Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality

Alexander Goldenshluger* Oleg Lepski†

Abstract

We address the problem of density estimation with $\mathbb{L}_p$–loss by selection of kernel estimators. We develop a selection procedure and derive corresponiding $\mathbb{L}_p$–risk oracle inequalities. It is shown that the proposed selection rule leads to the minimax estimator that is adaptive over a scale of the anisotropic Nikol'ski classes. The main technical tools used in our derivations are uniform bounds on the $\mathbb{L}_p$–norms of empirical processes developed recently in Goldenshluger and Lepski (2010).

Short Title: Selection of kernel density estimators
Keywords: density estimation, kernel estimators, $\mathbb{L}_p$–risk, oracle inequalities, adaptive estimation.
2000 AMS Subject Classification : 62G05, 62G20

1 Introduction

Let X be a random variable in $\mathbb{R}^d$ having density f with respect to the Lebesgue measure. We want to estimate f on the basis of the i.i.d. sample $\mathcal{X}_n = (X_1, \dots, X_n)$ drawn from f. By an estimator $\hat{f}$ we mean any measurable real function $\hat{f}(t) = \hat{f}(\mathcal{X}_n; t)$, $t \in \mathbb{R}^d$. Accuracy of an estimator $\hat{f}$ is measured by the $\mathbb{L}_s$–risk:

$$\mathcal{R}_s[\hat{f}, f] := \Big[\mathbb{E}_f \|\hat{f} - f\|_s^q\Big]^{1/q}, \quad s \in [1, \infty), \quad q \geq 1,$$

where $\mathbb{E}_f$ is the expectation with respect to the probability measure $\mathbb{P}_f$ of the observations $\mathcal{X}_n$. The objective is to develop an estimator of f with small $\mathbb{L}_s$–risk.

Kernel density estimates originate in Rosenblatt (1956) and Parzen (1962); this is one of the most popular techniques for estimating densities [Silverman (1986), Devroye and Györfi (1985)]. Let $K : \mathbb{R}^d \to \mathbb{R}$ be a fixed function such that $\int K(x)\mathrm{d}x = 1$ (we call such functions *kernels*). Given a *bandwidth vector* $h = (h_1, \dots, h_d)$, $h_i > 0$, the kernel estimator $\hat{f}_h$ of f is defined by

$$\hat{f}_h(t) = \frac{1}{nV_h} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} K_h(t - X_i), \tag{1}$$

*Department of Statistics, University of Haifa, 31905 Haifa, Israel. email: *goldensh@stat.haifa.ac.il*.
Supported by the ISF grant No. 389/07.
†Laboratoire d'Analyse, Topologie et Probabilités UMR CNRS 6632, Université de Provence, 39, rue F.Joliot Curie, 13453 Marseille, France. email: *lepski@cmi.univ-mrs.fr*

where $V_h := \prod_{i=1}^d h_i$, u/v for $u, v \in \mathbb{R}^d$ stands for the coordinate–wise division, and $K_h(\cdot) := V_h^{-1} K(\cdot/h)$. It is well–known that accuracy properties of $\hat{f}_h$ are determined by the choice of the bandwidth h, and *bandwidth selection* is the central problem in kernel density estimation. There are different approaches to the problem of bandwidth selection.

The *minimax approach* is based on the assumption that f belongs to a given class of densities $\mathbb{F}$, and accuracy of $\hat{f}_h$ is measured by its maximal $\mathbb{L}_s$–risk over the class $\mathbb{F}$,

$$\mathcal{R}_s[\hat{f}_h; \mathbb{F}] := \sup_{f \in \mathbb{F}} \mathcal{R}_s[\hat{f}_h; f].$$

Typically $\mathbb{F}$ is a class of smooth functions, e.g., the Hölder functional class. Then the bandwidth h is selected so that the maximal risk $\mathcal{R}_s[\hat{f}_h; \mathbb{F}]$ (or a reasonable upper bound on it) is minimized with respect to h. Such a choice leads to a deterministic bandwidth h depending on the sample size n, and on the underlying functional class $\mathbb{F}$. In many cases the resulting kernel estimator constructed in this way is *rate optimal* (or *optimal in order*) over the class $\mathbb{F}$. The minimax kernel density estimation with $\mathbb{L}_s$–risks on $\mathbb{R}^d$ was considered in Bretagnolle and Huber (1979), Ibragimov and Khasminskii (1980, 1981), Devroye and Györfi (1985), Hasminskii and Ibragimov (1990), Donoho et al. (1996), Kerkyacharian, Picard and Tribouley (1996), Juditsky and Lambert–Lacroix (2004), and Mason (2009) where further references can be found.

The *oracle approach* considers a set of kernel estimators $\mathcal{F}(\mathcal{H}) = \{\hat{f}_h, h \in \mathcal{H}\}$, and aims at a measurable data-driven choice $\hat{h} \in \mathcal{H}$ such that for every f from a *large* functional class the following $\mathbb{L}_s$*-risk oracle inequality* holds

$$\mathcal{R}_s[\hat{f}_{\hat{h}}; f] \leq C \inf_{h \in \mathcal{H}} \mathcal{R}_s[\hat{f}_h; f] + \delta_n. \tag{2}$$

Here C is a constant independent of f and n, and the remainder δ_n does not depend on f. Oracle inequalities with "small" remainder term δ_n and constant C close to one are of prime interest; they are key tools for establishing minimax and adaptive minimax results in estimation problems. To the best of our knowledge, oracle inequalities of the type (2) were established only in the cases $s = 1$ and $s = 2$. Devroye and Lugosi (1996, 1997, 2001) established oracle inequalities for $s = 1$. The case $s = 2$ was studied by Massart (2007, Chapter 7), Samarov and Tsybakov (2007), Rigollet and Tsybakov (2007) and Birgé (2008). The last cited paper contains a detailed discussion of recent developments in this area.

The contribution of this paper is two–fold. First, we propose a selection procedure for a set of kernel estimators, and establish the corresponding $\mathbb{L}_s$–risk, $s \in [1, \infty)$, oracle inequalies of the type (2). Second, we demonstrate that our selection rule leads to a minimax adaptive estimator over a scale of the anisotropic Nikolski's classes (see Section 3 below for the class definition).

More specifically, let $h^{\min} = \big(h_1^{\min}, \ldots, h_d^{\max}\big)$ and $h^{\max} = \big(h_1^{\max}, \ldots, h_d^{\max}\big)$ be two fixed vectors satisfying $0 < h_i^{\min} \leq h_i^{\max} \leq 1$, $\forall i$, and let

$$\mathcal{H} := \bigotimes_{i=1}^d \big[h_i^{\min}, h_i^{\max}\big]. \tag{3}$$

Consider the set of kernel estimators

$$\mathcal{F}(\mathcal{H}) = \{\hat{f}_h, h \in \mathcal{H}\}, \tag{4}$$

where $\hat{f}_h$ is given in (1). We propose a measurable choice $\hat{h} \in \mathcal{H}$ such that the resulting estimator $\hat{f} = \hat{f}_{\hat{h}}$ satisfies the following oracle inequality

$$\mathcal{R}_s[\hat{f}_{\hat{h}}; f] \;\le\; \inf_{h\in\mathcal{H}} \Big\{ \big(1 + 3\|K\|_1\big)\mathcal{R}_s[\hat{f}_h; f] + C_s(nV_h)^{-\gamma_s} \Big\} + \delta_{n,s}. \tag{5}$$

The constants C_s, γ_s, and the remainder term $\delta_{n,s}$ admit different expressions depending on the value of s.

- If $s \in [1, 2)$ then (5) holds for all densities f with $\gamma_s = 1 - \frac{1}{s}$, C_s depending on the kernel K only, and with

  $$\delta_{n,s} = c_1(\ln n)^{c_2} n^{1/s} \exp\big\{ -c_3 n^{2/s-1} \big\}$$

  for some constants c_i, $i = 1, 2, 3$.

- If $s \in [2, \infty)$ then (5) holds for all densities f uniformly bounded by a constant f_∞ with $\gamma_s = \frac{1}{2}$, C_s depending on K and f_∞ only, and with

  $$\delta_{n,s} = c_1(\ln n)^{c_2} n^{1/2} \exp\big\{ -c_3 V_{\max}^{-2/s} \big\}, \quad V_{\max} := V_{h^{\max}},$$

  for some constants c_i, $i = 1, 2, 3$. We emphasize that the proposed selection rule is fully data-driven and does not use information on the value of f_∞.

Thus the oracle inequality (5) holds with negligibly small (in terms of dependence on n) remander $\delta_{n,s}$ (by choice of $V_{\max}$ in the case $s \in [2, \infty)$). We stress that explicit non–asymptotic expressions for C_s, c_1, c_2 and c_3 are available. It is important to realize that the term $C_s(nV_h)^{-\gamma_s}$ is a tight upper bound on the stochastic error of the kernel estimator $\hat{f}_h$. This fact allows to derive rate optimal estimators that adapt to unknown smoothness of the density f. In particular, in Section 3 we apply our oracle inequalities in order to develop a rate optimal adaptive kernel estimator for the anisotropic Nikol'ski classes. Minimax estimation of densities from such classes was studied in Ibragimov and Khasminskii (1981), while the problem of adaptive estimation was not considered in the literature.

The paper is structured as follows. In Section 2 we define our selection rule and prove key oracle inequalities. Section 3 discusses adaptive rate optimal estimation of densitites for a scale of anisotropic Nikol'skii classes. Proofs of all results are given in Section 4.

