Journal of Machine Learning Research 18 (2018) 1-16 Submitted 12/17; Revised 05/18; Published 6/18

Rate of Convergence of k-Nearest-Neighbor Classification Rule

Maik Döring MAIK.DOERING@UNI-HOHENHEIM.DE
Institute of Applied Mathematics and Statistics
University of Hohenheim, 70599 Stuttgart, Germany,
Max Rubner Institute,76131 Karlsruhe, Germany

László Györfi GYORFI@CS.BME.HU
Department of Computer Science and Information Theory
Budapest University of Technology and Economics
1111 Budapest, Hungary

Harro Walk HARRO.WALK@T-ONLINE.DE
Institute of Stochastic and Applications
University of Stuttgart
70049 Stuttgart, Germany

Editor: John Shawe-Taylor

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 errors. Under a weak margin condition and under a modified Lipschitz condition or a local Lipschitz condition, tight upper bounds are presented such that one avoids the condition that the feature vector is bounded. The concept of modified Lipschitz condition is applied for discrete distributions, too. As a consequence of both concepts, we present the rate of convergence of L_2 error for the corresponding nearest neighbor regression estimate.

Keywords: rate of convergence, classification, error probability, k-nearest-neighbor rule

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:

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

where $sign(z) = 1$ for $z > 0$ and $sign(z) = -1$ for $z \leq 0$, 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 identical copies of (X, Y):

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

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

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

Then for any plug-in rule g derived from the regression estimate $\tilde{D}$ we have

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

where $\mathbb{I}$ denotes the indicator function (compare Theorem 2.2 in Devroye, Györfi and Lugosi 1996).

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 $(X_1, Y_1), \dots, (X_n, Y_n)$ according to increasing values of $\|X_i - x\|$, where $\| \cdot \|$ denotes the Euclidean norm. The reordered data sequence is denoted by

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

$X_{(n,k)}(x)$ is the k-th nearest neighbor of x. In this paper we assume that tie happens with probability 0. For instance when the distribution μ of X has a density f, this assumption is satisfied. In any case, by adding a randomizing component to X one can ensure that this assumption holds. Choose an integer k less than n, then the k-nearest-neighbor estimate of D is

$$D_{n,k}(x) = \frac{1}{k} \sum_{i=1}^{k} Y_{(n,i)}(x),$$

and the k-nearest-neighbor classification rule is

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

Concerning the properties of k-nearest-neighbor rule and the related literature see Biau and Devroye (2015).

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

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

Under some smoothness assumptions on D one could further upper bound the L_1 rate. For instance we may assume that D satisfies the *Lipschitz condition*: there is a constant C such that for any $x, z \in \mathbb{R}^d$

$$|D(x) - D(z)| \leq C\|x - z\|.$$

If D is Lipschitz continuous and X is bounded with the diameter M of the support of μ, then

$$\mathbb{E}\{|D(X) - D_{n,k}(X)|^2\} \leq c_1 M^2 (k/n)^{2/d} + c_2/k \tag{2}$$

with $d \geq 2$ (compare Chapter 6 in Györfi et al. 2002 and Liitiäinen, Corona and Lendasse 2010), so for $k = \lfloor c_3 n^{2/(d+2)} \rfloor$,

$$\mathbb{E}\{L(g_{n,k})\} - L^* \leq \sqrt{\mathbb{E}\{|D(X) - D_{n,k}(X)|^2\}} \leq c_4 n^{-1/(d+2)}. \tag{3}$$

However, according to Section 6.7 in Devroye, Györfi and Lugosi (1996) 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

$$B_0 = \{x; D(x) = 0\}. \tag{4}$$

This phenomenon has been discovered and investigated by Mammen and Tsybakov (1999); Tsybakov (2004); Audibert and Tsybakov (2007) and Kohler and Krzyżak (2007), who introduced the (weak) margin condition:

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

  $$\mathbb{P}\{0 < |D(X)| \leq t\} \leq c^* \cdot t^\alpha,$$

  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 (4). 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, X has a density f, the density f is bounded by f_{max} and $M^*(B_0)$ is finite, then Lemma 2 in Döring, Györfi and Walk (2015) implies that the weak margin condition holds with $\alpha = 1$. Notice that the Lipschitz condition implies $\alpha \leq 1$.

