Probab. Theory Relat. Fields (2009) 143:41–71
DOI 10.1007/s00440-007-0119-5

Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities

Alexander Goldenshluger · Oleg Lepski

Received: 19 April 2007 / Revised: 31 October 2007 / Published online: 30 November 2007

Abstract In this paper we study the problem of adaptive estimation of a multivariate function satisfying some structural assumption. We propose a novel estimation procedure that adapts simultaneously to unknown structure and smoothness of the underlying function. The problem of structural adaptation is stated as the problem of selection from a given collection of estimators. We develop a general selection rule and establish for it global oracle inequalities under arbitrary $\mathbb{L}_p$-losses. These results are applied for adaptive estimation in the additive multi-index model.

Keywords Structural adaptation · Oracle inequalities · Minimax risk · Adaptive estimation · Optimal rates of convergence

Mathematics Subject Classification (2000) 62G05 · 62G20

1 Introduction

1.1 Motivation

In this paper we study the problem of minimax adaptive estimation of an unknown function $F : \mathbb{R}^d \to \mathbb{R}$ in the multidimensional Gaussian white noise model

Supported by the ISF grant No. 389/07.

A. Goldenshluger (✉)
Department of Statistics, University of Haifa, 31905 Haifa, Israel
e-mail: goldensh@stat.haifa.ac.il

O. Lepski
Laboratoire d'Analyse, Topologie et Probabilités UMR CNRS 6632,
Université de Provence, 39, rue F.Joliot Curie, 13453 Marseille, France
e-mail: lepski@cmi.univ-mrs.fr

$$Y(dt) = F(t)\mathrm{d}t + \varepsilon W(\mathrm{d}t), \quad t = (t_1, \ldots, t_d) \in \mathcal{D}, \tag{1}$$

where $\mathcal{D} \supset [-1/2, 1/2]^d$ is an open interval in $\mathbb{R}^d$, W is the standard Brownian sheet in $\mathbb{R}^d$ and $0 < \varepsilon < 1$ is the noise level. Our goal is to estimate the function F on the set $\mathcal{D}_0 := [-1/2, 1/2]^d$ from the observation $\{Y(t),\ t \in \mathcal{D}\}$. We consider the observation set $\mathcal{D}$ which is larger than $\mathcal{D}_0$ in order to avoid discussion of boundary effects. We would like to emphasize that such assumptions are rather common in multivariate models (see, e.g., [5,12]).

To measure performance of estimators, we will use the risk function determined by the $\mathbb{L}_p$-norm $\|\cdot\|_p$, $1 \le p \le \infty$ on $\mathcal{D}_0$: for $F : \mathbb{R}^d \to \mathbb{R}$, $0 < \varepsilon < 1$, and for an arbitrary estimator $\tilde{F}$ based on the observation $\{Y(t),\ t \in \mathcal{D}\}$ we consider the risk

$$\mathcal{R}_p[\tilde{F}; F] = \mathbb{E}_F ||\tilde{F} - F||_p.$$

Here and in what follows $\mathbb{E}_F$ denotes the expectation with respect to the distribution $\mathbb{P}_F$ of the observation $\{Y(t),\ t \in \mathcal{D}\}$ satisfying (1).

We will suppose that $F \in \mathcal{G}_s$, where $\{\mathcal{G}_s, s \in \mathcal{S}\}$ is a collection of functional classes indexed by $s \in \mathcal{S}$. The choice of this collection is a delicate problem, and below we discuss it in detail.

For a given class $\mathcal{G}_s$ we define the maximal risk

$$\mathcal{R}_p[\tilde{F}; \mathcal{G}_s] = \sup_{g \in \mathcal{G}_s} \mathcal{R}_p[\tilde{F}; F], \tag{2}$$

and study asymptotics (as the noise level ε tends to 0) of the minimax risk

$$\inf_{\tilde{F}} \mathcal{R}_p[\tilde{F}; \mathcal{G}_s]$$

where $\inf_{\tilde{F}}$ denotes the infimum over all estimators of F. At this stage, we suppose that parameter s is known, and therefore the functional class $\mathcal{G}_s$ is fixed. In other words, we are interested in minimax estimation of F. The important remark in this context is that the minimax rate of convergence $\phi_\varepsilon(s)$ on $\mathcal{G}_s$ (the rate which satisfies $\phi_\varepsilon(s) \asymp \inf_{\tilde{F}} \mathcal{R}_p[\tilde{F}; \mathcal{G}_s]$) as well as the estimator attaining this rate (called the rate optimal estimator in asymptotic minimax sense) depend on the parameter s. This dependence restricts application of the minimax approach in practice. Therefore, our main goal is to construct an estimator which is independent of s and achieves the minimax rate $\phi_\varepsilon(s)$ simultaneously for all $s \in \mathcal{S}$. Such an estimator, if it exists, is called optimally adaptive on $\mathcal{S}$.

Let us discuss now the choice of the collection $\{\mathcal{G}_s, s \in \mathcal{S}\}$. It is well known that the main difficulty in estimation of multivariate functions is the curse of dimensionality: the best attainable rate of convergence of estimators becomes very slow, as the dimensionality grows. To illustrate this effect, suppose, for example, that the underlying function F belongs to $\mathcal{G}_s = \mathbb{H}_d(\alpha, L), s = (\alpha, L), \alpha > 0, L > 0$, where $\mathbb{H}_d(\alpha, L)$ is an isotropic Hölder ball of functions. We give the exact definition of this functional class later. Here we only mention that $\mathbb{H}_d(\alpha, L)$ consists of functions g with bounded

partial derivatives of order $\leq \lfloor \alpha \rfloor$ and such that, for all $x, y \in \mathcal{D}$,

$$\left| g(y) - P_g(x, y-x) \right| \leq L|x-y|^\alpha,$$

where $P_g(x, y-x)$ is the Taylor polynomial of order $\leq \lfloor \alpha \rfloor$ obtained by expansion of g around the point x, and $|\cdot|$ is the Euclidean norm in $\mathbb{R}^d$. The parameter α characterizes the isotropic (i.e., the same in each direction) smoothness of the function g.

If we use the risk (2), uniformly on $\mathbb{H}_d(\alpha, L)$ the rate of convergence of estimators cannot be asymptotically better than

$$\psi_{\varepsilon,d}(\alpha) = \begin{cases} \varepsilon^{2\alpha/(2\alpha+d)}, & p \in [1, \infty) \\ (\varepsilon\sqrt{\ln \varepsilon^{-1}})^{2\alpha/(2\alpha+d)} & p = \infty. \end{cases} \tag{3}$$

(cf. [3,16,26,29,30]). This is the minimax rate on $\mathbb{H}_d(\alpha, L)$: in fact, it can be achieved by a kernel estimator with properly chosen bandwidth and kernel. More general results on asymptotics of the minimax risks in estimation of multivariate functions can be found in Kerkyacharian et al. [21] and Bertin [3]. It is clear that if α is fixed then even for moderate d the estimation accuracy is very poor unless the noise level ε is unreasonably small.

This problem arises because the d-dimensional Hölder ball $\mathbb{H}_d(\alpha, L)$ is too massive. A way to overcome the curse of dimensionality is to consider models with smaller functional classes $\mathcal{G}_s$. Clearly, if the class of candidate functions F is smaller, the rate of convergence of estimators is faster. Note that the massiveness of a functional class can be described in terms of restrictions on its metric entropy. There are nevertheless several ways to do it.

1.2 Structural adaptation

In this paper we will follow the modeling strategy which consists in imposing additional structural assumptions on the function to be estimated. This approach was pioneered by Stone [31] who discussed the trade-off between flexibility and dimensionality of nonparametric models and formulated *the heuristic dimensionality reduction principle*. The main idea is to assume that even though F is a d-dimensional function, it has a simple structure such that F is effectively m-dimensional with $m < d$. The standard examples of structural nonparametric models are the following.

(i) *Single-index model.* Let e be a direction vector in $\mathbb{R}^d$, and assume that $F(x) = f(e^T x)$ for some unknown univariate function f.

(ii) *Additive model.* Assume that $F(x) = \sum_{i=1}^d f_i(x_i)$, where f_i are unknown univariate functions.

(iii) *Projection pursuit regression.* Let $e_1, \ldots, e_d$ be direction vectors in $\mathbb{R}^d$, and assume that $F(x) = \sum_{i=1}^d f_i(e^T x)$, where f_i are as in (ii).

(iv) *Multi-index model.* Let $e_1, \ldots, e_m$, $m < d$ are direction vectors and assume that $F(x) = f(e_1^T x, \ldots, e_m^T x)$ for some unknown m-dimensional function f.

In the first three examples the function F is effectively one-dimensional, while in the fourth one it is m-dimensional. The heuristic dimensionality reduction principle

by Stone [31] suggests that the optimal rate of convergence attainable in structural nonparametric models should correspond to the effective dimensionality of F.

Let us make the following important remark.

The estimation problem in the models of types (i), (iii) and (iv) can be viewed as the problem of adaptation to unknown structure (structural adaptation). Indeed, if the direction vectors are given then, after a linear transformation, the problem is reduced either to the estimation problem in the additive model (cases (i) and (iii)) or to the estimation of an m-variate function. This explains the form of minimax rate of convergence. The main problem however is to find an estimator that adjusts automatically to unknown direction vectors. For this purpose one can consider a family of estimators parameterized by direction vectors and to select an estimator from this family. Our approach to the problem of structural adaptation is based on selection of estimators from large parameterized collections.

1.3 $\mathbb{L}_p$-norm oracle inequalities

Suppose that we are given a collection of estimators $\{\hat{F}_\theta, \theta \in \Theta \subset \mathbb{R}^m\}$ based on the observation $\{Y(t),\ t \in \mathcal{D}\}$. In the previous examples parameter θ could be, for instance, the unknown matrix $E = (e_1, \dots, e_d)$ of the direction vectors, $\theta = E$, and $\hat{F}_E$ could be a kernel estimator constructed under hypothesis that E and the smoothness of the functional components are known (a kernel estimator with fixed bandwidth).

With each estimator $\hat{F}_\theta$ and unknown function F we associate the risk $\mathcal{R}_p[\hat{F}_\theta; F]$. The problem is to construct an estimator, say, $\hat{F}_*$ such that for all F obeying given smoothness conditions one has

$$\mathcal{R}_p[\hat{F}_*; F] \le \mathcal{L} \inf_{\theta \in \Theta} \mathcal{R}_p[\hat{F}_\theta; F], \tag{4}$$

where $\mathcal{L}$ is an absolute constant independent of F and ε. Following the modern statistical terminology we will call the inequality (4) *the $\mathbb{L}_p$-norm oracle inequality*.

Returning to our example with $\theta = E$ we observe that being established, the *L_p-norm oracle inequality* leads immediately to the minimax result for any given value of smoothness parameter (α, L). In particular, we can state that the estimator $\hat{F}_*$ is adaptive with respect to unknown structure.

It is important to realize that the same strategy allows to avoid dependence of estimation procedures on smoothness. To this end it is sufficient

- to consider $\theta = (E, \alpha, L)$ that leads to the collection of kernels estimators with the non-fixed bandwidth and orientation;
- to propose an estimator $\hat{F}_*$ based on this collection;
- to establish for this estimator *the L_p-norm oracle inequality* (4) for any $F \in L_2(\mathcal{D})$ (or on a bit smaller functional space).

Being realized, this program leads to an estimator that is adaptive with respect to unknown structure and unknown smoothness properties. It is important to note that such methods allow to estimate multivariate functions with high accuracy without sacrificing flexibility of modeling.

1.4 Objective of the paper

In this paper we state the problem of structural adaptation as the problem of selection from a given collection of estimators. For a collection of linear estimators satisfying rather mild assumptions we develop a novel general selection rule and establish for it the $\mathbb{L}_p$-norm oracle inequality (4). Similar ideas were used in Lepski and Levit [22], Kerkyacharian et al. [21], Juditsky et al. [20] for *pointwise* adaptation. However we emphasize that our work is the first where the $\mathbb{L}_p$-norm oracle inequality is derived directly without applying pointwise estimation results. It is precisely this fact that allows to obtain adaptive results for arbitrary $\mathbb{L}_p$-losses. The selection rule as well as the $\mathbb{L}_p$-norm oracle inequality are not related to any specific model, and they are applicable in a variety of setups where linear estimators are appropriate. We apply these general results to a specific collection of kernel estimators corresponding to a general structural model that we call *the additive multi-index model*.

The additive multi-index model includes models (i)–(iv) as special cases. This generalization is dictated by the following reasons. On the one hand, structural assumptions allow to improve the quality of statistical analysis. On the other hand, they can lead to inadequate modeling. Thus we seek a general structural model that still allows to gain in estimation accuracy. To our knowledge the additive multi-index model did not previously appear in the statistical literature. For this model we propose an estimation procedure that adapts simultaneously to unknown structure and smoothness of the underlying function. The adaptive results are obtained for $\mathbb{L}_\infty$-losses and for a scale of the Hölder type functional classes.

