E}[S_n \mid X_1, \ldots, X_n] - \mathbb{E}S_n) \ .$$

We prove Theorem 1 by showing that, for any $u, v \in \mathbb{R}$,

$$\mathbb{P}\{U_n \le u, V_n \le v\} \to \Phi\left(\frac{u}{\sigma_1}\right)\Phi\left(\frac{v}{\sigma_2}\right), \tag{4.2}$$

where Φ denotes the standard normal distribution function, and that

$$\mathrm{Var}(W_n) \to 0. \tag{4.3}$$

Györfi and Walk [16] proved that

$$\begin{aligned}&\left|\mathbb{P}\{U_n \le u, V_n \le v\} - \Phi\left(\frac{u}{\sigma_1}\right)\Phi\left(\frac{v}{\sigma_2}\right)\right| \\ &\le \mathbb{E}\left|\mathbb{P}\{U_n \le u \mid D_n\} - \Phi\left(\frac{u}{\sigma_1}\right)\right| + \left|\mathbb{P}\{V_n \le v\} - \Phi\left(\frac{v}{\sigma_2}\right)\right| \ .\end{aligned}$$

Thus, (4.2) holds if

$$\mathbb{P}\{U_n \le u \mid D_n\} \to \Phi\left(\frac{u}{\sigma_1}\right) \quad \text{in probability} \tag{4.4}$$

and

$$\mathbb{P}\{V_n \le v\} \to \Phi\left(\frac{v}{\sigma_2}\right). \tag{4.5}$$

Proof of (4.4).
Let's start with the decomposition

$$\begin{aligned}U_n &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i' m_n(X_i') - \mathbb{E}[Y_i' m_n(X_i') \mid D_n])\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i' m_n(X_i') - \mathbb{E}[Y_i' m_n(X_i') \mid D_n]) \ .\end{aligned}$$

Next we apply a Berry-Esseen type central limit theorem (see Theorem 14 in Petrov [20]). For a universal constant $c > 0$, we have

$$\left| \mathbb{P}\{U_n \le u \mid D_n\} - \Phi\left(\frac{u}{\sqrt{\mathbb{V}ar(Y_1' m_n(X_1') \mid D_n)}} \right) \right| \le \frac{c}{\sqrt{n}} \frac{\mathbb{E}[|Y_1' m_n(X_1')|^3 \mid D_n]}{\sqrt{\mathbb{V}ar(Y_1' m_n(X_1') \mid D_n)}^3} .$$

Since

$$\mathbb{E}[Y_1' m_n(X_1') \mid D_n] = \int m(x) m_n(x) \mu(dx) , \tag{4.6}$$

we have

$$\begin{aligned} \mathbb{V}ar(Y_1' m_n(X_1') \mid D_n) &= \mathbb{E}[{Y_1'}^2 m_n(X_1')^2 \mid D_n] - \mathbb{E}[Y_1' m_n(X_1') \mid D_n]^2 \\ &= \int M_2(x) m_n(x)^2 \mu(dx) - \left(\int m(x) m_n(x) \mu(dx) \right)^2 . \end{aligned}$$

We need to show that

$$\int M_2(x) m_n(x)^2 \mu(dx) \to \int M_2(x)^2 \mu(dx) \tag{4.7}$$

in probability and

$$\int m(x) m_n(x) \mu(dx) \to \int m(x)^2 \mu(dx) \tag{4.8}$$

in probability. Since $m_n(x) = Y_j$ if $x \in A_n(X_j)$, we get that

$$\begin{aligned} \int M_2(x) m_n(x)^2 \mu(dx) &= \sum_{j=1}^n \int_{A_n(X_j)} M_2(x) m_n(x)^2 \mu(dx) \\ &= \sum_{j=1}^n Y_j^2 \int_{A_n(X_j)} M_2(x) \mu(dx) . \end{aligned}$$

We use this to prove (4.7). Indeed,

$$\begin{aligned} &\int M_2(x) m_n(x)^2 \mu(dx) - \int M_2(x)^2 \mu(dx) \\ &= \sum_{j=1}^n Y_j^2 \int_{A_n(X_j)} M_2(x) \mu(dx) - \sum_{j=1}^n \int_{A_n(X_j)} M_2(x)^2 \mu(dx) \\ &= \sum_{j=1}^n (Y_j^2 - M_2(X_j)) \int_{A_n(X_j)} M_2(x) \mu(dx) \\ &\quad + \sum_{j=1}^n \int_{A_n(X_j)} M_2(x) (M_2(X_j) - M_2(x)) \mu(dx) . \end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[\left|\int M_2(x)m_n(x)^2\mu(dx) - \int M_2(x)^2\mu(dx)\right|\right] \\
&\leq \mathbb{E}\left[\left|\sum_{j=1}^{n}(Y_j^2 - M_2(X_j))\int_{A_n(X_j)} M_2(x)\mu(dx)\right|\right] \\
&\quad + \mathbb{E}\left[\left|\sum_{j=1}^{n}\int_{A_n(X_j)} M_2(x)(M_2(X_j) - M_2(x))\mu(dx)\right|\right],
\end{aligned}$$

