Mathematisches
Forschungsinstitut
Oberwolfach

Mitglied der Leibniz Gemeinschaft

Oberwolfach Preprints

OWP 2017 - 26

László Györfi and Harro Walk

Detecting Ineffective Features for Pattern Recognition

Mathematisches Forschungsinstitut Oberwolfach gGmbH
Oberwolfach Preprints (OWP) ISSN 1864-7596

Oberwolfach Preprints (OWP)

Starting in 2007, the MFO publishes a preprint series which mainly contains research results related to a longer stay in Oberwolfach. In particular, this concerns the Research in Pairs-Programme (RiP) and the Oberwolfach-Leibniz-Fellows (OWLF), but this can also include an Oberwolfach Lecture, for example.

A preprint can have a size from 1 - 200 pages, and the MFO will publish it on its website as well as by hard copy. Every RiP group or Oberwolfach-Leibniz-Fellow may receive on request 30 free hard copies (DIN A4, black and white copy) by surface mail.

Of course, the full copy right is left to the authors. The MFO only needs the right to publish it on its website *www.mfo.de* as a documentation of the research work done at the MFO, which you are accepting by sending us your file.

In case of interest, please send a **pdf file** of your preprint by email to *rip@mfo.de* or *owlf@mfo.de*, respectively. The file should be sent to the MFO within 12 months after your stay as RiP or OWLF at the MFO.

There are no requirements for the format of the preprint, except that the introduction should contain a short appreciation and that the paper size (respectively format) should be DIN A4, "letter" or "article".

On the front page of the hard copies, which contains the logo of the MFO, title and authors, we shall add a running number (20XX - XX).

We cordially invite the researchers within the RiP or OWLF programme to make use of this offer and would like to thank you in advance for your cooperation.

Imprint:

Mathematisches Forschungsinstitut Oberwolfach gGmbH (MFO)
Schwarzwaldstrasse 9-11
77709 Oberwolfach-Walke
Germany

Tel +49 7834 979 50
Fax +49 7834 979 55
Email admin@mfo.de
URL www.mfo.de

The Oberwolfach Preprints (OWP, ISSN 1864-7596) are published by the MFO.

DOI 10.14760/OWP-2017-26

Detecting ineffective features for pattern recognition*

László Györfi† Harro Walk‡

October 14, 2017

Abstract

For a binary classification problem, the hypothesis testing is studied, that a component of the observation vector is not effective, i.e., that component carries no information for the classification. We introduce nearest neighbor and partitioning estimates of the Bayes error probability, which result in a strongly consistent test.

AMS Classification: 62G10.

Key words and phrases: classification, Bayes error probability, dimension reduction, strongly consistent test

*This research was supported through the programme "Research in Pairs" by the Mathematisches Forschungsinstitut Oberwolfach in 2017.

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

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

1 The testing problem for classification

Pattern recognition in the case of two classes concerns that for a given random observation (feature) vector one has to decide on the binary valued, random label such that the probability of error is minimal. If the joint distribution of the observation vector and the label is known, then the optimal decision, called Bayes decision, can be derived. In statistical pattern recognition this distribution is unknown, instead we are given random samples, from which some estimates of the Bayes decision can be constructed. The rate of convergence of any pattern recognition rule is very sensitive to the dimension of the observation vector. Thus, the dimension reduction is crucial before constructing the pattern recognition rule.

Dimension reduction without loosing information means that the Bayes error probabilities based on the observation vector leaving out some components and based on the original observation vector, are equal. Thus, there is a nonparametric hypotheses testing problem, where the null hypothesis means that both Bayes error probabilities are equal. In this paper, we introduce two estimates (partitioning estimate and nearest neighbor estimate) of the difference of the Bayes errors. For a given threshold, the null hypothesis is accepted if the difference of the Bayes error estimates is less than the threshold, and otherwise rejected. For the random part of the estimates, we prove exponential concentration inequalities, and for the difference of the expectation of the estimates we show upper bounds. These results imply the strong consistency of the test, which means that with probability one after a random sample size neither the error of the first kind, nor the error of the second kind occurs.

Let the observation (feature) vector X take values in $\mathbb{R}^d$, and let its label Y be ± 1 valued. The task of statistical pattern recognition is to decide on Y given X, i.e., one aims to find a decision function g defined on the range of X such that $g(X) = Y$ with large probability. If g is an arbitrary decision function then its error probability is denoted by

$$L(g) = \mathbb{P}\{g(X) \neq Y\}.$$

Put

$$D(x) = \mathbb{E}\{Y \mid X = x\}.$$

It is well-known that the Bayes decision g^* minimizes the error probability:

$$g^*(x) = \text{sign } D(x)$$

and

$$L^* = \mathbb{P}\{g^*(X) \neq Y\} = \min_g L(g)$$

denotes its error probability. We have that

$$L(g) - L^* = \mathbb{E}\left\{\mathbb{I}_{\{g(X)\neq g^*(X)\}}|D(X)|\right\}, \tag{1}$$

where $\mathbb{I}$ denotes the indicator function, (cf. Theorem 2.2 in Devroye, Györfi, Lugosi [7]).

The Bayes decision cannot be constructed as long as the distribution of (X, Y) is unknown. Assume, that we observed data

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

consisting of independent and identically distributed copies of (X, Y). Devroye, Györfi, Lugosi [7] contains pattern recognition algorithms with strong universal consistency properties, which means that the error probability of these algorithms tends to the Bayes error probability with probability one for all distribution of (X, Y). However, the rate of convergence of the error probabilities heavily depends on regularity (smoothness) properties of the function D and on the dimension d. Detecting an ineffective feature, which in presence of the other features has no influence on L^*, allows reduction of the dimension from d to $d-1$. The project concerns a proposed hypothesis test on ineffectiveness of a specific feature. The test uses estimates of the difference of Bayes error probability with and without this feature.

Consider the test that the last component $X^{(d)}$ of the observation vector $X = (X^{(1)}, \ldots, X^{(d)})$ is ineffective. Let the transformation T be defined by

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

Neglecting the component $X^{(d)}$ from the observation vector

$$(X^{(1)}, \ldots, X^{(d)})$$

leads to the observation vector

$$\hat{X} = T(X) = (X^{(1)}, \ldots, X^{(d-1)})$$

with reduced dimension $d-1$.

For the notations

$$\hat{D}(\hat{X}) = \mathbb{E}\{Y \mid \hat{X}\}$$

and

$$\hat{L}^* = \mathbb{P}\{\hat{g}^*(\hat{X}) \neq Y\}$$

with

$$\hat{g}^*(\hat{x}) = \text{sign } \hat{D}(\hat{x}),$$

the classification null-hypothesis is defined by

$$\hat{L}^* = L^*. \tag{2}$$

The hypothesis (2) means that the component $X^{(d)}$ of the vector X carries no information, i.e., it has no predictive power.

The obvious solution of this problem would be that one estimates L^* and $\hat{L}^*$ from data, and accept the hypothesis (2) if the difference of the estimates is small. Unfortunately, for the time being there is no such estimate with fast rate of convergence.

We may modify the hypothesis (2) such that the Bayes error probability is replaced by the asymptotic error probability of the first nearest neighbor classification rule:

$$R_{NN} = \mathbb{E}\{\mathbb{E}\{Y \mid X\}(1 - \mathbb{E}\{Y \mid X\})\}.$$

(cf. Cover, Hart [4]). Because of

$$R_{NN} = \mathbb{E}\{Y\} - \mathbb{E}\{\mathbb{E}\{Y \mid X\}^2\},$$

the modified hypothesis is defined by

$$\mathbb{E}\{D(X)^2\} = \mathbb{E}\{\hat{D}(\hat{X})^2\}. \tag{3}$$

(Cf. De Brabanter et al. [3].) There are several nearest neighbor based estimates of $\mathbb{E}\{D(X)^2\}$ and $\mathbb{E}\{\hat{D}(\hat{X})^2\}$ with fast rate of convergence. (Cf. Devroye et al. [9], Devroye et al. [8], Evans and Jones [10], Ferrario and Walk [11], Liitiäinen et al. [17], [18], Liitiäinen et al. [19].) Therefore the problem of the hypothesis (3) is easier. The hypothesis (3) is equivalent to the regression hypothesis

$$D(X) = \hat{D}(\hat{X}) \tag{4}$$

a.s. Therefore (4) implies (2), i.e., if a component is ineffective for regression then it is ineffective for classification, too. The reverse is not true.

