Mathematisches
Forschungsinstitut
Oberwolfach

Mitglied der Leibniz Gemeinschaft

Oberwolfach Preprints

OWP 2017 - 25

MAIK DÖRING, LÁSZLÓ GYÖRFI AND HARRO WALK

Exact Rate of Convergence of k-Nearest-Neighbor Classification Rule

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-2017-25

Exact rate of convergence of k-nearest-neighbor classification rule*

Maik Döring † László Györfi ‡ Harro Walk §

October 10, 2017

Abstract

A binary classification problem is considered. The excess error probability of the k-nearest neighbor classification rule according to the error probability of the Bayes decision is revisited by a decomposition of the excess error probability into approximation and estimation error. Under a weak margin condition and under a modified Lipschitz condition, tight upper bounds are presented such that one avoids the condition that the feature vector is bounded.

AMS Classification: 62G10.

Key words and phrases: rate of convergence, classification, error probability, k-nearest neighbor rule

*The research of L. Györfi and of H. Walk was supported through the programme "Research in Pairs" by the Mathematisches Forschungsinstitut Oberwolfach in 2017. L. 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.

†Universität Hohenheim, maik.doering@uni-hohenheim.de

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

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

1 Introduction

Let the feature vector X take values in $\mathbb{R}^d$, and let its label Y be ± 1 valued. If g is an arbitrary decision function then its error probability is denoted by

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

Put

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

then the Bayes decision g^* minimizes the error probability:

$$g^*(x) = sign\, D(x)$$

and

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

denotes its error probability.

In the standard model of pattern recognition, we are given training labeled samples, which are independent and identically copies of (X, Y):

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

Based on these labeled samples, one can estimate the regression function D by $\tilde{D}$, and the corresponding plug-in classification rule g derived from $\tilde{D}$ is defined by

$$g(x) = sign\, \tilde{D}(x),$$

where $sign(x) = 1$ for $x > 0$ and $sign(x) = -1$ for $x \leq 0$. Then for any plug-in rule g derived from the regression estimate $\tilde{D}$ we have

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

where $\mathbb{I}$ denotes the indicator function (cf. Theorem 2.2 in Devroye, Györfi and Lugosi [3]).

In the sequel our focus lies on the rate of convergence of the excess error probability $\mathbb{E}\{L(g_{n,k})\} - L^*$, where $g_{n,k}$ is the k-nearest neighbor rule defined as follows. 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 $\|X_i - x\|$, where $\|\cdot\|$ denotes the Euclidean norm. 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. Let $S_{x,r}$ denote the closed Euclidean sphere centered at $x \in \mathbb{R}^d$ with radius $r > 0$. Choose an integer k less than n, then the k-nearest-neighbor estimate of D is

$$D_{n,k}(x) = \frac{1}{k} \sum_{i=1}^{k} Y_{(n,i)}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i \mathbb{I}_{\{X_i \in S_{x,\|x - X_{(n,k)}(x)\|}\}}}{k/n}, \tag{2}$$

and the k-nearest-neighbor classification rule is

$$g_{n,k}(x) = sign\, D_{n,k}(x). \tag{3}$$

Concerning the properties of k-nearest-neighbor rule and the related literature see Biau and Devroye [2].

The main aim of this paper is to show tight upper bounds on the excess error probability $\mathbb{E}\{L(g_{n,k})\} - L^*$ of the k-nearest-neighbor classification rule $g_{n,k}$.