and so

$$\begin{aligned}
&\mathbb{E}\left[\left|\int M_2(x)m_n(x)^2\mu(dx) - \int M_2(x)^2\mu(dx)\right|\right] \\
&\leq \sqrt{\mathbb{V}ar\left(\sum_{j=1}^{n}(Y_j^2 - M_2(X_j))\int_{A_n(X_j)} M_2(x)\mu(dx)\right)} \\
&\quad + \mathbb{E}\left[\sum_{j=1}^{n}\int_{A_n(X_j)} M_2(x)|M_2(X_j) - M_2(x)|\mu(dx)\right] \\
&\leq \sqrt{n\mathbb{E}\left[(Y_1^2 - M_2(X_1))^2\left(\int_{A_n(X_1)} M_2(x)\mu(dx)\right)^2\right]} \\
&\quad + n\mathbb{E}\left[\int_{A_n(X_1)} M_2(x)|M_2(X_1) - M_2(x)|\mu(dx)\right] \\
&\leq L^4\sqrt{n\mathbb{E}[\mu(A_n(X_1))^2]} + L^2 n\mathbb{E}\left[\int_{A_n(X_1)} |M_2(X_1) - M_2(x)|\mu(dx)\right]
\end{aligned}$$

To complete the proof of (4.7), it suffices to show that the sum above converges to zero as $n \to \infty$. To this end, note that Lemma 1 implies that

$$n\mathbb{E}\left[\mu(A_n(X_1))^2\right] \leq c_2/n \to 0\,,$$

and furthermore

$$\begin{aligned}
&n\mathbb{E}\left[\int_{A_n(X_1)} |M_2(X_1) - M_2(x)|\mu(dx)\right] \\
&= n\mathbb{E}\left[\int_{A_n(X_1)} |M_2(X_{1,n}(x)) - M_2(x)|\mu(dx)\right] \\
&= \mathbb{E}\left[\int |M_2(X_{1,n}(x)) - M_2(x)|\mu(dx)\right].
\end{aligned}$$

It remains to show that

$$\mathbb{E}\left[\int |M_2(X_{1,n}(x)) - M_2(x)|\mu(dx)\right] \to 0\,. \tag{4.9}$$

Fix any $\epsilon > 0$ and choose a bounded continuous function $\widetilde{M}_2$ such that

$$\int |M_2(x) - \widetilde{M}_2(x)|\mu(dx) < \epsilon\,.$$

Then, with $M_2^* = M_2 - \widetilde{M}_2$, one has

$$\begin{aligned}
&\mathbb{E}\left[\int |M_2(X_{1,n}(x)) - M_2(x)|\mu(dx)\right] \\
&\quad\leq\ \mathbb{E}\left[\int |\widetilde{M}_2(X_{1,n}(x)) - \widetilde{M}_2(x)|\mu(dx)\right] \\
&\quad+\ \mathbb{E}\left[\int |M_2^*(X_{1,n}(x))|\mu(dx)\right] + \int |M_2^*(x)|\mu(dx)\,.
\end{aligned} \tag{4.10}$$

The first term on the right-hand side converges to 0 by the dominated convergence theorem, since, by Lemma 6.1 in [15],

$$X_{1,n}(x) \to x \qquad \text{a.s. for } \mu\text{-almost all } x\,.$$

To bound the second term, we introduce some notation. A set $C \subset \mathbb{R}^d$ is a cone of angle $\pi/3$ centered at 0 if there exists an $x \in \mathbb{R}^d$ with $\|x\| = 1$ such that

$$C = \left\{y \in \mathbb{R}^d : \frac{(x,y)}{\|y\|} \geq \cos(\pi/6)\right\}\,.$$

Let γ_d be the minimal number of cones $C_1, \ldots, C_{\gamma_d}$ of angle $\pi/3$ centered at 0 such that their union covers $\mathbb{R}^d$. The second term on the right-hand side of (4.10) is bounded by

$$\gamma_d \int |M_2^*(x)|\mu(dx) \leq \gamma_d \epsilon$$

by Lemma 6.3 in [15]. Thus, (4.9) is proved and hence so is (4.7). For the proof of (4.8), we have that

$$\begin{aligned}
\int m(x)m_n(x)\mu(dx) &= \sum_{j=1}^{n} \int_{A_n(X_j))} m(x)m_n(x)\mu(dx) \\
&= \sum_{j=1}^{n} Y_j \int_{A_n(X_j))} m(x)\mu(dx).
\end{aligned} \tag{4.11}$$

