Theory of Adaptive Estimation

Oleg V Lepski

Theory of adaptive estimation

O. V. Lepski *

Institut de Mathématique de Marseille
Aix-Marseille Université, CNRS
39, rue F. Joliot-Curie
13453 Marseille, France
e-mail: oleg.lepski@univ-amu.fr

1. Introduction.

Let $\big(V^{(n)}, \mathfrak{A}^{(n)}, \mathbb{P}_f^{(n)}, f \in \mathfrak{F}\big)$, $n \in \mathbb{N}^*$, be a family of statistical experiments generated by observation $X^{(n)}$. It means that $X^{(n)}$ is a $V^{(n)}$-valued random variable defined on some probability space, and the probability law of $X^{(n)}$ belongs to the family $\big(\mathbb{P}_f^{(n)}, f \in \mathfrak{F}\big)$. Since the probability space on which $X^{(n)}$ is defined will play no role in the sequel we will just assume its existence.

Furthermore in this paper:

- $(\mathcal{D}, \mathfrak{D}, \mu)$ is a measurable space;
- $\mathfrak{F}$ is a set of functions $f : \mathcal{D} \to \mathbb{R}$. Typical examples of set $\mathfrak{F}$ are functional spaces, e.g., $\mathfrak{F} = \mathbb{L}_2(\mathbb{R}^d)$, $\mathbb{C}_b(\mathbb{R}^d)$, the set of all measurable real functions etc;
- $G : \mathfrak{F} \to \mathfrak{S}$, where $\mathfrak{S}$ is a set endowed with semi-metric ℓ.

The goal is to estimate $G(f), f \in \mathfrak{F}$, from observation $X^{(n)}$. By an estimator we mean any $X^{(n)}$-measurable $\mathfrak{S}$-valued mapping. Accuracy of an estimator $\tilde{G}$ is measured by the ℓ-risk

$$\mathcal{R}_n^{(\ell)}\big[\tilde{G}; G(f)\big] = \Big(\mathbb{E}_f^{(n)} \ \big[\ell\big(\tilde{G}, G(f)\big)\big]^q\Big)^{\frac{1}{q}}. \tag{1.1}$$

Here and later $\mathbb{E}_f^{(n)}$ denotes the mathematical expectation with respect to the probability measure $\mathbb{P}_f^{(n)}$ and the number $q \geq 1$ is supposed to be fixed. Recall that for any $X^{(n)}$-measurable map $T : V^{(n)} \to \mathbb{R}$

$$\mathbb{E}_f^{(n)}\big[T\big] = \int_{V^{(n)}} T(v)\mathbb{P}_f^{(n)}(\mathrm{d}v)$$

1.1. Examples of models.

In these notes we will consider the following statistical models.

Density Model. Let $\mathfrak{P}(\mathcal{D}, \mu)$ denote the set of all probability densities with respect to measure μ defined on $\mathcal{D}$ and let $\mathfrak{F} \subseteq \mathfrak{P}(\mathcal{D}, \mu)$.
Then the statistical experiment is generated by the observation $X^{(n)} = \big(X_1, \ldots, X_n\big), n \in \mathbb{N}^*$, where $X_i,\ i \in \mathbb{N}^*$, are i.i.d. random vectors possessing unknown density $f \in \mathfrak{F}$.

*This work has been carried out in the framework of the Labex Archimède (ANR-11-LABX-0033) and of the A*MIDEX project (ANR-11-IDEX-0001-02), funded by the "Investissements d'Avenir" French Government program managed by the French National Research Agency (ANR).

White Gaussian Noise Model. Let $\mathfrak{F} = \mathbb{L}_2\big(\mathcal{D}, \mu\big)$. Put $\widetilde{\mathfrak{D}} = \big\{B \in \mathfrak{D} : \ \mu(B) < \infty\big\}$ and let $\big(W(B),\ B \in \widetilde{\mathfrak{D}}\big)$ be the white noise with intensity μ.
Consider the statistical model generated by the observation $X^{(n)} = \big\{X_n(g),\ g \in \mathbb{L}_2\big(\mathcal{D}, \mu\big)\big\}$ where

$$X_n(g) = \int_{\mathcal{D}} f(t)g(t)\mu(\mathrm{d}t) + n^{-1/2}\int_{\mathcal{D}} g(t)W(\mathrm{d}t). \tag{1.2}$$

Recall also that for any $g \in \mathbb{L}_2\big(\mathcal{D}, \mu\big)$

$$X_n(g) \quad \backsim \quad \mathcal{N}\Big(\langle g, f\rangle, n^{-1}\langle g, g\rangle\Big), \tag{1.3}$$

where $\langle\cdot,\cdot\rangle$ is the inner product of $\mathbb{L}_2\big(\mathcal{D}, \mu\big)$, and $\mathcal{N}(\cdot,\cdot)$ denotes the normal law on $\mathbb{R}$.

1.2. Examples of estimation targets G.

Global estimation $G(f) = f$**.** The goal is to estimate the entire function f. Here $\mathfrak{S} = \mathfrak{F}$, and the accuracy of estimation is usually measured by the $\mathbb{L}_p$-risk on $D \subseteq \mathcal{D}$, i.e. $\ell(g_1, g_2) = \|g_1 - g_2\|_{p,D},\ 1 \le p \le \infty$, where

$$\|g\|^p_{p,D} = \int_D |g|^p\mu(\mathrm{d}t),\ p \in [1, \infty), \quad \|g\|_{\infty,D} = \sup_{t\in D}|g(t)|.$$

Pointwise estimation $G(f) = f(t_0),\ t_0 \in D$**.** Here $\mathfrak{S} = \mathbb{R}^1$ and $\ell(a, b) = |a - b|, a, b \in \mathbb{R}$, and $D \subseteq \mathcal{D}$. We present this estimation problem separately from the discussed below problems of estimation of functionals because it is often used in order to recover the underlying function itself.

Estimation of functionals. Here $\mathfrak{S} = \mathbb{R}^1$ and $\ell(a, b) = |a - b|, a, b \in \mathbb{R}$ and $D \subseteq \mathcal{D}$.

- Estimation of a derivative at a given point: $G(f) = f^{(k)}(t_0),\ t_0 \in D, k \in \mathbb{N}^*$;
- Estimation of norms: $G(f) = \|f\|_{p,D}, 1 \le p \le \infty$;
- Estimation of extreme points: $G(f) = \arg\max_{t\in D} f(t)$;
- Estimation of regular functionals, for example $G(f) = \int_D f^s(t)\mathrm{d}t, s \in \mathbb{N}^*$.

2. Minimax adaptive estimation

Let $\mathbb{F}$ be a given subset of $\mathfrak{F}$. For any estimator $\tilde{G}_n$ define its *maximal risk* on $\mathbb{F}$ by

$$\mathcal{R}^{(\ell)}_n\big[\tilde{G}_n; \mathbb{F}\big] = \sup_{f\in\mathbb{F}} \mathcal{R}^{(\ell)}_n\big[\tilde{G}_n; G(f)\big]$$

and the *minimax risk* on $\mathbb{F}$ is given by

$$\phi_n(\mathbb{F}) := \inf_{\tilde{G}_n} \mathcal{R}^{(\ell)}_n\big[\tilde{G}_n; \mathbb{F}\big], \tag{2.1}$$

where infimum is always taken over all possible estimators. An estimator whose maximal risk is proportional to $\phi_n(\mathbb{F})$ is called minimax on $\mathbb{F}$.

Let $\big\{\mathbb{F}_\vartheta, \vartheta \in \Theta\big\}$ be the collection of subsets of $\mathfrak{F}$, where ϑ is a nuisance parameter which may have very complicated structure (see examples below). Without further mentioning we will

consider only scales of functional classes for which a minimax on $\mathbb{F}_\vartheta$ estimator (usually depending on ϑ) exists for any $\vartheta \in \Theta$.

The problem of adaptive estimation can be formulated as follows: *is it possible to construct a single estimator $\hat{G}_n$ which is simultaneously minimax on each class $\mathbb{F}_\vartheta$, $\vartheta \in \Theta$, i.e. such that*

$$\limsup_{n\to\infty} \phi_n^{-1}(\mathbb{F}_\vartheta)\mathcal{R}_n^{(\ell)}\big[\hat{G}_n; \mathbb{F}_\vartheta\big] < \infty, \quad \forall \vartheta \in \Theta?$$

We refer to this question as *the problem of minimax adaptive estimation over the scale of classes* $\{\mathbb{F}_\vartheta,\ \vartheta \in \Theta\}$. If such estimator exists we will call it optimally-adaptive or rate-adaptive.

The first adaptive results were obtained in (11). Starting from this pioneering paper a variety of adaptive methods were proposed in different statistical models such as density and spectral density estimation, nonparametric regression, deconvolution model, inverse problems and many others. The interested reader can find a very detailed overview of this topic in (32). Here we only mention several methods allowing to construct optimally-adaptive estimators.

- Extension of Efroimovich-Pinsker method, (12; 14);
- Lepski method (27) and its extension Goldenshluger-Lepski method (18);
- Unbiased risk minimization, (20; 21);
- Wavelet thresholding, (10);
- Model selection, (1),(2);
- Aggregation of estimators, (37), (23), (43), (42), (3), (15);
- Exponential weights, (36), (9), (40);
- Risk hull method, (8);
- Blockwise Stein method, (4), (7), (39).

We will discuss existence of optimally-adaptive estimators in details later. Now let us provide some example of scales of functional classes over which the adaptation is studied.

2.1. Scales of functional classes.

2.1.1. Classes of smooth functions.

Let $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_d)$ denote the canonical basis of $\mathbb{R}^d$, $d \in \mathbb{N}^*$. For a function $T : \mathbb{R}^d \to \mathbb{R}^1$ and real number $u \in \mathbb{R}$ the first order difference operator with step size u in the direction of the variable x_j is defined by $\Delta_{u,j}T(x) = T(x + u\boldsymbol{e}_j) - T(x),\ j = 1, \dots, d$. By induction, the k-th order difference operator is

$$\Delta_{u,j}^k T(x) = \Delta_{u,j}\Delta_{u,j}^{k-1}T(x) = \textstyle\sum_{l=1}^k (-1)^{l+k}\binom{k}{l}\Delta_{ul,j}T(x).$$

Definition 1. *For given vectors $\vec{\beta} = (\beta_1, \dots, \beta_d) \in (0,\infty)^d$, $\vec{r} = (r_1, \dots, r_d) \in [1,\infty]^d$, and $\vec{L} = (L_1, \dots, L_d) \in (0,\infty)^d$ a function $T : \mathbb{R}^d \to \mathbb{R}^1$ is said to belong to anisotropic Nikolskii's class $\mathbb{N}_{\vec{r},d}\big(\vec{\beta}, \vec{L}\big)$ if $\|T\|_{r_j} \le L_j$ for all $j = 1, \dots, d$, and there exist natural numbers $k_j > \beta_j$ such that*

$$\big\|\Delta_{u,j}^{k_j} T\big\|_{r_j} \le L_j|u|^{\beta_j}, \quad \forall u \in \mathbb{R},\ \ \forall j = 1, \dots, d.$$

Let $\mathfrak{F} = \cup_{q\ge 1}^d \mathbb{L}_q(\mathbb{R}^d)$ and

$$\mathbb{F}_\vartheta = \mathbb{N}_{\vec{r},d}\big(\vec{\beta}, \vec{L}\big),\ \vartheta = \big(\vec{\beta}, \vec{r}, \vec{L}\big) \in \Theta \subseteq (0,\infty)^d \times [1,\infty]^d \times (0,\infty)^d,$$

where $\mathbb{N}_{\vec{r},d}\big(\vec{\beta}, \vec{L}\big)$ is anisotropic Nikolskii's class of functions on $\mathbb{R}^d, d \ge 1$.