Given the plug-in classification rule g derived from $\tilde{D}$, (1) implies that

$$\mathbb{E}\{L(g)\} - L^* \leq \mathbb{E}\{|D(X) - \tilde{D}(X)|\}.$$

Therefore we may get an upper bound on the rate of convergence of the excess error probability $\mathbb{E}\{L(g_{n,k})\} - L^*$ via the L_1 rate of convergence of the corresponding regression estimation. Then

$$\mathbb{E}\{L(g_{n,k})\} - L^* \leq \mathbb{E}\{|D(X) - D_{n,k}(X)|\}.$$

We may assume that D satisfies the *Lipschitz condition*: there is a constant C such that for any $x, z \in \mathbb{R}^d$

$$|D(x) - D(z)| \leq C\|x - z\|. \tag{4}$$

If D is Lipschitz continuous and X is bounded, then

$$\mathbb{E}\{|D(X) - D_{n,k}(X)|\} \quad \leq \quad c_1(k/n)^{1/d} + \sqrt{1/k}$$

with $d \geq 2$ (cf. Chapter 6 in Györfi et al. [6]), so for $k = c_3 n^{2/(d+2)}$,

$$\mathbb{E}\{L(g_{n,k})\} - L^* \leq c_4 n^{-1/(d+2)}. \tag{5}$$

However, according to Section 6.7 in Devroye, Györfi and Lugosi [3] the classification is easier than L_1 regression function estimation, since the rate of convergence of the error probability depends on the behavior of the function D in the neighborhood of the decision boundary

$$B_0 = \{x; D(x) = 0\}. \tag{6}$$

This phenomenon has been discovered and investigated by Mammen and Tsybakov [8], Tsybakov [13], Audibert and Tsybakov [1], and Kohler and Krzyżak [7], who introduced the (weak) margin condition:

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

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

  where $\alpha > 0$ and $c^* > 0$.

Denote by

$$B_{0,r} = \left\{x; \min_{z \in B_0} \|x - z\| \leq r\right\} \ (r > 0)$$

the closed r-neighborhood of the decision boundary B_0 defined by (6). Let λ be the Lebesgue measure and let $M^*(B_0)$ be the outer surface (Minkowski content) of the decision boundary B_0 defined by

$$M^*(B_0) = \lim_{r \downarrow 0} \frac{\lambda(B_{0,r} \setminus B_0)}{r}.$$

If D satisfies the Lipschitz condition, the density f of X is bounded by f_{max} and $M^*(B_0)$ is finite, then Lemma 2 in Döring, Györfi and Walk [4] implies $\alpha = 1$. Notice that the Lipschitz condition implies $\alpha \leq 1$.

In the analysis of classification rule we use conditions on the density f of X .

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

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

- The *weak density condition* means that there exist $c_{min} > 0$ and $\delta > 0$ such that for $f(x)r^d \leq \delta^d$,

$$\mu(S_{x,r}) \geq c_{min}^d f(x) r^d.$$

Kohler and Krzyżak [7] proved that under the margin condition, Lipschitz condition and strong density assumption, for choice

$$k_n = \lfloor (\log n)^2 n^{2/(d+2)} \rfloor, \tag{8}$$

the order of the upper bound is smaller than (5):

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

Gadat, Klein and Marteau [5] (comprehending also some classes of distributions with unbounded support) extended this bound such that under the margin condition, Lipschitz condition and the so called strong minimal mass assumption, for choice

$$k_n = \lfloor n^{2/(d+2)} \rfloor, \tag{9}$$

one has the order

$$n^{-\frac{1+\alpha}{d+2}}. \tag{10}$$

Audibert and Tsybakov [1] showed that, under the margin condition and the strong density assumption, (10) is the minimax optimal rate of convergence for the class of Lipschitz continuous D, i.e., (10) can be the lower bound for *any* classifier.

For higher order smoothness, one gets better rate of convergence. For weighted nearest neighbor classification including non-weighted k-nearest neighbor classification, Samworth [11], [12], with further references, considered the case when X is bounded, D is continuously differentiable with gradient $\nabla D(x) \neq 0$ for $x \in B_0$, the conditional densities of X given Y are twice differentiable and the density f of X satisfies the strong density assumption. Under some additional conditions on B_0, he in [12] derives the margin condition with $\alpha = 1$ and shows