2 Selection rule and oracle inequalities

Let $\mathcal{F}(\mathcal{H})$ be the set of kernel density estimators defined in (4). We want to select an estimator from the family $\mathcal{F}(\mathcal{H})$. For this purpose we need to impose some assumptions and establish notation that will be used in definition of our selection procedure.

2.1 Assumptions

The following assumptions on the kernel K will be used throughout the paper.

(K1) The kernel K satisfies the Lipschitz condition

$$|K(x) - K(y)| \le L_K|x - y|, \quad \forall x, y \in \mathbb{R}^d,$$

where $|\cdot|$ denotes the Euclidean distance. Moreover, K is compactly supported, and, without loss of generality, $\mathrm{supp}(K) \subseteq [-1/2, 1/2]^d$.

(K2) There exists a real number $\mathrm{k}_\infty < \infty$ such that $\|K\|_\infty \le \mathrm{k}_\infty$.

Assumptions (K1) and (K2) are rather standard in kernel density estimation. We note that Assumption (K1) can be weaken in several ways. For example, it suffices to assume that K belongs to the isotropic Hölder ball of functions $\mathbb{H}_d(\alpha, L_K)$ with any $\alpha > 0$ (in Assumption (K1) $\alpha = 1$).

Sometimes we will suppose that $f \in \mathbb{F}$, where

$$\mathbb{F} := \Big\{ p : \mathbb{R}^d \to \mathbb{R} : p \ge 0, \int p = 1,\ \|p\|_\infty \le \mathrm{f}_\infty < \infty \Big\},$$

and f_∞ is a fixed constant. Without loss of generality we assume that $\mathrm{f}_\infty \ge 1$.

2.2 Notation

For any $U : \mathbb{R}^d \to \mathbb{R}$ and $s \in [1, \infty)$ define

$$\rho_s(U) := \begin{cases} 4n^{1/s-1}\|U\|_s, & s \in [1,2), \\ n^{-1/2}\,\|U\|_2, & s = 2, \end{cases}$$

and if $s \in (2, \infty)$ then we set

$$\rho_s(U) \;:=\; c_s \Big\{ n^{-1/2} \Big(\int \Big[\int U^2(t-x) f(x) \mathrm{d}x \Big]^{s/2} \mathrm{d}t \Big)^{1/s} + 2n^{1/s-1}\|U\|_s \Big\},$$

where $c_s := 15s/\ln s$ is the best known constant in the Rosenthal inequality (Johnson, Schechtman and Zinn 1985). Observe that $\rho_s(U)$ depends on f when $s \in (2, \infty)$; hence we will also consider the empirical counterpart of $\rho_s(U)$:

$$\hat{\rho}_s(U) \;:=\; c_s \Big\{ n^{-1/2} \Big(\int \Big[\frac{1}{n} \sum_{i=1}^n U^2(t - X_i) \Big]^{s/2} \mathrm{d}t \Big)^{1/s} + 2n^{1/s-1}\|U\|_s \Big\}.$$

We put also

$$r_s(U) := \rho_s(U) \vee n^{-1/2}\|U\|_2, \quad \hat{r}_s(U) := \hat{\rho}_s(U) \vee n^{-1/2}\|U\|_2,$$

and

$$g_s(U) := \begin{cases} 32\rho_s(U), & s \in [1,2), \\ \frac{25}{3}\rho_2(U), & s = 2, \\ 32\hat{r}_s(U), & s > 2. \end{cases}$$

Armed with this notation we are ready to describe our selection rule.

2.3 Selection rule

The rule is based on auxiliary estimators $\{\hat{f}_{h,\eta}, h, \eta \in \mathcal{H}\}$ that are defined as follows: for every pair $h, \eta \in \mathcal{H}$ we let

$$\hat{f}_{h,\eta}(t) := \frac{1}{n} \sum_{i=1}^n [K_h * K_\eta](t - X_i),$$

where $*$ stands for the convolution on $\mathbb{R}^d$. Define also

$$\begin{aligned} m_s(h,\eta) &:= g_s(K_h) + g_s(K_h * K_\eta), \quad \forall h,\eta \in \mathcal{H}, \\ m_s^*(h) &:= \sup_{\eta\in\mathcal{H}} m_s(\eta,h), \quad \forall h\in\mathcal{H}. \end{aligned} \tag{6}$$

For every $h\in\mathcal{H}$ let

$$\hat{R}_h := \sup_{\eta\in\mathcal{H}} \big[\|\hat{f}_{h,\eta} - \hat{f}_\eta\|_s - m_s(h,\eta)\big]_+ + m_s^*(h). \tag{7}$$

Then the selected bandwidth $\hat{h}$ and the corresponding kernel density estimator are defined by

$$\hat{h} := \arg\inf_{h\in\mathcal{H}} \hat{R}_h, \quad \hat{f} = \hat{f}_{\hat{h}}. \tag{8}$$

The selection rule (6)–(8) is a refinement of the one introduced recently in Goldenshluger and Lepski (2008, 2009) for the Gaussian white noise model.

Remarks.

1. It is easy to check that Assumption (K1) implies that $\hat{R}_h$ and $m_s^*(h)$ are continuous random functions on the compact subset $\mathcal{H}\subset\mathbb{R}^d$. Thus, $\hat{h}$ exists and measurable (Jennrich 1969).

2. We call function $m_s(\cdot,\cdot)$ the *majorant*. In fact, if ξ_h and $\xi_{h,\eta}$ denote the stochastic errors of estimators $\hat{f}_h$ and $\hat{f}_{h,\eta}$ respectively, i.e., if

$$\begin{aligned} \xi_h(t) &:= \frac{1}{n}\sum_{i=1}^n \big[K_h(t-X_i) - \mathbb{E}_f K_h(t-X)\big], \\ \xi_{h,\eta}(t) &:= \frac{1}{n}\sum_{i=1}^n \big\{[K_h * K_\eta](t-X_i) - \mathbb{E}_f[K_h*K_\eta](t-X)\big\}, \end{aligned}$$

then it is seen from the proof of Theorems 1 and 2 below that $m_s(h,\eta)$ uniformly "majorates" $\|\xi_{h,\eta}-\xi_\eta\|_s$ in the sense that the expectation $\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}} \big[\|\xi_{h,\eta}-\xi_\eta\|_s - m_s(h,\eta)\big]_+^q$ is "small."

3. It is important to realize that majorant $m_s(h,\eta)$ does not depend on the density f to be estimated. The majorant is completely determined by kernel K and observations, and thus it is available to the statistician.

2.4 Oracle inequalities

Now we are in a position to establish oracle inequalities on the risk of the estimator $\hat{f} = \hat{f}_{\hat{h}}$ given by (7)–(8). Put

$$A_{\mathcal{H}} := \prod_{i=1}^d \big[1\vee \ln\big(h_i^{\max}/h_i^{\min}\big)\big], \quad B_{\mathcal{H}} := \big[1\vee\log_2\big(V_{\max}/V_{\min}\big)\big],$$

where from now on

$$V_{\min} := \prod_{i=1}^d h_i^{\min}, \qquad V_{\max} := \prod_{i=1}^d h_i^{\max}.$$

The next two statements, Theorem 1 and Theorem 2, provide oracle inequalities on the $\mathbb{L}_s$–risk of $\hat{f}$ in the cases $s\in[1,2]$ and $s\in(2,\infty)$ respectively.

Theorem 1. *Let Assumptions (K1) and (K2) hold.*

(i) *If $s \in [1,2)$ then for all f and $n \geq 4^{2s/(s-2)}$*

$$\mathcal{R}_s[\hat{f}; f] \;\leq\; \inf_{h\in\mathcal{H}} \Big[(1+3\|K\|_1)\mathcal{R}_s[\hat{f}_h, f] + C_1(nV_h)^{1/s-1}\Big] + C_2 A_{\mathcal{H}}^{4/q}\, n^{1/s} \exp\Big\{-\frac{2n^{2/s-1}}{37q}\Big\}. \tag{9}$$

(ii) *If $s = 2$ and $\mathrm{f}_\infty^2 V_{\max} + 4n^{-1/2} \leq 1/8$ then for all $f \in \mathbb{F}$*

$$\mathcal{R}_s[\hat{f}; f] \;\leq\; \inf_{h\in\mathcal{H}} \Big[(1+3\|K\|_1)\mathcal{R}_s[\hat{f}_h, f] + C_3(nV_h)^{-1/2}\Big] + C_4 A_{\mathcal{H}}^{4/q}\, n^{1/2} \exp\Big\{-\frac{1}{16q[\mathrm{f}_\infty^2 V_{\max} + 4n^{-1/2}]}\Big\}. \tag{10}$$

Here C_1 and C_3 are absolute constants, while C_2 and C_4 depend on L_K, k_∞, d and q only.