2.1.2. Functional classes with structure.

Structural models are usually used in estimation of multivariate functions in order to improve estimation accuracy and to overcome the curse of the dimensionality.

Single index structure. Let $\mathfrak{F}=\cup_{q\geq 1}^{d}\mathbb{L}_q(\mathbb{R}^d)$ and let $\mathbb{S}^{d-1}$, $d\geq 2$, denote the unit sphere in $\mathbb{R}^d$. Let also $\mathbb{N}_{r,1}\big(\beta,L\big)$, $r\geq 1,\beta>0$, $L>0$ be Nikolskii's class of functions on $\mathbb{R}^1$.
For any $\mathcal{S}\subseteq\mathbb{S}^{d-1}$ and any $r\geq 1,\beta>0$, $L>0$ introduce the following functional class

$$\mathcal{F}_r^{\text{single}}(\beta,L,\mathcal{S})=\Big\{f:\mathbb{R}^d\to\mathbb{R}^1:\ \ f(\cdot)=F\big(\omega^{\intercal}\,\cdot\,\big),\ F\in\mathbb{N}_{r,1}\big(\beta,L\big),\ \omega\in\mathcal{S}\Big\}.$$

The adaptive estimation over the collection

$$\mathbb{F}_{\vartheta}=\mathcal{F}_r^{\text{single}}(\beta,L,\mathcal{S}),\ \vartheta=\big(\beta,r,L,\mathcal{S}\big)\in\Theta\subseteq(0,\infty)\times[1,\infty]\times(0,\infty)\times\mathbb{S}^{d-1}$$

is called the estimation under the single index constraint.

Additive structure. Let as previously $\mathfrak{F}=\cup_{q\geq 1}^{d}\mathbb{L}_q(\mathbb{R}^d)$, $d\geq 2$, and let $\mathbb{N}_{r,1}\big(\beta,L\big)$, $r\geq 1,\beta>0$, $L>0$ denote Nikolskii's class of functions on $\mathbb{R}^1$.
For any $r\geq 1,\beta>0$, $L>0$ introduce the following functional class

$$\mathcal{F}_r^{\text{additive}}(\beta,L,\mathcal{S})=\Big\{f:\mathbb{R}^d\to\mathbb{R}^1:\ \ f(x)=\sum_{k=1}^{d}F_k\big(x_k\big),\ F_k\in\mathbb{N}_{r,1}\big(\beta,L\big)\Big\}.$$

The adaptive estimation over the collection

$$\mathbb{F}_{\vartheta}=\mathcal{F}_r^{\text{additive}}(\beta,L),\ \vartheta=\big(\beta,r,L\big)\in\Theta\subseteq(0,\infty)\times[1,\infty]\times(0,\infty)$$

is called the estimation under the additive constraint.

The functional classes introduced above are considered in the framework of Gaussian White Noise Model or more generally in nonparametric regression context.

Hypothesis of independence. The functional classes introduced below are used in the Density Model. Let $\mathcal{D}=\mathbb{R}^d, d\geq 2$, μ be the Lebesgue measure and recall that $\mathfrak{F}\subseteq\mathfrak{P}(\mathcal{D},\mu)$. At last, let $\mathcal{I}_d$ be the set of all subsets of $\{1,\dots,d\}$.
For any $I\in\mathcal{I}_d$ and any $x\in\mathbb{R}^d$ denote $x_I=\{x_i\in\mathbb{R},j\in I\}$, $\bar{I}=\{1,\dots,d\}\setminus I$, and set for any density $f\in\mathfrak{F}$

$$f_I(x_I)=\int_{\mathbb{R}^{\bar{I}}}f(x)\mathrm{d}x_{\bar{I}},\quad x_I\in\mathbb{R}^{|I|}.$$

If we denote the coordinates of the random vector X_i by $X_{i,1},\dots X_{i,d}$ we can assert that f_I is the marginal density of the random vector $X_{i,I}:=(X_{i,j},j\in I)$ whatever $i=1,\dots,n$. The latter is true because $X_i, i=1,\dots,n$, are identically distributed.

Let Π denote the set of all partitions of $\{1,\dots,d\}$. The independence hypothesis supposes that there exists a partition $\mathcal{P}$ such that the random vectors $X_{1,I}, I\in\mathcal{P}$, are mutually independent that means that

$$f(x)=\prod_{I\in\mathcal{P}}f_I(x_I),\quad\forall x\in\mathbb{R}^d.$$

For given vectors $\vec{\beta}=(\beta_1,\dots,\beta_d)\in(0,\infty)^d$, $\vec{r}=(r_1,\ \dots,r_d)\in[1,\infty]^d$, $\vec{L}=(L_1,\dots,L_d)\in(0,\infty)^d$ and a given partition $\mathcal{P}\in\Pi$ introduce the following functional class

$$\mathcal{F}_{\vec{r}}^{\text{indep}}\big(\vec{\beta},\vec{L},\mathcal{P}\big)=\Big\{f:\mathbb{R}^d\to\mathbb{R}_+:\ \ f(x)=\prod_{I\in\mathcal{P}}f_I(x_I),\ f_I\in\mathbb{N}_{r_I,|I|}\big(\beta_I,L_I\big),I\in\mathcal{P}\Big\}.$$

The adaptive estimation over the collection

$$\mathbb{F}_\vartheta = \mathcal{F}_{\vec{r}}^{\mathrm{indep}}\big(\vec{\beta}, \vec{L}, \mathcal{P}\big),\ \vartheta = \big(\vec{\beta}, \vec{r}, \vec{L}\big) \in \Theta \subseteq (0,\infty)^d \times [1,\infty]^d \times (0,\infty)^d \times \Pi$$

is called the estimation under hypothesis of independence.

2.2. Existence of adaptive estimators. Fundamental problem.

It is well-known that optimally-adaptive estimators do not always exist, see (26), (28), (13), (5). Formally nonexistence of optimally-adaptive estimator means that

$$\liminf_{n\to\infty} \inf_{\widetilde{G}_n} \sup_{\vartheta\in\{\vartheta_1,\vartheta_2\}} \phi_n^{-1}(\mathbb{F}_\vartheta)\mathcal{R}_n^{(\ell)}\big[\widetilde{G}_n; \mathbb{F}_\vartheta\big] = \infty, \quad \forall \vartheta_1, \vartheta_2 \in \Theta. \tag{2.2}$$

Indeed, since a minimax estimator on $\mathbb{F}_\vartheta$ exists for any $\vartheta \in \Theta$ we can assert that

$$0 < \liminf_{n\to\infty} \inf_{\widetilde{G}_n} \phi_n^{-1}(\mathbb{F}_\vartheta)\mathcal{R}_n^{(\ell)}\big[\widetilde{G}_n; \mathbb{F}_\vartheta\big] < \infty, \quad \forall \vartheta \in \Theta.$$

The latter result means that the optimal (from the minimax point of view) family of normalizations $\{\phi_n\big(\mathbb{F}_\vartheta\big), \vartheta \in \Theta\big\}$ is attainable for each value ϑ, while (2.2) shows that this family is unattainable by any estimation procedure simultaneously for any couple of elements from Θ. This, in its turn, implies that optimally-adaptive over the scale $\{\mathbb{F}_\vartheta, \vartheta \in \Theta\big\}$ does not exist.

However, the question of constructing a single estimator for all values of the nuisance parameter $\vartheta \in \Theta$ remains relevant. Hence, if (2.2) holds we need to find an attainable family of normalization and to prove its optimality. The realization of this program dates back to (27) where the notion of *adaptive rate of convergence* was introduced. Nowadays there exist several definitions of adaptive rate of convergence and corresponding to this notion criteria of optimality, see (27), (41), (25), (38). Here we present the simplest definition of the adaptive rate which is the following.

Definition 2. *A normalization family $\{\psi_n(\mathbb{F}_\vartheta), \vartheta \in \Theta\big\}$ is called adaptive rate of convergence over collection of functional classes $\{\mathbb{F}_\vartheta, \vartheta \in \Theta\big\}\}$ if*

$$\liminf_{n\to\infty} \inf_{\widetilde{G}_n} \sup_{\vartheta\in\{\vartheta_1,\vartheta_2\}} \psi_n^{-1}(\mathbb{F}_\vartheta)\mathcal{R}_n^{(\ell)}\big[\widetilde{G}_n; \mathbb{F}_\vartheta\big] > 0, \quad \forall \vartheta_1, \vartheta_2 \in \Theta, \tag{2.3}$$

and there exists an estimator $\widehat{G}_n$ such that

$$\limsup_{n\to\infty} \sup_{\vartheta\in\{\vartheta_1,\vartheta_2\}} \psi_n^{-1}(\mathbb{F}_\vartheta)\mathcal{R}_n^{(\ell)}\big[\widehat{G}_n; \mathbb{F}_\vartheta\big] < \infty, \quad \forall \vartheta_1, \vartheta_2 \in \Theta. \tag{2.4}$$

The sequence $\sup_{\vartheta\in\Theta}[\psi_n(\vartheta)/\varphi_n(\vartheta)]$ is called the price to pay for adaptation and the estimator $\widehat{G}_n$ is called an adaptive estimator.

Note that (2.4) is equivalent to

$$\limsup_{n\to\infty} \psi_n^{-1}(\mathbb{F}_\vartheta)\mathcal{R}_n^{(\ell)}\big[\widehat{G}_n; \mathbb{F}_\vartheta\big] < \infty, \quad \forall \vartheta \in \Theta$$

and, therefore, if (2.4) is fulfilled for any $n \in \mathbb{N}^*$ with

$$\psi_n(\vartheta) = c(\vartheta)\phi_n(\vartheta), \quad c(\vartheta) < \infty, \ \forall \vartheta \in \Theta,$$

then one can assert that $\widehat{G}_n$ is an *optimally-adaptive estimator.*

Example 1. *Consider univariate model (1.2), where $\mathcal{D}=[0,1]$ and μ is the Lebesgue measure. Let also $\mathbb{F}_\vartheta=\mathbb{N}_{\infty,1}(\beta,L)$, $\vartheta=(\beta,L)$, be the collection of Nikolskii's classes with $r=\infty$ (Hölder's classes). Let $b,\mathcal{L}>0$ be an arbitrary but a priori chosen numbers, and let $\Theta=(0,b]\times(0,\mathcal{L}]$. The goal is to estimate $G(f)=f(a)$ where $a\in(0,1)$ is a given point.*

The minimax rate of convergence for this problem is given by

$$\phi_n\big(\mathbb{N}_{\infty,1}(\beta,L)\big)=\big(L^{\frac{1}{\beta}}/n\big)^{\frac{\beta}{2\beta+1}},$$

while the adaptive rate of convergence is given, see (26), by

$$\psi_n\big(\mathbb{N}_{\infty,1}(\beta,L)\big)=\big(L^{\frac{1}{\beta}}\ln(n)/n\big)^{\frac{\beta}{2\beta+1}}.$$

We conclude that optimally-adaptive estimators do not exist in this estimation problem.