In the analysis of classification rule one may 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 strong density condition implies that the support of the density f has finite Lebesgue measure, and so this assumption is close to the condition that X is bounded. It is a very restrictive condition, excluding important densities like Gaussian densities.

Kohler and Krzyżak (2007) proved that under the margin condition, Lipschitz condition and strong density assumption, for choice

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

the order of the upper bound is smaller than (3):

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

Gadat, Klein and Marteau (2016) (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

$$k_n = \lfloor n^{2/(d+2)} \rfloor, \tag{5}$$

one has the order

$$n^{-\frac{1+\alpha}{d+2}}. \tag{6}$$

Audibert and Tsybakov (2007) showed that, under the margin condition and the strong density assumption, (6) is the minimax optimal rate of convergence for the class of Lipschitz continuous D, that is, (6) is the lower bound for *any* classifier.

Let $S_{x,r} = \{x' \in \mathbb{R}^d : \|x' - x\| \leq r\}$ and $S^o_{x,r} = \{x' \in \mathbb{R}^d : \|x' - x\| < r\}$ be the closed and open Euclidean ball, respectively, centered at $x \in \mathbb{R}^d$ with radius $r > 0$. In Chaudhuri and Dasgupta (2014) distribution-dependent rates of convergence are provided for the nearest neighbor classification rule in the framework of metric spaces. Therein a smoothness condition with respect to the distribution μ is introduced: For positive constants κ and L it is assumed that

$$\left| D(x) - \frac{1}{\mu(S_{x,r})} \int_{S_{x,r}} D(z)\mu(dz) \right| \leq L\mu(S^o_{x,r})^\kappa$$

for all $r > 0$ and x in the support of μ. Then the regression function D is called (κ, L)-smooth. Chaudhuri and Dasgupta (2014) revisited the order (6) using such a smoothness condition.

For higher order smoothness, one gets better rates of convergence. For weighted nearest neighbor classification including non-weighted k-nearest-neighbor classification, Samworth (2012a,b), 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, Samworth (2012b) derives the margin condition with $\alpha = 1$ and shows

$$\mathbb{E}\{L(g_{n,k})\} - L^* \leq \frac{c_5}{k} + c_6(k/n)^{4/d},$$

which implies the order

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

In Cannings, Berett and Samworth (2017) the order $n^{-\frac{4}{d+4}}$ is revisited combining tail and smoothness conditions. For the feature vector a moment condition instead of boundedness is required. Further it is assumed that in a neighborhood of the decision boundary the function D and the marginal feature density are twice continuously differentiable and that the latter density allows to control the error of a Taylor approximation even in this region. For feature values away from the decision boundary it is assumed that the marginal feature distribution fulfills the strong minimal mass assumption, see Gadat, Klein and Marteau (2016), and that the function D does not approach the decision boundary too fast:

$$\sup_{x\in\mathbb{R}^d\setminus B_{0,r}:f(x)\geq\delta}|D(x)|^{-1}=o(\delta^{-\tau})\text{ as }\delta\to 0\text{ for some }r>0\text{ and for every }\tau>0.$$

Interestingly, the analysis of empirical error minimization rules can avoid the condition that X is bounded, see Binev, Cohen, Dahmen and DeVore (2014) and Blaschzyk and Steinwart (2018).

Under the margin condition with $\alpha\leq 1$ $(d\geq 2)$ and the strong density assumption, Audibert and Tsybakov (2007) showed that the order

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

is the minimax optimal rate of convergence for the class of regression functions D, which have Lipschitz continuous gradients, that is, they are differentiable and the partial derivatives are Lipschitz continuous. Samworth (2012b) 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, attain this minimax rate.

2. Rate of Convergence of the Error Probability for k-NN Classifier

For most of the above cited results, the feature vector X is assumed to be bounded. Whenever the strong density assumption is used, it is implicitly assumed that the feature vector is bounded. 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 and the strong density assumption.

In order to have non-trivial rate of convergence of the classification error probability, one has to assume tail and smoothness conditions. We treat two concepts of combined tail and smoothness condition, under which we get the known minimax rate of convergence.