Similarly, the derivation for (4.7) implies that

$$\begin{aligned}
&\mathbb{E}\left[\left|\int m(x)m_n(x)\mu(dx) - \int m(x)^2\mu(dx)\right|\right] \\
&\leq L^2\sqrt{n\mathbb{E}[\mu(A_n(X_1))^2]} + Ln\mathbb{E}\left[\int_{A_n(X_1)} |m(X_1) - m(x)|\mu(dx)\right] \\
&\to 0,
\end{aligned}$$

and so (4.8) is proved, too. Thus,

$$\mathbb{V}ar(Y_1' m_n(X_1') \mid D_n) \to \sigma_1^2$$

in probability. Moreover,

$$\mathbb{E}[|Y_1' m_n(X_1')|^3 \mid D_n] \leq L^6.$$

These relations imply (4.4).

Proof of (4.3).
(4.6) and (4.11) imply that

$$\mathbb{E}[S_n \mid D_n] = \mathbb{E}[Y_1' m_n(X_1') \mid D_n] = \int m(x)m_n(x)\mu(dx) = \sum_{j=1}^{n} Y_j \int_{A_n(X_j)} m(x)\mu(dx)\,.$$

Hence

$$\mathbb{E}[S_n \mid X_1, \ldots, X_n] = \sum_{j=1}^{n} m(X_j) \int_{A_n(X_j)} m(x)\mu(dx) = \int m(x)m(X_{1,n}(x))\mu(dx).$$

We prove (4.3) by a slight extension of the proof of Theorem 2. Set

$$L_n := \sqrt{n}\int m(x)m(X_{1,n}(x))\mu(dx) = \sqrt{n}\sum_{j=1}^{n} m(X_j)\int_{A_n(X_j)} m(x)\mu(dx)\,.$$

Define $L_n^{(j)}$ as L_n when X_j is dropped. As in the proof of Theorem 2,

$$\mathbb{V}ar(W_n) = \mathbb{V}ar(L_n) \leq \mathbb{E}\left[\sum_{j=1}^{n}\left(L_n - L_n^{(j)}\right)^2\right] = n\mathbb{E}\left[\left(L_n - L_n^{(1)}\right)^2\right].$$

Then

$$L_n^{(1)} = \sqrt{n}\sum_{j=2}^{n} m(X_j)\int_{A_n'(X_j)} m(x)\mu(dx)\,,$$

and so

$$L_n - L_n^{(1)} = \sqrt{n} m(X_1) \int_{A_n(X_1)} m(x)\mu(dx) - \sqrt{n} \sum_{j=2}^{n} m(X_j) \int_{A_n'(X_j)\setminus A_n(X_j)} m(x)\mu(dx)$$
$$= \sqrt{n} \left(\int_{A_n(X_1)} m(X_{1,n}(x)) m(x)\mu(dx) - \int_{A_n(X_1)} m(X_{2,n}(x)) m(x)\mu(dx) \right),$$

where $X_{2,n}(x)$ denotes the second nearest neighbor of x among $X_1, \ldots, X_n$. Therefore

$$|L_n - L_n^{(1)}| \le \sqrt{n} L \int_{A_n(X_1)} |m(X_{1,n}(x)) - m(X_{2,n}(x))| \mu(dx)$$

by (2.2). Hence,

$$\mathbb{V}ar(W_n) \le L^2 \mathbb{E}\left[\left(n \int_{A_n(X_1)} |m(X_{1,n}(x)) - m(X_{2,n}(x))| \mu(dx) \right)^2 \right]. \tag{4.12}$$

As it is well known, for a real-valued random variable Z, by Hölder's inequality,

$$\mathbb{E}\left[Z^2\right] = \mathbb{E}\left[|Z|^{2/3} |Z|^{4/3}\right] \le \mathbb{E}\left[|Z|\right]^{2/3} \mathbb{E}\left[Z^4\right]^{1/3}. \tag{4.13}$$

