On the rate of convergence of local averaging plug-in classification rules under a margin condition *

Michael Kohler

Fachrichtung 6.1-Mathematik, Universität des Saarlandes, Postfach 151150,

D-66041 Saarbrücken, Germany, email: kohler@math.uni-sb.de

and

Adam Krzyżak

Department of Computer Science and Software Engineering, Concordia University, 1455

De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8, email:

krzyzak@cs.concordia.ca

January 22, 2006

Abstract

The rates of convergence of plug-in kernel, partitioning and nearest neighbors classification rules are analyzed. A margin condition, which measures how quickly the a posteriori probabilities cross the decision boundary, smoothness conditions on the a posteriori probabilities and boundedness of the feature vector are imposed. The rates of convergence of the plug-in classifiers shown in this paper are faster than previously known.

Key words and phrases: classification, statistical learning, plug-in classifiers, rate of con-

*Running title: *Rate of convergence of plug-in classification rules*

Please send correspondence and proofs to: Adam Krzyżak, Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec Canada H3G 1M8, email: krzyzak@cs.concordia.ca, phone: +1-514-848-2424, ext. 3007, fax: +1-514-848-2830.

vergence.

1 Introduction

Let $(X,Y),(X_1,Y_1),(X_2,Y_2),\ldots$ be i.i.d. random variables with values in $\mathbb{R}^d \times \{0,1\}$. In classification we want to predict Y given the value of X, i. e., we want to find a classifier $f:\mathbb{R}^d \to \{0,1\}$ such that the misclassification risk

$$\mathbf{P}\{f(X) \neq Y\}$$

is as small as possible. Denote by

$$m(x) = \mathbf{P}\{Y=1|X=x\} = \mathbf{E}\{Y|X=x\}$$

the a posteriori probability of Y given $X=x$. Then the Bayes classifier, i. e., the classification rule with the smallest misclassification risk

$$\mathbf{P}\{f^*(X) \neq Y\} = \min_{f:\mathbb{R}^d \to \{0,1\}} \mathbf{P}\{f(X) \neq Y\}$$

is given by

$$f^*(x) = \begin{cases} 1 & \text{if} \quad m(x) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

(cf. Devroye, Györfi and Lugosi (1996), Theorem 2.1.) In applications, the distribution of (X,Y), and hence also this optimal classifier are unknown. But often it is possible to observe a sample

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

of the underlying distribution, and then the task is to learn a classification rule $f_n(\cdot) = f_n(\cdot,\mathcal{D}_n) : \mathbb{R}^d \to \mathbb{R}$ from this data. For an introduction to pattern recognition and classification we refer the reader to the monographs Devroye, Györfi and Lugosi (1996) and Vapnik (1998).

In this paper we consider so–called plug-in classifiers, which estimate the regression function m by a regression estimate $m_n(\cdot) = m_n(\cdot, \mathcal{D}_n) : \mathbb{R}^d \to \mathbb{R}$ and define the classification rule by

$$f_n(x) = \begin{cases} 1 & \text{if} \quad m_n(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The rate of convergence of the difference between the misclassification risk of the plug-in classifier and the optimal misclassification risk is related to the error of the regression estimates. For a long time the following rather trivial bound was a main tool in this domain:

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \leq 2 \cdot \mathbf{E}\{|m_n(X) - m(X)|\} \leq 2 \cdot \sqrt{\mathbf{E}|m_n(X) - m(X)|^2}$$

(cf. Theorem 2.2 in Devroye, Györfi and Lugosi (1996)). It is well-known, that it is possible to construct for (p, C)-smooth regression functions (i. e., roughly speaking the functions that are p-times continuously differentiable) regression estimates such that the expected L_2 error

$$\mathbf{E}|m_n(X) - m(X)|^2 = \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)$$

converges to zero with the rate

$$n^{-2p/(2p+d)}$$

(cf. Stone (1982), or Györfi et al. (2002)). So for (p, C)-smooth a posteriori probabilities and suitably defined plug-in classifiers we have the bound

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \leq c_1 \cdot n^{-p/(2p+d)}$$

and for large p the rate of convergence of the so-called excess risk achieves rates up to $n^{-1/2}$.

Recently, it was shown that under so-called margin condition

$$\exists \bar{c}\ \exists \alpha > 0 : \mathbf{P}\{0 < |m(X) - \frac{1}{2}| < t\} \leq \bar{c} \cdot t^{\alpha} \tag{2}$$

for all $t > 0$, one can derive much better rates than $n^{-1/2}$. Corresponding results concerning classifiers based on empirical risk minimization can be found, e.g., in Audibert (2004), Mammen and Tsybakov (1999), Massart and Nédélec (2003) and Tsybakov and van de Geer (2005). For plug-in classifiers it was shown in in Audibert and Tsybakov (2005) that assuming margin condition (2) one gets

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \leq \left(\mathbf{E} \int |m_n(X) - m(X)|^2 \mathbf{P}_X(dx) \right)^{\frac{1+\alpha}{2+\alpha}} \tag{3}$$

(cf. Audibert and Tsybakov (2005), Lemma 5.2), which implies that for (p, C)–smooth regression functions and suitably defined plug-in classifiers we have the bound

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \leq c_2 \cdot n^{-\frac{2p\cdot(1+\alpha)}{(2p+d)\cdot(2+\alpha)}}, \tag{4}$$

so for large α we can get rates up to n^{-1}. Furthermore, it was shown that under the margin condition and by imposing restrictions on the distribution of X such as existence of a bounded density with respect to the Lebesgue-Borel measure one gets for estimates defined by minimizing the empirical risk on a special covering (these estimates are hard to compute in practice) for (p, C)-smooth regression function under the margin condition

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \leq c_3 \cdot n^{-\frac{p\cdot(1+\alpha)}{(2+\alpha)p+d}}. \tag{5}$$

