The Annals of Statistics
2014, Vol. 42, No. 1, 1–28
DOI: 10.1214/13-AOS1152

ADAPTIVE ESTIMATION UNDER SINGLE-INDEX CONSTRAINT IN A REGRESSION MODEL[1]

BY OLEG LEPSKI AND NORA SERDYUKOVA

Aix-Marseille Université and Universidad de Concepción

The problem of adaptive multivariate function estimation in the single-index regression model with random design and weak assumptions on the noise is investigated. A novel estimation procedure that adapts simultaneously to the unknown index vector and the smoothness of the link function by selecting from a family of specific kernel estimators is proposed. We establish a pointwise oracle inequality which, in its turn, is used to judge the quality of estimating the entire function ("global" oracle inequality). Both the results are applied to the problems of pointwise and global adaptive estimation over a collection of Hölder and Nikol'skii functional classes, respectively.

1. Introduction. This paper deals with multivariate functions estimation. For the proposed estimator we establish local as well as global oracle inequalities and show how to use them for deriving minimax adaptive results.

Model and setup. We observe $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ following

$$Y_i = F(X_i) + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1.1}$$

where $d \geq 2$, the noise $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. centered random variables satisfying a tail probability condition (Assumption 1), and the design points $\{X_i\}_{i=1}^n$ are independent random vectors with common density g with respect to the Lebesgue measure. The sequences $\{\varepsilon_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ are assumed to be independent. The density g is known, however, in Section 4 we discuss how to extend our results to the case of unknown design density.

In addition, we assume that the function $F : \mathbb{R}^d \to \mathbb{R}$ has a single-index structure, that is, there exist unknown $f : \mathbb{R} \to \mathbb{R}$ and $\theta^* \in \mathbb{R}^d$ such that

$$F(x) = f\bigl(x^\top \theta^*\bigr). \tag{1.2}$$

A minimal technical assumption about f is that it belongs to some Hölder ball, yet the knowledge of this ball will not be required for the proposed estimation procedure; see the discussion after Assumption 3 for more details.

Received November 2012; revised July 2013.

[1]Supported in part by the DFG FOR 916 Statistical Regularisation and Qualitative Constraints.

MSC2010 subject classifications. Primary 62G05; secondary 62G08, 62G20.

Key words and phrases. Adaptive estimation, lower bounds, minimax rate, nonparametric regression, oracle inequalities, single-index model, structural adaptation.

The paper aims at estimating the entire function F on $[-1/2, 1/2]^2$ or its value $F(t)$, $t \in [-1/2, 1/2]^2$, from the data $\{(X_i, Y_i)\}_{i=1}^n$ without any prior knowledge about the nuisance parameters $f(\cdot)$ and θ^*. The unit square is chosen for notation convenience; and all the results remain true when $[-1/2, 1/2]^2$ is replaced by an arbitrary bounded interval of $\mathbb{R}^2$.

Throughout the paper we adopt the following notation. The joint distribution of the sequence $\{(X_i, Y_i)\}_{i=1}^n$ will be denoted by $\mathbb{P}_F^{(n)}$, and those of $\{(X_i, \varepsilon_i)\}_{i=1}^n$ by $\mathbb{P}_{X,\varepsilon}^{(n)}$. In addition, $\mathbb{P}_X^{(n)}$ and $\mathbb{P}_\varepsilon^{(n)}$ stand for the marginal distributions of $\{X_i\}_{i=1}^n$ and $\{\varepsilon_i\}_{i=1}^n$, respectively.

To judge the quality of estimation, we use either the risk determined by the L_r norm, $\|\cdot\|_r$, on $[-1/2, 1/2]^2$ with $r \in [1, \infty)$:

$$\mathcal{R}_r^{(n)}(\widehat{F}, F) = \mathbb{E}_F^{(n)} \|\widehat{F} - F\|_r, \tag{1.3}$$

a "global" risk; or the "pointwise" risk defined as follows:

$$\mathcal{R}_{r,t}^{(n)}(\widehat{F}, F) = \big(\mathbb{E}_F^{(n)} |\widehat{F}(t) - F(t)|^r\big)^{1/r}, \qquad t \in [-1/2, 1/2]^2. \tag{1.4}$$

Here $\widehat{F}(\cdot)$ is an estimator, that is, an $\{(X_i, Y_i)\}_{i=1}^n$-measurable function, and $\mathbb{E}_F^{(n)}$ denotes the mathematical expectation with respect to $\mathbb{P}_F^{(n)}$.

All the results established in the paper, except the lower bound given in Theorem 4, are obtained for $d = 2$. The principal difficulties with the case of arbitrary dimension are commented in Remark 2. It is noteworthy that the single-index modeling, even if $d = 2$, is a direct generalization of the univariate regression model. Therefore, our results, mainly presented in Section 2.2, generalize in several directions the existing ones obtained for the univariate random design regression (see the discussion after Theorem 5).

Main assumptions. Let us formulate the principal assumptions used in the sequel. They are imposed on the distributions of the design and noise variables as well as on the approximation property of the link function.

ASSUMPTION 1. The random variable ε_1 has a symmetric distribution with density p with respect to the Lebesgue measure. Moreover, there exist $\Upsilon > 0$, $\Omega \in (0, 1]$, and $\omega > 0$ such that

$$p \in \mathfrak{P} = \left\{\ell : \mathbb{R} \to \mathbb{R}_+ \,\middle|\, \int_x^\infty \ell(y)\,\mathrm{d}y \le \Upsilon e^{-\Omega x^\omega}\ \forall x \ge 0\right\}.$$

The assumption holds, for example, for the Gaussian, Laplace or, more generally, for the symmetrized Weibull distribution. In the following, the functional class $\mathfrak{P}$ is considered as fixed.

ASSUMPTION 2. There exists $\underline{g} \in (0, 1)$ such that $\inf_{x \in [-3,3]^2} g(x) \ge \underline{g}$.

The assumption holds obviously if the design points are uniformly distributed on any bounded Borel set containing $[-3, 3]^2$. The imposed condition is "fitted" to the estimation over $[-1/2, 1/2]^2$ that explains the set $[-3, 3]^2$. When estimating over a rectangle $[a, b] \times [c, e] \in \mathbb{R}^2$, the infimum should be taken over $[a-5/2, b+5/2] \times [c-5/2, e+5/2]$. If M from Assumption 3 below is known, the above condition can be relaxed to $[a-2, b+2] \times [c-2, e+2]$. We also remark that independently of the values a, b, c, e Assumption 2 is fulfilled if $g \in \mathbb{C}(\mathbb{R}^2)$ and $g(x) > 0$ for any $x \in \mathbb{R}^2$.

ASSUMPTION 3. There exist $\beta_0 \in (0, 1)$ and $M > 0$ such that

$$f \in \mathbb{F}(\beta_0, M) = \left\{ U : \mathbb{R} \to \mathbb{R} \middle| \|U\|_\infty + \sup_{y_1, y_2 \in \mathbb{R}} \frac{|U(y_1) - U(y_2)|}{|y_1 - y_2|^{\beta_0}} \le M \right\}.$$

The latter assumption guarantees that the link function is smooth. However, it is important to emphasize that β_0 and M are not supposed to be known a priori. In particular, they are not involved in our estimation procedure. On the other hand, both the parameters restrict the minimal sample size needed to justify the theoretical results involved. Set for any $n \in \mathbb{N}^*$

$$h_{\min} = n^{-1} \ln^{1+2/\omega}(n), \qquad \mathfrak{h} = \sqrt{n^{-1} \ln^{1+1/\omega}(n)}. \tag{1.5}$$

In the sequel it will be assumed that $n \ge n_0$, where

$$n_0 = \inf\{ m \in \mathbb{N}^* | (M \vee 1) \max\{ \mathfrak{h}^{\beta_0}, \ln^{1/\omega}(n) h_{\min}^{\beta_0} \} \le 1 \ \forall n \ge m \}. \tag{1.6}$$

To finish this section, we remark that all the presented results remain true if one assumes that $f \in \mathbb{F}(0, M)$, that is, is uniformly bounded, and M is known.

Objectives. For clarity of presentation, it is assumed that the index vector $\theta^* \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1}$ stands for the unite sphere in $\mathbb{R}^d$. However, in Section 2.1.4 it is shown that our results can be extended to the case $\theta^* \in \mathbb{R}^2$.

The goal of our studies is at least threefold. We first seek an estimation procedure $\widehat{F}(t)$, $t \in [-1/2, 1/2]^2$, for F which could be applicable to any function F satisfying assumption (1.2). Moreover, we would like to bound the risk of this estimator uniformly over the set $\mathbb{F}(\beta_0, M) \times \mathbb{S}^1$. More precisely, we want to establish for $\widehat{F}(t)$ the so-called local oracle inequality—at any point $t \in [-1/2, 1/2]^2$ the risk of $\widehat{F}(t)$ should be bounded as follows:

$$\mathcal{R}_{r,t}^{(n)}(\widehat{F}, F) \le C_r A_{f,\theta^*}^{(n)}(t) \qquad \forall f \in \mathbb{F}(\beta_0, M), \ \forall \theta^* \in \mathbb{S}^1. \tag{1.7}$$

Here $A_{f,\theta^*}^{(n)}(\cdot)$ is completely determined by the function f, vector θ^* and observations number n, while C_r is a constant independent of F and n.

After being established, the local oracle inequality allows deriving minimax adaptive results for the function estimation at a given point. Indeed, let $\{\mathbb{F}(\gamma),$

$\gamma \in \Gamma\}$ be a collection of functional classes such that $\bigcup_{\gamma\in\Gamma} \mathbb{F}(\gamma) \subseteq \mathbb{F}(\beta_0, M)$. For any $\gamma \in \Gamma$ define

$$\phi_n(\gamma) = \inf_{\widetilde{F}} \sup_{(f,\theta^*)\in\mathbb{F}(\gamma)\times\mathbb{S}^1} \mathcal{R}_{r,t}^{(n)}(\widetilde{F}, F),$$

where the infimum is taken over all possible estimators. The quantity $\phi_n(\gamma)$ is the minimax risk on $\mathbb{F}(\gamma) \times \mathbb{S}^1$. In the framework of minimax adaptive estimation, the task is to construct an estimator F^* such that for any $\gamma \in \Gamma$

$$\sup_{(f,\theta^*)\in\mathbb{F}(\gamma)\times\mathbb{S}^1} \mathcal{R}_{r,t}^{(n)}(F^*, F) \asymp \phi_n(\gamma), \qquad n \to \infty. \tag{1.8}$$

The estimator F^* satisfying (1.8) is called optimally rate adaptive over the collection $\{\mathbb{F}(\gamma), \gamma \in \Gamma\}$. Subsequently, let (1.7) be proved; and let for any $\gamma \in \Gamma$

$$\sup_{(f,\theta^*)\in\mathbb{F}(\gamma)\times\mathbb{S}^1} A_{f,\theta^*}^{(n)}(t) \asymp \phi_n(\gamma), \qquad n \to \infty.$$

Then one can assert that the estimator $\widehat{F}$ is adaptive over $\{\mathbb{F}(\gamma), \gamma \in \Gamma\}$.

Thus, the first step is to prove (1.7). To the best of our knowledge, such results do not exist in the context of regression with random design not only under the single-index constraint, but also in univariate regression.

Next, (1.7) is applied to minimax adaptive estimation over Hölder classes, $\{\mathbb{F}(\gamma) = \mathbb{H}(\beta, L), \gamma = (\beta, L)\}$; see Section 2.2 for pertinent definitions. We will find the minimax rate over $\mathbb{H}(\beta, L) \times \mathbb{S}^1$ and prove that $\widehat{F}$ achieves it, that is, is optimally rate adaptive. This result is quite surprising because, if θ^* is fixed, say, $\theta^* = (1, 0)^\top$, it is well known that an optimally adaptive estimator does not exist; see Lepskiĭ (1990) for the Gaussian white noise model, Brown and Low (1996) for density estimation, and Gaïffas (2007) for regression.

Local oracle inequality (1.7) allows us to bound from above the "global" risk as well. Indeed, for any $r \geq 1$, in view of Jensen's inequality and Fubini's theorem, $[\mathcal{R}_r^{(n)}(\widehat{F}, F)]^r \leq \mathbb{E}_F^{(n)} \|\widehat{F} - F\|_r^r = \|\mathcal{R}_{r,\cdot}^{(n)}(\widehat{F}, F)\|_r^r$ and, therefore,

$$\mathcal{R}_r^{(n)}(\widehat{F}, F) \leq C_r \|A_{f,\theta^*}^{(n)}\|_r. \tag{1.9}$$

Inequality (1.9) is called the global oracle inequality, and in the considered framework it supplies new results. As local oracle inequality (1.7) is a powerful tool for deriving minimax adaptive results in pointwise estimation, so inequality (1.9) can be used for constructing adaptive estimators of F.

We will consider a collection of Nikol'skii classes $\mathbb{N}_p(\beta, L)$ (see Definition 2), where $\beta, L > 0$ and $1 \leq p < \infty$. When considering these classes, we aim at estimating functions with inhomogeneous smoothness. This means that the underlying function can be very regular on some parts of its domain and rather irregular on the other sets. We will compute bounds for

$$\sup_{(f,\theta^*)\in\mathbb{N}_p(\beta,L)\times\mathbb{S}^1} \|A_{f,\theta^*}^{(n)}\|_r$$

and show that, if $(2\beta + 1)p < r$, the rate of convergence is the minimax rate over $\mathbb{N}_p(\beta, L) \times \mathbb{S}^1$. This means that our estimator $\widehat{F}$ is optimally rate adaptive over the collection $\{\mathbb{N}_p(\beta, L) \times \mathbb{S}^1, \beta > 0, L > 0\}$ whenever $(2\beta + 1)p < r$. In the case $(2\beta + 1)p \geq r$, we will show that the latter bound differs from the bound on the minimax risk by a logarithmic factor. Following the contemporary language, we say that $\widehat{F}$ is "nearly" adaptive. The construction of an optimally rate adaptive over the entire range of the Nikol'skii classes estimator under the single-index constraint (1.2) is an open question.

