Sharp Adaptive Estimation of Linear Functionals

Jussi Klemelä,
Universität Heidelberg
and
Alexandre B. Tsybakov
Université Paris VI

July 21, 2001

Abstract

We consider estimation of a linear functional $T(f)$ where f is an unknown function observed in Gaussian white noise. We find asymptotically sharp adaptive estimators on various scales of smoothness classes in multidimensional situation. The results allow to evaluate explicitly the effect of dimension and to treat general scales of classes. Furthermore, we establish a connection between sharp adaptation and optimal recovery. Namely, we propose a scheme that reduces the construction of sharp adaptive estimators on a scale of functional classes to a solution of the corresponding optimization problem.

Mathematics Subject Classifications: 62G05, 62G20
Key Words: Adaptive curve estimation, Bandwidth selection, Exact constants in nonparametric smoothing, Gaussian white noise, Kernel estimation, Minimax risk.
Short title: Estimation of Linear Functionals

1 Introduction

Adaptation is now commonly considered as a crucial element of curve estimation procedures. The literature on adaptive estimation suggests various methods, starting from classical cross-validation or C_p criterion and ending with more recent techniques such as wavelet shrinkage or the method of Lepski. A more complete review of the existing approaches and further references can be found in Donoho at al. (1995), Jones, Marron and Sheather (1996), Lepski and Spokoiny (1997), Härdle et al. (1998), Tsybakov (1998), Barron, Birgé and Massart (1999), Nemirovski (2000).

How to choose a method of adaptation which is optimal in a certain sense? Comparing the rates of convergence does not suffice for this purpose. In fact, it is proved in the literature that most of the adaptive estimates attain optimal rates (exactly or up to a logarithmic factor), and thus the rate criterion does not allow to distinguish between them. This suggests to study exact asymptotics of the estimation error.

Let $f : \mathbf{R}^d \to \mathbf{R}$ be the unknown function to be estimated. Intuitively, the aim of adaptation would be to select the estimator which has the smallest risk among all estimators for every f. Unfortunately, this is not possible. We have either to restrict the class of estimators, considering, for example, kernel, spline or orthogonal series estimators, and to mimic the best estimator in this class for fixed f, or to restrict the class of functions f (usually, assuming that f has some smoothness which is unknown) and to adapt among all estimators, but in a minimax sense. Restricting the class of estimators disagrees with our initial wish to seek optimality among all estimators. To satisfy it, the approach starting from smoothness classes of f and using the minimax seems more relevant. Note that, for curve estimation problems, it is often not a big limitation to assume that f belongs to some class of functions $\mathcal{F}_\nu$, where ν is an *unknown* smoothness parameter.

There exist several results on exact asymptotics in minimax adaptation: Efroimovich and Pinsker (1984), Gobubev and Nussbaum (1990), Nemirovski (2000), Cavalier and Tsybakov (2000) (estimation of f in the L_2 -norm), Lepski (1992b), Tsybakov (1998) (estimation in sup-norm), Lepski and Spokoiny (1997), Tsybakov (1998), Lepski and Levit (1998) (estimation at a fixed point). These papers consider the one-dimensional case ($d = 1$). Recently some first results on multidimensional exact constants appeared: Lepski and Levit (1999) study the estimation of analytical functions in d dimensions and Efromovich (2000) extends the L_2-results of Efroimovich and Pinsker (1984) to multivariate case.

We call the collection $\mathcal{F} = \{\mathcal{F}_\nu\}_{\nu \in B}$, where B is a given set of indices ν, the *scale of classes*. A typical form of $\mathcal{F}_\nu$ is

$$\mathcal{F}_\nu = \mathcal{F}_{s,L} = \Big\{ f : \mathbf{R}^d \to \mathbf{R} \,\Big|\, \rho_s(f) \le L \Big\} \tag{1}$$

where $\nu = (s, L)$, $\rho_s(\cdot)$ is a given functional, usually a semi-norm (for example, the Hölder or Sobolev semi-norm), $s > 0$ is a smoothness parameter (for example, the number of derivatives) and $L > 0$ is the radius of the ball $\mathcal{F}_\nu$.

In this paper we consider the estimation of f at a fixed point of $\mathbf{R}^d$, or, more generally, the estimation of some linear functional $T(f)$ with values in $\mathbf{R}$. Ibragimov and Hasminskii (1981, 1984), Stone (1980), Sacks and Ylvisaker (1981), Donoho and Liu (1991), Donoho and Low (1992), Donoho (1994b) obtained optimal rates of convergence and linear minimax estimates in this problem for various examples of semi-norms $\rho_s(\cdot)$ and classes $\mathcal{F}_\nu$. In particular, as noticed by Donoho and Low (1992), the optimal rates can be expressed in terms of renormalization exponents related to the functionals ρ_s and T. It is shown in these papers that the optimal rate has the form ε^κ (where $\kappa = \kappa(\rho_s, T) > 0$ is an exponent depending on ρ_s and T) if the observations $Y_\varepsilon(t)$ follow the *Gaussian white noise model*:

$$dY_\varepsilon(t) = f(t)dt + \varepsilon dW(t), \quad t \in \mathbf{R}^d, \tag{2}$$

where W is the standard Brownian sheet in $\mathbf{R}^d$ and $0 < \varepsilon < 1$ is a small parameter. The use of the Gaussian white noise model has recently become standard in the literature on nonparametric estimation: it approximates asymptotically (in the sense of convergence of experiments) some common models with discrete observations, such as nonparametric regression or density estimation [Brown and Low (1996a), Nussbaum (1996)]. In such an approximation $\varepsilon \sim 1/\sqrt{n}$ where n is the number of discrete observations. Up to our knowledge, the equivalence results are now available only in dimension $d = 1$. Also, the equivalence is valid only for smoothness s large enough. Nevertheless, this does not restrict extension of our results to other nonparametric models, since this can be done directly, without equivalence considerations. For example, results for density estimation similar to ours and obtained by a direct method are actually available [see Butucea (2000)].

For an estimator T_ε based on the observation $Y_\varepsilon(t)$ consider the maximal risk

$$\mathcal{R}_{\varepsilon,\nu}(T_\varepsilon) = \sup_{f\in\mathcal{F}_\nu} E_f\left(|T_\varepsilon - T(f)|^p\right)$$

where $p > 0$ and E_f denotes the expectation w.r.t. the distribution of the observations when the underlying function is f. For the adaptive setup, ν is unknown, and the minimax approach consists in looking for estimators T_ε^* such that the supremum of the normalized risk $\sup_{\nu\in B}\varphi_{\varepsilon,\nu}^{-p}\mathcal{R}_{\varepsilon,\nu}(T_\varepsilon^*)$ is as small as possible, where $\varphi_{\varepsilon,\nu}$ is the rate of convergence. As shown by Lepski (1990, 1992a), Efromovich and Low (1994), Brown and Low (1996b), the last expression does not tend to 0 asymptotically as $\varepsilon \to 0$ if $\varphi_{\varepsilon,\nu}$ equals the optimal rate ε^κ. The correct rate for adaptation deteriorates to $\varphi_{\varepsilon,\nu} = (\varepsilon\sqrt{\log(1/\varepsilon)})^\kappa$, i.e. the best we can guarantee in terms of the rate is

$$\limsup_{\varepsilon\to 0}\sup_{\nu\in B}\varphi_{\varepsilon,\nu}^{-p}\mathcal{R}_{\varepsilon,\nu}(T_\varepsilon^*) < \infty,$$

except for the upper boundary of B where the normalization ε^κ can be maintained [see Lepski (1992a), Tsybakov (1998)]. The results of Lepski (1990, 1992a), Brown and Low (1996b) are proved for the case where $\mathcal{F}_\nu$ are Hölder classes of functions, $T(f) = f(0)$ and $d = 1$. Efromovich and Low (1994) considered more general linear functionals and Tsybakov (1998) proved the result for the Sobolev classes and $d = 1$. Following the scheme of Lepski (1992a) or of Tsybakov (1998), it is not difficult to show that in the general situation with $d \geq 1$ and Hölder or Sobolev classes of functions the correct asymptotic normalization $\varphi_{\varepsilon,\nu}$ in the risk $\sup_{\nu\in B}\varphi_{\varepsilon,\nu}^{-p}\mathcal{R}_{\varepsilon,\nu}(T_\varepsilon^*)$ is of the form $(\varepsilon\sqrt{\log(1/\varepsilon)})^\kappa$, up to a boundary effect.

Here we do not go into details of these results about the rates. For a more comprehensive discussion, the definition of adaptive rate and the lower bounds see Tsybakov (1998). Below we assume as given the normalization $\varphi_{\varepsilon,\nu} = (\varepsilon\sqrt{\log(1/\varepsilon)})^\kappa$ for the risk, where κ is the exponent of the optimal rate that is calculated as in Ibragimov and Hasminskii (1984), Donoho and Liu (1991). Our aim is to find the exact asymptotical constant c_ν in the expression for the minimax "adaptive" risk and to construct the adaptive estimator T_ε^* that attains this constant.

Such an estimator T_ε^* will be called sharp adaptive on the scale of classes $\{\mathcal{F}_\nu\}$. The examples of sharp adaptive estimators of functionals are known for the case where $d = 1$, $T(f) = f(0)$. The first example has been given by Lepski and Spokoiny (1997) who considered the Hölder scale of classes with smoothness $0 < s \leq 2$. Tsybakov (1998) obtained sharp adaptive estimators for the Sobolev scale of classes where s takes discrete values without upper restriction on s and with fixed L. His setup is somewhat different from the one considered here and his results cannot be formally deduced from ours. In the present paper we assume that s belongs to a bounded interval: this allows, in particular, a unified treatment of the Hölder and Sobolev cases (in the Hölder case, if s is large, we cannot guarantee the necessary assumptions and the optimal solutions for the kernels are not explicitly known). Also, the Gaussian model here differs from those in Lepski and Spokoiny (1997) and Tsybakov (1998): we consider the observations on $\mathbf{R}^d$, while in those papers the observations are on $[0, 1]$. From the mathematical point of view, the difference is not significant between considering functions on $\mathbf{R}$ and periodic functions on $[0, 1]$ (as in Tsybakov (1998)) or neglecting the boundary effects on $[0, 1]$ (as in Lepski and Spokoiny (1997)). However, working with the infinite interval of observations leads to more transparent notation. In practice we always have finite intervals, but if they are large enough they can be approximately considered as infinite. This is commonly done in the literature on signal processing (cf. a discussion in Donoho and Low (1992)). Lepski and Levit (1998, 1999) considered the Gaussian white noise model on the infinite interval and obtained sharp adaptation results for the case where $\mathcal{F}_\nu$ are classes of analytical or supersmooth functions.