One has

$$\begin{aligned}
&\mathbb{E}\left[n \int_{A_n(X_1)} |m(X_{1,n}(x)) - m(X_{2,n}(x))| \mu(dx) \right] \\
&\le \mathbb{E}\left[n \int_{A_n(X_1)} |m(X_{1,n}(x)) - m(x)| \mu(dx) \right] + \mathbb{E}\left[n \int_{A_n(X_1)} |m(X_{2,n}(x)) - m(x)| \mu(dx) \right] \\
&= \mathbb{E}\left[\int |m(X_{1,n}(x)) - m(x)| \mu(dx) \right] + \mathbb{E}\left[\int |m(X_{2,n}(x)) - m(x)| \mu(dx) \right] \\
&\to 0
\end{aligned} \tag{4.14}$$

as $n \to \infty$, where the latter can be shown as the limit relation (4.9). Furthermore

$$\begin{aligned}
\mathbb{E}\left[\left(n \int_{A_n(X_1)} |m(X_{1,n}(x)) - m(X_{2,n}(x))| \mu(dx) \right)^4 \right] &\le 16 L^4 \mathbb{E}\left[n^4 \mu(A_n(X_1))^4 \right] \\
&\le 16 L^4 c_4
\end{aligned} \tag{4.15}$$

by (2.2) and Lemma 1. With the notation

$$Z = n \int_{A_n(X_1)} |m(X_{1,n}(x)) - m(X_{2,n}(x))| \mu(dx)$$

(4.12), (4.13), (4.14) and (4.15) imply (4.3).

Proof of (4.5).
For

$$V_n = \frac{\sum_{j=1}^{n} V_{n,j}}{\sqrt{n}}$$

with

$$V_{n,j} = n(Y_j - m(X_j)) \int_{A_n(X_j)} m(x)\mu(dx) ,$$

notice that the triangular array $V_{n,j}$, $n = 1, 2, \ldots$, $j = 1, \ldots, n$ is (row-wise) exchangeable, for which there is a classical central limit theorem:

Theorem 4. *(Blum et al. [3], Weber [21]) Let $\{V_{n,j}\}$ be a triangular array of exchangeable random variables with zero mean and finite variance. Assume that*

(i)

$$\mathbb{E}[V_{n,1} V_{n,2}] = o(1/n) ,$$

(ii)

$$\lim_{n\to\infty} \max\{|V_{n,j}|; j = 1, \ldots, n\}/\sqrt{n} = 0$$

in probability,

(iii)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} V_{n,j}^2 = \sigma^2$$

in probability.

Then

$$\frac{\sum_{j=1}^{n} V_{n,j}}{\sqrt{n}}$$

is asymptotically normal with mean zero and variance σ^2.

Condition (i) of Theorem 4 is satisfied since

$$\mathbb{E}[V_{n,1} V_{n,2}] = 0.$$

Condition (ii) of Theorem 4 follows from (2.2), Lemma 1 and Jensen's inequality:

$$\begin{aligned} n\mathbb{E}\left[\max_j \mu(A_n(X_j)\right] &\leq n\mathbb{E}\left[\left(\sum_j \mu(A_n(X_j)^3\right)^{1/3}\right] \\ &\leq n\left(\mathbb{E}\left[\sum_j \mu(A_n(X_j)^3\right]\right)^{1/3} \\ &\leq n\left(n\frac{c_3}{n^3}\right)^{1/3} \\ &= o(\sqrt{n})\,. \end{aligned}$$

Condition (iii) in Theorem 4 is fulfilled if

$$\lim_{n\to\infty} \mathbb{E}[V_{n,1}^2] = \sigma_2^2 \tag{4.16}$$

and

$$\mathbb{V}ar\left(\frac{1}{n}\sum_{j=1}^{n} V_{n,j}^2\right) \to 0. \tag{4.17}$$

We have that

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}[V_{n,1}^2] &= \lim_{n\to\infty} n^2\mathbb{E}\left[(Y_1 - m(X_1))^2\left(\int_{A_n(X_1)} m(x)\mu(dx)\right)^2\right] \\ &= \lim_{n\to\infty} n^2\mathbb{E}\left[(Y_1 - m(X_1))^2 m(X_1)^2\mu(A_n(X_1))^2\right] \\ &= \lim_{n\to\infty} n^2\mathbb{E}\left[(M_2(X_1)m(X_1)^2 - m(X_1)^4)\mu(A_n(X_1))^2\right]\,. \end{aligned} \tag{4.18}$$