1.5 Connection to other works

Structural models. The heuristic dimensionality reduction principle holds for the additive model (ii) [31], and for the projection pursuit regression model (iii) that includes as a particular case the single-index model (i) (see [5,9]). In particular, it was shown that in these models the asymptotics of the risk (2) with $p = 2$ and with $\mathcal{G}_s$, $s = (\alpha, L)$, where $\mathcal{G}_s$ is either Hölder or Sobolev ball, is given by $\psi_{\varepsilon,1}(\alpha)$. As we see, the accuracy of estimation in such models corresponds to the one-dimensional rate ($d = 1$).

Further results and references on estimation in models (i)–(iv) can be found, e.g., in Nicoleris and Yatracos [28], Györfi et al. [11, Chap. 22], and Ibragimov [17]. Let us briefly discuss the results obtained.

- The estimators providing the rate mentioned above depend heavily on the use of $\mathbb{L}_2$-losses ($p = 2$) in the risk definition. As a consequence, all proposed constructions cannot be used for any other types of loss functions.
- Except for the paper by Golubev [9], where an estimator independent of the parameter $s = (\alpha, L)$ was proposed for the model (i), all other estimators depend explicitly on the prior information on smoothness of the underlying function.
- As far as we know no minimax results have been obtained for the model (iv). One can guess that asymptotics of the risk (2) is given by $\psi_{\varepsilon,m}(\alpha)$ which is much better then the d-dimensional rate $\psi_{\varepsilon,d}(\alpha)$ since $m < d$.

It is also worth mentioning that there is a vast literature on estimation of vectors e_i, when f_i are treated as nonparametric nuisance parameters; (see, e.g., [12–15] and references therein.)

Oracle approach. To understand the place of the oracle approach within the theory of nonparametric estimation let us quote Johnstone [19]:

> *"Oracle inequalities are neither the beginning nor the end of a theory, but when available, are informative tools."*

Indeed, oracle inequalities are very powerful tools for deriving minimax and minimax adaptive results. The aim of the oracle approach can be formulated as follows: given a collection of different estimators based on available data, select the best estimator from the family (*model selection*) (see, e.g., [1]), or find the best convex/linear combination of the estimators from the family (*convex/linear aggregation*) (see [27,33]). The formal definition of the oracle requires specification of the collection of estimators and the criterion of optimality.

The majority of oracle procedures described in the literature use the $\mathbb{L}_2$-risk as the criterion of optimality. The following methods can be cited in this context: penalized likelihood estimators, unbiased risk estimators, blockwise Stein estimators, risk hull estimators and so on (see [1,4,10] and references therein). The most general results in the framework of $\mathbb{L}_2$-risk aggregation theory were obtained by Nemirovski [27] who showed how to aggregate arbitrary estimators.

Other oracle procedures were developed in the context of pointwise estimation; see, e.g., [2,8,24] for the univariate case, and [21,22] for the multivariate case. Moreover [21,23] show how to derive $\mathbb{L}_p$-norm oracle inequalities from pointwise oracle inequalities. Although these $\mathbb{L}_p$-norm oracle inequalities allow to derive minimax results on rather complicated functional spaces, they do not lead to sharp adaptive results.

Finally we mention the $\mathbb{L}_1$-norm oracle approach developed by Devroye and Lugosi [6] in context of density estimation.

The rest of the paper is organized as follows. In Sect. 2 we present our general selection rule and establish the *key oracle inequality*. Section 3 is devoted to adaptive estimation in the additive multi-index model. The proofs of the mains results are given in Sect. 4. Auxiliary results are postponed to Appendix.

2 General selection rule

2.1 Preliminaries

In what follows $\|\cdot\|_p$ stands for the $\mathbb{L}_p(\mathcal{D})$-norm, while $\|\cdot\|_{p,q}$ denotes the $\mathbb{L}_{p,q}(\mathcal{D}\times\mathcal{D}_0)$-norm:

$$\|G\|_{p,q}=\left(\int\limits_{\mathcal{D}}\left(\int\limits_{\mathcal{D}_0}|G(t,x)|^p\mathrm{d}t\right)^{q/p}\mathrm{d}x\right)^{1/q},\quad p,q\in[1,\infty]$$

with usual modification when $p = \infty$ and/or $q = \infty$. We write also $|\cdot|$ for the Euclidean norm, and it will be always clear from the context which Euclidean space is meant.

Let $\Theta \subset \mathbb{R}^m$. Assume that we are given a parameterized family of kernels $\mathcal{K} = \{K_\theta(\cdot,\cdot), \theta \in \Theta\}$, where $K_\theta : \mathcal{D} \times \mathcal{D}_0 \to \mathbb{R}$. Consider the collection of linear estimators of F associated with family $\mathcal{K}$:

$$\mathcal{F}(\mathcal{K}) = \left\{ \hat{F}_\theta(x) = \int K_\theta(t,x) Y(\mathrm{d}t),\ \theta \in \Theta \right\}.$$

Our goal is to propose a measurable choice from the collection $\{\hat{F}_\theta, \theta \in \Theta\}$ such that the risk of the selected estimator will be as close as possible to $\inf_{\theta\in\Theta} \mathcal{R}_p[\hat{F}_\theta, F]$.

Let

$$B_\theta(x) := \int K_\theta(t,x) F(t)\mathrm{d}t - F(x), \quad Z_\theta(x) := \int K_\theta(t,x) W(\mathrm{d}t); \tag{5}$$

then $\hat{F}_\theta(x) - F(x) = B_\theta(x) + \varepsilon Z_\theta(x)$, so that $B_\theta(\cdot)$ and $\varepsilon Z_\theta(\cdot)$ are the bias and the stochastic error of the estimator $\hat{F}_\theta$ respectively. We assume that the family $\mathcal{K}$ of kernels satisfies the following conditions.

(K0) For every $x \in \mathcal{D}_0$ and $\theta \in \Theta$, $\mathrm{supp}\{K_\theta(\cdot,x)\} \subseteq \mathcal{D}$,

$$\int K_\theta(t,x)\mathrm{d}t = 1, \quad \forall (x,\theta) \in \mathcal{D}_0 \times \Theta, \tag{6}$$

$$\sigma(\mathcal{K}) := \sup_{\theta\in\Theta} \|K_\theta\|_{2,\infty} < \infty, \tag{7}$$

$$M(\mathcal{K}) := \sup_{\theta\in\Theta} \left\{ \sup_x \|K_\theta(\cdot,x)\|_1 \vee \sup_t \|K_\theta(t,\cdot)\|_1 \right\} < \infty. \tag{8}$$

Remark 1 Conditions (6) and (7) are absolutely standard in the context of kernel estimation, and only condition (8) has to be discussed. First we note that (8) is rather mild. In particular, if the collection $\mathcal{K}$ contains positive kernels such that $K_\theta(t,x) = K_\theta(t-x)$ then $M(\mathcal{K}) = 1$. The quantity $M(\mathcal{K})$ will appear in the expression of the constant $\mathcal{L}$ in the $\mathbb{L}_p$-*norm oracle inequality* (4).

(K1) For any $\theta, \nu \in \Theta$

$$\int K_\theta(t,y) K_\nu(y,x)\mathrm{d}y = \int K_\theta(y,x) K_\nu(t,y)\mathrm{d}y \quad \forall (x,t) \in \mathcal{D}_0 \times \mathcal{D}. \tag{9}$$

Remark 2 Assumption K1 is crucial for the construction of our estimation procedure, and it restricts the collection of kernels to be used. We note that property (9) is trivially fulfilled for convolution kernels $K_\theta(t,x) = K_\theta(t-x)$ that correspond to the standard kernel estimators:

$$\int K_\theta(t-y) K_\nu(y-x)\mathrm{d}y = \int K_\theta(y-x) K_\nu(t-y)\mathrm{d}y.$$

The next example describes a collection of kernels corresponding to the single-index model.

Example Let $K : \mathbb{R}^d \to \mathbb{R}$, $\int K(t)\mathrm{d}t = 1$, E be an orthogonal matrix with the first vector-column equal to e. Define for all $h \in \mathbb{R}^d_+$

$$K_h(t) = \left[\prod_{i=1}^d h_i\right]^{-1} K\left(\frac{t_1}{h_1}, \ldots, \frac{t_d}{h_d}\right).$$

Denote $\mathcal{H} = \{h \in \mathbb{R}^d_+ : h = (h_1, h_{\max}, \ldots, h_{\max}), h_1 \in [h_{\min}, h_{\max}]\}$, where the bandwidth range $[h_{\min}, h_{\max}]$ is supposed to be fixed. The collection of the kernels corresponding to the single-index model is

$$\mathcal{K} = \Big\{K_\theta(t, x) = K_h[E^T(t - x)], \;\; \theta = (E, h) \in \Theta = \mathcal{E} \times \mathcal{H} \subset \mathbb{R}^d\Big\},$$

where $\mathcal{E}$ is the set of all $d \times d$ orthogonal matrices.

Clearly, $M(\mathcal{K}) = \|K\|_1$ so that K0 is fulfilled if $\|K\|_1 < \infty$. Assumption K1 is trivially fulfilled because $K_\theta(t, x) = K_\theta(t - x)$.

For $\theta, \nu \in \Theta$ we define

$$K_{\theta,\nu}(t, x) := \int K_\theta(t, y)K_\nu(y, x)\mathrm{d}y, \tag{10}$$

and let

$$\hat{F}_{\theta,\nu}(x) := \int K_{\theta,\nu}(t, x)Y(\mathrm{d}t), \quad x \in \mathcal{D}_0.$$

Observe that $K_{\theta,\nu} = K_{\nu,\theta}$ in view of (9), so that indeed $\hat{F}_{\theta,\nu} \equiv \hat{F}_{\nu,\theta}$. This property is heavily exploited in the sequel, since the statistic $\hat{F}_{\theta,\nu}$ is an auxiliary estimator used in our construction. We have

$$\begin{aligned}\hat{F}_{\theta,\nu}(x) - F(x) \;&=\; \int K_{\theta,\nu}(t, x)F(t)\mathrm{d}t - F(x) + \varepsilon \int K_{\theta,\nu}(t, x)W(\mathrm{d}t)\\ &=: B_{\theta,\nu}(x) + \varepsilon Z_{\theta,\nu}(x).\end{aligned} \tag{11}$$

The next simple result is a basic tool for construction of our selection procedure.

Lemma 1 *Let Assumption K0 hold; then for any* $F \in \mathbb{L}_2(\mathcal{D}) \cap \mathbb{L}_p(\mathcal{D})$

$$\sup_{\nu\in\Theta} \|B_{\theta,\nu} - B_\nu\|_p \le M(\mathcal{K})\|B_\theta\|_p \quad \forall\theta \in \Theta. \tag{12}$$

Proof By definition of $B_{\theta,\nu}$, B_ν and by the Fubini theorem

$$\begin{aligned} B_{\theta,\nu}(x) - B_\nu(x) &= \int K_{\theta,\nu}(t,x)F(t)\mathrm{d}t - \int K_\nu(t,x)F(t)\mathrm{d}t \\ &= \int K_\nu(y,x)\left[\int K_\theta(t,y)F(t)\mathrm{d}t - F(y)\right]\mathrm{d}y \\ &= \int K_\nu(y,x)B_\theta(y)\mathrm{d}y. \end{aligned}$$

The statement of the lemma follows from the general theorem about boundedness of integral operators on $\mathbb{L}_p$-spaces (see, e.g., [7, Theorem 6.18]) and (8). □

2.2 Selection rule

In order to present the basic idea underlying construction of the selection rule we first discuss the noise-free version ($\varepsilon = 0$) of the estimation problem.

Idea of construction (*ideal case* $\varepsilon = 0$). In this situation

$$\mathcal{F}(\mathcal{K}) = \left\{\hat{F}_\theta(\cdot) = \int K_\theta(t,\cdot)F(t)\mathrm{d}t \quad \forall\theta \in \Theta\right\}.$$

so that $\hat{F}_\theta$ can be viewed as a kernel-type approximation (smoother) of F. Note that the risk $\mathcal{R}_p[\hat{F}_\theta; F] = \|\hat{F}_\theta - F\|_p = \|B_\theta\|_p$ represents the quality of approximation. Let $\hat{F}_{\theta_*}$ be a smoother from $\mathcal{F}(\mathcal{K})$ with the minimal approximation error, i.e.

$$\theta_* = \arg\inf_{\theta\in\Theta} \mathcal{R}_p[\hat{F}_\theta; F].$$

Suppose that $\mathcal{K}$ satisfies Assumptions K0 and K1. Based on this collection we want to select a smoother, say $\hat{F}_{\hat{\theta}} \in \mathcal{F}(\mathcal{K})$, that is "as good as" $\hat{F}_{\theta_*}$, i.e., the smoother satisfying $\mathbb{L}_p$-oracle inequality (4).