The most challenging problem of the adaptive theory is to understand how the existence/nonexistence of optimally-adaptive estimators depends on the statistical model, underlying estimation problem (mapping G), loss functional ℓ, and the collection of considered classes. An attempt to provide such classification was undertaken in (27)–(28), but the sufficient conditions found there for both the existence and the nonexistence of optimally-adaptive estimators turned out to be too restrictive.

PROBLEM: Find necessary and sufficient conditions of the existence of optimally-adaptive estimators, i.e. the existence of an estimator $\hat{G}_n$ satisfying the following property:

$$\limsup_{n\to\infty}\phi_n^{-1}\big(\mathbb{F}_\vartheta\big)\mathcal{R}_n^{(\ell)}\big[\hat{G}_n;\mathbb{F}_\vartheta\big]<\infty,\quad\forall\vartheta\in\Theta.$$

This problem stated in (27) thirty years ago remains unsolved.

It is important to realize that answers to the formulated problem may be different even if the statistical model and the collection of functional classes are the same and estimation problems have "similar nature".

Example 2. *Consider univariate model (1.2), where $\mathcal{D}=[0,1]$ and μ is the Lebesgue measure. Let also $\mathbb{F}_\vartheta=\mathbb{N}_{\infty,1}(\beta,L)$, $\vartheta=(\beta,L)$, be the collection of Nikolskii's classes with $r=\infty$ (Hölder's classes). Let $b,\mathcal{L}>0$ be an arbitrary but a priory chosen numbers and let $\Theta=(0,b]\times(0,\mathcal{L}]$. Set*

$$G_\infty(f)=\|f\|_{\infty,[0,1]},\qquad G_2(f)=\|f\|_{2,[0,1]}.$$

The optimally-adaptive estimator of $G_\infty(\cdot)$, was constructed in (29). On the other hand, there is no optimally-adaptive estimator for $G_2(\cdot)$, see (6).

2.3. Adaptive estimation via oracle approach.

Let $\mathcal{G}=\big\{\widehat{G}_\mathfrak{h},\mathfrak{h}\in\mathfrak{H}\big\}$ be a family of estimators built from the observation $X^{(n)}$. The goal is to propose a data-driven (based on $X^{(n)}$) selection procedure from the collection $\mathcal{G}$ and establish for it ℓ-oracle inequality.

More precisely we want to construct a $\mathfrak{H}$-valued random element $\hat{\mathfrak{h}}$ completely determined by the observation $X^{(n)}$ and to prove that for any $n\ge1$

$$\mathcal{R}_n^{(\ell)}\big[\widehat{G}_{\hat{\mathfrak{h}}};G(f)\big]\le\inf_{\mathfrak{h}\in T}U_n^{(\ell)}(f,\mathfrak{h})+r_n,\quad\forall f\in\mathfrak{F}.\tag{2.5}$$

We call (2.5) an *ℓ-oracle inequality*. Here $r_n \to 0, n \to \infty$ is a given sequence which may depend on $\mathfrak{F}$ and the family of estimators $\mathcal{G}$ only. As to the quantity $U_n^{(\ell)}(\cdot,\cdot)$, it is explicitly expressed, and for some particular problems one can prove the inequality (2.5) with

$$U_n^{(\ell)}(f,\mathfrak{h}) = C\mathcal{R}_n^{(\ell)}\big[\widehat{G}_{\mathfrak{h}}; G(f)\big], \tag{2.6}$$

where C is a constant which may depend on $\mathfrak{F}$ and the family of estimators $\mathcal{G}$ only.

Historically, the inequality (2.5) with $U_n^{(\ell)}(\cdot,\cdot)$ given in (2.6) was called the oracle inequality. The latter means that the "oracle" knowing the true parameter f can construct the estimator $\widehat{G}_{\mathfrak{h}(f)}$ which provides the minimal over the collection $\mathcal{G}$ risk for any $f \in \mathfrak{F}$, that is

$$\mathfrak{h}(f): \quad \mathcal{R}_n^{(\ell)}\big[\widehat{G}_{\mathfrak{h}(f)}; G(f)\big] = \inf_{\mathfrak{h}\in\mathfrak{H}} \mathcal{R}_n^{(\ell)}\big[\widehat{G}_{\mathfrak{h}}; G(f)\big].$$

Since $\mathfrak{h}(f)$ depends on unknown f the estimator $\widehat{G}_{\mathfrak{h}(f)}$, called oracle estimator, is not an estimator in usual sense and, therefore, cannot be used. The goal is to construct the estimator $\widehat{G}_{\hat{\mathfrak{h}}}$ which "mimics" the oracle one.
It is worth noting that the ℓ-oracle inequality with $U_n^{(\ell)}(\cdot,\cdot)$ given in (2.6) is not always available, and this is the reason why we deal with more general definition given by (2.5).

The important remark is that inequality (2.5) provides a very simple criterion allowing to assert that the selected estimator $\widehat{G}_{\hat{\mathfrak{h}}}$ is optimally-adaptive or adaptive with respect to the scale of functional classes $\big\{\mathbb{F}_\vartheta, \vartheta \in \Theta\big\}$. Indeed, let us assume that

(i) $r_n \le C \inf_{\vartheta\in\Theta} \phi_n\big(\mathbb{F}_\vartheta\big)$ for some $C > 0$ (verified for all known problems);
(ii) $\exists \vartheta \mapsto \mathfrak{h}(\vartheta)$ and $c(\vartheta) > 0$ such that

$$\sup_{f\in\mathbb{F}_\vartheta} U_n^{(\ell)}\big(f, \mathfrak{h}(\vartheta)\big) \le c(\vartheta)\phi_n\big(\mathbb{F}_\vartheta\big), \ \forall \vartheta \in \Theta.$$

Hence we deduce from (2.5) for any $\vartheta \in \Theta$

$$\sup_{f\in\mathbb{F}_\vartheta} \mathcal{R}_n^{(\ell)}\big[\widehat{G}_{\hat{\mathfrak{h}}}; G(f)\big] \le \sup_{f\in\mathbb{F}_\vartheta} U_n^{(\ell)}\big(f, \mathfrak{h}(\vartheta)\big) + r_n \le \big(c(\theta) + C\big)\phi_n\big(\mathbb{F}_\vartheta\big),$$

and, therefore, we can assert that $\widehat{G}_{\hat{\mathfrak{h}}}$ is optimally-adaptive. If (i) and (ii) hold with $\psi_n\big(\mathbb{F}_\vartheta\big)$ instead of $\phi_n\big(\mathbb{F}_\vartheta\big)$, where $\psi_n\big(\mathbb{F}_\vartheta\big)$ is the adaptive rate of convergence, we can state that $\widehat{G}_{\hat{\mathfrak{h}}}$ is an adaptive estimator.

3. Universal selection rule and ℓ-oracle inequality.

Our objective now is to propose a data-driven selection rule from family of estimators satisfying few very general assumptions and to establish for it ℓ-oracle inequality (2.5). It is important to emphasize that we provide an explicit expression of the functional $U_n^{(\ell)}(\cdot,\cdot)$ that allows us to derive various adaptive results from the unique oracle inequality. The proposed approach can be viewed as a generalization of several estimation procedures developed by the author and his collaborators during last twenty years, see (30), (24), (22), (16), (17), (18), (31), (19), and (32).

3.1. Assumptions.

Let $\mathfrak{H}_n, n \in \mathbb{N}^*$, be a sequence of countable subsets of $\mathfrak{H}$. Let $\{\widehat{G}_{\mathfrak{h}}, \mathfrak{h} \in \mathfrak{H}\}$ and $\{\widehat{G}_{\mathfrak{h},\eta},\ \mathfrak{h}, \eta \in \mathfrak{H}\}$ be the families of $X^{(n)}$-measurable $\mathfrak{S}$-valued mappings possessing the properties formulated below.

Both $\widehat{G}_h$ and $\widehat{G}_{\mathfrak{h},\eta}$ depend usually on n but we will omit this dependence for the sake of simplicity of notations.

Let $\varepsilon_n \to 0$, $n \to \infty$, and δ_n, $n \to \infty$, be two given sequences. Suppose there exist collections of $\mathfrak{S}$-valued functionals $\{\Lambda_{\mathfrak{h}}(f),\ \mathfrak{h} \in \mathfrak{H}\}$, $\{\Lambda_{\mathfrak{h},\eta}(f),\ \mathfrak{h},\eta \in \mathfrak{H}\}$ and a collection of *positive* $X^{(n)}$-measurable random variables $\Psi_n = \{\Psi_n(\mathfrak{h}),\ \mathfrak{h} \in \mathfrak{H}\}$ for which the following conditions hold. (The functionals $\Lambda_{\mathfrak{h}}$ and $\Lambda_{\mathfrak{h},\eta}$ may depend on n (not necessarily) but we will omit this dependence in the notations.)

$\mathbf{A}^{\mathrm{permute}}$. *For any* $f \in \mathfrak{F}$ *and* $n \geq 1$

$$\text{either } (\mathbf{i}) \quad \widehat{G}_{\mathfrak{h},\eta}(f) = \widehat{G}_{\eta,\mathfrak{h}}(f), \quad \forall \eta,\mathfrak{h} \in \mathfrak{H};$$
$$\text{or } (\mathbf{ii}) \quad \sup_{\mathfrak{h},\eta\in\mathfrak{H}_n} \ell\big(\Lambda_{\mathfrak{h},\eta}(f), \Lambda_{\eta,\mathfrak{h}}(f)\big) \leq \delta_n.$$

$\mathbf{A}^{\mathrm{upper}}$. *For any* $f \in \mathfrak{F}$ *and* $n \geq 1$

$$(\mathbf{i}) \quad \mathbb{E}_f^{(n)}\bigg(\sup_{\mathfrak{h}\in\mathfrak{H}_n} \Big[\ell\big(\widehat{G}_{\mathfrak{h}}, \Lambda_{\mathfrak{h}}(f)\big) - \Psi_n(\mathfrak{h})\Big]_+^q \bigg) \leq \varepsilon_n^q;$$

$$(\mathbf{ii}) \quad \mathbb{E}_f^{(n)}\bigg(\sup_{\mathfrak{h},\eta\in\mathfrak{H}_n} \Big[\ell\big(\widehat{G}_{\mathfrak{h},\eta}, \Lambda_{\mathfrak{h},\eta}(f)\big) - \big\{\Psi_n(\mathfrak{h}) \wedge \Psi_n(\eta)\big\}\Big]_+^q \bigg) \leq \varepsilon_n^q.$$

Some remarks are in order.

1) The assumption $\mathbf{A}^{\mathrm{permute}}(\mathbf{i})$ was called in (18) *commutativity property*. The selection rule presented in the next section was proposed in (34) and ℓ-oracle inequality was established under assumptions $\mathbf{A}^{\mathrm{upper}}(\mathbf{i})$ and $\mathbf{A}^{\mathrm{permute}}$. However, it turned out that for some estimator collections the assumption $\mathbf{A}^{\mathrm{permute}}(\mathbf{i})$ is not verified. So our main objective is to prove the same (up to absolute constants) ℓ-oracle inequality under assumptions $\mathbf{A}^{\mathrm{permute}}(\mathbf{ii})$ and $\mathbf{A}^{\mathrm{upper}}$.