All presented results are completely new. The adaptive estimation under the L_r loss and single-index constraint, except the case $r = 2$ in Gaïffas and Lecué (2007), was not studied. Note, however, that the cited result was obtained under the Gaussian errors model and over the Hölder classes that do not admit the consideration of functions with inhomogeneous smoothness.

Remarks. It turns out that the adaptation to the unknown θ^* and $f(\cdot)$ can be viewed as selecting from a special family of kernel estimators in the spirit of that of Lepskiĭ (1990), Kerkyacharian, Lepski and Picard (2001), Goldenshluger and Lepski (2008). However, our selection rule is quite different from the aforementioned proposals, and it allows us to solve the problem of minimax adaptive estimation under the L_r losses over a collection of Nikol'skii classes.

It is worth mentioning that the single-index model is particularly popular in econometrics [see, e.g., Horowitz (1998), Maddala (1983)]. The estimation, nevertheless, is usually performed under smoothness assumptions on the link function. One usually uses the L_2 losses, and the available methodology is based on these restrictions. To the best of our knowledge, the only exceptions are Golubev (1992) for the minimax estimation under the projection pursuit constraints, and Goldenshluger and Lepski (2009) for adaptation to unknown smoothness and structure.

Organization of the paper. In Section 2.1 we present our selection rule and establish for it local and global oracle inequalities. Section 2.2 is devoted to the application of these results to minimax adaptive estimation. The proofs of the main results are given in Section 3; Section 4 discusses an unknown design density, and the proofs of lemmas are moved to the supplementary material [Lepski and Serdyukova (2014)].

2. Main results. In this section we motivate and explain our procedure and prove the local and global oracle inequalities. Then we apply these results to adaptive estimation over a collection of Hölder classes (pointwise estimation) and over a collection of Nikol'skii classes (estimating the entire function with the accuracy of an estimator measured under the L_r risk).

2.1. *Oracle approach*. Let $\mathcal{K}:\mathbb{R}\to\mathbb{R}$ be a function (kernel) satisfying $\int\mathcal{K}=1$. With any such $\mathcal{K}$, any $z\in\mathbb{R}$, $h\in(0,1]$ and any $f\in\mathbb{F}(\beta_0,M)$, we associate the quantity

$$\Delta_{\mathcal{K},f}(h,z)=\sup_{\delta\le h}\left|\frac{1}{\delta}\int\mathcal{K}\left(\frac{u-z}{\delta}\right)[f(u)-f(z)]\,\mathrm{d}u\right|.$$

Note that the kernel smoother $\delta^{-1}\int\mathcal{K}([u-z]/\delta)f(u)\,\mathrm{d}u$ can be understood as an approximation of the function f at the point z. Thus, $\Delta_{\mathcal{K},f}(h,z)$ is a monotonous approximation error provided by this kernel smoother. In particular, under Assumption 3, we have $\Delta_{\mathcal{K},f}(h,z)\to 0$ as $h\to 0$.

In what follows, $\|\mathcal{K}\|_p$, $1\le p\le\infty$, denotes the L_p norm of $\mathcal{K}$ and we will assume that the kernel $\mathcal{K}$ satisfies the following condition.

ASSUMPTION 4. (1) $\operatorname{supp}(\mathcal{K})\subseteq[-1/2,1/2]$, $\int\mathcal{K}=1$, $\mathcal{K}$ is symmetric;
(2) there exists $Q>0$ such that $|\mathcal{K}(u)-\mathcal{K}(v)|\le Q|u-v|$ $\forall u,v\in\mathbb{R}$.

2.1.1. *Oracle estimator*. For any $y\in\mathbb{R}$, denote the Hardy–Littlewood maximal function of $\Delta_{\mathcal{K},f}(h,\cdot)$ [see, e.g., Wheeden and Zygmund (1977)] by

$$\Delta^*_{\mathcal{K},f}(h,y)=\sup_{a>0}\frac{1}{2a}\int_{y-a}^{y+a}\Delta_{\mathcal{K},f}(h,z)\,\mathrm{d}z.$$

Clearly, $\Delta^*_{\mathcal{K},f}(h,\cdot)\ge\Delta_{\mathcal{K},f}(h,\cdot)$ for any $f\in\mathbb{F}(\beta_0,M)$. Now, let us define the oracle estimator. For any $y\in\mathbb{R}$ and $h_{\min}$ defined in (1.5), set

$$h^*_{\mathcal{K},f}(y)=\sup\bigl\{h\in[h_{\min},1]\,|\,\sqrt{nh}\,\Delta^*_{\mathcal{K},f}(h,y)\le\|\mathcal{K}\|_\infty\sqrt{\ln(n)}\bigr\}. \tag{2.1}$$

Note that $\Delta^*_{\mathcal{K},f}(h,\cdot)\le M\|\mathcal{K}\|_1h^{\beta_0}$ for any $f\in\mathbb{F}(\beta_0,M)$ and any $h>0$. Hence, $\sqrt{nh_{\min}}\,\Delta^*_{\mathcal{K},f}(h_{\min},\cdot)\le\|\mathcal{K}\|_1\sqrt{\ln(n)}$ for any $n\ge n_0$ in view of (1.6). Next, Assumption 4(2) implies that $\Delta^*_{\mathcal{K},f}(\cdot,y)$ is continuous, hence,

$$\text{either}\quad \sqrt{nh^*_{\mathcal{K},f}(y)}\,\Delta^*_{\mathcal{K},f}\bigl(h^*_{\mathcal{K},f}(y),y\bigr)=\|\mathcal{K}\|_\infty\sqrt{\ln(n)}, \tag{2.2}$$

$$\text{or}\quad \sqrt{nh}\,\Delta^*_{\mathcal{K},f}(h,y)\le\|\mathcal{K}\|_\infty\sqrt{\ln(n)}\qquad\forall h\in[h_{\min},1]. \tag{2.3}$$

Here we have also used that $\|\mathcal{K}\|_1\le\|\mathcal{K}\|_\infty$ in view of Assumption 4(1).

The quantity similar to $h^*_{\mathcal{K},f}$ first appeared in Lepski, Mammen and Spokoiny (1997) for estimating univariate functions with inhomogeneous smoothness. Some years later, this idea was further developed for multivariate function estimation; see Kerkyacharian, Lepski and Picard (2001), Goldenshluger and Lepski (2008) and the more detailed discussion of the oracle approach therein. Following their lead, we advance it for the estimation under the single-index constraint. The basic idea behind our selection rule is simple.

For any $(\theta, h) \in \mathbb{S}^1 \times [h_{\min}, 1]$, define the matrix

$$E_{(\theta,h)} = \begin{pmatrix} h^{-1}\theta_1 & h^{-1}\theta_2 \\ -\theta_2 & \theta_1 \end{pmatrix}, \qquad \det(E_{(\theta,h)}) = h^{-1}$$

and consider the family of kernel estimators with $K(u, v) = \mathcal{K}(u)\mathcal{K}(v)$ so that

$$\mathcal{F} = \left\{ \widehat{F}_{(\theta,h)}(\cdot) = \frac{\det(E_{(\theta,h)})}{n} \sum_{i=1}^{n} \frac{K(E_{(\theta,h)}(X_i - \cdot))}{g(X_i)} Y_i, (\theta, h) \in \mathbb{S}^1 \times [h_{\min}, 1] \right\}.$$

We remark that Assumptions 2 and 4(1) assure well-definiteness of $\widehat{F}_{(\theta,h)}$ because $K(E_{(\theta,h)}(x - t)) = 0\ \forall x \in [-3/2, 3/2]^2$ and $\forall t \in [-1/2, 1/2]^2$.

The choice $\theta = \theta^*$ and $h = h^* := h^*_{\mathcal{K},f}(t^\top\theta^*)$ leads to the *oracle estimator* $\widehat{F}_{(\theta^*,h^*)}$. Note that $\widehat{F}_{(\theta^*,h^*)}$ is not an estimator in the usual sense because it depends on the function F to be estimated [more precisely, on (f, θ^*) which determines F]. The meaning of $\widehat{F}_{(\theta^*,h^*)}$ is explained by the following result based on the straightforward application of Rozenthal's inequality.

PROPOSITION 1. *For any $(f, \theta^*) \in \mathbb{F}(\beta_0, M) \times \mathbb{S}^1$, $r \geq 1$ and $n \geq n_0$,*

$$\mathcal{R}^{(n)}_{r,t}(\widehat{F}_{(\theta^*,h^*)}, F) \leq c \ln^{1/2}(n)\big[nh^*_{\mathcal{K},f}(t^\top\theta^*)\big]^{-1/2} \qquad \forall t \in [-1/2, 1/2]^2,$$

where $c > 0$ is a numerical constant independent of n.

This result indicates that the "oracle" knows the exact value of θ^* and the optimal, up to $\ln(n)$, trade-off h^* between the approximation error induced by $\Delta^*_{\mathcal{K},f}(h^*, \cdot)$ and the stochastic error of the kernel estimator from $\mathcal{F}$ with bandwidth h^*. It explains why the "oracle" chooses the "estimator" $\widehat{F}_{(\theta^*,h^*)}$. Below we propose a "real," based on the observation, estimator $\widehat{F}$, which mimics the oracle—for any $(f, \theta^*) \in \mathbb{F}(\beta_0, M) \times \mathbb{S}^1$, $r \geq 1$ and $n \geq n_0$,

$$\mathcal{R}^{(n)}_{r,t}(\widehat{F}, F) \leq c' \ln^{1/2}(n)\big[nh^*_{\mathcal{K},f}(t^\top\theta^*)\big]^{-1/2} \qquad \forall t \in [-1/2, 1/2]^2,$$

where c' is an absolute constant independent of n and the underlying function F. The latter result is a local oracle inequality. The construction of the estimator $\widehat{F}$ is based on the data-driven selection from the family $\mathcal{F}$.

2.1.2. *Selection rule.* For any $\theta, \nu \in \mathbb{S}^1$ and any $h \in [h_{\min}, 1]$, define

$$\overline{E}_{(\theta,h)(\nu,h)} = \begin{pmatrix} \dfrac{(\theta_1 + \nu_1)}{2h(1 + |\nu^\top\theta|)} & \dfrac{(\theta_2 + \nu_2)}{2h(1 + |\nu^\top\theta|)} \\ -\dfrac{(\theta_2 + \nu_2)}{2(1 + |\nu^\top\theta|)} & \dfrac{(\theta_1 + \nu_1)}{2(1 + |\nu^\top\theta|)} \end{pmatrix},$$

where

$$E_{(\theta,h)(\nu,h)} = \begin{cases} \overline{E}_{(\theta,h)(\nu,h)}, & \nu^\top\theta \geq 0, \\ \overline{E}_{(-\theta,h)(\nu,h)}, & \nu^\top\theta < 0, \end{cases} \qquad \frac{1}{4h} \leq \det(E_{(\theta,h)(\nu,h)}) \leq \frac{1}{2h}.$$

A kernel estimator associated with the matrix $E_{(\theta,h)(\nu,h)}$ is defined by

$$\widehat{F}_{(\theta,h)(\nu,h)}(\cdot) = \frac{\det(E_{(\theta,h)(\nu,h)})}{n} \sum_{i=1}^{n} \frac{K(E_{(\theta,h)(\nu,h)}(X_i - \cdot))}{g(X_i)} Y_i. \tag{2.4}$$

The definition of $\widehat{F}_{(\theta,h)(\nu,h)}$ is legitimate because $K(E_{(\theta,h)}(x-t)) = 0$ $\forall x \in [-5/2, 5/2]^2$ and $\forall t \in [-1/2, 1/2]^2$.

For any $u_1, u_2 \in \mathbb{R}$, set $K_{\mathfrak{h}}(u_1, u_2) = \mathfrak{h}^{-2}\mathcal{K}(u_1/\mathfrak{h})\mathcal{K}(u_2/\mathfrak{h})$ and define

$$\widehat{F}(v) = n^{-1} \sum_{i=1}^{n} g^{-1}(X_i) K_{\mathfrak{h}}(X_i - v) Y_i, \qquad \widehat{F}_\infty = 2\|\widehat{F}\|_\infty + 2C_5(n),$$

where $\|\widehat{F}\|_\infty = \sup_{v \in [-5/2,5/2]^2} |\widehat{F}(v)|$ and $\mathfrak{h}$ is defined in (1.5). Put also

$$\mathrm{TH}(\eta) = 2\big[\|\mathcal{K}\|_\infty^2 \sqrt{\ln(n)} + \widehat{F}_\infty C_1(n) + C_2(n)\big](\eta n)^{-1/2}, \qquad \eta \in (0, 1].$$

The quantities $C_1(n)$, $C_2(n)$ and $C_5(n)$ are listed in Section 3.1.