To select $\hat{\theta}$ we suggest the following rule

$$\hat{\theta} = \arg\inf_{\theta\in\Theta}\{\sup_{\nu\in\Theta}\|\hat{F}_{\theta,\nu} - \hat{F}_\nu\|_p\}.$$

Let us compute the approximation error of the selected smoother $\hat{F}_{\hat{\theta}}$. By the triangle inequality

$$\begin{aligned} \|B_{\hat{\theta}}\|_p = \|\hat{F}_{\hat{\theta}} - F\|_p &\le \|\hat{F}_{\hat{\theta}} - \hat{F}_{\hat{\theta},\theta_*}\|_p + \|\hat{F}_{\hat{\theta},\theta_*} - \hat{F}_{\theta_*}\|_p + \|\hat{F}_{\theta_*} - F\|_p \\ &= \|B_{\hat{\theta}} - B_{\hat{\theta},\theta_*}\|_p + \|B_{\hat{\theta},\theta_*} - B_{\theta_*}\|_p + \|B_{\theta_*}\|_p. \end{aligned} \tag{13}$$

In view of Assumption K1 and (12) the first term on the right hand side of (13) does not exceed $M(\mathcal{K})\|B_{\theta_*}\|_p$. To bound the second term we use the definition of $\hat{\theta}$ and (12):

$$\|B_{\hat\theta,\theta_*} - B_{\theta_*}\|_p \le \sup_{\nu\in\Theta} \|B_{\hat\theta,\nu} - B_\nu\|_p \le \sup_{\nu\in\Theta} \|B_{\theta_*,\nu} - B_\nu\|_p \le M(\mathcal{K})\|B_{\theta_*}\|_p.$$

Combining these bounds we obtain from (13) that

$$\mathcal{R}_p[\hat F_{\hat\theta}; F] \le (2M(\mathcal{K}) + 1)\|B_{\theta_*}\|_p = (2M(\mathcal{K}) + 1) \inf_{\theta\in\Theta} \mathcal{R}_p[\hat F_\theta; F].$$

Therefore in the ideal situation $\varepsilon = 0$, the $\mathbb{L}_p$-oracle inequality (4) holds with $\mathcal{L} = 2M(\mathcal{K}) + 1$.

Example (*continuation*) We suppose additionally that there exists a positive integer l such that

$$\int t^{\boldsymbol{k}} K(t)\mathrm{d}t = 0, \quad |\boldsymbol{k}| = 1, \dots, l,$$

where $\boldsymbol{k} = (k_1, \dots, k_d)$ is the multi-index, $k_i \ge 0$, $|\boldsymbol{k}| = k_1 + \cdots + k_d$, $t^{\boldsymbol{k}} = t_1^{k_1} \cdots t_d^{k_d}$ for $t = (t_1, \dots, t_d)$. Let e be the true direction vector in the model (i). After rotation described by the matrix E for any $h \in \mathcal{H}$ we have

$$\|B_{\theta_*}\|_p \le \left\| \int K(u)[f(\cdot + h_1 u) - f(\cdot)]\mathrm{d}u \right\|_p .$$

If there exists $0 < \alpha < l + 1$, $L > 0$ such that $f \in \mathbb{H}_1(\alpha, L)$ then

$$\|B_{\theta_*}\|_p \le cLh_1^\alpha \quad \forall h_1 \in [h_{\min}, h_{\max}], \tag{14}$$

where c a numerical constant depending on K only. It is evident that when there is no noise in the model, the best choice of h_1 is $h_{\min}$.

Idea of construction (*real case* $\varepsilon > 0$) When the noise is present, we use the same selection procedure with additional control of the noise contribution by its maximal value. Similarly to the ideal case our selection rule is based on the statistics $\{\sup_{\nu\in\Theta} \|\hat F_{\theta,\nu} - \hat F_\nu\|_p,\ \theta \in \Theta\}$. Note that

$$\begin{aligned} \|\hat F_{\theta,\nu} - \hat F_\nu\|_p &\le \|B_{\theta,\nu} - B_\nu\|_p + \varepsilon\|Z_{\theta,\nu} - Z_\nu\|_p \\ &\le \|B_{\theta,\nu} - B_\nu\|_p + \varepsilon \sup_x |\tilde\sigma_{\theta,\nu}(x)| \sup_{\theta,\nu} \|\tilde Z_{\theta,\nu}\|_p, \end{aligned} \tag{15}$$

where $Z_{\theta,\nu}(\cdot)$ and $Z_\nu(\cdot)$ are given in (11) and (5) respectively, and

$$\sigma^2_{\theta,\nu}(x) := \mathbb{E}|Z_{\theta,\nu}(x) - Z_\nu(x)|^2 = \|K_{\theta,\nu}(\cdot, x) - K_\nu(\cdot, x)\|_2^2,\ x \in \mathcal{D}_0, \tag{16}$$

$$\tilde\sigma_{\theta,\nu}(x) := \max\{\sigma_{\theta,\nu}(x)\,,\ 1\}$$

$$\tilde Z_{\theta,\nu}(x) := \tilde\sigma^{-1}_{\theta,\nu}(x)[Z_{\theta,\nu}(x) - Z_\nu(x)]. \tag{17}$$

Remark 3 In what follows we will be interested in large deviation probability for the maximum of the process $Z_{\theta,\nu}(x) - Z_\nu(x)$. Typically the variance $\sigma_{\theta,\nu}(x)$ of this process tends to infinity as $\varepsilon \to 0$; therefore in the most interesting examples $\tilde{\sigma}_{\theta,\nu}(x) = \sigma_{\theta,\nu}(x)$, and $\tilde{Z}_{\theta,\nu}(x)$ has unit variance. However, for an abstract collection of the kernels, it can happen that $\sigma_{\theta,\nu}(x)$ is very small, for example, if K_θ approaches the delta-function. That is why we truncate the variance from below by 1.

In the ideal case we deduced from (12) that

$$[M(\mathcal{K})]^{-1} \sup_{\nu\in\Theta} \|\hat{F}_{\theta,\nu} - \hat{F}_\nu\|_p \le \|B_\theta\|_p \quad \forall\theta \in \Theta, \tag{18}$$

i.e., the left hand side can be considered as a lower estimator of the bias. In the case of $\varepsilon > 0$ we would like to guarantee the same property with high probability.

This leads to the following control of the stochastic term. Let $\delta \in (0, 1)$, and let $\varkappa_p = \varkappa_p(\mathcal{K}, \delta)$ be the minimal positive real number such that

$$\mathbb{P}\left\{\sup_{\theta\in\Theta} \|\tilde{Z}_\theta(\cdot)\|_p \ge \varkappa_p\right\} + \mathbb{P}\left\{\sup_{(\theta,\nu)\in\Theta\times\Theta} \|\tilde{Z}_{\theta,\nu}(\cdot)\|_p \ge \varkappa_p\right\} \le \delta, \tag{19}$$

where similarly to (16) and (17) we set

$$\tilde{Z}_\theta(x) := \sigma_\theta^{-1}(x) Z_\theta(x),$$
$$\sigma_\theta^2(x) := \mathbb{E}|Z_\theta(x)|^2 = \|K_\theta(\cdot, x)\|_2^2.$$

The constant $\varkappa_p$ controls deviation of $\|\tilde{Z}_{\theta,\nu}\|_p$ as well as the deviation of standardized stochastic terms of all estimators from the collection $\mathcal{F}(\mathcal{K})$. We immediately obtain from (15), (16) and (19) that

$$\hat{B}_\theta(p) := [M(\mathcal{K})]^{-1} \sup_{\nu\in\Theta}\left[\|\hat{F}_{\theta,\nu} - \hat{F}_\nu\|_p - \varepsilon\varkappa_p \sup_x \tilde{\sigma}_{\theta,\nu}(x)\right] \le \|B_\theta\|_p \quad \forall\theta \in \Theta, \tag{20}$$

with probability larger than $1 - \delta$.

Thus, similarly to (18), $\hat{B}_\theta(p)$ is a lower estimator of the $\mathbb{L}_p$-norm of the bias of the estimator $\hat{F}_\theta$. This leads us to the following selection procedure.

Selection rule Define

$$\hat{\theta} = \hat{\theta}(\delta) := \arg\inf_{\theta\in\Theta}\left\{\hat{B}_\theta(p) + \varkappa_p(\mathcal{K}, \delta)\,\varepsilon \sup_x \sigma_\theta(x)\right\}, \tag{21}$$

and put finally

$$\hat{F}(\delta) = \hat{F}_{\hat{\theta}}.$$

Remark 4 The choice of $\hat{\theta}$ is very natural. Indeed, in view of (20) for any $\theta \in \Theta$ with high probability

$$\hat{B}_\theta(p) + \varkappa_p \varepsilon \sup_x \sigma_\theta(x) \le \|B_\theta\|_p + \varkappa_p \varepsilon \sup_x \sigma_\theta(x).$$

On the other hand, under rather general assumptions (see Sect. 2.4)

$$\|B_\theta\|_p + \varepsilon \varkappa_p \sup_x \sigma_\theta(x) \le C \mathcal{R}_p[\hat{F}_\theta; F],$$

where C is an absolute constant, independent of F and ε. Therefore with high probability

$$\hat{B}_{\hat{\theta}}(p) + \varkappa_p \varepsilon \sup_x \sigma_{\hat{\theta}}(x) \le C \inf_{\theta \in \Theta} \mathcal{R}_p[\hat{F}_\theta; F].$$

Thus in order to establish the $\mathbb{L}_p$-norm oracle inequality it suffices to majorate the risk of the estimator $\hat{F}_{\hat{\theta}}$ by $\hat{B}_{\hat{\theta}}(p) + \varkappa_p \varepsilon \sup_x \sigma_{\hat{\theta}}(x)$ and to choose $\delta = \delta(\varepsilon)$ tending to zero at an appropriate rate.

2.3 Basic result

The next theorem establishes the basic result of this paper.

Theorem 1 *Let Assumptions K*0 *and K*1 *hold, and suppose that*

(I) $\hat{\theta}$ *defined in* (21) *is measurable with respect to the observation* $\{Y(t), t \in \mathcal{D}\}$, *and* $\hat{\theta}$ *belongs to* Θ;

(II) *the events in* (19) *belong to the* σ*-algebra generated by the observation* $\{Y(t), t \in \mathcal{D}\}$.

Let $\delta \in (0, 1)$, $\varkappa_p$ *be defined in* (19), *and* F *be such that* (I) *and* (II) *hold. Then*

$$\mathbb{E}_F \|\hat{F}(\delta) - F\|_p \le [3 + 2M(\mathcal{K})] \inf_{\theta \in \Theta} \left\{ \|B_\theta\|_p + \varkappa_p \varepsilon \left[\sup_x \sigma_\theta(x) \right] \right\} + r(\delta), \quad (22)$$

where

$$r(\delta) := \|F\|_\infty [1 + M(\mathcal{K})]\delta + \sigma(\mathcal{K}) \delta^{1/2} [\mathbb{E}|\zeta|^2]^{1/2},$$

$\sigma(\mathcal{K})$ *is defined in* (7), $\zeta := \sup_{x,\theta} |\tilde{Z}_\theta(x)|$, *and* $\mathbb{E}$ *denotes expectation with respect to the Wiener measure.*

Remark 5 In order to verify measurability of $\hat{\theta}$ and the condition (II) we need to impose additional assumptions on the collection of kernels $\mathcal{K}$. These assumptions should guarantee smoothness properties of the sample paths of Gaussian processes $\{\tilde{Z}_\theta(x), (x, \theta) \in \mathcal{D}_0 \times \Theta\}$ and $\{\tilde{Z}_{\theta,\nu}(x), (x, \theta, \nu) \in \mathcal{D}_0 \times \Theta \times \Theta\}$. It is well-known

(see, e.g., [25]) that such properties for Gaussian processes can be described in terms of their covariance structures. In our particular case, the covariance structure is entirely determined by the collection of kernels $\mathcal{K}$. These fairly general conditions on $\mathcal{K}$ are given in Sect. 2.4.

To ensure that $\hat{\theta} \in \Theta$ we need not only smoothness conditions on the stochastic processes involved in the procedure description, but also conditions on smoothness of F. It is sufficient to suppose that F belongs to some isotropic Hölder ball, and this will be always assumed in the sequel. This hypothesis also guarantees that F is uniformly bounded, which, in turn, implies boundedness of the remainder term $r(\delta)$. It is important to note that neither the procedure nor the inequality (22) depend on parameters of this ball.