2) For many statistical models and problems

$$\Lambda_{\mathfrak{h}}(f) = \mathbb{E}_f^{(n)}\big(\widehat{G}_{\mathfrak{h}}\big), \qquad \Lambda_{\mathfrak{h},\eta}(f) = \mathbb{E}_f^{(n)}\big(\widehat{G}_{\mathfrak{h},\eta}\big).$$

In this case $\ell\big(\widehat{G}_{\mathfrak{h}}, \Lambda_{\mathfrak{h}}(f)\big)$ and $\ell\big(\widehat{G}_{\mathfrak{h},\eta}, \Lambda_{\mathfrak{h},\eta}(f)\big)$ can be viewed as stochastic errors related to the estimators $\widehat{G}_{\mathfrak{h}}$ and $\widehat{G}_{\mathfrak{h},\eta}$ respectively. Hence, following the terminology used in (33) we can say that $\{\Psi_n(\mathfrak{h}),\ \mathfrak{h} \in \mathfrak{H}\}$ and $\{\Psi_n(\mathfrak{h}) \wedge \Psi_n(\eta),\ \mathfrak{h},\eta \in \mathfrak{H}\}$ are upper functions of level ε_n for the collection of corresponding stochastic errors. Often the collection $\{\Psi_n(\mathfrak{h}),\ \mathfrak{h} \in \mathfrak{H}\}$ is not random. This is typically the case when a statistical problem is studied in the framework of white gaussian noise or regression model.

3) We consider countable $\mathfrak{H}_n$ in order not to discuss of the measurability of the supremum inside the mathematical expectation appearing in the assumption $\mathbf{A}^{\mathrm{upper}}$. The developed in the next section theory remains valid for any parameter set over which the corresponding supremum is $X^{(n)}$-measurable.

3.2. Universal selection rule and corresponding ℓ-oracle inequality.

Our objective is to propose the selection rule from an arbitrary collection $\mathcal{G}(\mathfrak{H}_n) = \{\widehat{G}_h, \mathfrak{h} \in \mathfrak{H}_n\}$ satisfying hypotheses $\mathbf{A}^{\mathrm{permute}}$ and $\mathbf{A}^{\mathrm{upper}}$, and establish for it the ℓ-oracle inequality (2.5). Define for any $\mathfrak{h} \in \mathfrak{H}_n$

$$\widehat{R}_n(\mathfrak{h}) = \sup_{\eta\in\mathfrak{H}_n} \Big[\ell\big(\widehat{G}_{\eta}, \widehat{G}_{\mathfrak{h},\eta}\big) - 2\Psi_n(\eta)\Big]_+.$$

Let $\hat{\mathfrak{h}}^{(n)} \in \mathfrak{H}_n$ be an arbitrary $X^{(n)}-$measurable random element satisfying

$$\widehat{R}_n\big(\hat{\mathfrak{h}}^{(n)}\big) + 2\Psi_n\big(\hat{\mathfrak{h}}^{(n)}\big) \le \inf_{\mathfrak{h}\in\mathfrak{H}_n}\big\{\widehat{R}_n(\mathfrak{h}) + 2\Psi_n(\mathfrak{h})\big\} + \varepsilon_n.$$

Our final estimator is $\widehat{G}_{\hat{\mathfrak{h}}^{(n)}}$. In order to bound from above its risk introduce the following notation: for any $f \in \mathbb{F}$, $\mathfrak{h} \in \mathfrak{H}_n$ and $n \ge 1$

$$\mathcal{B}^{(n)}(f,\mathfrak{h}) = \ell\big(\Lambda_{\mathfrak{h}}(f), G(f)\big) + 2\sup_{\eta\in\mathfrak{H}_n}\ell\big(\Lambda_{\mathfrak{h},\eta}(f), \Lambda_\eta(f)\big), \quad \psi_n(f,\mathfrak{h}) = \Big[\mathbb{E}_f^{(n)}\big\{\Psi_n^q(\mathfrak{h})\big\}\Big]^{\frac{1}{q}}.$$

Theorem ((34)). *Let* $\mathbf{A}^{permute}(\mathbf{i})$ *and* $\mathbf{A}^{upper}$ *be fulfilled. Then, for any* $f \in \mathfrak{F}$ *and* $n \ge 1$

$$\mathcal{R}_n^{(\ell)}\big[\widehat{G}_{\hat{\mathfrak{h}}^{(n)}}; G(f)\big] \le \inf_{\mathfrak{h}\in\mathfrak{H}_n}\big\{\mathcal{B}^{(n)}(f,\mathfrak{h}) + 5\psi_n(f,\mathfrak{h})\big\} + 6\varepsilon_n.$$

Thus, ℓ-oracle inequality is established with $r_n = 6\varepsilon_n$ and

$$U_n^{(\ell)}(f,\mathfrak{h}) = \mathcal{B}^{(n)}(f,\mathfrak{h}) + 5\psi_n(f,\mathfrak{h}).$$

Our goal now is to prove the following result.

Theorem. *Let* $\mathbf{A}^{permute}(\mathbf{ii})$ *and* $\mathbf{A}^{upper}$ *be fulfilled. Then, for any* $f \in \mathfrak{F}$ *and* $n \ge 1$

$$\mathcal{R}_n^{(\ell)}\big[\widehat{G}_{\hat{\mathfrak{h}}^{(n)}}; G(f)\big] \le \inf_{\mathfrak{h}\in\mathfrak{H}_n}\big\{\mathcal{B}^{(n)}(f,\mathfrak{h}) + 9\psi_n(f,\mathfrak{h})\big\} + 10\varepsilon_n + \delta_n.$$

Thus, ℓ-oracle inequality is established with $r_n = 10\varepsilon_n + \delta_n$ and

$$U_n^{(\ell)}(f,\mathfrak{h}) = \mathcal{B}^{(n)}(f,\mathfrak{h}) + 9\psi_n(f,\mathfrak{h}).$$

Proof. We break the proof into three short steps and for the simplicity of notations we will write $\hat{\mathfrak{h}}$ instead of $\hat{\mathfrak{h}}^{(n)}$. Set

$$\xi_1 = \sup_{\eta\in\mathfrak{H}_n}\Big[\ell\big(\widehat{G}_\eta, \Lambda_\eta\big) - \Psi_n(\eta)\Big]_+, \quad \xi_2 = \sup_{\mathfrak{h},\eta\in\mathfrak{H}_n}\Big[\ell\big(\widehat{G}_{\mathfrak{h},\eta}, \Lambda_{\mathfrak{h},\eta}\big) - \big\{\Psi_n(\mathfrak{h}) \wedge \Psi_n(\eta)\big\}\Big]_+.$$

1) Our first goal is to prove that for any $\mathfrak{h}, \eta \in \mathfrak{H}_n$

$$\ell\big(\widehat{G}_{\mathfrak{h}}, \widehat{G}_{\mathfrak{h},\eta}\big) \le \widehat{R}_n(\eta) + 6\Psi_n(\mathfrak{h}) + 2\xi_1 + 2\xi_2 + \delta_n. \tag{3.1}$$

Indeed, the following chain od inequalities is obtained from the triangle inequality

$$\begin{aligned}\ell\big(\widehat{G}_{\mathfrak{h}}, \widehat{G}_{\mathfrak{h},\eta}\big) &\le \ell\big(\widehat{G}_{\mathfrak{h}}, \Lambda_{\mathfrak{h}}\big) + \ell\big(\Lambda_{\mathfrak{h}}, \widehat{G}_{\mathfrak{h},\eta}\big)\\ &\le \ell\big(\widehat{G}_{\mathfrak{h}}, \Lambda_{\mathfrak{h}}\big) + \ell\big(\Lambda_{\mathfrak{h}}, \Lambda_{\mathfrak{h},\eta}\big) + \ell\big(\widehat{G}_{\mathfrak{h},\eta}, \Lambda_{\mathfrak{h},\eta}\big)\\ &\le \ell\big(\Lambda_{\mathfrak{h}}, \Lambda_{\mathfrak{h},\eta}\big) + 2\Psi_n(\mathfrak{h}) + \xi_1 + \xi_2. \end{aligned}\tag{3.2}$$

Similarly, taking into account $\mathbf{A}^{\mathrm{permute}}(\mathbf{ii})$ we get

$$\begin{aligned}\ell\big(\Lambda_{\mathfrak{h}}, \Lambda_{\mathfrak{h},\eta}\big) &\le \ell\big(\Lambda_{\mathfrak{h}}, \Lambda_{\eta,\mathfrak{h}}\big) + \delta_n\\ &\le \ell\big(\widehat{G}_{\mathfrak{h}}, \Lambda_{\mathfrak{h}}\big) + \ell\big(\widehat{G}_{\mathfrak{h}}, \widehat{G}_{\eta,\mathfrak{h}}\big) + \ell\big(\widehat{G}_{\eta,\mathfrak{h}}, \Lambda_{\eta,\mathfrak{h}}\big) + \delta_n\\ &\le \ell\big(\widehat{G}_{\mathfrak{h}}, \widehat{G}_{\eta,\mathfrak{h}}\big) + 2\Psi_n(\mathfrak{h}) + \xi_1 + \xi_2 + \delta_n. \end{aligned}\tag{3.3}$$

It remains to note that in view of the definition of $\widehat{R}_n(\cdot)$

$$\ell\big(\widehat{G}_{\mathfrak{h}},\widehat{G}_{\eta,\mathfrak{h}}\big)\le 2\Psi_n(\mathfrak{h})+\Big[\ell\big(\widehat{G}_{\mathfrak{h}},\widehat{G}_{\eta,\mathfrak{h}}\big)-2\Psi_n(\mathfrak{h})\Big]_+\le 2\Psi_n(\mathfrak{h})+\widehat{R}_n(\eta).$$

This together with (3.2) and (3.3) implies (3.1).