Theorem 2. *Let Assumptions (K1) and (K2) hold, $f \in \mathbb{F}$, $s \in (2,\infty)$, and assume that for some $C_1 = C_1(K,s,d) > 1$*

$$nV_{\min} > C_1, \quad V_{\max} \geq 1/\sqrt{n}.$$

If $n \geq C_2$ for some constant C_2 depending on L_K, k_∞, f_∞, d, and s only, then

$$\mathcal{R}_s[\hat{f}; f] \;\leq\; \inf_{h\in\mathcal{H}} \Big[(1+3\|K\|_1)\mathcal{R}_s[\hat{f}_h, f] + C_3 \mathrm{f}_\infty^{1/2}(nV_h)^{-1/2}\Big] + C_4 A_{\mathcal{H}}^{4/q} B_{\mathcal{H}}^{1/q} n^{1/2}\Big[\exp\{-C_5 b_{n,s}\} + \exp\Big\{-C_6 \mathrm{f}_\infty^{-1} V_{\max}^{-2/s}\Big\}\Big], \tag{11}$$

where $b_{n,s} := n^{4/s-1}$ if $s \in (2,4)$, and $b_{n,s} := [\mathrm{f}_\infty V_{\max}^{4/s}]^{-1}$ if $s \geq 4$. The constants C_i, $i = 3,\ldots,6$ depend on L_K, k_∞, d, q and s only.

Remarks.

1. All constants appearing in Theorems 1 and 2 can be expressed explicitly [see Lemmas 1 and 2 below and corresponding results in Goldenshluger and Lepski (2010) for details].

2. We will show that for given h the expected value of the stochastic error of the estimator $\hat{f}_h$, i.e. $(\mathbb{E}\|\xi_h\|_s^q)^{1/q}$, admits the upper bound of the order $O((nV_h)^{1/s-1})$ when $s \in [1,2)$, and $O((nV_h)^{-1/2})$ when $s \in (2,\infty)$. It is also obvious, that

$$\mathcal{R}_s[\hat{f}_h; f] \;\leq\; \|B_h\|_s + (\mathbb{E}_f\|\xi_h\|_s^q)^{1/q}\,,$$

where $B_h(f,t) := \int K_h(t-x)f(x)\mathrm{d}x - f(t)$, $t \in \mathbb{R}^d$. Thus, our estimator attains, up to a constant and reminder term, the minimum of the sum of the bias and the upper bound on the stochastic error. This form of the oracle inequality is convenient for deriving minimax and minimax adaptive results [see Section 3]. Indeed, bounds on the bias and the stochastic error are usually developed separately and require completely different techniques.

3. We note that $A_{\mathcal{H}} \leq O([\ln n]^d)$ and $B_{\mathcal{H}} \leq O(\ln n)$ for any set $\mathcal{H} \subset [0,1]^d$ such that $h_i^{\min} \geq O(n^{-c})$, $c > 0$, $\forall i = 1,\ldots,d$. If $s \in (2,\infty)$, and if the set of considered bandwidths $\mathcal{H}$ is such that $V_{\max} = [\varkappa \ln n]^{-s/2}$ for some $\varkappa > 0$ then the second term on the right hand side of (10) and (11) can be made negligibly small by choice of constant $\varkappa$. Observe that

conditions ensuring consistency of $\hat{f}_h$ are $nV_h \to \infty$ and $V_h \to 0$ as $n \to \infty$; thus the requirement $V_{\max} = [\varkappa \ln n]^{-s/2}$ is not restrictive. Note also that in the case $s \in [1,2)$ the second term on the right hand side of (9) is exponentially small in n for any $\mathcal{H}$.

4. The condition $V_{\max} \geq 1/\sqrt{n}$ is imposed only for the sake of convenience in presentation of our results. Clearly, we would like to have the set $\mathcal{H}$ as large as possible; hence consideration of vectors $h^{\max}$ such that $V_{\max} = V_{h^{\max}} \leq 1/\sqrt{n}$ has no much sense.

5. It should be also mentioned that if for $s \in [1,2)$ we impose additional conditions on f [e.g., such as the domination condition in Donoho et al. (1996, p. 514)], then the order of stochastic error of $\hat{f}_h$ can be improved to $O((nV_h)^{-1/2})$. It is well–known that smoothness condition alone is not sufficient for consistent density estimation on $\mathbb{R}^d$ with $\mathbb{L}_1$–losses (Ibragimov and Khasminskii 1981).

2.5 $\mathbb{L}_s$–risk oracle inequalities

As it was mentioned above, the oracle inequalities of Theorems 1 and 2 are useful for derivation of adaptive rate optimal estimators. Moreover, they are established under very mild assumptions on the density f. However, traditionally oracle inequalities compare the risk of a proposed estimator to the risk of the best estimator in the given family, cf. (2). The natural question is whether an $\mathbb{L}_s$*–risk oracle inequality* of the type (2) can be derived from the results of Theorems 1 and 2. In this section we provide an answer to this question. We will be mostly interested in finding minimal assumptions on the underlying density f that are sufficient for establishing the $\mathbb{L}_s$–risk oracle inequality. It will be shown that this problem is directly related to establishing a lower bound on the term $(\mathbb{E}_f \|\xi_h\|_s^q)^{1/q}$.

Let $\mu \in (0,1)$ and $\nu > 0$ be fixed real numbers. Denote by $\mathbb{F}_{\mu,\nu}$ the set of all probability densities p satisfying the following condition:

$$\exists\, B \in \mathcal{B}(\mathbb{R}^d):\ \mathrm{mes}(B) \leq \nu,\quad \int_B p \geq \mu.$$

Here $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ–algebra on $\mathbb{R}^d$, and $\mathrm{mes}(\cdot)$ is the Lebesgue measure on $\mathbb{R}^d$.

Below we will assume that $f \in \mathbb{F}_{\mu,\nu}$ for some μ and ν. This condition is very weak. For example, if $\mathcal{F}$ is a set of densitites such that either (i) $\mathcal{F}$ is a totally bounded subset of $\mathbb{L}_1(\mathbb{R}^d)$; or (ii) the family of probability measures $\{\mathbb{P}_f, f \in \mathcal{F}\}$ is tight, then for any $\mu \in (0,1)$ there exists $0 < \nu < \infty$ such that $\mathcal{F} \subseteq \mathbb{F}_{\mu,\nu}$. The statement (i) is a consequence of the Kolmogorov-Riesz compactness theorem.

Theorem 3. *Let $s \in [2,\infty)$ and suppose that assumptions of Theorem 1(ii) and Theorem 2 are fulfilled. If $s > 2$ then assume additionally that $f \in \mathbb{F}_{\mu,\nu}$ for some μ and ν, and*

$$V_{\max} \leq 2^{-1}\mu \left[\frac{||K||_2}{||K||_1}\right]^2.$$

If $n \geq C_1 = C_1(L_K, \mathrm{k}_\infty, \mathrm{f}_\infty, d, s)$ then there exist a constant $C_0 > 0$ ($C_0 = C_0(K)$ if $s = 2$, and $C_0 = C_0(K, \mu, \nu, s)$ if $s > 2$) such that

$$\mathcal{R}_s[\hat{f}; f] \leq C_0 \inf_{h \in \mathcal{H}} \mathcal{R}_s[\hat{f}_{\hat{h}}; f] + C_2 A_{\mathcal{H}}^{4/q} B_{\mathcal{H}}^{1/q} n^{1/2}\Big[\exp\{-C_3 b_{n,s}\} + \exp\Big\{-C_4 \mathrm{f}_\infty^{-1} V_{\max}^{-2/s}\Big\}\Big],$$

where $b_{n,s} := n^{4/s-1}$ if $s \in (2,4)$, and $b_{n,s} := [\mathrm{f}_\infty V_{\max}^{4/s}]^{-1}$ if $s \geq 4$. The constants C_i depend on L_K, k_∞, d, q and s only.

The proof indicates that Theorem 3 follows from the fact that for any $s \in [2, \infty)$ one has

$$[\mathbb{E}_f \|\xi_h\|_s^q]^{1/q} \geq c(nV_h)^{-1/2}, \quad \forall h, \tag{12}$$

where $c > 0$ is a constant. This lower bound holds under very weak conditions on the density f (for arbitrary f is $s = 2$ and $f \in \mathbb{F}_{\mu,\nu}$ if $s > 2$). In order to prove the similar $\mathbb{L}_s$–risk oracle inequality in the case $s \in [1, 2)$ it would be sufficient to show that $[\mathbb{E}_f \|\xi_h\|_s^q]^{1/q} \geq c(nV_h)^{-1+1/s}$ for any h. However, the last lower bound cannot hold in such generality as (12). In particular, according to remark 5 after Theorem 2, $[\mathbb{E}_f \|\xi_h\|_s^q]^{1/q} \leq c(nV_h)^{-1/2}$ for all h under a tail domination condition (e.g., for compactly supported densities). Under such a domination condition the corresponding $\mathbb{L}_s$–risk oracle inequality can be easily established using the same arguments as in the proof of Theorem 3.

3 Adaptive estimation of densities with anisotropic smoothness

In this section we illustrate the use of oracle inequalities of Theorems 1 and 2 for derivation of adaptive rate optimal density estimators.

We start with the definition of the *anisotropic Nikol'skii class of functions.*

Definition 1. *Let $\alpha = (\alpha_1, \ldots, \alpha_d)$, $\alpha_i > 0$ and $L > 0$. We say that density $f : \mathbb{R}^d \to \mathbb{R}$ belongs to the anisotropic Nikol'ski class $N_{s,d}(\alpha, L)$ of functions if*

(i) $\|D_i^{\lfloor \alpha_i \rfloor} f\|_s \leq L$, *for all* $i = 1, \ldots, d$;

(ii) *for all* $i = 1, \ldots, d$, *and all* $z \in \mathbb{R}^1$

$$\left\{ \int \left| D_i^{\lfloor \alpha_i \rfloor} f(t_1, \ldots, t_i + z, \ldots, t_d) - D_i^{\lfloor \alpha_i \rfloor} f(t_1, \ldots, t_i, \ldots, t_d) \right|^s \mathrm{d}t \right\}^{1/s} \leq L|z|^{\alpha_i - \lfloor \alpha_i \rfloor}.$$

Here $D_i^k f$ denotes the kth order partial derivative of f with respect to the variable t_i, and $\lfloor \alpha_i \rfloor$ is the largest integer strictly less than α_i.