Remark 6 Our procedure and the basic oracle inequality depend on the design parameter δ. The choice of this parameter is a delicate problem. On the one hand, in order to reduce the remainder term we should choose δ as small as possible. On the other hand, in view of the definition, $\varkappa_p = \varkappa_p(\delta) \to \infty$ as $\delta \to 0$. Note that we cannot minimize the right hand side of (22) with respect to δ because this leads to δ depending on the unknown function F. Fortunately, the same assumptions from Sect. 2.4 ensure that up to an absolute constant

$$\inf_{\theta\in\Theta}\left\{\|B_\theta\|_p + \varkappa_p\varepsilon\left[\sup_x \sigma_\theta(x)\right]\right\} \gtrsim \varepsilon. \tag{23}$$

The form of the remainder term $r(\delta)$ together with (23) suggests that δ should depend on ε, for example, $\delta = \delta(\varepsilon) = \varepsilon^a$, $a > 1$. Such a choice under assumptions from Sect. 2.4 allows to show that

$$\varkappa_p(\delta) = \varkappa_p(\delta(\varepsilon)) = \begin{cases} C(p), & p \in [1,\infty), \\ \sqrt{C(\infty)\ln(1/\varepsilon)}, & p = \infty, \end{cases} \tag{24}$$

where $C(p)$, $p \in [1,\infty]$, are absolute constants, independent of ε.

Although the inequality (22) is not stated in the form of the $\mathbb{L}_p$-norm oracle inequality, it can be helpful (in view of (24)) for deriving adaptive minimax results. To demonstrate this we return to the single-index model.

Example (*continuation*) Remind that $\theta = (E, h)$ and note that

$$\sigma_\theta^2(x) = \sigma_{E,h}^2(x) = \left[h_1 h_{\max}^{d-1}\right]^{-2} \int K_h^2[E^T(t-x)]\mathrm{d}t = \left[h_1 h_{\max}^{d-1}\right]^{-1} \|K\|_2^2$$

does not depend on E and x. Fix $\delta = \varepsilon^a$ and let $\hat{F}_\varepsilon$ be the estimator $\hat{F}(\varepsilon^a)$ satisfying (22). Then (22) takes the form

$$\mathbb{E}_F\|\hat{F}_\varepsilon - F\|_p \le (3+2\|K\|_1)\inf_{E,h}\left[\|B_{E,h}\|_p + \varepsilon\varkappa_p(\varepsilon^a)\sup_x \sigma_{E,h}(x)\right] + O(\varepsilon^a)$$
$$\le (3+2\|K\|_1)\inf_h\left[\inf_E\|B_{E,h}\|_p + \varepsilon\varkappa_p(\varepsilon^a)\left[h_1 h_{\max}^{d-1}\right]^{-1/2}\|K\|_2\right] + O(\varepsilon^a)$$
$$\le (3+2\|K\|_1)\inf_{h_1}\left[cLh_1^\alpha + \varkappa_p(\varepsilon^a)\left[h_1 h_{\max}^{d-1}\right]^{-1/2}\|K\|_2\right] + O(\varepsilon^a).$$

The last inequality follows from (14). Taking into account (24), choosing $h_{\max} > 0$ independent of ε, $h_{\min} = \varepsilon^2$, and minimizing the last inequality with respect to $h_1 \in [h_{\min}, h_{\max}]$ we obtain for all $\alpha > 0$, $L > 0$

$$\sup_{f\in\mathbb{H}_1(\alpha,L)} \mathbb{E}_F\|\hat{F}_\varepsilon - F\|_p \le C_p(L, h_{\max}, K)\begin{cases} \varepsilon^{2\alpha/(2\alpha+1)}, & p\in[1,\infty) \\ \left[\varepsilon\sqrt{\ln(1/\varepsilon)}\right]^{2\alpha/(2\alpha+1)}, & p=\infty. \end{cases}$$

It remains to note that $\hat{F}_\varepsilon$ does not depend on (α, L), and attains in view of the last inequality the minimax rate of convergence for all values of (α, L) simultaneously. It means that $\hat{F}_\varepsilon$ is optimally adaptive on the scale of Hölder balls.

2.4 Key oracle inequality

In this section we discuss the choice of δ which leads to the *key oracle inequality*. This inequality is suitable for deriving minimax and minimax adaptive results with minimal technicalities. In particular, we will use it for adaptive estimation in the additive multi-index model.

In order to establish the *key oracle inequality* we need to impose additional conditions on the collection of kernels $\mathcal{K}$. In particular, these conditions should guarantee the bounds (24) for $\varkappa_p(\delta(\varepsilon))$. In the case $p = \infty$ such conditions are rather mild and standard; they are related to deviation of supremum of Gaussian processes and therefore can be expressed through smoothness of their covariance functions [25]. As for the case $p < \infty$, we need to establish bounds on large deviation probabilities of the $\mathbb{L}_p$-norm of Gaussian processes. It requires additional assumptions on the collection of the kernels. Moreover, such bounds cannot be directly obtained from the existing results. We note nevertheless that (24) for the case $p < \infty$ can be shown under fairly general assumptions, and this will be the subject of a forthcoming paper. From now on we restrict ourselves with the case $p = \infty$.

In the end of this section we discuss the connection between the *key oracle inequality* and the $\mathbb{L}_\infty$-*norm oracle inequality* of type (4).

Assumptions We suppose that the set Θ has the following structure.

(A) $\Theta = \Theta_1 \times \Theta_2$ where $\Theta_1 = \{\theta^1, \dots, \theta^N\}$ is a finite set, and $\Theta_2 \subset \mathbb{R}^m$ is a compact subset of $\mathbb{R}^m$ contained in the Euclidean ball of radius R. Without loss of generality we assume that $R \ge 1$.

Remark 7 Assumption A allows to consider both discrete and continuous parameter sets. In particular, the case of empty Θ_2 corresponds to selection from a finite set of

estimators. This setup is often considered within the framework of the oracle approach. In order to emphasize dependence of kernels K_θ on $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$, we sometimes write $K_{(\theta_1,\theta_2)}$ instead of K_θ.

(B) There exists M_0 such that $F \in \mathbb{H}_d(M_0)$, where

$$\mathbb{H}_d(M_0) = \left\{ g : \ g \in \bigcup_{\alpha>0, L>0} \mathbb{H}_d(\alpha, L), \ \|g\|_\infty \le M_0 \right\}.$$

Remark 8 Assumption B is necessary for verification of the condition (I) of Theorem 1. It is also needed for deriving the key oracle inequality from Theorem 1 since it allows to bound uniformly the remainder term in (22).

We emphasize that our procedure does not depend on M_0. Finally note that $\mathbb{H}_d(M_0)$ is a huge set of functions(a bit smaller than the space of all bounded continuous functions), i.e., Assumption B is not restrictive at all.

(K2) Denote $U := \mathcal{D}_0 \times \Theta_2$. There exist positive constants $\bar{L}$, and $\gamma \in (0, 1]$ such that

$$\sup_{\theta_1\in\Theta_1} \sup_{u,u'\in U} \frac{\|K_{(\theta_1,\theta_2)}(\cdot, x) - K_{(\theta_1,\theta_2')}(\cdot, x')\|_2}{|u - u'|^\gamma} \le \bar{L},$$

where $u = (x, \theta_2)$, and $u' = (x', \theta_2')$. Without loss of generality we assume that $\bar{L} \ge 1$.

Remark 9 Assumption K2 ensures that sample paths of the processes $\{\tilde{Z}_\theta(x), (x, \theta) \in \mathcal{D}_0 \times \Theta\}$ and $\{\tilde{Z}_{\theta,\nu}(x), (x, \theta, \nu) \in \mathcal{D}_0 \times \Theta \times \Theta\}$ belong with probability one to the isotropic Hölder spaces $\mathbb{H}_{m+d}(\tau)$ and $\mathbb{H}_{2m+d}(\tau)$ with regularity index $0 < \tau < \gamma$ [25, Sect. 15]. In particular, it is sufficient for fulfillment of conditions (I) and (II) of Theorem 1.

Choice of δ. Now we are ready to state the upper bound on the risk of our estimator (21) under Assumptions A, B, K0–K2. Define

$$C_{\mathcal{K}} := M(\mathcal{K})\bar{L}R$$

Theorem 2 *Let Assumptions A, B, K0–K2 hold, and assume that there exists $a > 0$ such that*

$$\delta_* := \min\left\{ \frac{1}{N}, C_{\mathcal{K}}^{-(2m+d)/\gamma}, \varepsilon^2[\sigma(\mathcal{K})]^{-2} \right\} \ge \varepsilon^a. \tag{25}$$

Let $\hat{F}_* = \hat{F}(\delta_*)$ be the estimator of Sect. 2 associated with the choice $\delta = \delta_*$. Then there exists a constant $C_1 \ge M_0$ depending on d, m and γ only such that

$$\mathbb{E}_F\|\hat{F}_* - F\|_\infty \le [3 + 2M(\mathcal{K})] \inf_{\theta\in\Theta} \left\{ \|B_\theta\|_\infty + C_1\varepsilon\sqrt{\ln \varepsilon^{-1}} \sup_x \sigma_\theta(x) \right\}. \tag{26}$$

Remark 10 Typically in nonparametric setups $\bar{L} \sim \varepsilon^{-a_1}$, $\sigma(\mathcal{K}) \sim \varepsilon^{-a_2}$ for some $a_1, a_2 > 0$. If N grows not faster than ε^{-a_3}, then (25) holds.

$\mathbb{L}_\infty$-norm oracle inequality. Finally we show how the $\mathbb{L}_\infty$-norm oracle inequality (4) can be obtained from Theorem 2.

Theorem 3 *Assume that there exists a constant $C_2 > 0$ such that*

$$\mathbb{E}\|Z_\theta(\cdot)\|_\infty \ge C_2\sqrt{\ln(1/\varepsilon)}\,\sup_x \sigma_\theta(x), \quad \forall \theta \in \Theta, \tag{27}$$

and let $\hat{F}_*$ be the estimator from Theorem 2. Then

$$\mathcal{R}_\infty[\hat{F}_*; F] \le \mathcal{L} \inf_{\theta\in\Theta} \mathcal{R}_\infty[\hat{F}_\theta; F],$$

where $\mathcal{L} = 3[3 + 2M(\mathcal{K})]\max\{1, C_1/C_2\}$.

Remark 11 The condition (27) seems to be necessary in order to have the constant $\mathcal{L}$ independent of ε. In fact, (27) is an assumption on the collection of kernels $\mathcal{K}$. To verify this condition one can use the Sudakov lower bound on the expectation of the maximum of a Gaussian process (see, e.g., [25, Sect. 14]).

The proof of Theorem 3 is an immediate consequence of Theorem 2, (27), and the following auxiliary result that is interesting in its own right.

Lemma 2 *Let $\tilde{F}(\cdot) = \int S(t,\cdot)Y(dt)$ be a linear estimator of $F(\cdot)$. Denote by $B_S(\cdot)$ and $\varepsilon Z_S(\cdot)$ the bias and the stochastic part of $\tilde{F}(\cdot) - F(\cdot)$ respectively. Then for any $F \in \mathbb{L}_p(\mathcal{D}) \cap \mathbb{L}_2(\mathcal{D})$ and $p \in [1, \infty]$*

$$\frac{1}{3}\{\|B_S\|_p + \varepsilon\mathbb{E}\|Z_S\|_p\} \le \mathcal{R}_p[\tilde{F}; F] \le \|B_S\|_p + \varepsilon\mathbb{E}\|Z_S\|_p. \tag{28}$$

3 Adaptive estimation in additive multi-index model

In this section we apply the *key oracle inequality* of Theorem 2 to adaptive estimation in the additive multi-index model.

3.1 Problem formulation

We impose the following structural assumption on the function F in the model (1).

Let $\mathcal{I}$ denote the set of all partitions of $(1, \dots, d)$, and for $\eta > 0$ let

$$\mathcal{E}_\eta = \{E = (e_1, \dots, e_d) : e_i \in \mathbb{S}^{d-1}, \quad |\det(E)| \ge \eta\}.$$

For any $I \in \mathcal{I}$ and $E \in \mathcal{E}_\eta$ let $E_1, \dots, E_{|I|}$ be the corresponding partition of columns of E.

(F) *Let $I = (I_1, \dots, I_{|I|}) \in \mathcal{I}$, and $E \in \mathcal{E}_\eta$. There exist functions $f_i : \mathbb{R}^{|I_i|} \to \mathbb{R}$, $i = 1, \dots, |I|$ such that*

$$F(t) = \sum_{i=1}^{|I|} f_i\left(E_i^T t\right).$$

Assumption F states that the unknown function F can be represented as a sum of $|I|$ unknown functions f_i, $i = 1, \dots, |I|$, where f_i is $|I_i|$-dimensional after an unknown linear transformation. Note that partition I is also unknown. The assumption that $|\det(E)| \geq \eta$ is chosen for technical reasons; note that our estimation procedure does not require knowledge of the value of this parameter.

Later on the functions f_i will be supposed to be smooth; in particular, we will assume that all f_i's belong to an isotropic Hölder ball (see the next definition).