2) Let $\mathfrak{h}\in\mathfrak{H}_n$ be fixed. We have in view of the definition of $\widehat{R}_n(\cdot)$

$$\ell\big(\widehat{G}_{\hat{\mathfrak{h}}},\widehat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}\big)\le 2\Psi_n(\hat{\mathfrak{h}})+\Big[\ell\big(\widehat{G}_{\hat{\mathfrak{h}}},\widehat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}\big)-2\Psi_n(\hat{\mathfrak{h}})\Big]_+\le 2\Psi_n(\hat{\mathfrak{h}})+\widehat{R}_n(\mathfrak{h}). \tag{3.4}$$

Here we have also used that $\hat{\mathfrak{h}}\in\mathfrak{H}_n$ by its definition.
Applying (3.1) with $\eta=\hat{\mathfrak{h}}$ we obtain

$$\ell\big(\widehat{G}_{\mathfrak{h}},\widehat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}\big)\le\widehat{R}_n\big(\hat{\mathfrak{h}}\big)+6\Psi_n(\mathfrak{h})+2\xi_1+2\xi_2+\delta_n. \tag{3.5}$$

We get from (3.4), (3.5) and the definition of $\hat{\mathfrak{h}}$

$$\begin{aligned}\ell\big(\widehat{G}_{\hat{\mathfrak{h}}},\widehat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}\big)+\ell\big(\widehat{G}_{\mathfrak{h}},\widehat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}\big)&\le\widehat{R}_n(\hat{\mathfrak{h}})+2\Psi_n(\hat{\mathfrak{h}})+\widehat{R}_n(\mathfrak{h})+6\Psi_n(\mathfrak{h})+2\xi_1+2\xi_2+\delta_n\\&\le 2\widehat{R}_n(\mathfrak{h})+8\Psi_n(\mathfrak{h})+2\xi_1+2\xi_2+\varepsilon_n+\delta_n.\end{aligned} \tag{3.6}$$

3) We have in view of the triangle inequality for any $\mathfrak{h}\in\mathfrak{H}_n$

$$\widehat{R}_n(\mathfrak{h})\le\sup_{\eta\in\mathfrak{H}_n}\ell\big(\Lambda_{\mathfrak{h},\eta}(f),\Lambda_\eta(f)\big)+\xi_1+\xi_2. \tag{3.7}$$

Thus, we obtain from (3.6) and (3.7) for any $\mathfrak{h}\in\mathfrak{H}_n$

$$\begin{aligned}&\ell\big(\widehat{G}_{\hat{\mathfrak{h}}},\widehat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}\big)+\ell\big(\widehat{G}_{\mathfrak{h}},\widehat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}\big)\\&\qquad\le 2\sup_{\eta\in\mathfrak{H}_n}\ell\big(\Lambda_{\mathfrak{h},\eta}(f),\Lambda_\eta(f)\big)+8\Psi_n(\mathfrak{h})+4\xi_1+4\xi_2+\varepsilon_n+\delta_n.\end{aligned} \tag{3.8}$$

Obviously for any $\mathfrak{h}\in\mathfrak{H}_n$

$$\ell\big(\widehat{G}_{\mathfrak{h}},G(f)\big)\le\ell\big(\Lambda_{\mathfrak{h}}(f),G(f)\big)+\Psi_n(\mathfrak{h})+\xi_1.$$

By the triangle inequality it yields together with (3.8) for any $\mathfrak{h}\in\mathfrak{H}_n$

$$\ell\big(\widehat{G}_{\hat{\mathfrak{h}}},G(f)\big)\le\mathcal{B}^{(n)}(f,\mathfrak{h})+9\Psi_n(\mathfrak{h})+5\xi_1+4\xi_2+\varepsilon_n+\delta_n,\quad\forall f\in\mathfrak{F}.$$

Taking into account the hypothesis $\mathbf{A}^{\rm upper}$ we get for any $\mathfrak{h}\in\mathfrak{H}_n$ and any $f\in\mathfrak{F}$

$$\Big\{\mathbb{E}_f^{(n)}\Big[\ell\big(\widehat{G}_{\hat{\mathfrak{h}}},G(f)\big)\Big]^q\Big\}^{\frac{1}{q}}\le\mathcal{B}^{(n)}(f,\mathfrak{h})+9\psi_n(f,\mathfrak{h})+10\varepsilon_n+\delta_n.$$

Noting that the left hand side of the obtained inequality is independent of $\mathfrak{h}$ we come to the assertion of the theorem. ∎

We finish this section with simple but very useful (in minimax and minimax adaptive estimation) consequence of Theorems 3.2–3.2.
Set for any $\mathbb{F}\subseteq\mathfrak{F}$

$$\gamma_n(\mathbb{F})=\inf_{\mathfrak{h}\in\mathfrak{H}}\sup_{f\in\mathbb{F}}\big[\mathcal{B}^{(n)}(f,\mathfrak{h})+\psi_n(f,\mathfrak{h})\big].$$

The quantity $\gamma_n(\mathbb{F})$ is often called *bias-variance tradeoff*.

Corollary 1. *Let $\mathbf{A}^{upper}$ be fulfilled. Assume also that either $\mathbf{A}^{permute}$(**i**) holds or $\mathbf{A}^{permute}$(**ii**) is verified with $\delta_n=\varepsilon_n$. Then, for any $\mathbb{F}\subseteq\mathfrak{F}$ and $n\ge1$*

$$\mathcal{R}_n^{(\ell)}\big[\widehat{G}_{\hat{\mathfrak{h}}^{(n)}};\mathbb{F}\big]\le 9\gamma_n(\mathbb{F})+11\varepsilon_n.$$

The proof of the corollary is elementary and can be omitted.

4. Examples of estimator collections satisfying assumption $\mathbf{A}^{\mathbf{permute}}$.

4.1. Estimator collections in Density model.

First example. Let $\mathcal{D}=\mathbb{R}^d, d\geq 1$ and μ be the Lebesgue measure. Let $K:\mathbb{R}^d\to\mathbb{R}$ be a function from $\mathbb{L}_1\big(\mathbb{R}^d\big)$, and $\int_{\mathbb{R}}K=1$. Let $\mathfrak{H}\subseteq(0,1]^d$, and define for any $\mathfrak{h}=(\mathfrak{h}_1,\ldots,\mathfrak{h}_d)\in\mathfrak{H}$

$$K_{\mathfrak{h}}(t)=V_{\mathfrak{h}}^{-1}K\big(t_1/\mathfrak{h}_1,\ldots,t_d/\mathfrak{h}_d\big),\ t\in\mathbb{R}^d,\quad V_{\mathfrak{h}}=\prod_{j=1}^{d}\mathfrak{h}_j. \tag{4.1}$$

Introduce the following estimator collection

$$\mathcal{G}=\Big\{\widehat{G}_{\mathfrak{h}}(x)=n^{-1}\sum_{i=1}^{n}K_{\mathfrak{h}}\big(X_i-x\big),\ x\in\mathbb{R}^d,\ \mathfrak{h}\in\mathfrak{H}\Big\}. \tag{4.2}$$

The estimator $\widehat{G}_{\mathfrak{h}}(\cdot)$ is called the kernel estimator with bandwidth $\mathfrak{h}$. Kernel estimators are used in estimating the underlying density at a given point as well as in estimating of entire f. Also, they are used as a building block for constructing estimators of many functionals of density mentioned in Section 1.2. Selection from the family $\mathcal{G}$, usually referred to as bandwidth selection, is one of the central problems in nonparametric density estimation.

Set for any $\mathfrak{h}\in\mathfrak{H}$

$$\Lambda_{\mathfrak{h}}(f,\cdot)=\mathbb{E}_f^{(n)}\Big[\widehat{G}_{\mathfrak{h}}(\cdot)\Big]=\int_{\mathcal{D}}K_{\mathfrak{h}}(t-\cdot)f(t)\mathrm{d}t$$

and consider two possible constructions of the collection $\widehat{G}_{\mathfrak{h},\eta}(\cdot),\mathfrak{h},\eta\in\mathfrak{H}$.

Construction based on the convolution product. Define $K_{\mathfrak{h},\eta}:\mathbb{R}^d\to\mathbb{R}$ by

$$K_{\mathfrak{h},\eta}(\cdot)=\int_{\mathbb{R}^d}K_{\eta}(\cdot-t)K_{\mathfrak{h}}(t)\mathrm{d}t=:\big[K_{\mathfrak{h}}*K_{\eta}\big](\cdot).$$

and set

$$\widehat{G}_{\mathfrak{h},\eta}(\cdot)=n^{-1}\sum_{i=1}^{n}K_{\mathfrak{h},\eta}\big(X_i-\cdot\big),\quad \Lambda_{\mathfrak{h},\eta}(f,\cdot)=\mathbb{E}_f^{(n)}\Big[\widehat{G}_{\mathfrak{h},\eta}(\cdot)\Big].$$

Since obviously $K_{\mathfrak{h},\eta}\equiv K_{\eta,\mathfrak{h}}$ we can assert that $\widehat{G}_{\mathfrak{h},\eta}\equiv\widehat{G}_{\eta,\mathfrak{h}}$ and, therefore the assumptions $\mathbf{A}^{\mathrm{permute}}(\mathbf{i})$ and $\mathbf{A}^{\mathrm{permute}}(\mathbf{ii})$ are both fulfilled.

Construction based on the coordinatewise maximum. Define $K_{\mathfrak{h},\eta}:\mathbb{R}^d\to\mathbb{R}$ by

$$K_{\mathfrak{h},\eta}(\cdot)=K_{\mathfrak{h}\vee\eta}(\cdot),\quad \mathfrak{h}\vee\eta=\big(\mathfrak{h}_1\vee\eta_1,\ldots,\mathfrak{h}_d\vee\eta_d\big),$$

and set

$$\widehat{G}_{\mathfrak{h},\eta}(\cdot)=n^{-1}\sum_{i=1}^{n}K_{\mathfrak{h},\eta}\big(X_i-\cdot\big),\quad \Lambda_{\mathfrak{h},\eta}(f,\cdot)=\mathbb{E}_f^{(n)}\Big[\widehat{G}_{\mathfrak{h},\eta}(\cdot)\Big].$$

Since obviously $K_{\mathfrak{h},\eta}\equiv K_{\eta,\mathfrak{h}}$ we can assert that $\widehat{G}_{\mathfrak{h},\eta}\equiv\widehat{G}_{\eta,\mathfrak{h}}$ and, therefore the assumptions $\mathbf{A}^{\mathrm{permute}}(\mathbf{i})$ and $\mathbf{A}^{\mathrm{permute}}(\mathbf{ii})$ are both fulfilled.

Second example. Consider now the estimator collection related to the density estimation under hypothesis of independence presented in Section 2.1.2.

Here is previously $\mathcal{D}=\mathbb{R}^d, d\geq 2$, μ is the Lebesgue measure. Recall that $\mathfrak{F}\subseteq\mathfrak{P}(\mathcal{D},\mu)$, $\mathcal{I}_d$ is the set of all subsets of $\{1,\dots,d\}$ and Π denotes the set of all partitions of $\{1,\dots,d\}$.