The functional classes $N_{s,d}(\alpha, L)$ were considered in approximation theory by Nikol'skii; see, e.g., Nikol'skii (1969). Minimax estimation of densities from the class $N_{s,d}(\alpha, L)$ was considered in Ibragimov and Khasminskii (1981). We refer also to Kerkyacharian, Lepski and Picard (2001) where the problem of adaptive estimation over a scale of classes $N_{s,d}(\alpha, L)$ was treated for the Gaussian white noise model.

Consider the following family of kernel estyimators. Let u be an integrable, compactly supported function on $\mathbb{R}$ such that $\int u(y)\mathrm{d}y = 1$. As in Kerkyacharian, Lepski and Picard (2001), for some integer number l we put

$$u_l(y) := \sum_{k=1}^{l} \binom{l}{k} (-1)^{k+1} \frac{1}{k} u\left(\frac{y}{k}\right),$$

and define

$$K(t) := \prod_{i=1}^{d} u_l(t_i), \quad t = (t_1, \ldots, t_d). \tag{13}$$

The kernel K constructed in this way is bounded and compactly supported, and it is easily verified that

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

where $k = (k_1, \dots, k_d)$ is the multi–index, $k_i \geq 0$, $|k| = k_1 + \cdots + k_d$, and $t^k = t_1^{k_1} \cdots t_d^{k_d}$ for $t = (t_1, \dots, t_d)$.

For fixed $\alpha = (\alpha_1, \dots, \alpha_d)$ let $\bar{\alpha}$ be defined by the relation $1/\bar{\alpha} = \sum_{i=1}^d (1/\alpha_i)$. Define also

$$\varphi_{n,s}(\bar{\alpha}) := L^{-\gamma_s/(\bar{\alpha}+\gamma_s)} n^{-\gamma_s\bar{\alpha}/(\bar{\alpha}+\gamma_s)}, \quad \gamma_s := \begin{cases} 1 - 1/s, & s \in (1, 2], \\ 1/2, & s \in (2, \infty). \end{cases}$$

Theorem 4. *Let $\mathcal{F}(\mathcal{H})$ be the family of kernel estimators defined in (1), (3) and (4) that is associated with the kernel (13). Let $\hat{f}$ denote the estimator given by selection according to our rule (6)–(8) from the family $\mathcal{F}(\mathcal{H})$.*

(i) *Let $s \in (1, 2)$, and assume that $h_i^{\min} = 1/n$ and $h_i^{\max} = 1$, $\forall i = 1, \dots, d$. Then for any class $N_{s,d}(\alpha, L)$ such that $\max_{i=1,\dots,d} \lfloor \alpha_i \rfloor \leq l - 1$, $L > 0$ one has*

$$\limsup_{n\to\infty} \Big\{ [\varphi_{n,s}(\bar{\alpha})]^{-1} \mathcal{R}_s[\hat{f}; N_{s,d}(\alpha, L)] \Big\} \leq C < \infty.$$

(ii) *Let $s \in [2, \infty)$, and assume that $h_i^{\min} = \varkappa_1/n$ and $h_i^{\max} = [\varkappa_2 \ln n]^{-s/(2d)}$, $\forall i = 1, \dots, d$ for some constants $\varkappa_1$ and $\varkappa_2$. Then for any class $N_{s,d}(\alpha, L)$ such that $\max_{i=1,\dots,d} \lfloor \alpha_i \rfloor \leq l - 1$, $L > 0$ one has*

$$\limsup_{n\to\infty} \Big\{ [\varphi_{n,s}(\bar{\alpha})]^{-1} \mathcal{R}_s[\hat{f}; N_{s,d}(\alpha, L)] \Big\} \leq C < \infty.$$

It is well–known that $\varphi_{n,s}(\bar{\alpha})$ is the minimax rate of convergence in estimation of densitites from the class $N_{s,d}(\alpha, L)$ [see Ibragimov and Khasminskii (1981) and Hasminskii and Ibragimov (1990)]. Therefore Theorem 4 shows that our estimator $\hat{f}$ is adaptive minimax over a scale of the classes $N_{s,d}(\alpha, L)$.

4 Proofs

First we recall that accuracy of estimators $\hat{f}_h$ and $\hat{f}_{h,\eta}$, $h, \eta \in \mathcal{H}$ is characterized by the bias and stochastic error given by

$$\begin{aligned} B_h(f,t) &:= \int K_h(t-x) f(x)\mathrm{d}x - f(t), \\ \xi_h(t) &:= \frac{1}{n}\sum_{i=1}^n \big[K_h(t - X_i) - \mathbb{E}_f K_h(t-X)\big], \end{aligned}$$

and

$$\begin{aligned} B_{h,\eta}(f,t) &:= \int [K_h * K_\eta](t-x) f(x)\mathrm{d}x - f(t), \\ \xi_{h,\eta}(t) &:= \frac{1}{n}\sum_{i=1}^n \big\{[K_h * K_\eta](t - X_i) - \mathbb{E}_f [K_h * K_\eta](t-X)\big\}. \end{aligned}$$

respectively.

The proofs extensively use results from Goldenshluger and Lepski (2010); in what follows for the sake of brevity we refer to this paper as GL (2010).

4.1 Auxiliary results

We start with two auxiliary lemmas that establish probability and moment bounds on $\mathbb{L}_s$–norms of the processes ξ_h and $\xi_{h,\eta}$. Proofs of these results are given in Appendix.

Lemma 1. *Let Assumptions (K1) and (K2) hold.*

(i) *If $s \in [1,2)$ then for all $n \geq 4^{2s/(2-s)}$ one has*

$$\Big\{\mathbb{E}_f \sup_{h\in\mathcal{H}} \Big[\|\xi_h\|_s - 32\rho_s(K_h)\Big]_+^q\Big\}^{1/q} \;\leq\; \delta_{n,s}^{(1)} \;:=\; C_1 A_{\mathcal{H}}^{2/q}\, n^{1/s} \exp\Big\{-\frac{2n^{2/s-1}}{37q}\Big\}, \tag{14}$$

$$\Big\{\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}} \Big[\|\xi_{h,\eta}\|_s - 32\rho_s(K_h * K_\eta)\Big]_+^q\Big\}^{1/q} \;\leq\; \delta_{n,s}^{(2)}$$
$$:= C_2 A_{\mathcal{H}}^{4/q}\, n^{1/s} \exp\Big\{-\frac{2n^{2/s-1}}{37q}\Big\}. \tag{15}$$

(ii) *Let $f \in \mathbb{F}$, and assume that $8[\mathrm{f}_\infty^2 V_{\max} + 4n^{-1/2}] \leq 1$; then for all $f \in \mathbb{F}$ one has*

$$\Big\{\mathbb{E}_f \sup_{h\in\mathcal{H}} \Big[\|\xi_h\|_2 - \frac{25}{3}\rho_2(K_h)\Big]_+^q\Big\}^{1/q} \;\leq\; \delta_{n,2}^{(1)}$$
$$:=\; C_3 A_{\mathcal{H}}^{2/q}\, n^{1/2} \exp\Big\{-\frac{1}{16q[V_{\max}\mathrm{f}_\infty^2 + 4n^{-1/2}]}\Big\}, \tag{16}$$

$$\Big\{\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}} \Big[\|\xi_{h,\eta}\|_2 - \frac{25}{3}\rho_2(K_h * K_\eta)\Big]_+^q\Big\}^{1/q} \;\leq\; \delta_{n,2}^{(2)}$$
$$:=\; C_4 A_{\mathcal{H}}^{4/q}\, n^{1/2} \exp\Big\{-\frac{1}{16q[\mathrm{f}_\infty^2 V_{\max} + 4n^{-1/2}]}\Big\}. \tag{17}$$

The constants C_i, $i = 1, \ldots, 4$ depend on L_K, k_∞, d and q only.