Definition 1 A function $f : \mathcal{T} \to \mathbb{R}$, $\mathcal{T} \subset \mathbb{R}^s$, is said to belong to the Hölder ball $\mathbb{H}_s(\beta, L)$ if f has continuous partial derivatives of all orders $\leq l$ satisfying the Hölder condition with exponent $\alpha \in (0, 1]$:

$$\begin{aligned} &\|D^{\boldsymbol{k}} f\|_\infty \leq L, \quad \forall |\boldsymbol{k}| = 0, \dots, l; \\ &\left| f(z) - \sum_{j=0}^{l} \frac{1}{j!} \sum_{|\boldsymbol{k}|=j} D^{\boldsymbol{k}} f(t)(z-t)^{\boldsymbol{k}} \right| \leq L|z-t|^\beta, \quad \forall z, t \in \mathcal{T}, \end{aligned}$$

where $\beta = l + \alpha$, $\boldsymbol{k} = (k_1, \dots, k_s)$ is a multi-index, $k_i \geq 0$, $|\boldsymbol{k}| = k_1 + \cdots + k_s$, $t^{\boldsymbol{k}} = t_1^{k_1} \cdots t_s^{k_s}$ for $t = (t_1, \dots, t_s)$, and $D^{\boldsymbol{k}} = \partial^{|\boldsymbol{k}|} / \partial t_1^{k_1} \cdots \partial t_s^{k_s}$.

The described structure includes models (i)–(iv).

1. *Single-index model.* Let $F(t) = f(e^T t)$ for some unknown $f : \mathbb{R}^d \to \mathbb{R}$ and $e \in \mathbb{S}^{d-1}$. In order to express the single-index model in terms of assumption F, we set $E = (e_1, \dots, e_d)$ with $e_1, e_2, \dots, e_d$ being an orthogonal basis of $\mathbb{R}^d$ such that $e_1 = e$. In this case we can set $I = (I_1, I_2)$ with $I_1 = \{1\}$, $I_2 = \{2, \dots, d\}$ and $f_1 = f$, $f_2 \equiv 0$.
2. *Additive model.* Let $F(t) = \sum_{i=1}^d f_i(x_i)$ for unknown $f_i : \mathbb{R}^d \to \mathbb{R}$. Here E is the $d \times d$ identity matrix, and $I = (I_1, \dots, I_d)$, $I_i = \{i\}$.
3. *Projection pursuit model.* Let $F(t) = \sum_{i=1}^d f_i(e_i^T t)$ for unknown $f_i : \mathbb{R}^d \to \mathbb{R}^1$ and unknown linearly independent direction vectors $e_1, \dots, e_d \in \mathbb{S}^{d-1}$. Here $E = (e_1, \dots, e_d)$, $I = (I_1, \dots, I_d)$, $I_i = \{i\}$.
4. *Multi-index model.* Let $F(t) = f(e_1^T t, \dots, e_m^T t)$ for unknown direction vectors $e_1, \dots, e_m \in \mathbb{S}^{d-1}$, and unknown function $f : \mathbb{R}^m \to \mathbb{R}^1$. We define $E = (e_1, \dots, e_d)$, where $(e_{m+1}, \dots, e_d)$ is the orthogonal basis of the orthogonal complement to the subspace $\operatorname{span}\{e_1, \dots, e_m\}$. In this case we set $I = (I_1, I_2)$, $I_1 = (1, \dots, m)$, $I_2 = (m+1, \dots, d)$, and $f_1 = f$, $f_2 \equiv 0$.

Definition 2 We say that function F belongs to the class $\mathbb{F}_{I,E}(\beta, L)$, $\beta > 0$, $L > 0$ if

(i) Assumption F is fulfilled with partition $I = (I_1, \dots, I_{|I|}) \in \mathcal{I}$ and matrix $E \in \mathcal{E}_\eta$;

(ii) there exist positive real numbers β_i and L such that $f_i \in \mathbb{H}_{|I_i|}(\beta_i, L)$, $i = 1, \ldots, |I|$;
(iii) For all $i = 1, \ldots, |I|$

$$\beta = \frac{\beta_i}{|I_i|}. \tag{29}$$

Remark 12 The meaning of condition (iii) is that smoothness of functions f_i is related to their dimensionality in such a way that the effective smoothness of all functional components is the same. This condition does not restrict generality as smoothness of a sum of functions is determined by the worst smoothness of summands. In particular, if (29) is not fulfilled then the results of this section hold with $\beta = \min_{i=1,\ldots,|I|} \beta_i/|I_i|$.

Let $\tilde{F}$ be an estimator of $F \in \mathbb{F}_{I,E}(\beta, L)$; accuracy of $\tilde{F}$ is measured by the maximal risk

$$\mathcal{R}_\infty[\tilde{F}; \mathbb{F}_{I,E}(\beta, L)] := \sup_{F \in \mathbb{F}_{I,E}(\beta,L)} \mathbb{E}_F \|\tilde{F} - F\|_\infty.$$

Proposition 1 (Minimax lower bound) *Let* $\varphi_\varepsilon(\beta) = [\varepsilon\sqrt{\ln(1/\varepsilon)}]^{2\beta/(2\beta+1)}$. *Then*

$$\liminf_{\varepsilon\to 0} \inf_{\tilde{F}} \varphi_\varepsilon^{-1}(\beta)\, \mathcal{R}_\infty[\tilde{F}; \mathbb{F}_{I,E}(\beta, L)] > 0, \quad I \in \mathcal{I},\ E \in \mathcal{E}_\eta,$$

where inf *is taken over all possible estimators* $\tilde{F}$.

Remark 13 The appearance of the univariate rate $\varphi_\varepsilon(\beta)$ in the lower bound is not surprising since $2\beta/(2\beta + 1) = 2\beta_i/(2\beta_i + |I_i|)$, $i = 1, \ldots, |I|$ in view of (29). It is worth mentioning that $\varphi_\varepsilon(\beta) = \psi_{\varepsilon,|I_i|}(\beta_i)$ is the minimax rate of convergence in estimation of each component f_i [cf. (3)].

The proof of Proposition 1 is absolutely standard and is omitted. Obviously, the accuracy of estimation under the additive multi-index model cannot be better than the accuracy of estimation of one component provided that all other components are identically zero. Since E is fixed, the problem is reduced to estimating $|I_i|$-variate function of smoothness β_i in the model (1). In this case the lower bound is well-known and given by $\psi_{\varepsilon,|I_i|}(\beta_i)$. It remains to note that $\psi_{\varepsilon,|I_i|}(\beta_i)$ does not depend on i and coincides with $\varphi_\varepsilon(\beta)$ in view of (29).

Below we propose an estimator that attains the rate $\varphi_\varepsilon(\beta)$ simultaneously over $\mathbb{F}_{I,E}(\beta, L)$, $I \in \mathcal{I}$, $E \in \mathcal{E}_\eta$, $0 < \beta \le \beta_{\max} < \infty$, $L > 0$, i.e., the optimally adaptive estimator.

3.2 Kernel construction

To construct a family of kernel estimators let us consider the idealized situation when both the partition $I = (I_1, \ldots, I_{|I|}) \in \mathcal{I}$ and $E \in \mathcal{E}_\eta$ are known.

(G) *Let* $g : [-1/2, 1/2] \to \mathbb{R}$ *be a univariate kernel satisfying the following conditions*

(i) $\int g(x)\mathrm{d}x = 1, \quad \int g(x)x^k\mathrm{d}x = 0, \ k = 1, \ldots, \ell;$
(ii) $g \in \mathbb{C}^1$.

Fix a bandwidth $h = (h_1, \ldots, h_d)$, $h_{\min} \le h_i \le h_{\max}$ and put

$$G_0(t) = \prod_{i=1}^{d} g(t_i)$$
$$G_{i,h}(t) = \prod_{j \in I_i} \frac{1}{h_j} g\left(\frac{t_j}{h_j}\right) \prod_{j \notin I_i} g(t_j), \quad i = 1, \ldots, |I|.$$

Now we define the kernel associated with partition I, matrix E, and bandwidth h. Fix $\theta = (I, E, h) \in \Theta = \mathcal{I} \times \mathcal{E}_\eta \times [h_{\min}, h_{\max}]^d$, and let

$$K_\theta(t) = |\det(E)| \sum_{i=1}^{|I|} G_{i,h}(E^T t) - (|I| - 1)|\det(E)| G_0(E^T t). \tag{30}$$

3.3 Properties of the kernel

First we state evident properties of the kernel K_θ.

Lemma 3 *For any $\theta \in \Theta$*

$$\int K_\theta(t)\mathrm{d}t = 1$$
$$\|K_\theta\|_1 \le (2|I| - 1)\|g\|_1^d.$$
$$\|K_\theta\|_2 \le |\det(\mathrm{E})|^{1/2}\|g\|_2^d \left(\sum_{i=1}^{|I|} \prod_{j \in I_i} h_j^{-1/2} + |I| - 1\right). \tag{31}$$

The proof follows straightforwardly from (30).
Next lemma establishes approximation properties of K_θ. Put for any $x \in \mathcal{D}_0$

$$B_\theta(x) = \int K_\theta(t - x)F(t)\mathrm{d}t - F(x).$$

Clearly, $B_\theta(\cdot)$ is the bias of the estimator associated with kernel K_θ.

Lemma 4 *Let $F \in \mathbb{F}_{I,E}(\beta, L)$, and let Assumption G hold with $\ell = \max_i \lfloor \beta_i \rfloor$. Then*

$$\|B_\theta\|_\infty \le L \sum_{i=1}^{|I|} \|g\|_1^{|I_i|} \sum_{j \in I_i} h_j^{\beta_i}. \tag{32}$$

Remark 14 Lemmas 3 and 4 allow to derive an upper bound on the accuracy of estimation on the class $\mathbb{F}_{I,E}(\beta, L)$ for given I and E. Indeed, the typical balance equation for the bandwidth selection takes the form

$$\varepsilon\sqrt{\ln(1/\varepsilon)}\|K_\theta\|_2 = \|B_\theta\|_\infty.$$

Therefore using the upper bounds in (32) and (31) we arrive to the optimal choice of bandwidth given by $h = h^* = (h_1^*, \dots, h_d^*)$,

$$h_j^* = \left(\frac{\varepsilon}{L}\sqrt{\ln(1/\varepsilon)}\right)^{2/(2\beta_i+|I_i|)} \left(\frac{\|g\|_2}{\|g\|_1}\right)^{2d/(2\beta_i+|I_i|)}, \quad j \in I_i, \ i = 1, \dots, |I|. \quad (33)$$

If $\hat{F}_\theta(x) = \int K_\theta(t-x)Y(\mathrm{d}t)$ is a kernel estimator with $\theta = (I, E, h_*)$ then we have the following upper bound on its $\mathbb{L}_\infty$-risk:

$$\mathcal{R}_\infty[\hat{F}_\theta; \mathbb{F}_{I,E}(\beta, L)] \le CL^{1/(2\beta+1)}\varphi_\varepsilon(\beta), \quad (34)$$

where C is an absolute constant. Thus, in view of Proposition 1, $\varphi_\varepsilon(\beta)$ is the minimax rate of convergence on the class $\mathbb{F}_{I,E}(\beta, L)$. We stress that construction of minimax estimator $\hat{F}_\theta$ requires knowledge of all parameters of the functional class: I, E, β and L.

3.4 Optimally adaptive estimator

Let $h_{\min} = \varepsilon^2$ and $h_{\max} = \varepsilon^{2/[(2\beta_{\max}+1)d]}$ for some $\beta_{\max} > 0$. Consider the collection of kernels $\mathcal{K} = \{K_\theta(\cdot), \theta = (I, E, h) \in \Theta\}$ where $K_\theta(\cdot)$ is defined in (30). The corresponding collection of estimators is given by

$$\mathcal{F}(\mathcal{K}) = \left\{\hat{F}_\theta(x) = \int K_\theta(t-x)Y(\mathrm{d}t), \quad \theta \in \Theta\right\}.$$

Based on the collection $\mathcal{F}(\mathcal{K})$ we define the estimator $\hat{F}_*$ following the selection rule (21) with the choice of $\delta = \varepsilon^a$ where $a = 24d^3 + 12d^2$.

Theorem 4 *Suppose that Assumption G holds with $\ell = \lfloor d\beta_{\max}\rfloor$. Then for any $I \in \mathcal{I}$, $E \in \mathcal{E}_\eta$, $0 < \beta \le \beta_{\max}$, and $L > 0$*

$$\limsup_{\varepsilon\to 0} \varphi_\varepsilon^{-1}(\beta)\, \mathcal{R}_\infty[\hat{F}_*; \mathbb{F}_{I,E}(\beta, L)] \le CL^{1/(2\beta+1)},$$

where C depends on d, $\beta_{\max}$, and the kernel g only.

Combining the results of Theorem 4 and Proposition 1 we obtain that the estimator $\hat{F}_*$ is optimally adaptive on the scale of functional classes $\mathbb{F}_{I,E}(\beta, L)$. Thus this estimator adjusts automatically to unknown structure as well as to unknown smoothness.