Set $\mathcal{H}_n = \{h_k = 2^{-k}, k \in \mathbb{N}^0\} \cap [2^{-1}h_{\min}, 1]$ and let for any $\theta \in \mathbb{S}^1$ and $h \in \mathcal{H}_n$,

$$R_t^{(1)}(\theta, h) = \sup_{\eta \in \mathcal{H}_n : \eta \le h} \Big[\sup_{\nu \in \mathbb{S}^1} \big|\widehat{F}_{(\theta,\eta)(\nu,\eta)}(t) - \widehat{F}_{(\nu,\eta)}(t)\big| - \mathrm{TH}(\eta)\Big]_+,$$

$$R_t^{(2)}(h) = \sup_{\eta \in \mathcal{H}_n : \eta \le h} \Big[\sup_{\theta \in \mathbb{S}^1} \big|\widehat{F}_{(\theta,h)}(t) - \widehat{F}_{(\theta,\eta)}(t)\big| - \mathrm{TH}(\eta)\Big]_+.$$

Subsequently, define $(\hat{\theta}, \hat{h})$ as a solution of the following minimization problem:

$$\begin{aligned} &R_t^{(1)}(\hat{\theta}, \hat{h}) + R_t^{(2)}(\hat{h}) + \mathrm{TH}(\hat{h}) \\ &\quad = \inf_{(\theta,h) \in \mathbb{S}^1 \times \mathcal{H}_n} \big[R_t^{(1)}(\theta, h) + R_t^{(2)}(h) + \mathrm{TH}(h)\big]. \end{aligned} \tag{2.5}$$

Then our final estimator is $\widehat{F}(t) = \widehat{F}_{(\hat{\theta},\hat{h})}(t)$, where $(\hat{\theta}, \hat{h})$ is obtained by minimizing (2.5).

REMARK 1. We note that Assumption 4(2) guarantees that all random fields involved in the description of selection rule (2.5) are continuous on $\mathbb{S}^1$. Moreover, the set $\mathcal{H}_n$ is finite. Thus, $(\hat{\theta}, \hat{h})$ is $\{(X_i, Y_i)\}_{i=1}^n$-measurable and $(\hat{\theta}, \hat{h}) \in \mathbb{S}^1 \times \mathcal{H}_n$ [see Jennrich (1969)].

REMARK 2. Our selection rule (2.5) is defined in the case $d = 2$. The main difficulty in extending it to $d > 2$ consists in the construction of the matrix $E_{(\theta,h)(\nu,h)}$ for any vectors $\theta, \nu \in \mathbb{S}^{d-1}$. Indeed, analyzing the proof of Theorem 1, we remark that the following properties should be fulfilled:

$$E_{(\theta,h)(\nu,h)} \in \mathcal{E}_{a,A}, \qquad E_{(\theta,h)(\nu,h)} = \pm E_{(\nu,h)(\theta,h)} \qquad \forall \theta, \nu \in \mathbb{S}^{d-1}, \forall h \in \mathcal{H}_n,$$

where the class of matrices $\mathcal{E}_{a,A}$ is defined in (3.2). If $d = 2$, these requirements hold. However, we were not able to construct a class of matrices obeying latter restrictions in the dimension strictly larger than 2. Note, nevertheless, that if such a class would be found, our results could be extended to $d > 2$ without any additional consideration.

2.1.3. *Local and global oracle inequalities.* We reinforce restriction (1.6) on the minimal sample size n. Let $n_1 \geq 1$ be defined as follows:

$$n_1 = \inf\{m \in \mathbb{N}^* : (n\mathfrak{h}^2)^{-1/2} C_3(n) \leq 1/2 \ \forall n \geq m\}, \tag{2.6}$$

where $\mathfrak{h}$ is defined in (1.5) and $C_3(n)$ is given at the beginning of Section 3.1. All our results below will be proved under the condition $n \geq n_0 \vee n_1$.

First, we note that n_1 is well-defined since $(n\mathfrak{h}^2)^{-1/2} C_3(n) \to 0$ as $n \to \infty$. Next, contrary to restriction (1.6) that relates the sample size n to the quantities β_0 and M from Assumption 3, restriction (2.6) links the minimal value of n with the quantity $\underline{g}$ appearing in Assumption 2.

THEOREM 1. *For any $(f, \theta^*) \in \mathbb{F}(\beta_0, M) \times \mathbb{S}^1$, $r \geq 1$ and $n \geq n_0 \vee n_1$,*

$$\mathcal{R}_{r,t}^{(n)}(\widehat{F}_{(\hat{\theta},\hat{h})}, F) \leq c_1 \left[\frac{\ln(n)}{n h^*_{\mathcal{K},f}(t^\top \theta^*)}\right]^{1/2} + c_2 n^{-1/2} \qquad \forall t \in [-1/2, 1/2]^2.$$

The constants c_1 and c_2 are independent of n and F and their explicit expressions can be extracted from the proof of the theorem.

As already mentioned, the global oracle inequality is obtained by integrating the local oracle inequality. Indeed, for any $r \geq 1$, using Jensen's inequality and Fubini's theorem, we have $\mathcal{R}_r^{(n)}(\widehat{F}, F) \leq \|\mathcal{R}_{r,\cdot}^{(n)}(\widehat{F}, F)\|_r$ so

$$\mathcal{R}_r^{(n)}(\widehat{F}, F) \leq c_1 \left\{\int_{[-1/2,1/2]^2} \left[\frac{\ln(n)}{n h^*_{\mathcal{K},f}(t^\top \theta^*)}\right]^{r/2} \mathrm{d}t\right\}^{1/r} + c_2 n^{-1/2}.$$

Integration by substitution yields

$$\int_{[-1/2,1/2]^2} \left[\frac{\ln(n)}{n h^*_{\mathcal{K},f}(t^\top \theta^*)}\right]^{r/2} \mathrm{d}t \leq \int_{-1/2}^{1/2} \left[\frac{\ln(n)}{n h^*_{\mathcal{K},f}(z)}\right]^{r/2} \mathrm{d}z,$$

that leads to the following bound.

THEOREM 2. *For any $(f, \theta^*) \in \mathbb{F}(\beta_0, M) \times \mathbb{S}^1$, $r \geq 1$ and $n \geq n_0 \vee n_1$,*

$$\mathcal{R}_r^{(n)}(\widehat{F}_{(\hat{\theta},\hat{h})}, F) \leq c_1 \left\| \frac{\ln(n)}{n h^*_{\mathcal{K},f}} \right\|_{r/2}^{1/2} + c_2 n^{-1/2}.$$

2.1.4. *Extension to the case $\theta^* \notin \mathbb{S}^1$.* Define $f_{\theta^*}(t) = f(|\theta^*|_2 t)$, $\vartheta^* = \theta^*/|\theta^*|_2$ and let $F_{\theta^*}(t) := f_{\theta^*}(t^\top \vartheta^*)$. Obviously, for all $t \in \mathbb{R}^2$ we have $f_{\theta^*}(t^\top \vartheta^*) = f(t^\top \theta^*)$ that implies $F_{\theta^*}(\cdot) \equiv F(\cdot)$ so the estimation of F is equivalent to the estimation of F_{θ^*}. Because $\vartheta \in \mathbb{S}^1$, Theorems 1 and 2 are applicable. To this end, it suffices to replace f by f_{θ^*} in the definition of $h^*_{\mathcal{K},f}(\cdot)$. In general, however, there is no universal way of expressing $h^*_{\mathcal{K},f_{\theta^*}}(\cdot)$ via $h^*_{\mathcal{K},f}(\cdot)$, although in particular cases, mainly in adaptive estimation over classes of smooth functions, it is often possible.

2.2. *Adaptive estimation.* In this section we first apply the local oracle inequality given in Theorem 1 to the problem of pointwise adaptive estimation over a collection of Hölder classes. Next, we study adaptive estimation under the L_r losses over a collection of Nikol'skii classes. The corresponding result is deduced from the global oracle inequality proved in Theorem 2.

Assume throughout this section that the kernel $\mathcal{K}$ obeys additionally Assumption 5 below; we then introduce the following notation: for any $a > 0$, let m_a be the maximal integer strictly less than a.

ASSUMPTION 5. There exists $\mathbf{b} > 0$ such that

$$\int z^j \mathcal{K}(z)\,\mathrm{d}z = 0 \qquad \forall j = 1, \dots, m_{\mathbf{b}}.$$

2.2.1. *Pointwise adaptive estimation.* We start with some definitions.

DEFINITION 1. Let $\beta > 0$ and $L > 0$. A function $\ell : \mathbb{R} \to \mathbb{R}$ belongs to the Hölder class $\mathbb{H}(\beta, L)$ if ℓ is m_β-times continuously differentiable, $\|\ell^{(m)}\|_\infty \le L$ for all $m \le m_\beta$, and

$$\left|\ell^{(m_\beta)}(u+h) - \ell^{(m_\beta)}(u)\right| \le L h^{\beta - m_\beta} \qquad \forall u \in \mathbb{R}, h > 0.$$

The aim is to estimate the function $F(t)$ at a given point $t \in [-1/2, 1/2]^2$ under the assumption that $F \in \mathbb{F}(\mathbf{b}) := \bigcup_{\beta \le \mathbf{b}} \bigcup_{L>0} \mathbb{F}_2(\beta, L)$, where

$$\mathbb{F}_d(\beta, L) = \{F : \mathbb{R}^d \to \mathbb{R} | F(z) = f(z^\top \theta), f \in \mathbb{H}(\beta, L), \theta \in \mathbb{S}^{d-1}\},$$

the constant $\mathbf{b}$ is from Assumption 5, and $d \ge 2$ is the dimension. We will see that $\mathbf{b}$ can be an arbitrary number but it must be chosen a priori.

THEOREM 3. *Let $\mathbf{b} > 0$ be fixed; and let additionally Assumptions 4 and 5 hold. Then, for any $\beta \le \mathbf{b}$, $L > 0$, $r \ge 1$ and $t \in [-1/2, 1/2]^2$,*

$$\sup_{F \in \mathbb{F}_2(\beta, L)} \mathcal{R}^{(n)}_{r,t}(\widehat{F}_{(\hat\theta, \hat h)}, F) \le \varkappa_1 \psi_n(\beta, L),$$

where $\psi_n(\beta, L) = L^{1/(2\beta+1)}[n^{-1}\ln(n)]^{\beta/(2\beta+1)}$ and $\varkappa_1$ is independent of n.

The proof of the theorem is based on the evaluation of the uniform over $\mathbb{H}_d(\beta, L)$ lower bound for $h^*_{\mathcal{K},f}(\cdot)$ and on the application of Theorem 1. We note that a similar upper bound for the minimax risk appeared in Goldenshluger and Lepski (2008) in the framework of Gaussian white noise model, but the estimation procedure used there is different from our selection rule.

The main question, however, is if $\psi_n(\beta, L)$ coincides with the minimax rate for any given value of β and L? To answer it, we need some additional assumptions on the densities of the noise variable ε_1 and design variable X_1.

ASSUMPTION 6. There exist $\mathfrak{q}, \mathfrak{Q} > 0$ such that, for any $\upsilon_1, \upsilon_2 \in [-\mathfrak{q}, \mathfrak{q}]$,

$$\int_{\mathbb{R}} p(y + \upsilon_1) p(y + \upsilon_2) p^{-1}(y)\, \mathrm{d}y \le 1 + \mathfrak{Q}|\upsilon_1 \upsilon_2|.$$

It is easy to see that the density of the normal law $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$, obeys the aforementioned assumption. In general, this assumption is fulfilled if the density p is regular and decreases rapidly at infinity. More precisely, if the Fisher information corresponding to the density p is finite and the function $\int [p'(y + \cdot)]^2 p^{-1}(y)\, \mathrm{d}y$ is continuous at zero, Assumption 6 is verified.

ASSUMPTION 7. There exist $\mathfrak{g} > 0$ and $\varpi > 1$ such that, for all $x \in \mathbb{R}^d$, $g(x) \le (1 + |x|_2^{\varpi})^{-1} \mathfrak{g}$. Here $|\cdot|_2$ is the Euclidean vector norm on $\mathbb{R}^d$.

We remark that the imposed assumption is very weak and holds for the majority of probability distributions used in statistical applications.

THEOREM 4. *Let Assumptions 6 and 7 be fulfilled. Then, for any* $t \in [-1/2, 1/2]^d$, $d \ge 2$, $r \ge 1$, $\beta, L > 0$, *and any* $n \in \mathbb{N}^*$ *large enough*,

$$\inf_{\widetilde{F}} \sup_{F \in \mathbb{F}_d(\beta, L)} \mathcal{R}^{(n)}_{r,t}(\widetilde{F}, F) \ge \varkappa_2 \psi_n(\beta, L),$$

where the infimum is over all possible estimators. Here $\varkappa_2$ *is a numerical constant independent of* n *and* L, *and* $\psi_n(\beta, L)$ *is defined in Theorem 3.*

To the best of our knowledge, this lower bound is new. It is worth mentioning that Assumption 6 is close to being necessary. One can give examples where this condition does not hold and Theorem 4 is not true anymore.

Theorems 3 and 4 indicate that the estimator $\widehat{F}_{(\hat{\theta}, \hat{h})}$ is minimax adaptive with respect to the collection $\{\mathbb{F}_d(\beta, L), \beta \le \mathbf{b}, L > 0\}$. As already mentioned, this result is quite surprising. Indeed, if, for example, $\theta = (1, 0)^\top$, that is, is known, then $\mathbb{F}(\beta, L) = \mathbb{H}(\beta, L)$, and the considered estimation problem reduces to estimation of f at a point in the univariate regression model. As it is shown in Gaïffas (2007), an adaptive estimator over $\{\mathbb{H}(\beta, L), \beta \le \mathbf{b}, L > 0\}$ does not exist and a price for adaption appears. The latter means that the asymptotic bound on the minimax risk provided by the adaptive estimator differs from the minimax rate of convergence by some factor. This factor for the majority of known results is $\ln(n)$.