Here we find sharp adaptive estimators of linear functionals for the general problem of dimension $d \geq 1$, classes (1) with some general functional $\rho_s(\cdot)$ and both s and L unknown, and a functional T satisfying some assumptions that are relevant for the non-regular case where the "$\sqrt{n}$-consistent" estimation is not possible . The main examples are $T(f) = f(0)$ or $T(f)$ being a partial derivative of f at a point.

We consider a general framework that makes transparent the connection between sharp adaptation and optimal estimation of linear functionals (optimal recovery). An explicit scheme is proposed that reduces the construction of sharp adaptive estimators to a solution of the corresponding optimal recovery (OR) problem. Donoho (1994a,b) was the first to point out a connection between OR and nonparametric statistics. He showed that the OR argument can be used to get exact asymptotics of linear minimax risks in estimation problems. More recently Lepski and Tsybakov

(2000) proved that by means of OR one can construct asymptotically sharp minimax nonparametric tests. The present paper describes one more field of application of OR: construction of sharp adaptive estimators. Our conclusion can be formulated as follows: it is possible to construct sharp adaptive estimators of linear functionals by action of the Lepski-type selection procedure (with properly chosen thresholds) over families of linear estimates with optimal recovery kernels.

2 Assumptions and preliminaries

Let $s > 0$ and let ρ_s be a functional defined on a subset $\mathcal{D}$ of the space of all functions $f : \mathbf{R}^d \to \mathbf{R}$. Suppose in the sequel that $\mathcal{D} = \{f : \ \rho_s(f) < \infty\}$. We assume the following conditions on the functional ρ_s.

Assumption 1.

(i) *The functional ρ_s is convex, nonnegative and symmetric, i.e. $\rho_s(f) = \rho_s(-f)$, and $\rho_s(f) \not\equiv 0$,*

(ii) *$\rho_s(af(b\cdot)) = ab^s\rho_s(f(\cdot))$ for any $a \geq 0$, $b > 0$, $f \in \mathcal{D}$.*

Furthermore, we assume that the functional T satisfies the following conditions.
Assumption 2.

(i) *T is a linear functional on $\mathcal{D}$.*

(ii) *There exists $r \geq 0$ such that $T(af(b\,\cdot)) = ab^rT(f(\cdot))$ for $a \geq 0$, $b > 0$, $f \in \mathcal{D}$.*

(iii) *The modulus of continuity is well-defined:*

$$\omega_{s,L}(\varepsilon) \stackrel{\text{def}}{=} \sup\{T(f) : \|f\|_2 \leq \varepsilon,\ \rho_s(f) \leq L\} < \infty$$

for all $s > r$, $L > 0$, $\varepsilon > 0$ where $\|\cdot\|_2$ is the $L_2(\mathbf{R}^d)$-norm.

Assumptions 1(ii) and 2(ii) are usual renormalization assumptions (see Donoho and Low (1992) for discussion and examples).

As described by Donoho (1994a,b), Donoho and Liu (1991), Donoho and Low (1992), the minimax estimation of functionals from random noisy data is closely related to the deterministic problem of minimax optimal recovery that considers estimation from observations in non-random noise. These papers show that, by calibrating the algorithms of optimal recovery one can construct linear minimax estimators for the statistical estimation of linear functionals and asymptotic minimax estimators for the statistical estimation of functions with supremum loss. We refer to these papers for a detailed discussion. Here we show that by calibrating the algorithms of optimal recovery, one can construct a family of linear estimators such that choosing one of these linear estimators with a certain data-based decision rule will result in a sharp adaptive estimation procedure. Next we give a brief summary of the results on optimal recovery that will be used below.

By the generalized Weierstrass theorem, under the Assumptions 1 and 2 there exists a function $g_{s,L,\varepsilon}$ which attains the supremum of the modulus of continuity, i.e.

$$T(g_{s,L,\varepsilon}) = \omega_{s,L}(\varepsilon) \tag{3}$$

[cf. Gabushin (1970), Micchelli and Rivlin (1977), Arestov (1989)]. These authors show that the extremal problem

$$\max T(f) \text{ subject to } \begin{cases} \|f\|_2 \le 1, \\ \rho_s(f) \le 1, \end{cases} \tag{4}$$

is related to the optimal recovery problem: find a function K_s such that

$$\sup_{\rho_s(f)\le 1,\ \|f-g\|_2\le 1} \left| \int K_s g - T(f) \right| = \inf_K \sup_{\rho_s(f)\le 1,\ \|f-g\|_2\le 1} \left| \int K g - T(f) \right| \stackrel{\text{def}}{=} E(s). \tag{5}$$

In particular, Theorems 6, 8 and 11 in Micchelli and Rivlin (1977) and Theorems 2.4 - 2.5 in Arestov (1989) show that under Assumptions 1 and 2 there exists $K_s \in L_2(\mathbf{R}^d)$ that satisfies (5) and, moreover,

$$E(s) = T(g_{s,1,1}) = \sup_{\rho_s(f)\le 1} \left| \int K_s f - T(f) \right| + \|K_s\|_2. \tag{6}$$

The property (6) plays crucial role in our argument. In the sequel K_s denotes the optimal recovery kernel, i.e. the function in $L_2(\mathbf{R}^d)$ satisfying (6). Note that if

$T(f) = f(0)$ and $0 < \int g_{s,1,1} < \infty$, the kernel K_s has a particular form $K_s = K_s^0$ where

$$K_s^0 = g_{s,1,1} / \int g_{s,1,1}.$$

This can be shown in a simple way (see e.g. Lemma 1 of Lepski and Tsybakov (2000)). For general functionals T similar condition usually holds:

$$K_s = C g_{s,1,1} \tag{7}$$

where the constant $C > 0$ depends only on s, r, d. In fact, Assumptions 1 - 2 and the renormalization argument entail that $\omega_{s,L}(\varepsilon)$ is of power law form: $\omega_{s,L}(\varepsilon) = \omega_{s,1}(1) L^{2(r+d)/(2s+d)} \varepsilon^{\kappa}$ where

$$\kappa = \kappa(s) = 2(s-r)/(2s+d).$$

Hence, as in Donoho (1994b), Donoho and Liu [1991, Section 4.3] and Donoho and Low [1992, Section 8], one gets (7).

We assume that the observations $Y_\varepsilon(t),\ t \in \mathbf{R}^d$, are obtained from the Gaussian white noise model (2). As follows from Donoho and Liu (1991), Donoho and Low (1992), Donoho (1994b), the linear minimax estimator of $T(f)$ under the mean squared risk on the class of functions $\mathcal{F}_{s,L}$ is the kernel estimator with properly rescaled optimal recovery kernel K_s and the bandwidth

$$h_l(s, L, \varepsilon) = (\varepsilon/L)^{2/(2s+d)}. \tag{8}$$

The rate of convergence of the linear minimax estimator is respectively ε^{κ}.

Our definition of sharp adaptive estimator starts from the family of kernel estimators with optimal kernels K_s, though with the bandwidths different from those of the linear minimax case.

We suppose that the following continuity condition on the kernel holds.

Assumption 3.

The optimal recovery kernel K_s satisfies (7) and $\|K_{s'} - K_s\|_2 \to 0$, as $s' \to s$, for any $s > r$.

Let the scale of classes $\{\mathcal{F}_{s,L}\}_{(s,L)\in B}$ be defined by (1) with

$$B = \{(s, L) : \ s_* \le s \le s^*,\ L_* \le L \le L^*\}$$

where $r < s_* < s^* < \infty$, $0 < L_* < L^* < \infty$. This means that we are certain that $f \in \mathcal{F}_{s,L}$ for some $L \in [L_*, L^*]$ and $s \in [s_*, s^*]$. The values r, s_*, s^* are supposed to be known but L_* and L^* can be unknown: we do not need L_*, L^* for the construction of our sharp adaptive estimators.

Define the bias constant

$$b_{s,s'} = \sup_{\rho_s(f) \le 1} \left| \int K_{s'} f - T(f) \right|.$$

The following boundedness and continuity condition on the bias constant will be assumed.

Assumption 4.

(i) *For any $r < r' < s^*$ there exists a positive constant $b_{max}(r', s^*)$ such that*

$$b_{s,s'} \le b_{max}(r', s^*), \quad \forall\ r' \le s' \le s \le s^*.$$

(ii)

$$\limsup_{\delta \to 0} \sup_{s,s' \in [s_*, s^*]: |s-s'| \le \delta} \frac{b_{s,s'}}{b_{s,s}} \le 1.$$

Clearly, the main interest of our construction is in the case where the solution $g_{s,L,\varepsilon}$ and the kernel K_s can be expressed explicitly. In this case Assumptions 3 – 4 can be checked directly, see the examples in Section 5.

3 Results

For any $h > 0$ denote $K_{s,h}(\cdot) = h^{-d-r} K_s(\cdot/h)$ where K_s is defined in Section 2. Consider kernel estimators of the form $\int K_{s,h}(t) dY_\varepsilon(t)$ where h is a suitably chosen bandwidth. Denote

$$h(s, \varepsilon) = \varepsilon^{2/(2s+d)}, \tag{9}$$

and introduce the "effective noise level under adaptation":

$$\tilde{\varepsilon} = \tilde{\varepsilon}(s) = \varepsilon d_\varepsilon(s) = \left(\lambda(s) \varepsilon^2 \log \varepsilon^{-1} \right)^{1/2},$$

where

$$d_\varepsilon(s) = \left(\lambda(s)\log\varepsilon^{-1}\right)^{1/2},$$

$$\lambda(s) = 2p(2r+d)\left(\frac{1}{2s+d} - \frac{1}{2s^*+d}\right).$$

We use the bandwidth computed at the effective noise level:

$$h(s,\tilde\varepsilon(s)) = \left(\lambda(s)\varepsilon^2\log\varepsilon^{-1}\right)^{1/(2s+d)}.$$

This bandwidth is by a logarithmic factor in order than the bandwidth (8) of linear minimax estimate.