We note that traditionally any structural assumption is understood as the existence of the structure. Mathematically in our case it means that the underlying function belongs to the union of classes $\mathbb{F}_{I,E}(\beta, L)$ with respect to $I \in \mathcal{I}$ and $E \in \mathcal{E}_\eta$, i.e.,

$$F \in \mathbb{F}(\beta, L) = \bigcup_{I\in\mathcal{I}, E\in\mathcal{E}_\eta} \mathbb{F}_{I,E}(\beta, L).$$

Next theorem shows that our estimation procedure is optimally adaptive on the scale of functional classes $\mathbb{F}(\beta, L)$, $0 < \beta \le \beta_{\max}$, $L > 0$.

Theorem 5 *Suppose that Assumption G holds with $\ell = \lfloor d\beta_{\max} \rfloor$. Then for any $0 < \beta \le \beta_{\max}$, and $L > 0$*

$$\limsup_{\varepsilon\to 0} \varphi_\varepsilon^{-1}(\beta)\, \mathcal{R}_\infty[\hat{F}_*; \mathbb{F}(\beta, L)] \le C L^{1/(2\beta+1)},$$

where C depends on d, $\beta_{\max}$, and the kernel g only.

Theorem 4 follows immediately from Theorem 5. Proposition 1 together with Theorem 5 shows that in terms of rates of convergence there is no price to pay for adaptation with respect to unknown structure.

4 Proofs of Theorems 1, 2 and 5

Proof of Theorem 1. Define the random event

$$A = A_1 \cap A_2 := \left\{\omega : \sup_{\theta\in\Theta} \|\tilde{Z}_\theta\|_p \le \varkappa_p\right\} \cap \left\{\omega : \sup_{(\theta,\nu)\in\Theta\times\Theta} \|\tilde{Z}_{\theta,\nu}\|_p \le \varkappa_p\right\}.$$

1^0. First, we observe that

$$\hat{B}_\theta(p)\, 1(A) \le \|B_\theta\|_p, \quad \forall\theta \in \Theta. \tag{35}$$

Indeed, in view of Lemma 1 on the set A

$$\begin{aligned}
\|B_\theta\|_p &\ge \sup_{\nu\in\Theta} \frac{1}{\|K_\nu\|_{1,\infty}} \left\| \int K_\nu(t,x) B_\theta(t) \mathrm{d}t \right\|_p \\
&\ge [M(\mathcal{K})]^{-1} \sup_{\nu\in\Theta} \left(\|\hat{F}_{\theta,\nu} - \hat{F}_\nu\|_p - \varepsilon \|Z_{\theta,\nu} - Z_\nu\|_p \right) \\
&\ge [M(\mathcal{K})]^{-1} \sup_{\nu\in\Theta} \left[\|\hat{F}_{\theta,\nu} - \hat{F}_\nu\|_p - \varkappa_p \varepsilon \sup_x \tilde{\sigma}_{\theta,\nu}(x) \right] = \hat{B}_\theta(p),
\end{aligned}$$

where we have also used definition of A and the fact that

$$\hat{F}_{\theta,\nu}(x) - \hat{F}_\nu(x) = \int K_\nu(t,x) B_\theta(t) \mathrm{d}t + \varepsilon [Z_{\theta,\nu}(x) - Z_\nu(x)].$$

2^0. Second, we note that for any $\theta, \nu \in \Theta$

$$\begin{aligned}\sup_x \sigma_{\theta,\nu}(x) &= \|K_{\theta,\nu} - K_\nu\|_{2,\infty} \le \|K_{\theta,\nu}\|_{2,\infty} + \|K_\nu\|_{2,\infty} \\ &\le \|K_\theta\|_{1,\infty}\|K_\nu\|_{2,\infty} + \|K_\nu\|_{2,\infty} \le [1 + M(\mathcal{K})]\,\|K_\nu\|_{2,\infty} \\ &= [1 + M(\mathcal{K})] \sup_x \sigma_\nu(x).\end{aligned}$$

Here we have used the inequality $\|K_{\theta,\nu}\|_{2,\infty} \le \|K_\theta\|_{1,\infty}\|K_\nu\|_{2,\infty}$ which follows from the Minkowski integral inequality.

The Cauchy–Schwarz inequality and (6) yield $\sigma_\nu(x) \ge (\mathrm{mes}\{\mathcal{D}\})^{-1/2}$ for all x and ν. This implies without loss of generality that for any $\theta, \nu \in \Theta$

$$\sup_x \tilde\sigma_{\theta,\nu}(x) \le [1 + M(\mathcal{K})] \sup_x \sigma_\nu(x). \tag{36}$$

3^0. Now define

$$\theta_* := \arg\inf_{\theta\in\Theta} \{\, \|B_\theta\|_p + \varkappa_p \varepsilon \sup_x \sigma_\theta(x) \,\},$$

and let $\hat F_* = \hat F_{\theta_*}$. We write

$$\begin{aligned}\|\hat F - F\|_p 1(A) \le \|\hat F_{\theta_*} - F\|_p\, 1(A) + \|\hat F_{\theta_*} - \hat F_{\hat\theta,\theta_*}\|_p\, 1(A) \\ +\ \|\hat F_{\hat\theta} - \hat F_{\hat\theta,\theta_*}\|_p\, 1(A),\end{aligned} \tag{37}$$

and note that

$$\|\hat F_{\theta_*} - F\|_p\, 1(A) \le \|B_{\theta_*}\|_p + \varkappa_p\varepsilon \sup_x \sigma_{\theta_*}(x) = \inf_{\theta\in\Theta} \{\, \|B_\theta\|_p + \varkappa_p\varepsilon \sup_x \sigma_\theta(x) \,\}. \tag{38}$$

Furthermore,

$$\begin{aligned}\|\hat F_{\theta_*} - \hat F_{\hat\theta,\theta_*}\|_p\, 1(A) &\le M(\mathcal{K})\hat B_{\hat\theta}(p)\, 1(A) + \varkappa_p\varepsilon \sup_x \tilde\sigma_{\hat\theta,\theta_*}(x) \\ &\le M(\mathcal{K})\hat B_{\hat\theta}(p)\, 1(A) + [1 + M(\mathcal{K})]\varkappa_p\varepsilon \sup_x \sigma_{\theta_*}(x),\end{aligned}$$

where the first inequality follows from definition of $\hat B_\theta(p)$; the second inequality is a consequence of (8) and (36). Similarly,

$$\begin{aligned}\|\hat F_{\hat\theta} - \hat F_{\hat\theta,\theta_*}\|_p 1(A) &\le M(\mathcal{K})\hat B_{\theta_*}(p)\, 1(A) + \varkappa_p\varepsilon \sup_x \tilde\sigma_{\theta_*,\hat\theta}(x) \\ &\le M(\mathcal{K})\hat B_{\theta_*}(p)\, 1(A) + [1 + M(\mathcal{K})]\varkappa_p\varepsilon \sup_x \sigma_{\hat\theta}(x).\end{aligned}$$

Now using (21) and (35) we obtain

$$
\begin{aligned}
&[\|\hat{F}_{\theta_*} - \hat{F}_{\hat{\theta},\theta_*}\|_p + \|\hat{F}_{\hat{\theta}} - \hat{F}_{\hat{\theta},\theta_*}\|_p]1(A) \\
&\quad \le [1 + M(\mathcal{K})] \left\{ [\hat{B}_{\hat{\theta}}(p) + \hat{B}_{\theta_*}(p)]1(A) + \varkappa_p \varepsilon \sup_x \sigma_{\hat{\theta}}(x) + \varkappa_p \varepsilon \sup_x \sigma_{\theta_*}(x) \right\} \\
&\quad \le 2[1 + M(\mathcal{K})] \left\{ \|B_{\theta_*}\|_p + \varkappa_p \varepsilon \sup_x \sigma_{\theta_*}(x) \right\}.
\end{aligned}
$$

Then (37) and (38) lead to

$$
\|\hat{F} - F\|_p 1(A) \le [3 + 2M(\mathcal{K})] \inf_{\theta \in \Theta} \left\{ \|B_\theta\|_p + \varkappa_p \varepsilon \sup_x \sigma_\theta(x) \right\}. \tag{39}
$$

4^0. In order to complete the proof it suffices to bound $\|\hat{F} - F\|_p 1(A^c)$. Note that by our choice of $\varkappa_p$ (see (19)), $\mathbb{P}(A^c) \le \delta$. Moreover

$$
\begin{aligned}
\|\hat{F} - F\|_p 1(A^c) &\le \left(\sup_{\theta \in \Theta} \|B_\theta\|_p + \sup_{\theta \in \Theta} \|Z_\theta(\cdot)\|_p \right) 1(A^c) \\
&\le \|F\|_\infty [1 + M(\mathcal{K})] 1(A^c) + \sigma(\mathcal{K}) \zeta \, 1(A^c),
\end{aligned}
$$

where $\sigma(\mathcal{K})$ is defined in (7), and $\zeta := \sup_{x,\theta} |\tilde{Z}_\theta(x)|$. Therefore

$$
\begin{aligned}
\mathbb{E}\|\hat{F} - F\|_p 1(A^c) &\le \|F\|_\infty [1 + M(\mathcal{K})] \mathbb{P}(A^c) + \sigma(\mathcal{K}) [\mathbb{E}\zeta^2]^{1/2} \mathbb{P}^{1/2}(A^c) \\
&\le \|F\|_\infty [1 + M(\mathcal{K})] \delta + \sqrt{\delta}\, \sigma(\mathcal{K}) [\mathbb{E}|\zeta|^2]^{1/2}
\end{aligned}
$$

where we have used (19). Combining this inequality with (39) we complete the proof. □

Proof of Theorem 2. 1^0. First we show that Assumptions A, B, and K2 imply conditions (I) and (II) of Theorem 1.

Indeed, Assumption K2 ensures that sample paths of the processes $\{\tilde{Z}_\theta(x), (x, \theta) \in \mathcal{D}_0 \times \Theta\}$ and $\{\tilde{Z}_{\theta,\nu}(x), (x, \theta, \nu) \in \mathcal{D}_0 \times \Theta \times \Theta\}$ belong with probability one to the isotropic Hölder spaces $\mathbb{H}_{m+d}(\tau)$ and $\mathbb{H}_{2m+d}(\tau)$ with regularity index $0 < \tau < \gamma$ [25, Sect. 15]. Thus the condition (II) is fulfilled.

Moreover, together with Assumption B this implies that for any $F \in \mathbb{H}_d(M_0)$ sample paths of the process $\hat{F}_{\theta,\nu}(x) - \hat{F}_\nu(x)$ belong with probability one to the isotropic Hölder space $\mathbb{H}_{2m+d}(\tau')$ on $\mathcal{D}_0 \times \Theta \times \Theta$ with some regularity index $0 < \tau' < \gamma$. This, in turn, shows that for any $F \in \mathbb{H}_d(M_0)$ sample paths of the process

$$
\sup_{\nu \in \Theta} \|\hat{F}_{\theta,\nu} - \hat{F}_\nu\|_p
$$

belong to $\mathbb{H}_m(\tau')$ on Θ. Then condition (I) holds in view of Assumption A and Jennrich [18].

2^0. It follows from Lemma 6 in Appendix that for any $\varkappa \geq 1 + \sqrt{(2m+d)/\gamma}$

$$\mathbb{P}\left\{\sup_{\theta\in\Theta} \|\tilde{Z}_\theta(\cdot)\|_\infty \geq \varkappa\right\} + \mathbb{P}\left\{\sup_{(\theta,\nu)\in\Theta\times\Theta} \|\tilde{Z}_{\theta,\nu}(\cdot)\|_\infty \geq \varkappa\right\} \leq N^2[c_1 M(\mathcal{K})\bar{L}R\varkappa]^{(2m+d)/\gamma} \exp\{-\varkappa^2/2\},$$

where c_1 is an absolute constant. By definition of $\varkappa$ we obtain that

$$\exp\{\varkappa^2/2\} \leq N^2[c_1 M(\mathcal{K})\bar{L}R\varkappa]^{(2m+d)/\gamma}\delta_*^{-1} \tag{40}$$

which, in turn, implies

$$\begin{aligned}\varkappa &\leq \left[2\ln\delta_*^{-1} + 4\ln N + \frac{2(2m+d)}{\gamma}\ln C_{\mathcal{K}} + \frac{2m+d}{\gamma}(\ln\varkappa^2 + c_2)\right]^{1/2} \\ &\leq \sqrt{c_3 \ln\varepsilon^{-1}} =: \bar{\varkappa},\end{aligned} \tag{41}$$

where c_3 depends on $(2m+d)/\gamma$ only; here we have used (25).