In addition, we would like to note that the assertion of Theorem 4 is proved for arbitrary dimension.

2.2.2. *Adaptive estimation under the L_r losses.* We begin by defining the relevant functional classes.

DEFINITION 2. Let $\beta > 0$, $L > 0$ and $p \in [1, \infty)$ be fixed. A function $\ell : \mathbb{R} \to \mathbb{R}$ belongs to the Nikol'skii class $\mathbb{N}_p(\beta, L)$ if ℓ is m_β-times continuously differentiable and

$$\left(\int_{\mathbb{R}} |\ell^{(m)}(t)|^p \,\mathrm{d}t\right)^{1/p} \le L \qquad \forall m = 0, \dots, m_\beta,$$

$$\left(\int_{\mathbb{R}} |\ell^{(m_\beta)}(t+h) - \ell^{(m_\beta)}(t)|^p \,\mathrm{d}t\right)^{1/p} \le L h^{\beta - m_\beta} \qquad \forall h > 0.$$

It is also assumed that $\mathbb{N}_p(\beta, L) = \mathbb{H}(\beta, L)$ if $p = \infty$.

Here, the target of estimation is the entire function F under the assumption that $F \in \mathbb{F}_p(\mathbf{b}) := \bigcup_{\beta \le \mathbf{b}} \bigcup_{L > 0} \mathbb{F}_{2,p}(\beta, L)$, where

$$\mathbb{F}_{d,p}(\beta, L) = \{F : \mathbb{R}^d \to \mathbb{R} | F(z) = f(z^\top \theta), f \in \mathbb{N}_p(\beta, L), \theta \in \mathbb{S}^{d-1}\}.$$

Let us briefly discuss the applicability of Theorem 2 requiring $f \in \mathbb{F}(\beta_0, M)$. To this end, we assume that $\beta p > 1$. The latter assumption is standard for estimating functions with inhomogeneous smoothness [see, e.g., Donoho et al. (1995), Lepski, Mammen and Spokoiny (1997), Kerkyacharian, Lepski and Picard (2008)]. If $\beta p > 1$, the embedding $\mathbb{N}_p(\beta, L) \subset \mathbb{H}(\beta - 1/p, cL)$ with an absolute constant $c > 0$ guarantees that $f \in \mathbb{F}(\beta_0, M)$ with $\beta_0 = \beta - 1/p$ and $M = cL$.

THEOREM 5. *Let $\mathbf{b} > 0$ be fixed, and let Assumptions 4 and 5 hold. Then, for any $L > 0$, $p > 1$, $p^{-1} < \beta \le \mathbf{b}$ and $r \ge 1$,*

$$\sup_{F \in \mathbb{F}_{2,p}(\beta, L)} \mathcal{R}_r^{(n)}(\widehat{F}_{(\hat\theta, \hat h)}, F) \le \varkappa_3 \varphi_n(\beta, L, p),$$

where $\varkappa_3$ is independent of n, and

$$\varphi_n(\beta, L, p) = \begin{cases} L^{1/(2\beta+1)}\left(n^{-1}\ln(n)\right)^{\beta/(2\beta+1)}, & (2\beta+1)p > r, \\ L^{1/(2\beta+1)}\left(n^{-1}\ln(n)\right)^{\beta/(2\beta+1)} \ln^{1/r}(n), & (2\beta+1)p = r, \\ L^{(1/2-1/r)/(\beta-1/p+1/2)}\left(n^{-1}\ln(n)\right)^{(\beta-1/p+1/r)/(2\beta-2/p+1)}, & (2\beta+1)p < r. \end{cases}$$

Note that $\mathbb{F}_{2,p}(\beta, L) \supset \mathbb{N}_p(\beta, L)$. Indeed, the class $\mathbb{N}_p(\beta, L)$ can be viewed as a class of functions F satisfying $F(\cdot) = f(\theta^\top \cdot)$ with $\theta = (1, 0)^\top$. Then, the problem of estimating such (2-variate) functions reduces to the estimation of univariate regression functions.

There are at least two observations arising in view of the latter remark. First, the upper bound of Theorem 5 generalizes the results for the univariate regression [Donoho et al. (1995), Delyon and Juditsky (1996), Baraud (2002), Kerkyacharian and Picard (2004), Kulik and Raimondo (2009), Zhang, Wong and Zheng (2002)] in several directions. In particular, the majority of the papers treat the Gaussian errors or the errors having exponential moment. An exception is Baraud (2002), where some results are obtained under a very weak assumption on the noise (weaker than our Assumption 1). Nevertheless, these results are available only if $p = r = 2$.

Next, the rate of convergence for the latter problem, which can be found in Chesneau (2007), is also the lower bound for the minimax risk defined on $\mathbb{F}_{2,p}(\beta, L)$. With the proviso that $\beta p > 1$, the rate of convergence is given by

$$\phi_n(\beta, L, p) = \begin{cases} L^{1/(2\beta+1)} n^{-\beta/(2\beta+1)}, \\ \qquad (2\beta+1)p > r, \\ L^{1/(2\beta+1)} (n^{-1}\ln(n))^{\beta/(2\beta+1)}, \\ \qquad (2\beta+1)p = r, \\ L^{(1/2-1/r)/(\beta-1/p+1/2)} (n^{-1}\ln(n))^{(\beta-1/p+1/r)/(2\beta-2/p+1)}, \\ \qquad (2\beta+1)p < r. \end{cases}$$

The minimax rate of convergence in the case $(2\beta+1)p = r$ is not known, hence, the rate presented in the middle line above is only the lower asymptotic bound for the minimax risk.

Thus, the proposed estimator $\widehat{F}_{(\hat{\theta},\hat{h})}$ is adaptive whenever $(2\beta+1)p < r$. In the case $(2\beta+1)p \geq r$, we loose only a logarithmic factor with respect to the optimal rate and, as mentioned in the Introduction, the construction of an adaptive estimator over the collection $\{\mathbb{F}_{2,p}(\beta, L), \beta > 0, L > 0\}$ in this case remains an open problem. In view of the latter remark, we conjecture that the presented lower bound is correct and, therefore, the upper bound result has to be improved.

3. Proofs. We now list the quantities that are involved in the description of the selection rule that led to the adaptive estimator $\widehat{F}_{(\hat{\theta},\hat{h})}$.

3.1. *Important quantities.* Let $\tau = (\Omega^{-1}(4r+1)\ln(n))^{1/\omega}$. Set

$$c_1(n) = 730 \ln\big(16n^2 \underline{g}^{-1/2}[12Q + \sqrt{2}]\big) + 8r\ln(n) + 394,$$

$$c_2(n) = 730 \ln\big(16n^2 \tau \underline{g}^{-1/2}[12Q + \sqrt{2}]\big) + 8r\ln(n) + 394,$$

$$c_3(n) = 365 \ln\big(5n^2 Q \underline{g}^{-1/2}\big) + 8r\ln(n) + 197,$$

$$c_4(n) = 365 \ln\big(5n^2 \tau Q \underline{g}^{-1/2}\big) + 8r\ln(n) + 197.$$

With $\mathfrak{h}$ given in (1.5) and $\sigma^2 = \sup_{p\in\mathfrak{P}} \int_{\mathbb{R}} x^2 p(x)\,\mathrm{d}x$, we define

$$C_1(n) = 2\sqrt{2}\underline{g}^{-1/2}\|\mathcal{K}\|_\infty^2\sqrt{c_1(n)} + (8/3)c_1(n)\big(\ln(n)\big)^{-(2+\omega)/(2\omega)}\underline{g}^{-1}\|\mathcal{K}\|_\infty^2,$$
$$C_2(n) = 2\sqrt{2}(\sigma\vee 1)\underline{g}^{-1/2}\|\mathcal{K}\|_\infty^2\sqrt{c_2(n)} + (8/3)c_2(n)\big(\ln(n)\big)^{-1/2}\underline{g}^{-1}\|\mathcal{K}\|_\infty^2\big(\Omega^{-1}(4r+1)\big)^{1/\omega},$$
$$C_3(n) = 2\sqrt{2}\underline{g}^{-1/2}\|\mathcal{K}\|_\infty^2\sqrt{c_3(n)} + (8/3)\underline{g}^{-1}\|\mathcal{K}\|_\infty^2 c_3(n)\big(n\mathfrak{h}^2\big)^{-1/2},$$
$$C_4(n) = 2\sqrt{2}(\sigma\vee 1)\underline{g}^{-1/2}\|\mathcal{K}\|_\infty^2\sqrt{c_4(n)} + (8/3)\tau c_4(n)\big(n\mathfrak{h}^2\big)^{-1/2}\underline{g}^{-1}\|\mathcal{K}\|_\infty^2,$$
$$C_5(n) = \|\mathcal{K}\|_1^2 + \big(n\mathfrak{h}^2\big)^{-1/2}C_4(n) + 1/2.$$

In spite of the cumbersome expressions, it is easy to see that

$$\sup_{n\geq 3}\frac{C_i(n)}{\sqrt{\ln(n)}} =: C_i < \infty, \qquad i=1,2, \qquad \sup_{n\geq 3} C_5(n) =: C_5 < \infty. \tag{3.1}$$

3.2. *Proof of Theorem* 1. To begin with we present upper bounds for the approximation errors of the estimators involved (Lemma 1) and their stochastic errors (Lemma 2). Lemma 3 allows us to proceed without knowledge of M from Assumption 3. The proofs of the later two results are essentially based on Proposition 1 of Lepski (2013). The detailed proofs of these technical results are moved to the supplementary material [Lepski and Serdyukova (2014)].

3.2.1. *Auxiliary results*. For any $\theta,\nu\in\mathbb{S}^1$ and $h\in[2^{-1}h_{\min},1]$, denote

$$S_{(\theta,h)(\nu,h)}(t) = \det(E_{(\theta,h)(\nu,h)})\int K\big(E_{(\theta,h)(\nu,h)}(x-t)\big)F(x)\,\mathrm{d}x,$$
$$S_{(\theta,h)}(t) = \det(E_{(\theta,h)})\int K\big(E_{(\theta,h)}(x-t)\big)F(x)\,\mathrm{d}x.$$

For ease of notation, we write $h_f^* = h_{\mathcal{K},f}^*(t^\top\theta^*)$.

LEMMA 1. *Grant Assumption* 4. *Then, for any* $\nu\in\mathbb{S}^1$ *and any bandwidths* $\eta, h\in[2^{-1}h_{\min},1]$ *satisfying* $\eta\leq h\leq 2^{-1}h_f^*$, *one has*

$$\big|S_{(\theta^*,h)(\nu,h)}(t) - S_{(\nu,h)}(t)\big| \leq 2\big(h_f^*\big)^{-1/2}\|\mathcal{K}\|_\infty^2\sqrt{n^{-1}\ln(n)},$$
$$\big|S_{(\nu,h)}(t) - S_{(\nu,\eta)}(t)\big| \leq 2\big(h_f^*\big)^{-1/2}\|\mathcal{K}\|_\infty^2\sqrt{n^{-1}\ln(n)},$$
$$\big|S_{(\theta^*,h)}(t) - F(t)\big| \leq \big(h_f^*\big)^{-1/2}\|\mathcal{K}\|_\infty\sqrt{n^{-1}\ln(n)}.$$

Let $\mathcal{E}_{a,A}$ with $a\in(0,1]$, $A\geq 1$, be a set of 2×2 matrices satisfying

$$\big|\det(E)\big| \leq A, \qquad |E|_\infty \leq (\sqrt{2a})^{-1}\big|\det(E)\big|. \tag{3.2}$$

Here $|E|_\infty = \max_{i,j} |E_{i,j}|$ denotes the matrix sup norm. Set, $\forall E \in \mathcal{E}_{a,A}$,

$$J(x, E) = \sqrt{|\det(E)|}K\big(E(x-t)\big)g^{-1}(x), \qquad x \in \mathbb{R}^2$$

and consider the following random fields defined on $\mathcal{E}_{a,A}$:

$$\eta_{n,t}(E) = n^{-1/2}\sum_{i=1}^{n}\{J(X_i, E)F(X_i) - \mathbb{E}_X^{(n)}[J(X_i, E)F(X_i)]\},$$

$$\xi_{n,t}(E) = n^{-1/2}\sum_{i=1}^{n} J(X_i, E)\varepsilon_i.$$

Denote finally by $\mathcal{E}_*$ the set of matrices $\mathcal{E}_{a,A}$ with $a = 1/8$ and $A = h_{\min}^{-1}$. In what follows, we denote by $\|F\|_\infty = \sup_{x\in[-5/2,5/2]^2} |F(x)|$.

LEMMA 2. *Grant Assumptions 1–4. Then, for any $n \geq 3$ and any $r \geq 1$,*

$$\mathbb{P}_{X,\varepsilon}^{(n)}\Big\{\sup_{E\in\mathcal{E}_*}[|\eta_{n,t}(E)| + |\xi_{n,t}(E)|] \geq C_1(n)\|F\|_\infty + C_2(n)\Big\} \leq (8+\Upsilon)n^{-4r}.$$

The expressions for $C_1(n)$ and $C_2(n)$ are given in Section 3.1.

LEMMA 3. *Grant Assumptions 1–4. Then, for any $n \geq n_0 \vee n_1$,*

$$\sup_{\theta^*\in\mathbb{S}^1}\ \sup_{f\in\mathbb{F}(\beta_0,M)} \mathbb{P}_F^{(n)}\{\widehat{F}_\infty \notin [\|F\|_\infty, 3M + 4C_5(n)]\} \leq (8+\Upsilon)n^{-4r}.$$

The numbers n_0, n_1 are defined in (1.6) and $C_5(n)$ is defined in Section 3.1.