For $g = -g^*$, (1) implies that

$$(1 - L^*) - L^* = \mathbb{E}\left\{|D(X)|\right\}.$$

Therefore

$$L^* = \frac{1}{2}(1 - \mathbb{E}\left\{|D(X)|\right\}),$$

and similarly

$$\hat{L}^* = \frac{1}{2}(1 - \mathbb{E}\{|\hat{D}(\hat{X})|\}).$$

Thus,

$$\hat{L}^* - L^* = \frac{1}{2}\left(\mathbb{E}\{|D(X)|\} - \mathbb{E}\{|\hat{D}(\hat{X})|\}\right). \tag{5}$$

For an estimate T_n of the functional $\mathbb{E}\{|D(X)|\} - \mathbb{E}\{|\hat{D}(\hat{X})|\}$ and a sequence of positive thresholds $a_n \to 0$, introduce a test such that accept the null hypothesis (2) if

$$T_n \leq a_n,$$

and reject otherwise.

For suitable choice of a_n, we have to show the strong consistency of this test:

(I) under the alternative hypothesis, prove that $\liminf_n T_n > 0$ a.s. for any distribution of (X, Y),

(II) under the null hypothesis (2), find a sequence of positive thresholds a_n such that

$$\sum_{n=1}^{\infty} \mathbb{P}\{T_n > a_n\} < \infty.$$

In order to verify (I) and (II), we plan to investigate the following problems:

(i) prove that $T_n \to \mathbb{E}\{|D(X)|\} - \mathbb{E}\{|\hat{D}(\hat{X})|\}$ a.s. for any distribution of (X, Y),

(ii) derive a concentration inequality for $T_n - \mathbb{E}\{T_n\}$,

(iii) under the null hypothesis (2) and some condition on the regression function D, calculate the rate of convergence of $\mathbb{E}\{T_n\}$.

Based on (5), we create two candidate statistics: a partitioning-based resubstitution statistic and a k-nearest-neighbor-based splitting data statistic and . For both estimates, introduce the corresponding plug-in estimates such that compare the performances. In the analysis of plug-in estimates one usually assumes some conditions:

- D satisfies the Lipschitz condition if for any $x, z \in \mathbb{R}^d$,

$$|D(x) - D(z)| \leq C\|x - z\|, \tag{6}$$

  where $\|\cdot\|$ denotes the Euclidean norm.

- The *margin condition* means that for all $0 < t \leq 1$,

$$\mathbb{E}\left\{\mathbb{I}_{\{|D(X)|\leq t\}}|D(X)|\right\} \leq c^* t^{1+\alpha}. \tag{7}$$

- The *strong density condition* means that for $f(x) > 0$,

$$f(x) \geq f_{min} > 0.$$

- The *modified Lipschitz condition* means that for any $x, z \in \mathbb{R}^d$

$$|D(x) - D(z)| \leq C^* \mu(S_{x,\|x-z\|})^{1/d}. \tag{8}$$

2 A partitioning-based resubstitution estimate

Introduce some notations such that the partition of $\mathbb{R}^d$ is $\mathcal{P}_n = \{A_{n,(j,l)},\, j, l = 1, 2, \dots\}$ and the partition of $\mathbb{R}^{d-1}$ is $\hat{\mathcal{P}}_n = \{\hat{A}_{n,j},\, j = 1, 2, \dots\}$ with

$$A_{n,(j,l)} = \hat{A}_{n,j} \times D_{n,l},$$

where $\{D_{n,l},\, l = 1, 2, \dots\}$ is a partition of $\mathbb{R}$.

Introduce the notations

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i \in A\}} Y_i,\ A \subset \mathbb{R}^d$$

and

$$\hat{\nu}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{\hat{X}_i \in A\}} Y_i,\ A \subset \mathbb{R}^{d-1}.$$

The partitioning classification rule g_n is defined by

$$g_n(x) = \text{sign } \nu_n(A_{n,(j,l)}) \text{ if } x \in A_{n,(j,l)}. \tag{9}$$

Then the *plug-in partitioning error estimate* is defined by

$$\bar{L}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{g_n(X_i') \neq Y_i'\}}. \tag{10}$$

One can show that

$$\mathbb{V}ar(\bar{L}_n) = \mathbb{E}\{(\bar{L}_n - \mathbb{E}\{L(g_n)\})^2\} \leq \frac{c}{n},$$

where c is a universal constant. Thus

$$\mathbb{E}\{|\bar{L}_n - L^*|\} \leq O(1/\sqrt{n}) + \mathbb{E}\{L(g_n)\} - L^*. \tag{11}$$

Kohler and Krzyżak [16] proved that under the margin condition, Lipschitz condition and strong density assumption and for choice

$$h_n = n^{-1/(d+2)}, \tag{12}$$

one gets that

$$\mathbb{E}\{L(g_n)\} - L^* \leq O\left(n^{-\frac{1+\alpha}{d+2}}\right). \tag{13}$$

Let

$$\bar{T}_n = \sum_j \left(\sum_l |\nu_n(A_{n,(j,l)})| - \left| \sum_l \nu_n(A_{n,(j,l)}) \right| \right). \tag{14}$$

be the *resubstitution partitioning error estimate*. Notice that

$$\bar{T}_n = L_n - \hat{L}_n, \tag{15}$$

where

$$L_n = \sum_{j,l} |\nu_n(A_{n,(j,l)})| = \sum_{A \in \mathcal{P}_n} |\nu_n(A)| \tag{16}$$

is the estimate of $\mathbb{E}\{|D(X)|\}$, and

$$\hat{L}_n = \sum_j |\hat{\nu}_n(A_{n,j})| = \sum_{\hat{A} \in \hat{\mathcal{P}}_n} |\hat{\nu}_n(\hat{A})| \tag{17}$$

is the estimate of $\mathbb{E}\{|\hat{D}(\hat{X})|\}$.

For cubic partition of cell size h_n, $h_n \to 0$ and $nh_n^d \to \infty$ imply that

$$L_n \to \mathbb{E}\{|D(X)|\}$$

a.s. and

$$\hat{L}_n \to \mathbb{E}\{|\hat{D}(\hat{X})|\}$$

a.s. (cf. Theorem 23.1 in Györfi et al. [14]). Therefore

$$\bar{T}_n \to \mathbb{E}\left\{|D(X)|\right\} - \mathbb{E}\{|\hat{D}(\hat{X})|\}$$

a.s.

Concerning the resubstitution error estimate for partitioning rule the following inequalities are known (see Sec. 23.2 in Devroye, Györfi and Lugosi [7]): for arbitrary partition

$$\mathbb{V}ar(L_n) \leq \frac{4}{n} \qquad \text{and} \qquad \mathbb{V}ar(\hat{L}_n) \leq \frac{4}{n},$$

which implies that

$$\mathbb{V}ar(\bar{T}_n) \leq \frac{16}{n}.$$

For cubic partition with $h_n \to 0$ and $nh_n^{2d} \to \infty$, Györfi and Horváth [13] and Pintér [20] proved the asymptotic normality of $L_n - \mathbb{E}L_n$ and of $\hat{L}_n - \mathbb{E}\hat{L}_n$. Furthermore, without any condition the McDiarmid inequality implies that

$$\mathbb{P}\{|L_n - \mathbb{E}L_n| > \epsilon\} \leq 2e^{-n\epsilon^2/8} \qquad \text{and} \qquad \mathbb{P}\{|\hat{L}_n - \mathbb{E}\hat{L}_n| > \epsilon\} \leq 2e^{-n\epsilon^2/8}.$$

Thus

$$\mathbb{P}\{|\bar{T}_n - \mathbb{E}\bar{T}_n| > \epsilon\} \leq 4e^{-n\epsilon^2/32}. \tag{18}$$

In the next theorem we bound the expectation of the estimates of the Bayes error probability.