Now we bound the remainder term in (22). It follows from Lemma 6 that for any $\lambda \geq 1 + \sqrt{(d+m)/\gamma}$ one has

$$\begin{aligned}\mathbb{E}|\zeta|^2 &= \int_0^\infty 2t\mathbb{P}(\zeta > t)\mathrm{d}t \leq 2\lambda + 2\int_\lambda^\infty tN[c_4\bar{L}Rt]^{(d+m)/\gamma} e^{-t^2/2}\mathrm{d}t \\ &\leq 2\lambda + 2N[c_4\bar{L}R]^{(d+m)/\gamma} e^{-\lambda^2/4}\int_0^\infty t^{1+(d+m)/\gamma} e^{-t^2/4}\mathrm{d}t.\end{aligned}$$

If we choose $\lambda = \sqrt{2}\bar{\varkappa}$ and apply (40), we get

$$\mathbb{E}|\zeta|^2 \leq 2\sqrt{2}\bar{\varkappa} + c_5 N^{-1}\delta_* \leq c_6 \ln\delta_*^{-1}.$$

Using (25) and the fact that $\sigma(\mathcal{K}) \geq c_7$ we finally obtain $r(\delta_*) \leq M_0[1 + M(\mathcal{K})]\varepsilon + c_8\varepsilon\sqrt{\ln\varepsilon^{-1}}$ which yields (26). □

Proof of Theorem 5. 1^0. In order to apply the result of Theorem 2 we have to verify Assumption K2 for the collection of kernels defined in (30). Recall that $\theta = (I, E, h)$, and in notation of Assumptions A and K2, $\theta = (\theta_1, \theta_2)$, where $\theta_1 = I \in \Theta_1 = \mathcal{I}$, and $\theta_2 = (E, h) \in \Theta_2 = \mathcal{E}_\eta \times [h_{\min}, h_{\max}]^d$.

We deduce from (30) and Assumption G(ii) that $K_\theta(t)$ is continuously differentiable in θ_2 and t, and

$$\sup_{\theta_2\in\Theta_2}\sup_{t\in\mathcal{D}} |\nabla_{\theta_2,t}\, K_\theta(t)| \leq \tilde{L}h_{\min}^{-3d},$$

where $\tilde{L}$ is an absolute constant depending only on d and $\|g\|_\infty$. Taking into account that $h_{\min} = \varepsilon^2$ we arrive to Assumption K2 with

$$\bar{L} = \tilde{L}\varepsilon^{-6d}, \quad \text{and} \quad \gamma = 1/2. \tag{42}$$

2^0. In view of (42), assumption (25) is verified.

3^0. Fix β and L and assume that $F \in \mathbb{F}(\beta, L)$. By definition of the class $F \in \mathbb{F}(\beta, L)$ there exist $I_* \in \mathcal{I}$ and $E_* \in \mathcal{E}_\eta$ such that $F \in \mathbb{F}_{I_*,E_*}(\beta, L)$. Let h_* be given by (33). Then from (26) and (34)

$$\begin{aligned}
\mathbb{E}_F\|\hat{F}_* - F\|_\infty &\le [3 + 2M(\mathcal{K})] \inf_{(I,E,h)\in\Theta} \left\{ \|B_{I,E,h}\|_\infty + C_1\varepsilon\sqrt{\ln \varepsilon^{-1}} \sup_x \sigma_{I,E,h}(x) \right\} \\
&\le [3 + 2M(\mathcal{K})] \left\{ \|B_{I_*,E_*,h_*}\|_\infty + C_1\varepsilon\sqrt{\ln \varepsilon^{-1}} \sup_x \sigma_{I_*,E_*,h_*}(x) \right\} \\
&\le 2[3 + 2M(\mathcal{K})](C_1 \vee 1) C L^{1/(2\beta+1)} \varphi_\varepsilon(\beta),
\end{aligned}$$

where C is the constant appearing in (34). □

Appendix

Proof of Lemma 2. Only the left hand side inequality should be proved. First we note that

$$\|f\|_p = \sup\left\{ \left| \int \phi f \right| : \ \|\phi\|_q = 1 \right\}$$

[7, p. 188]. Thus we have for $p < \infty$

$$\begin{aligned}
\mathbb{E}_F\|\tilde{F} - F\|_p &= \mathbb{E}_F\|B_S + \varepsilon Z_S\|_p \\
&= \mathbb{E}_F \sup_{g:\|g\|_q\le 1} \int [B_S(x) + \varepsilon Z_S(x)] g(x)\mathrm{d}x \\
&\ge \mathbb{E}_F \int [B_S(x) + \varepsilon Z_S(x)] g_*(x)\mathrm{d}x,
\end{aligned}$$

where $g_*(x) = \|B_S\|_p^{-p/q} |B_S(x)|^{p-1} \operatorname{sign}\{B_S(x)\}$. Therefore

$$\mathbb{E}_F\|B_S + \varepsilon Z_S\|_p \ge \int B_S(x) g_*(x)\mathrm{d}x + \mathbb{E}\int Z_S(x) g_*(x)\mathrm{d}x = \|B_S\|_p. \tag{43}$$

On the other hand, by the triangle inequality $\mathbb{E}_F\|B_S + \varepsilon Z_S\|_p \ge \varepsilon\mathbb{E}\|Z_S\|_p - \|B_S\|_p$. Multiplying the both side of (43) by 2 and summing up with the last inequality we obtain (28).

If $p = \infty$ then for any $x_0 \in \mathcal{D}_0$ one has $\mathbb{E}\|B_\theta + \varepsilon Z_\theta\|_\infty \ge \pm\mathbb{E}[B_\theta(x_0) + \varepsilon Z_\theta(x_0)] = \pm B_\theta(x_0)$, and therefore $\mathbb{E}\|B_\theta + \varepsilon Z_\theta\|_\infty \ge \|B_\theta\|_\infty$. □

Proof of Lemma 4. We will use the following notation: for any vector $t \in \mathbb{R}^d$, and partition $I = (I_1, \dots, I_{|I|})$ we will write $t_{(i)} = (t_j, j \in I_i)$. Throughout the proof without loss of generality we assume that E is the $d \times d$ identity matrix.

Using the fact that $F(t) = \sum_{i=1}^{|I|} f_i(E_i^T t)$ we have

$$\int K_\theta(t-x)F(t)\mathrm{d}t = \sum_{i=1}^{|I|}\sum_{j=1}^{|I|}\int G_{j,h}(t-x)f_i(t_{(i)})\mathrm{d}t - (|I|-1)\sum_{i=1}^{|I|}\int G_0(t-x)f_i(t_{(i)})\mathrm{d}t.$$

Note that for all $i = 1, \dots, |I|$

$$\int G_0(t-x)f_i(t_{(i)})\mathrm{d}t = \int\left[\prod_{j\in I_i} g(t_j - x_j)\right] f_i(t_{(i)})\mathrm{d}t_{(i)}$$

$$\int G_{i,h}(t-x)f_i(t_{(i)})\mathrm{d}t = \int\left[\prod_{j\in I_i}\frac{1}{h_j}g\left(\frac{t_j-x_j}{h_j}\right)\right] f_i(t_{(i)})\mathrm{d}t_{(i)}$$

$$\int G_{j,h}(t-x)f_i(t_{(i)})\mathrm{d}t = \int\left[\prod_{j\in I_i} g(t_j - x_j)\right] f_i(t_{(i)})\mathrm{d}t_{(i)}, \quad j \neq i.$$

Combining these equalities we obtain

$$\int K_\theta(t-x)F(t)\mathrm{d}t = \int\left[\prod_{j\in I_i}\frac{1}{h_j}g\left(\frac{t_j-x_j}{h_j}\right)\right] f_i(t_{(i)})\mathrm{d}t_{(i)},$$

and

$$\begin{aligned} B_\theta(x) &= \sum_{i=1}^{|I|}\int\left[\prod_{j\in I_i}\frac{1}{h_j}g\left(\frac{t_j-x_j}{h_j}\right)\right][f_i(t_{(i)}) - f_i(x_{(i)})]\mathrm{d}t_{(i)} \\ &= \sum_{i=1}^{|I|}\int\left[\prod_{j\in I_i}\frac{1}{h_j}g\left(\frac{t_j-x_j}{h_j}\right)\right] \\ &\quad\times\left[f_i(t_{(i)}) - f_i(x_{(i)}) - \sum_{s=1}^{l_i}\frac{1}{s!}\sum_{|\boldsymbol{k}|=s} D^{\boldsymbol{k}} f_i(x_{(i)})(t_{(i)} - x_{(i)})^{\boldsymbol{k}}\right]\mathrm{d}t_{(i)}, \end{aligned}$$

where the last equality follows from the fact that

$$\int \prod_{j\in I_i} \frac{1}{h_j} g\Big(\frac{t_j - x_j}{h_j}\Big)(t_{(i)} - x_{(i)})^{\boldsymbol{k}} \mathrm{d}t_{(i)} = 0, \quad \forall |\boldsymbol{k}| : |\boldsymbol{k}| = 1, \dots, l_i, \ \ i = 1, \dots, |I|,$$

see Assumption G(i). Because $f_i \in H_{|I_i|}(\beta_i, L_i)$, we obtain

$$|B_\theta(x)| \le \sum_{i=1}^{|I|} L_i \int \left| \prod_{j\in I_i} \frac{1}{h_j} g\left(\frac{t_j - x_j}{h_j}\right) \right| |t_{(i)} - x_{(i)}|^{\beta_i} dt_{(i)} \le \sum_{i=1}^{|I|} L_i \|g\|_1^{|I_i|} \sum_{j\in I_i} h_j^{\beta_i}.$$

as claimed. □

We quote the following result from Talagrand [32] that is repeatedly used in the proof of Lemma 6 below.

Lemma 5 *Consider a centered Gaussian process $(X_t)_{t\in T}$. Let $\sigma^2 = \sup_{t\in T} EX_t^2$. Consider the intrinsic semi-metric ρ_X on T given by $\rho_X^2(s,t) = \mathbb{E}(X_s - X_t)^2$. Assume that for some constant $A > \sigma$, some $v > 0$ and some $0 \le \varepsilon_0 \le \sigma$ we have*

$$\varepsilon < \varepsilon_0 \ \Rightarrow \ N(T, \rho_X, \varepsilon) \le \left(\frac{A}{\varepsilon}\right)^v,$$

where $N(T, \rho_X, \varepsilon)$ is the smallest number of balls of radius ε needed to cover T. Then for $u \ge \sigma^2[(1 + \sqrt{v})/\varepsilon_0]$ we have

$$\mathbb{P}\left(\sup_{t\in T} X_t \ge u\right) \le \left(\frac{KAu}{\sqrt{v}\sigma^2}\right)^v \Phi\left(\frac{u}{\sigma}\right),$$

where K is universal constant, and $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_u^\infty e^{-s^2/2} \mathrm{d}s$.

Lemma 6 *Let Assumptions A, K0 and K2 hold. Then for any $\varkappa \ge 1 + \sqrt{\frac{d+m}{\gamma}}$ one has*

$$\mathbb{P}\left\{\sup_{\theta\in\Theta} \|\tilde{Z}_\theta(\cdot)\|_\infty \ge \varkappa\right\} \le N[C_1 \bar{L} R \varkappa]^{(d+m)/\gamma} \exp\{-\varkappa^2/2\}, \tag{44}$$

where C_1 is an absolute constant.

Furthermore, for any $\varkappa \ge 1 + \sqrt{\frac{d+2m}{\gamma}}$ one has

$$\mathbb{P}\left\{\sup_{(\theta,\nu)\in\Theta\times\Theta} \|\tilde{Z}_{\theta,\nu}(\cdot)\|_\infty \ge \varkappa\right\} \le N^2[C_2 M(\mathcal{K}) \bar{L} R \varkappa]^{(d+2m)/\gamma} \exp\{-\varkappa^2/2\}, \tag{45}$$

where C_2 is an absolute constant.

Proof 1^0. First we prove (44). Recall our notation:

$$Z_\theta(x) = \int K_\theta(t,x)W(\mathrm{d}t), \quad \sigma_\theta(x) = \|K_\theta(\cdot,x)\|_2, \quad \tilde{Z}_\theta(x) = \sigma_\theta^{-1}(x)Z_\theta(x).$$

By Assumption A, $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$. Because the set Θ_1 is finite, throughout the proof we keep $\theta_1 \in \Theta_1$ fixed. For brevity, we will write $\theta = (\theta_1, \theta_2)$, $\theta' = (\theta_1, \theta_2')$, $u = (x, \theta_2)$, $u' = (x', \theta_2')$. Also with a slight abuse of notation we write $Z(u)$, $\tilde{Z}(u)$ and $\sigma(u)$ for $Z_\theta(x)$, $\tilde{Z}_\theta(x)$ and $\sigma_\theta(x)$ respectively. The same notation with u replaced by u' will be used for the corresponding quantities depending on u'.