$$\mathbb{E}\{L(g_{n,k})\} - L^* \leq \frac{c_7}{k} + c_8(k/n)^{4/d},$$

which implies in the order

$$n^{-\frac{4}{d+4}}. \tag{11}$$

Under the margin condition with $\alpha \leq 1$ ($d \geq 2$) and the strong density assumption, Audibert and Tsybakov [1] showed that the order

$$n^{-\frac{2(1+\alpha)}{d+4}} \tag{12}$$

is the minimax optimal rate of convergence for the class of regression functions D, which have Lipschitz continuous gradients, i.e., they are differentiable and the partial derivatives are Lipschitz continuous. Samworth [12] showed that under the assumptions together with Lipschitz continuity of the density function f several weighted nearest neighbor classifiers, particularly the non-weighted k-nearest neighbor classifiers, can attain this minimax rate.

2 Main result

For most of the above cited results, the feature vector X is assumed to be bounded. Therefore, they exclude the classical parametric discrimination problem, where the conditional distribution of X given Y are multidimensional Gaussian distributions. Next, we revisit these bounds such that our main aim is to avoid the condition that X is bounded.

In order to have non-trivial rate of convergence of the classification error probability, one has to assume tail and smoothness conditions. We introduce a new concept of combined tail and smoothness condition, under which we get the known results on the rate of convergence.

Introduce the *modified Lipschitz condition*: there is a constant C^* such that for any $x, z \in \mathbb{R}^d$

$$|D(x) - D(z)| \leq C^* \mu(S_{x,\|x-z\|})^{1/d}. \tag{13}$$

The main result (Theorem 1) establishes rate of convergence under the modified Lipschitz condition such that it extends and sharpens the result of Kohler and Krzyżak [7].

Theorem 1. *Assume that D satisfies the weak margin condition with $0 < \alpha \leq 1$ and the modified Lipschitz condition. If $d \geq 2$, then*

$$\mathbb{E}\{L(g_{n,k})\} - L^* = O(1/k^{(1+\alpha)/2}) + O((k/n)^{(\alpha+1)/d}),$$

and the choice (9) yields the order (10).

Because of (1), we have the following decomposition of the excess error probability:

$$\mathbb{E}\{L(g_{n,k})\} - L^* = \mathbb{E}\left\{\int_{\{sign\, D_{n,k}(x) \neq sign\, D(x)\}} |D(x)|\mu(dx)\right\} \leq I_{n,k} + J_{n,k},$$

where

$$I_{n,k} = \mathbb{E}\left\{\int_{\{sign\, \bar{D}_{\|x-X_{(n,k)}\|}(x) \neq sign\, D(x)\}} |D(x)|\mu(dx)\right\}$$

and

$$J_{n,k} = \mathbb{E}\left\{\int_{\{sign\, D_{n,k}(x) \neq sign\, \bar{D}_{\|x-X_{(n,k)}\|}(x)\}} |D(x)|\mu(dx)\right\}$$

with

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

$I_{n,k}$ is called approximation error, while $J_{n,k}$ is the estimation error.

We split Theorem 1 into three lemmas such that Lemmas 1 and 2 are on the estimation error, while Lemma 3 is on the approximation error.

Introduce the notations

$$\bar{D}_r(x) = \mathbb{E}\{D_{n,k}(x) \mid \|x - X_{(n,k)}(x)\| = r\}$$

and

$$N_{x,r} = \frac{\bar{D}_r(x)^2}{1 - \bar{D}_r(x)^2} \qquad (r > 0).$$

Put

$$\bar{J}_{n,k} = \mathbb{E}\Big\{ \int |D(x)| \Phi \left(-\sqrt{k \cdot N_{x,\|x-X_{(n,k)}(x)\|}} \right) \mu(dx) \Big\},$$

where Φ stands for the standard Gaussian distribution function.

Lemma 1. *We have that*

$$|J_{n,k} - \bar{J}_{n,k}| \leq \mathbb{E}\Big\{ \int \frac{c|D(x)|}{\sqrt{k} + k^2 |\bar{D}_{\|x-X_{(n,k)}(x)\|}(x)|^3} \mu(dx) \Big\},$$

with a universal constant $c > 0$.