Theorem 1. *Assume that D satisfies the weak margin condition with $0 < \alpha \leq 1$ and the Lipschitz condition, the strong density assumption is satisfied and X is bounded. Then*

$$\begin{aligned}\sum_{A\in\mathcal{P}_n} |\nu(A)| &\leq \mathbb{E}\{L_n\} \\ &\leq \sum_{A\in\mathcal{P}_n} |\nu(A)| + O(1/(nh_n^d)^{1/2}) \left(O(1/(nh_n^d)^{\alpha/2}) + O(h_n^\alpha)\right).\end{aligned}$$

Proof. The Jensen inequality implies the lower bound:

$$\mathbb{E}\{L_n\} = \sum_{A\in\mathcal{P}_n} \mathbb{E}\{|\nu_n(A)|\} \geq \sum_{A\in\mathcal{P}_n} |\mathbb{E}\{\nu_n(A)\}| = \sum_{A\in\mathcal{P}_n} |\nu(A)|.$$

For the upper bound, we use Lemma 5.8 in Devroye and Györfi [6]:

Lemma 1. *Let $Z_1, \dots, Z_n$ be i.i.d. zero mean random variables with variance $\sigma^2 > 0$ and with $\rho = \mathbb{E}|Z_1|^3 < \infty$. Then*

$$\sup_a \left| \mathbb{E} \left| \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n Z_i - a \right| - \mathbb{E}|N - a| \right| \leq \frac{c\rho\sigma^{-3}}{\sqrt{n}},$$

where c is a universal constant and N is $N(0,1)$. For the notation

$$\psi(|a|) = \mathbb{E}|N - a|,$$

observe that

$$|a| \leq \psi(|a|) = |a| - 2|a|\Phi(-|a|) + 2\varphi(-|a|) \leq |a| + 2\varphi(-|a|),$$

where Φ and φ are the standard normal distribution function and density function, respectively.

We apply Lemma 1 for

$$Z_1 = \mathbb{I}_{\{X_1 \in A\}} Y_1 - \mathbb{E}\{\mathbb{I}_{\{X_1 \in A\}} Y_1\}.$$

We have that

$$\sigma_A^2 := \mathbb{V}ar(Z_1) = \mathbb{E}\{(\mathbb{I}_{\{X_1 \in A\}} Y_1 - \mathbb{E}\{\mathbb{I}_{\{X_1 \in A\}} Y_1\})^2\}$$

and

$$\begin{aligned} \rho := \mathbb{E}|Z_1|^3 &= \mathbb{E}\{|\mathbb{I}_{\{X_1 \in A\}} Y_1 - \mathbb{E}\{\mathbb{I}_{\{X_1 \in A\}} Y_1\}|^3\} \\ &\leq 4\mathbb{E}\{|\mathbb{I}_{\{X_1 \in A\}} Y_1 - \mathbb{E}\{\mathbb{I}_{\{X_1 \in A\}} Y_1\}|^2\} = 4\sigma_A^2. \end{aligned}$$

Then Lemma 1 implies that

$$\begin{aligned} \frac{\sqrt{n}}{\sigma_A}\mathbb{E}\{|\nu_n(A)|\} &= \frac{\sqrt{n}}{\sigma_A}\mathbb{E}\{|\nu_n(A) - \nu(A) + \nu(A)|\} \\ &= \mathbb{E}\left\{ \left| \frac{\sqrt{n}}{\sigma_A}(\nu_n(A) - \nu(A)) + \frac{\sqrt{n}}{\sigma_A}\nu(A) \right| \right\} \\ &\leq \psi\left(\frac{\sqrt{n}}{\sigma_A}|\nu(A)| \right) + \frac{c\rho}{\sigma_A^3\sqrt{n}}, \end{aligned}$$

therefore

$$\begin{aligned} \mathbb{E}\{|\nu_n(A)|\} &\leq \frac{\sigma_A}{\sqrt{n}}\psi\left(\frac{\sqrt{n}}{\sigma_A}|\nu(A)| \right) + \frac{c\rho}{\sigma_A^2 n} \\ &\leq |\nu(A)| + 2\frac{\sigma_A}{\sqrt{n}}\varphi\left(-\frac{\sqrt{n}}{\sigma_A}|\nu(A)| \right) + \frac{c\rho}{\sigma_A^2 n} \\ &\leq |\nu(A)| + \sqrt{\frac{2}{\pi}}\sqrt{\frac{\mu(A)}{n}} e^{-\frac{n\nu(A)^2}{2\mu(A)}} + \frac{4c}{n}. \end{aligned}$$

Thus,

$$\mathbb{E}\{L_n\} \leq \sum_{A\in\mathcal{P}_n} |\nu(A)| + \sqrt{\frac{2}{\pi}} \sum_{A\in\mathcal{P}_n} \sqrt{\frac{\mu(A)}{n}} e^{-\frac{n\nu(A)^2}{2\mu(A)}} + \frac{4c\sum_{A\in\mathcal{P}_n,\mu(A)>0} 1}{n}.$$

Since X is bounded, one has

$$\frac{\sum_{A\in\mathcal{P}_n,\mu(A)>0} 1}{n} = O\left(1/(nh_n^d)\right).$$

Put

$$\bar{D}_n(x) = \frac{\nu(A)}{\mu(A)} \text{ if } x \in A.$$

The Jensen inequality and the strong density assumption imply that

$$\begin{aligned}
\sum_{A\in\mathcal{P}_n} \sqrt{\frac{\mu(A)}{n}} e^{-\frac{n\nu(A)^2}{2\mu(A)}} &= \sum_{A\in\mathcal{P}_n} \sqrt{\frac{\mu(A)}{n}} e^{-n\int_A \bar{D}_n(x)^2\mu(dx)/2} \\
&\leq \sum_{A\in\mathcal{P}_n} \frac{1}{\sqrt{n\mu(A)}} \int_A e^{-n\mu(A)\bar{D}_n(x)^2/2}\mu(dx) \\
&\leq \sum_{A\in\mathcal{P}_n} \frac{1}{\sqrt{nf_{min}h_n^d}} \int_A e^{-nf_{min}h_n^d\bar{D}_n(x)^2/2}\mu(dx) \\
&= \frac{1}{\sqrt{f_{min}nh_n^d}} \int e^{-f_{min}(\sqrt{nh_n^d}\bar{D}_n(x))^2/2}\mu(dx).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\int e^{-f_{min}(\sqrt{nh_n^d}|\bar{D}_n(x)|)^2/2}\mu(dx) \\
&\leq \int e^{-f_{min}(\sqrt{nh_n^d}|D(x)|)^2/8}\mu(dx) + \int \mathbb{I}_{\left\{|D(x)|/2\geq|\bar{D}_n(x)|\right\}}\mu(dx).
\end{aligned}$$