(4.18) follows from

$$
\begin{aligned}
& n^2 \Big| \mathbb{E}\left[(Y_1 - m(X_1))^2 \left(\int_{A_n(X_1)} m(x)\mu(dx) \right)^2 \right] \\
& - \mathbb{E}\left[(Y_1 - m(X_1))^2 m(X_1)^2 \mu(A_n(X_1))^2 \right] \Big| \\
& \le n^2 4L^2 \mathbb{E}\left[\left| \left(\int_{A_n(X_1)} m(x)\mu(dx) \right)^2 - m(X_1)^2 \mu(A_n(X_1))^2 \right| \right] \\
& \le n^2 8L^3 \mathbb{E}\left[\left| \int_{A_n(X_1)} m(x)\mu(dx) - m(X_1)\mu(A_n(X_1)) \right| \mu(A_n(X_1)) \right] \\
& = n^2 8L^3 \mathbb{E}\left[\left| \frac{\int_{A_n(X_1)} m(x)\mu(dx)}{\mu(A_n(X_1))} - m(X_1) \right| \mu(A_n(X_1))^2 \right] \\
& \le n^2 8L^3 \sqrt{ \mathbb{E}\left[\left| \frac{\int_{A_n(X_1)} m(x)\mu(dx)}{\mu(A_n(X_1))} - m(X_1) \right|^2 \right] } \sqrt{ \mathbb{E}\left[\mu(A_n(X_1))^4 \right] } \\
& \le 8L^3 \sqrt{c_4} \sqrt{ \mathbb{E}\left[\left| \frac{\int_{A_n(X_1)} m(x)\mu(dx)}{\mu(A_n(X_1))} - m(X_1) \right|^2 \right] } .
\end{aligned}
$$

The expression on the right-hand side converges to zero. To show this, fix an arbitrary $\epsilon > 0$ and choose a decomposition $m = m^* + m^{**}$ such that m^* is Lipschitz continuous with bounded support and $\mathbb{E}[m^{**}(X)^2] < \epsilon$. Then it suffices to show the limit relation for m^*. But this follows from the fact that $\operatorname{diam}(A_n(X_1)) \to 0$ in probability (Devroye, Györfi, Lugosi, and Walk [8, Section 5]). Lemma 2 implies that

$$
\mathbb{E}\left[n^2 \mu(A_n(X_1))^2 \mid X_1 \right] \to \alpha(d) \quad \text{with probability one.} \tag{4.19}
$$

Set

$$
Z_n = (M_2(X_1) m(X_1)^2 - m(X_1)^4) \mathbb{E}\left[n^2 \mu(A_n(X_1))^2 \mid X_1 \right] .
$$

By (2.2) and Lemma 1 for $k = 4$ together with Jensen's inequality for conditional expectations we obtain

$$
\mathbb{E}[Z_n^2] \le L^8 c_4
$$

and thus uniform integrability of $\{Z_n\}$, i.e.,

$$
\lim_{K \to \infty} \sup_n \mathbb{E}[Z_n \mathbb{1}_{\{Z_n > K\}}] = 0.
$$

Then (4.19) yields

$$\begin{aligned}
&n^2\mathbb{E}\Big[(M_2(X_1)m(X_1)^2 - m(X_1)^4)\mu(A_n(X_1))^2\Big] \\
&= \mathbb{E}\Big[(M_2(X_1)m(X_1)^2 - m(X_1)^4)\mathbb{E}\Big[n^2\mu(A_n(X_1))^2 \mid X_1\Big]\Big] \\
&\to \alpha(d)\mathbb{E}\Big[M_2(X_1)m(X_1)^2 - m(X_1)^4\Big] \\
&= \sigma_2^2 \,,
\end{aligned}$$

verifying (4.16).

One may check (4.17) similarly to (4.3). Indeed, put

$$L_n := \frac{1}{n}\sum_{j=1}^{n} V_{n,j}^2 = n\sum_{j=1}^{n}(Y_j - m(X_j))^2\left(\int_{A_n(X_j)} m(x)\mu(dx)\right)^2 .$$

Thus,

$$\begin{aligned}
&|L_n - L_n^{(1)}| \\
&\le n(Y_1 - m(X_1))^2\left(\int_{A_n(X_1)} m(x)\mu(dx)\right)^2 \\
&+ n\sum_{j=2}^{n}(Y_j - m(X_j))^2\left|\left(\int_{A_n(X_j)} m(x)\mu(dx)\right)^2 - \left(\int_{A_n'(X_j)} m(x)\mu(dx)\right)^2\right| .
\end{aligned}$$

Therefore

$$\begin{aligned}
&|L_n - L_n^{(1)}| \\
&\le 4L^4 n\mu(A_n(X_1))^2 \\
&+ 4L^2 n\sum_{j=2}^{n}(Y_j - m(X_j))^2\left|\int_{A_n(X_j)} m(x)\mu(dx) + \int_{A_n'(X_j)} m(x)\mu(dx)\right| \\
&\cdot\left|\int_{A_n'(X_j)\setminus A_n(X_j)} m(x)\mu(dx)\right| \\
&\le 4L^4 n\mu(A_n(X_1))^2 + 8L^4 n\sum_{j=2}^{n}\mu(A_n'(X_j))\mu(A_n'(X_j)\setminus A_n(X_j)) \\
&\le 4L^4 n\mu(A_n(X_1))^2 + 8L^4 n\left(\max_{j=2,\dots n}\mu(A_n'(X_j))\right)\mu(A_n(X_1)) \,,
\end{aligned}$$