If, in addition, X has a density with respect to the Lebesgue-Borel measure bounded away from zero and infinity, then it was shown that suitably defined local polynomial kernel plug-in classifiers (these estimates are easy to compute in practice) satisfy

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \leq c_4 \cdot n^{-\frac{p(1+\alpha)}{2p+d}} \tag{6}$$

(cf. Theorems 4.3 and 3.3 in Audibert and Tsybakov (2005)). In (6) one can get for large α rates better than n^{-1}. However, the assumption on the distribution of X somewhat limits the value of this result. The margin condition (2) measures how quickly the a posteriori probability crosses the decision boundary $\{x : m(x) = 1/2\}$. It depends on the distribution of X and on the steepness of the regression function near the decision boundary. But if we require as in (6) that X have a density with respect to the Lebesgue-Borel measure bounded away from zero, then for $p \geq 1$ the class of distributions of (X, Y) which yield m (p, C)-smooth and which fulfill the margin condition for $\alpha > 1$ is very narrow.

In this paper we improve the bound (4) without assuming existence of a density of X. As main result we prove bounds which are for $d \cdot \alpha > 2$ better than (4) but not as good as the one in (5) for kernel, partitioning and nearest neighbor plug-in classification rules and $p \leq 1$. In case a density of X exists which is bounded away from zero we provide for kernel and partitioning plug-in classifiers simple proofs of (6) for $p \leq 1$. In contrast to Audibert and Tsybakov (2005) this result does not require that the density of X is bounded away from infinity. The main results are formulated in Section 2 and proven in Section 3.

2 Main results

In the sequel we make the following three assumptions on the distribution of (X, Y):

(A1) There exists $\bar{c} > 0$ and $\alpha > 0$ such that for all $\delta > 0$ we have

$$\mathbf{E}\left\{|m(X) - \frac{1}{2}| \cdot 1_{\{|m(X)-\frac{1}{2}| \leq \delta\}}\right\} \leq \bar{c} \cdot \delta^{1+\alpha}.$$

(A2) $X \in [0,1]^d$ *a.s.*

(A3) There exists $0 < p \leq 1$ and $C > 0$ such that $m(x) = \mathbf{E}\{Y|X = x\}$ is (p, C)–smooth, i.e.,

$$|m(x) - m(z)| \leq C \cdot \|x - z\|^p \quad \text{for all } x, z \in [0,1]^d,$$

where $\|x - z\|$ is the Euclidean norm of $x - z$.

Note that the margin condition (A1) is slightly weaker than the margin condition (2).

First we consider the Nadaraya-Watson kernel estimate (Nadaraya (1964) and Watson (1964)) defined by

$$m_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) \cdot Y_i}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} \quad (\text{with } \frac{0}{0} = 0)$$

with naive kernel $K : \mathbb{R}^d \to \mathbb{R}_+$ given by $K(u) = 1_{\{\|u\| \le 1\}}$ and bandwidth $h_n > 0$. The plug-in classifier is then defined by (1).

Theorem 1 *Assume that the distribution of* (X, Y) *satisfies (A1), (A2) and (A3) and that the plug-in kernel classification rule is defined as above with bandwidth*

$$h_n = n^{-\frac{1}{d+p\cdot(3+\alpha)}}.$$

Then

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \le c_5 \cdot n^{-\frac{p\cdot(1+\alpha)}{p\cdot(3+\alpha)+d}}.$$

If we add restrictions on the distribution of X, we can improve the rate of convergence above:

Theorem 2 *Assume that the distribution of* (X, Y) *satisfies (A1), (A2), (A3) and, in addition, assume that* X *has a density with respect to the Lebesgue-Borel measure which is bounded away from zero. Let the plug-in kernel classification rule be defined as above with bandwidth*

$$h_n = n^{-\frac{1}{2p+d}}.$$

Then

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \le c_6 \cdot n^{-\frac{p\cdot(1+\alpha)}{2p+d}}.$$

Next we consider plug-in classifiers based on partitioning estimates. Let $\mathcal{P}_n$ be a cubic partition of $\mathbb{R}^d$ into cubes with side-length $h_n > 0$. For $x \in \mathbb{R}^d$ let $A_n(x)$ be that cube $A_{j,n} \in \mathcal{P}_n$ with $x \in A_{j,n}$. Then the partitioning estimate with partition $\mathcal{P}_n$ is defined by

$$m_n(x) = \frac{\sum_{i=1}^n 1_{A_n(x)}(X_i) \cdot Y_i}{\sum_{i=1}^n 1_{A_n(x)}(X_i)} \quad (\text{with } \frac{0}{0} = 0).$$

Let $f_n(x)$ be the corresponding plug-in classification rule defined by (1). In the same way as Theorem 1 we will show

Theorem 3 *Assume that the distribution of* (X, Y) *satisfies (A1), (A2) and (A3) and that the plug-in partitioning classification rule is defined as above with cubes of side-length*

$$h_n = n^{-\frac{1}{d+p\cdot(3+\alpha)}}.$$

Then

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \leq c_7 \cdot n^{-\frac{p\cdot(1+\alpha)}{p\cdot(3+\alpha)+d}}.$$

Under restrictions on the distribution of X, we get again a better rate of convergence than above:

Theorem 4 *Assume that the distribution of* (X, Y) *satisfies (A1), (A2), (A3) and, in addition, assume that* X *has a density with respect to the Lebesgue-Borel measure which is bounded away from zero. Let the plug-in partitioning classification rule be defined as above with cubes of side-length*

$$h_n = n^{-\frac{1}{2p+d}}.$$

Then

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \leq c_8 \cdot n^{-\frac{p\cdot(1+\alpha)}{2p+d}}.$$

Next we consider nearest neighbor regression estimates. For $x \in \mathbb{R}^d$ let

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

be a permutation of $\mathcal{D}_n$ such that

$$\|x - X_{(1)}(x)\| \leq \dots \leq \|x - X_{(n)}(x)\|.$$

In case of ties, i.e., in case $\|x - X_i\| = \|x - X_j\|$, we assume that the data point with the smaller index comes before the other data point. For $k_n \in \{1, \dots, n\}$ the k_n–nearest neighbor estimate is defined by

$$m_n(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i)}(x).$$

Let $f_n(x)$ be the corresponding plug-in classifier.