Let G be the distribution function of $|D(X)|$. Put

$$H(s) = c^* s^\alpha$$

and

$$w(s) = e^{-f_{min}(\sqrt{nh_n^d}s)^2/8},$$

with

$$w'(s) \leq 0.$$

Because of the margin condition, we have that

$$G(s) \leq H(s).$$

Thus, by partial integration,

$$\begin{aligned}
\int e^{-f_{min}(\sqrt{nh_n^d}|D(x)|)^2/8}\mu(dx) &= \int_0^1 w(s)G(ds) \\
&\leq \int_0^1 w(s)H'(s)ds \\
&= c^*\alpha \int_0^1 e^{-f_{min}(\sqrt{nh_n^d}s)^2/8} s^{\alpha-1} ds \\
&\leq const \int_0^\infty e^{-u} u^{(\alpha-2)/2} du/(nh_n^d)^{\alpha/2} \\
&= O(1/(nh_n^d)^{\alpha/2}). \qquad (19)
\end{aligned}$$

The Lipschitz condition and the margin condition imply that

$$\begin{aligned}
\int \mathbb{I}_{\left\{|D(x)|/2 \geq |\bar{D}_n(x)|\right\}}\mu(dx) &\leq \int \mathbb{I}_{\left\{|D(x)|/2 < |D(x)-\bar{D}_n(x)|\right\}}\mu(dx) \\
&\leq \int \mathbb{I}_{\left\{|D(x)|/2 < C\sqrt{d}h_n\right\}}\mu(dx) \\
&= O(h_n^\alpha). \qquad (20)
\end{aligned}$$

By these relations the theorem is proved. □

Corollary 1. *Assume $d \geq 2$. Under the conditions of Theorem 1 and under the null hypothesis we have that*

$$\mathbb{E}\{\bar{T}_n\} \leq O(h_n^{1+\alpha}) + O(1/(nh_n^d)^{1/2}) \left(O(1/(nh_n^d)^{\alpha/2}) + O(h_n^\alpha)\right).$$

Proof. Theorem 1 implies that

$$\begin{aligned}
&\mathbb{E}\{\bar{T}_n\} \\
&\leq \sum_{A\in\mathcal{P}_n} |\nu(A)| - \sum_{\hat{A}\in\hat{\mathcal{P}}_n} |\hat{\nu}(\hat{A})| + O(1/(nh_n^d)^{1/2}) \left(O(1/(nh_n^d)^{\alpha/2}) + O(h_n^\alpha)\right).
\end{aligned}$$

Therefore, under the null hypothesis we have to show that

$$\sum_{A\in\mathcal{P}_n} |\nu(A)| - \mathbb{E}\left\{|D(X)|\right\} - \left(\sum_{\hat{A}\in\hat{\mathcal{P}}_n} |\hat{\nu}(\hat{A})| - \mathbb{E}\{|\hat{D}(\hat{X})|\}\right) \leq O(h_n^{1+\alpha}).$$

Because of

$$\sum_{A\in\mathcal{P}_n} |\nu(A)| - \mathbb{E}\left\{|D(X)|\right\} \leq 0,$$

we upper bound

$$\mathbb{E}\{|\hat{D}(\hat{X})|\} - \sum_{\hat{A}\in\hat{\mathcal{P}}_n} |\hat{\nu}(\hat{A})| = \sum_{\hat{A}\in\hat{\mathcal{P}}_n} \left(\int_{\hat{A}} |\hat{D}(\hat{x})| \hat{\mu}(d\hat{x}) - \left| \int_{\hat{A}} \hat{D}(\hat{x}) \hat{\mu}(d\hat{x}) \right| \right).$$

Introduce the notations

$$B_0 = \{\hat{x} : \hat{D}(\hat{x}) = 0\} \text{ and } B_+ = \{\hat{x} : \hat{D}(\hat{x}) \geq 0\} \text{ and } B_- = \{\hat{x} : \hat{D}(\hat{x}) < 0\}.$$

If $\hat{A} \cap B_0 = \emptyset$, then

$$\int_{\hat{A}} |\hat{D}(\hat{x})| \hat{\mu}(d\hat{x}) = \left| \int_{\hat{A}} \hat{D}(\hat{x}) \hat{\mu}(d\hat{x}) \right|,$$

otherwise

$$\begin{aligned} \int_{\hat{A}} \hat{D}(\hat{x}) \hat{\mu}(d\hat{x}) &= \int_{\hat{A}\cap B_+} |\hat{D}(\hat{x})| \hat{\mu}(d\hat{x}) - \int_{\hat{A}\cap B_-} |\hat{D}(\hat{x})| \hat{\mu}(d\hat{x}) \\ &= \int_{\hat{A}} |\hat{D}(\hat{x})| \hat{\mu}(d\hat{x}) - 2 \int_{\hat{A}\cap B_-} |\hat{D}(\hat{x})| \hat{\mu}(d\hat{x}). \end{aligned}$$

Thus,

$$\begin{aligned} &\sum_{\hat{A}\in\hat{\mathcal{P}}_n} \left(\int_{\hat{A}} |\hat{D}(\hat{x})| \hat{\mu}(d\hat{x}) - \left| \int_{\hat{A}} \hat{D}(\hat{x}) \hat{\mu}(d\hat{x}) \right| \right) \\ &\leq 2 \sum_{\hat{A}\in\hat{\mathcal{P}}_n, \hat{A}\cap B_0 \neq \emptyset} \int_{\hat{A}\cap B_-} |\hat{D}(\hat{x})| \hat{\mu}(d\hat{x}). \end{aligned}$$

If $\hat{x} \in \hat{A} \cap B_0$, then the Lipschitz condition implies that

$$|\hat{D}(\hat{x})| \leq \bar{C} h_n,$$

and so from the margin condition one gets

$$\begin{aligned} \sum_{\hat{A}\in\hat{\mathcal{P}}_n} \left(\int_{\hat{A}} |\hat{D}(\hat{x})| \hat{\mu}(d\hat{x}) - \left| \int_{\hat{A}} \hat{D}(\hat{x}) \hat{\mu}(d\hat{x}) \right| \right) &\leq 2\mathbb{E}\{|\hat{D}(\hat{X})| \mathbb{I}_{\{|\hat{D}(\hat{X})| \leq \bar{C} h_n\}}\} \\ &= O(h_n^{1+\alpha}). \end{aligned}$$

□

Now, we summarize the consequences for the testing problem.
Concerning (ii), choose

$$b_n = \ln n/\sqrt{n}.$$

Then (18) implies that

$$\sum_{n=1}^{\infty} \mathbb{P}\{|T_n - \mathbb{E}\{T_n\}| > b_n\} < \infty.$$

For (iii), the problem left to find c_n such that

$$\mathbb{E}\{T_n\} \leq c_n,$$

which is done in Corollary 1. Put

$$c_n = h_n^{1+\alpha} + 1/(nh_n^d)^{1/2} \left(1/(nh_n^d)^{\alpha/2} + h_n^{\alpha}\right),$$

which results in the threshold a_n of a strong consistent test:

$$a_n = \ln n(1/\sqrt{n} + c_n).$$

3 k-nearest-neighbor-based splitting data estimate

In this section we consider two nearest-neighbor-based estimates. We fix $x \in \mathbb{R}^d$, and reorder the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ according to increasing values of $\|X_i - x\|$. The reordered data sequence is denoted by

$$(X_{(n,1)}(x), Y_{(n,1)}(x)), \ldots, (X_{(n,n)}(x), Y_{(n,n)}(x)).$$

$X_{(n,k)}(x)$ is the k-th nearest neighbor of x. The tie breaking is done by indices, i.e., if X_i and X_j are equidistant from x, then X_i is declared "closer" if $i < j$. In this paper we assume that the distribution μ of X has a density f, therefore tie happens with probability 0. Choose an integer k less than n, then the k-nearest-neighbor classification rule is

$$g_{n,k}(x) = \text{sign } D_n(x). \tag{21}$$

Concerning the properties of k-nearest-neighbor estimate and k-nearest-neighbor rule see Biau and Devroye [2] and Györfi et al. [14].