We will introduce a sufficiently fine grid on $[s_*, s^*]$ and a statistic $\hat s$ having values on this grid. To each point of the grid we assign a linear kernel estimator. The statistic $\hat s$ will choose one of these estimators. Namely, we consider the grid

$$S = \{s_1, \ldots, s_m\},$$

where

$$r' < s_1 < \cdots < s_m < s^*$$

with a fixed r' satisfying $r < r' \le s_*$, and we assume that there exist $k_2 > k_1 > 0$ and $\gamma_1 \ge \gamma > 1$ such that

$$k_1(\log\varepsilon^{-1})^{-\gamma_1} \le s_{i+1} - s_i \le k_2(\log\varepsilon^{-1})^{-\gamma}, \quad i = 0,\ldots,m-1, \tag{10}$$

where $s_0 = r',\ s_m - s^* = o(1)$, as $\varepsilon \to 0$. Note that the same grid S can be used for different values s_*, provided $s_* > r'$. In this sense the exact knowledge of s_* is not required for the construction of the estimator.

For any $s \in S$ introduce the linear kernel estimator of the functional $T(f)$:

$$T_{s,\varepsilon} = \int K_{s,h(s,\tilde\varepsilon(s))}(t)dY_\varepsilon(t).$$

The sharp adaptive estimator has the form $T_{\hat s,\varepsilon}$ where $\hat s$ is a suitably chosen statistic. To define $\hat s$ we follow the approach used in different statistical models starting from the paper of Lepski (1990). That is, the statistic $\hat s$ is defined as the largest of those s-values in the grid for which the estimator $T_{s,\varepsilon}$ does not differ significantly from the estimators corresponding to the smaller s-values. We choose

$$\hat s = \max\{s \in S : |T_{s,\varepsilon} - T_{s',\varepsilon}| \le \eta(s') \text{ for all } s' \in S,\ s' \le s\}$$

with the threshold

$$\eta(s) = d_\varepsilon(s)\,\sigma_s = \tilde{\varepsilon}(s)^{2(s-r)/(2s+d)}\|K_s\|_2,$$

where σ_s is the standard deviation of $T_{s,\varepsilon}$,

$$\sigma_s = \varepsilon\|K_{s,h(s,\tilde{\varepsilon}(s))}\|_2 = \varepsilon h^{-r-d/2}(s,\tilde{\varepsilon}(s))\|K_s\|_2. \tag{11}$$

Finally, define the estimator of $T(f)$ as

$$T^*_\varepsilon = T_{\hat{s},\varepsilon}. \tag{12}$$

The next theorem is the main result of this paper. It states that the estimator T^*_ε is sharp adaptive and that the exact asymptotical constant c_ν for the minimax adaptive risk is given by the expression

$$c_\nu = c_{\nu,s^*} = L^{(2r+d)/(2s+d)}T(g_{s,1,1})\left[2p(2r+d)\left(\frac{1}{2s+d} - \frac{1}{2s^*+d}\right)\right]^{(s-r)/(2s+d)}. \tag{13}$$

To formulate the theorem we introduce, for any $\psi > 0,\ p > 0$, the normalized risk:

$$\mathcal{R}_{\varepsilon,\nu}(T_\varepsilon, \psi) = \sup_{f\in\mathcal{F}_\nu} E_f\left(\psi^{-p}\,|T_\varepsilon - T(f)|^p\right)$$

and denote

$$\psi_\nu = c_\nu\varphi_{\varepsilon,\nu} = c_\nu\left(\varepsilon^2\log\varepsilon^{-1}\right)^{(s-r)/(2s+d)}. \tag{14}$$

(Recall that we defined $\varphi_{\varepsilon,\nu} = (\varepsilon\sqrt{\log\varepsilon^{-1}})^\kappa$ and $\kappa = 2(s-r)/(2s+d)$.) The normalizing factor ψ_ν may be expressed as the value of the modulus of continuity at the effective noise level $\tilde{\varepsilon}$. Indeed, by standard renormalization argument, $g_{s,L,\tilde{\varepsilon}}(\cdot) = a g_{s,1,1}(b\,\cdot)$ where $a = Lb^{-s}$ and $b = (L/\tilde{\varepsilon})^{2/(2s+d)}$. Thus,

$$\begin{aligned}
\omega_{s,L}\left(\tilde{\varepsilon}\right) &= T\left(g_{s,L,\tilde{\varepsilon}}\right) = ab^r T\left(g_{s,1,1}\right) \\
&= \tilde{\varepsilon}^{2(s-r)/(2s+d)}L^{(2r+d)/(2s+d)}T(g_{s,1,1}) \\
&= c_\nu\left(\varepsilon^2\log\varepsilon^{-1}\right)^{(s-r)/(2s+d)} = \psi_\nu.
\end{aligned} \tag{15}$$

Theorem 1 *Let Assumptions 1-4 hold, let $p > 0$ and denote $B_q = [s_*, q]\times[L_*, L^*]$, where $s_* < q < s^*$. Then the estimator T^*_ε defined in (12) is sharp adaptive:*

$$\sup_{s_*<q<s^*}\ \limsup_{\varepsilon\to 0}\ \sup_{\nu\in B_q}\mathcal{R}_{\varepsilon,\nu}(T^*_\varepsilon, \psi_\nu) \le 1, \tag{16}$$

and

$$\sup_{s_*<q<s^*} \liminf_{\varepsilon\to 0} \inf_{T_\varepsilon} \sup_{\nu\in B_q} \mathcal{R}_{\varepsilon,\nu}(T_\varepsilon,\psi_\nu) \geq 1. \tag{17}$$

Here $\inf_{T_\varepsilon}$ *denotes the infimum over all estimators.*

Proofs of Theorem 1 and of the further results are given in Section 7.

Remark 1. Since s and L are not fixed, it is more precise to call $c_\nu = c_{\nu,s^*}$ the "optimal normalizing function" rather than the optimal constant. An insight on the structure of this function has been first given by Lepski (1992a), Theorem 8, where he shows that

$$c_{\nu,s^*} \asymp \left(\frac{1}{2s+1} - \frac{1}{2s^*+1}\right)^{(s-r)/(2s+1)}$$

(he considers the case $d = 1$ and the Hölder scale of classes, but it is not hard to extend his result to our multivariate setting). Here we specify the exact value of c_{ν,s^*} which contains, of course, the same factor, but also turns out to contain another factor expressed in terms of optimal recovery solutions (cf. (13)).

For $s = s^*$ we have $c_{\nu,s^*} = 0$. Thus $\mathcal{R}_{\varepsilon,\nu}(T^*_\varepsilon,\psi_\nu)$ is not defined for the single point $s = s^*$. This explains why the set B_q appears in place of B in Theorem 1: indeed, the difference between (16) – (17) and analogous expressions with $\lim_{\varepsilon\to 0}\sup_{\nu\in B}$ is really minor but we have to use the form (16) – (17) in order to exclude the point $s = s^*$. It is possible to construct an adaptive estimator that has the property given in Theorem 1 and attains for $s = s^*$ a faster rate, without the logarithmic factor: $\varepsilon^{2(s^*-r)/(2s^*+d)}$ (cf. Lepski (1992a), Theorem 8). Such an estimator is defined similarly to T^*_ε , but with the enlarged grid $\{s_1,\ldots,s_m,s^*\}$ and with $T_{s,\varepsilon}$ replaced by

$$\tilde{T}_{s,\varepsilon} = \begin{cases} \int K_{s,h(s,\tilde{\varepsilon}(s))}(t)dY_\varepsilon(t) \text{ for } s = s_1,\ldots,s_m, \\ \int K_{s,h(s,\varepsilon)}(t)dY_\varepsilon(t) \text{ for } s = s^*. \end{cases}$$

Also one should impose an assumption on the rate of approximation of s^* by s_m as $\varepsilon \to 0$. The effect of improving the rate at a single boundary point s^* is discussed for example by Lepski (1992a) and Tsybakov (1998). This issue is of minor importance, although involving more technical details, and we do not pursue it here.

The estimator T^*_ε depends on the value s^* which is not always available. One can propose a sub-optimal modification of T^*_ε that does not depend on s^*. It is obtained by putting formally $s^* = \infty$ in the definition for T_ε^*. In other words, we replace in all the formulas $\lambda(s)$ by

$$\tilde{\lambda}(s) = \frac{2p(2r+d)}{2s+d},$$

and we set

$$\tilde{c}_\nu = L^{(2r+d)/(2s+d)} T(g_{s,1,1}) \left(\frac{2p(r+d)}{2s+d} \right)^{(s-r)/(2s+d)},$$

$\tilde{\psi}_\nu = \tilde{c}_\nu \left(\varepsilon^2 \log \varepsilon^{-1}\right)^{(s-r)/(2s+d)}$. We also assume that the grid $S = \{s_1, \ldots, s_m\}$ is extended to the right beyond s^*:

$$r' < s_1 < \cdots < s_m = s_{\max}$$

where $s_{\max} > s^*$ and s_i satisfy (10). Here $s_{\max}$ is an arbitrary large fixed number.

Let $\tilde{T}_\varepsilon$ be the estimate defined as T_ε^* with the change of $\lambda(s)$ to $\tilde{\lambda}(s)$ and with the extended grid S as above. Note that $\tilde{T}_\varepsilon$ is completely data-driven: the dependence on s^* disappears.

Theorem 2 *Let Assumptions 1-4 hold and let $p > 0$. Then*

$$\limsup_{\varepsilon \to 0} \sup_{\nu \in B} \mathcal{R}_{\varepsilon,\nu}(\tilde{T}_\varepsilon, \tilde{\psi}_\nu) \le 1.$$

Proof of Theorem 2 is omitted: it follows the same lines as that of Theorem 1, with a minor modification concerning the extension of the grid beyond s^* (this is done as in the proof of Theorem 3 below). Comparing with that proof, we observe that Theorem 2 remains valid if one takes $s_{\max} \to \infty$ as $\varepsilon \to 0$, but not faster than $\log(1/\varepsilon)$.

Note that in view of Theorems 1 and 2 the asymptotical risk of the estimator $\tilde{T}_\varepsilon$ can be larger than that of T_ε^* on any set $\mathcal{F}_s$ (with $s_* \le s < s^*$) at most by a factor of

$$\frac{\tilde{c}_\nu}{c_\nu} = \left(\frac{2s^* + d}{2(s^* - s)} \right)^{(s-r)/(2s+d)}.$$

This factor is particularily close to 1 if s is fixed and s^* gets large. Thus, for s^* large enough and fixed s the behaviour of $\tilde{T}_\varepsilon$ and T_ε^* is similar, but $\tilde{T}_\varepsilon$ has an advantage since it does not depend on s^*. Although, for s close to s^* the estimator $\tilde{T}_\varepsilon$ is much less efficient than T_ε^*.