- The *modified Lipschitz condition* means that there is a constant C^* such that for any $x,z\in\mathbb{R}^d$
$$|D(x)-D(z)|\leq C^*\mu(S^o_{x,\|x-z\|})^{1/d}.$$

Obviously, this definition is equivalent to the corresponding symmetric definition, i.e.,

$$|D(x)-D(z)|\leq C^*\min\{\mu(S^o_{x,\|x-z\|})^{1/d},\mu(S^o_{z,\|x-z\|})^{1/d}\}.$$

In fact, for (κ,L)-smooth regression function D, Chaudhuri and Dasgupta (2014) already introduced the modified Lipschitz condition, where $\kappa=1/d$ and $L=C^*$. The modified

Lipschitz condition is a universal condition not assuming the existence of a density, it holds for any pair of distribution μ and function D of practical interest.

The main result (Theorem 1) establishes rate of convergence under the modified Lipschitz condition by a decomposition of the excess error into approximation and estimation errors such that it extends and sharpens the result of Kohler and Krzyżak (2007) by avoiding the use of the strong density assumption. Furthermore, Theorem 7b in Chaudhuri and Dasgupta (2014) is closely related to Theorem 1 below.

Theorem 1 *Assume that tie happens with probability* 0*,* D *satisfies the weak margin condition with* $0 < \alpha \le 1$ *and the modified Lipschitz condition. Then*

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

and the choice (5) yields the order (6).

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

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

where

$$I_{n,k} = \mathbb{E}\left\{\int_{\{|\overline{D}_{n,k}(x)-D(x)|\ge|D(x)|/2\}} |D(x)|\mu(dx)\right\}$$

and

$$J_{n,k} = \mathbb{E}\left\{\int_{\{|D_{n,k}(x)-\overline{D}_{n,k}(x)|\ge|D(x)|/2\}} |D(x)|\mu(dx)\right\}$$

with

$$\overline{D}_{n,k}(x) = \mathbb{E}\{D_{n,k}(x) \mid X_1, \dots, X_n\} = \frac{1}{k}\sum_{i=1}^{k} D(X_{(n,i)}(x)). \tag{7}$$

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

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

Lemma 2 *If* D *satisfies the weak margin condition with* $0 < \alpha \le 1$*, then*

$$J_{n,k} = O(1/k^{(1+\alpha)/2}). \tag{8}$$

Lemma 3 *Assume that tie happens with probability* 0*. If* D *satisfies the weak margin condition with* $0 < \alpha \le 1$ *and the modified Lipschitz condition holds, then*

$$I_{n,k} \le e^{-(1-\log 2)k} + O((k/n)^{(\alpha+1)/d}). \tag{9}$$

The proofs of these lemmas are in Section 4.

The concept of modified Lipschitz condition can be applied for discrete distributions, too. As an example, assume that the values of X are positive integers:

$$\mathbb{P}\{X = j\} = p_j.$$

For the classical example for slow rate of convergence, put $Y = D(X)$, that means the function D takes ±1 values. Then $L^* = 0$ and D satisfies the Lipschitz condition with $C = 2$. As in the proof of Theorem 7.2 in Devroye, Györfi and Lugosi (1996), for any classifier g_n, the rate of convergence of $\mathbb{E}\{L(g_n)\}$ to zero can be arbitrarily slow by appropriate choice of the distribution $\{p_i\}$ of large tail.

Consider this discrete case with arbitrary function D such that the modified Lipschitz condition has the form

$$|D(j) - D(j')| \leq C^* \mu([j - |j - j'| + 1, j + |j - j'| - 1]). \tag{10}$$

Next we show that under (10) and for any distribution $\{p_i\}$, even with large tail, the slow rate of convergence is excluded. Apply the k-NN rule with tie-breaking by indices, such that the k-NN estimate of D has the form

$$D_{n,k}(j) = \frac{1}{k} \sum_{i=1}^{k} Y_{(n,i)}(j).$$

Proposition 4 *Under the modified Lipschitz condition (10),*

$$\mathbb{E}\{L(g_{n,k})\} - L^* \leq \frac{2 \max_{\{0 \leq z\}} z e^{-z^2/8}}{\sqrt{k}} + e^{-3k/14} + 4C^* k/n.$$

The modified Lipschitz condition is an implicit condition, it is used in the proof of Lemma 3 in Section 4. We show how to extend this proof starting from a second concept such that we avoid the boundedness of X again. One can check that the Lipschitz condition and the strong density condition imply both concepts. However, as we mentioned earlier, the strong density condition is close to the condition, that X is bounded.