Lemma 2. *Let Assumptions (K1) and (K2) hold, $f \in \mathbb{F}$, $s > 2$, and assume that*

$$n \geq C_1, \quad nV_{\min} > C_2, \quad V_{\max} \geq 1/\sqrt{n}.$$

Then the following statements hold:

$$\Big\{\mathbb{E}_f \sup_{h\in\mathcal{H}} \Big[\|\xi_h\|_s - 32\,\hat{r}_s(K_h)\Big]_+^q\Big\}^{1/q} \;\leq\; \delta_{n,s}^{(1)}$$
$$:=\; C_3 A_{\mathcal{H}}^{2/q}\, B_{\mathcal{H}}^{1/q}\, n^{1/2} \exp\Big\{-\frac{C_4}{\mathrm{f}_\infty V_{\max}^{2/s}}\Big\}, \tag{18}$$

$$\Big\{\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}} \Big[\|\xi_{h,\eta}\|_s - 32\,\hat{r}_s(K_h * K_\eta)\Big]_+^q\Big\}^{1/q} \;\leq\; \delta_{n,s}^{(2)}$$
$$:=\; C_5 A_{\mathcal{H}}^{4/q}\, B_{\mathcal{H}}^{1/q}\, n^{1/2} \exp\Big\{-\frac{C_6}{\mathrm{f}_\infty V_{\max}^{2/s}}\Big\}. \tag{19}$$

In addition, for any $H_1 \subseteq \mathcal{H}$ and $H_2 \subseteq \mathcal{H}$

$$\mathbb{E}_f \sup_{h\in H_1} [\hat{r}_s(K_h)]^q \;\le\; (1+8c_s)^q \sup_{h\in H_1} [r_s(K_h)]^q + \; C_7 A_{\mathcal{H}}^2\, B_{\mathcal{H}}\, n^{q(s-2)/(2s)} \exp\big\{-C_8 b_{n,s}\big\}, \tag{20}$$

$$\mathbb{E}_f \sup_{(h,\eta)\in H_1\times H_2} [\hat{r}_s(K_h * K_\eta)]^q \;\le\; (1+8c_s)^q \sup_{(h,\eta)\in H_1\times H_2} [r_s(K_h * K_\eta)]^q + C_9 A_{\mathcal{H}}^4\, B_{\mathcal{H}}\, n^{q(s-2)/(2s)} \exp\big\{-C_{10} b_{n,s}\big\}, \tag{21}$$

where $b_{n,s} := n^{4/s-1}$ if $s \in (2,4)$, and $b_{n,s} := [\mathrm{f}_\infty V_{\max}^{4/s}]^{-1}$ if $s\in[4,\infty)$. The constants C_i, $i=2,\dots,10$ depend on L_K, k_∞, d, q and s only, while C_1 depends also on f_∞.

4.2 Proof of Theorems 1 and 2

The proofs of both theorems (which we break in several steps) follow along the same lines.

1^0. First we show that for any $h,\eta\in\mathcal{H}$

$$B_{h,\eta}(f,x) = B_\eta(f,x) + \int K_\eta(y-x) B_h(f,y)\mathrm{d}y \tag{22}$$

$$= B_h(f,x) + \int K_h(y-x) B_\eta(f,y)\mathrm{d}y. \tag{23}$$

Indeed, by the Fubini theorem

$$\begin{aligned}\int [K_h * K_\eta](t-x) f(t)\mathrm{d}t &= \int \Big[\int K_h(t-y)K_\eta(y-x)\mathrm{d}y\Big] f(t)\mathrm{d}t \\ &= \int \Big[\int K_h(t-y) f(t)\mathrm{d}t - f(y)\Big] K_\eta(y-x)\mathrm{d}y + \int K_\eta(y-x) f(y)\mathrm{d}y \\ &= \int K_\eta(y-x) f(y)\mathrm{d}y + \int K_\eta(y-x) B_h(f,y)\mathrm{d}y.\end{aligned}$$

Subtracting $f(x)$ from the both sides of the last equality we come to (22); (23) follows similarly.

2^0. Let $m_s(\cdot,\cdot)$ and $m_s^*(\cdot)$ be given by (6), and define

$$\delta_{n,s} := \Big\{\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}} \big[\|\xi_{h,\eta} - \xi_\eta\|_s - m_s(h,\eta)\big]_+^q\Big\}^{1/q}. \tag{24}$$

Let $\hat{f} = \hat{f}_{\hat{h}}$ be the estimator defined in (7)-(8). Our first goal is to prove that

$$\mathcal{R}_s[\hat{f}; f] \;\le\; \inf_{h\in\mathcal{H}} \Big\{\big(1+3\|K\|_1\big)\,\mathcal{R}_s[\hat{f}_h; f] + 2\big(\mathbb{E}_f\big[m_s^*(h)\big]^q\big)^{1/q}\Big\} + 3\delta_{n,s}. \tag{25}$$

By the triangle inequality for any $\eta\in\mathcal{H}$

$$\|\hat{f}_{\hat{h}} - f\|_s \le \|\hat{f}_{\hat{h}} - \hat{f}_{\hat{h},\eta}\|_s + \|\hat{f}_{\hat{h},\eta} - \hat{f}_\eta\|_s + \|\hat{f}_\eta - f\|_s, \tag{26}$$

and we are going to bound the first two terms on the right hand side.

Define

$$\bar{B}_h(f) := \sup_{\eta\in\mathcal{H}} \Big\| \int K_\eta(t-\cdot) B_h(f,t)\mathrm{d}t \Big\|_s, \quad h\in\mathcal{H}.$$

We have for any $h\in\mathcal{H}$

$$\begin{aligned}
\hat{R}_h - m_s^*(h) &= \sup_{\eta\in\mathcal{H}} \Big[\|\hat{f}_{h,\eta}-\hat{f}_\eta\|_s - m_s(h,\eta)\Big] \\
&\leq \sup_{\eta\in\mathcal{H}} \Big[\|B_{h,\eta}(f,\cdot)-B_\eta(f,\cdot)\|_s + \|\xi_{h,\eta}-\xi_\eta\|_s - m_s(h,\eta)\Big] \\
&\leq \bar{B}_h(f) + \sup_{\eta\in\mathcal{H}} \Big[\|\xi_{h,\eta}-\xi_\eta\|_s - m_s(h,\eta)\Big]_+ =: \bar{B}_h(f)+\zeta.
\end{aligned}$$

Here the second line is by the triangle inequality and the third line is by (22) and definition of $\bar{B}_h(f)$. Therefore for any $h\in\mathcal{H}$ one has

$$\hat{R}_h \leq \bar{B}_h(f) + m_s^*(h) + \zeta. \tag{27}$$

By (22), (23) for any $h,\eta\in\mathcal{H}$

$$\begin{aligned}
\|\hat{f}_{h,\eta}-\hat{f}_h\|_s &\leq \|B_{h,\eta}(f,\cdot)-B_h(f,\cdot)\|_s + \|\xi_{h,\eta}-\xi_h\|_s \\
&\leq \bar{B}_\eta(f) + \zeta + \sup_{\eta\in\mathcal{H}} m_s(\eta,h) \\
&= \bar{B}_\eta(f) + m_s^*(h) + \zeta \;\leq\; \bar{B}_\eta(f) + \hat{R}_h + \zeta,
\end{aligned}$$

where the last inequality is by definition of $\hat{R}_h$. In particular, letting $h=\hat{h}$ we have that for any $\eta\in\mathcal{H}$

$$\begin{aligned}
\|\hat{f}_{\hat{h},\eta}-\hat{f}_{\hat{h}}\|_s &\leq \bar{B}_\eta(f) + \hat{R}_{\hat{h}} + \zeta \\
&\leq \bar{B}_\eta(f) + \hat{R}_\eta + \zeta \;\leq\; 2\bar{B}_\eta(f) + m_s^*(\eta) + 2\zeta,
\end{aligned} \tag{28}$$

where we have used that $\hat{R}_{\hat{h}} \leq \hat{R}_\eta$, $\forall\eta\in\mathcal{H}$ and (27).

Furthermore, for any $\eta\in\mathcal{H}$

$$\begin{aligned}
\|\hat{f}_{\hat{h},\eta}-\hat{f}_\eta\|_s &= \|\hat{f}_{\hat{h},\eta}-\hat{f}_\eta\|_s - m_s(\hat{h},\eta) + m_s(\hat{h},\eta) \\
&\leq \hat{R}_{\hat{h}} \leq \hat{R}_\eta \leq \bar{B}_\eta(f) + m_s^*(\eta) + \zeta,
\end{aligned} \tag{29}$$

where the first inequality is by definition of $\hat{R}_h$, the second inequality is by the definition of $\hat{h}$, and the last inequality follows from (27).

Combining (26), (28) and (29) we get for any $\eta\in\mathcal{H}$ that

$$\begin{aligned}
\|\hat{f}_{\hat{h}}-f\|_s &\leq \|\hat{f}_{\hat{h}}-\hat{f}_{\hat{h},\eta}\|_s + \|\hat{f}_{\hat{h},\eta}-\hat{f}_\eta\|_s + \|\hat{f}_\eta - f\|_s \\
&\leq \|\hat{f}_\eta - f\|_s + 3\bar{B}_\eta(f) + 2m_s^*(\eta) + 3\zeta.
\end{aligned}$$

Taking this expression to the power q, computing the expectation and using the fact that $[\mathbb{E}_f|\zeta|^q]^{1/q}=\delta_{n,s}$ we obtain

$$\mathcal{R}_s[\hat{f};f] \leq \inf_{h\in\mathcal{H}} \Big\{ \mathcal{R}_s[\hat{f}_h;f] + 3\bar{B}_h(f) + 2\big(\mathbb{E}_f\big[m_s^*(h)\big]^q\big)^{1/q}\Big\} + 3\delta_{n,s}. \tag{30}$$

By the Young inequality $\|\bar{B}_h(f)\|_s \leq \big(\sup_{\eta\in\mathcal{H}}\|K_\eta\|_1\big)\|B_h(\cdot,f)\|_s = \|K\|_1\|B_h(\cdot,f)\|_s$. In addition, see (36),

$$\|B_h(\cdot,f)\|_s \leq \mathcal{R}_s[\hat{f}_h;f], \;\; \forall h\in\mathcal{H}$$

Combining this with (30) we complete the proof of (25).