A useful modification of Theorem 1 consists in constructing a grid on the values of smoothing parameter h and not on the s-values as above. To get correspondence with the s-grid satisfying (10), the h-grid should have a geometrical character. This means that an "economic" choice of h (among a logarithmic number of possible candidates) is in fact sufficient to attain sharp asymptotic adaptivity: increasing the cardinality of the grid or passing to the choice of h in a continuum of values complicates the procedure but does not improve the result. We set $h_0 = h(r', \tilde{\varepsilon}(r'))$ and define the sequence $\{h_i\}$ by the recursion

$$h_{i+1} = h_i(1 + \alpha(h_i)) \tag{18}$$

where $\alpha(h_i)$ is a slowly varying function of h_i. It will be sufficient to consider

$$\alpha(h) = (\log(1/h))^{-\gamma_0} \tag{19}$$

where $\gamma_0 > 0$ is a constant. Given the grid

$$\mathcal{H} = \{h_1, h_2, \ldots, h_m\},$$

with $m = \max\{i : \ h_i < h_{\max}\},\ h_{\max} \geq \varepsilon^{2/(2s^*+d)}$, consider the bandwidth

$$\hat{h} = \max\{h \in \mathcal{H} : |T_\varepsilon(h) - T_\varepsilon(h')| \leq \eta_{h'}, \text{ for all } h' \leq h,\ h' \in \mathcal{H}\}$$

where $T_\varepsilon(h) = \int K_{s(h),h}(t) dY_\varepsilon(t)$, and

$$\eta_h = \varepsilon h^{-r-d/2} \sqrt{p(2r+d) \log \frac{h_{\max}}{h}}\ \|K_{s(h)}\|_2$$

with

$$s(h) = \left(\frac{\log \varepsilon}{\log h} - \frac{d}{2} \right).$$

Define

$$T_\varepsilon^* = T_\varepsilon(\hat{h}). \tag{20}$$

We now state an analogue of Theorems 1 and 2 for the estimator (20).

Theorem 3 *Let Assumptions 1-4 be satisfied, $p > 0$, and let T_ε^* be defined in (20).*

(i) If $h_{\max} = \varepsilon^{2/(2s^*+d)}$ then (16) holds, i.e. the estimator (20) is sharp adaptive.

(ii) If $h_{\max} = 1$ *then the estimator (20) satisfies*

$$\lim_{\varepsilon \to 0} \sup_{\nu \in B} \sup \mathcal{R}_{\varepsilon,\nu}(T_\varepsilon^*, \tilde{\psi}_\nu) \leq 1.$$

We observe that the result of Theorem 3 (ii) is robust to the choice of $h_{\max}$. Inspection of the proof shows that $h_{\max} = 1$ can be replaced by $h_{\max} = h^*$ for any positive h^*, and other choices of $h_{\max} > \varepsilon^{2/(2s^*+d)}$ are acceptable as well. We also note that a special case of the above construction is given by Lepski and Spokoiny (1997). They considered a grid on h-values defined in (18), (19) with $\gamma_0 = 1/4$ and for the particular case of Hölder scale of classes with $0 < s \leq 2, r = 0, d = 1$. Their Theorem 3.3 in the proper form follows from Theorem 3 modulo the fact that their procedure is slightly different: it includes an additional factor $(1 + \alpha(h))$ in the threshold η_h.

Remark 2. (*Pointwise and spatial adaptivity.*) Application of our adaptive procedure to the particular functionals $T(f) = f(x)$ for each point x of an interval where f is defined gives an estimator of f on this interval. Therefore, a special case of our results is sharp pointwise adaptivity for estimation of a whole function f on an interval. Arguing as in Lepski, Mammen, and Spokoiny (1997), Goldenshluger and Nemirovskii (1997) one can deduce spatial adaptivity of such an estimator of a function from its pointwise adaptivity.

Remark 3. (*Data-dependent kernels.*) One of the most studied topics in non-parametric curve estimation is the choice of the smoothing parameter in kernel estimation. We do not restrict ourselves to kernel estimators but find asymptotically sharp adaptive estimator among all estimators. Remark that nevertheless this estimator turns out to be a kernel one, with smoothing parameter selected in a data-dependent way. What is more, our results suggest that, to attain optimality, not only the smoothing parameter but also the kernel function of this estimator should be chosen in a data-dependent way.

4 Other statistical models

It is possible to modify the proposed estimator for other types of observations than the Gaussian white noise model. Consider some examples.

Density estimation. Let $X_1, \dots, X_n$ be i.i.d. observations with density f on $\mathbf{R}^d$ and consider estimation of $T(f) = f(x)$ for some $x \in \mathbf{R}^d$. Construct a preliminary estimator $\hat{f}_n(x) > 0$ for $f(x)$ and consider the family of kernel estimators

$$T_{s,n} = \frac{1}{n} \sum_{i=1}^{n} K_{s,h(s,\tilde{\varepsilon}(s))}(X_i - x),$$

where K_s, h and $\tilde{\varepsilon}(s)$ are defined as above where one substitutes $\varepsilon = (\hat{f}_n(x)/n)^{1/2}$. The adaptive procedure is defined as in Section 3: it is a data-driven selection of an appropriate member of this family. But again, in the definition of the threshold $\eta(s)$ of the adaptive procedure one should set $\varepsilon = (\hat{f}_n(x)/n)^{1/2}$. Sharp adaptation properties of this adaptive density estimator on the Sobolev scale of classes are proved by Butucea (2000). Butucea (2001) presents a large simulation study showing a successful behavior of the adaptive procedure for different densities f. In particular, the procedure is quite robust to the choice of the preliminary estimator $\hat{f}_n$.

Nonparametric regression. Consider the nonparametric regression model

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \dots, n,$$

where $X_i = i/n - 1/2$ are equispaced regressors on the interval $[-1/2, 1/2]$, f is an unknown regression function and ξ_i are i.i.d. random variables such that $E(\xi_i) = 0$, $E(\xi_i^2) = \sigma^2 > 0$, satisfying some additional moment conditions. Let again $T(f) = f(x)$. An adaptive procedure analogous to ours can be suggested similar to the density case. Construct a preliminary estimator $\hat{\sigma}$ of σ, and consider the family of linear kernel estimators

$$T_{s,n} = \frac{1}{n} \sum_{i=1}^{n} Y_i K_{s,h(s,\tilde{\varepsilon}(s))}(X_i - x),$$

where K_s, h and $\tilde{\varepsilon}$ are defined as in Section 3, with $\varepsilon = \hat{\sigma}/\sqrt{n}$. Finally, apply the thresholding procedures of Section 3 to get $\hat{s}$ (or $\hat{h}$), with the above definition of ε used to compute the threshold $\eta(s)$. The adaptive estimator is then $T_{\hat{s},n}$. We conjecture that $T_{\hat{s},n}$ has sharp optimality properties, as in Theorems 1 - 3.

5 Examples

In this section we give examples of classes $\mathcal{F}_\nu$ (respectively, functionals ρ_s) that satisfy the assumptions of Section 2 and allow explicit construction of sharp adaptive estimators.

1. Sobolev classes. Let $\beta > d/2$ and denote $s = \beta - d/2$. Define the Sobolev semi-norm ρ_s by

$$\rho_s^2(f) = (2\pi)^d \int_{\mathbf{R}^d} \|\omega\|^{2\beta} \left|\hat{f}(\omega)\right|^2 d\omega$$

where the Fourier transform of a function $f \in L_1(\mathbf{R}^d)$ is

$$\hat{f}(\omega) = \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} f(x) \exp(-ix^T \omega) dx$$

and $\|\cdot\|$ denotes the Euclidean norm in $\mathbf{R}^d$. Note that if β is an integer,

$$\rho_s^2(f) = \sum_{|\alpha|=\beta} \int_{\mathbf{R}^d} \left|f^{(\alpha)}\right|^2$$

where

$$f^{(\alpha)}(x) = i^{|\alpha|} \int_{\mathbf{R}^d} \omega^\alpha \hat{f}(\omega) \exp(ix^T \omega)\, d\omega$$

and $\alpha = (\alpha_1, \dots, \alpha_d)$ is a multi-index, $|\alpha| = \alpha_1 + \dots + \alpha_d$, $\omega^\alpha = \omega_1^{\alpha_1} \cdots \omega_d^{\alpha_d}$, for $\omega = (\omega_1, \dots, \omega_d)$.

Consider the estimation of the functional $T(f) = f^{(\alpha_0)}(0)$ where α_0 is a multi-index, $|\alpha_0| = r$ and $r \geq 0$ is an integer, $s > r$. The kernel K_s is then obtained as a renormalized version of the basic kernel

$$\tilde{K}_s(x) = (2\pi)^{-d} i^r \int_{\mathbf{R}^d} \omega^{\alpha_0} (1 + \|\omega\|^{2\beta})^{-1} \exp(ix^T \omega) d\omega. \tag{21}$$

Note that $\tilde{K}_s$ is always real-valued: it is the directional derivative corresponding to the multi-index α_0 of the function whose Fourier transform is $(2\pi)^{-d}(1 + \|\omega\|^{2\beta})^{-1}$. Introduce the constant

$$C_* = \left[\frac{1}{2\beta} B\left(1 + \frac{2r+d}{2\beta}, 1 - \frac{2r+d}{2\beta}\right) (2\pi)^{-d} \int_{S_d} \xi^{2\alpha_0} d\mu(\xi)\right]^{1/2}$$

where $S_d = \{x \in \mathbf{R}^d : \|x\| = 1\}$ for $d = 2, 3, \dots$, $S_1 = [-1, 1]$, μ is the Lebesgue measure of S_d so that $\mu(S_d) = 2\pi^{d/2}/\Gamma(d/2)$, $d = 1, 2, \dots$, and $B(\cdot, \cdot)$ denotes the beta-function.