which implies that

$$
\begin{aligned}
&\mathbb{V}ar\left(\frac{1}{n}\sum_{j=1}^{n} V_{n,j}^2\right) \\
&\leq n\mathbb{E}\left[\left(L_n - L_n^{(1)}\right)^2\right] \\
&\leq 32L^8 n^3 \mathbb{E}\left[\mu(A_n(X_1))^4\right] \\
&+ 128L^8 n^3 \sqrt{\mathbb{E}\left[\max_{j=2,\ldots n} \mu(A'_n(X_j))^4\right]} \sqrt{\mathbb{E}\left[\mu(A_n(X_1))^4\right]} \\
&\leq 32L^8 c_4/n + 128L^8 n \sqrt{\mathbb{E}\left[\sum_{j=2}^{n} \mu(A'_n(X_j))^4\right]} \sqrt{c_4}
\end{aligned}
$$

by Lemma 1. Noticing that

$$
\mathbb{E}\left[\sum_{j=2}^{n} \mu(A'_n(X_j))^4\right] = (n-1)\mathbb{E}\left[\mu(A'_n(X_2))^4\right] = O(n^{-3})
$$

by Lemma 1, we obtain (4.17).

Proof of Theorem 3

As we mentioned in the proof (4.4), for given D_n, S_n is an average of i.i.d. random variables bounded by L^2. Therefore, by the Hoeffding inequality, one has

$$
\mathbb{P}\{|U_n| > \epsilon \mid D_n\} \leq 2e^{-n\epsilon^2/(2L^4)} .
$$

For the term V_n, apply the extension of the Efron-Stein inequality for the centered higher moments, which is a slight modification of Theorem 15.5 in Boucheron et al. [4]:

Lemma 4. *Let $Z = (Z_1, \ldots, Z_n)$ be a collection of independent random variables taking values in some measurable set A and denote by $Z^{(i)} = (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$ the collection with the i-th random variable dropped. Let $f : A^n \to \mathbb{R}$ be a measurable real-valued function and the function $g_i : A^{n-1} \to \mathbb{R}$ is obtained from f by dropping the i-th*

argument, $i = 1,\dots,n$. Then for any integer $q \geq 1$,

$$\mathbb{E}\Big[(f(Z)-\mathbb{E}f(Z))^{2q}\Big] \leq (cq)^q\Bigg(\mathbb{E}\Bigg[\Bigg(\sum_{i=1}^n\Big(f(Z)-g_i(Z^{(i)})\Big)^2\Bigg)^q\Bigg] + \mathbb{E}\Bigg[\Bigg(\sum_{i=1}^n\mathbb{E}\Big[\Big(f(Z)-g_i(Z^{(i)})\Big)^2 \mid Z_1,\dots,Z_{i-1},Z_{i-1},\dots,Z_n\Big]\Bigg)^q\Bigg]\Bigg), \tag{4.20}$$

with a universal constant $c < 5.1$.

Proof. If $Z_1,\dots,Z_n,Z_1',\dots,Z_n'$ are i.i.d. and

$$Z'^{(i)} = (Z_1,\dots,Z_{i-1},Z_i',Z_{i+1},\dots,Z_n)$$

then from Theorem 15.5 in [4] one gets

$$\mathbb{E}\Big[(f(Z)-\mathbb{E}f(Z))_+^{2q}\Big] \leq (2\kappa q)^q\mathbb{E}\Big[(V^+)^q\Big]\,,$$

and

$$\mathbb{E}\Big[(f(Z)-\mathbb{E}f(Z))_-^{2q}\Big] \leq (2\kappa q)^q\mathbb{E}\left[(V^-)^q\right]\,,$$

with $\kappa = \sqrt{e}/(2(\sqrt{e}-1)) < 1.271$ and with

$$\begin{aligned} V^+ &\leq \sum_{i=1}^n\mathbb{E}\Big\{(f(Z)-f(Z'^{(i)}))^2 \mid Z_1,\dots,Z_n\Big\} \\ &\leq 2\sum_{i=1}^n\Big((f(Z)-g_i(Z^{(i)}))^2+\mathbb{E}\Big[(g_i(Z^{(i)})-f(Z'^{(i)}))^2 \mid Z_1,\dots,Z_n\Big]\Big) \end{aligned}$$

and

$$V^- \leq 2\sum_{i=1}^n\Big((f(Z)-g_i(Z^{(i)}))^2+\mathbb{E}\Big[(g_i(Z^{(i)})-f(Z'^{(i)}))^2 \mid Z_1,\dots,Z_n\Big]\Big)\,.$$