Theorem 5 *Assume that the distribution of (X,Y) satisfies (A1), (A2) and (A3) and that the plug-in nearest neighbor classification rule is defined as above with*

$$k_n = \left\lceil \log(n) \cdot n^{\frac{2p}{(3+\alpha)\cdot p+d}} \right\rceil.$$

Then

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \leq c_9 \cdot \log(n)^r \cdot n^{-\frac{p\cdot(1+\alpha)}{p\cdot(3+\alpha)+d}}$$

for some $r > 0$.

Under restrictions on the distribution of X, we will show again a better rate of convergence:

Theorem 6 *Assume that the distribution of (X,Y) satisfies (A1), (A2), (A3) and, in addition, assume that X has a density with respect to the Lebesgue-Borel measure which is bounded away from zero. Let the plug-in nearest neighbor classification rule be defined as above with*

$$k_n = \log^2(n) \cdot n^{\frac{2p}{2p+d}}.$$

Then

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \leq c_{10} \cdot \log(n)^{\frac{2p}{d}\cdot(1+\alpha)} \cdot n^{-\frac{p\cdot(1+\alpha)}{2p+d}}.$$

Remark 1. For $p \leq 1$ and $d \cdot \alpha > 2$ the rate of convergence in Theorems 1 and 3 is better than in (4), but worse than in (6). However, as already mentioned in the introduction, these theorems do not require the existence of a density of X with respect to the Lebesgue-Borel measure, and therefore in Theorems 1 and 3 there is no contradiction between the margin condition and smoothness of the regression function for large α and large p.

Remark 2. For large α the rates of convergence in Theorems 1, 3 and 5 approach the parametric rate n^{-1}. In Theorems 2, 4 and 6 the rate of convergence is even better than n^{-1} for large α, but because of the restrictions on the distribution of X the class of distributions of (X, Y) satisfying the assumptions of Theorems 2, 4 and 6 is rather narrow for α large.

Remark 3. In the proofs of the above theorems we analyze the rate of convergence of the pointwise error of local averaging estimates. This pointwise error was also analyzed in Devroye (1981, 1982), Greblicki, Krzyżak and Pawlak (1984), Györfi (1981), Mack (1981) and Walk (2001).

3 Proofs

3.1 Proof of Theorem 1

By Theorem 2.2 in Devroye, Györfi and Lugosi (1996) we have for any $\delta_n > 0$

$$\begin{aligned}
&\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \\
&= \mathbf{E}\left\{|2m(X) - 1| \cdot 1_{\{f_n(X) \neq f^*(X)\}}\right\} \\
&= 2 \cdot \mathbf{E}\left\{|m(X) - 1/2| \cdot 1_{\{|m(X)-1/2| \leq \delta_n\}} \cdot 1_{\{f_n(X) \neq f^*(X)\}}\right\} \\
&\quad +2 \cdot \mathbf{E}\left\{|m(X) - 1/2| \cdot 1_{\{|m(X)-1/2| > \delta_n\}} \cdot 1_{\{f_n(X) \neq f^*(X)\}}\right\}.
\end{aligned}$$

If $f_n(X) \neq f^*(X)$ then $|m(X) - 1/2| \leq |m_n(X) - m(X)|$. Using the margin condition $(A1)$ we conclude

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \leq 2\bar{c} \cdot \delta_n^{1+\alpha} + \mathbf{P}\{|m_n(X) - m(X)| > \delta_n\}.$$

Put

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) \cdot m(X_i)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}.$$

Then

$$\mathbf{P}\{|m_n(X) - m(X)| > \delta_n\}$$

$$\leq \mathbf{P}\{|m_n(X) - \hat{m}_n(X)| > \delta_n/2\} + \mathbf{P}\{|\hat{m}_n(X) - m(X)| > \delta_n/2\}.$$

First we bound

$$\begin{aligned}
&\mathbf{P}\{|m_n(X) - \hat{m}_n(X)| > \delta_n/2\} \\
&= \int \mathbf{P}\{|m_n(x) - \hat{m}_n(x)| > \delta_n/2\} \mathbf{P}_X(dx) \\
&= \int \mathbf{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^n \frac{K(\frac{x-X_i}{h_n})}{\sum_{j=1}^n K(\frac{x-X_j}{h_n})} \cdot (Y_i - m(X_i)) \right| > \frac{\delta_n}{2n} \right\} \mathbf{P}_X(dx).
\end{aligned}$$

Using $K^2(u) = K(u)$ $(u \in \mathbb{R}^d)$ we get by Hoeffding inequality (cf., e.g., Lemma A.3 in Györfi et al. (2002))

$$\begin{aligned}
&\mathbf{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^n \frac{K(\frac{x-X_i}{h_n})}{\sum_{j=1}^n K(\frac{x-X_j}{h_n})} \cdot (Y_i - m(X_i)) \right| > \frac{\delta_n}{2n} \middle| X_1, \ldots, X_n \right\} \\
&\leq 2 \cdot \exp\left(- \frac{2 \cdot n \cdot (\delta_n/(2n))^2}{\frac{1}{n} \sum_{i=1}^n \frac{K^2(\frac{x-X_i}{h_n})}{\left(\sum_{j=1}^n K(\frac{x-X_j}{h_n})\right)^2}} \right) \\
&= 2 \cdot \exp\left(-\frac{1}{2} \sum_{j=1}^n K(\frac{x-X_j}{h_n}) \cdot \delta_n^2 \right) \\
&\leq 2 \cdot I_{\{\sum_{j=1}^n K(\frac{x-X_j}{h_n}) < \frac{1}{2} \cdot n \cdot \mathbf{P}_X(S_{x,h_n}) - \log^2(n)\}} \\
&\quad +2 \cdot \exp\left(-\frac{1}{4} \cdot n \cdot \mathbf{P}_X(S_{x,h_n}) \cdot \delta_n^2 \right) \cdot \exp\left(\frac{\log^2(n) \cdot \delta_n^2}{2} \right).
\end{aligned}$$