3^0. Lemmas 1 and 2 lead to an upper bound on the quantity $\delta_{n,s}$ given in (24). Indeed, by definition of $m_s(\cdot,\cdot)$ [see (6)] we have

$$\begin{aligned}
\delta_{n,s} &= \Big\{\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}} \big[\|\xi_{h,\eta}-\xi_\eta\|_s - m_s(h,\eta)\big]_+^q\Big\}^{1/q} \\
&\leq \Big\{\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}} \big[\|\xi_{h,\eta}\|_s - g_s(K_h*K_\eta)\big]_+^q\Big\}^{1/q} + \Big\{\mathbb{E}_f \sup_{h\in\mathcal{H}} \big[\|\xi_h\|_s - g_s(K_h)\big]_+^q\Big\}^{1/q} \\
&\leq \delta_{n,s}^{(1)} + \delta_{n,s}^{(2)}, \qquad (31)
\end{aligned}$$

where expressions for $\delta_{n,s}^{(1)}$ and $\delta_{n,s}^{(2)}$ depending on the value of $s\in[1,\infty)$ are given in (14)–(15), (16)–(17), and (18)–(19).

In order to apply (25) it remains to to bound $\{\mathbb{E}_f[m_s^*(h)]^q\}^{1/q}$.

4^0. We start with the case $s\in[1,2)$. Here, by definiton,

$$\begin{aligned}
m_s^*(h) &= \sup_{\eta\in\mathcal{H}} m_s(\eta,h) = g_s(K_h) + \sup_{\eta\in\mathcal{H}} g_s(K_\eta*K_h) \\
&= 128 n^{1/s-1}\big(\|K_h\|_s + \sup_{\eta\in\mathcal{H}}\|K_h*K_\eta\|_s\big) \leq 128[1+\|K\|_1](nV_h)^{1/s-1}.
\end{aligned}$$

Therefore applying (25), and taking into account (31), (14) and (15) we come to the statement (i) of Theorem 1.

The statement (ii) of Theorem 1 dealing with the case $s=2$ follows similarly by application of (25) and (31), (16) and (17). This completes the proof of Theorem 1.

5^0. Now consider the case $s\in(2,\infty)$. Because

$$\begin{aligned}
m_s^*(h) &= \sup_{\eta\in\mathcal{H}} m_s(\eta,h) = g_s(K_h) + \sup_{\eta\in\mathcal{H}} g_s(K_\eta*K_h) \\
&= 32\hat{r}_s(K_h) + 32\sup_{\eta\in\mathcal{H}}\hat{r}_s(K_\eta*K_h), \qquad (32)
\end{aligned}$$

it suffices to bound from above $[\mathbb{E}_f|\hat{r}_s(K_h)|^q]^{1/q}$ and $[\mathbb{E}_f\sup_{\eta\in\mathcal{H}}|\hat{r}_s(K_h*K_\eta)|^q]^{1/q}$. Using (20) of Lemma 2 with $H_1=\{h\}$ we have

$$[\mathbb{E}_f|\hat{r}_s(K_h)|^q]^{1/q} \leq c_1 r_s(K_h) + c_2 A_{\mathcal{H}}^{2/q} B_{\mathcal{H}}^{1/q} n^{(s-2)/(2s)}\exp\{-c_3 b_{n,s}\}.$$

In addition, by the Young inequality

$$\begin{aligned}
\rho_s(K_h) &= c_s n^{-1/2}\|K_h^2*f\|_{s/2}^{1/2} + n^{1/s-1}\|K_h\|_s \\
&\leq c_s n^{-1/2}\|K_h\|_2\|\sqrt{f}\|_s + (nV_h)^{-1+1/s}\|K\|_s \\
&\leq c_s \mathrm{f}_\infty^{1/2}\|K\|_2(nV_h)^{-1/2} + \|K\|_s(nV_h)^{-1+1/s} \ \leq\ c_4\mathrm{f}_\infty^{1/2}(nV_h)^{-1/2};
\end{aligned}$$

hence

$$[\mathbb{E}_f|\hat{r}_s(K_h)|^q]^{1/q} \leq c_5\mathrm{f}_\infty^{1/2}(nV_h)^{-1/2} + c_2 A_{\mathcal{H}}^{2/q} B_{\mathcal{H}}^{1/q} n^{(s-2)/(2s)}\exp\{-c_3 b_{n,s}\}. \qquad (33)$$

Now, applying (21) with $H_1=\{h\}$ and $H_2=\mathcal{H}$ we obtain

$$\big[\mathbb{E}_f\sup_{\eta\in\mathcal{H}}|\hat{r}_s(K_h*K_\eta)|^q\big]^{1/q} \leq c_6\sup_{\eta\in\mathcal{H}} r_s(K_h*K_\eta) + c_7 A_{\mathcal{H}}^{4/q} B_{\mathcal{H}}^{1/q} n^{(s-2)/(2s)}\exp\{-c_8 b_{n,s}\}.$$

In addition, similarly to the above

$$\begin{aligned}\sup_{\eta\in\mathcal{H}}\rho_s(K_h*K_\eta) &\le \sup_{\eta\in\mathcal{H}}\Big\{c_s n^{-1/2}\|K_h*K_\eta\|_2\|\sqrt{f}\|_s+n^{-1+1/s}\|K_h*K_\eta\|_s\Big\}\\ &\le c_8\mathrm{f}_\infty^{1/2}\sup_{\eta\in\mathcal{H}}[n(V_h\vee V_\eta)]^{1/2}\le\ c_9\mathrm{f}_\infty^{1/2}(nV_h)^{-1/2}.\end{aligned}$$

Therefore the last two bounds yield

$$\big[\mathbb{E}_f\sup_{\eta\in\mathcal{H}}|\hat{r}_s(K_h*K_\eta)|^q\big]^{1/q}\ \le\ c_{10}\mathrm{f}_\infty^{1/2}(nV_h)^{-1/2}+c_7A_{\mathcal{H}}^{4/q}B_{\mathcal{H}}^{1/q}n^{(s-2)/(2s)}\exp\{-c_8b_{n,s}\}.$$

This along with (33) and (32) results in

$$[\mathbb{E}_f|m_s^*(K_h)|^q]^{1/q}\ \le\ c_{11}\mathrm{f}_\infty^{1/2}(nV_h)^{-1/2}+c_{12}A_{\mathcal{H}}^{4/q}B_{\mathcal{H}}^{1/q}n^{(s-2)/(2s)}\exp\{-c_{13}b_{n,s}\}.$$

Combining this bound with (18), (19) and (31), and applying (25) we complete the proof of Theorem 2. ∎

4.3 Proof of Theorem 3

Throughout the proof we denote by $c_0,c_1,\dots,$ the positive constants depending only on the kernel K, the index s and the quantity f_∞. We divide the proof in several steps.

1^0. Let us prove that for any $q\ge1$ and $h\in\mathcal{H}$

$$3\mathcal{R}_s[\hat{f}_h;f]\ge\|B_h(f)\|_s+\mathbb{E}_f\|\xi_h\|_s. \tag{34}$$

Indeed, in view of the Jensen inequality for any $q\ge1$

$$\mathcal{R}_s[\hat{f}_h;f]\ge\mathbb{E}_f\|\hat{f}_h-f\|_s=\mathbb{E}_f\|B_h(f)+\xi_h\|_s. \tag{35}$$

Denote by $\mathbb{B}_p(1),1\le p\le\infty$, the unit ball in $\mathbb{L}_p\big(\mathbb{R}^d\big)$. By the duality argument

$$\mathbb{E}_f\|B_h(f)+\xi_h\|_s=\mathbb{E}_f\sup_{\ell\in\mathbb{B}_r(1)}\int\ell(t)\big[B_h(f,t)+\xi_h(t)\big]\mathrm{d}t,\quad r=\frac{s}{s-1}.$$

Let $\ell_0\in\mathbb{B}_r(1)$ be such that $||B_h(f)||_s=\int\ell_0(t)B_h(f,t)\mathrm{d}t$; then

$$\mathbb{E}_f\|B_h(f)+\xi_h\|_s\ge\mathbb{E}_f\int\ell_0(t)\big[B_h(f,t)+\xi_h(t)\big]\mathrm{d}t=\|B_h(f)\|_s. \tag{36}$$

Here we have used that $\mathbb{E}_f\xi_h(t)=0,\ \forall t\in\mathbb{R}^d$. We also have by the triangle inequality

$$\mathbb{E}_f\|B_h(f)+\xi_h\|_s\ge\mathbb{E}_f\|\xi_h\|_s-\|B_h(f)\|_s. \tag{37}$$

Summing up the inequalities in (36) and (37) we get

$$\mathbb{E}_f\|B_h(f)+\xi_h\|_s\ge2^{-1}\mathbb{E}_f\|\xi_h\|_s. \tag{38}$$

Thus, in view of (36) and (38) for any $\alpha\in(0,1)$

$$\mathbb{E}_f\|B_h(f)+\xi_h\|_s\ge(1-\alpha)\|B_h(f)\|_s+2^{-1}\alpha\mathbb{E}_f\|\xi_h\|_s. \tag{39}$$

Choosing $\alpha = 2/3$ we arrive to (34) in view of (35).