Therefore, c_r-inequality implies

$$\begin{aligned} &\mathbb{E}\Big[(f(Z)-\mathbb{E}f(Z))^{2q}\Big] \\ &\leq 2(2\kappa q)^q 2^{q-1}\mathbb{E}\Bigg[\Bigg(\sum_{i=1}^n(f(Z)-g_i(Z^{(i)}))^2\Bigg)^q+\Bigg(\sum_{i=1}^n\mathbb{E}\Big[(g_i(Z^{(i)})-f(Z'^{(i)}))^2 \mid Z_1,\dots,Z_n\Big]\Bigg)^q\Bigg]\,. \end{aligned}$$

By the equality

$$\mathbb{E}\Big[(g_i(Z^{(i)}) - f(Z'^{(i)}))^2 \mid Z_1,\dots,Z_n\Big] = \mathbb{E}\Big[(g_i(Z^{(i)}) - f(Z))^2 \mid Z_1,\dots,Z_{i-1},Z_{i-1},\dots,Z_n\Big] ,$$

the lemma is proved. □

Notice that

$$m_n(x) = \sum_{j=1}^{n} Y_j I_{\{x\in A_n(X_j)\}} .$$

Then

$$L_n := \mathbb{E}[S_n \mid D_n] = \int m(x)m_n(x)\mu(dx) = \sum_{j=1}^{n} Y_j \int_{A_n(X_j)} m(x)\mu(dx) .$$

Consider now L_n as a function of n i.i.d. vectors $(X_1,Y_1),\dots,(X_n,Y_n)$. Define $L_n^{(i)}$ as L_n when the pair (X_i,Y_i) is dropped. As in the proof of Theorem 1

$$L_n - L_n^{(i)} = \int_{A_n(X_i)} (Y_{1,n}(x) - Y_{2,n}(x))m(x)\mu(dx) ,$$

where $Y_{2,n}(x)$ denotes the label of the second nearest neighbor $X_{2,n}(x)$ of x among $X_1,\dots,X_n$. Thus,

$$\begin{aligned}\left(L_n - L_n^{(i)}\right)^2 &= \left(\int_{A_n(X_i)} (Y_{1,n}(x) - Y_{2,n}(x))m(x)\mu(dx)\right)^2 \\ &\le (2L^2)^2 \left(\mu(A_n(X_i))\right)^2 .\end{aligned}$$

(4.20) implies that

$$\begin{aligned}\mathbb{E}[|L_n - \mathbb{E}[L_n]|^{2q}] \le (cq)^q (2L^2)^{2q} \Bigg(&\mathbb{E}\left[\left(\sum_{i=1}^{n} \mu(A_n(X_i))^2\right)^q\right] \\ &+ \mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{E}[\mu(A_n(X_i))^2 \mid X_1,\dots,X_{i-1},X_{i-1},\dots,X_n]\right)^q\right]\Bigg) . \qquad (4.21)\end{aligned}$$

Because of

$$\sum_{i=1}^{n} \mu(A_n(X_i)) = 1,$$

the Jensen inequality implies that

$$\left(\sum_{i=1}^{n} \mu(A_n(X_i))^2\right)^q \le \sum_{i=1}^{n} \mu(A_n(X_i))^{q+1} ,$$

and so from Lemma 1 we get

$$\mathbb{E}\left[\left(\sum_{i=1}^{n}\mu(A_n(X_i))^2\right)^q\right] \le \mathbb{E}\left[\sum_{i=1}^{n}\mu(A_n(X_i))^{q+1}\right] \le n^{-q}(q+1)!\,. \tag{4.22}$$

Apply the Jensen inequality twice and Lemma 1:

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{E}[\mu(A_n(X_i))^2 \mid X_1,\dots,X_{i-1},X_{i-1},\dots,X_n]\right)^q\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} n\mathbb{E}[\mu(A_n(X_i))^2 \mid X_1,\dots,X_{i-1},X_{i-1},\dots,X_n]\right)^q\right] \\
&\le \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Big(n\mathbb{E}[\mu(A_n(X_i))^2 \mid X_1,\dots,X_{i-1},X_{i-1},\dots,X_n]\Big)^q\right] \\
&= \mathbb{E}\Big[\Big(n\mathbb{E}[\mu(A_n(X_1))^2 \mid X_2,\dots,X_n]\Big)^q\Big] \\
&\le n^{-q}\mathbb{E}\Big[(n\mu(A_n(X_1)))^{2q}\Big] \\
&\le n^{-q}(2q)!\,.
\end{aligned} \tag{4.23}$$