3.2.2. *Proof of Theorem 1.* In view of Jensen's inequality, an upper bound for $\mathcal{R}_{r,t}^{(n)}$, $r \geq 2$, will suffice to complete the proof.

Let $h^* \in \mathcal{H}_n$ be a bandwidth such that $2h^* \leq h_f^* < 4h^*$. Introduce the following random events:

$$\mathcal{A} = \{R_t^{(1)}(\theta^*, h^*) + R_t^{(2)}(h^*) = 0\}, \qquad \mathcal{B} = \{\widehat{F}_\infty \in [\|F\|_\infty, 3M + 4C_5(n)]\}$$

and let $\overline{\mathcal{A}}$ and $\overline{\mathcal{B}}$ denote the events complimentary to $\mathcal{A}$ and $\mathcal{B}$, respectively. The proof is split into three steps.

Risk computation under $\mathcal{A}\cap\mathcal{B}$. First, the following inclusion holds:

$$\mathcal{A} \subseteq \{\hat{h} \geq h^*\}. \tag{3.3}$$

Indeed, the definition of the couple $(\hat{\theta}, \hat{h})$ yields

$$\begin{aligned}1_{\mathcal{A}}\,\mathrm{TH}(h^*) &= 1_{\mathcal{A}}\{R_t^{(1)}(\theta^*, h^*) + R_t^{(2)}(h^*) + \mathrm{TH}(h^*)\}\\ &\geq 1_{\mathcal{A}}\{R_t^{(1)}(\hat{\theta}, \hat{h}) + R_t^{(2)}(\hat{h}) + \mathrm{TH}(\hat{h})\} \geq 1_{\mathcal{A}}\,\mathrm{TH}(\hat{h}).\end{aligned}$$

It remains to note that the mapping $\eta \mapsto \mathrm{TH}(\eta)$ is decreasing so inclusion (3.3) follows. Next, the triangle inequality yields

$$
\begin{aligned}
\big|\widehat{F}_{(\hat\theta,\hat h)}(t) - F(t)\big| &\le \big|\widehat{F}_{(\theta^*,h^*)}(t) - F(t)\big| + \big|\widehat{F}_{(\hat\theta,\hat h)}(t) - \widehat{F}_{(\hat\theta,h^*)}(t)\big| \\
&\quad + \big|\widehat{F}_{(\theta^*,h^*)(\hat\theta,h^*)}(t) - \widehat{F}_{(\hat\theta,h^*)}(t)\big| \\
&\quad + \big|\widehat{F}_{(\theta^*,h^*)(\hat\theta,h^*)}(t) - \widehat{F}_{(\theta^*,h^*)}(t)\big|.
\end{aligned} \tag{3.4}
$$

1^0. We have in view of (3.3) and the definition of $R_t^{(2)}$ that

$$
1_{\mathcal{A}}\big|\widehat{F}_{(\hat\theta,\hat h)}(t) - \widehat{F}_{(\hat\theta,h^*)}(t)\big| \le 1_{\mathcal{A}}\big[R_t^{(2)}(\hat h) + \mathrm{TH}(h^*)\big]. \tag{3.5}
$$

The definition of $R_t^{(1)}(\cdot,\cdot)$ implies that

$$
\begin{aligned}
1_{\mathcal{A}}\big|\widehat{F}_{(\theta^*,h^*)(\hat\theta,h^*)}(t) - \widehat{F}_{(\hat\theta,h^*)}(t)\big| &\le 1_{\mathcal{A}}\big[R_t^{(1)}(\theta^*,h^*) + \mathrm{TH}(h^*)\big] \\
&= 1_{\mathcal{A}}\,\mathrm{TH}(h^*).
\end{aligned} \tag{3.6}
$$

Note that $E_{(\theta,h)(\nu,h)} = \pm E_{(\nu,h)(\theta,h)}$, for any θ, ν and h. Hence,

$$
\widehat{F}_{(\theta^*,h^*)(\hat\theta,h^*)}(\cdot) \equiv \widehat{F}_{(\hat\theta,h^*)(\theta^*,h^*)}(\cdot),
$$

because $\mathcal{K}$ is symmetric. The latter observation, inclusion (3.3) and the definition of $R_t^{(1)}$ yield

$$
\begin{aligned}
1_{\mathcal{A}}\big|\widehat{F}_{(\theta^*,h^*)(\hat\theta,h^*)}(t) - \widehat{F}_{(\theta^*,h^*)}(t)\big| &= 1_{\mathcal{A}}\big|\widehat{F}_{(\hat\theta,h^*)(\theta^*,h^*)}(t) - \widehat{F}_{(\theta^*,h^*)}(t)\big| \\
&\le 1_{\mathcal{A}}\big[R_t^{(1)}(\hat\theta,\hat h) + \mathrm{TH}(h^*)\big].
\end{aligned} \tag{3.7}
$$

From (3.4), (3.5), (3.6) and (3.7), we obtain that

$$
\begin{aligned}
1_{\mathcal{A}}\big|\widehat{F}_{(\hat\theta,\hat h)}(t) - F(t)\big| &\le 1_{\mathcal{A}}\big[R_t^{(1)}(\hat\theta,\hat h) + R_t^{(2)}(\hat h)\big] + 3\,\mathrm{TH}(h^*) \\
&\quad + \big|\widehat{F}_{(\theta^*,h^*)}(t) - F(t)\big|.
\end{aligned}
$$

In addition, the definition of $(\hat\theta,\hat h)$ guarantees that

$$
\begin{aligned}
R_t^{(1)}(\hat\theta,\hat h) + R_t^{(2)}(\hat h) &\le R_t^{(1)}(\hat\theta,\hat h) + R_t^{(2)}(\hat h) + \mathrm{TH}(\hat h) \\
&\le R_t^{(1)}(\theta^*,h^*) + R_t^{(2)}(h^*) + \mathrm{TH}(h^*).
\end{aligned}
$$

We then obtain

$$
1_{\mathcal{A}}\big|\widehat{F}_{(\hat\theta,\hat h)}(t) - F(t)\big| \le 4\,\mathrm{TH}(h^*) + \big|\widehat{F}_{(\theta^*,h^*)}(t) - F(t)\big|. \tag{3.8}
$$

Note also that, for any $\eta \in \mathcal{H}_n$,

$$
\begin{aligned}
1_{\mathcal{B}}\,\mathrm{TH}(\eta) &\le 2\big[\|\mathcal{K}\|_\infty^2\sqrt{\ln(n)} + (3M + 4C_5)C_1(n) + C_2(n)\big](\eta n)^{-1/2} \\
&\le C_6\sqrt{(\eta n)^{-1}\ln(n)},
\end{aligned}
$$

where $C_6 = 2\|\mathcal{K}\|_\infty^2 + 2(3M + 4C_5)C_1 + 2C_2$ and C_1, C_2 and C_5 are defined in (3.1). Because $\mathrm{TH}(h^*) \le \mathrm{TH}(h_f^*/4)$, this bound and (3.8) yield

$$1_{\mathcal{A}\cap\mathcal{B}}\big|\widehat{F}_{(\hat\theta,\hat h)}(t) - F(t)\big| \le 8C_6\sqrt{\frac{\ln(n)}{nh_f^*}} + \big|\widehat{F}_{(\theta^*,h^*)}(t) - F(t)\big|. \tag{3.9}$$

2^0. Note that $E_{(\theta,h)(\nu,h)}, E_{(\theta,h)} \in \mathcal{E}_*$, for any $\theta, \nu \in \mathbb{S}^1$, $h \in [h_{\min}, 1]$. Set

$$\widehat{F}(E,t) = \frac{\det(E)}{n}\sum_{i=1}^n K\big(E(X_i - t)\big)g^{-1}(X_i)Y_i, \qquad E \in \mathcal{E}_*.$$

The following "approximation + stochastic part" decomposition of $\widehat{F}(E,t)$ will be useful in the sequel:

$$\begin{aligned}\widehat{F}(E,t) &= \det(E)\int K\big(E(x-t)\big)F(x)\,\mathrm{d}x \\ &\quad + \sqrt{n^{-1}\det(E)}\big[\eta_{n,t}(E) + \xi_{n,t}(E)\big],\end{aligned} \tag{3.10}$$

where $\eta_{n,t}(E)$ and $\xi_{n,t}(E)$ are defined before the statement of Lemma 2. Hence

$$\begin{aligned}&\big|\widehat{F}_{(\theta^*,h^*)}(t) - F(t)\big| \\ &\quad\le \big|S_{(\theta^*,h^*)}(t) - F(t)\big| + \sqrt{n^{-1}\det(E_{(\theta^*,h^*)})}\big|\eta_{n,t}(E_{(\theta^*,h^*)}) + \xi_{n,t}(E_{(\theta^*,h^*)})\big|.\end{aligned}$$

Taking into account that $\det(E_{(\theta^*,h^*)}) = (h^*)^{-1} \le 4(h_f^*)^{-1}$ in view of the definition of h^* and using the third assertion of Lemma 1, we obtain

$$\begin{aligned}&\big|\widehat{F}_{(\theta^*,h^*)}(t) - F(t)\big| \\ &\quad\le \big(nh_f^*\big)^{-1/2}\big[\sqrt{\ln(n)}\|\mathcal{K}\|_\infty + 2\big|\eta_{n,t}(E_{(\theta^*,h^*)}) + \xi_{n,t}(E_{(\theta^*,h^*)})\big|\big].\end{aligned}$$

Applying the Rosenthal inequality to $\eta_{n,t}(E_{(\theta^*,h^*)}) + \xi_{n,t}(E_{(\theta^*,h^*)})$ which is a sum of centered independent random variables, from (3.9) we obtain

$$\big\{\mathbb{E}_F^{(n)}\big|\widehat{F}_{(\hat\theta,\hat h)}(t) - F(t)\big|^r 1_{\mathcal{A}\cap\mathcal{B}}\big\}^{1/r} \le \tilde c_0\sqrt{\big(nh_f^*\big)^{-1}\ln(n)}, \tag{3.11}$$

where $\tilde c_0$ is independent of F and n.

Risk computation under $\overline{\mathcal{B}}$. Because $f \in \mathbb{F}(\beta_0, M)$ and $nh_{\min} > 1$, we have

$$\big|\widehat{F}_{(\hat\theta,\hat h)}(t) - F(t)\big| \le n\left\{M\big(1 + \underline{g}^{-1}\|K\|_\infty\big) + \underline{g}^{-1}\|K\|_\infty n^{-1}\sum_{i=1}|\varepsilon_i|\right\}.$$

Hence, in view of the Rosenthal inequality, we obtain

$$\big[\mathbb{E}_F^{(n)}\big|\widehat{F}_{(\hat\theta,\hat h)}(t) - F(t)\big|^{2r}\big]^{1/(2r)} \le \tilde c_1 n, \tag{3.12}$$

where $\tilde{c}_1$ is independent of F and n.

The use of the Cauchy–Schwarz inequality together with the statement of Lemma 3 leads to the following bound:

$$\{\mathbb{E}_F^{(n)}|\widehat{F}_{(\hat{\theta},\hat{h})}(t)-F(t)|^r 1_{\overline{\mathcal{B}}}\}^{1/r} \le \tilde{c}_1 n[\mathbb{P}_F^{(n)}(\overline{\mathcal{B}})]^{1/(2r)} \le \tilde{c}_1(8+\Upsilon)^{1/(2r)}n^{-1}. \tag{3.13}$$

Risk computation under $\overline{\mathcal{A}}\cap\mathcal{B}$. We note that

$$\mathbb{P}_F^{(n)}\{\overline{\mathcal{A}}\cap\mathcal{B}\} \le \mathbb{P}_F^{(n)}\{R_t^{(1)}(\theta^*,h^*)>0,\mathcal{B}\} + \mathbb{P}_F^{(n)}\{R_t^{(2)}(h^*)>0,\mathcal{B}\}.$$

1^0. First, let us bound from above $\mathbb{P}_F^{(n)}\{R_t^{(1)}(\theta^*,h^*)>0,\mathcal{B}\}$. We have

$$\{R_t^{(1)}(\theta^*,h^*)>0\} = \bigcup_{\eta\in\mathcal{H}_n:\,\eta\le h^*}\Big\{\sup_{\nu\in\mathbb{S}^1}|\widehat{F}_{(\theta^*,\eta)(\nu,\eta)}(t)-\widehat{F}_{(\nu,\eta)}(t)| > \mathrm{TH}(\eta)\Big\}$$

and, therefore,

$$\mathbb{P}_F^{(n)}\{R_t^{(1)}(\theta^*,h^*)>0,\mathcal{B}\} \le \sum_{k:\,2^{-1}h_{\min}\le 2^{-k}\le h^*}\mathbb{P}_F^{(n)}\Big\{\sup_{\nu\in\mathbb{S}^1}|\widehat{F}_{(\theta^*,2^{-k})(\nu,2^{-k})}(t)-\widehat{F}_{(\nu,2^{-k})}(t)| > \mathrm{TH}(2^{-k}),\mathcal{B}\Big\}. \tag{3.14}$$