Proposition 1 *Let $r \geq 0$ be an integer and $s > r$. Then for the Sobolev semi-norm ρ_s and $T(f) = f^{(\alpha_0)}(0)$, $|\alpha_0| = r$, the extremal function $g_{s,1,1}$ is given by $g_{s,1,1}(x) =$*

$(-1)^r a\tilde{K}_s(bx)$ *where* $\tilde{K}_s$ *is defined in (21),* $a = C_*^{-1}b^{-s}$ *and* $b = \left(\frac{2(s-r)}{2r+d}\right)^{1/(2s+d)}$. *Furthermore,*

$$K_s(x) = b^{r+d}\tilde{K}_s(bx),$$
$$\|K_s\|_2 = C_*\left(\frac{2(s-r)}{2r+d}\right)^{(s+r+d)/(2s+d)},$$
$$T(g_{s,1,1}) = C_*\left(\frac{2r+d}{2(s-r)}\right)^{(s-r)/(2s+d)}\frac{2s+d}{2r+d}.$$

For the one-dimensional case ($d = 1$) the extremal function $g_{s,1,1}$ was found by Taikov (1968).

The kernel $\tilde{K}_s$ can be expressed in terms of the Bessel functions. Thus, if $r = 0$, we have

$$\tilde{K}_s(x) = (2\pi)^{-d/2}\|x\|^{1-d/2}\int_0^\infty \frac{t^{d/2}}{1+t^{2\beta}}J_{(d-1)/2}(t\|x\|)dt$$

where J_n is the ordinary Bessel function of order n. For $r \neq 0$ the kernel $\tilde{K}_s$ is the corresponding directional derivative of the right hand side of the last equality.

Proposition 2 *Let* ρ_s *be the Sobolev semi-norm, and let* $T(f)$ *be as in Proposition 1. Then Assumptions 1-4 are satisfied with any* s_*, s^*, r' *such that* $r < r' \le s_* < s^* < \infty$, $r' > (s^* - r - d)/2$.

2. Taylor and Hölder classes. The Taylor classes are defined by (1) with

$$\rho_s(f) = \sup_{x\neq 0}\|x\|^{-s}\left|f(x) - \sum_{i=0}^{\lceil s-1\rceil}\sum_{|\alpha|=i}\frac{x^\alpha}{\alpha!}f^{(\alpha)}(0)\right| \tag{22}$$

where $s > 0$, $f^{(\alpha)}(0)$ is the partial derivative corresponding to the multi-index α and $\alpha! = \alpha_1!\cdots\alpha_d!$. Consider the estimation of the functional $T(f) = f(0)$ when f is in a Taylor class. This problem was studied in a non-adaptive setting of linear minimax estimation by Legostaeva and Shiryaev (1971), Sacks and Ylvisaker (1981). The extremal function $g_{s,1,1}$ of the maximization problem (4) is given by the following proposition which is a multivariate generalization of their results. This function is a renormalized version of the basic kernel

$$\tilde{K}_s(x) = (1 - \|x\|^s)\, I(\|x\| \le 1). \tag{23}$$

Proposition 3 *Let $r = 0,\ 0 < s \le 2$. Then, for ρ_s defined in (22) and $T(f) = f(0)$ the extremal function $g_{s,1,1}$ is given by $g_{s,1,1}(x) = a\tilde{K}_s(bx)$ where $\tilde{K}_s$ is defined in (23), $a = b^{-s}$, and*

$$b = \left(\frac{2s^2\mu(S_d)}{(2s+d)(s+d)d}\right)^{1/(2s+d)}.$$

Furthermore,

$$K_s(x) = \left(\frac{\mu(S_d)s}{(s+d)d}\right)^{-2s/(2s+d)} \left(\frac{2s}{2s+d}\right)^{d/(2s+d)} \tilde{K}_s(bx),$$

$$\|K_s\|_2 = \frac{(s+d)d}{\mu(S_d)s}\left(\frac{2s^2\mu(S_d)}{(2s+d)(s+d)d}\right)^{(s+d)/(2s+d)},$$

$$T(g_{s,1,1}) = \left(\frac{(2s+d)(s+d)d}{2s^2\mu(S_d)}\right)^{s/(2s+d)}.$$

For $0 < s \le 1$ the Hölder semi-norm is defined by

$$\rho_s(f) = \sup_{x,y\in\mathbf{R^d},x\neq y} \|x\|^{-s}\,|f(x) - f(y)| \tag{24}$$

and the Hölder classes are defined by (1) with ρ_s as in (24).

Proposition 4 *The result of Proposition 3 remains valid if ρ_s is the Hölder semi-norm and $0 < s \le 1$.*

Proposition 5 *Let $T(f) = f(0)$. Then Assumptions 1 - 4 are satisfied with any $0 < s_* < s^* \le 2$ for the Taylor classes and with any $0 < s_* < s^* \le 1$ for the Hölder classes.*

The Hölder semi-norm can be also defined for $s > 1$. However, explicit expressions for the solution $g_{s,1,1}$ and the kernel K_s are not generally known for $s > 1$, even in the dimension $d = 1$. This does not allow to construct sharp adaptive estimators on the Hölder scale with $s^* > 1$. Thus, for $d = 1, s > 1$ the Hölder semi-norm is

$$\rho_s(f) = \sup_{x,y\in\mathbf{R},x\neq y} \frac{|f^{(l)}(x) - f^{(l)}(y)|}{|x-y|^{s-l}}$$

where $l = \lfloor s \rfloor$. Explicit solutions of the extremal problem (4) with this semi-norm are available only for $0 < s \le 1$ and $s = 2$ [see Fuller (1961), Gabushin (1968), Korostelev (1994), Leonov (1997, 1999), Zhao (1997)].

6 Lemmas

The following lemma gives a bound for the bias of a kernel estimator.

Lemma 1 *Let $h > 0$ and $s', s > r$. Then under Assumptions 1 and 2,*

$$\sup_{f \in \mathcal{F}_{s,L}} |E_f T_{s',\varepsilon} - T(f)| \leq L h^{s-r} b_{s,s'}.$$

Proof. Using Assumptions 1 and 2, we have

$$\begin{aligned}
\sup_{f \in \mathcal{F}_{s,L}} \left| E_f \int K_{s',h} dY_\varepsilon - T(f) \right| &= \sup_{\rho_s(f) \leq L} \left| \int K_{s',h} f - T(f) \right| \\
&= \sup_{\rho_s(f) \leq L} \left| \int K_{s'}(x) h^{-r} f(hx) dx - T(f) \right| \\
&= L h^{s-r} \sup_{\rho_s(f) \leq L} \left| \int K_{s'}(x) \frac{f(hx)}{Lh^s} dx - T\left(\frac{f(h\,\cdot)}{Lh^s} \right) \right| \\
&= L h^{s-r} \sup_{\rho_s(f) \leq 1} \left| \int K_{s'} f - T(f) \right|
\end{aligned}$$

since

$$\left\{ \frac{f(h\,\cdot)}{Lh^s} : \rho_s(f) \leq L \right\} = \{ f : \rho_s(f) \leq 1 \}.$$

□

Recall that $h_l(s, L, \varepsilon) = (\varepsilon/L)^{2/(2s+d)}$. Define the bias term:

$$B(s, L, \varepsilon) = L h_l^{s-r}(s, L, \varepsilon) b_{s,s}$$

and the standard deviation term:

$$R(s, L, \varepsilon) = \varepsilon h_l^{-r-d/2}(s, L, \varepsilon) \|K_s\|_2.$$

Lemma 2 *Suppose that Assumptions 1 and 2 hold and let $g_{s,L,\varepsilon}$ be defined by (3), $\varepsilon > 0$, $L > 0$, $s > r$. Then*

(i) $\|g_{s,L,\varepsilon}\|_2 = \varepsilon$,

(ii) if $\rho_s(g) \leq L$ *and* $T(g) = T(g_{s,L,\varepsilon})$, *then* $\int g g_{s,L,\varepsilon} \geq \|g_{s,L,\varepsilon}\|_2^2$,

(iii) $\psi_\nu = B(s, L, \tilde{\varepsilon}(s)) + R(s, L, \tilde{\varepsilon}(s))$, *where* ψ_ν *is defined by (14).*

Proof. Assertion (i) means that the solution $g_{s,L,\varepsilon}$ is attained at the boundary of the set of restrictions, while (ii) follows from the usual duality argument (see Gabushin (1970), Micchelli and Rivlin (1977), Arestov (1989)). An elementary proof of (i) and (ii) for the case $T(f) = f(0)$ is given in Lepski and Tsybakov (2000). It can be easily extended to our general case. Let us prove (iii). In view of (6),

$$T(g_{s,1,1}) = b_{s,s} + \|K_s\|_2. \tag{25}$$

Now, by standard renormalization argument, $g_{s,L,\varepsilon}(x) = a g_{s,1,1}(bx)$ where

$$a = Lb^{-s} = \varepsilon^{2s/(2s+d)} L^{d/(2s+d)}, \ b = (L/\varepsilon)^{2/(2s+d)} = h_l^{-1}(s, L, \varepsilon).$$

Thus,

$$ab^r = Lh_l^{s-r}(s, L, \varepsilon) = \varepsilon h_l^{-r-d/2}(s, L, \varepsilon),$$

and, by (25),

$$T(g_{s,L,\varepsilon}) = ab^r T(g_{s,1,1}) = B(s, L, \varepsilon) + R(s, L, \varepsilon).$$

This and (15) yield (iii). □

For $s > s' > 0$ define

$$\begin{aligned} d_\varepsilon(s', s) &= \left[2p(\kappa(s) - \kappa(s')) \log \varepsilon^{-1}\right]^{1/2} \\ &= \left[2p(2r+d)\left(\frac{1}{2s'+d} - \frac{1}{2s+d}\right) \log \varepsilon^{-1}\right]^{1/2}. \end{aligned} \tag{26}$$