Let $\mathbf{K}:\mathbb{R}^1\to\mathbb{R}^1$ be a univariate kernel, that is $\mathbf{K}\in\mathbb{L}_1(\mathbb{R}^1)$ and $\int_{\mathbb{R}^1}\mathbf{K}=1$.
For any $h=(0,1]^d$ and any $I\in\mathcal{I}_d$ set

$$K_{h_I}(u)=V_{h_I}^{-1}\prod_{j\in I}\mathbf{K}\big(u_j/h_j\big),\qquad V_{h_I}=\prod_{j\in I}h_j.$$

Since the independence hypothesis assumes that there exists a partition $\mathcal{P}$ such that

$$f(x)=\prod_{I\in\mathcal{P}}f_I(x_I),\quad\forall x\in\mathbb{R}^d,$$

the idea is to estimate each marginal density by kernel method and to use the product of these estimators as the final one. Thus, define for any $x\in\mathbb{R}^d$, $\mathfrak{h}\in\mathfrak{H}$ and any $I\in\mathcal{I}_d$

$$\widehat{f}_{h_I}(x_I)=n^{-1}\sum_{i=1}^{n}K_{h_I}\big(X_{I,i}-x_I\big)$$

and introduce the following family of estimators

$$\mathcal{G}=\Big\{\widehat{G}_{\mathfrak{h}}(x)=\prod_{I\in\mathcal{P}}\widehat{f}_{h_I}(x_I),\ \ x\in\mathbb{R}^D,\ \mathfrak{h}=(h,\mathcal{P})\in[0,1]^d\times\Pi=:\mathfrak{H}\Big\}.$$

Let $*$ denote the convolution operator on $\mathbb{R}$. Set for any $x\in\mathbb{R}^d$, $h,h'\in(0,1]^d$ and any $I\in\mathcal{I}_d$

$$\big[K_{h_{\mathbf{I}}}\star K_{h'_I}\big]=\prod_{j\in I}\big[\mathbf{K}_{h_j}*\mathbf{K}_{h'_j}\big]$$

and introduce

$$\widehat{f}_{h_I,h'_I}\big(x_I\big)=n^{-1}\sum_{i=1}^{n}\big[K_{h_I}\star K_{h'_I}\big]\big(X_{I,i}-x_I\big),$$

Let us endow the set Π with the operation "$\diamond$" putting for any $\mathcal{P},\mathcal{P}'\in\Pi$

$$\mathcal{P}\diamond\mathcal{P}'=\{I\cap I'\neq\emptyset,\ \ I\in\mathcal{P},\ I'\in\mathcal{P}'\}\in\Pi.$$

Introduce for any $\mathfrak{h},\eta\in\mathfrak{H}$ the estimator

$$\widehat{G}_{\mathfrak{h},\eta}(x)=\prod_{I\in\mathcal{P}\diamond\mathcal{P}'}\widehat{f}_{h_I,h'_I}\big(x_I\big),\ \ x\in\mathbb{R}^d.$$

Obviously $\widehat{G}_{\mathfrak{h},\eta}\equiv\widehat{G}_{\eta,\mathfrak{h}}$ and, therefore the assumptions $\mathbf{A}^{\mathrm{permute}}(\mathbf{i})$ is fulfilled. On the other hand, see (31), functionals $\Lambda_{\mathfrak{h}}$ and $\Lambda_{\mathfrak{h},\eta}$ are so complicated that the verification of $\mathbf{A}^{\mathrm{permute}}(\mathbf{ii})$ is not seemed possible. We are not even sure that it holds with sufficiently small δ_n.

Third example. Let us now consider the family of estimators which appears in adaptive estimation under following structural assumption Let $\mathcal{D}=\mathbb{R}^2$ and μ is the Lebesgue measure. Let $\mathfrak{Q}$ denote the set of all 2×2 rotational matrices and $\mathfrak{P}_1^{sym}$ denote the set of all symmetric probability densities on $\mathbb{R}^1$. Set

$$\mathcal{A}=\big\{a:\mathbb{R}^2\to\mathbb{R}^1:\quad a(\cdot,\cdot)=a_1(\cdot)a_2(\cdot),\ a_1,a_2\in\mathfrak{P}_1^{sym}\big\},$$

and assume that there exist $a \in \mathcal{A}$ and $M \in \mathfrak{Q}$ such that $f(\cdot) = a(M^T \cdot)$.
The latter means that

$$X_i = M\xi_i, \quad i = 1, \dots, n,$$

where ξ_i, $i = 1, \dots, n$, are i.i.d. random vectors with common density a.

If M is known then $\xi_i = M^T X_i, \dots, \xi_n = M^T X_n$ are observable i.i.d. random vectors with *independent coordinates*. Indeed, the density of ξ_1 is $a_1(\cdot)a_2(\cdot)$. Hence the estimation of a is the estimation under hypothesis of independence, which as it was mentioned above allows to improve the accuracy of estimation of the density a, and, therefore, of the density f as well. However, if M is unknown, the sequence $\xi_i = M^T X_i, \dots, \xi_n = M^T X_n$ is not observable anymore and the estimation of f can be viewed as the problem of adaptation to unknown rotation of coordinate system.

Let the kernel $\mathbf{K} : \mathbb{R}^1 \to \mathbb{R}^1$ be the same as in previous example and set $K_h(\cdot) = h^{-1}\mathbf{K}(\cdot/h)$, $h \in (0, 1]$. Later on $Q \in \mathfrak{Q}$ will be presented as

$$Q = (q, q_\perp) = \begin{pmatrix} q_1 & -q_2 \\ q_2 & q_1 \end{pmatrix},$$

where $q, q_\perp \in \mathbb{S}^1$. Set for any $\mathfrak{h} := (h, Q) \in [0, 1] \times \mathfrak{Q}$ and $x \in \mathbb{R}^2$

$$\widehat{G}_\mathfrak{h}(x) = \Big[n^{-1}\sum_{k=1}^n K_h\big(q^T(X_k - x)\big)\Big]\Big[n^{-1}\sum_{k=1}^n K_h\big(q_\perp^T(X_k - x)\big)\Big].$$

and introduce the following family of estimators.

$$\mathcal{G} = \big\{\widehat{G}_\mathfrak{h}(x),\ x \in \mathbb{R}^2,\ \mathfrak{h} \in \mathfrak{H} \subseteq [0, 1] \times \mathfrak{Q}\big\}.$$

In order to construct estimator $\widehat{G}_{\mathfrak{h},\eta}(\cdot)$, $\mathfrak{h}, \eta \in \mathfrak{H}$ we will need the following notation.

Set for any $Q, D \in \mathfrak{Q}$

$$p(D, Q) = q^T d_\perp, \quad \pi(D, Q) = q^T d.$$

Set also $\mathcal{K}_h(t) = K_h(t_1)K_h(t_2), t \in \mathbb{R}^2$, $h \in (0, 1]$, and let

$$\Gamma = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \Omega = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

Define, see (35), for any $\mathfrak{h} = (h, Q) \in \mathfrak{H}$ and $\eta = (\varkappa, D) \in \mathfrak{H}$

$$\widehat{G}_{\mathfrak{h},\eta}(x) = \frac{1}{n(n-1)}\sum_{k,l=1, k\neq l}^n \mathcal{K}_{h\vee\varkappa}\big(p(D, Q)\Omega\Gamma X_k + \pi(D, Q)X_l - \Omega\Gamma QD\Omega x\big)$$

and let

$$\Lambda_{\mathfrak{h},\eta}(f, \cdot) = \mathbb{E}_f^{(n)}\Big[\widehat{G}_{\mathfrak{h},\eta}(\cdot)\Big].$$

Note that for any $D, Q \in \mathfrak{Q}$

$$p(D, Q) = -p(Q, D), \quad \pi(D, Q) = \pi(Q, D), \quad DQ = QD. \tag{4.3}$$

Obviously $\widehat{G}_{\mathfrak{h},\eta}(\cdot) \neq \widehat{G}_{\eta,\mathfrak{h}}(\cdot)$ and, therefore, the assumption $\mathbf{A}^{\mathrm{permute}}(\mathbf{i})$ is not verified.
On the other hand

$$\Lambda_{\mathfrak{h},\eta}(f, \cdot) = \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} \mathcal{K}_{h\vee\varkappa}\big(p(D, Q)\Omega\Gamma u + \pi(D, Q)v - \Omega\Gamma QD\Omega x\big)f(u)f(v)\mathrm{d}u\mathrm{d}v.$$

Since $f(\cdot)=a(M^T\cdot)$ and a is symmetric, f is symmetric function as well and we have

$$\begin{aligned}\Lambda_{\mathfrak{h},\eta}(f,\cdot)&=\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\mathcal{K}_{h\vee\varkappa}\big(-p(D,Q)\Omega\Gamma u+\pi(D,Q)v-\Omega\Gamma QD\Omega x\big)f(u)f(v)\mathrm{d}u\mathrm{d}v\\&=\int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\mathcal{K}_{h\vee\varkappa}\big(p(Q,D)\Omega\Gamma u+\pi(Q,D)v-\Omega\Gamma DQ\Omega x\big)f(u)f(v)\mathrm{d}u\mathrm{d}v\\&=\Lambda_{\eta,\mathfrak{h}}(f,\cdot).\end{aligned}$$

To get the penultimate equality we have used (4.3). We conclude that the assumption $\mathbf{A}^{\mathrm{permute}}(\mathbf{ii})$ holds with any δ_n whatever the semi-metric ℓ is considered.

4.2. Estimator collections in White Gaussian Noise Model.

First example. Let $\mathcal{D}$ be a set endowed with the Borel measure μ and $\mu(\mathcal{D})<\infty$. Recall that the observation $X^{(n)}=\{X_n(g),g\in\mathbb{L}_2(\mathcal{D},\mu)\}$ is given in (1.2).
Let $\{\psi_{\boldsymbol{k}},\boldsymbol{k}\in\mathbb{M}\}$ be an orthonormal basis in $\mathbb{L}_2\big(\mathcal{D},\mu\big)$ and let $\mathfrak{H}=\big\{\mathfrak{h}=\big(\mathfrak{h}_{\boldsymbol{k}},\boldsymbol{k}\in\mathbb{M}\big)\big\}$ be a given subset of l_2. Introduce for any $t,x\in\mathcal{D}$

$$K_{\mathfrak{h}}(t,x)=\sum_{\boldsymbol{k}\in\mathbb{M}}\mathfrak{h}_{\boldsymbol{k}}\psi_{\boldsymbol{k}}(t)\psi_{\boldsymbol{k}}(x),\ \mathfrak{h}\in\mathfrak{H},$$

and consider the following estimation collection

$$\mathcal{G}=\Big\{\widehat{G}_{\mathfrak{h}}(x)=X_n\big(K(\cdot,x)\big),x\in\mathcal{D},\ \mathfrak{h}\in\mathfrak{H}\Big\}.$$

The estimator $\widehat{G}_{\mathfrak{h}}(\cdot)$ is used in estimation of unknown f under $\mathbb{L}_2$-loss, that is $\mathfrak{S}=\mathfrak{F}$, $G(f)=f$ and $\ell(f,g)=\|f-g\|_{2,\mathcal{D}}$, $f,g\in\mathfrak{F}\subset\mathbb{L}_2(\mathcal{D},\mu)$. Let

$$\Lambda_{\mathfrak{h}}(f,\cdot)=\mathbb{E}_f^{(n)}\Big[\widehat{G}_{\mathfrak{h}}(\cdot)\Big]=\int_{\mathcal{D}}K_{\mathfrak{h}}(t,\cdot)f(t)\mu(\mathrm{d}t)=\sum_{\boldsymbol{k}\in\mathbb{M}}\mathfrak{h}_{\boldsymbol{k}}\psi_{\boldsymbol{k}}(\cdot)\int_{\mathcal{D}}\psi_{\boldsymbol{k}}(t)f(t)\mu(\mathrm{d}t).$$

Denoting $\boldsymbol{k}$-th Fourier coefficient of f by $f_{\boldsymbol{k}}$ we get

$$\Lambda_{\mathfrak{h}}(f,\cdot)=\sum_{\boldsymbol{k}\in\mathbb{M}}\mathfrak{h}_{\boldsymbol{k}}f_{\boldsymbol{k}}\psi_{\boldsymbol{k}}(\cdot).$$

In particular, in view of Parseval's identity

$$\big\|\Lambda_{\mathfrak{h}}(f)-f\big\|_{2,\mathcal{D}}=\sum_{\boldsymbol{k}\in\mathbb{M}}\big(\mathfrak{h}_{\boldsymbol{k}}-1\big)^2f_{\boldsymbol{k}}^2.$$