If we choose δ_n such that $\delta_n \leq 1/\log(n)$ we get

$$\begin{aligned}
&\mathbf{P}\{|m_n(X) - \hat{m}_n(X)| > \delta_n/2\} \\
&\leq \int \mathbf{P}\left\{ \frac{1}{n} \sum_{j=1}^n K(\frac{x-X_j}{h_n}) - \mathbf{P}_X(S_{x,h_n}) < -\frac{1}{2} \cdot \mathbf{P}_X(S_{x,h_n}) - \frac{\log^2(n)}{n} \right\} \mathbf{P}_X(dx) \\
&\quad +4 \cdot \int \exp\left(-\frac{1}{4} \cdot n \cdot \mathbf{P}_X(S_{x,h_n}) \cdot \delta_n^2 \right) \mathbf{P}_X(dx).
\end{aligned}$$

Using

$$K(\frac{x-X_j}{h_n}) = I_{\{X_j \in S_{x,h_n}\}} \quad \text{and} \quad \mathbf{V}(K(\frac{x-X_j}{h_n})) \leq \mathbf{P}_X(S_{x,h_n})$$

(where S_{x,h_n} denotes the (closed) ball of radius h_n around x) we can bound the probability in the first integral by Bernstein inequality (cf., e.g., Lemma A.2 in Györfi et al. (2002)). This yields

$$
\begin{aligned}
&\mathbf{P}\left\{\frac{1}{n}\sum_{j=1}^{n} K(\frac{x-X_j}{h_n}) - \mathbf{P}_X(S_{x,h_n}) < -\frac{1}{2}\cdot \mathbf{P}_X(S_{x,h_n}) - \frac{\log^2(n)}{n}\right\} \\
&\leq 2\cdot \exp\left(-\frac{n\cdot(\frac{1}{2}\mathbf{P}_X(S_{x,h_n})+\frac{\log^2(n)}{n})^2}{2\mathbf{P}_X(S_{x,h_n})+2\cdot(\frac{1}{2}\mathbf{P}_X(S_{x,h_n})+\frac{\log^2(n)}{n})\cdot(1-0)/3}\right) \\
&\leq 2\cdot \exp\left(-\frac{n\cdot(\frac{1}{2}\mathbf{P}_X(S_{x,h_n})+\frac{\log^2(n)}{n})}{4+2/3}\right) \\
&\leq 2\cdot \exp\left(-\frac{14}{3}\cdot \log^2(n)\right).
\end{aligned}
$$

To bound the second integral we use inequality (5.1) in Györfi et al. (2002) and get

$$
\begin{aligned}
&4\cdot \int \exp\left(-\frac{1}{4}\cdot n\cdot \mathbf{P}_X(S_{x,h_n})\cdot \delta_n^2\right)\mathbf{P}_X(dx) \\
&= \frac{16}{n\cdot\delta_n^2}\cdot \int \frac{1}{4}\cdot n\cdot \mathbf{P}_X(S_{x,h_n})\cdot\delta_n^2\cdot \exp\left(-\frac{1}{4}\cdot n\cdot \mathbf{P}_X(S_{x,h_n})\cdot \delta_n^2\right)\cdot \frac{1}{\mathbf{P}_X(S_{x,h_n})}\mathbf{P}_X(dx) \\
&\leq \frac{16\cdot \max_{u\in\mathbb{R}_+} u\cdot e^{-u}}{n\cdot\delta_n^2}\int \frac{1}{\mathbf{P}_X(S_{x,h_n})}\mathbf{P}_X(dx) \qquad (7) \\
&\leq \frac{16}{e}\cdot \frac{c_{11}}{n\cdot \delta_n^2\cdot h_n^d}. \qquad (8)
\end{aligned}
$$

Putting together the above results we get

$$
\mathbf{P}\{|m_n(X)-\hat{m}_n(X)|>\delta_n/2\} \leq 2\cdot\exp\left(-\frac{14}{3}\cdot\log^2(n)\right)+\frac{16}{e}\cdot\frac{c_{11}}{n\cdot\delta_n^2\cdot h_n^d}
$$

provided we choose δ_n such that $\delta_n \leq 1/\log(n)$.

So it remains to bound

$$
\mathbf{P}\{|\hat{m}_n(X)-m(X)|>\delta_n/2\} = \int \mathbf{P}\{|\hat{m}_n(x)-m(x)|>\delta_n/2\}\mathbf{P}_X(dx).
$$

Fix $x\in[0,1]^d$ and define the event $B_n(x)$ by

$$
B_n(x) = \left\{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) > 0\right\}.
$$

By triangle inequality and (p, C)–smoothness of m we have

$$
\begin{aligned}
& |\hat{m}_n(x) - m(x)| \\
&= \left| \frac{\sum_{i=1}^n (m(X_i) - m(x)) \cdot K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} \right| \cdot 1_{B_n(x)} + m(x) 1_{B_n(x)^c} \\
&\leq \frac{\sum_{i=1}^n |m(X_i) - m(x)| \cdot K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} \cdot 1_{B_n(x)} + m(x) 1_{B_n(x)^c} \\
&\leq C h_n^p + m(x) 1_{B_n(x)^c},
\end{aligned}
$$

where the last inequality follows from the fact that

$$
K\left(\frac{x - X_i}{h_n}\right) \neq 0 \quad \text{implies} \quad \|x - X_i\| \leq h_n. \tag{9}
$$