Lemma 3 *Let Assumptions 1 - 4 hold. Let $s, s' \in [r', q]$, $r < r' < q$, $s' < s$, $L \in [L_*, L^*]$, and denote $\nu = (s, L)$. Then there exist positive constants $D_1, \ldots, D_5$ (that can depend only on $s_*, s^*, L_*, L^*, r, q, d, p$) such that*

$$\frac{\psi_{s',L}}{\psi_\nu} \le D_1 \exp\left\{\frac{1}{2p} d_\varepsilon^2(s', s)\right\}, \tag{27}$$

$$\frac{\psi_{s',L}}{\psi_\nu} \ge D_2 \left(\varepsilon^2 \log \frac{1}{\varepsilon}\right)^{(\kappa(s')-\kappa(s))/2}, \tag{28}$$

$$\frac{h(s, \tilde\varepsilon(s))}{h(s', \tilde\varepsilon(s'))} \ge D_3 \exp\left\{\frac{2}{(2s^*+d)^2}(s - s') \log \varepsilon^{-1}\right\}, \tag{29}$$

and

$$D_4 \le \frac{\psi_{s,L}}{\eta(s)} \le D_5. \tag{30}$$

Proof. From (15) we have $\psi_{s,L} = \tilde{\varepsilon}^{2(s-r)/(2s+d)} L^{(2r+d)/(2s+d)} T(g_{s,1,1})$ where $\tilde{\varepsilon} = \varepsilon d_\varepsilon(s, s^*)$. Thus,

$$\frac{\psi_{s',L}}{\psi_\nu} = \exp\left\{\frac{1}{2p}\, d_\varepsilon^2(s', s)\right\} \frac{d_\varepsilon^{2(s'-r)/(2s'+d)}(s', s^*)}{d_\varepsilon^{2(s-r)/(2s+d)}(s, s^*)}\, \frac{T(g_{s',1,1})}{T(g_{s,1,1})} L^{\frac{2r+d}{2s'+d} - \frac{2r+d}{2s+d}}.$$

Note that $\|K_s\|_2 > 0$ for every s, since otherwise $\|g_{s,1,1}\|_2 = 0$ (by (7)) which contradicts Lemma 2 (i). Also, by Assumption 3, $\|K_s\|_2$ is a continuous function of s on the interval $[r', s^*]$ for $r' > r$. Hence

$$\inf_{s\in[r',s^*]} \|K_s\|_2 > 0, \quad \sup_{s\in[r',s^*]} \|K_s\|_2 < \infty. \tag{31}$$

Combining (31) with (25) and taking into account Assumption 4(i) we get

$$\inf_{s\in[r',s^*]} T(g_{s,1,1}) > 0, \quad \sup_{s\in[r',s^*]} T(g_{s,1,1}) < \infty. \tag{32}$$

Now, for $s_0 < s_1$,

$$d_\varepsilon^2(s_0, s_1) = \frac{4p(2r+d)}{(2s_0+d)(2s_1+d)}\,(s_1 - s_0)\log\varepsilon^{-1} \tag{33}$$

and thus, for $s \in [r', q]$,

$$\frac{4p(2r+d)}{(2s^*+d)^2}\,(s^* - q)\log\varepsilon^{-1} \le d_\varepsilon^2(s, s^*) \le \frac{4p}{2r+d}\,(s^* - r)\log\varepsilon^{-1}.$$

Therefore,

$$D_6\left(\log\varepsilon^{-1}\right)^{(\kappa(s')-\kappa(s))/2} \le \frac{d_\varepsilon^{2(s'-r)/(2s'+d)}(s', s^*)}{d_\varepsilon^{2(s-r)/(2s+d)}(s, s^*)} \le D_7\left(\log\varepsilon^{-1}\right)^{(\kappa(s')-\kappa(s))/2}$$

for $D_6, D_7 > 0$. Observing that $(\kappa(s') - \kappa(s))/2 < 0$ and using (32) we obtain (27) and (28). To prove the bound (29) it is enough to note that

$$\frac{h(s, \tilde{\varepsilon}(s))}{h(s', \tilde{\varepsilon}(s'))} = \exp\left\{\frac{4(s-s')}{(2s+d)(2s'+d)}\left(\log\varepsilon^{-1} - \frac{1}{2}\,\log\log\varepsilon^{-1}\right)\right\} \frac{\lambda(s)^{1/(2s+d)}}{\lambda(s')^{1/(2s'+d)}}. \tag{34}$$

The bounds (30) follow from the equality

$$\frac{\psi_{s,L}}{\eta(s)} = L^{(2r+d)/(2s+d)} \frac{T(g_{s,1,1})}{\|K_s\|_2}$$

in view of (31) - (32). □

We need some exponential bounds for the stochastic part of the estimator. Define

$$Z_s = \varepsilon \int K_{s,h(s,\tilde{\varepsilon}(s))}(t)\,dW(t). \tag{35}$$

Lemma 4 *Let σ_s be as defined in (11). Then for $u > 0$, $p \geq 0$,*

$$E\left[|Z_s|^p I(|Z_s| \geq u)\right] \leq D(p)\left(\sigma_s^p + u^p\right)\exp\left\{-\frac{u^2}{2\sigma_s^2}\right\}.$$

where $D(p) > 0$ is a constant depending only on p.

Proof of this lemma is straightforward since $Z_s \sim \mathcal{N}(0, \sigma_s^2)$.

7 Proofs of the results

In the following we denote $C, C', C_1, C_2, \ldots$ positive constants that can depend only on $s_*, s^*, L_*, L^*, r, q, d, p$. These constants may be different in different occasions.

Proof of the upper bound in Theorem 1.

Here we prove the bound (16). Consider $\nu = (s, L) \in B_q$ and define $s^- = s^-(s)$ by

$$s^- = s - \frac{\log\log\log(1/\varepsilon)}{\log(1/\varepsilon)}.$$

We have

$$\sup_{f\in\mathcal{F}_\nu} E_f\left(\psi_\nu^{-p}\left|T_\varepsilon^* - T(f)\right|^p\right) = R_{\varepsilon,\nu}^- + R_{\varepsilon,\nu}^+$$

where

$$R_{\varepsilon,\nu}^- = \sup_{f\in\mathcal{F}_\nu} E_f\left(\psi_\nu^{-p}\left|T_\varepsilon^* - T(f)\right|^p I\left(\hat{s} < s^-\right)\right),$$

$$R_{\varepsilon,\nu}^+ = \sup_{f\in\mathcal{F}_\nu} E_f\left(\psi_\nu^{-p}\left|T_\varepsilon^* - T(f)\right|^p I\left(\hat{s} \geq s^-\right)\right).$$

To show (16), we will prove that

$$\lim_{\varepsilon\to 0}\sup_{\nu\in B_q} R_{\varepsilon,\nu}^- = 0 \tag{36}$$

and

$$\limsup_{\varepsilon\to 0}\sup_{\nu\in B_q} R_{\varepsilon,\nu}^+ \leq 1. \tag{37}$$

Proof of (36). Let $s \in [s_*, q]$, $s' \in S$, $s' < s^-$, $L \in [L_*, L^*]$. Let $f \in \mathcal{F}_\nu$, $\nu = (s, L)$. Then for sufficiently small ε, using Lemma 1, (29) and the fact that $s' < s$, we get

$$|E_f T_{s',\varepsilon} - T(f)| \leq L h^{s-r}(s', \tilde{\varepsilon}(s')) b_{s,s'} \leq L h^{s-r}(s, \tilde{\varepsilon}(s)) b_{s,s'}. \tag{38}$$

By definition,

$$h^{s-r}(s,\tilde{\varepsilon}(s)) = \tilde{\varepsilon}^{\kappa(s)}(s,\tilde{\varepsilon}(s)). \tag{39}$$

where $\kappa(s) = 2(s-r)/(2s+d)$. Comparing this to (15) and using the inequality $b_{s,s'} \leq b_{max}(r', s^*)$ (Assumption 4(i)), we get

$$Lh^{s-r}(s,\tilde{\varepsilon}(s))b_{s,s'} \leq C_1\psi_\nu. \tag{40}$$

Then

$$|T_{s',\varepsilon} - T(f)| \leq |E_f T_{s',\varepsilon} - T(f)| + |Z_{s'}| \leq C_1\psi_\nu + |Z_{s'}|$$

where $Z_{s'}$ is defined in (35) as a stochastic error of the kernel estimator. Using this we find

$$R^-_{\varepsilon,\nu} \leq \sum_{s'\in S, s'<s^-} \sup_{f\in\mathcal{F}_\nu} E_f\left(\psi_\nu^{-p}\left|T_{s',\varepsilon} - T(f)\right|^p I\left(\hat{s} = s'\right)\right) \leq g_1(\nu) + g_2(\nu)$$

where

$$g_1(\nu) = C \sum_{s'\in S, s'<s^-} \sup_{f\in\mathcal{F}_\nu} P_f\left(\hat{s} = s'\right)\left(1 + \psi_\nu^{-1}\tau(s')\right)^p \tag{41}$$

and

$$g_2(\nu) = C \sum_{s'\in S, s'<s^-} E\left[\left(1 + \psi_\nu^{-1}\left|Z_{s'}\right|\right)^p I\left(|Z_{s'}| \geq \tau(s')\right)\right] \tag{42}$$

with

$$\tau(s') = \sigma_{s'}\left[d_\varepsilon(s',s) + \left(\log\varepsilon^{-1}\right)^{1/4}\right]$$

where $\sigma_{s'}$ is defined in (11) and $d_\varepsilon(s', s)$ is defined in (26).

Let us prove that the probability of underestimating largely the value of s by the statistic $\hat{s}$ is small, uniformly over $f \in \mathcal{F}_\nu$.

Lemma 5 *Let $s \in [s_*, q]$, $s' \in S$, $s' < s^-$, $L \in [L_*, L^*]$ and $\nu = (s, L)$. Then for every $\delta > 0$ there exists $\varepsilon_0 = \varepsilon_0(\delta) > 0$ independent of s, s', L and such that for all $0 < \varepsilon < \varepsilon_0$ we have*

$$\sup_{f\in\mathcal{F}_\nu} P_f\left(\hat{s} = s'\right) \leq Cm\exp\left\{-\frac{1}{2}\, d^2_\varepsilon(s', s^*)(1-\delta)\right\}.$$

Proof. Since $\mathrm{Card}(S) = m$, we have

$$\begin{aligned}\sup_{f\in\mathcal{F}_\nu} P_f\left(\hat{s}=s'\right) &\le \sum_{s''\in S, s''\le s'} \sup_{f\in\mathcal{F}_\nu} P_f\left(\left|T_{\bar{s'},\varepsilon}-T_{s'',\varepsilon}\right| > \eta(s'')\right)\\ &\le m \max_{s''\in S, s''\le s'} \sup_{f\in\mathcal{F}_\nu} P_f\left(\left|T_{\bar{s'},\varepsilon}-T_{s'',\varepsilon}\right| > \eta(s'')\right) \qquad (43)\end{aligned}$$

where $\bar{s'} = \bar{s'}(s')$ is the smallest element of S greater than s'. Let $f \in \mathcal{F}_\nu$. Arguing as in (38) and (40), and using (28), for sufficiently small ε we get

$$\left|E_f T_{\bar{s'},\varepsilon} - T(f)\right| + \left|E_f T_{s'',\varepsilon} - T(f)\right| \le C_1\psi_\nu \le C_1\gamma_\varepsilon\psi_{s'',L}$$

where

$$\gamma_\varepsilon = \gamma_\varepsilon(s,s'') = D_2^{-1}\left(\varepsilon^2\log\frac{1}{\varepsilon}\right)^{(\kappa(s)-\kappa(s''))/2} \le D_2^{-1}\exp\left\{-C(\kappa(s)-\kappa(s''))\log(1/\varepsilon)\right\}$$

$$\le D_2^{-1}\exp\left\{-C(\kappa(s)-\kappa(s^-))\log(1/\varepsilon)\right\} \le D_2^{-1}\exp\left\{-C'\log\log\log(1/\varepsilon)\right\}.$$

Note that $\gamma_\varepsilon(s,s'') \to 0$ as $\varepsilon \to 0$ uniformly in $s, s'' \in [r', q]$.

