A nearest neighbor estimate of the residual variance *

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

June 21, 2017

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. The asymptotic variance does not depend on the smoothness of the density of X or of the regression function. A non-asymptotic concentration inequality is also proved. We apply the new estimate for 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 and Pompeu Fabra University, gabor.lugosi@upf.edu

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

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 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 asymptotic variance is bounded by a constant that is independent of the dimension. We also establish a non-asymptotic sub-Gaussian concentration inequality. We apply these results for 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 $X = (X^{(1)}, \dots, X^{(d)})$ 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). For reasons explained below it is convenient to assume that we have $2n$ samples split into two halves as

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

such that $(X,Y),(X_1,Y_1),\ldots,(X_n,Y_n),(X'_1,Y'_1),\ldots,(X'_n,Y'_n)$ are i.i.d.

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 [13].) 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 [5] Devroye, Schäfer, Györfi, and Walk [8], Evans and Jones [10], Liitiäinen, Corona, and Lendasse [15], [16], Liitiäinen, Verleysen, Corona, and Lendasse [17], and Ferrario and Walk [11]. 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 [5] is strongly universal 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. This 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 concentration inequalities. These are the main results of the paper. In Section 3 we describe the variable selection method based on the results of Section 2 and examine its properties. Finally, the proofs are presented in Section 4.

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 construct the nearest-neighbor (1-NN) regression function estimator as follows. Let $X_{1,n}(x)$ be the first nearest neighbor of x among $X_1, \ldots, X_n$ 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 as 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 splitting 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 [5], 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 $S_{x,r}$ 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 $S_{0,1}$. Define $\overline{1} = (1, 0, 0, \ldots, 0) \in \mathbb{R}^d$ and let $\overline{S} = S_{\overline{1},1} \bigcup S_{V,\|V\|}$. Introduce

the random variable

$$W = \frac{\lambda(\overline{S})}{\lambda(S_{0,1})} .$$

Define

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

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

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

Devroye, Györfi, Lugosi, and Walk [6] proved that $1 \le \alpha(d) \le 2$. 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 $n \ge 1$,*

$$\mathbb{V}ar(S_n) \le \frac{33 \cdot L^4}{n}.$$

We believe that a non-asymptotic exponential analog of Theorem 2 also holds. In particular, we conjecture that there exists an absolute constant c such that, if μ has a density and $|Y| < L$, then for all $n \ge 1$,

$$\mathbb{P}\{|S_n - \mathbb{E}S_n| > \epsilon\} \le ce^{-n\epsilon^2/(cL^4)} . \tag{2.3}$$

Unfortunately, we are unable to prove such an inequality. However, we have the following weaker version in which the exponent gets worse as the dimension grows.

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 $\mathbb{R}^d$.

Theorem 3. *Assume that μ has a density and that $|Y| < L$. Then for all $n \geq 1$,*

$$\mathbb{P}\{|S_n - \mathbb{E}S_n| > \epsilon\} \leq 4e^{-n\epsilon^2/(121L^4\gamma_d)} .$$

We prove Theorems 1, 2 and 3 in Section 4.

Application in dimension reduction

In standard nonparametric regression design, one considers a finite number of real-valued features $X^{(i)}$, $i \in I \subset \{1, \ldots, d\}$, and evaluates whether these suffice to explain Y. In case they suffice for the given explanatory task, 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 $(Y, X^{(i)} : i \in I)$. The first phase of any regression estimation process therefore relies on estimates of L^* (even *before* a regression estimate is picked).

For dimension reduction, one needs, in general, to test the hypothesis

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

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 et al. [4]).

Introduce the notation

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

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

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

and

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

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

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

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

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

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

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

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

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

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 the consistency result of Devroye, Ferrario, Györfi, and Walk [5], 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

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

If $\widehat{m}$ and f are Lipschitz continuous and f is bounded away from 0, then, by Devroye, Ferrario, Györfi, and Walk [5],

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

when $d \ge 2$ and

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

when $d \ge 3$.

Thus, under the null hypothesis (3.2),

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

for $d \geq 2$.