In the framework of the second concept we assume that μ has a density f satisfying a mild condition:

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

If v_d denotes the volume of the unit ball $S_{0,1}$, then the Lebesgue density theorem implies that for all f and for almost all x with respect to the Lebesgues measure,

$$\lim_{r \downarrow 0} \frac{\mu(S_{x,r})}{v_d r^d} = f(x),$$

therefore, for small $r > 0$,

$$\mu(S_{x,r}) \approx f(x) v_d r^d.$$

Thus, the weak density condition requires a bit more than the Lebesgue density theorem.

- The *local Lipschitz condition* means that there exists a constant $\bar{C}$ such that for any $x, z \in \mathbb{R}^d$ with $f(x) > 0$

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

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. One may have a symmetric definition meaning that there exists a constant $\bar{C}$ such that for $\min\{f(x), f(z)\} > 0$

$$|D(x) - D(z)| \leq \overline{C} \min\{f(x), f(z)\}^{1/d} \|x - z\|.$$

However, this definition has no additional advantage, just makes the derivation more involved. Because all three sets $\{(x,z); f(x) = 0\}$, $\{(x,z); f(z) = 0\}$, $\{(x,z); f(x) = 0 or f(z) = 0\}$ have zero product measure $\mu \otimes \mu$, both definitions are equivalent. Notice that the local Lipschitz condition is satisfied for any pair of density f and function D of practical interest.

Theorem 5 below states that under the local Lipschitz condition together with the weak density condition instead of the modified Lipschitz condition, the assertion of Theorem 1 remains valid. It will be shown in Section 4 by a modification of the proof of Lemma 3.

Theorem 5 *Assume that μ has a density such that the weak density condition holds. Furthermore, suppose that D satisfies the weak margin condition with $0 < \alpha \leq 1$ and the local Lipschitz condition. Then*

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

and the choice (5) yields the order (6).

On the one hand, note that the conditions of Theorem 1 don't imply the conditions of Theorem 5, because for Theorem 1, even the existence of a density is not required. On the other hand, the local Lipschitz and the weak density conditions imply that the modified Lipschitz condition holds in a neighborhood of x. It means that for all $x, z \in \mathbb{R}^d$ with $f(x) > 0$ and with $\|x - z\| \leq \delta f(x)^{-1/d}$, one has

$$|D(x) - D(z)| \leq \overline{C} f(x)^{1/d} \|x - z\| \leq \overline{C} \mu(S^o_{x,\|x-z\|})^{1/d} / c_{min}.$$

3. Rate of Convergence of the L_2 Error for k-NN Regression Estimator

In this section we summarize the consequences of the previous section for k-NN regression estimation such that we prove (2) without assuming that X is bounded. Usually, (2) is proved such that after applying the Lipschitz condition one investigates

$$\mathbb{E}\{\|X_{(n,k)}(X) - X\|^2\},$$

which involves the condition that X is bounded. Compare Theorem 14.5 in Biau and Devroye (2015), Theorem 6.2 in Györfi et al. (2002) and Theorem 3.2 in Liitiäinen, Corona and Lendasse (2010), also Theorem 2 in Kohler, Krzyżak and Walk (2006), where a moment condition on X is assumed.

Theorem 6 *Put*

$$\sigma^2(x) = \mathbb{E}\left\{(Y - D(X))^2 \mid X = x\right\}.$$

If tie happens with probability 0*, D satisfies the modified Lipschitz condition and $k/n \to 0$, then*

$$\int \mathbb{E}\{(D_{n,k}(x) - D(x))^2\}\mu(dx) \leq (\mathbb{E}\{\sigma^2(X)\} + o(1))/k + 2C^{*2}(k/n)^{2/d}.$$

The proof of Theorem 6 is at the end of Section 4. Similarly to Theorem 5, in Theorem 6 the modified Lipschitz condition can be replaced by the local Lipschitz condition together with the weak density condition.

4. Proofs

In this section we present the proofs of Lemmas 2 and 3, hence Theorem 1, and of Proposition 4, and of Theorems 5 and 6.