Assume additional data

$$\mathcal{D}'_n = \{(X'_1, Y'_1), \ldots, (X'_n, Y'_n)\}$$

independently of $\mathcal{D}_n$. Then the *plug-in nearest neighbor error estimate* is defined by

$$\tilde{L}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{g_{n,k}(X'_i) \neq Y'_i\}}. \tag{22}$$

One can show that

$$\mathbb{V}ar(\tilde{L}_n) = \mathbb{E}\{(\tilde{L}_n - \mathbb{E}\{L(g_{n,k})\})^2\} \leq \frac{c_d}{n},$$

where c_d depends only on the dimension d. Thus

$$\mathbb{E}\{|\tilde{L}_n - L^*|\} \leq O(1/\sqrt{n}) + \mathbb{E}\{L(g_{n,k})\} - L^*. \tag{23}$$

Kohler and Krzyżak [16] proved that under the margin condition, Lipschitz condition and strong density assumption, for choice

$$k = k_n = \lfloor (\log n)^2 n^{2/(d+2)} \rfloor, \tag{24}$$

the upper bound is of order

$$(\log n)^{\frac{2(1+\alpha)}{d}} n^{-\frac{1+\alpha}{d+2}}.$$

Gadat, Klein and Marteau [12] extended this bound such that under the margin condition, Lipschitz condition and the so called strong minimal mass assumption, for choice $k_n = \lfloor n^{2/(d+2)} \rfloor$, one has the order

$$n^{-\frac{1+\alpha}{d+2}}. \tag{25}$$

Audibert and Tsybakov [1] showed that, under the margin condition and the strong density assumption, (25) is the minimax optimal rate of convergence for the class of Lipschitz continuous D, i.e., (25) can be the lower bound for *any* classifier.

Our aim is to construct a test statistic, which is a consistent estimate of the functional $\mathbb{E}\{|D(X)|\} - \mathbb{E}\{|\hat{D}(\hat{X})|\}$. The functional $\mathbb{E}\{|D(X)|\} - \mathbb{E}\{|\hat{D}(\hat{X})|\}$ depends on the regression function D. Therefore a functional estimate will depend on nonparametric regression estimate of D. The k-nearest-neighbor estimate of D is

$$D_{n,k}(x) = D_n(x) = \frac{1}{k} \sum_{i=1}^{k} Y_{(n,i)}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i \mathbb{I}_{\{X_i \in S_{x, \|x - X_{(n,k)}(x)\|}\}}}{k/n}, \tag{26}$$

One estimates the regression function D by the k-nearest-neighbor regression estimate D_n from the samples $\mathcal{D}_n$, while the k-nearest-neighbor regression estimate $\hat{D}_n$ is the estimate of $\hat{D}$ from the samples

$$\hat{\mathcal{D}}_n = \{(\hat{X}_1, Y_1), \ldots, (\hat{X}_n, Y_n)\}.$$

The *splitting data nearest neighbor error estimate* is defined as follows:

$$T'_n = L_n - \hat{L}_n. \tag{27}$$

where

$$L_n = \frac{1}{n}\sum_{i=1}^{n} |D_n(X'_i)|$$

and

$$\hat{L}_n = \frac{1}{n}\sum_{i=1}^{n} |\hat{D}_n(\hat{X}'_i)|.$$

Next we show (I):

$$T'_n \to \mathbb{E}\{|D(X)|\} - \mathbb{E}\{|\hat{D}(\hat{X})|\} \tag{28}$$

a.s.

Notice that

$$\mathbb{E}\{L_n \mid \mathcal{D}_n\} = \int |D_n(x)|\mu(dx)$$

and

$$\mathbb{E}\{L_n\} = \mathbb{E}\left\{\int |D_n(x)|\mu(dx)\right\}.$$

Theorem 2. *One has that*

$$\mathbb{P}\{||L_n - \mathbb{E}\{L_n\}| > \epsilon\} \leq 6e^{-n\epsilon^2/(128\gamma_d^2)},$$

where γ_d is the minimal number of cones of angel $\pi/3$ centered at 0 such that their union covers $\mathbb{R}^d$.

Proof. Consider the following decomposition

$$\begin{aligned} L_n - \mathbb{E}\{L_n\} = {} & L_n - \mathbb{E}\{L_n \mid \mathcal{D}_n\} \\ & + \int |D_n(x)|\mu(dx) - \mathbb{E}\left\{\int |D_n(x)|\mu(dx)\right\}. \end{aligned} \tag{29}$$

The Hoeffding inequality implies that

$$\mathbb{P}\left\{|L_n - \mathbb{E}\left\{L_n \mid \mathcal{D}_n\right\}| > \epsilon\right\} \leq 2e^{-2n\epsilon^2}. \tag{30}$$

A modification of the proof of Theorem 23.7 in Györfi et al. [14] results in

$$\mathbb{P}\left\{\left|\int |D_n(x)|\mu(dx) - \mathbb{E}\left\{\int |D_n(x)|\mu(dx)\right\}\right| > \epsilon\right\} \leq 4e^{-n\epsilon^2/(32\gamma_d^2)}. \tag{31}$$

In order to show (31), we prove that, for an appropriate constant C_n,

$$\mathbb{P}\left\{\left|\int |D_n(x)|\mu(dx) - C_n\right| > \epsilon\right\} \leq 4e^{-n\epsilon^2/(32L^2\gamma_d^2)}.$$

Define $\rho_n(x)$ as the solution of the equation

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

Note that the condition that for each x the distribution of the random variable $\|X - x\|$ is absolutely continuous implies that the solution always exists. (This is the only point in the proof where we use this assumption.) Also define

$$D_n^*(x) = \frac{1}{k_n}\sum_{j=1}^{n} Y_j I_{\{\|X_j - x\| < \rho_n(x)\}}.$$

The basis of the proof is the following decomposition:

$$|D_n(x)| \leq |D_n(x) - D_n^*(x)| + |D_n^*(x)|.$$

For the first term on the right-hand side observe that, denoting $R_n(x) = \|X_{(k_n,n)}(x) - x\|$,

$$\begin{aligned} |D_n^*(x) - D_n(x)| &= \frac{1}{k_n}\left|\sum_{j=1}^{n} Y_j I_{\{X_j \in S_{x,\rho_n(x)}\}} - \sum_{j=1}^{n} Y_j I_{\{X_j \in S_{x,R_n(x)}\}}\right| \\ &\leq \frac{1}{k_n}\sum_{j=1}^{n}\left|I_{\{X_j \in S_{x,\rho_n(x)}\}} - I_{\{X_j \in S_{x,R_n(x)}\}}\right|. \end{aligned}$$

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

$$|D_n^*(x) - D_n(x)| \leq \left|\frac{1}{k_n}\sum_{j=1}^{n} I_{\{X_j \in S_{x,\rho_n(x)}\}} - 1\right| = |E_n^*(x) - 1|,$$

where E_n^* is defined as D_n^* with Y replaced by the constant random variable $Y = 1$. Thus,

$$|D_n(x)| \leq |E_n^*(x) - 1| + |D_n^*(x)|. \tag{32}$$

Next we get an exponential bound for the second term on the right-hand side of (32) by McDiarmid's inequality. Fix an arbitrary realization of the data $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and replace (x_i, y_i) by $(\widehat{x}_i, \widehat{y}_i)$, changing the value of $D_n^*(x)$ to $D_{ni}^*(x)$. Then

$$\left| \int |D_n^*(x)| \mu(dx) - \int |D_{ni}^*(x)| \mu(dx) \right| \quad \leq \quad \int |D_n^*(x) - D_{ni}^*(x)| \mu(dx).$$