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 conditions on the density f may be weakened if X is bounded. In particular, if m is C-Lipschitz and X is bounded, then

$$\begin{aligned} n^{1/d}|\mathbb{E}S_n - \mathbb{E}[m(X)^2]| &= n^{1/d}|\mathbb{E}[m(X)m_n(X)] - \mathbb{E}[m(X)^2]| \\ &= n^{1/d}|\mathbb{E}[m(X)m(X_{1,n}(X))] - \mathbb{E}[m(X)^2]| \\ &= n^{1/d} LC\mathbb{E}\|X_{1,n}(X) - X\| \\ &= O(1) \,. \end{aligned}$$

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 [7], Biau and Györfi [1], Gretton and Györfi [12].

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

$$\mathbb{P}\left\{|T_n - \mathbb{E}T_n| > \epsilon\right\} \leq 8e^{-n\epsilon^2/(484L^4\gamma_d)} \,.$$

Therefore, under the null hypothesis

$$\sum_{n=1}^{\infty} \mathbb{P}\{T_n > a_n\} \leq \sum_{n=1}^{\infty} \left(\mathbb{P}\{T_n - \mathbb{E}T_n > \omega_n \cdot n^{-1/2}\} + \mathbb{1}_{\{\mathbb{E}T_n > \omega_n \cdot n^{-2/d}\}}\right) < \infty$$

whenever $\omega_n \geq 23L^2\sqrt{\gamma_d}\sqrt{\ln n}$ for all n 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 all d variables or even for sets of variables. In such cases, one faces a *multiple testing* problem. In order to analyze such multiple testing procedures, one needs a uniform control over the fluctuations of the test statistic. It is for this reason why it would be important to prove a non-asymptotic concentration inequality as the one conjectured above in (2.3).

Proofs

We prove the variance bound of Theorem 2 first. The proof relies of the following version of the Efron-Stein inequality, see, for example, [3, Theorem 3.1].

Lemma 1. *(Efron-Stein inequality.) Let $X = (X_1, \ldots, X_n)$ be a collection of independent random variables taking values in some measurable set $\mathcal{X}$ and denote by $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ the collection with the i-th random variable dropped. Let $f : \mathcal{X}^n \to \mathbb{R}$ and $g : \mathcal{X}^{n-1} \to \mathbb{R}$ be measurable real-valued functions. Then*

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

Proof of Theorem 2

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

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

where

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

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

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

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, we have

$$\mathbb{V}ar(L_n) \leq 4nL^4 \mathbb{E}\left[\mu(A_n(X_1))^2\right] .$$

Observe that

$$\mathbb{E}\left[\mu(A_n(X_1))^2\right] = \mathbb{P}\{X_{n+1}, X_{n+2} \in A_n(X_1)\} .$$

It suffices to prove that

$$\begin{aligned}
&\mathbb{P}\{X_{n+1}, X_{n+2} \in A_n(X_1)\} \\
&\quad \leq \quad 4\mathbb{P}\{X_{n+1} \text{ and } X_{n+2} \text{ are the nearest neighbors of } X_1 \text{ among } X_2, \ldots, X_{n+2}\} ,
\end{aligned} \tag{4.1}$$

because this implies

$$n\mathbb{E}\left[\mu(A_n(X_1))^2\right] \leq \frac{4n}{\binom{n+1}{2}} ,$$

leading to

$$\mathbb{V}ar\left(\mathbb{E}[S_n \mid D_n]\right) \leq \frac{32L^4}{n} ,$$

and therefore to the desired bound

$$\mathbb{V}ar(S_n) \le \frac{33L^4}{n} .$$

In order to prove (4.1), note that

$$\mathbb{P}\{X_{n+1}, X_{n+2} \in A_n(X_1)\} = \mathbb{E}\Big[(1 - \mu(B_{X,\|X-X_1\|} \cup B_{X',\|X'-X_1\|}))^{n-1}\Big]$$

and that

$$\begin{aligned}
&\mathbb{P}\{X_{n+1} \text{ and } X_{n+2} \text{ are the nearest neighbors of } X_1 \text{ among } X_2, \ldots, X_{n+2}\} \\
&= \mathbb{E}\Big[(1 - \max\{\mu(B_{X_1,\|X-X_1\|}), \mu(B_{X_1,\|X'-X_1\|})\})^{n-1}\Big] .
\end{aligned}$$