Proof of Lemma 2. For a fixed x, Proposition 8.1 in Biau and Devroye (2015) says the following: given $X_1, \dots, X_n$, the random pairs

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

are independent, and

$$\mathbb{E}\{Y_{(n,i)}(x) - D(X_{(n,i)}(x)) \mid X_1, \dots, X_n\} = 0.$$

Therefore, the Hoeffding inequality implies that

$$\begin{aligned}
&\mathbb{P}\{|D_{n,k}(x) - \overline{D}_{n,k}(x)| \geq |D(x)|/2 \mid X_1, \dots, X_n\} \\
&= \mathbb{P}\left\{\left|\frac{1}{k}\sum_{i=1}^{k}(Y_{(n,i)}(x) - D(X_{(n,i)}(x)))\right| \geq |D(x)|/2 \mid X_1, \dots, X_n\right\} \\
&\leq 2e^{-k|D(x)|^2/8}.
\end{aligned}$$

Thus,

$$J_{n,k} \leq 2\int |D(x)|e^{-k|D(x)|^2/8}\mu(dx).$$

The weak margin condition

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

implies by use of partial integration with respect to $G(s)$ that

$$\begin{aligned}\int |D(x)|e^{-k|D(x)|^2/8}\mu(dx) &= \int_0^1 se^{-ks^2/8}G(ds)\\ &= e^{-k/8} - \int_0^1 e^{-ks^2/8}[1-ks^2/4]G(s)ds\\ &\leq e^{-k/8} + \frac{c^*}{4}\int_0^1 e^{-ks^2/8}ks^{2+\alpha}ds\\ &\leq e^{-k/8} + \frac{c^*}{4}k^{-(\alpha+1)/2}\int_0^\infty e^{-u^2/8}u^{2+\alpha}du\\ &= O(k^{-(\alpha+1)/2}).\end{aligned}$$

Thus, (8) is obtained. ■

Proof of Lemma 3. For $U_1, \dots, U_n$ i.i.d. uniformly distributed on $[0,1]$, let $U_{(1,n)}, \dots, U_{(n,n)}$ denote the corresponding order statistic. If tie happens with probability 0, then for any fixed x, $\mu(S_{x,r})$ is continuous in r, which implies that $\mu(S_{x,\|x-X\|})$ is uniformly distributed on $[0,1]$. From Section 1.2 in Biau and Devroye (2015) for any fixed x we have that

$$\mu(S_{x,\|x-X_{(n,k)}(x)\|}) \overset{\mathcal{D}}{=} U_{(k,n)}. \tag{11}$$

Because of

$$\begin{aligned}|D(x) - \overline{D}_{n,k}(x)| &= \left|D(x) - \frac{1}{k}\sum_{i=1}^k D(X_{(n,i)}(x))\right|\\ &\leq \frac{1}{k}\sum_{i=1}^k |D(x) - D(X_{(n,i)}(x))|\end{aligned}$$

the modified Lipschitz condition together with (11) implies that

$$\begin{aligned}&\mathbb{P}\left\{|D(x)|/2 < |D(x) - \overline{D}_{n,k}(x)|\right\}\\ &\leq \mathbb{P}\left\{|D(x)|/2 < C^*\frac{1}{k}\sum_{i=1}^k \mu(S_{x,\|x-X_{(n,i)}(x)\|})^{1/d}\right\}\\ &\leq \mathbb{P}\left\{|D(x)|/2 < C^*\mu(S_{x,\|x-X_{(n,k)}(x)\|})^{1/d}\right\}\\ &= \mathbb{P}\left\{|D(x)|/2 < C^*U_{(k,n)}^{1/d}\right\}\\ &= \mathbb{P}\left\{|D(x)|^d/(2C^*)^d < U_{(k,n)}\right\}.\end{aligned} \tag{12}$$

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

$$\begin{aligned}
&\mathbb{P}\left\{|D(x)|/2 < |D(x) - \overline{D}_{n,k}(x)|\right\} \\
&\leq \mathbb{P}\left\{\sum_{i=1}^{n} \mathbb{I}_{\{U_i \leq |D(x)|^d/(2C^*)^d\}} < k\right\} \\
&\leq \mathbb{I}_{\{|D(x)|^d/(2C^*)^d \geq 2k/n\}} \mathbb{P}\left\{\sum_{i=1}^{n} \mathbb{I}_{\{U_i \leq |D(x)|^d/(2C^*)^d\}} < \frac{n}{2}|D(x)|^d/(2C^*)^d\right\} \\
&+ \mathbb{I}_{\{|D(x)|^d/(2C^*)^d < 2k/n\}} \\
&\leq \mathbb{I}_{\{|D(x)|^d/(2C^*)^d \geq 2k/n\}} e^{-\frac{1-\log 2}{2} n|D(x)|^d/(2C^*)^d} + \mathbb{I}_{\{|D(x)|^d/(2C^*)^d < 2k/n\}} \\
&\leq e^{-(1-\log 2)k} + \mathbb{I}_{\{|D(x)|^d/(2C^*)^d < 2k/n\}}, \qquad (13)
\end{aligned}$$