But $|D_n^*(x) - D_{ni}^*(x)|$ is bounded by $2/k_n$ and can differ from zero only if $\|x - x_i\| < \rho_n(x)$ or $\|x - \widehat{x}_i\| < \rho_n(x)$. Observe that $\|x - x_i\| < \rho_n(x)$ or $\|x - \widehat{x}_i\| < \rho_n(x)$ if and only if $\mu(S_{x,\|x-x_i\|}) < k_n/n$ or $\mu(S_{x,\|x-\hat{x}_i\|}) < k_n/n$. But the measure of such x's is bounded by $2 \cdot \gamma_d k_n/n$ by Lemma 6.2 in Györfi et al. [14]. Therefore,

$$\sup_{x_1, y_1, \ldots, x_n, y_n, \widehat{x}_i, \widehat{y}_i} \int |D_n^*(x) - D_{ni}^*(x)| \mu(dx) \leq \frac{2}{k_n} \frac{2 \cdot \gamma_d k_n}{n} = \frac{4\gamma_d}{n}$$

and, by by McDiarmid's inequality,

$$\mathbb{P}\left\{ \left| \int |D_n^*(x)| \mu(dx) - \mathbb{E} \int |D_n^*(x)| \mu(dx) \right| > \frac{\epsilon}{2} \right\} \leq 2e^{-n\epsilon^2/(32\gamma_d^2)}.$$

Finally, we need a bound for the first term on the right-hand side of (32). This probability may be bounded by McDiarmid's inequality exactly in the same way as for the second term, obtaining

$$\mathbb{P}\left\{ \left| \int |E_n^*(x) - 1| \mu(dx) - \mathbb{E} \int |E_n^*(x) - 1| \mu(dx) \right| > \frac{\epsilon}{2} \right\} \leq 2e^{-n\epsilon^2/(32\gamma_d^2)},$$

and the proof of (31) is completed. (30) and (31) yield the theorem. □

Moreover, Theorem 6.1 in Györfi et al. [14] implies

$$\left| \mathbb{E}\left\{ \int |D_n(x)| \mu(dx) \right\} - \int |D(x)| \mu(dx) \right| \leq \mathbb{E}\left\{ \int |D_n(x) - D(x)| \mu(dx) \right\} \to 0,$$

i.e.,

$$|\mathbb{E}\{L_n\} - \mathbb{E}\{|D(X)|\}| \to 0$$

This together with Theorem 2 implies

$$L_n \to \mathbb{E}\left\{|D(X)|\right\}$$

a.s. Analogously

$$\hat{L}_n \to \mathbb{E}\left\{|\hat{D}(\hat{X})|\right\}$$

a.s. Thus, (28) is verified.

Next we consider the expectation $\mathbb{E}\left\{L_n\right\} = \mathbb{E}\left\{\int |D_n(x)|\mu(dx)\right\}$ of the estimate.

Theorem 3. *Assume that D satisfies the weak margin condition with $0 < \alpha \leq 1$ and the modified Lipschitz condition. Then*

$$\begin{aligned}
&\int |\mathbb{E}\{D_n(x)\}|\mu(dx) \\
&\leq \mathbb{E}\left\{\int |D_n(x)|\mu(dx)\right\} \\
&\leq \int |\mathbb{E}\{D_n(x)\}|\mu(dx) \\
&+ O(1/k^{1/2})\left(O(1/k^{\alpha/2}) + O((k/n)^{\alpha/d})\right) + O((k/n)^{1/d}).
\end{aligned}$$

Proof. The Jensen inequality implies the lower bound. Introduce the notation

$$\begin{aligned}
\bar{D}_{\|x-X_{(n,k)}(x)\|}(x) &= \mathbb{E}\{D_{n,k}(x) \mid \|x - X_{(n,k)}(x)\|\} \\
&= \frac{\mathbb{E}\{Y_1 \mathbb{I}_{\{X_1 \in S_{x,\|x-X_{(n,k)}(x)\|}\}} \mid \|x - X_{(n,k)}(x)\|\}}{k/n} \\
&= \frac{\int_{S_{x,\|x-X_{(n,k)}(x)\|}} D(z)\mu(dz)}{k/n}. \qquad (33)
\end{aligned}$$

We show that for given $\|x - X_{(n,k)}(x)\|$,

$$\sqrt{k}(D_{n,k}(x) - \bar{D}_{\|x-X_{(n,k)}(x)\|}(x)) \xrightarrow{\mathcal{D}} N(0,1) \qquad (34)$$

in probability, and

$$\sqrt{k}(\bar{D}_{\|x-X_{(n,k)}(x)\|}(x) - \mathbb{E}D_{n,k}(x)) \xrightarrow{\mathcal{D}} N(0, D(x)^2), \qquad (35)$$

which imply that

$$\sqrt{k}(D_{n,k}(x) - \mathbb{E}D_{n,k}(x)) \xrightarrow{\mathcal{D}} N(0, 1 + D(x)^2). \qquad (36)$$

(Cf. the proof of Theorem 1 in Györfi and Walk [15].) Because of (26),

$$
\begin{aligned}
&D_{n,k}(x) - \bar{D}_{\|x-X_{(n,k)}(x)\|}(x) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \mathbb{I}_{\{X_i \in S_{x,\|x-X_{(n,k)}(x)\|}\}} - \int_{S_{x,\|x-X_{(n,k)}(x)\|}} D(z)\mu(dz)}{k/n}.
\end{aligned}
$$

Given $\|x - X_{(n,k)}(x)\|$, $D_{n,k}(x) - \bar{D}_{\|x-X_{(n,k)}(x)\|}(x)$ is an average of i.i.d. random variables with mean zero. Therefore

$$
\frac{D_{n,k}(x) - \bar{D}_{\|x-X_{(n,k)}(x)\|}(x)}{\sqrt{\mathbb{E}\{(D_{n,k}(x) - \bar{D}_{\|x-X_{(n,k)}(x)\|}(x))^2 \mid \|x - X_{(n,k)}(x)\|\}}} \xrightarrow{\mathcal{D}} N(0,1) \quad (37)
$$

in probability. We show this asymptotic normality with remainder term such that apply Berry-Esseen type central limit theorem. Firstly, calculate the asymptotic variance.

$$
\begin{aligned}
&\mathbb{E}\{(D_{n,k}(x) - \bar{D}_{\|x-X_{(n,k)}(x)\|}(x))^2 \mid \|x - X_{(n,k)}(x)\|\} \\
&= \frac{\mathbb{E}\left\{\left(\frac{Y_1 \mathbb{I}_{\{X_1 \in S_{x,\|x-X_{(n,k)}(x)\|}\}} - \int_{S_{x,\|x-X_{(n,k)}(x)\|}} D(z)\mu(dz)}{k/n}\right)^2 \mid \|x - X_{(n,k)}(x)\|\right\}}{n} \\
&= \frac{\mathbb{E}\left\{\mathbb{I}_{\{X_1 \in S_{x,\|x-X_{(n,k)}(x)\|}\}} \mid \|x - X_{(n,k)}(x)\|\right\} - \left(\int_{S_{x,\|x-X_{(n,k)}(x)\|}} D(z)\mu(dz)\right)^2}{k^2/n} \\
&= \frac{\mu(S_{x,\|x-X_{(n,k)}(x)\|}) - \left(\int_{S_{x,\|x-X_{(n,k)}(x)\|}} D(z)\mu(dz)\right)^2}{k^2/n}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\frac{\mu(S_{x,\|x-X_{(n,k)}(x)\|}) - \mu(S_{x,\|x-X_{(n,k)}(x)\|})^2}{k/n} \\
&\le k\mathbb{E}\{(D_{n,k}(x) - \bar{D}_{\|x-X_{(n,k)}(x)\|}(x))^2 \mid \|x - X_{(n,k)}(x)\|\} \\
&\le \frac{\mu(S_{x,\|x-X_{(n,k)}(x)\|})}{k/n}.
\end{aligned}
$$