(4.1) follows from

$$\begin{aligned}
&\mathbb{P}\Big\{\max\{\mu(B_{X_1,\|X-X_1\|}), \mu(B_{X_1,\|X'-X_1\|})\} \le z\Big\} \\
&= \mathbb{P}\Big\{\mu(B_{X_1,\|X-X_1\|}) \le z, \mu(B_{X_1,\|X'-X_1\|}) \le z\Big\} \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X_1,\|X-X_1\|}) \le z, \mu(B_{X_1,\|X'-X_1\|}) \le z \mid X_1\Big\}\Big] \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X_1,\|X-X_1\|}) \le z \mid X_1\Big\}\mathbb{P}\Big\{\mu(B_{X_1,\|X'-X_1\|}) \le z \mid X_1\Big\}\Big] \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X_1,\|X-X_1\|}) \le z \mid X_1\Big\}^2\Big] \\
&= z^2
\end{aligned}$$

and from

$$\begin{aligned}
&\mathbb{P}\Big\{\mu(B_{X,\|X-X_1\|} \cup B_{X',\|X'-X_1\|}) \le z\Big\} \\
&\ge \mathbb{P}\Big\{2\max\{\mu(B_{X,\|X-X_1\|}), \mu(B_{X',\|X'-X_1\|})\} \le z\Big\} \\
&= \mathbb{P}\Big\{\mu(B_{X,\|X-X_1\|}) \le z/2, \mu(B_{X',\|X'-X_1\|}) \le z/2\Big\} \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X,\|X-X_1\|}) \le z/2, \mu(B_{X',\|X'-X_1\|}) \le z/2 \mid X_1\Big\}\Big] \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X,\|X-X_1\|}) \le z/2 \mid X_1\Big\}\mathbb{P}\Big\{\mu(B_{X',\|X'-X_1\|}) \le z/2 \mid X_1\Big\}\Big] \\
&= \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X,\|X-X_1\|}) \le z/2 \mid X_1\Big\}^2\Big] \\
&\ge \mathbb{E}\Big[\mathbb{P}\Big\{\mu(B_{X,\|X-X_1\|}) \le z/2 \mid X_1\Big\}\Big]^2 \\
&= \mathbb{P}\Big\{\mu(B_{X,\|X-X_1\|}) \le z/2\Big\}^2 \\
&= z^2/4 .
\end{aligned}$$

Proof of Theorem 1

In the proof of Theorem 1 we use the following two lemmas on the measure of Voronoi cells.

Lemma 2. *(Devroye, Györfi, Lugosi, and Walk [6]). Assume that μ has a density. Then there for each $k = 1, 2, \ldots$ there exists a positive constant c_k such that*

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

Lemma 3. *(Devroye, Györfi, Lugosi, and Walk [6]). 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).

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 [14] 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 [18]). 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,

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

and so

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

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 2 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

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

On the right-hand side, the first term converges to 0 by of the dominated convergence theorem, since, by Lemma 6.1 in [13],

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

The second term is bounded by

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

by Lemma 6.3 in [13], where γ_d is introduced in Section 2. Thus, (4.9) is proved and hence so is (4.7). For the proof of (4.8), we have that

$$\begin{aligned}
\int m(x)m_n(x)\mu(dx) &= \sum_{j=1}^{n} \int_{A_n(X_j))} m(x)m_n(x)\mu(dx) \\
&= \sum_{j=1}^{n} Y_j \int_{A_n(X_j))} m(x)\mu(dx). \qquad (4.10)
\end{aligned}$$

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 (4.10) 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,

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

Then

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

and so

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

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,

$$\mathbb{V}ar(W_n) \le L^2 \mathbb{E}\left[\left(n \int_{A_n(X_1)} |m(X_{1,n}(x)) - m(X_{2,n}(x))| \mu(dx) \right)^2 \right] \, . \tag{4.11}$$

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

$$\mathbb{E}\left[Z^2\right] = \mathbb{E}\left[|Z|^{2/3}|Z|^{4/3}\right] \le \mathbb{E}\left[|Z|\right]^{2/3} \mathbb{E}\left[Z^4\right]^{1/3} . \tag{4.12}$$