where the third inequality follows from Chernoff's exponential inequality. Applying the weak margin condition, we get (9) by

$$\begin{aligned}
I_{n,k} &= \int |D(x)| \mathbb{P}\left\{|D(x)|/2 < |D(x) - \overline{D}_{n,k}(x)|\right\} \mu(dx) \\
&\leq e^{-(1-\log 2)k} + O((k/n)^{(\alpha+1)/d}). \qquad (14)
\end{aligned}$$

■

Proof of Proposition 4. The Hoeffding inequality implies

$$\begin{aligned}
&\mathbb{P}\{|D_{n,k}(j) - \overline{D}_{n,k}(j)| \geq |D(j)|/2 \mid X_1, \ldots, X_n\} \\
&= \mathbb{P}\left\{\left|\frac{1}{k}\sum_{i=1}^{k}(Y_{(n,i)}(j) - D(X_{(n,i)}(j)))\right| \geq |D(j)|/2 \mid X_1, \ldots, X_n\right\} \\
&\leq 2e^{-k|D(j)|^2/8}, \quad j \in \mathbb{N},
\end{aligned}$$

from which one gets a rough non-trivial upper bound on the estimation error:

$$\begin{aligned}
\mathbb{E}\left\{\int_{\{|D_{n,k}(x) - \overline{D}_{n,k}(x)| \geq |D(x)|/2\}} |D(x)| \mu(dx)\right\} &\leq 2\sum_{j=1}^{\infty} p_j |D(j)| e^{-k|D(j)|^2/8} \\
&\leq \frac{2\max_{\{0 \leq z\}} z e^{-z^2/8}}{\sqrt{k}}.
\end{aligned}$$

Concerning the approximation error, the modified Lipschitz condition (10) implies

$$\begin{aligned}
|\overline{D}_{n,k}(j) - D(j)| &\leq \frac{1}{k}\sum_{i=1}^{k} |D(X_{(n,i)}(j)) - D(j)| \\
&\leq C^* \frac{1}{k}\sum_{i=1}^{k} \mu\left([j - |X_{(n,i)}(j) - j| + 1, j + |X_{(n,i)}(j) - j| - 1]\right) \\
&\leq C^* \mu\left([j - |X_{(n,k)}(j) - j| + 1, j + |X_{(n,k)}(j) - j| - 1]\right).
\end{aligned}$$

Without loss of generality assume that $C^* > 1/2$ and $|D(j)| > 0$. Introduce the notation

$$\ell_j^* := \min\left\{\ell \in \mathbb{N}; \mu\left([j-\ell+1, j+\ell-1]\right) \geq \frac{|D(j)|}{2C^*}\right\}.$$

Because of

$$0 < \frac{|D(j)|}{2C^*} < 1,$$

ℓ_j^* is well defined. Put

$$A_j = \left[j-\ell_j^*+1, j+\ell_j^*-1\right].$$

Thus

$$\begin{aligned}
&\mathbb{E}\left\{\int_{\{|\overline{D}_{n,k}(x)-D(x)|\geq|D(x)|/2\}} |D(x)|\mu(dx)\right\} \\
&\leq \sum_{j=1}^{\infty} p_j|D(j)|\mathbb{P}\left\{\mu\left([j-|X_{(n,k)}(j)-j|+1, j+|X_{(n,k)}(j)-j|-1]\right) \geq \frac{|D(j)|}{2C^*}\right\} \\
&= \sum_{j=1}^{\infty} p_j|D(j)|\mathbb{P}\left\{|X_{(n,k)}(j)-j| \geq \ell_j^*\right\}.
\end{aligned}$$

If μ_n denotes the empirical distribution for $X_1, \ldots, X_n$, then

$$\begin{aligned}
\mathbb{E}\left\{\int_{\{|\overline{D}_{n,k}(x)-D(x)|\geq|D(x)|/2\}} |D(x)|\mu(dx)\right\} &\leq \sum_{j=1}^{\infty} p_j|D(j)|\mathbb{P}\left\{\sum_{i=1}^{n}\mathbb{I}_{|X_i-j|<\ell_j^*} \leq k\right\} \\
&= \sum_{j=1}^{\infty} p_j|D(j)|\mathbb{P}\left\{\mu_n(A_j) \leq k/n\right\}.
\end{aligned}$$