For i.i.d. uniformly distributed $U_1, \ldots, U_n$, let $U_{(1,n)}, \ldots, U_{(n,n)}$ denote the corresponding order statistic. From Section 1.2 in Biau and Devroye [2] we have that

$$
\mu(S_{x,\|x-X_{(n,k)}(x)\|}) \stackrel{\mathcal{D}}{=} U_{(k,n)}. \quad (38)
$$

Therefore

$$\begin{aligned}\frac{U_{(k,n)}-U_{(k,n)}^2}{\mathbb{E}U_{(k,n)}} &\overset{\mathcal{D}}{\leq} k\mathbb{E}\{(D_{n,k}(x)-\bar{D}_{\|x-X_{(n,k)}(x)\|}(x))^2 \mid \|x-X_{(n,k)}(x)\|\} \\ &\overset{\mathcal{D}}{\leq} \frac{U_{(k,n)}}{\mathbb{E}U_{(k,n)}},\end{aligned}$$

which together with $k\to\infty$ and $k/n\to 0$ implies that

$$k\mathbb{E}\{(D_{n,k}(x)-\bar{D}_{\|x-X_{(n,k)}(x)\|}(x))^2 \mid \|x-X_{(n,k)}(x)\|\}\to 1$$

in probabilty (cf. Theorem 1.4 in Biau and Devroye [2]). With the notation

$$Z_1 = Y_1\mathbb{I}_{\{X_1\in S_{x,\|x-X_{(n,k)}(x)\|}\}} - \int_{S_{x,\|x-X_{(n,k)}(x)\|}} D(z)\mu(dz)$$

the Berry-Esseen inequality says that

$$\begin{aligned}&\left|\mathbb{P}\{\sqrt{k}(D_{n,k}(x)-\bar{D}_{\|x-X_{(n,k)}(x)\|}(x))\leq z \mid \|x-X_{(n,k)}(x)\|\}-\Phi(z)\right| \\ &\leq \frac{c\frac{\mathbb{E}\{|Z_1|^3|\|x-X_{(n,k)}(x)\|\}}{\mathbb{E}\{|Z_1|^2|\|x-X_{(n,k)}(x)\|\}^{3/2}}}{\sqrt{n}\,(1+z^3)},\end{aligned}$$

with the universal constant $c>0$. We have that

$$\mathbb{E}\{|Z_1|^3 \mid \|x-X_{(n,k)}(x)\|\}\leq 4\mathbb{E}\{|Z_1|^2 \mid \|x-X_{(n,k)}(x)\|\}$$

The numerator of the right hand side Berry-Esseen inequality is less than

$$\begin{aligned}&\frac{4c}{\mathbb{E}\{|Z_1|^2 \mid \|x-X_{(n,k)}(x)\|\}^{1/2}} \\ &\leq \frac{4c}{\left(\mu(S_{x,\|x-X_{(n,k)}(x)\|})-\mu(S_{x,\|x-X_{(n,k)}(x)\|})^2\right)^{1/2}} \\ &= \left(\frac{k/n}{\mu(S_{x,\|x-X_{(n,k)}(x)\|})-\mu(S_{x,\|x-X_{(n,k)}(x)\|})^2}\right)^{1/2}\frac{4c}{(k/n)^{1/2}}\end{aligned}$$

in probability. Therefore

$$\begin{aligned}&\left|\mathbb{P}\{\sqrt{k}(D_{n,k}(x)-\bar{D}_{\|x-X_{(n,k)}(x)\|}(x))\leq z \mid \|x-X_{(n,k)}(x)\|\}-\Phi(z)\right| \\ &\leq \frac{4c}{\sqrt{k}\,(1+z^3)}\left(\frac{k/n}{\mu(S_{x,\|x-X_{(n,k)}(x)\|})-\mu(S_{x,\|x-X_{(n,k)}(x)\|})^2}\right)^{1/2} \\ &\approx \frac{4c}{\sqrt{k}\,(1+z^3)},\end{aligned}$$

and (34) is proved with a remainder term. For (35), we need

$$\sqrt{k}\frac{\int_{S_{x,\|x-X_{(n,k)}(x)\|}} D(z)\mu(dz) - \mathbb{E}\int_{S_{x,\|x-X_{(n,k)}(x)\|}} D(z)\mu(dz)}{k/n}$$
$$\xrightarrow{\mathcal{D}} N(0, D(x)^2).$$

From the proof of Lemma 6.1 in Biau and Devroye [2]) one gets that

$$\sqrt{k}\left(\frac{U_{(k,n)}}{k/n} - 1\right) \xrightarrow{\mathcal{D}} N(0,1). \tag{39}$$

(39) has been proved by the representation

$$U_{(k,n)} \stackrel{\mathcal{D}}{=} \frac{\sum_{i=1}^{k} E_i}{\sum_{i=1}^{n+} E_i},$$

where $E_1, \ldots, E_{n+1}$ are i.i.d. exponentially distributed random variables. Then

$$\sqrt{k}\left(\frac{U_{(k,n)}}{k/n} - 1\right) \stackrel{\mathcal{D}}{=} \sqrt{k}\left(\frac{1}{k}\sum_{i=1}^{k} E_i - 1\right) + \sqrt{k}\frac{1}{k}\sum_{i=1}^{k} E_i\left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n+1} E_i} - 1\right),$$

which together with Berry-Esseen inequality implies that

$$\int_0^\infty \left|\mathbb{P}\left\{\sqrt{k}\left(\frac{U_{(k,n)}}{k/n} - 1\right) \geq z\right\} - \Phi(-z)\right| dz$$
$$\leq \int_0^\infty \frac{c}{\sqrt{k}\,(1+z^3)} dz + O(\sqrt{k/n}).$$

From (39) we get that

$$\sqrt{k}\left(\frac{D(x)U_{(k,n)}}{k/n} - D(x)\right) = \sqrt{k}\left(\frac{D(x)\mu(S_{x,\|x-X_{(n,k)}(x)\|})}{k/n} - D(x)\right)$$
$$\xrightarrow{\mathcal{D}} N(0, D(x)^2).$$

Therefore, for (35), we need

$$\sqrt{k}\left(\frac{\int_{S_{x,\|x-X_{(n,k)}(x)\|}} D(z)\mu(dz) - D(x)\mu(S_{x,\|x-X_{(n,k)}(x)\|})}{k/n}\right) \to 0$$

in L_1 with certain rate of convergence. By the mean value theorem, there exists a random variable $Z_{x,n}$ taking values in $S_{x,\|x-X_{(n,k)}(x)\|}$ such that

$$D(Z_{x,n})\mu(S_{x,\|x-X_{(n,k)}(x)\|}) = \int_{S_{x,\|x-X_{(n,k)}(x)\|}} D(z)\mu(dz).$$

The modified Lipschitz condition implies that

$$\begin{aligned}
&\sqrt{k}\frac{\mathbb{E}\left\{\left|\int_{S_{x,\|x-X_{(n,k)}(x)\|}} D(z)\mu(dz) - D(x)\mu(S_{x,\|x-X_{(n,k)}(x)\|})\right|\right\}}{k/n} \\
&\le \sqrt{k}\frac{\mathbb{E}\left\{|D(Z_{x,n}) - D(x))|\,\mu(S_{x,\|x-X_{(n,k)}(x)\|})\right\}}{k/n} \\
&\le \sqrt{k}C^*\frac{\mathbb{E}\left\{\mu(S_{x,\|x-X_{(n,k)}(x)\|})^{1+1/d}\right\}}{k/n} \\
&\le \sqrt{k}C^*(k/n)^{1/d}. \qquad (40)
\end{aligned}$$