One has

$$\begin{aligned}
&\mathbb{E}\left[n \int_{A_n(X_1)} |m(X_{1,n}(x)) - m(X_{2,n}(x))| \mu(dx)\right] \\
&\le \mathbb{E}\left[n \int_{A_n(X_1)} |m(X_{1,n}(x)) - m(x)| \mu(dx)\right] + \mathbb{E}\left[n \int_{A_n(X_1)} |m(X_{2,n}(x)) - m(x)| \mu(dx)\right] \\
&= \mathbb{E}\left[\int |m(X_{1,n}(x)) - m(x)| \mu(dx)\right] + \mathbb{E}\left[\int |m(X_{2,n}(x)) - m(x)| \mu(dx)\right] \\
&\to 0
\end{aligned} \tag{4.13}$$

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

$$\begin{aligned}
\mathbb{E}\left[\left(n \int_{A_n(X_1)} |m(X_{1,n}(x)) - m(X_{2,n}(x))| \mu(dx)\right)^4\right] &\le 16 L^4 \mathbb{E}\left[n^4 \mu(A_n(X_1))^4\right] \\
&\le 16 L^4 c_4
\end{aligned} \tag{4.14}$$

by (2.2) and Lemma 2. With the notation

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

(4.11), (4.12), (4.13) and (4.14) 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. [2], Weber [19].) 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)

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

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 2 and Jensen's inequality:

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

Condition (iii) in Theorem 4 is fulfilled if

$$\lim_{n\to\infty} \mathbb{E}[V_{n,1}^2] = \sigma_2^2 \tag{4.15}$$

and

$$\mathbb{V}ar\left(\frac{1}{n}\sum_{j=1}^{n} V_{n,j}^2\right) \to 0. \tag{4.16}$$

We have that

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}[V_{n,1}^2] &= \lim_{n\to\infty} n^2\mathbb{E}\left[(Y_1 - m(X_1))^2\left(\int_{A_n(X_1)} m(x)\mu(dx)\right)^2\right] \\ &= \lim_{n\to\infty} n^2\mathbb{E}\left[(Y_1 - m(X_1))^2 m(X_1)^2\mu(A_n(X_1))^2\right] \qquad (4.17) \\ &= \lim_{n\to\infty} n^2\mathbb{E}\left[(M_2(X_1)m(X_1)^2 - m(X_1)^4)\mu(A_n(X_1))^2\right].\end{aligned}$$

(4.17) follows from

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

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 [6, Section 5]). Lemma 3 implies that

$$\mathbb{E}\left[n^2\mu(A_n(X_1))^2 \mid X_1\right] \to \alpha(d) \quad \text{with probability one.} \qquad (4.18)$$

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 2 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 (4.18) 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 (4.15).

One may check (4.16) 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 2. 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 2, we obtain (4.16).

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

$$\mathbb{P}\{|S_n - \mathbb{E}[S_n \mid D_n]| \geq \epsilon \mid D_n\} \leq 2e^{-n\epsilon^2/(2L^4)} . \tag{4.19}$$

Note that

$$M_n := \mathbb{E}[S_n \mid D_n] = \int m_n(x) m(x) \mu(dx)$$

and

$$m_n(x) = \sum_{j=1}^{n} Y_j \mathbb{1}_{\{X_j \in S_{x,R_n(x)}\}}$$

where $R_n(x) = \|X_{(1,n)}(x) - x\|$. Define $\rho_n(x)$ as the solution of the equation

$$\frac{1}{n} = \mu(S_{x,\rho_n(x)}) .$$

By the assumption that X has a density, the solution always exists. Put

$$m_n^*(x) = \sum_{j=1}^{n} Y_j \mathbb{1}_{\{\|X_j - x\| < \rho_n(x)\}} \,.$$

Define $M_n^{(j)}$, $m_n^{(j)}(x)$, $m^{*(j)}_n(x)$ as M_n, $m_n(x)$, $m^*{}_n(x)$, respectively, when (X_j, Y_j) is replaced by $(\widehat{X}_j, \widehat{Y}_j)$ $(j = 1, \ldots, n)$, where $(X_1, Y_1), \ldots, (X_n, Y_n), (\widehat{X}_1, \widehat{Y}_1), \ldots, (\widehat{X}_n, \widehat{Y}_n)$ are i.i.d. random vectors. Further define $m'^{(j)}_n(x)$ as $m_n(x)$ when (X_j, Y_j) is omitted. We have