Lemma 2. *Under the conditions of Theorem 1, we have that*

$$\bar{J}_{n,k} = O(1/k^{(1+\alpha)/2}) + O((k/n)^{(\alpha+1)/d}),$$

and for the error term,

$$\mathbb{E}\Big\{ \int \frac{|D(x)|}{\sqrt{k} + k^2 |\bar{D}_{\|x-X_{(n,k)}(x)\|}(x)|^3} \mu(dx) \Big\}$$
$$= O(1/k^{(1+\alpha)/2})/\sqrt{k} + O((k/n)^{(\alpha+1)/d})/\sqrt{k}.$$

Lemma 3. *Under the conditions of Theorem 1, we have that*

$$I_{n,k} \leq e^{-(1-\log 2)k} + O((k/n)^{(\alpha+1)/d}).$$

Remark. The modified Lipschitz condition is used in the proofs of Lemmas 2 and 3 in Section 3. We show how to extend these prooofs from other conditions such that avoid the boundedness of X again. One can check that the Lipschitz condition and the strong density assumption imply the modified Lipschitz condition. However, the strong density assumption implies that the support of μ has finite Lebesgue measure. The *local Lipschitz condition* means that for any $x, z \in \mathbb{R}^d$

$$|D(x) - D(z)| \leq \bar{C} f(x)^{1/d} \|x - z\|. \tag{14}$$

For the local Lipschitz condition the Lipschitz factor is proportional to $f(x)^{1/d}$. Thus, the fluctuation of D is small if the density is small. At the end of Section 3 we show that under the local Lipschitz condition and the weak density condition, the proofs of Lemmas 2 and 3 can be modified.

3 Proofs

Proof of Lemma 1

We show the following: For fixed $x \in \mathbb{R}^d$ and $r > 0$, under $0 < \bar{D}_r(x)$ we have that

$$|\mathbb{P}\{D_{n,k}(x) \leq 0 \mid \|x - X_{(n,k)}(x)\| = r\} - \Phi\left(-\sqrt{k \cdot N_{x,r}}\right)| \\ \leq \frac{c}{\sqrt{k}(1-\bar{D}_r(x)^2)^{3/2} + k^2 \cdot |\bar{D}_r(x)|^3}, \tag{15}$$

which implies the lemma. (The case $\bar{D}_r(x) \leq 0$ and $D_{n,k}(x) > 0$ is completely analogous.)

The density of X exists, therefore the conditional distribution of

$$(X_{(n,1)}(x), Y_{(n,1)}(x)), \dots, (X_{(n,k)}(x), Y_{(n,k)}(x))$$

given $\|x - X_{(n,k)}(x)\| = r$ and the distribution of nearest neighbor ordering of the i.i.d. random variables

$$(\tilde{X}_{(r,1)}(x), \tilde{Y}_{(r,1)}(x)), \dots, (\tilde{X}_{(r,k)}(x), \tilde{Y}_{(r,k)}(x))$$

are the same, where the conditional distribution of Y given X and the conditional distribution of $\tilde{Y}$ given $\tilde{X}$ are equal, and the distribution of $\tilde{X}$ is the restriction of μ to the sphere $S_{x,r}$. Therefore

$$\bar{D}_r(x) = \mathbb{E}\left\{\frac{1}{k}\sum_{i=1}^{k} \tilde{Y}_{(r,i)}(x)\right\} = \frac{\int_{S_{x,r}} D(\tilde{x})\mu(d\tilde{x})}{\mu(S_{x,r})}. \tag{16}$$