Hence

$$
\begin{aligned}
\mathbf{P}\{|\hat{m}_n(x) - m(x)| > \delta_n/2\} &\leq \mathbf{P}\left\{m(x) 1_{B_n(x)^c} > \delta_n/2 - C \cdot h_n^p\right\} \\
&\leq \mathbf{P}\left\{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) = 0\right\}
\end{aligned}
$$

provided we choose δ_n such that

$$
\delta_n > 2 \cdot C \cdot h_n^p.
$$

From this we conclude

$$
\begin{aligned}
\mathbf{P}\{|\hat{m}_n(X) - m(X)| > \delta_n/2\} &\leq \int \mathbf{P}\left\{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) = 0\right\} \mathbf{P}_X(dx) \\
&= \int (1 - \mathbf{P}_X(S_{x,h_n}))^n \, \mathbf{P}_X(dx) \\
&\leq \int e^{-n \cdot \mathbf{P}_X(S_{x,h_n})} \mathbf{P}_X(dx) \\
&\leq \max_{u \in \mathbb{R}_+} u \cdot e^{-u} \cdot \int \frac{1}{n \cdot \mathbf{P}_X(S_{x,h_n})} \mathbf{P}_X(dx) \\
&\leq \frac{c_{11}}{n \cdot h_n^d}
\end{aligned}
$$

where the last inequality follows from inequality (5.1) in Györfi et al. (2002).

Thus we get for any δ_n satisfying $2 \cdot C \cdot h_n^p < \delta_n \leq 1/\log(n)$

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\}$$
$$\leq 2\bar{c} \cdot \delta_n^{1+\alpha} + \frac{16}{e} \cdot \frac{c_{11}}{n \cdot h_n^d \cdot \delta_n^2} + 2 \cdot \exp(-\frac{14}{3} \cdot \log^2(n)) + \frac{c_{11}}{n \cdot h_n^d}.$$

With $\delta_n = 4 \cdot C \cdot h_n^p$ and $h_n = n^{-1/(d+p\cdot(3+\alpha))}$ we get the desired result. □

3.2 Proof of Theorem 2

Set

$$\delta_n = 2C \cdot h_n^p = 2C \cdot n^{-\frac{p}{2p+d}}$$

and for $x \in \mathbb{R}^d$ let $B_n(x)$ be the event that

$$\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) > \frac{1}{2} \cdot n \cdot \mathbf{P}_X(S_{x,h_n})$$

where S_{x,h_n} denotes the (closed) ball of radius h_n around x. Because of

$$\mathbf{E}K\left(\frac{x - X_1}{h_n}\right) = \mathbf{P}_X(S_{x,h_n}) \quad \text{and} \quad \mathbf{V}\left(K\left(\frac{x - X_1}{h_n}\right)\right) \leq \mathbf{P}_X(S_{x,h_n})$$

we get by Bernstein inequality (cf., e.g., Lemma A.2 in Györfi et al. (2002))

$$\begin{aligned}
\mathbf{P}(B_n(x)^c) &= \mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) - \mathbf{P}_X(S_{x,h_n}) \leq -\frac{1}{2} \cdot \mathbf{P}_X(S_{x,h_n})\right\} \\
&\leq \exp\left(-\frac{n \cdot \left(\frac{1}{2} \cdot \mathbf{P}_X(S_{x,h_n})\right)^2}{2\mathbf{P}_X(S_{x,h_n}) + 2 \cdot \frac{1}{2} \cdot \mathbf{P}_X(S_{x,h_n}) \cdot (1-0)/3}\right) \\
&= \exp\left(-\frac{n \cdot \mathbf{P}_X(S_{x,h_n})}{8 + 4/3}\right) \\
&\leq \exp(-c_{12} \cdot n \cdot h_n^d).
\end{aligned}$$

Using this together with Theorem 2.2 in Devroye, Györfi and Lugosi (1996) we get

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\}$$

$$= \mathbf{E}\left\{|2m(X) - 1| \cdot 1_{\{f_n(X) \neq f^*(X)\}}\right\}$$

$$\leq 2 \cdot \mathbf{E}\left\{|m(X) - 1/2| \cdot 1_{\{|m(X)-1/2| \leq \delta_n\}} \cdot 1_{\{f_n(X) \neq f^*(X)\}}\right\}$$

$$+\sum_{j=1}^{\infty} 2\cdot \mathbf{E}\left\{|m(X)-1/2|\cdot 1_{\{2^{j-1}\delta_n<|m(X)-1/2|\leq 2^j\delta_n\}}\cdot 1_{\{f_n(X)\neq f^*(X)\}}\cdot 1_{B_n(x)}\right\}$$
$$+\exp(-c_{12}\cdot n\cdot h_n^d).$$

If $f_n(X)\neq f^*(X)$ then $|m(X)-\frac{1}{2}|\leq |m_n(X)-m(X)|$, from which we conclude

$$\begin{aligned}
&\mathbf{P}\{f_n(X)\neq Y\}-\mathbf{P}\{f^*(X)\neq Y\}\\
&\leq 2\cdot \mathbf{E}\left\{|m(X)-1/2|\cdot 1_{\{|m(X)-1/2|\leq\delta_n\}}\right\}\\
&\quad+\sum_{j=1}^{\infty} 2\cdot \mathbf{E}\Bigg\{|m(X)-1/2|\cdot 1_{\{|m(X)-1/2|\leq 2^j\delta_n\}}\\
&\qquad\qquad\cdot\mathbf{P}\left\{|m_n(X)-m(X)|>2^{j-1}\cdot\delta_n|X,X_1,\dots,X_n\right\}\cdot 1_{B_n(x)}\Bigg\}\\
&\quad+\exp(-c_{12}\cdot n\cdot h_n^d).
\end{aligned}$$

Using the margin condition (A1) we get

$$2\cdot \mathbf{E}\left\{|m(X)-1/2|\cdot 1_{\{|m(X)-1/2|\leq\delta_n\}}\right\}\leq 2\bar{c}\cdot\delta_n^{1+\alpha}.$$