Consider the random process $\{Z(u), u \in U\}$, $U := \mathcal{D}_0 \times \Theta_2$. Clearly, it has zero mean and variance $\mathbb{E}Z^2(u) = \sigma^2(u)$. Let ρ_Z denote the intrinsic semi-metric of $\{Z(u), u \in U\}$; then

$$\begin{aligned}\rho_Z(u,u') &:= [\mathbb{E}|Z(u) - Z(u')|^2]^{1/2} \\ &= \|K_{(\theta_1,\theta_2)}(\cdot,x) - K_{(\theta_1,\theta_2')}(\cdot,x')\|_2 \\ &\le \bar{L}|u-u'|^\gamma,\end{aligned}$$

where the last inequality follows from Assumption K2.

Now consider the random process $\{\tilde{Z}(u), u \in U\}$. Let $\underline{\sigma} = \inf_{u\in U}\sigma(u)$; then

$$\begin{aligned}\rho_{\tilde{Z}}(u,u') &:= \left[\mathbb{E}|\tilde{Z}(u) - \tilde{Z}(u')|^2\right]^{1/2} \\ &= \left[\mathbb{E}\left|\frac{Z(u)}{\sigma(u)} - \frac{Z(u')}{\sigma(u')}\right|^2\right]^{1/2} \\ &\le \frac{1}{\sigma(u)}\rho_Z(u,u') + \sigma(u')\left|\frac{1}{\sigma(u)} - \frac{1}{\sigma(u')}\right| \\ &\le \underline{\sigma}^{-1}\left[\rho_Z(u,u') + |\sigma(u) - \sigma(u')|\right] \\ &\le 2\underline{\sigma}^{-1}\rho_Z(u,u') \le 2(\mathrm{mes}\{\mathcal{D}\})^{1/2}\bar{L}|u-u'|^\gamma. \end{aligned} \tag{46}$$

Here we have taken into account that $\underline{\sigma} \ge (\mathrm{mes}\{\mathcal{D}\})^{-1/2}$, and

$$\begin{aligned}|\sigma(u) - \sigma(u')| &= |\,\|K_\theta(\cdot,x)\|_2 - \|K_{\theta'}(\cdot,x')\|_2\,| \\ &\le \|K_\theta(\cdot,x) - K_{\theta'}(\cdot,x')\|_2 = \rho_Z(u,u').\end{aligned}$$

It follows from (46) that the covering number $N(U, \rho_{\tilde{Z}}, \eta)$ of the index set $U = \mathcal{D}_0 \times \Theta_2$ with respect to the intrinsic semi-metric $\rho_{\tilde{Z}}$ does not exceed $[c_1\bar{L}R\eta^{-1}]^{(d+m)/\gamma}$, where c_1 is an absolute constant. Then using the exponential inequality of Lemma 5 [with $v = (d+m)/\gamma$, $A = c_1\bar{L}R$ and $\sigma = \varepsilon_0 = 1$], and summing over all $\theta_1 \in \Theta_1$ we obtain (44).

2^0. Now we turn to the proof of (45). We recall that

$$Z_{\theta,\nu}(x) - Z_\nu(x) = \int \big[K_{\theta,\nu}(t,x) - K_\nu(t,x)\big]\, W(\mathrm{d}t),$$
$$\sigma_{\theta,\nu}(x) = \|K_{\theta,\nu}(\cdot,x) - K_\nu(\cdot,x)\|_2,$$

where $K_{\theta,\nu}(\cdot,\cdot)$ is defined in (10). We keep $\theta_1, \nu_1 \in \Theta_1$ fixed, and denote $\theta = (\theta_1, \theta_2)$, $\theta' = (\theta_1, \theta_2')$, $\nu = (\nu_1, \nu_2)$, $\nu' = (\nu_1, \nu_2')$. We also denote $V = \mathcal{D}_0 \times \Theta_2 \times \Theta_2$, $v = (\theta, \nu, x)$, $v'(\theta', \nu', x')$, and consider the Gaussian random processes $\{\zeta(v),\, v \in V\}$ and $\{\tilde\zeta(v),\, v \in V\}$, where

$$\zeta(v) = Z_{\theta,\nu}(x) - Z_\nu(x), \quad \tilde\zeta(v) = \tilde\sigma_{\theta,\nu}^{-1}(x)[Z_{\theta,\nu}(x) - Z_\nu(x)].$$

Let ρ_ζ and $\rho_{\tilde\zeta}$ be the intrinsic semi-metrics of these processes. Similarly to (46), it is straightforward to show that $\rho_{\tilde\zeta}(v, v') \le 2\rho_\zeta(v, v')$, and our current goal is to bound $\rho_\zeta(v, v')$ from above.

We have

$$\begin{aligned}
\rho_\zeta(v,v') &= \Big[\mathbb{E}|\zeta(v) - \zeta(v')|^2\Big]^{1/2} \\
&= \|K_{\theta,\nu}(\cdot,x) - K_\nu(\cdot,x) - K_{\theta',\nu'}(\cdot,x') + K_{\nu'}(\cdot,x')\|_2 \\
&\le \|K_\nu(\cdot,x) - K_{\nu'}(\cdot,x')\|_2 + \|K_{\theta,\nu}(\cdot,x) - K_{\theta',\nu'}(\cdot,x')\|_2 \\
&\le \|K_\nu(\cdot,x) - K_{\nu'}(\cdot,x')\|_2 + \|K_{\theta,\nu}(\cdot,x) - K_{\theta,\nu'}(\cdot,x')\|_2 \\
&\quad + \|K_{\theta,\nu'}(\cdot,x') - K_{\theta',\nu'}(\cdot,x')\|_2 =: J_1 + J_2 + J_3.
\end{aligned}$$

By Assumption K2

$$J_1 \le \bar L |v - v'|^\gamma.$$

In order to bound J_2 we recall that

$$K_{\theta,\nu}(t,x) - K_{\theta,\nu'}(t,x') = \int K_\theta(t,y)\big[K_\nu(y,x) - K_{\nu'}(y,x')\big]\,\mathrm{d}y.$$

Then applying the general theorem about the boundedness of integral operators on $\mathbb{L}_p$-spaces (see, e.g., [7, Theorem 6.18]) and using (8) we obtain that

$$J_2 \le M(\mathcal{K})\|K_\nu(\cdot,x) - K_{\nu'}(\cdot,x')\|_2 \le M(\mathcal{K})\bar L|v - v'|^\gamma,$$

where the last inequality follows from Assumption K2. It is shown similarly that $J_3 \le M(\mathcal{K})\bar L|v - v'|^\gamma$. Combining upper bounds for J_1, J_2 and J_3 we get $\rho_\zeta(v, v') \le [1 + 2M(\mathcal{K})]\bar L|v - v'|^\gamma$, and finally

$$\rho_{\tilde\zeta}(v, v') \le 2[1 + 2M(\mathcal{K})]\bar L|v - v'|^\gamma. \tag{47}$$

It follows from (47) that the covering number $N(V, \rho_{\tilde{\zeta}}, \eta)$ of the index set $V = \mathcal{D}_0 \times \Theta_2 \times \Theta_2$ with respect to the intrinsic semi-metric $\rho_{\tilde{\zeta}}$ does not exceed $[c_2 M(\mathcal{K}) \bar{L} R \eta^{-1}]^{(d+2m)/\gamma}$, where c_2 is an absolute constant. Then noting that $\sup_v \mathrm{var}(\tilde{\zeta}(v)) \le 1$, using the exponential inequality of Lemma 5 [with $v = (d+2m)/\gamma$, $A = c_2 M(\mathcal{K}) \bar{L} R$ and $\sigma = \varepsilon_0 = 1$], and summing over all $(\theta_1, \nu_1) \in \Theta_1 \times \Theta_1$ we obtain (45).

References

1. Barron, A., Birgé, L., Massart, P.: Risk bounds for model selection via penalization. Probab. Theory Relat. Fields **113**, 301–413 (1999)
2. Belomestny, D., Spokoiny, V.: Local likelihood modeling via stagewise aggregation. WIAS preprint No. 1000 (2004). www.wias-berlin.de
3. Bertin, K.: Asymptotically exact minimax estimation in sup-norm for anisotropic Hölder balls. Bernoulli **10**, 873–888 (2004)
4. Cavalier, L., Golubev, G.K., Picard, D., Tsybakov, A.B.: Oracle inequalities for inverse problems. Ann. Stat. **30**, 843–874 (2002)
5. Chen, H.: Estimation of a projection-pursuit type regression model. Ann. Stat. **19**, 142–157 (1991)
6. Devroye, L., Lugosi, G.: Combinatorial Methods in Density Estimation. Springer, New York (2001)
7. Folland, G.B.: Real Analysis, 2nd edn. Wiley, New York (1999)
8. Goldenshluger, A., Nemirovski, A.: On spatially adaptive estimation of nonparametric regression. Math. Methods Stat. **6**, 135–170 (1997)
9. Golubev, G.K.: Asymptotically minimax estimation of a regression function in an additive model. Probl. Inform. Transm. **28**, 101–112 (1992)
10. Golubev, G.K.: The method of risk envelopes in the estimation of linear functionals (Russian). Probl. Inform. Transm. **40**, 53–65 (2004)
11. Györfi, L., Kohler, M., Krzyzak, A., Walk, H.: A Distribution-Free Theory of Nonparametric Regression. Springer, New York (2002)
12. Hall, P.: On projection-pursuit regression. Ann. Stat. **17**, 573–588 (1989)
13. Hristache, M., Juditsky, A., Spokoiny, V.: Direct estimation of the index coefficient in a single-index model. Ann. Stat. **29**, 595–623 (2001)
14. Hristache, M., Juditsky, A., Polzehl, J., Spokoiny, V.: Structure adaptive approach for dimension reduction. Ann. Stat. **29**, 1537–1566 (2001)
15. Huber, P.: Projection pursuit. With discussion. Ann. Stat. **13**, 435–525 (1985)
16. Ibragimov, I.A., Khasminskii, R.Z.: Bounds for the quality of nonparametric estimation of regression. Theory Probab. Appl. **27**, 81–94 (1982)
17. Ibragimov, I.A.: Estimation of multivariate regression. Theory Probab. Appl. **48**, 256–272 (2004)
18. Jennrich, R.: Asymptotic properties of non-linear least squares estimators. Ann. Math. Stat. **40**, 633–643 (1969)
19. Johnstone, I.M.: Oracle inequalities and nonparametric function estimation. In: Proceedings of the International Congress of Mathematicians, Vol. III (Berlin, 1998). Doc. Math., Extra vol. III, pp. 267–278 (1998)
20. Juditsky, A., Lepski, O., Tsybakov, A.: Statistical estimation of composite functions. Manuscript (2006)
21. Kerkyacharian, G., Lepski, O., Picard, D.: Nonlinear estimation in anisotropic multi-index denoising. Probab. Theory Relat. Fields **121**, 137–170 (2001)
22. Lepski, O.V., Levit, B.Y.: Adaptive nonparametric estimation of smooth multivariate functions. Math. Methods Stat. **8**, 344–370 (1999)
23. Lepski, O., Mammen, E., Spokoiny, V.: Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimators with variable bandwidth selectors. Ann. Stat. **25**, 929–947 (1997)
24. Lepski, O.V., Spokoiny, V.G.: Optimal pointwise adaptive methods in nonparametric estimation. Ann. Stat. **25**, 2512–2546 (1997)
25. Lifshits, M.: Gaussian Random Functions. Kluwer, Dordrecht (1995)
26. Nemirovski, A.S.: Nonparametric estimation of smooth regression functions. Soviet J. Comput. Syst. Sci. **23**(6), 1–11 (1985); translated from Izv. Akad. Nauk SSSR Tekhn. Kibernet. 235 (3), 50–60 (1985)(Russian)

27. Nemirovski, A.: Topics in Non-parametric Statistics. Lectures on probability theory and statistics (Saint-Flour, 1998), Lecture Notes in Mathematics, vol. 1738, pp. 85–277. Springer, Berlin (2000)
28. Nicoleris, T., Yatracos, Y.: Rates of convergence of estimators, Kolmogorov's entropy and the dimensionality reduction principle in regression. Ann. Stat. **25**, 2493–2511 (1997)
29. Nussbaum, M.: Nonparametric estimation of a regression function that is smooth in a domain in $\mathbb{R}^k$. Theory Probab. Appl. **31**, 108–115 (1987)
30. Stone, C.J.: Optimal global rates of convergence for nonparametric regression. Ann. Stat. **10**, 1040–1053 (1982)
31. Stone, C.J.: Additive regression and other nonparametric models. Ann. Stat. **13**, 689–705 (1985)
32. Talagrand, M.: Sharper bounds for Gaussian and empirical processes. Ann. Probab. **22**, 28–76 (1994)
33. Tsybakov, A.: Optimal rates of aggregation. In: Scholkopf, B., Warmuth, M. (eds) Computational Learning Theory and Kernel Machines. Lectures Notes in Artificial Intelligence, vol. 2777, pp. 303–313. Springer, Heidelberg (2003)