Set for any $\mathfrak{h},\eta\in\mathfrak{H}$

$$K_{\mathfrak{h},\eta}(t,x)=\int_{\mathcal{D}}K_{\mathfrak{h}}(t,y)K_{\eta}(y,x)\mu(\mathrm{d}y),\quad t,x\in\mathcal{D}$$

and put for any $x\in\mathcal{D}$

$$\widehat{G}_{\mathfrak{h},\eta}(x)=X_n\Big(K_{\mathfrak{h},\eta}(\cdot,x)\Big).$$

Noting that that for any $t,x\in\mathcal{D}$

$$K_{\mathfrak{h},\eta}(t,x)=\sum_{\boldsymbol{k}\in\mathbb{M}}\sum_{j\in\mathbb{M}}\mathfrak{h}_{\boldsymbol{k}}\eta_j\psi_{\boldsymbol{k}}(t)\psi_j(x)\int_{\mathcal{D}}\psi_{\boldsymbol{k}}(y)\psi_j(y)\mu(\mathrm{d}t)=\sum_{\boldsymbol{k}\in\mathbb{M}}\mathfrak{h}_{\boldsymbol{k}}\eta_m\psi_{\boldsymbol{k}}(t)\psi_m(x)$$

we can assert that $K_{\mathfrak{h},\eta} \equiv K_{\eta,\mathfrak{h}}$. It implies $\widehat{G}_{\mathfrak{h},\eta} \equiv \widehat{G}_{\eta,\mathfrak{h}}$ and, therefore

$$\Lambda_{\mathfrak{h},\eta} := \mathbb{E}_f^{(n)}\Big[\widehat{G}_{\mathfrak{h},\eta}\Big] \equiv \mathbb{E}_f^{(n)}\Big[\widehat{G}_{\eta,\mathfrak{h}}\Big] =: \Lambda_{\eta,\mathfrak{h}}.$$

Hence, the assumptions $\mathbf{A}^{\mathrm{permute}}(\mathbf{i})$ and $\mathbf{A}^{\mathrm{permute}}(\mathbf{ii})$ are both fulfilled.

Second example. Here and later $D = R^d$, $d \geq 1$, μ is the Lebesgue measure and $X^{(n)} = \{X_n(g), g \in \mathbb{L}_2(\mathbb{R}^d, \mu)\}$ is given in (1.2).
Let $b > 0$ be given and denote by $\mathfrak{H}(b)$ the set of all Borel functions $\mathfrak{h} : (-b, b)^d \to (0, 1]^d$.
As previously let $K : \mathbb{R}^d \to \mathbb{R}$, $K \in \mathbb{L}_1(\mathbb{R}^d)$ be a function satisfying $\int K = 1$.

With any $\mathfrak{h} \in \mathfrak{H}(b)$ we associate the function

$$K_{\mathfrak{h}(x)}(t, x) = V_{\mathfrak{h}}^{-1}(x) K\left(\frac{t - x}{\mathfrak{h}(x)}\right), \quad t \in \mathbb{R}^d, \ x \in (-b, b)^d,$$

where $V_{\mathfrak{h}}(x) = \prod_{i=1}^d \mathfrak{h}_i(x)$ and $\mathfrak{h}(\cdot) = (\mathfrak{h}_1(\cdot), \ldots, \mathfrak{h}_d(\cdot))$.

Consider the family of estimators

$$\mathcal{G} = \Big\{\widehat{G}_{\mathfrak{h}(x)}(x) = X_n\big(K_{\mathfrak{h}(x)}(\cdot, x)\big), \ \mathfrak{h} \in \mathfrak{H}(b), \ x \in (-b, b)^d\Big\}. \tag{4.4}$$

The estimators from this collection are called *kernel estimators with varying bandwidth*. Let

$$\Lambda_{\mathfrak{h}(\cdot)}(f, \cdot) = \mathbb{E}_f^{(n)}\Big[\widehat{G}_{\mathfrak{h}(\cdot)}(\cdot)\Big] = \int_{\mathbb{R}^d} K_{\mathfrak{h}(\cdot)}(t, \cdot) f(t) \mu(\mathrm{d}t).$$

For any $\mathfrak{h}, \eta \in \mathfrak{H}(b)$ set

$$\widehat{G}_{\mathfrak{h}(x)\vee\eta(x)}(x) = X_n\big(K_{\mathfrak{h}(x)\vee\eta(x)}(\cdot, x)\big), \ x \in (-b, b)^d,$$

where as previously $\mathfrak{h}(\cdot) \vee \eta(\cdot) = \big(\mathfrak{h}_1(\cdot) \vee \eta_1(\cdot), \ldots, \mathfrak{h}_d(\cdot) \vee \eta_d(\cdot)\big)$. Let also

$$\widehat{G}_{\mathfrak{h}(\cdot)\vee\eta(\cdot)}(\cdot) = \mathbb{E}_f^{(n)}\Big[\widehat{G}_{\mathfrak{h}(\cdot)}\Big] = \int_{\mathbb{R}^d} K_{\mathfrak{h}(\cdot)}(t, \cdot) f(t) \mu(\mathrm{d}t).$$

Since obviously $K_{\mathfrak{h}\vee\eta} \equiv K_{\eta\vee\mathfrak{h}}$ for any $\mathfrak{h}, \eta \in \mathfrak{H}(b)$ we can assert that both assumptions $\mathbf{A}^{\mathrm{permute}}(\mathbf{i})$ and $\mathbf{A}^{\mathrm{permute}}(\mathbf{ii})$ are fulfilled whatever the semi-metric ℓ is considered.

5. One example of estimator collection satisfying assumption $\mathbf{A}^{\mathbf{upper}}$.

In this section we continue to consider the estimator family given in (4.4). Our objective here is to find $\mathfrak{H}_n \subset \mathfrak{H}(b)$ and $\{\Psi_n(\mathfrak{h}), \mathfrak{h} \in \mathfrak{H}_n\}$ for which assumption $\mathbf{A}^{\mathrm{upper}}$ can be checked in the case, where ℓ is $\mathbb{L}_p$-norm on $(-b, b)^d$, $1 \leq p < \infty$.

Set for any $\mathfrak{h} \in \mathfrak{H}(b)$

$$\xi_{\mathfrak{h}(x)}(x) = \int_{\mathbb{R}^d} K_{\mathfrak{h}(x)}(t, x) W(\mathrm{d}t), \quad x \in (-b, b)^d$$

and note that in view of (1.2)

$$\ell\big(\widehat{G}_{\mathfrak{h}}, \Lambda_{\mathfrak{h}}(f)\big) = n^{-\frac{1}{2}} \big\|\xi_{\mathfrak{h}}\big\|_{p,(-b,b)^d}.$$

We remark that $\xi_{\mathfrak{h}(\cdot)}(\cdot)$ is independent of f and n. Hence, assumption $\mathbf{A}^{\rm upper}$ will be checked if we find $\mathfrak{H}_n$ and non random $\{\Psi^*_n(\mathfrak{h}), \mathfrak{h}\in\mathfrak{H}_n\}$ such that

$$\mathbb{E}\left(\sup_{\mathfrak{h}\in\mathfrak{H}_n}\left[\left\|\xi_{\mathfrak{h}}\right\|_{p,(-b,b)^d}-\Psi^*_n(\mathfrak{h})\right]^q_+\right)\le\varepsilon^q_n n^{\frac{q}{2}}; \tag{5.1}$$

$$\mathbb{E}\left(\sup_{\mathfrak{h},\eta\in\mathfrak{H}_n}\left[\left\|\xi_{\mathfrak{h}\vee\eta}\right\|_{p,(-b,b)^d}-\left\{\Psi^*_n(\mathfrak{h})\wedge\Psi^*_n(\eta)\right\}\right]^q_+\right)\le\varepsilon^q_n n^{\frac{q}{2}}. \tag{5.2}$$

Here and later $\mathbb{E}$ denotes the mathematical expectation with respect to the law of W. Also, furthermore, we will assume that

$$K(x)=\prod_{i=1}^d\mathcal{K}(x_i),\ \ \forall x\in\mathbb{R}^d,$$

where $\mathcal{K}:\mathbb{R}^1\to\mathbb{R}^1$ such that $\int\mathcal{K}=1$, $\mathrm{supp}(\mathcal{K})\subset[-1,1]$ and for some $M>0$

$$|\mathcal{K}(s)-\mathcal{K}(t)|\le M|s-t|,\ \ \forall s,t\in\mathbb{R}.$$

5.1. Functional classes of bandwidths

Let $\alpha_n\to 0, n\to\infty$, be given sequence and let

$$\omega_n=e^{-\sqrt{|\ln(\alpha_n)|}},\quad \Omega_n=e^{\ln^2(\alpha_n)}.$$

Set $H_n=\{h_s=e^{-s}, s\in\mathbb{N}\}\cap(0,\omega_n]$ and denote by $\mathfrak{H}_{1,n}$ the set of all measurable functions defined on $(-b,b)^d$ and taking values in H^d_n. Obviously $\mathfrak{H}_{1,n}\subset\mathfrak{H}(b)$.
Put for any $\mathfrak{h}\in\mathfrak{H}_{1,n}$ and any $\mathbf{s}=(s_1,\dots,s_d)\in\mathbb{N}^d$

$$\Upsilon_{\mathbf{s}}[\mathfrak{h}]=\cap_{j=1}^d\Upsilon_{s_j}[\mathfrak{h}_j],\quad \Upsilon_{s_j}[\mathfrak{h}_j]=\left\{x\in(-b,b)^d:\ \ \mathfrak{h}_j(x)=h_{s_j}\right\}.$$

Let $\tau\in(0,1)$ and $L>0$ be given constants. Define

$$\mathfrak{H}_n(\tau,L)=\left\{\mathfrak{h}\in\mathfrak{H}_{1,n}:\ \sum_{\mathbf{s}\in\mathbb{N}^d}\mu^\tau\left(\Upsilon_{\mathbf{s}}[\mathfrak{h}]\right)\le L\right\}.$$

Set $\mathbb{N}_p=\left\{\lfloor p\rfloor+1,\lfloor p\rfloor+2,\dots\right\}$ and introduce

$$\mathfrak{H}_{2,n}=\bigcup_{r\in\mathbb{N}_p}\mathfrak{H}_n(r),\qquad \mathfrak{H}_n(r)=\left\{\mathfrak{h}\in\mathfrak{H}_{1,n}:\ \ \left\|V_{\mathfrak{h}}^{-\frac{1}{2}}\right\|_{\frac{rp}{r-p},(-b,b)^d}\le\Omega_n\right\}.$$

We will establish (5.1) and (5.2) with $\mathfrak{H}_n=\mathfrak{H}^*_n(\tau,L):=\mathfrak{H}_{2,n}\cap\mathfrak{H}_n(\tau,L)$.

5.2. Verification of (5.1).

For any $\mathfrak{h}\in\mathfrak{H}_{2,n}$ define

$$\mathbb{N}_{p,n}(\mathfrak{h})=\mathbb{N}_p\cap\left[r_n(\mathfrak{h}),\infty\right),\quad r_n(\mathfrak{h})=\inf\left\{r\in\mathbb{N}_p:\ \ \mathfrak{h}\in\mathfrak{H}_n(r)\right\}.$$

Obviously $r_n(\mathfrak{h}) < \infty$ for any $\mathfrak{h} \in \mathfrak{H}_{2,n}$. For any $\mathfrak{h} \in \mathfrak{H}_{2,n}$ define

$$\Psi_n(\mathfrak{h}) = \inf_{r \in \mathbb{N}_{p,n}(\mathfrak{h})} C(r,\tau,L) \left\| V_{\mathfrak{h}}^{-\frac{1}{2}} \right\|_{\frac{rp}{r-p},(-b,b)^d},$$

where $C(r,\tau,L), \tau \in (0,1),\ L > 0$, can be found in (33), Section 3.2.2. Here we only mention that $C(r,\tau,L)$ is finite for any given r,τ,L but $\lim_{r\to\infty} C(r,\tau,L) = \infty$.
Note also that the condition $\mathfrak{h} \in \mathfrak{H}_{2,n}$ guarantees that $\Psi_n(\mathfrak{h}) < \infty$.