Introduce the notation

$$Z_i = -\tilde{Y}_{(r,i)}(x).$$

Then

$$\mathbb{P}\{D_{n,k}(x) \leq 0 \mid \|x - X_{(n,k)}(x)\| = r\} \\ = \mathbb{P}\left\{\sum_{i=1}^{k} Z_i \geq 0\right\}$$

$$= \mathbb{P}\left\{ \frac{\sum_{i=1}^k (Z_i - \mathbb{E}\{Z_i\})}{\sqrt{k \mathbb{V}ar(Z_1)}} \geq -\frac{\sqrt{k}\mathbb{E}\{Z_1\}}{\sqrt{\mathbb{V}ar(Z_1)}} \right\}.$$

Because of

$$\mathbb{E}\{Z_1\} = -\bar{D}_r(x) < 0$$

and

$$\mathbb{V}ar(Z_1) = \mathbb{E}\{|Z_1|^2\} - (\mathbb{E}\{Z_1\})^2 = 1 - \bar{D}_r(x)^2$$

we have that

$$\frac{\mathbb{E}\{Z_1\}}{\sqrt{\mathbb{V}ar(Z_1)}} = -\frac{\bar{D}_r(x)}{\sqrt{1 - \bar{D}_r(x)^2}} = -\sqrt{N_{x,r}}$$

Therefore the central limit theorem implies that

$$\begin{aligned} &\mathbb{P}\{D_{n,k}(x) \leq 0 \mid \|x - X_{(n,k)}(x)\| = r\} \\ &= \mathbb{P}\left\{ -\frac{\sum_{i=1}^k (Z_i - \mathbb{E}\{Z_i\})}{\sqrt{k \mathbb{V}ar(Z_1)}} \leq -\sqrt{k N_{x,r}} \right\} \\ &\approx \Phi\left(-\sqrt{k N_{x,r}}\right). \end{aligned}$$

Notice that it is only an approximation. In order to make bounds out of the normal approximation, we refer to Berry-Esseen type central limit theorem (see Theorem 14 in Petrov [10]). Thus,

$$\begin{aligned} &\left| \mathbb{P}\{D_{n,k}(x) \leq 0 \mid \|x - X_{(n,k)}(x)\| = r\} - \Phi\left(-\sqrt{k N_{x,r}}\right) \right| \\ &\leq \frac{c \frac{\mathbb{E}\{|Z_1|^3\}}{\mathbb{V}ar(Z_1)^{3/2}}}{\sqrt{k}\left(1 + \left(\sqrt{k N_{x,r}}\right)^3\right)}, \end{aligned}$$

with the universal constant $30.84 \geq c > 0$ (cf. Michel [9]). Because of $|Z_1| = 1$ we get that

$$c \frac{\mathbb{E}\{|Z_1|^3\}}{\mathbb{V}ar(Z_1)^{3/2}} = \frac{c}{\left(1 - \bar{D}_r(x)^2\right)^{3/2}},$$

hence

$$\begin{aligned} &\left| \mathbb{P}\{D_{n,k}(x) \leq 0 \mid \|x - X_{(n,k)}(x)\| = r\} - \Phi\left(-\sqrt{k N_{x,r}}\right) \right| \\ &\leq \frac{c}{\sqrt{k}(1 - \bar{D}_r(x)^2)^{3/2} + k^2 \cdot |\bar{D}_r(x)|^3} \end{aligned}$$

□

Proof of Lemma 2

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

$$\mu(S_{x,\|x-X_{(n,k)}(x)\|}) \stackrel{\mathcal{D}}{=} U_{(k,n)}. \tag{17}$$

Introduce the abbreviation

$$\bar{D}(x) = \bar{D}_{\|x-X_{(n,k)}(x)\|}(x).$$

Then

$$\begin{aligned}
&\bar{J}_{n,k}\\
&\le \mathbb{E}\Big\{ \int |D(x)| \Phi\left(-\sqrt{k}|\bar{D}(x)|\right) \mu(dx) \Big\}\\
&= \mathbb{E}\Big\{ \int |D(x)| \left(\mathbb{I}_{\{|\bar{D}(x)| \ge |D(x)|/2\}} + \mathbb{I}_{\{|\bar{D}(x)| < |D(x)|/2\}} \right) \Phi\left(-\sqrt{k}|\bar{D}(x)|\right) \mu(dx) \Big\}\\
&\le \int |D(x)| \Phi\left(-\sqrt{k}|D(x)|/2\right) \mu(dx) + \int |D(x)| \mathbb{P}\left\{ |\bar{D}(x)| < |D(x)|/2 \right\} \mu(dx).
\end{aligned}$$