Thus, denoting by $\varsigma_n=\sup_{E\in\mathcal{E}_*}[|\eta_{n,t}(E)|+|\xi_{n,t}(E)|]$ and using (3.10) together with the first assertion of Lemma 1, we obtain, for any $k:2^{-k}\le h^*$,

$$\begin{aligned}\sup_{\nu\in\mathbb{S}^1}&|\widehat{F}_{(\theta^*,2^{-k})(\nu,2^{-k})}(t)-\widehat{F}_{(\nu,2^{-k})}(t)| \\ &\le 2(h_f^*)^{-1/2}\|\mathcal{K}\|_\infty^2\sqrt{n^{-1}\ln(n)}+2\sqrt{2^k}n^{-1/2}\varsigma_n \\ &\le 2\|\mathcal{K}\|_\infty^2\sqrt{2^k n^{-1}\ln(n)}+2\sqrt{2^k n^{-1}}\varsigma_n.\end{aligned} \tag{3.15}$$

Here we have also used that $2^{-1}h_f^*\ge 2^{-k}$. Note also that

$$1_{\mathcal{B}}\,\mathrm{TH}(\eta) \ge 2\|\mathcal{K}\|_\infty^2\sqrt{\frac{\ln(n)}{\eta n}}+\frac{2}{\sqrt{\eta n}}(C_1(n)\|F\|_\infty+C_2(n)) \tag{3.16}$$

and, therefore, we obtain from (3.15), for any k satisfying $2^{-k}\le h^*$,

$$\begin{aligned}\mathbb{P}_F^{(n)}&\Big\{\sup_{\nu\in\mathbb{S}^1}|\widehat{F}_{(\theta^*,2^{-k})(\nu,2^{-k})}(t)-\widehat{F}_{(\nu,2^{-k})}(t)| > \mathrm{TH}(2^{-k}),\mathcal{B}\Big\} \\ &\le \mathbb{P}_{X,\varepsilon}^{(n)}\{\varsigma_n\ge\|F\|_\infty C_1(n)+C_2(n)\} \le (8+\Upsilon)n^{-4r},\end{aligned}$$

in view of Lemma 2. This bound and (3.14) yield

$$\mathbb{P}_F^{(n)}\{R_t^{(1)}(\theta^*, h^*) > 0, \mathcal{B}\} \le (8+\Upsilon)\log_2(n)n^{-4r}. \tag{3.17}$$

2^0. Now, let us bound from above $\mathbb{P}_F^{(n)}\{R_t^{(2)}(h^*) > 0, \mathcal{B}\}$. We have

$$\{R_t^{(2)}(h^*) > 0\} = \bigcup_{\eta\in\mathcal{H}_n\,:\,\eta\le h^*} \Big\{\sup_{\theta\in\mathbb{S}^1} |\widehat{F}_{(\theta,h^*)}(t) - \widehat{F}_{(\theta,\eta)}(t)| > \mathrm{TH}(2^{-k})\Big\}$$

and, hence,

$$\begin{aligned}\mathbb{P}_F^{(n)}&\{R_t^{(2)}(h^*) > 0, \mathcal{B}\}\\ &\le \sum_{k\,:\,2^{-1}h_{\min}\le 2^{-k}\le h^*} \mathbb{P}_F^{(n)}\Big\{\sup_{\theta\in\mathbb{S}^1} |\widehat{F}_{(\theta,h^*)}(t) - \widehat{F}_{(\theta,2^{-k})}(t)| \\ &\qquad\qquad > \mathrm{TH}(2^{-k}), \mathcal{B}\Big\}.\end{aligned} \tag{3.18}$$

Similar to estimate (3.15), with the use of (3.10) and the second assertion of Lemma 1, we obtain, for any k satisfying $2^{-k} \le h^*$, that

$$\sup_{\theta\in\mathbb{S}^1} |\widehat{F}_{(\theta,h^*)}(t) - \widehat{F}_{(\theta,2^{-k})}(t)| \le 2\|\mathcal{K}\|_\infty^2\sqrt{2^k n^{-1}\ln(n)} + 2\sqrt{2^k n^{-1}}\varsigma_n.$$

For any k satisfying $2^{-k} \le h^*$, bound (3.16) and Lemma 2 yield

$$\begin{aligned}\mathbb{P}_F^{(n)}&\Big\{\sup_{\theta\in\mathbb{S}^1} |\widehat{F}_{(\theta,h^*)}(t) - \widehat{F}_{(\theta,2^{-k})}(t)| > \mathrm{TH}(2^{-k}), \mathcal{B}\Big\}\\ &\le \mathbb{P}_{X,\varepsilon}^{(n)}\{\varsigma_n \ge \|F\|_\infty C_1(n) + C_2(n)\} \le (8+\Upsilon)n^{-4r}.\end{aligned}$$

Together with (3.18), the latter bound gives

$$\mathbb{P}_F^{(n)}\{R_t^{(2)}(h^*) > 0, \mathcal{B}\} \le (8+\Upsilon)\log_2(n)n^{-4r}. \tag{3.19}$$

Thus, we obtain from (3.17) and (3.19) that

$$\mathbb{P}_F^{(n)}(\overline{\mathcal{A}}\cap\mathcal{B}) \le 2(8+\Upsilon)\log_2(n)n^{-4r}.$$

Subsequently, this bound and (3.12) yield

$$\{\mathbb{E}_F^{(n)}|\widehat{F}_{(\hat\theta,\hat h)}(t) - F(t)|^r 1_{\overline{\mathcal{A}}\cap\mathcal{B}}\}^{1/r} \le \tilde c_1 n[\mathbb{P}_F^{(n)}(\overline{\mathcal{A}}\cap\mathcal{B})]^{1/(2r)} \le \tilde c_2 n^{-1/2}, \tag{3.20}$$

where $\tilde c_2$ is independent of F and n.

The assertion of the theorem follows from (3.11), (3.13) and (3.20).

3.3. *Proof of Theorem 3.* Using the standard computation of the bias of kernel estimators, under Assumptions 4 and 5, we get, for any $f \in \mathbb{H}(\beta, L)$ and any $z \in \mathbb{R}$,

$$\Delta_{\mathcal{K},f}(h,z) \le \frac{Lh^\beta 2^{-\beta}\|\mathcal{K}\|_\infty}{(1+\beta)m_\beta!} \le \|\mathcal{K}\|_\infty Lh^\beta.$$

Since the right-hand side of the latter inequality is independent of z, we have $\Delta^*_{\mathcal{K},f}(h,z) \le \|\mathcal{K}\|_\infty Lh^\beta$. This implies $h^*_{\mathcal{K},f}(z) \ge (L^{-2}n^{-1}\ln(n))^{1/(2\beta+1)}$, for any $z \in \mathbb{R}$, so the assertion of the theorem follows from Theorem 1.

3.4. *Proof of Theorem* 4. We start this section with an auxiliary result used in the proof of the second assertion of the theorem. It was established in Kerkyacharian, Lepski and Picard (2008), Corollary 2 of Proposition 5, and, for convenience, we formulate it as Lemma 4 below.

3.4.1. *Auxiliary result.* The result cited below concerns a lower bound for estimators of an arbitrary mapping in the framework of an abstract statistical model. We do not present it in full generality and below a version reduced to the estimation at a given point is provided.

Let $\mathcal{F}$ be a nonempty class of functions; and let $F:\mathbb{R}^d \to \mathbb{R}$ be an unknown function from model defined in (1.1)–(1.2). The aim is to estimate the functional $F(t)$, $t \in [-1/2, 1/2]^d$.

Introduce the following notation. For any given $F, G \in \mathcal{F}$, set

$$Z(F,G) = \prod_{i=1}^{n}\left[\frac{p(Y_i - F(X_i))}{p(Y_i - G(X_i))}\right].$$

LEMMA 4. *Assume that, for any sufficiently large* $n \ge 1$, *there exist a positive integer* N_n, $c > 1$ *and functions* $F_0, \dots, F_{N_n} \in \mathcal{F}$ *such that*

$$\left|F_j(t) - F_0(t)\right| = \lambda_n \qquad \forall j = 1, \dots, N_n, \tag{3.21}$$

$$\mathbb{E}_{F_0}^{(n)}\left(\frac{1}{N_n}\sum_{j=1}^{N_n} Z(F_j, F_0)\right)^2 \le c. \tag{3.22}$$

Then, for $r \ge 1$ *and any* $t \in [-1/2, 1/2]^d$,

$$\inf_{\widetilde{F}}\sup_{F\in\mathcal{F}}\left(\mathbb{E}_F^{(n)}\left|\widetilde{F}(t) - F(t)\right|^r\right)^{1/r} \ge \frac{1}{2}\left[1 - \sqrt{\frac{c-1}{c+3}}\right]\lambda_n.$$

3.4.2. *Proof of Theorem* 4. The proof is based on the construction of $F_0, \dots, F_{N_n}$ satisfying conditions (3.21)–(3.22) of Lemma 4.

1^0. First, we construct $F_0, \dots, F_{N_n}$ and verify (3.21). Let $w:\mathbb{R}\to\mathbb{R}$ be a function such that $\operatorname{supp}(w) \subset (-1/2, 1/2)$, $w \in \mathbb{H}(\beta, 1)$ and $w(0) \neq 0$. Set $h = (\mathfrak{a}(L^2 n)^{-1}\ln(n))^{1/(2\beta+1)}$, where $\mathfrak{a} > 0$ will be chosen later, and define

$$f(z) = Lh^{\beta} w(zh^{-1}), \qquad z \in \mathbb{R}. \tag{3.23}$$

For $b > 0$, put $N_n = n^b$ assuming without loss of generality that N_n is an integer. The value of b will be determined later in order to satisfy (3.22).

Let $\{\vartheta_j, j = 1, \dots, N_n\} \subset \mathbb{S}^{d-1}$ be defined as follows:

$$\vartheta_j = (\theta_j^{(1)}, \theta_j^{(2)}, 0, \dots, 0)^{\top}, \qquad \theta_j^{(1)} = \cos(j/N_n), \theta_j^{(2)} = \sin(j/N_n).$$

Finally, we set

$$F_0 \equiv 0 \quad \text{and} \quad F_j(x) = f\big(\vartheta_j^\top (x-t)\big), \qquad j = 1, \dots, N_n. \tag{3.24}$$

Obviously, f defined by (3.23) belongs to $\mathbb{H}(\beta, L)$, so all F_i are in the class $\mathcal{F} = \mathbb{F}_d(\beta, L)$. Moreover, for any $i = 1, \dots, N_n$,

$$\begin{aligned}\big|F_j(t) - F_0(t)\big| &= \big|w(0)\big| L^{1/(2\beta+1)}\big(\mathfrak{a} n^{-1}\ln(n)\big)^{\beta/(2\beta+1)} \\ &= \big|w(0)\big|\mathfrak{a}^{\beta/(2\beta+1)}\psi_n(\beta, L).\end{aligned}$$

We see that (3.21) holds with $\lambda_n = |w(0)|\mathfrak{a}^{\beta/(2\beta+1)}\psi_n(\beta, L)$.

2^0. It is noteworthy that

$$\begin{aligned}&\mathbb{E}_{F_0}^{(n)}\Bigg[\frac{1}{N_n}\sum_{j=1}^{N_n} Z(F_j, F_0)\Bigg]^2 \\ &\quad = \frac{1}{N_n^2}\sum_{j=1}^{N_n}\mathbb{E}_{F_0}^{(n)}\big[Z^2(F_j, F_0)\big] + \frac{1}{N_n^2}\sum_{\substack{j,k=1,\\ j\neq k}}^{N_n}\mathbb{E}_{F_0}^{(n)}\big[Z(F_j, F_0)Z(F_k, F_0)\big].\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{E}_{F_0}^{(n)}\big[Z^2(F_j, F_0)\big] &= \bigg\{\int_{\mathbb{R}^{d+1}} \frac{p^2(y - F_j(x))}{p(y)} g(x)\,\mathrm{d}x\,\mathrm{d}y\bigg\}^n, \\ \mathbb{E}_{F_0}^{(n)}\big[Z(F_j, F_0)Z(F_k, F_0)\big] &= \bigg\{\int_{\mathbb{R}^{d+1}} \frac{p(y - F_j(x))p(y - F_k(x))}{p(y)} g(x)\,\mathrm{d}x\,\mathrm{d}y\bigg\}^n.\end{aligned}$$

Because $\lim_{n\to\infty}\sup_{j=1,\dots,N_n}\|F_j\|_\infty = 0$, we have in view of Assumptions 6 and 7, for all n large enough,

$$\begin{aligned}&\int_{\mathbb{R}^{d+1}}\bigg[\frac{p^2(y - F_j(x))}{p(y)}\bigg] g(x)\,\mathrm{d}x\,\mathrm{d}y \\ &\qquad \le 1 + \mathfrak{Q}\int_{\mathbb{R}^d} F_j^2(x) g(x)\,\mathrm{d}x \\ &\qquad \le 1 + \mathfrak{Q}\mathfrak{g}\int_{\mathbb{R}^d} F_j^2(x)\big(1 + |x|_2^{\varpi}\big)^{-1}\,\mathrm{d}x,\end{aligned} \tag{3.25}$$

$$\begin{aligned}&\int_{\mathbb{R}^{d+1}}\bigg[\frac{p(y - F_j(x))p(y - F_k(x))}{p(y)}\bigg] g(x)\,\mathrm{d}x\,\mathrm{d}y \\ &\qquad \le 1 + \mathfrak{Q}\int_{\mathbb{R}^d}\big|F_j(x)F_k(x)\big| g(x)\,\mathrm{d}x \\ &\qquad \le 1 + \mathfrak{Q}\mathfrak{g}\int_{\mathbb{R}^d}\big|F_j(x)F_k(x)\big|\big(1 + |x|_2^{\varpi}\big)^{-1}\,\mathrm{d}x.\end{aligned} \tag{3.26}$$