For the decomposition

$$\mathbb{P}\left\{\mu_n(A_j) \leq k/n\right\} \leq \mathbb{I}_{\mu(A_j)\leq 2k/n} + \mathbb{I}_{\mu(A_j)>2k/n}\mathbb{P}\left\{\mu_n(A_j) \leq k/n\right\},$$

apply the Bernstein inequality:

$$\begin{aligned}
\mathbb{I}_{\mu(A_j)>2k/n}\mathbb{P}\left\{\mu_n(A_j) \leq k/n\right\} &= \mathbb{I}_{\mu(A_j)>2k/n}\mathbb{P}\left\{\mu_n(A_j)-\mu(A_j) \leq k/n-\mu(A_j)\right\} \\
&\leq \mathbb{I}_{\mu(A_j)>2k/n}\mathbb{P}\left\{\mu_n(A_j)-\mu(A_j) \leq -\mu(A_j)/2\right\} \\
&\leq \mathbb{I}_{\mu(A_j)>2k/n}e^{-\frac{n\mu(A_j)^2/4}{2(\mu(A_j)+\mu(A_j)/6)}} \\
&= \mathbb{I}_{\mu(A_j)>2k/n}e^{-3n\mu(A_j)/28} \\
&\leq e^{-3k/14}.
\end{aligned}$$

The definition of A_j implies

$$\mu(A_j) \geq \frac{|D(j)|}{2C^*}.$$

Therefore,

$$\begin{aligned}\sum_{j=1}^{\infty} p_j |D(j)| \mathbb{I}_{\mu(A_j)\leq 2k/n} &\leq \sum_{j=1}^{\infty} p_j |D(j)| \mathbb{I}_{|D(j)|/(2C^*)\leq 2k/n}\\ &\leq 4C^* k/n.\end{aligned}$$

These bounds imply the bound on the approximation error:

$$\mathbb{E}\left\{\int_{\{|\overline{D}_{n,k}(x)-D(x)|\geq |D(x)|/2\}} |D(x)|\mu(dx)\right\} \leq e^{-3k/14} + 4C^* k/n.$$

■

Proof of Theorem 5. Again we use the decomposition (7). Lemma 2 with unchanged proof yields (8). Under the local Lipschitz condition and the weak density condition, we have to prove (9), i.e., (14). Let $\delta > 0$ be from the definition of weak density assumption. We have that

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

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 (12) and (13). 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

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

■

Proof of Theorem 6. With the notation (7), we have

$$\mathbb{E}\{(D_{n,k}(x) - D(x))^2\} = \mathbb{E}\{(D_{n,k}(x) - \overline{D}_{n,k}(x))^2\} + \mathbb{E}\{(\overline{D}_{n,k}(x) - D(x))^2\}.$$

We show that

$$\int \mathbb{E}\left\{\left(D_{n,k}(x) - \overline{D}_{n,k}(x)\right)^2\right\} \mu(dx) = (\mathbb{E}\{\sigma^2(X)\} + o(1))/k \tag{15}$$

and

$$\mathbb{E}\left\{\left(\overline{D}_{n,k}(x) - D(x)\right)^2\right\} \leq 2C^{*2}(k/n)^{2/d}, \tag{16}$$

which imply the assertion of the theorem.

In the proof of Lemma 2 we mentioned that for given $X_1, \dots, X_n$, the random variable $D_{n,k}(x) - \overline{D}_{n,k}(x)$ is an average of independent random variables with mean zero, therefore

$$\begin{aligned}
&\mathbb{E}\left\{(D_{n,k}(x) - \overline{D}_{n,k}(x))^2 \mid X_1, \dots, X_n\right\} \\
&= \mathbb{E}\left\{\left(\frac{1}{k}\sum_{i=1}^k (Y_{(n,i)}(x) - D(X_{(n,i)}(x)))\right)^2 \mid X_1, \dots, X_n\right\} \\
&= \frac{1}{k^2}\sum_{i=1}^k \mathbb{E}\left\{\left(Y_{(n,i)}(x) - D(X_{(n,i)}(x))\right)^2 \mid X_1, \dots, X_n\right\} \\
&= \frac{1}{k^2}\sum_{i=1}^k \sigma^2(X_{(n,i)}(x)).
\end{aligned}$$