Next we fix $j\in\mathbb{N}$, assume that $B_n(X)$ holds and bound

$$\mathbf{P}\{|m_n(X)-m(X)|>2^{j-1}\cdot\delta_n|X,X_1,\dots,X_n\}.$$

Set

$$\hat{m}_n(x)=\frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\cdot m(X_i)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}.$$

On $B_n(X)$ we have

$$\sum_{i=1}^n K\left(\frac{X-X_i}{h_n}\right)>0,$$

and together with the (p,C)–smoothness of m this implies

$$\begin{aligned}
|\hat{m}_n(X)-m(X)| &= \left|\frac{\sum_{i=1}^n (m(X_i)-m(X))\cdot K\left(\frac{X-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{X-X_i}{h_n}\right)}\right|\\
&\leq \frac{\sum_{i=1}^n |m(X_i)-m(X)|\cdot K\left(\frac{X-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{X-X_i}{h_n}\right)}\\
&\leq \sup_{u,v\in[0,1]^d,\|u-v\|\leq h_n}|m(u)-m(v)|\\
&\leq Ch_n^p=\frac{\delta_n}{2},
\end{aligned}$$

from which we conclude that on $B_n(X)$ we have

$$
\begin{aligned}
&\mathbf{P}\{|m_n(X) - m(X)| > 2^{j-1} \cdot \delta_n | X, X_1, \dots, X_n\} \\
&\le \mathbf{P}\{|m_n(X) - \hat{m}_n(X)| + \frac{\delta_n}{2} > 2^{j-1} \cdot \delta_n | X, X_1, \dots, X_n\} \\
&\le \mathbf{P}\{|m_n(X) - \hat{m}_n(X)| > 2^{j-2} \cdot \delta_n | X, X_1, \dots, X_n\} \\
&= \mathbf{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \frac{K\left(\frac{X-X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{X-X_j}{h_n}\right)} (Y_i - m(X_i)) \right| > 2^{j-2} \frac{\delta_n}{n} | X, X_1, \dots, X_n \right\}.
\end{aligned}
$$

By Hoeffding inequality (cf., e.g., Lemma A.3 in Györfi et al. (2002)) we can bound the last term by

$$
2\exp\left(- \frac{2n \cdot (2^{j-2}\delta_n/n)^2}{\frac{1}{n}\sum_{i=1}^{n} \frac{K^2\left(\frac{X-X_i}{h_n}\right)}{(\sum_{j=1}^{n} K\left(\frac{X-X_j}{h_n}\right))^2}} \right) \le 2\exp\left(-2^{2j-3}\delta_n^2 \sum_{j=1}^{n} K\left(\frac{X-X_j}{h_n}\right) \right)
$$

(where the last inequality follows from $K^2(u) = K(u)$ $(u \in \mathbb{R}^d)$), which, in turn, is bounded on $B_n(X)$ by

$$
\begin{aligned}
2\exp\left(-2^{2j-4}\delta_n^2 n \cdot c_{13} \cdot h_n^d\right) &= 2\exp\left(-2^{2j-4} 4C^2 n^{-2p/(2p+d)} n \cdot c_{13} \cdot n^{-d/(2p+d)}\right) \\
&= 2\exp(-c_{14} \cdot 2^{2j}).
\end{aligned}
$$

Putting together the above results and applying the margin condition (A2) again yields

$$
\begin{aligned}
&\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \\
&\le 2\bar{c} \cdot \delta_n^{1+\alpha} + 4 \sum_{j=1}^{\infty} \mathbf{E}\left\{ |m(X) - 1/2| \cdot 1_{\{|m(X)-1/2| \le 2^j \delta_n\}} \cdot \exp(-c_{14} \cdot 2^{2j}) \right\} \\
&\quad + \exp(-c_{12} \cdot n \cdot h_n^d) \\
&\le 2\bar{c} \cdot \delta_n^{1+\alpha} + 4\bar{c} \cdot \delta_n^{1+\alpha} \sum_{j=1}^{\infty} 2^{j(1+\alpha)} \exp(-c_{14} \cdot 2^{2j}) + \exp(-c_{12} \cdot n \cdot h_n^d) \\
&\le c_{15} \cdot n^{-\frac{p(1+\alpha)}{2p+d}}.
\end{aligned}
$$

□

3.3 Proofs of Theorems 3 and 4

The proofs are similar to the proofs of Theorems 1 and 2, therefore we give only the outline of the proofs.

Let $K_n(x,z) = 1_{A_n(x)}(z)$. Then the partitioning estimate is given by

$$m_n(x) = \frac{\sum_{i=1}^n K_n(x,X_i) \cdot Y_i}{\sum_{i=1}^n K_n(x,X_i)}$$

and with

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n K_n(x,X_i) \cdot m(X_i)}{\sum_{i=1}^n K_n(x,X_i)}$$

we get for $\delta_n > 2C \cdot h_n^p$ as in the proof of Theorem 1:

$$\begin{aligned}
&\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \\
&\leq 2\bar{c} \cdot \delta_n^{1+\alpha} + \mathbf{P}\left\{|m_n(X) - \hat{m}_n(X)| > \delta_n/2\right\} + \mathbf{P}\left\{|\hat{m}_n(X) - m(X)| > \delta_n/2\right\} \\
&\leq 2\bar{c} \cdot \delta_n^{1+\alpha} + 2 \cdot \exp\left(-\frac{14}{3} \cdot \log^2(n)\right) + 4 \cdot \int \exp\left(-\frac{1}{4} \cdot n \cdot \mathbf{P}_X(A_n(x)) \cdot \delta_n^2\right) \mathbf{P}_X(dx) \\
&\quad + \int \mathbf{P}\left\{\sum_{i=1}^n K_n(x,X_i) = 0\right\} \mathbf{P}_X(dx).
\end{aligned}$$