Set $\theta_{j\perp} = (-\sin(j/N_n), \cos(j/N_n))^\top$ and $\vartheta_{j\perp} = (\theta_{j\perp}^\top, 0, \dots, 0)^\top \in \mathbb{S}^{d-1}$. Denote for all $j = 1, \dots, N_n$ by $\Theta_j^\top$ the orthogonal matrix $(\vartheta_j, \vartheta_{j\perp}, \mathbf{e}_3, \dots, \mathbf{e}_d)$, where $\mathbf{e}_s$, $s = 3, \dots, d$, are the canonical basis vectors in $\mathbb{R}^d$. Integration by substitution with $\Theta_j x = v$ gives

$$\begin{aligned}\int_{\mathbb{R}^d} F_j^2(x)\big(1 + |x|_2^{\varpi}\big)^{-1}\,\mathrm{d}x &= L^2 h^{2\beta} \int_{\mathbb{R}^d} w^2\big[h^{-1}\big(v_1 - \vartheta_j^\top t\big)\big]\big(1 + |v|_2^{\varpi}\big)^{-1}\,\mathrm{d}v \\ &\le C_\varpi L^2 \|w\|_2^2 h^{2\beta+1} = \mathfrak{a} C_\varpi \|w\|_2^2 n^{-1}\ln(n),\end{aligned}$$

where we have denoted $C_\varpi = \int_{\mathbb{R}^{d-1}}(1 + |\mathrm{v}|_2^{\varpi})^{-1}\,\mathrm{dv}$ and $\mathrm{v} = (v_2, \dots, v_d)^\top$. For n sufficiently large, this bound, together with (3.25), leads to

$$\sup_{j=1,\dots,N_n} \mathbb{E}_{F_0}^{(n)}\big[Z^2(F_j, F_0)\big] \le n^{\mathfrak{a}\mathfrak{Q}\mathfrak{g}C_\varpi \|w\|_2^2}. \tag{3.27}$$

For any $j \neq k$, set $\Theta_{j,k}^\top = (\vartheta_j, \vartheta_k, \mathbf{e}_3, \dots, \mathbf{e}_d)$. By changing of variables with $\Theta_{j,k}x = v$, we have

$$\begin{aligned}&\int_{\mathbb{R}^d} \big|F_j(x)F_k(x)\big|\big(1 + |x|_2^{\varpi}\big)^{-1}\,\mathrm{d}x \\ &\qquad = \big|\det(\Theta_{j,k})\big|^{-1} L^2 h^{2\beta} \int_{\mathbb{R}^d} \frac{|w[h^{-1}(v_1 - \vartheta_j^\top t)]w[h^{-1}(v_2 - \vartheta_k^\top t)]|}{1 + |\Theta_{j,k}^{-1} v|_2^{\varpi}}\,\mathrm{d}v \\ &\qquad \le \big|\det(\Theta_{j,k})\big|^{-1} c_\varpi L^2 h^{2\beta+2} \|w\|_1^2,\end{aligned}$$

where $c_\varpi = \int_{\mathbb{R}^{d-2}}(1 + |\mathrm{v}|_2^{\varpi})^{-1}\,\mathrm{dv}$ and $\mathrm{v} = (v_3, \dots, v_d)^\top$. Note that

$$\begin{aligned}\big|\det(\Theta_{j,k})\big| &= \big|\cos(j/N_n)\sin(k/N_n) - \cos(k/N_n)\sin(j/N_n)\big| \\ &= \big|\sin\big((k-j)/N_n\big)\big| \ge \sin(1/N_n) > (2N_n)^{-1}\end{aligned}$$

for sufficiently large n. We obtain

$$\int_{\mathbb{R}^d} \big|F_j(x)F_k(x)\big|\big(1 + |x|_2^{\varpi}\big)^{-1}\,\mathrm{d}x \le 2\mathfrak{a} c_\varpi \|w\|_1^2 n^{-1}\ln(n) N_n h.$$

Hence, choosing $b < 1/(2\beta + 1)$, we obtain, for all n large enough, that

$$\sup_{j\neq k;\, j,k=1,\dots,N_n} \int_{\mathbb{R}^d} \big|F_j(x)F_k(x)\big|\big(1 + |x|_2^{\varpi}\big)^{-1}\,\mathrm{d}x \le 2\mathfrak{a} c_\varpi \|w\|_1^2 n^{-1}. \tag{3.28}$$

We have in view of (3.26) and (3.28)

$$\sup_{j\neq k;\, j,k=1,\dots,N_n} \mathbb{E}_{F_0}^{(n)}\big\{Z(F_j, F_0)Z(F_k, F_0)\big\} \le e^{2\mathfrak{a}\mathfrak{Q}\mathfrak{g}c_\varpi \|w\|_1^2} \tag{3.29}$$

and, hence, (3.27) and (3.29) give

$$\mathbb{E}_{F_0}^{(n)}\left(\frac{1}{N_n}\sum_{j=1}^{N_n} Z(F_j, F_0)\right)^2 \le n^{-b+\mathfrak{a}\mathfrak{Q}\mathfrak{g}C_\varpi \|w\|_2^2} + e^{2\mathfrak{a}\mathfrak{Q}\mathfrak{g}c_\varpi \|w\|_1^2}.$$

Choosing $\mathfrak{a} = b(\mathfrak{Q}\mathfrak{g}C_{\varpi}\|w\|_2^2)^{-1}$, we see that (3.22) holds with the constant $c = 1 + e^{2\mathfrak{a}\mathfrak{Q}\mathfrak{g}c_{\varpi}\|w\|_1^2}$. Since c appearing in (3.22) is chosen independently of L, the assertion of the theorem follows from Lemma 4.

3.4.3. *Proof of Theorem* 5. In the proof we exploit the ideas from Lepski, Mammen and Spokoiny (1997). Moreover, our considerations are, to a great degree, based on the technical result of Lemma 5 below. Its proof is moved to the supplementary material [Lepski and Serdyukova (2014)].

LEMMA 5. *Grant Assumptions* 4 *and* 5. *Then, for any* $\mathfrak{p} > 1$, $0 < s \le \mathbf{b}$ *and* $\mathcal{Q} > 0$, *we have*

$$\sup_{g \in \mathbb{N}_{\mathfrak{p}}(s,\mathcal{Q})} \big\|\Delta^*_{\mathcal{K},g}(h,\cdot)\big\|_{\mathfrak{p}} \le 2\tau_{\mathfrak{p}}\mathcal{Q}h^s\|\mathcal{K}\|_\infty\big[2^{s\mathfrak{p}} - 1\big]^{-1/\mathfrak{p}} \qquad \forall h > 0.$$

Here $\tau_{\mathfrak{p}}$ *is a dependent only of* $\mathfrak{p}$ *constant from the* $(\mathfrak{p},\mathfrak{p})$-*strong maximal inequality*; *see Wheeden and Zygmund* (1977) *for more details.*

PROOF OF THEOREM 5. It is sufficient to prove the theorem in the case $r \ge p$ only. Indeed, let us recall that the risk $\mathcal{R}_r^{(n)}(\cdot,\cdot)$ is described by the L_r norm on $[-1/2, 1/2]$, therefore,

$$\mathcal{R}_r^{(n)}(\cdot,\cdot) \le \mathcal{R}_p^{(n)}(\cdot,\cdot), r \le p.$$

Hence, the case $r \le p$ can be reduced to the case $r = p$.

In view of Theorem 2, in order to obtain the assertion of the theorem, it suffices to bound from above $\|(nh^*_{\mathcal{K},f})^{-1}\ln(n)\|_{r/2}^{1/2}$.

Set $\Gamma_0 = \{y \in [-1/2, 1/2]: h^*_{\mathcal{K},f}(y) = 1\}$ and $\Gamma_k = \{y \in [-1/2, 1/2]: h^*_{\mathcal{K},f}(y) \in (2^{-k}, 2^{-k+1}] \cap [h_{\min}, 1]\}$, for $k = 1, 2, \ldots.$ In what follows, the integration over the empty set is supposed to be zero. We have

$$\left\|\frac{\ln(n)}{nh^*_{\mathcal{K},f}}\right\|_{r/2}^{r/2} = \sum_{k\ge1}\int_{\Gamma_k}\left(\frac{\ln(n)}{nh^*_{\mathcal{K},f}(y)}\right)^{r/2}\mathrm{d}y + \int_{\Gamma_0}\left(\frac{\ln(n)}{nh^*_{\mathcal{K},f}(y)}\right)^{r/2}\mathrm{d}y.$$

For simplicity of notation, we denote by $\bar{c}_i$, $i \ge 1$, constants independent of n, f and L.

The definition of Γ_0 implies

$$\int_{\Gamma_0}\left(\frac{\ln(n)}{nh^*_{\mathcal{K},f}(y)}\right)^{r/2}\mathrm{d}y \le \bar{c}_1\big[n^{-1}\ln(n)\big]^{r/2}. \tag{3.30}$$

We have in view of (2.2), for any $k \ge 1$,

$$\Delta^*_{\mathcal{K},f}\big(h^*_{\mathcal{K},f}(y), y\big) = \left[\frac{\|\mathcal{K}\|_\infty^2\ln(n)}{nh^*_{\mathcal{K},f}(y)}\right]^{1/2} \qquad \forall y \in \Gamma_k. \tag{3.31}$$

Let $0 \le q_k \le r$ be a sequence whose choice will be done later. We obtain from (3.31)

$$
\begin{aligned}
&\sum_{k\ge1}\int_{\Gamma_k}\left(\frac{\ln(n)}{nh^*_{\mathcal{K},f}(y)}\right)^{r/2}\mathrm{d}y\\
&\qquad\le \bar c_2\sum_{k\ge1}\left(\frac{\ln(n)}{n2^{-k}}\right)^{(r-q_k)/2}\int_{\Gamma_k}\left(\Delta^*_{\mathcal{K},f}(2^{1-k},y)\right)^{q_k}\mathrm{d}y \qquad (3.32)\\
&\qquad\le \bar c_2\sum_{k\ge1}\left(\frac{\ln(n)}{n2^{-k}}\right)^{(r-q_k)/2}\int\left(\Delta^*_{\mathcal{K},f}(2^{1-k},y)\right)^{q_k}\mathrm{d}y =: \Xi.
\end{aligned}
$$

To get the first inequality, we have used that $\Delta^*_{\mathcal{K},f}(\cdot,y)$ is a monotonically increasing function.

The computation of the quantity on the right-hand side of (3.32), including the choice of $(q_k, k\ge1)$, will be done differently in dependence on β, p and r.

1^0. *Case* $(2\beta+1)p>r$. Put $h^*=[L^{-2}n^{-1}\ln(n)]^{1/(2\beta+1)}$ and choose $q_k=p$ if $2^{-k}\le h^*$ and $q_k=0$ if $2^{-k}>h^*$. By applying Lemma 5 with $\mathfrak{p}=p$, $s=\beta$ and $\mathcal{Q}=L$, we get

$$
\begin{aligned}
\Xi &\le \bar c_3L^p\sum_{k:\,2^{-k}\le h^*}\left(\frac{\ln(n)}{n2^{-k}}\right)^{(r-p)/2}2^{-k\beta p}+\bar c_4\left(\frac{\ln(n)}{nh^*}\right)^{r/2}\\
&\le \bar c_5\left[L^p\left(n^{-1}\ln(n)\right)^{(r-p)/2}\sum_{k:\,2^{-k}\le h^*}2^{-k[\beta p-(r-p)/2]}+\left(\frac{\ln(n)}{nh^*}\right)^{r/2}\right]. \qquad (3.33)
\end{aligned}
$$

Because in the considered case $\beta p-\frac{r-p}{2}>0$, we obtain

$$
\Xi\le\bar c_6\left[L^p\left(n^{-1}\ln(n)\right)^{(r-p)/2}(h^*)^{\beta p-(r-p)/2}+\left(\frac{\ln(n)}{nh^*}\right)^{r/2}\right].
$$

It remains to note that h^* is chosen by balancing two terms on the right-hand side of the latter inequality. It yields

$$
\Xi\le\bar c_7L^{r/(2\beta+1)}\left(n^{-1}\ln(n)\right)^{(r\beta)/(2\beta+1)}. \tag{3.34}
$$

The argument in the case $(2\beta+1)p>r$ is completed with the use of Theorem 2, (3.30) and (3.34).

2^0. *Case* $(2\beta+1)p=r$. Put $h^*=1$ and choose $q_k=p$ for all $k\ge1$. Repeating the computations that led to (3.33), we get

$$
\Xi\le\bar c_8\ln(n)L^p\left(n^{-1}\ln(n)\right)^{(r-p)/2}. \tag{3.35}
$$

Here we have used that $\beta p-(r-p)/2=0$ and that the summation in (3.32) runs over k such that $2^{-k}\ge h_{\min}$, since otherwise $\Gamma_k=\varnothing$. It remains to note that the equality $(2\beta+1)p=r$ is equivalent to $p/r=1/(2\beta+1)$ and $(r-p)/2r=\beta/(2\beta+1)$. The assertion of the theorem in the case $(2\beta+1)p=r$ follows now from Theorem 2, (3.30) and (3.35).