Using (30) we obtain $\psi_{s'',L} \le D_5\eta(s'')$ and thus

$$\begin{aligned}\left|T_{\bar{s'},\varepsilon}-T_{s'',\varepsilon}\right| &\le \left|E_f T_{\bar{s'},\varepsilon} - T(f)\right| + \left|E_f T_{s'',\varepsilon} - T(f)\right| + \left|Z_{\bar{s'}} - Z_{s''}\right|\\ &\le C_3\gamma_\varepsilon\eta(s'') + \left|Z_{\bar{s'}} - Z_{s''}\right|.\end{aligned}$$

Hence, for sufficiently small ε,

$$P_f\left(\left|T_{\bar{s'},\varepsilon}-T_{s'',\varepsilon}\right| > \eta(s'')\right) \le P_f\left(\left|Z_{\bar{s'}} - Z_{s''}\right| > \eta(s'')(1-C_3\gamma_\varepsilon)\right). \qquad (44)$$

Denote $h_0 = h(s'', \tilde{\varepsilon}(s''))$, $h_1 = h(\bar{s'}, \tilde{\varepsilon}(\bar{s'}))$, $\bar{K}_0 = K_{s''}$, $\bar{K}_1 = K_{\bar{s'}}$, and, as usually, $\bar{K}_{i,h_i} = h_i^{-r-d}\bar{K}_i(\cdot/h_i)$. Now $Z_{\bar{s'}} - Z_{s''} \sim \mathcal{N}(0, \varepsilon^2 A^2)$ where

$$A = \left\|\bar{K}_{0,h_0} - \bar{K}_{1,h_1}\right\|_2 \le \left\|\bar{K}_{0,h_0} - \bar{K}_{0,h_1}\right\|_2 + \left\|\bar{K}_{0,h_1} - \bar{K}_{1,h_1}\right\|_2 \stackrel{\text{def}}{=} A_1 + A_2.$$

Now

$$\begin{aligned}A_1^2 &= h_0^{-2r-d}\int\left[\bar{K}_0(x) - \left(\frac{h_0}{h_1}\right)^{r+d}\bar{K}_0\left(\frac{h_0}{h_1}x\right)\right]^2 dx\\ &= h_0^{-2r-d}\left[\int\bar{K}_0^2 + \left(\frac{h_0}{h_1}\right)^{2r+d}\int\bar{K}_0^2 - 2\left(\frac{h_0}{h_1}\right)^{r+d}\int\bar{K}_0(x)\bar{K}_0\left(\frac{h_0}{h_1}x\right)dx\right].\end{aligned}$$

Denote

$$\tilde{g}(x) = \left(\frac{h_0}{h_1}\right)^{-r} g_{s'',1,1}\left(\frac{h_0}{h_1}\,x\right).$$

Then, because $h_0 < h_1$,

$$\rho_{s''}(\tilde{g}) = \left(\frac{h_0}{h_1}\right)^{s''-r} \rho_{s''}(g_{s'',1,1}) \le 1$$

and $T(\tilde{g}) = T(g_{s'',1,1})$. Lemma 2 (ii) gives

$$\int g_{s'',1,1}\tilde{g} \ge \int g^2_{s'',1,1}$$

and therefore, by (7),

$$\left(\frac{h_0}{h_1}\right)^{-r} \int \bar{K}_0(x)\bar{K}_0\left(\frac{h_0}{h_1}\,x\right) dx \ge \int \bar{K}_0^2.$$

Thus

$$A_1^2 \le h_0^{-2r-d}\left[1 - \left(\frac{h_0}{h_1}\right)^{2r+d}\right]\int \bar{K}_0^2 \le h_0^{-2r-d}\int \bar{K}_0^2. \tag{45}$$

Also,

$$A_2^2 = h_1^{-2r-d}\left\|\bar{K}_0 - \bar{K}_1\right\|_2^2 = h_0^{-2r-d}\left\|\bar{K}_0\right\|_2^2 \gamma'_\varepsilon \tag{46}$$

where

$$\gamma'_\varepsilon = \gamma'_\varepsilon(\bar{s'}, s'') = \left(\frac{h_0}{h_1}\right)^{2r+d} \frac{\left\|\bar{K}_0 - \bar{K}_1\right\|_2^2}{\left\|\bar{K}_0\right\|_2^2} \longrightarrow 0$$

as $\varepsilon \to 0$. Indeed, to prove this consider the two cases: (i) $\bar{s'} - s'' \ge (\log\varepsilon^{-1})^{-1/2}$ and (ii) $0 < \bar{s'} - s'' < (\log\varepsilon^{-1})^{-1/2}$. If (i) holds then, by (29), $h_0/h_1 \le D_3^{-1}\exp(-C\sqrt{\log\varepsilon^{-1}})$, and, using (31), we get $\gamma'_\varepsilon \le C\exp(-C\sqrt{\log\varepsilon^{-1}})$. If (ii) holds, then (29) and the inequality $\bar{s'} > s''$ entail $h_0/h_1 \le C$, while $\left\|\bar{K}_0 - \bar{K}_1\right\|_2 \le \Omega(\bar{s'} - s'') \le \Omega((\log\varepsilon^{-1})^{-1/2})$ where $\Omega(\delta) = \sup\{\|K_s - K_{s'}\|_2 : \ |s - s'| \le \delta,\ s, s' \in [r', s^*]\}$. Note that $\Omega(\delta) \to 0$, as $\delta \to 0$, by the uniform continuity of the function $F(s,s') = \|K_s - K_{s'}\|_2$ on $[r', s^*] \times [r', s^*]$ which follows from Assumption 3. Thus, in both cases (i) and (ii) we have $\gamma'_\varepsilon(\bar{s'}, s'') \to 0$ as $\varepsilon \to 0$, uniformly in $\bar{s'}, s'' \in [r', s^*]$.

From (45) and (46) it follows that for sufficiently small ε,

$$\varepsilon^2 A^2 \le \varepsilon^2 h^{-2r-d}(s'', \tilde{\varepsilon}(s''))\|K_{s''}\|_2^2(1 + \gamma'_\varepsilon) = \sigma^2_{s''}(1 + \gamma'_\varepsilon).$$

Now, for sufficiently small ε,

$$\frac{\eta^2(s'')}{\varepsilon^2 A^2} \geq \frac{\eta^2(s'')}{\sigma_{s''}^2(1+\gamma'_\varepsilon)} = \frac{d_\varepsilon^2(s'',s^*)}{1+\gamma'_\varepsilon} \geq \frac{d_\varepsilon^2(s',s^*)}{1+\gamma'_\varepsilon}.$$

Thus, using Lemma 4 we get

$$\begin{aligned} P_f\left(|Z_{\bar{s}'} - Z_{s''}| > \eta(s'')(1-C_3\gamma_\varepsilon)\right) &\leq C_4 \exp\left\{-\frac{1}{2\varepsilon^2 A^2}\,\eta^2(s'')(1-C_3\gamma_\varepsilon)^2\right\} \\ &\leq C_4 \exp\left\{-\frac{1}{2}\,d_\varepsilon^2(s',s^*)(1-C_3\gamma_\varepsilon)^2(1+\gamma'_\varepsilon)^{-1}\right\}. \end{aligned}$$

Comparing this to (43) and (44) we get the lemma. □

Lemma 6 *Let g_1 be defined in (41). Then,*

$$\lim_{\varepsilon\to 0}\sup_{\nu\in B_q} g_1(\nu) = 0.$$

Proof. Let $s \in [s_*, q]$, $s' \in S$, $s' < s^-$, $L \in [L_*, L^*]$, $\nu = (s, L)$. By definitions (see also (33)),

$$\frac{\tau(s')}{\eta(s')} = \frac{d_\varepsilon(s',s) + (\log\varepsilon^{-1})^{1/4}}{d_\varepsilon(s',s^*)} \leq C_1.$$

Next, by (30),

$$\frac{\eta(s')}{\psi_\nu} \leq C_2\,\frac{\psi_{s',L}}{\psi_\nu}$$

Combining the two previous inequalities and using (27), we get

$$\frac{\tau(s')}{\psi_\nu} \leq C_3\,\frac{\psi_{s',L}}{\psi_\nu} \leq C_4 \exp\left\{\frac{1}{2p}\,d_\varepsilon^2(s',s)\right\}. \tag{47}$$

Using this and Lemma 5, we find

$$\begin{aligned} g_1(\nu) &\leq C_5 \sum_{s'\in S, s'<s^-} \sup_{f\in\mathcal{F}_\nu} P_f\left(\hat{s}=s'\right)\left(\frac{\tau(s')}{\psi_\nu}\right)^p \\ &\leq C_6\, m \sum_{s'\in S, s'<s^-}\left[\exp\left\{\frac{1}{2}\,d_\varepsilon^2(s',s)\right\}\exp\left\{-\frac{1}{2}\,d_\varepsilon^2(s',s^*)(1-\delta)\right\}\right] \\ &\leq C_7\, m^2 \exp\left\{-c(\delta)\log\varepsilon^{-1}\right\} \end{aligned}$$

where

$$c(\delta) = p(2r+d)\left(\frac{1}{2s+d} - \frac{\delta}{2r+d} - \frac{1-\delta}{2s^*+d}\right).$$

Choose $\delta = \frac{(s^*-q)(2r+d)}{2(s^*-r)(2q+d)}$. Then $c(\delta) \geq C_8 > 0$. Moreover, in view of (10),

$$m = O([\log(1/\varepsilon)]^{\gamma_1}), \quad \varepsilon \to 0, \tag{48}$$

and the lemma follows. □

Lemma 7 *Let g_2 be defined in (42). Then,*

$$\lim_{\varepsilon\to 0} \sup_{\nu\in B_q} g_2(\nu) = 0.$$

Proof. Let $s \in [s_*, q]$, $s' \in S$, $s' < s^-$, $L \in [L_*, L^*]$, $\nu = (s, L)$. Now,

$$\frac{\tau^2(s')}{\sigma^2_{s'}} = \left[d_\varepsilon(s', s) + \left(\log \varepsilon^{-1}\right)^{1/4}\right]^2 \geq \left(\log \varepsilon^{-1}\right)^{1/2}.$$

This, together with Lemma 4 and (47), (48) yields

$$\begin{aligned} g_2(\nu) &\leq C_1 \sum_{s'\in S, s'<s^-} \left[1 + \psi_\nu^{-p}\left(\sigma^p_{s'} + \tau^p(s')\right)\right] \exp\left\{-\frac{\tau^2(s')}{2\sigma^2_{s'}}\right\} \\ &\leq C_2\, m \, \exp\left\{-\frac{1}{2}\left(\log \varepsilon^{-1}\right)^{1/2}\right\} \longrightarrow 0 \end{aligned}$$

uniformly in $\nu \in B_q$, as $\varepsilon \to 0$. □

Lemmas 6 and 7 imply (36).