In view (34), the assertion of the theorem will follow from the statement of Theorem 2 if we will show that

$$\mathbb{E}_f \|\xi_h\|_s \geq c_0 (nV_h)^{-1/2}.$$

2^0. Let $b > 0$ be a constant to be specified, and put $a = b^{-1}\sqrt{nV_h}$. By duality

$$\mathbb{E}_f \|\xi_h\|_s = \mathbb{E}_f \sup_{\ell \in \mathbb{B}_r(1)} \int \ell(t)\xi_h(t)\mathrm{d}t, \quad r = \frac{s}{s-1}. \tag{40}$$

Define the random event $\mathcal{A} = \{a\xi_h \in \mathbb{B}_2(1)\}$, and note that if $\mathcal{A}$ occurs then by the Hölder inequality

$$ag\xi_h \in \mathbb{B}_r(1), \quad \forall g \in \mathbb{B}_{\frac{2r}{2-r}}(1). \tag{41}$$

Remind that $s \geq 2$ implies $r \in [1,2]$, and if $r = s = 2$ the we formally put $\frac{2r}{2-r} = \infty$.

If the event $\mathcal{A}$ occurs then $\mathbb{B}_r(1) \supseteq \{ag\xi_h : g \in \mathbb{B}_{\frac{2r}{2-r}}(1)\}$. Therefore, by (40) and (41)

$$\begin{aligned}
\mathbb{E}_f \|\xi_h\|_s &\geq a\mathbb{E}_f \left[\mathbb{I}(\mathcal{A}) \sup_{g \in \mathbb{B}_{\frac{2r}{2-r}}(1)} \int g(t)\xi_h^2(t)\mathrm{d}t \right] \geq a \sup_{g \in \mathbb{B}_{\frac{2r}{2-r}}(1)} \mathbb{E}_f \left[\mathbb{I}(\mathcal{A}) \int g(t)\xi_h^2(t)\mathrm{d}t \right] \\
&= a \sup_{g \in \mathbb{B}_{\frac{2r}{2-r}}(1)} \int g(t) \left[\mathbb{E}_f\, \mathbb{I}(\mathcal{A})\xi_h^2(t) \right] \mathrm{d}t \;=\; a \left\| \mathbb{E}_f \xi_h^2(\cdot)\, \mathbb{I}(\mathcal{A}) \right\|_{\frac{2s}{s+2}} \\
&\geq a \left(\left\| \mathbb{E}_f \xi_h^2(\cdot) \right\|_{\frac{2s}{s+2}} - \left\| \mathbb{E}_f \xi_h^2(\cdot)\, \mathbb{I}(\bar{\mathcal{A}}) \right\|_{\frac{2s}{s+2}} \right),
\end{aligned} \tag{42}$$

where $\bar{\mathcal{A}}$ is the event complementary to $\mathcal{A}$.

Now consider separately two cases: $s = 2$ and $s > 2$.

3^0. If $s = 2$ we get from (42)

$$\mathbb{E}_f \|\xi_h\|_2 \geq a \left[\int \mathbb{E}_f \xi_h^2(t)\mathrm{d}t - \mathbb{E}_f \left\{ \|\xi_h\|_2^2\, \mathbb{I}\left(\|\xi_h\|_2 \geq \frac{b}{\sqrt{nV_h}} \right) \right\} \right]. \tag{43}$$

Note that

$$\mathbb{E}_f \xi_h^2(t) = n^{-1} \int K_h^2(t-x)f(x)\mathrm{d}x - n^{-1} \left[\int K_h(t-x)f(x)\mathrm{d}x \right]^2 \tag{44}$$

and, therefore,

$$\int \mathbb{E}_f \xi_h^2(t)\mathrm{d}t = \frac{\|K\|_2^2}{nV_h} - n^{-1} \int \left[\int K_h(t-x)f(x)\mathrm{d}x \right]^2 \mathrm{d}t.$$

The application of Young's inequality yields

$$\int \left[\int K_h(t-x)f(x)\mathrm{d}x \right]^2 \mathrm{d}t \leq \|K_h\|_1^2\, \|f\|_2^2 \leq \|K\|_1^2\, \mathrm{f}_\infty. \tag{45}$$

Here we have used that $f \in \mathbb{F}$. Thus, we obtain, in view of $V_h \leq V_{\max} \leq 1/8$ [see assumption of the part (ii) of Theorem 1]

$$\int \mathbb{E}_f \xi_h^2(t)\mathrm{d}t \;\; \geq \;\; \frac{\|K\|_2^2}{nV_h} - \frac{\|K\|_1^2\, \mathrm{f}_\infty}{n} \geq c_1 (nV_h)^{-1}. \tag{46}$$

It follows from Theorem 1 of GL (2010) that for any $x \geq 2$

$$\mathbb{P}\left\{\|\xi_h\|_2 \geq \frac{x\|K\|_2}{\sqrt{nV_h}}\right\} \leq e^{c_2(1-x)} \tag{47}$$

and, therefore, putting $b = y\|K\|_2,\ y \geq 2$, we obtain

$$\mathbb{E}_f\left\{\|\xi_h\|_2^2\, \mathbb{I}\left(\|\xi_h\|_2 \geq \frac{y\|K\|_2}{\sqrt{nV_h}}\right)\right\} \leq 2\|K\|_2^2(nV_h)^{-1}\int_y^\infty xe^{c_2(1-x)}\mathrm{d}x. \tag{48}$$

Choosing y sufficiently large in order to make latter integral less than $c_1/(4\|K\|_2^2)$ we obtain from (43), (46) and (48)

$$\mathbb{E}_f\|\xi_h\|_2 \geq c_3(nV_h)^{-1/2}.$$

The theorem is proved in the case $s = 2$.

4^0. Return now to the case $s > 2$. Note first that

$$\left\|\mathbb{E}_f\xi_h^2(\cdot)\right\|_{\frac{2s}{s+2}} \geq \left(\int_B \left|\mathbb{E}_f\xi_h^2(t)\right|^{\frac{2s}{s+2}}\mathrm{d}t\right)^{\frac{s+2}{2s}} \geq \nu^{\frac{2-s}{2s}}\int_B \mathbb{E}_f\xi_h^2(t)\mathrm{d}t. \tag{49}$$

The last relation is obtained by Hölder inequality. Taking into account that $\int_B f(t)\mathrm{d}t \geq \mu$, we get, using (44) and (45),

$$\int_B \mathbb{E}_f\xi_h^2(t)\mathrm{d}t \geq \frac{\mu\|K\|_2^2}{nV_h} - \frac{\|K\|_1^2\,\mathrm{f}_\infty}{n} \geq c_4\mu(nV_h)^{-1}. \tag{50}$$

Here we have used that $V_h \leq 2^{-1}\mu\|K\|_2^2/\|K\|_1^2$. On the other hand

$$\mathbb{E}_f\xi_h^2(\cdot)\,\mathbb{I}(\bar{\mathcal{A}}) \leq \left\{\mathbb{E}_f\left[\xi_h(\cdot)\right]^{\frac{4s}{s+2}}\right\}^{\frac{s+2}{2s}}\left\{\mathbb{P}(\bar{\mathcal{A}})\right\}^{\frac{s-2}{2s}}$$

and, therefore,

$$\left\|\mathbb{E}_f\xi_h^2(\cdot)\,\mathbb{I}(\bar{\mathcal{A}})\right\|_{\frac{2s}{s+2}} \leq \left\{\mathbb{E}_f\left(\|\xi_h\|_{\frac{4s}{s+2}}\right)^{\frac{4s}{s+2}}\right\}^{\frac{s+2}{2s}}\left\{\mathbb{P}(\bar{\mathcal{A}})\right\}^{\frac{s-2}{2s}}. \tag{51}$$

We derive from Theorem 1 in GL (2010) that there exists c_5 such that

$$\mathbb{E}_f\left(\|\xi_h\|_{\frac{4s}{s+2}}\right)^{\frac{4s}{s+2}} \leq c_5(nV_h)^{-\frac{2s}{s+2}}. \tag{52}$$

Putting $b = x||K||_2,\ x \geq 2$, we have in view of (47)

$$\left\{\mathbb{P}(\bar{\mathcal{A}})\right\}^{\frac{s-2}{2s}} \leq e^{\frac{c_2(1-x)(s-2)}{2s}}.$$

It leads together with (51) and (52) to the following estimate.

$$\left\|\mathbb{E}_f\xi_h^2(\cdot)\,\mathbb{I}(\bar{\mathcal{A}})\right\|_{\frac{2s}{s+2}} \leq c_6(nV_h)^{-1}e^{\frac{c_2(1-x)(s-2)}{2s}}. \tag{53}$$

We obtain finally from (42), (49), (50) and (53)

$$\mathbb{E}_f\|\xi_h\|_s \geq (x\|K\|_2)^{-1}\big(nV_h\big)^{-1/2}\left(c_4\mu\,\nu^{\frac{2-s}{2s}} - c_6e^{\frac{c_2(1-x)(s-2)}{2s}}\right).$$

It remains to choose x sufficiently large and we come to the assertion of the theorem in the case $s > 2$. ∎

4.4 Proof of Theorem 4

Let $f \in N_{s,d}(\alpha, L)$. It easily checked [see, e.g., Proposition 3 in Kerkyacharian, Lepski and Picard (2001)] that bias of the estimator $\hat{f}_h$ is bounded as follows

$$\|B_h(f,\cdot)\|_s \le C_1(u,d,s) L \sum_{j=1}^{d} h_j^{\alpha_j}.$$

Moreover, $\big\{\mathbb{E}_f \|\xi_h\|_s^q\big\}^{1/q} \le C_2(nV_h)^{-\gamma_s}$. If we set the "oracle bandwidth" $h^* := (h_1^*, \dots, h_d^*)$ so that

$$[h_j^*]^{\alpha_j} := \left[\frac{C_2}{C_1}\right]^{\bar{\alpha}/(\gamma_s+\bar{\alpha})} L^{-\bar{\alpha}/(\gamma_s+\bar{\alpha})} n^{-\gamma_s\bar{\alpha}/(\gamma_s+\bar{\alpha})}, \quad j = 1, \dots, d$$

then $h^* \in \mathcal{H}$ and $\hat{f}_{h^*} \in \mathcal{F}(\mathcal{H})$ for large enough n. Hence, for any $f \in N_{s,d}(\alpha, L)$ we have that $\mathcal{R}_s[\hat{f}_{h^*}; f] \le C_3 \varphi_{n,s}(\bar{\alpha})$. Then we apply oracle inequalities of Theorems 1 and 2. Observe that by choice of constant $\varkappa_2$ in definition of $h^{\max}$ we guarantee that the remainder terms are negligibly small in terms of dependence on n as compared with the first terms in (10) and (11). This fact leads to the statement of the theorem. ∎

5 Appendix

Proofs of Lemmas 1 and 2 follow directly from general uniform bounds on norms of empirical processes established in GL (2010). In our proofs below we use notation and terminology of the aforecited paper.