The weak margin condition with α means that

$$G(t) := \mathbb{P}\{0 < |D(X)| \le t\} \le c^* \cdot t^\alpha, \; 0 \le t \le 1.$$

This implies that

$$\begin{aligned}
&\int |D(x)| \Phi\left(-\sqrt{k}|D(x)|/2\right) \mu(dx) = \int_0^1 s \Phi\left(-\sqrt{k}s/2\right) G(ds)\\
&= s\Phi\left(-\sqrt{k}s/2\right) G(s)\Big|_0^1 - \int_0^1 \left[\Phi\left(-\sqrt{k}s/2\right) - s\frac{\sqrt{k}}{2}\Phi'\left(-\sqrt{k}s/2\right) \right] G(s)ds\\
&\le \Phi\left(-\sqrt{k}/2\right) + \int_0^{\sqrt{k}} \frac{u}{2}\Phi'\left(-u/2\right) c^* u^\alpha du k^{-(\alpha+1)/2} = O(k^{-(\alpha+1)/2}).
\end{aligned}$$

We have

$$\begin{aligned}
\mathbb{P}\left\{ |\bar{D}(x)| < |D(x)|/2 \right\} &\le \mathbb{P}\left\{ |D(x)|/2 < |D(x)| - |\bar{D}(x)| \right\}\\
&\le \mathbb{P}\left\{ |D(x)|/2 < |D(x) - \bar{D}(x)| \right\}.
\end{aligned} \tag{18}$$

The modified Lipschitz condition together with (17) implies that

$$\begin{aligned}
&\mathbb{P}\left\{|D(x)|/2 < |D(x) - \bar{D}(x)|\right\} \\
&\le \mathbb{P}\left\{|D(x)|/2 < C^* \mu(S_{x,\|x-X_{(n,k)}(x)\|})^{1/d}\right\} \\
&= \mathbb{P}\left\{|D(x)|/2 < C^* U_{(k,n)}^{1/d}\right\} \\
&= \mathbb{P}\left\{|D(x)|^d/(2C^*)^d < U_{(k,n)}\right\}. \qquad (19)
\end{aligned}$$

Without loss of generality, assume that $C^* \ge 1/2$. Then

$$\begin{aligned}
&\mathbb{P}\left\{|D(x)|/2 < |D(x) - \bar{D}(x)|\right\} \\
&\le \mathbb{P}\left\{\sum_{i=1}^n \mathbb{I}_{\{U_i \le |D(x)|^d/(2C^*)^d\}} < k\right\} \\
&\le \mathbb{I}_{\{|D(x)|^d/(2C^*)^d \ge 2k/n\}} \mathbb{P}\left\{\sum_{i=1}^n \mathbb{I}_{\{U_i \le |D(x)|^d/(2C^*)^d\}} < \frac{n}{2}|D(x)|^d/(2C^*)^d\right\} \\
&+ \mathbb{I}_{\{|D(x)|^d/(2C^*)^d < 2k/n\}} \\
&\le \mathbb{I}_{\{|D(x)|^d/(2C^*)^d \ge 2k/n\}} e^{-\frac{1-\log 2}{2} n|D(x)|^d/(2C^*)^d} + \mathbb{I}_{\{|D(x)|^d/(2C^*)^d < 2k/n\}} \\
&\le e^{-(1-\log 2)k} + \mathbb{I}_{\{|D(x)|^d/(2C^*)^d < 2k/n\}}, \qquad (20)
\end{aligned}$$

where the third inequality follows from Chernoff's exponential inequality. Applying the weak margin condition, we get

$$\begin{aligned}
&\int |D(x)| \mathbb{P}\left\{|\bar{D}(x)| < |D(x)|/2\right\} \mu(dx) \\
&\le \int |D(x)| \mathbb{P}\left\{|D(x)|/2 < |D(x) - \bar{D}(x)|\right\} \mu(dx) \\
&\le e^{-(1-\log 2)k} + O((k/n)^{(\alpha+1)/d}). \qquad (21)
\end{aligned}$$