3^0. *Case* $(2\beta+1)p<r$. Set $q_k=r$ if $2^{-k}\le h^*$ and $q_k=p$ if $2^{-k}>h^*$, where h^* will be chosen later. The following embedding holds [see page 62 in Besov, Il'in and Nikol'skiĭ (1979)]: $\mathbb{N}_p(\beta,L)\subseteq\mathbb{N}_r(\beta-1/p+1/r,c_6L)$. Thus, by applying Lemma 5 with $\mathfrak{p}=r$, $s=\beta-1/p+1/r$ and $\mathcal{Q}=c_6L$, we obtain

$$\begin{aligned}\Xi_1 &:= \sum_{k:2^{-k}\le h^*}\left(\frac{\ln(n)}{n2^{-k}}\right)^{(r-q_k)/2}\int(\Delta^*_{\mathcal{K},f}(2^{1-k},y))^{q_k}\,\mathrm{d}y\\ &= \sum_{k:2^{-k}\le h^*}\int(\Delta^*_{\mathcal{K},f}(2^{1-k},y))^{r}\,\mathrm{d}y\le\bar{c}_9L^r(h^*)^{\beta r-r/p+1}.\end{aligned}\tag{3.36}$$

By applying the same lemma with $\mathfrak{p}=r$, $s=\beta$ and $\mathcal{Q}=L$, we get

$$\begin{aligned}\Xi_2 &:= \sum_{k:2^{-k}>h^*}\left(\frac{\ln(n)}{n2^{-k}}\right)^{(r-q_k)/2}\int(\Delta^*_{\mathcal{K},f}(2^{1-k},y))^{q_k}\,\mathrm{d}y\\ &= \bar{c}_{10}L^p(n^{-1}\ln(n))^{(r-p)/2}\sum_{k:2^{-k}>h^*}2^{-k[\beta p-(r-p)/2]}\\ &\le \bar{c}_{11}L^p(n^{-1}\ln(n))^{(r-p)/2}(h^*)^{\beta p-(r-p)/2}.\end{aligned}\tag{3.37}$$

Here we have used that $\beta p-(r-p)/2<0$. In view of (3.36) and (3.37), we choose h^* from the equality $L^r(h^*)^{\beta r-r/p+1}=L^p(n^{-1}\ln(n))^{(r-p)/2}(h^*)^{\beta p-(r-p)/2}$, so that $h^*=(L^{-2}n^{-1}\ln(n))^{1/(2\beta-2/p+1)}$. Finally, we obtain that

$$\Xi\le\bar{c}_{12}L^{(r(1/2-1/r))/(\beta-1/p+1/2)}(n^{-1}\ln(n))^{(r(\beta-1/p+1/r))/(2\beta-2/p+1)}.\tag{3.38}$$

The assertion of the theorem in the case $(2\beta+1)p<r$ follows now from Theorem 2, (3.30) and (3.38). □

4. Unknown design density. In this section we briefly comment on the case when the design density g is unknown. We provide changes to be done in the selection rule and in the presentation of the main result established in Theorems 1 and 2. We also explain basic ideas related to the proofs of the new results.

In the context of the unknown design density, it is standard practice to use a plug-in estimator. This idea goes back to the Nadaraya–Watson estimator and the problem considered in the paper is not an exception.

Suppose that an additional independent of an $\{X_i\}_{i=1}^n$ sample, say, $\{\widetilde{X}_i\}_{i=1}^n$, is available. Alternatively, one can split the sample into two nonoverlapping parts. Let $\widetilde{\mathbb{P}}_g^{(n)}$ stand for the probability law of $\{\widetilde{X}_i\}_{i=1}^n$. We reinforce Assumption 2 by the following condition: $g\in\mathbb{G}\subset\{\ell:\mathbb{R}^2\to\mathbb{R}:\|\ell\|_\infty\le\bar{g}\}$, where $\bar{g}<\infty$ and $\|\cdot\|_\infty$

denotes the supremum norm on $\mathbb{R}^2$. It is noteworthy that the constants $\underline{g}$ and $\bar{g}$ are both unknown.

Assume that based on $\{\widetilde{X}_i\}_{i=1}^n$ we can construct an estimator of the design density g, say, $\hat{g}$, having the following property. There exists a positive sequence $a_n \downarrow 0$ as $n \to \infty$ such that, for all sufficiently large n,

$$\sup_{g \in \mathbb{G}} \widetilde{\mathbb{P}}_g^{(n)} \{ \|\hat{g} - g\|_\infty \ge a_n \} \le n^{-4r}. \tag{4.1}$$

Denote by $\Delta = [-3, 3]^2$ the interval from Assumption 2 and introduce $\underline{\hat{g}} = \inf_{x \in \Delta} \hat{g}(x)$ and $\underline{\hat{g}}_n = \underline{\hat{g}} \vee b_n$, where b_n tends to zero rather slowly. Theoretically, b_n can be chosen arbitrary, but a compromise allowing to keep our results under reasonable sample size is $b_n = \ln^{-3}(n)$.

Changes in the selection rule (2.5) *and in the oracle inequalities.*

1^0. In the definition of estimators $\widehat{F}_{(\theta,h)}(\cdot)$ and $\widehat{F}_{(\theta,h)(\nu,h)}(\cdot)$, the unknown now values $g(X_i)$, $i = 1, \dots, n$, should be replaced by their truncated estimators $\hat{g}(X_i) \vee b_n$, $i = 1, \dots, n$.

2^0. In the definition of all constants presented in Section 3.1, the unknown now value $\underline{g}^{-1}$ has to be replaced by $8\underline{\hat{g}}^{-2}\|\hat{g}\|_\infty$.

3^0. Let $\widehat{\mathrm{TH}}(\cdot)$ be obtained from $\mathrm{TH}(\cdot)$ by the replacement indicated in 2^0. Then one should use $\mathrm{TH}^{(\mathrm{new})}(\eta) = \widehat{\mathrm{TH}}(\eta) + 2a_n \underline{\hat{g}}^{-1} \|\mathcal{K}\|_1^2 \widehat{F}_\infty$ in (2.5).

4^0. The right-hand sides of the local and global oracle inequalities established in Theorems 1 and 2 will additionally contain a term ca_n, where c is a numerical constant independent of F, g and the sample size n.

Sketch of the proof of the new version of Theorem 1.

1. Denote by $\mathcal{C}$ the event $\{\|\hat{g} - g\|_\infty \le a_n\}$. Similar to the proof given in the step *Risk computation under* $\overline{\mathcal{B}}$ of Theorem 1, the computations under the event $\overline{\mathcal{C}}$ lead to the same reminder term in the oracle inequality.

2. All computations under the event $\mathcal{C}$ are done conditionally with respect to $\{\widetilde{X}_i\}_{i=1}^n$. It allows us to treat the estimator $b_n \vee \hat{g}(\cdot)$ as nonrandom.

2.1. Analyzing the proof of "probabilistic" Lemmas 2 and 3, we see that the results remain valid if we replace the function g in the denominator of all expressions by an arbitrary function bounded from below and above on Δ. Thus, the use of $b_n \vee \hat{g}(\cdot)$ in place of $g(\cdot)$ under $\mathcal{C}$ is eligible. That leads to the similar assertion where $\underline{g}^{-1/2}$ is substituted with $\underline{\hat{g}}_n^{-1} \|g\|_\infty^{-1/2}$ in all the constants involved. The latter quantity can be bounded under $\mathcal{C}$ by $2\underline{\hat{g}}^{-1} \|g\|_\infty^{-1/2}$, for n large enough.

However, such quantities cannot be used in the definition of the threshold directly, because they incorporate the unknown $\|g\|_\infty$. Nevertheless, under the event $\mathcal{C}$, $\|g\|_\infty$ can be bounded by $2\|\hat{g}\|_\infty$. Moreover, we remark that all the quantities listed in Section 3.1 are increasing functions of $\underline{g}^{-1}$ so we can replace $\underline{g}^{-1}$ by the upper bound $8\underline{\hat{g}}^{-2}\|\hat{g}\|_\infty$ available under $\mathcal{C}$. It explains 2^0.

2.2. The replacement of g by $b_n \vee \hat{g}$ leads to an additional "approximation error" bounded from above by $A(g) := \sup_{x\in\Delta} |g(x)[b_n \vee \hat{g}(x)]^{-1} - 1|$. This quantity should be added to $\widehat{\mathrm{TH}}(\cdot)$ in order to preserve the proof of Theorem 1. Since $A(g)$ depends on g, one should instead use its upper bound $a_n \underline{\hat{g}}^{-1}$ which is available under $\mathcal{C}$. It gives 3^0 for n large enough.

2.3. The use of $\mathrm{TH}^{(\mathrm{new})}(\cdot)$ in place of $\mathrm{TH}(\cdot)$ leads to an additional term $2a_n \underline{\hat{g}}^{-1} \|\mathcal{K}\|_1^2 (3M + 4C_5^{(\mathrm{new})})$ in (3.9). It explains 4^0 and completes the sketch of the proof of Theorem 1. Since the global oracle inequality is obtained by the integration of the local one over bounded interval of $\mathbb{R}^2$, the assertion of Theorem 2 remains the same up to the term ca_n.

The additional assumption about g and an example of an estimator obeying (4.1). If we suppose that $\mathbb{G} \subseteq \mathbb{H}_2(\gamma, R)$, where $\mathbb{H}_2(\gamma, R)$ is an isotropic Hölder class of two-variate functions, then $a_n = [n^{-1}\ln(n)]^{\gamma/(2(\gamma+1))}$, and this rate is attainable by a kernel estimator with properly chosen kernel and bandwidth. This yields together with 4^0 that if $\gamma > 2\mathbf{b}$, the adaptive results established in Theorems 3 and 5 remain unchangeable.

Another possibility is to suppose that $\mathbb{G}$ is a parametric family of densities. In this case, ca_n can be viewed as a reminder term.

Acknowledgments. The authors wish to thank an Associate Editor and anonymous referees for their comments that substantially improved the paper.

SUPPLEMENTARY MATERIAL

Proofs of lemmas for "Adaptive estimation under single-index constraint in a regression model" (DOI: 10.1214/13-AOS1152SUPP; .pdf). We provide detailed proofs of the auxiliary results (Lemmas 1–3 and 5) for the paper.

REFERENCES

BARAUD, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Stat.* **6** 127–146 (electronic). MR1918295

BESOV, O. V., IL'IN, V. P. and NIKOL'SKIĬ, S. M. (1979). *Integral Representations of Functions and Imbedding Theorems. Vol. II*. Wiley, New York. MR0521808

BROWN, L. D. and LOW, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24** 2524–2535. MR1425965

CHESNEAU, C. (2007). Regression with random design: A minimax study. *Statist. Probab. Lett.* **77** 40–53. MR2339017

DELYON, B. and JUDITSKY, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** 215–228. MR1400080

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: Asymptopia? *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 301–369. MR1323344

GAÏFFAS, S. (2007). On pointwise adaptive curve estimation based on inhomogeneous data. *ESAIM Probab. Stat.* **11** 344–364 (electronic). MR2339297

GAÏFFAS, S. and LECUÉ, G. (2007). Optimal rates and adaptation in the single-index model using aggregation. *Electron. J. Stat.* **1** 538–573. MR2369025

GOLDENSHLUGER, A. and LEPSKI, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli* **14** 1150–1190. MR2543590

GOLDENSHLUGER, A. and LEPSKI, O. (2009). Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields* **143** 41–71. MR2449122

GOLUBEV, G. K. (1992). Asymptotically minimax estimation of a regression function in an additive model. (In Russian) *Problemy Peredachi Informatsii* **28** 3–15; translation in *Probl. Inf. Transm.* **28** (1992) 101–112. MR1178413

HOROWITZ, J. L. (1998). *Semiparametric Methods in Econometrics*. *Lecture Notes in Statistics* **131**. Springer, New York. MR1624936

JENNRICH, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* **40** 633–643. MR0238419

KERKYACHARIAN, G., LEPSKI, O. and PICARD, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields* **121** 137–170. MR1863916

KERKYACHARIAN, G., LEPSKI, O. and PICARD, D. (2008). Nonlinear estimation in anisotropic multi-index denoising. Sparse case. *Theory Probab. Appl.* **52** 58–77. MR2354574

KERKYACHARIAN, G. and PICARD, D. (2004). Regression in random design and warped wavelets. *Bernoulli* **10** 1053–1105. MR2108043

KULIK, R. and RAIMONDO, M. (2009). Wavelet regression in random design with heteroscedastic dependent errors. *Ann. Statist.* **37** 3396–3430. MR2549564

LEPSKI, O. (2013). Upper functions for positive random functionals. I. General setting and Gaussian random functions. *Math. Methods Statist.* **22** 1–27. MR3040410

LEPSKI, O. V., MAMMEN, E. and SPOKOINY, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25** 929–947. MR1447734

LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 459–466. MR1091202

LEPSKI, O. and SERDYUKOVA, N. (2014). Supplement to "Adaptive estimation under single-index constraint in a regression model." DOI:10.1214/13-AOS1152SUPP.

MADDALA, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. *Econometric Society Monographs in Quantitative Economics* **3**. Cambridge Univ. Press, Cambridge. MR0799154

WHEEDEN, R. L. and ZYGMUND, A. (1977). *Measure and Integral*: *An Introduction to Real Analysis*. *Pure and Applied Mathematics* **43**. Dekker, New York. MR0492146

ZHANG, S., WONG, M.-Y. and ZHENG, Z. (2002). Wavelet threshold estimation of a regression function with random design. *J. Multivariate Anal.* **80** 256–284. MR1889776

LABORATOIRE D'ANALYSE
TOPOLOGIE, PROBABILITÉS UMR 7353
AIX-MARSEILLE UNIVERSITÉ
39, RUE F. JOLIOT CURIE
13453 MARSEILLE
FRANCE
E-MAIL: Oleg.Lepski@cmi.univ-mrs.fr

DEPARTAMENTO DE ESTADÍSTICA
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CONCEPCIÓN
AVDA. ESTEBAN ITURRA S/N—BARRIO UNIVERSITARIO
CONCEPCIÓN, REGIÓN VIII
CHILE
E-MAIL: Nora.Serdyukova@gmail.com