These limit relations imply the asymptotic normality with remainder term

$$\begin{aligned}
&\int_0^\infty \left|\mathbb{P}\left\{\sqrt{k}\,(D_{n,k}(x) - \mathbb{E}D_{n,k}(x)) \ge z\right\} - \Phi\left(-z/\sqrt{1+D(x)^2}\right)\right| dz \\
&\le \int_0^\infty \frac{2c}{\sqrt{k}\,(1+z^3)}dz + O(\sqrt{k/n}) + \sqrt{k}O((k/n)^{1/d}).
\end{aligned}$$

Thus, by Lemma 1 and its proof we get that

$$\begin{aligned}
&\sqrt{\frac{k}{1+D(x)^2}}\mathbb{E}\{|D_{n,k}(x)|\} \\
&= \sqrt{\frac{k}{1+D(x)^2}}\mathbb{E}\{|D_{n,k}(x) - \mathbb{E}D_{n,k}(x) + \mathbb{E}D_{n,k}(x)|\} \\
&\le \psi\left(\sqrt{\frac{k}{1+D(x)^2}}|\mathbb{E}D_{n,k}(x)|\right) + \frac{c}{\sqrt{k}} + O(\sqrt{k/n}) + \sqrt{k}O((k/n)^{1/d}).
\end{aligned}$$

Therefore

$$
\begin{aligned}
&\mathbb{E}\{|D_{n,k}(x)|\} \\
&\le \sqrt{\frac{1+D(x)^2}{k}}\psi\left(\sqrt{\frac{k}{1+D(x)^2}}|\mathbb{E}D_{n,k}(x)|\right) + \frac{c}{k} + O(1/\sqrt{n}) + O((k/n)^{1/d}) \\
&\le |\mathbb{E}D_{n,k}(x)| + \frac{4}{\sqrt{k}}\varphi\left(-\sqrt{k}|\mathbb{E}D_{n,k}(x)|/\sqrt{2}\right) + O(1/k) + O(1/\sqrt{n}) + O((k/n)^{1/d}) \\
&= |\mathbb{E}D_{n,k}(x)| + \frac{4}{\sqrt{k}}e^{-k|\mathbb{E}D_{n,k}(x)|^2/4} + O(1/k) + O((k/n)^{1/d}),
\end{aligned}
$$

noticing that $O(1/\sqrt{n})$ is comprehended by other O-terms for all pairs (k,n). Similarly to (19) and (20), the margin condition and the modified Lipschitz condition imply that

$$
\begin{aligned}
&\int e^{-(\sqrt{k}|\mathbb{E}D_{n,k}(x)|)^2/4}\mu(dx) \\
&\le \int e^{-(\sqrt{k}|D(x)|)^2/16}\mu(dx) + \int \mathbb{I}_{\left\{|D(x)|/2\ge|\mathbb{E}D_{n,k}(x)|\right\}}\mu(dx) \\
&\le O(1/k^{\alpha/2}) + O((k/n)^{\alpha/d}).
\end{aligned}
$$

□

Corollary 2. *Assume $d \ge 2$. Under the conditions of Theorem 3 and under the null hypothesis we have that*

$$
\mathbb{E}\{T_n'\} \le O(1/k^{1/2})\left(O(1/k^{\alpha/2}) + O((k/n)^{\alpha/d})\right) + O((k/n)^{1/d}).
$$

Proof. Theorem 3 implies that

$$
\begin{aligned}
\mathbb{E}\{T_n'\} \le &\int |\mathbb{E}\{D_n(x)\}|\mu(dx) - \int |\mathbb{E}\{\hat{D}_n(\hat{x})\}|\hat{\mu}(d\hat{x}) \\
&+ O(1/k^{1/2})\left(O(1/k^{\alpha/2}) + O((k/n)^{\alpha/d})\right) + O((k/n)^{1/d}).
\end{aligned}
$$

Therefore, under the null hypothesis we show that

$$
\begin{aligned}
&\int |\mathbb{E}\{D_n(x)\}|\mu(dx) - \int |D(x)|\mu(dx) \\
&- \left(\int |\mathbb{E}\{\hat{D}_n(\hat{x})\}|\hat{\mu}(d\hat{x}) - \int |\hat{D}(\hat{x})|\hat{\mu}(d\hat{x})\right) \\
&\le O((k/n)^{1/d}).
\end{aligned}
$$

Because of

$$\int |\mathbb{E}\{D_n(x)\}|\mu(dx) - \int |D(x)|\mu(dx) \leq 0,$$

we prove the upper bound

$$\int |\hat{D}(\hat{x})|\hat{\mu}(d\hat{x}) - \int |\mathbb{E}\{\hat{D}_n(\hat{x})\}|\hat{\mu}(d\hat{x}) \leq O((k/n)^{1/d}). \tag{41}$$

We show the equivalent assertion

$$\int |D(x)|\mu(dx) - \int |\mathbb{E}\{D_n(x)\}|\mu(dx) \leq O((k/n)^{1/d}) \tag{42}$$

using the notations in the proof of Theorem 3. Noticing

$$\mathbb{E}\left\{\mu(S_{x,\|x-X_{(n,k)}(x)\|})\right\} = k/n$$

obtained by (38), we have

$$\begin{aligned}
&|D(x)| - |\mathbb{E}\{D_n(x)\}| \\
&\leq |\mathbb{E}\{D_n(x)\} - D(x)| \\
&= \left|\mathbb{E}\{\bar{D}_{\|x-X_{(n,k)}(x)\|}(x)\} - \mathbb{E}\left\{D(x)\frac{\mu(S_{x,\|x-X_{(n,k)}(x)\|})}{k/n}\right\}\right| \\
&= \frac{\left|\mathbb{E}\left\{\int_{S_{x,\|x-X_{(n,k)}(x)\|}} D(z)\mu(dz) - D(x)\mu(S_{x,\|x-X_{(n,k)}(x)\|})\right\}\right|}{k/n} \\
&\leq C^*(k/n)^{1/d}
\end{aligned}$$

by (40), and thus (42) and (41). □

Now, we summarize the consequences for the testing problem. Concerning (ii), choose

$$b_n = \ln n/\sqrt{n}. \tag{43}$$

Then Theorem 2 implies that

$$\sum_{n=1}^{\infty} \mathbb{P}\{|T'_n - \mathbb{E}\{T'_n\}| > b_n\} < \infty. \tag{44}$$

For (iii), the problem is to find c_n such that

$$\mathbb{E}\{T'_n\} \leq c_n,$$

which is done in Corollary 2. Therefore, we get

$$c_n = (k/n)^{1/d} + 1/k^{1/2}\left(1/k^{\alpha/2} + (k/n)^{\alpha/d}\right) + (k/n)^{1/d},$$

which results in the threshold a_n of a strong consistent test:

$$a_n = \ln n(1/\sqrt{n} + c_n).$$

References

[1] J-Y. Audibert and A. B. Tsybakov, Fast learning rates for plug-in classifiers. *Annals of Statistics*, 35:608–633, 2007.

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

[3] 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.

[4] Cover, T. M. and Hart, P. E.: Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13:21–27, 1967.

[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. and Györfi, L. *Nonparametric Density Estimation: The L_1 View*. John Wiley, New York, 1985.

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

[8] Devroye, L., Györfi, L., Lugosi, G and Walk, H.: A nearest neighbor estimate of the residual variance. 2017. (submitted)

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

[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] S. Gadat, T. Klein and C. Marteau, Classification with the nearest neighbor rule in general finite dimensional space. *Annals of Statistics*, 44:982–1009, 2016.

[13] Györfi, L. and Horváth, M.: On the asymptotic normality of a resubstitution error estimate, in *Advances in Data Science and Classification*, A. Rizzi, M. Vichi, H. H. Bock (Eds.), Springer, pp. 197-204, 1998.

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

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

[16] M. Kohler and A. Krzyżak, On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE Transactions on Information Theory* 53:1735–1742, 2007.

[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] Pintér, M.: On the rate of convergence of error estimates for the partitioning classification rule. *Theoretical Computer Science*, 284:181–196, 2002.