Now using the similar argument as in (7) we get

$$\begin{aligned}
\int \mathbf{P}\left\{\sum_{i=1}^n K_n(x,X_i) = 0\right\} \mathbf{P}_X(dx) &= \int (1 - \mathbf{P}_X(A_n(x)))^n \mathbf{P}_X(dx) \\
&\leq \int \exp\left(-n \cdot \mathbf{P}_X(A_n(x))\right) \mathbf{P}_X(dx) \\
&\leq \frac{1}{e \cdot n} \cdot \int \frac{1}{\mathbf{P}_X(A_n(x))} \mathbf{P}_X(dx)
\end{aligned}$$

and

$$4 \cdot \int \exp\left(-\frac{1}{4} \cdot n \cdot \mathbf{P}_X(A_n(x)) \cdot \delta_n^2\right) \mathbf{P}_X(dx) \leq \frac{16}{e \cdot n \cdot \delta_n^2} \cdot \int \frac{1}{\mathbf{P}_X(A_n(x))} \mathbf{P}_X(dx).$$

Let $A_{1,n}, \ldots, A_{N_n,n}$ be those sets of the partition that overlap with $[0,1]^d$. Then $N_n \leq c_{16}/h_n^d$ and we have

$$\int \frac{1}{\mathbf{P}_X(A_n(x))} \mathbf{P}_X(dx) = \sum_{j=1}^{N_n} \int_{A_{j,n}} \frac{1}{\mathbf{P}_X(A_{j,n})} \mathbf{P}_X(dx) \leq N_n \leq c_{16}/h_n^d.$$

From this Theorem 3 follows as in the proof of Theorem 1.

Similarly Theorem 4 follows from the proof of Theorem 2 if we replace $K((x - X_i)/h_n)$ by $K_n(x, X_i)$ and S_{x,h_n} by $A_n(x)$. $\square$

3.4 Proof of Theorem 5

As in the proof of Theorem 1 we have

$$\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \leq 2\bar{c} \cdot \delta_n^{1+\alpha} + \mathbf{P}\left\{|m_n(X) - m(X)| > \delta_n\right\}.$$

Put

$$\hat{m}_n(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} m(X_{(i)}(x)).$$

Application of Hoeffding inequality (cf., e.g., Lemma A.3 in Györfi et al. (2002)) conditioned on X, X_1, …, X_n yields

$$\begin{aligned}
&\mathbf{P}\{|m_n(X) - m(X)| > \delta_n\} \\
&\leq \mathbf{P}\{|m_n(X) - \hat{m}_n(X)| > \delta_n/2\} + \mathbf{P}\{|\hat{m}_n(X) - m(X)| > \delta_n/2\} \\
&\leq 2 \cdot \exp\left(-2 \cdot k_n \cdot (\delta_n/2)^2\right) + \mathbf{P}\{|\hat{m}_n(X) - m(X)| > \delta_n/2\}.
\end{aligned}$$

Fix $x \in \mathbb{R}^d$. By (p, C)–smoothness of m we have

$$\begin{aligned}
|\hat{m}_n(x) - m(x)| &\leq \frac{1}{k_n} \sum_{i=1}^{k_n} |m(X_{(i)}(x)) - m(x)| \\
&\leq \frac{1}{k_n} \sum_{i=1}^{k_n} C \cdot \|X_{(i)} - x\|^p \\
&\leq C \cdot \|X_{(k_n)} - x\|^p.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbf{P}\{|\hat{m}_n(X) - m(X)| > \delta_n/2\} &= \int \mathbf{P}\{|\hat{m}_n(x) - m(x)| > \delta_n/2\} \mathbf{P}_X(dx) \\
&\leq \int \mathbf{P}\left\{\|X_{(k_n)} - x\| > \left(\frac{\delta_n}{2C}\right)^{1/p}\right\} \mathbf{P}_X(dx).
\end{aligned}$$

Put $\epsilon_n = (\delta_n/(2C))^{1/p}$ and set $p_{x,n} = \mathbf{P}_X(S_{x,\epsilon_n})$. Then

$$\begin{aligned}
\mathbf{P}\left\{\|X_{(k_n)} - x\| > \epsilon_n\right\} &= \mathbf{P}\left\{\sum_{i=1}^n 1_{S_{x,\epsilon_n}}(X_i) < k_n\right\} \\
&= \mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^n 1_{S_{x,\epsilon_n}}(X_i) - p_{x,n} < \frac{k_n}{n} - p_{x,n}\right\} \\
&\leq 1_{\{p_{x,n} < \frac{4k_n}{n}\}} + \mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^n 1_{S_{x,\epsilon_n}}(X_i) - p_{x,n} < -\left(\frac{k_n}{n} + \frac{p_{x,n}}{2}\right)\right\}.
\end{aligned}$$

By Bernstein inequality (cf., e.g., Lemma A.2 in Györfi et al. (2002)) we can bound the probability on the right-hand side above by

$$\begin{aligned}
2 \cdot \exp\left(-\frac{n \cdot (k_n/n + p_{x,n}/2)^2}{2 \cdot p_{x,n}(1 - p_{x,n}) + 2 \cdot (k_n/n + p_{x,n}/2)/3}\right) &\leq 2 \cdot \exp\left(-\frac{n \cdot (k_n/n + p_{x,n}/2)}{4 + 2/3}\right) \\
&\leq 2 \cdot \exp\left(-\frac{3}{14} \cdot k_n\right).
\end{aligned}$$

We conclude

$$\begin{aligned}
&\mathbf{P}\{|\hat{m}_n(X) - m(X)| > \delta_n/2\} \\
&\leq 2 \cdot \exp\left(-\frac{3}{14} \cdot k_n\right) + \mathbf{P}_X\left(\left\{x \,:\, \mathbf{P}_X(S_{x,\epsilon_n}) < 4 \cdot \frac{k_n}{n}\right\}\right).
\end{aligned}$$