$$\begin{aligned}
&\sum_{j=1}^{n} \left(M_n - M_n^{(j)} \right)^2 \\
&= \sum_{j=1}^{n} \left(\int m_n(x) m(x) \mu(dx) - \int m_n^{(j)}(x) m(x) \mu(dx) \right)^2 \\
&\le L \sup_{j=1,\ldots,n} \int |m_n(x) - m_n^{(j)}(x)| \mu(dx) \cdot \sum_{j=1}^{n} \left| \int (m_n(x) - m_n^{(j)}(x)) m(x) \mu(dx) \right| . \qquad (4.20)
\end{aligned}$$

The bounding of the supremum term on the right-hand side may be done by an easy modification of the proof of Theorem 23.7 in Györfi, Kohler, Krzyżak, and Walk [13]. For $j = 1, \ldots n$, we have

$$\begin{aligned}
&\int |m_n(x) - m_n^{(j)}(x)| \mu(dx) \\
&\le \int |m_n^*(x) - m^{*(j)}_n(x)| \mu(dx) + \int |m_n(x) - m_n^*(x)| \mu(dx) + \int |m_n^{(j)}(x) - m^{*(j)}_n(x)| \mu(dx) \,.
\end{aligned}$$

$|m_n^*(x) - m^{*(j)}_n(x)|$ is bounded by $2L$ and can differ from zero only if $\|x - X_j\| < \rho_n(x)$ or $\|x - \widehat{X}_j\| < \rho_n(x)$. Observe that $\|x - X_j\| < \rho_n(x)$ or $\|x - \widehat{X}_j\| < \rho_n(x)$ if and only if $\mu(S_{x,\|x-X_j\|}) < 1/n$ or $\mu(S_{x,\|x-\widehat{X}_j\|}) < 1/n$. The measure of such x's is bounded by $2 \cdot \gamma_d / n$ by [13, Lemma 6.2], and therefore

$$\int |m_n^*(x) - m^{*(j)}_n(x)| \mu(dx) \le \frac{4L\gamma_d}{n} \,.$$

Further

$$\begin{aligned}
|m_n^*(x) - m_n(x)| &= \left| \sum_{j=1}^{n} Y_j \mathbb{1}_{\{X_j \in S_{x,\rho_n(x)}\}} - \sum_{j=1}^{n} Y_j \mathbb{1}_{\{X_j \in S_{x,R_n(x)}\}} \right| \\
&\le L \sum_{j=1}^{n} \left| \mathbb{1}_{\{X_j \in S_{x,\rho_n(x)}\}} - \mathbb{1}_{\{X_j \in S_{x,R_n(x)}\}} \right| .
\end{aligned}$$

By considering the cases $\rho_n(x) \le R_n(x)$ and $\rho_n(x) > R_n(x)$ one gets that $\mathbb{1}_{\{X_j \in S_{x,\rho_n(x)}\}} - \mathbb{1}_{\{X_j \in S_{x,R_n(x)}\}}$ have the same sign for each j. It follows that

$$|m_n^*(x) - m_n(x)| \quad \le \quad L \left| \sum_{j=1}^{n} \mathbb{1}_{\{X_j \in S_{x,\rho_n(x)}\}} - 1 \right| = L|M_n^*(x) - 1| \,,$$

where M_n^* is defined as m_n^* with Y replaced by the constant random variable 1. Thus, as before

$$\int |m_n(x) - m_n^*(x)| \mu(dx) \le L \int |M_n^*(x) - 1| \mu(dx) \le \frac{4L\gamma_d}{n} \,.$$

Analogously,

$$\int |m_n^{(j)}(x) - m^{*(j)}_n(x)| \mu(dx) \le \frac{4L\gamma_d}{n} \,.$$

Therefore

$$\sup_{j=1,\ldots,n} \int |m_n(x) - m_n^{(j)}(x)| \mu(dx) \le \frac{12L\gamma_d}{n} \,. \tag{4.21}$$