Proof of Lemma 1. The statement is a direct consequence of Theorem 4 of Section 3.3 in GL (2010).

To apply this theorem one should verify Assumptions (W1), (W4), and (L) for the following classes of weights $\mathcal{W}^{(1)} = \{w = n^{-1}K_h : h \in \mathcal{H}\}$ and $\mathcal{W}^{(2)} = \{w = n^{-1}(K_h * K_\eta) : (h,\eta) \in \mathcal{H}\times\mathcal{H}\}$. The sets $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$ are considered as images of $\mathcal{H}$ and $\mathcal{H}\times\mathcal{H}$ under transformations $h \mapsto n^{-1}K_h$ and $(h,\eta) \mapsto n^{-1}(K_h * K_\eta)$ respectively. The sets $\mathcal{H}$ and $\mathcal{H}\times\mathcal{H}$ are equipped with the distances

$$\mathrm{d}_1(h,h') = c_1 \max_{i=1,\dots,d} \ln\left(\frac{h_i \vee h_i'}{h_i \wedge h_i'}\right), \quad \mathrm{d}_2[(h,h'),(\eta,\eta')] := c_2\{\mathrm{d}_1(h,h') \vee \mathrm{d}_1(\eta,\eta')\},$$

where c_1 and c_2 are appropriate constants depending on k_∞, L_K and d only [see formulae (9.1)-(9.2) in GL (2010)]. With this notation Lemma 9 of GL (2010) shows that Assumption (L) holds for both $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$. Moreover, Assumption (W1) holds trivially both for $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$ with $\mu_* = V_{\max}$ and $\mu_* = 2^d V_{\max}$ repsectively. Moreover, Assumption (W4) for both $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$ follows from formula (9.8) in GL (2010). Thus all conditions of Theorem 4 are fulfilled.

(i). We apply this theorem with $z = 1$ and $\epsilon = 1$. We need to evaluate the constant $T_{3,\epsilon}$ for $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$. If $N_{\mathcal{H},\mathrm{d}_1}(\epsilon)$ denotes the minimal number of balls in the metric d_1 needed to cover $\mathcal{H}$, then formula (9.8) from GL (2010) shows that $N_{\mathcal{H},\mathrm{d}_1}(1/8) \le c_3 A_{\mathcal{H}}$, where c_3 depends on d only. Similarly, $N_{\mathcal{H}\times\mathcal{H},\mathrm{d}_2}(1/8) \le c_4 A_{\mathcal{H}}^2$. In addition, for

$$L_{\mathcal{H},\mathrm{d}_1}(\epsilon) := \sum_{k=1}^{\infty} \exp\left\{2\ln N_{\mathcal{H},\mathrm{d}_1}(\epsilon 2^{-k}) - (9/16)2^k k^{-2}\right\}$$

we have $L_{\mathcal{H},\mathrm{d}_1}(1) \le c_5 A_{\mathcal{H}}$. Similarly, $L_{\mathcal{H}\times\mathcal{H},\mathrm{d}_2}(1) \le c_6 A^2_{\mathcal{H}}$. Combining these bounds we come to the statement (i).

(ii). The second statement follows exactly in the same way from the above considerations. Theorem 4 of GL (2010) is again applied with $z = 1$ and $\epsilon = 1$. ∎

Proof of Lemma 2. The proof is by application of Theorem 7 from GL (2010). We need to calculate several quantities.

We start with the class $\mathcal{W}^{(1)}$. Here for $\vartheta_0^{(1)} = 10 c_s \mathrm{f}_\infty (L_K\sqrt{d})^{d/2}$ we have

$$\begin{aligned} C^*_{\xi,1}(y) &= 1 + 2\vartheta_0^{(1)}\big\{\sqrt{y}\big(V_{\max}^{1/s} + n^{-1/(2s)}\big) + y n^{-1/s}\big\} \\ &\le 1 + 2\vartheta_0^{(1)}\big\{2\sqrt{y}V_{\max}^{1/s} + yV_{\max}^{2/s}\big\}, \end{aligned}$$

where we have used that $V_{\max} \ge 1/\sqrt{n}$. If we set $y = \bar{y} := [4V_{\max}^{2/s}(\vartheta_0^{(1)} \vee 1)]^{-1}$ then $C^*_{\xi,1}(\bar{y}) \le 4$. We apply Theorem 7 with $\epsilon = 1$ and $y = \bar{y}$. Condition $nV_{\min} > C_1 = [256 c_s^2]^{(s\wedge 4)/(s\wedge 4-2)}$ implies that

$$\bar{u}_1(\gamma) = 4\big[1 - 8c_s (nV_{\min})^{1/(s\wedge 4)-1/2}\big]^{-1} \le 8.$$

Moreover, we note that condition $\bar{y} \le y_*^{(1)}$ follows from definition of $\bar{y}$ and $n \ge C_2$. In addition, $\tilde{T}^{(1)}_{1,\epsilon} \le cA^2_{\mathcal{H}} B_{\mathcal{H}}$. These facts imply (18) and (20).

The bounds (19) and (21) for $\mathcal{W}^{(2)}$ follow from similar computations. ∎

References

Birgé, L. (2008). Model selection for density estimation with $\mathbb{L}_2$–loss. `arXiv:0808.1416v2, http://arxiv.org`

Bretagnolle, J. and Huber, C. (1979). Estimation des densités: risque minimax. *Z. Wahrsch. Verw. Gebiete* **47**, 119–137.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation. The* $\mathbb{L}_1$ *View.* Wiley & Sons, New York.

Devroye, L. and Lugosi, G. (1996). A universally acceptable smoothing factor for kernel density estimation. *Ann. Statist.* **24**, 2499–2512.

Devroye, L. and Lugosi, G. (1997). Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *Ann. Statist.* **25**, 2626–2637.

Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation.* Springer, New York.

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

Goldenshluger A. and Lepski, O. (2008). Universal pointwise selection rule in multivariable function estimation. *Bernoulli* **14**, 1150-1190.

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

Goldenshluger, A. and Lepski, O. (2010). Uniform bounds for norms of sums of independent random functions. `arXiv:0904.1950v2`, `http://arxiv.org`. To appear in *Ann. Probab.*

Hasminskii, R. and Ibragimov, I. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory. *Ann. Statist.* **18**, 999–1010.

Ibragimov, I. A. and Khasminski, R. Z. (1980). An estimate of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **98**, 61–85 (in Russian).

Ibragimov, I. A. and Khasminski, R. Z. (1981). More on estimation of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **108**, 72–88 (in Russian).

Jennrich, R. (1969). Asymptotic properties of non–linear least squares estimators. *Ann. Statist.* **40**, 633–643.

Johnson, W. B., Schechtman, G. and Zinn, J. (1985). Best constants in moment inequalities for linear combinations of independent and exchangeable random variables. *Ann. Probab.* **13**, 234–253.

Juditsky, A. and Lambert–Lacroix, S. (2004). On minimax density estimation on $\mathbb{R}$. *Bernoulli* **10**, 187-220.

Kerkyacharian, G., Lepski, O. and Picard, D. (2001). Nonlinear estimation in anisotropic multi–index denoising. *Probab. Theory Related Fields* **121**, 137–170.

Kerkyacharian, G., Picard, D. and Tribouley, K. (1996). L^p adaptive density estimation. *Bernoulli* **2**, 229–247.

Massart, P. (2007). *Concentration Inequalities and Model Selection.* Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. Lecture Notes in Mathematics, 1896. Springer, Berlin.

Mason, D. M. (2009). Risk bounds for kernel density estimators. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **363**, 66–104. Available at `http://www.pdmi.ras.ru/znsl/`

Nikol'skii, S. M. (1969). *Priblizhenie Funktsii Mnogikh Peremennykh i Teoremy Vlozheniya.* (in Russian). [Approximation of functions of several variables and imbedding theorems.] Nauka, Moscow, 1969.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Ann. Math. Statist.* **33**, 1065-1076.

Rigollet, Ph. and Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16**, 260–280.

Rosenblatt, M (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832-837.

Samarov, A. and Tsybakov, A. (2007). Aggregation of density estimators and dimension reduction. *Advances in Statistical Modeling and Inference*, 233–251, Ser. Biostat., 3, World Sci. Publ., Hackensack, NJ.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London.