The error term can be managed similarly:

$$\begin{aligned}
&\mathbb{E}\left\{\int \frac{|D(x)|}{\sqrt{k} + k^2|\bar{D}(x)|^3} \mu(dx)\right\} \\
&= \mathbb{E}\left\{\int \left(\mathbb{I}_{\{|\bar{D}(x)| \ge |D(x)|/2\}} + \mathbb{I}_{\{|\bar{D}(x)| < |D(x)|/2\}}\right) \frac{|D(x)|}{\sqrt{k} + k^2|\bar{D}(x)|^3} \mu(dx)\right\} \\
&\le \frac{1}{\sqrt{k}} \int \frac{|D(x)|}{1 + (\sqrt{k}|D(x)|/2)^3} \mu(dx)
\end{aligned}$$

$$+\frac{1}{\sqrt{k}}\int |D(x)|\mathbb{P}\left\{|\bar{D}(x)|<|D(x)|/2\right\}\mu(dx).$$

For the first term of the right hand side, we have the bound

$$\begin{aligned}
\int \frac{|D(x)|}{1+(\sqrt{k}|D(x)|)^3}\mu(dx) &= \int_0^1 \frac{s}{1+\left(\sqrt{k}s\right)^3}G(ds)\\
&= \frac{s}{1+\left(\sqrt{k}s\right)^3}G(s)\Big|_0^1\\
&- \int_0^1 \frac{1+\left(\sqrt{k}s\right)^3-3s\sqrt{k}\left(\sqrt{k}s\right)^2}{\left(1+\left(\sqrt{k}s\right)^3\right)^2}G(s)ds\\
&\le O(k^{-3/2})+\int_0^1 \frac{3\left(\sqrt{k}s\right)^3}{\left(1+\left(\sqrt{k}s\right)^3\right)^2}cs^{\alpha}ds\\
&\le O(k^{-3/2})+3ck^{-(1+\alpha)/2}\int_0^{\sqrt{k}}\frac{u^{1+\alpha}u^2}{(1+u^3)^2}du\\
&= O(k^{-(\alpha+1)/2}).
\end{aligned}$$

For the second term of the right hand side, apply (21). □

Proof of Lemma 3

We have that

$$\begin{aligned}
I_{n,k} &= \int \mathbb{E}\left\{\mathbb{I}_{\{sign\,\bar{D}_{\|x-X_{(n,k)}(x)\|}(x)\neq sign\,D(x)\}}\cdot|D(x)|\right\}\mu(dx)\\
&\le \int \mathbb{P}\left\{|\bar{D}_{\|x-X_{(n,k)}(x)\|}(x)-D(x)|\ge|D(x)|\right\}\cdot|D(x)|\mu(dx)\\
&\le e^{-(1-\log 2)k}+O((k/n)^{(\alpha+1)/d}),
\end{aligned}$$

as a conclusion by (21). □

Proof of the Remark

Under the local Lipschitz condition and the weak density condition, we have to prove (21). Let $\delta>0$ be from the definition of weak density assumption.