(4.21), (4.22) and (4.23) imply that

$$\begin{aligned}
\mathbb{P}\{|V_n| \ge \epsilon\} &= \mathbb{P}\{|L_n - \mathbb{E}[L_n]| \ge \epsilon\} \\
&\le \frac{\mathbb{E}[|L_n - \mathbb{E}[L_n]|^{2q}]}{\epsilon^{2q}} \\
&\le 2\epsilon^{-2q}(cq)^q(2L^2)^{2q}n^{-q}(2q)! \\
&\le 2\epsilon^{-2q}(cq)^q(2L^2)^{2q}(2q)^{2q}e^{-2q/3}n^{-q} \\
&\le 2\left(\frac{q^3}{n\epsilon^2/(42L^4)}\right)^q,
\end{aligned}$$

because $c \cdot 4 \cdot 4 \cdot e^{-2/3} < 42$. We assume that $n\epsilon^2/(42eL^4) \ge 1$, otherwise the bound (2.3) is trivial. Put

$$q = \lfloor [n\epsilon^2/(42eL^4)]^{1/3} \rfloor \ge 1\,.$$

Thus,

$$\mathbb{P}\{|V_n| \ge \epsilon\} \le 2\left(\frac{\lfloor [n\epsilon^2/(42eL^4)]^{1/3}\rfloor^3}{n\epsilon^2/(42L^4)}\right)^{\lfloor [n\epsilon^2/(42eL^4)]^{1/3}\rfloor} \le 2e^{-n^{1/3}\epsilon^{2/3}/(42eL^4)^{1/3}+1}\,.$$

Acknowledgements. We thank the referees for their thorough reading of the manuscript, for spotting some errors in the proof and for their suggestions of improving the presentation.

References

[1] Biau, G. and Devroye, L.: *Lectures on the Nearest Neighbor Method*, Springer–Verlag, New York, 2015.

[2] Biau, G. and Györfi, L.: On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51:3965–3973, 2005.

[3] Blum, J. R., Chernoff, H., Rosenblatt, M. and Teicher, H.: Central limit theorems for interexchangeable processes. *Canadian J.of Mathematics*, 10:222–229, 1958.

[4] Boucheron, S., Lugosi, G., and Massart, P.: *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[5] De Brabanter, K., Ferrario, P. G. and Györfi, L.: Detecting ineffective features for nonparametric regression. In *Regularization, Optimization, Kernels, and Support Vector Machines*, ed. by J. A. K. Suykens, M. Signoretto, A. Argyriou, pp. 177–194, Chapman & Hall/CRC Machine Learning and Pattern Recognition Series, 2014.

[6] Devroye, L., Ferrario, P., Györfi, L. and Walk, H.: Strong universal consistent estimate of the minimum mean squared error. In *Empirical Inference - Festschrift in Honor of Vladimir N. Vapnik*, ed. by B. Schölkopf, Z. Luo, and V. Vovk, pp. 143–160, Springer, Heidelberg, 2013.

[7] L. Devroye, L. Györfi and G.Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer–Verlag, New York, 1996.

[8] Devroye, L., Györfi, L., Lugosi, G. and Walk, H.: On the measure of Voronoi cells. *Journal of Applied Probability*, 54:394–408, 2017.

[9] Devroye, L. and Lugosi, G.: Almost sure classification of densities. *J. Nonparametr. Stat.*, 14:675-698, 2002.

[10] Devroye, L., Schäfer, D., Györfi, L. and Walk, H.: The estimation problem of minimum mean squared error. *Statistics and Decisions*, 21:15–28, 2003.

[11] Efron, B. and Stein, C.: The jackknife estimate of variance. *Annals of Statistics*, 9:586–596, 1981.

[12] Evans, D. and Jones, A. J.: Non-parametric estimation of residual moments and covariance. *Proceedings of the Royal Society*, A 464:2831–2846, 2008.

[13] Ferrario, P. G. and Walk, H.: Nonparametric partitioning estimation of residual and local variance based on first and second nearest neighbors. *Journal of Nonparametric Statistics*, 24:1019–1039, 2012.

[14] Gretton, A. and Györfi, L.: Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.

[15] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H.: *A Distribution-Free Theory of Nonparametric Regression*. Springer–Verlag, New York, 2002.

[16] Györfi, L. and Walk, H.: On the asymptotic normality of an estimate of a regression functional. *Journal of Machine Learning Research*, 16:1863–1877, 2015.

[17] Liitiäinen, E., Corona, F. and Lendasse, A.: On nonparametric residual variance estimation. *Neural Processing Letters*, 28:155–167, 2008.

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

[19] Liitiäinen, E., Verleysen, M, Corona, F. and Lendasse, A.: Residual variance estimation in machine learning. *Neurocomputing*, 72:3692–3703, 2009.

[20] Petrov, V. V.: *Sums of Independent Random Variables*. Springer-Verlag, Berlin, 1975.

[21] Weber, N. C.: A martingale approach to central limit theorems for exchangeable random variables. *Journal of Applied Probability*, 17:662–673, 1980.