Furthermore, as in the proof of (4.3)

$$\begin{aligned}
&\sum_{j=1}^{n} \left| \int (m_n(x) - m_n^{(j)}(x)) m(x) \mu(dx) \right| \\
&\le 2 \sum_{j=1}^{n} \left| \int (m_n(x) - m'^{(j)}_n(x)) m(x) \mu(dx) \right| \\
&= 2 \sum_{j=1}^{n} \left| \int \sum_{i=1}^{n} Y_i \mathbb{1}_{\{x \in A_n(X_i)\}} m(x) \mu(dx) - \int \sum_{i \in \{1,\ldots,n\} \setminus \{j\}} Y_i \mathbb{1}_{\{x \in A_n^{(j)}(X_i)\}} m(x) \mu(dx) \right|,
\end{aligned}$$

where $\{A_n^{(j)}(X_i), i \in \{1,\dots,n\} \setminus \{j\}\}$ is the Voronoi partition, when X_j is omitted from the sample. Let $Y_{2,n}(x)$ be the label of $X_{2,n}(x)$. Then

$$\begin{aligned}
&\sum_{j=1}^{n} \left| \int (m_n(x) - m_n^{(j)}(x)) m(x) \mu(dx) \right| \\
&\leq 2 \sum_{j=1}^{n} \left| \sum_{i=1}^{n} Y_i \int_{A_n(X_i)} m(x)\mu(dx) - \sum_{i \in \{1,\dots,n\}\setminus\{j\}} Y_i \int_{A_n^{(j)}(X_i)} m(x)\mu(dx) \right| \\
&= 2 \sum_{j=1}^{n} \left| \int_{A_n(X_j)} Y_{1,n}(x) m(x)\mu(dx) - \sum_{i \in \{1,\dots,n\}\setminus\{j\}} \int_{A_n^{(j)}(X_i)\setminus A_n(X_i)} Y_{2,n}(x) m(x)\mu(dx) \right| \\
&= 2 \sum_{j=1}^{n} \left| \int_{A_n(X_j)} Y_{1,n}(x) m(x)\mu(dx) - \int_{A_n(X_j)} Y_{2,n}(x) m(x)\mu(dx) \right| \\
&\leq 2 \sum_{j=1}^{n} \int_{A_n(X_j)} |Y_{1,n}(x) - Y_{2,n}(x)| \cdot |m(x)| \mu(dx) \\
&\leq 4L^2 \,. \qquad (4.22)
\end{aligned}$$

(4.20), (4.21), (4.22) yield

$$\sum_{j=1}^{n} \left(M_n - M_n^{(j)} \right)^2 \leq \frac{48 L^4 \gamma_d}{n} \,.$$

Thus, by the bounded differences inequality

$$\mathbb{P}\{|\mathbb{E}[S_n \mid D_n] - \mathbb{E}[S_n]| \geq \epsilon\} \leq 2e^{-n\epsilon^2/(96L^4\gamma_d)} \,. \qquad (4.23)$$

(4.19) and (4.23) imply that

$$\begin{aligned}
\mathbb{P}\{|S_n - \mathbb{E}[S_n]| \geq \epsilon\} &\leq \mathbb{P}\{|S_n - \mathbb{E}[S_n \mid D_n]| \geq \epsilon/11\} + \mathbb{P}\{|\mathbb{E}[S_n \mid D_n] - \mathbb{E}[S_n]| \geq 10\epsilon/11\} \\
&\leq 2e^{-n\epsilon^2/(2L^4 11^2)} + 2e^{-n\epsilon^2 10^2/(96L^4 11^2 \gamma_d)} \\
&\leq 4e^{-n\epsilon^2/(121 L^4 \gamma_d)} \,,
\end{aligned}$$

where we applied that $\gamma_d \geq 2$.

References

[1] 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.

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

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

[4] 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.

[5] 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.

[6] Devroye, L., Györfi, L., Lugosi, G. and Walk, H.: On the measure of Voronoi cells. Journal of Applied Probability, to appear, 2016.

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

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

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

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

[11] 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.

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

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

[14] 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.

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

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

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

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

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