Under these conditions, by (18) we have that

$$\begin{aligned}
&\int |D(x)|\mathbb{P}\left\{|\bar{D}(x)| < |D(x)|/2\right\} \mu(dx) \\
&\leq \int |D(x)|\mathbb{P}\left\{|D(x)|/2 < |D(x) - \bar{D}(x)|\right\} \mu(dx) \\
&\leq \int |D(x)|\mathbb{P}\left\{|D(x)|/2 < \bar{C} f(x)^{1/d} \|x - X_{(n,k)}(x)\|\right\} \mu(dx) \\
&\leq \int |D(x)|\mathbb{P}\left\{|D(x)|/2 < \bar{C} \mu(S_{x,\|x-X_{(n,k)}(x)\|})^{1/d}/c_{min}\right\} \mu(dx) \\
&+ \int |D(x)|\mathbb{P}\left\{f(x)^{1/d}\|x - X_{(n,k)}(x)\| > \delta\right\} \mu(dx).
\end{aligned}$$

The first term of the right hand side is

$$e^{-(1-\log 2)k} + O((k/n)^{(\alpha+1)/d})$$

by the weak margin condition according to (19) and (20). For the second term, we note

$$\begin{aligned}
&\mathbb{P}\left\{f(x)^{1/d}\|x - X_{(n,k)}(x)\| > \delta\right\} \\
&= \mathbb{P}\left\{\|x - X_{(n,k)}(x)\| > \delta/f(x)^{1/d}\right\} \\
&= \mathbb{P}\left\{\sum_{i=1}^{n} \mathbb{I}_{\left\{X_i \in S_{x,\delta/f(x)^{1/d}}\right\}} < k\right\} \\
&\leq \mathbb{I}_{\left\{\mu(S_{x,\delta/f(x)^{1/d}}) \geq 2k/n\right\}} \mathbb{P}\left\{\sum_{i=1}^{n} \mathbb{I}_{\left\{X_i \in S_{x,\delta/f(x)^{1/d}}\right\}} < \frac{n}{2}\mu(S_{x,\delta/f(x)^{1/d}})\right\} \\
&+ \mathbb{I}_{\left\{\mu(S_{x,\delta/f(x)^{1/d}}) < 2k/n\right\}} \\
&\leq \mathbb{I}_{\left\{\mu(S_{x,\delta/f(x)^{1/d}}) \geq 2k/n\right\}} e^{-\frac{1-\log 2}{2} n\mu(S_{x,\delta/f(x)^{1/d}})} + \mathbb{I}_{\left\{\mu(S_{x,\delta/f(x)^{1/d}}) < 2k/n\right\}},
\end{aligned}$$

the latter by Chernoff's exponential inequality. The weak density assumption yields

$$\mathbb{I}_{\left\{\mu(S_{x,\delta/f(x)^{1/d}}) < 2k/n\right\}} \leq \mathbb{I}_{\left\{c_{min}^d \delta^d < 2k/n\right\}}.$$

Thus the second term is bounded by

$$e^{-(1-\log 2)k} + \mathbb{I}_{\left\{c_{min}^d \delta^d < 2k/n\right\}} = e^{-(1-\log 2)k},$$

as soon as

$$c_{min}^d \delta^d \geq 2k/n.$$

□

References

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

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

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

[4] M. Döring, L. Györfi, H. Walk, Exact rate of convergence of kernel-based classification rule. In *Challenges in Statistics and Data Mining*, ed. by S. Matwin, J. Mielniczuk, pp. 71–91, Springer series: Studies in Computational Intelligence, 2015.

[5] 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.

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

[7] 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.

[8] E. Mammen and A. B. Tsybakov, Smooth discrimination analysis. *Annals of Statistics*, 27:1808–1829, 1999.

[9] R. Michel, On the constant in the non-uniform version of the Berry-Esseen theorem. *Z. Wahrsch. Verw. Gebiete*, 55:109–117, 1981.

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

[11] R. J. Samworth, Optimal weighted nearest neighbor classifiers. *Annals of Statistics*, 40:2733–2763, 2012.

[12] R. J. Samworth, Supplement to "Optimal weighted nearest neighbor classifiers". arXiv:1101.5783, 2012.

[13] A. B. Tsybakov, Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.