Choose a covering of $[0,1]^d$ by balls $S_{x_1,\epsilon_n/2}, \dots, S_{x_{N_n},\epsilon_n/2}$ with radius $\epsilon_n/2$ and such that the number N_n of balls is as small as possible. Then

$$N_n \leq c_{17} \cdot \epsilon_n^{-d}$$

and we get

$$\begin{aligned}
&\mathbf{P}_X\left(\left\{x \,:\, \mathbf{P}_X(S_{x,\epsilon_n}) < 4 \cdot \frac{k_n}{n}\right\}\right) \\
&\leq \sum_{k=1}^{N_n} \mathbf{P}_X\left(\left\{x \in S_{x_k,\epsilon_n/2} \,:\, \mathbf{P}_X(S_{x,\epsilon_n}) < 4 \cdot \frac{k_n}{n}\right\}\right) \\
&\leq \sum_{k=1}^{N_n} \mathbf{P}_X\left(\left\{x \in S_{x_k,\epsilon_n/2} \,:\, \mathbf{P}_X(S_{x_k,\epsilon_n/2}) < 4 \cdot \frac{k_n}{n}\right\}\right) \\
&= \sum_{k=1}^{N_n} \mathbf{P}_X\left(\left\{x \in S_{x_k,\epsilon_n/2}\right\}\right) 1_{\left\{\mathbf{P}_X(S_{x_k,\epsilon_n/2}) < 4 \cdot \frac{k_n}{n}\right\}}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{N_n} \mathbf{P}_X \left(S_{x_k,\epsilon_n/2}\right) 1_{\left\{\mathbf{P}_X(S_{x_k,\epsilon_n/2})<4\cdot \frac{k_n}{n}\right\}} \\
&\leq N_n \cdot \frac{4k_n}{n} = c_{18} \cdot \frac{k_n}{n} \cdot \delta_n^{-d/p},
\end{aligned}
$$

where we have used the fact that $x \in S_{x_k,\epsilon_n/2}$ implies $S_{x_k,\epsilon_n/2} \subseteq S_{x,\epsilon_n}$.

Gathering the above results we get

$$
\begin{aligned}
&\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \\
&\leq 2\bar{c} \cdot \delta_n^{1+\alpha} + 2 \cdot \exp\left(-2 \cdot k_n \cdot (\delta_n/2)^2\right) + 2 \cdot \exp\left(-\frac{3}{14} \cdot k_n\right) + c_{18} \cdot \frac{k_n}{n} \cdot \delta_n^{-d/p}.
\end{aligned}
$$

With

$$
k_n = \left\lceil \log(n) \cdot n^{\frac{2p}{(3+\alpha)\cdot p+d}} \right\rceil
$$

and

$$
\delta_n = \left(\frac{k_n}{n}\right)^{p/((1+\alpha)\cdot p+d)}
$$

we get the desired result. □

3.5 Proof of Theorem 6

Set

$$
\delta_n = c_{19}(k_n/n)^{p/d}.
$$

Since X has a density with respect for the Lebesgue-Borel measure bounded away from zero we get with ϵ_n as in the proof of Theorem 5 for each $x \in [0,1]^d$

$$
\mathbf{P}_X(S_{x,\epsilon_n}) \geq c_{20} \cdot \left(\frac{\delta_n}{2C}\right)^{d/p} \geq 4 \cdot \frac{k_n}{n}
$$

provided we choose c_{19} sufficiently large. Using this and proceeding otherwise as in the proof of Theorem 5 we get

$$
\begin{aligned}
&\mathbf{P}\{f_n(X) \neq Y\} - \mathbf{P}\{f^*(X) \neq Y\} \\
&\leq 2\bar{c} \cdot \delta_n^{1+\alpha} + 2 \cdot \exp\left(-2 \cdot k_n \cdot (\delta_n/2)^2\right) + 2 \cdot \exp\left(-\frac{3}{14} \cdot k_n\right).
\end{aligned}
$$

With

$$k_n = \log^2(n) \cdot n^{\frac{2p}{2p+d}}$$

the result follows. $\square$

Acknowledgement

The authors wish to thank L. Györfi for pointing out an error in an early version of this manuscript.

References

[1] Audibert, J.-Y. (2004). Classification using Gibbs estimators under complexity and margin assumptions. Preprint.

[2] Audibert, J.-Y. and Tsybakov, A. B. (2005). Fast learning rates for plug-in classifiers under the margin condition. Preprint.

[3] Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Annals of Statistics*, **9**, 1310–1319.

[4] Devroye, L. (1982). Necessary and sufficient conditions for pointwise convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **61**, pp. 467–481.

[5] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York.

[6] Greblicki, W., Krzyżak, A., and Pawlak, M. (1984). Distribution-free pointwise consistency of kernel regression estimate, *Annals of Statistics*, **12**, 1570-1575.

[7] Györfi, L. (1981). The rate of convergence of $k_n - NN$ regression estimates and classification rules. *IEEE Transactions on Information Theory*, **27**, 362-364.

[8] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics, Springer.

[9] Mack, Y. P. (1981). Local properties of k–NN regression estimates. *SIAM J. Alg. Disc. Math.* **2**, pp. 311-323.

[10] Mammen, E. and Tsybakov, A. B. (1999). Smooth discriminant analysis. *Annals of Statistics* **27**, pp. 1808–1829.

[11] Massart, P. and Nédélec, E. (2003). Risk bounds for statistical learning. Preprint.

[12] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications* **9**, pp. 141–142.

[13] Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**, 1040–1053.

[14] Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics* **32**, pp. 135–166.

[15] Tsybakov, A. B. and van de Geer, S. (2005). Square root penalty: adaptation to the margin in classification and in edge estimation. *Annals of Statistics* **33**, pp. 1203-1224.

[16] V. N. Vapnik (1998). *Statistical Learning Theory.* John Wiley & Sons.

[17] Walk, H. (2001). Strong universal pointwise consistency of recursive regression estimates. *Annals of the Institute of Statistical Mathematics* **53**, pp. 691–707.

[18] Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A* **26**, pp. 359–372.