Problems 6.3 and 6.4 in Györfi et al. (2002) together with $k/n \to 0$ imply that

$$\int \mathbb{E}\{(D_{n,k}(x) - \overline{D}_{n,k}(x))^2\}\mu(dx) = \frac{\mathbb{E}\{\sigma^2(X)\} + o(1)}{k},$$

and so (15) is verified. Concerning (16), the modified Lipschitz condition implies that

$$\begin{aligned}\left(\overline{D}_{n,k}(x)-D(x)\right)^2 &= \left(\frac{1}{k}\sum_{i=1}^{k}(D(x)-D(X_{(n,i)}(x)))\right)^2\\ &\le \left(\frac{1}{k}\sum_{i=1}^{k}|D(x)-D(X_{(n,i)}(x))|\right)^2\\ &\le C^{*2}\left(\frac{1}{k}\sum_{i=1}^{k}\mu(S_{x,\|x-X_{(n,i)}(x)\|})^{1/d}\right)^2\\ &\le C^{*2}\mu(S_{x,\|x-X_{(n,k)}(x)\|})^{2/d}.\end{aligned}$$

Thus,

$$\mathbb{E}\left\{\left(\overline{D}_{n,k}(x)-D(x)\right)^2\right\} \le C^{*2}\mathbb{E}\left\{U_{(k,n)}^{2/d}\right\}.$$

If $d \ge 2$, then the Jensen inequality implies

$$\mathbb{E}\left\{U_{(k,n)}^{2/d}\right\} \le (k/n)^{2/d},$$

while for $d = 1$, one has

$$\mathbb{E}\left\{U_{(k,n)}^{2}\right\} = \mathbb{V}ar(U_{(k,n)}) + \mathbb{E}\left\{U_{(k,n)}\right\}^2 \le k/n^2 + (k/n)^2.$$

■

Acknowledgments

The authors would like to thank the Editor and Reviewers for careful reading and comments. These comments and suggestions have been very helpful for revising and improving the manuscript.

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.

References

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. The Annals of Statistics, 35(2):608—633, 2007.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method.* Springer—Verlag, Cham, 2015.

Peter Binev, Albert Cohen, Wofgang Dahmen and Ronald DeVore, Classification algorithms using adaptive partitioning. *Annals of Statistics* 42:2141—2163, 2014.

Iingrid Blaschzyk and Ingo Steinwart, Improved classification rates under refined margin conditions. *Electronic Journal of Statistics*, 12:793—823, 2018.

Timothy I. Cannings, Thomas B. Berrett and Richard J. Samworth. Local nearest neighbor classification with applications to semi-supervised learning. *arXiv: 1704.00642*, 2017.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence and Kilian Q. Weinberger, editors, *Neural Information Processing Systems.* 27:3437—3445,*arXiv: 1407.0067v2*, 2014.

Luc Devroye, László Györfi and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer—Verlag, New York, 1996.

Maik Döring, László Györfi and Harro Walk. Exact rate of convergence of kernel-based classification rule. In Stan Matwin and Jan Mielniczuk, editors, *Challenges in Statistics and Data Mining.* Studies in Computational Intelligence, 605:71—91, Springer, Cham, 2015.

Sébastien Gadat, Thierry Klein and Clément Marteau. Classification with the nearest neighbor rule in general finite dimensional space. *The Annals of Statistics*, 44(3):982—1009, 2016.

László Györfi, Michael Kohler, Adem Krzyżak and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer—Verlag, New York, 2002.

Michael Kohler and Adam Krzyżak. On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE Transactions on Information Theory*, 53(5):1735—1742, 2007.

Michael Kohler, Adam Krzyżak and Harro Walk. Rate of convergence for partitioning and nearest neighbor regression estimates with unbounded data. *Journal of Multivariate Analysis*, 97:311—323, 2006.

Elia Liitiäinen, Francesco Corona and Amaury Lendasse. Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis*, 101:811—823, 2010.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808—1829, 1999.

Richard J. Samworth. Optimal weighted nearest neighbor classifiers. *The Annals of Statistics*, 40(5):2733—2763, 2012a.

Richard J. Samworth. Supplement to Samworth (2012a). arXiv: 1101.5783, 2012b.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135—166. 2004.