Theorem ((33), Corollary 1.). *For any $\tau \in (0,1)$ and any $q \geq 1$ one can find $n(\tau,q)$ such that for any $n \geq n(\tau,q)$*

$$\mathbb{E}\left\{ \sup_{\mathfrak{h} \in \mathfrak{H}_n^*(\tau,L)} \left[\left\| \xi_{\mathfrak{h}} \right\|_{p,(-b,b)^d} - \Psi_n(\mathfrak{h}) \right]_+ \right\}^q \leq \left(c\alpha_n \right)^q,$$

where c depends on $\mathcal{K}, p, q, b$ and d only.

Choosing $\alpha_n = c^{-1}\varepsilon_n\sqrt{n}$ we can assert that is (5.1) is verified with any $\Psi *_n (\cdot) \geq \Psi_n(\cdot)$.

5.3. Verification of (5.2).

The verification of (5.2) is mostly based on two facts.
First, the following result has been proved in (32), Lemma 1.

Lemma 1. *For any $d \geq 1$, $\tau \in (0, 1/d)$, $L > 0$ there exist $n(\tau, d, L)$ such that for all $n \geq n(\tau, L, d)$*

$$\mathfrak{h} \vee \eta \in \mathfrak{H}_n\big(d\tau, (2L)^d\big), \quad \forall\, \mathfrak{h}, \eta \in \mathfrak{H}_n(\tau, L).$$

Hence, setting

$$\Psi_n^*(\mathfrak{h}) = \inf_{r \in \mathbb{N}_{p,n}(\mathfrak{h})} C^*(r,\tau,L) \left\| V_{\mathfrak{h}}^{-\frac{1}{2}} \right\|_{\frac{rp}{r-p},(-b,b)^d},$$

where $C^*(r,\tau,L) = C(r,\tau,L) \vee C\big(r, d\tau, (2L)^d\big)$, we can assert that statement of Theorem 5.2 remains true for $\Psi_n^*(\cdot)$ as well if $\tau > 1/d$. It follows from the the fact that $\Psi_n^*(\cdot) \geq \Psi_n(\cdot)$.

Moreover, in view of Theorem 5.2 for all n large enough

$$\mathbb{E}\left\{ \sup_{\mathfrak{h} \in \mathfrak{H}_n^*(d\tau,(2L)^d)} \left[\left\| \xi_{\mathfrak{h}} \right\|_{p,(-b,b)^d} - \Psi_n^*(\mathfrak{h}) \right]_+ \right\}^q \leq \left(c\alpha_n \right)^q. \tag{5.3}$$

Since in view Lemma 1 if $\tau > 1/d$

$$\sup_{\mathfrak{h},\eta \in \mathfrak{H}_n^*(\tau,L)} \left[\left\| \xi_{\mathfrak{h}\vee\eta} \right\|_{p,(-b,b)^d} - \Psi_n^*(\mathfrak{h} \vee \eta) \right]_+ \leq \sup_{\rho \in \mathfrak{H}_n^*(d\tau,(2L)^d)} \left[\left\| \xi_{\rho} \right\|_{p,(-b,b)^d} - \Psi_n^*(\rho) \right]_+$$

we deduce from (5.3)

$$\mathbb{E}\left\{ \sup_{\mathfrak{h},\eta \in \mathfrak{H}_n^*(\tau,L)} \left[\left\| \xi_{\mathfrak{h}\vee\eta} \right\|_{p,(-b,b)^d} - \Psi_n^*(\mathfrak{h} \vee \eta) \right]_+ \right\}^q \leq \left(c\alpha_n \right)^q, \tag{5.4}$$

It remains to note that for any $1 \leq t < \infty$ and any $\mathfrak{h} \in \mathfrak{H}$

$$\left\| V_{\mathfrak{h}\vee\eta}^{-\frac{1}{2}} \right\|_{t,(-b,b)^d} \leq \left\| V_{\mathfrak{h}}^{-\frac{1}{2}} \right\|_{t,(-b,b)^d} \bigwedge \left\| V_{\eta}^{-\frac{1}{2}} \right\|_{t,(-b,b)^d},$$

that implies

$$\Psi_n^*(\mathfrak{h} \vee \eta) \leq \Psi_n^*(\mathfrak{h}) \wedge \Psi_n^*(\eta), \quad \forall \mathfrak{h}, \eta \in \mathfrak{H}. \tag{5.5}$$

The inequality (5.2) follows now from (5.4) and (5.5) if one chooses $\alpha_n = c^{-1}\varepsilon_n\sqrt{n}$.

The author is grateful to A. Goldenshluger who read the manuscript and made useful comments.

References

[1] A. Barron, L. Birgé, and P. Massart, Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** (1999), 301–413.

[2] L. Birgé, and P. Massart, Gaussian model selection. *J. Eur. Math. Soc.* **3** (2001)., no. 3, 203-268.

[3] F. Bunea, A.B. Tsybakov, and M. H. Wegkamp, Aggregation for Gaussian regression. *Ann. Statist.*. **35** (2007), no. 4, 1674–1697.

[4] T.T. Cai, Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.* **27** (1999), no. 3, 898–924.

[5] T.T. Cai, and M.G Low, On adaptive estimation of linear functionals. *Ann. Statist.* **33** (2005), no. 5, 2311–2343.

[6] T.T. Cai, and M.G Low, Optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **34** (2006), no. 5,, 2298–2325.

[7] L. Cavalier and A.B. Tsybakov, Penalized blockwise Stein's method, monotone oracle and sharp adaptive estimation. *Math. Methods Statist.* **10** (2001), 247–282.

[8] L. Cavalier and G.K. Golubev, Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. Statist.* **34** (2006), 1653–1677.

[9] A. Dalalyan and A.B. Tsybakov, Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning* **72** (2008), 39–61.

[10] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian and D. Picard, Density estimation by wavelet thresholding. *Ann. Statist.* **24** (1996), 508–539.

[11] S.Yu. Efroimovich and M.S. Pinsker An adaptive algorithm of nonparametric filtering. *Automat.Remote Control* **45** (1984), 58–65.

[12] S.Yu. Efroimovich, Non-parametric estimation of the density with unknown smoothness. *Theory Probab. Appl.* **30** (1986), 557–568.

[13] S.Yu. Efroimovich and M.G. Low, Adaptive estimates of linear functionals. *Probab. Theory and Related Fields* **98** (1994), 261–275.

[14] S.Yu. Efroimovich,. Adaptive estimation of and oracle inequalities for probability densities and characteristic functions. *Ann. Statist.* **36** (2008), 1127–1155.

[15] A. Goldenshluger, A universal procedure for aggregating estimators. *Ann. Statist.* **37** (2009), no. 1, 542–568.

[16] A. Goldenshluger and O.V. Lepski, Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probability Theory and Related Fields* **143** (2009), 41–71.

[17] A. Goldenshluger and O.V. Lepski, Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** (2011), 1608–1632.

[18] A. Goldenshluger and O.V. Lepski, General selection rule from the family of linear estimators. *Theory Probab. Appl.* **57** (2012), no. 2, 257–277.

[19] A. Goldenshluger and O.V. Lepski, On adaptive minimax density estimation on $\mathbb{R}^d$. *Probability Theory and Related Fields*, **159** (2014), 479–543.

[20] G.K. Golubev, Non-parametric estimation of smooth probability densities. *Probl. Inform. Transm.* **1** (1992), 52–62.

[21] G.K. Golubev, and M. Nussbaum, An adaptive spline estimate in nonparametric regression model. *Theory Probab. Appl.* **37** (1992), no. 3, 553–560.

[22] A.B Iouditski, O.V. Lepski and A.B. Tsybakov, Nonparametric estimation of composite functions. *Ann. Statist.*, **37** (2009), no. 3, 1360–1440.

[23] A. Juditsky and A. Nemirovski, Functional aggregation for nonparametric regression. *Ann. Statist.* **28** (2000), no. 3, 681–712.

[24] G. Kerkyacharian, O.V. Lepski and D. Picard, Nonlinear estimation in anisotropic multi–

index denoising. *Probability Theory and Related Fields*, **121** (2001), 137–170.
[25] N. Klutchnikoff, Pointwise adaptive estimation of a multivariate function *Math. Methods Statist.* **23** (2014), no. 2, 132–150.
[26] O.V. Lepskii, One problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** (1990), 459–470.
[27] O.V. Lepskii, Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** (1991), no. 4, 682–697.
[28] O.V. Lepskii, Asymptotically minimax adaptive estimation. II. Statistical model without optimal adaptation. Adaptive estimators. *Theory Probab. Appl.* **37** (1992), no. 3, 468–481.
[29] O.V. Lepskii, On problems of adaptive estimation in white Gaussian noise. In *Topics in nonparametric estimation*, pp 87–106. Adv. Soviet Math. 12, Amer. Math. Soc., Providence, RI, 1992.
[30] O.V. Lepskii and B.Ya. Levit, Adaptive minimax estimation of infinitely differentiable functions. *Math. Methods Statist.*, **7** (1998), no. 2, 123–156.
[31] O.V. Lepskii, Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure. *Ann. Statist.*, **41** (2013), no. 2, 1005–1034.
[32] O.V. Lepskii, Adaptive estimation over anisotropic functional classes via oracle approach. *Ann. Statist.*, **43** (2015), no. 3, 1178–1242.
[33] O.V. Lepskii, Upper functions for $\mathbb{L}_p$-norm of gaussian random fields. *Bernoulli*, **22** (2016), no. 2, 732–773.
[34] O.V. Lepskii, A new approach to estimator selection. *Bernoulli*, **24** (2018), no. 4A, 2776–2810.
[35] O.V. Lepskii and G. Rebelles, Structural adaptation in the density model. *Math. Stat. Learn*, (2021) to appear.
[36] G. Leung and A. R. Barron, Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52** (2006), no. 8, 3396–3410.
[37] A. S. Nemirovski, Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour)* 85–277. Lecture Notes in Math. 1738, Springer, Berlin, 1998.
[38] G. Rebelles, Pointwise adaptive estimation of a multivariate density under independence hypothesis. *Bernoulli* **21** (2015), no. 4, 1984–2023.
[39] P. Rigollet, Adaptive density estimation using the blockwise Stein method. *Bernoulli* **12** (2006), 351–370.
[40] P. Rigollet and A.B. Tsybakov, Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** (2011), no. 2, 731–771.
[41] A. Tsybakov, Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.* **26** (1998), 2420–2469.
[42] A. Tsybakov, Optimal rate of aggregation. *Proc. COLT.* Lecture Notes in Artificial Intelligence, **2777** (2003), 303–313.
[43] M.H. Wegkamp, Model selection in nonparametric regression. *Ann. Statist.* **31** (2003), 252–273.