Proof of (37). Let $s \in [s_*, q]$, $L \in [L_*, L^*]$, $\nu = (s, L)$. Let $\bar{s} = \bar{s}(s)$ be defined by

$$\bar{s} = s - \frac{(2s + d)\log L}{2\log(1/\varepsilon) - \log\log(1/\varepsilon) + 2\log L}.$$

In other words, $\bar{s}$ is chosen so that

$$\left(L^{-2}\varepsilon^2 \log \varepsilon^{-1}\right)^{1/(2s+d)} = \left(\varepsilon^2 \log \varepsilon^{-1}\right)^{1/(2\bar{s}+d)}. \tag{49}$$

Let $s^+ \in S$ be the largest grid point $\leq \bar{s}$. Denote $\mathcal{S}_1 = \mathcal{S}_1(s) = \{s' \in S : s^- \leq s' \leq s^+\}$ and $\mathcal{S}_2 = \mathcal{S}_2(s) = \{s' \in S : s^+ < s' \leq q\}$. Assume that ε is small enough, so that $s^- < s^+$. We have

$$R^+_{\varepsilon,\nu} = \sup_{f\in\mathcal{F}_\nu} E_f\left(\psi_\nu^{-p} \left|T^*_\varepsilon - T(f)\right|^p I\left(\hat{s} \in \mathcal{S}_1 \cup \mathcal{S}_2\right)\right).$$

Let $s' \in \mathcal{S}_1$ and $f \in \mathcal{F}_\nu$. Using successively Lemma 1, the fact that $s' \leq \bar{s}$ and (49) we get

$$\begin{aligned}
|E_f T_{s',\varepsilon} - T(f)| &\leq Lh^{s-r}(s', \tilde{\varepsilon}(s'))b_{s,s'} \\
&= L\lambda(s')^{(s-r)/(2s'+d)} \left(\varepsilon^2 \log \varepsilon^{-1}\right)^{(s-r)/(2s'+d)} b_{s,s'} \\
&\leq L\lambda(s')^{(s-r)/(2s'+d)} \left(\varepsilon^2 \log \varepsilon^{-1}\right)^{(s-r)/(2\bar{s}+d)} b_{s,s'} \\
&= L\lambda(s')^{(s-r)/(2s'+d)} \left(L^{-2}\varepsilon^2 \log \varepsilon^{-1}\right)^{(s-r)/(2s+d)} b_{s,s'} \\
&= \Lambda(s,s') L h_l^{s-r}(s, L, \tilde{\varepsilon}(s)) b_{s,s'} \qquad (50)
\end{aligned}$$

where $\Lambda(s,s') = \lambda(s')^{(s-r)/(2s'+d)}\lambda(s)^{-(s-r)/(2s+d)}$. Note that

$$|s - s'| \leq \frac{C \log\log\log(1/\varepsilon)}{\log(1/\varepsilon)}, \quad \forall s' \in \mathcal{S}_1. \qquad (51)$$

This and the uniform continuity of $\Lambda(s,s')$ in $s, s' \in [s_*, q]$ yields that $\Lambda(s,s') \leq 1 + \gamma_{\varepsilon 1}$. (Here and later we denote $\gamma_{\varepsilon i}, i = 1, 2, \ldots$ the functions of ε that can depend only on $s_*, s^*, L_*, L^*, r, q, d, p$ and such that $\lim_{\varepsilon\to 0} \gamma_{\varepsilon i} = 0$.) Next, using Assumption 4(ii) and (51), for every $s' \in \mathcal{S}_1$, $s \in [s_*, s^*]$, we get $b_{s,s'} \leq b_{s,s}(1+\gamma_{\varepsilon 2})$. These remarks, (50) and Lemma 2 (iii), yield

$$|E_f T_{s',\varepsilon} - T(f)| \leq L h_l^{s-r}(s, L, \tilde{\varepsilon}(s)) b_{s,s}(1 + \gamma_{\varepsilon 3})$$

$$= B(s, L, \tilde{\varepsilon}(s))(1 + \gamma_{\varepsilon 3}) \leq \psi_\nu(1 + \gamma_{\varepsilon 3}), \quad \forall s' \in \mathcal{S}_1. \qquad (52)$$

From (27) and (51) we have $\psi_{s',L}/\psi_\nu \leq (\log\log(1/\varepsilon))^{C_1}$, $\forall s' \in \mathcal{S}_1$. This and (30) entail

$$\eta(s') \leq D_4^{-1}\psi_{s',L} \leq C_2(\log\log(1/\varepsilon))^{C_1}\psi_\nu, \quad \forall s' \in \mathcal{S}_1,$$

and

$$\begin{aligned}
\frac{\sigma_{s'}}{\psi_\nu} &\leq C_3(\log\log(1/\varepsilon))^{C_1}\frac{\sigma_{s'}}{\eta(s')} = \frac{C_3(\log\log(1/\varepsilon))^{C_1}}{d_\varepsilon(s', s^*)} \\
&\leq \frac{C_3(\log\log(1/\varepsilon))^{C_1}}{d_\varepsilon(q, s^*)} \leq \frac{C_4}{\log^{1/4}(1/\varepsilon)}, \quad \forall s' \in \mathcal{S}_1. \qquad (53)
\end{aligned}$$

Note also that, since $s^+ \leq \bar{s}$, the argument similar to (50), (52) and Assumption 3 yield

$$\eta(s^+) = \left(\lambda(s^+)\varepsilon^2 \log \varepsilon^{-1}\right)^{(s^+-r)/(2s^++d)} \|K_{s^+}\|_2$$

$$
\begin{aligned}
&\leq \lambda(s^+)^{(s^+-r)/(2s^++d)}(\varepsilon^2\log\varepsilon^{-1})^{(\bar{s}-r)/(2\bar{s}+d)}\|K_{s^+}\|_2(1+\gamma_{\varepsilon 4})\\
&= \lambda(s^+)^{(s^+-r)/(2s^++d)}\sqrt{\varepsilon^2\log\varepsilon^{-1}}\left(\frac{\varepsilon^2\log\varepsilon^{-1}}{L^2}\right)^{-(r+d/2)/(2s+d)}\|K_{s^+}\|_2(1+\gamma_{\varepsilon 4})\\
&\leq \tilde{\varepsilon}(s)h_l^{-r-d/2}(s,L,\tilde{\varepsilon}(s))\|K_s\|_2(1+\gamma_{\varepsilon 5})\\
&= R(s,L,\tilde{\varepsilon}(s))(1+\gamma_{\varepsilon 5}). \qquad (54)
\end{aligned}
$$

(For the first inequality in this display we used the fact that, by (10), $|\bar{s}-s^+| \leq k_2[\log(1/\varepsilon)]^{-\gamma}, \gamma > 1$, and thus $(\varepsilon^2\log\varepsilon^{-1})^{(s^+-r)/(2s^++d)-(\bar{s}-r)/(2\bar{s}+d)} \leq (1+\gamma_{\varepsilon 4})$.)

Now we are ready for the main argument of the proof. Let first $\hat{s} = s' \in \mathcal{S}_1$. Then, in view of (52),

$$
|T_\varepsilon^* - T(f)| = |T_{s',\varepsilon} - T(f)| \leq |E_f T_{s',\varepsilon} - T(f)| + |Z_{s'}| \leq \psi_\nu(1+\gamma_{\varepsilon 3}) + |Z_{s'}|. \qquad (55)
$$

Next, let $\hat{s} = s' \in \mathcal{S}_2$. Then, using successively the definition of $\hat{s}$, (52), (54) and Lemma 2 (iii), we get

$$
\begin{aligned}
|T_\varepsilon^* - T(f)| &\leq |T_{s',\varepsilon} - T_{s^+,\varepsilon}| + |T_{s^+,\varepsilon} - T(f)|\\
&\leq \eta(s^+) + |E_f T_{s^+,\varepsilon} - T(f)| + |Z_{s^+}|\\
&\leq R(s,L,\tilde{\varepsilon}(s))(1+\gamma_{\varepsilon 5}) + B(s,L,\tilde{\varepsilon}(s))(1+\gamma_{\varepsilon 3}) + |Z_{s^+}|\\
&\leq \psi_\nu(1+\gamma_{\varepsilon 6}) + |Z_{s^+}|.
\end{aligned}
$$

This and (55) entail

$$
\begin{aligned}
E_f\left(\psi_\nu^{-p}|T_\varepsilon^* - T(f)|^p I(\hat{s} \in \mathcal{S}_1 \cup \mathcal{S}_2)\right) &\leq \sum_{s' \in \mathcal{S}_1} E_f\left(\left(1+\gamma_{\varepsilon 3} + \psi_\nu^{-1}|Z_{s'}|\right)^p I(\hat{s} = s')\right)\\
&+ \sum_{s' \in \mathcal{S}_2} E_f\left(\left(1+\gamma_{\varepsilon 6} + \psi_\nu^{-1}|Z_{s^+}|\right)^p I(\hat{s} = s')\right). \qquad (56)
\end{aligned}
$$

Applying Lemma 4 and (53) we get, for any $s' \in \mathcal{S}_1$,

$$
\begin{aligned}
&E_f\left(\left(1+\gamma_{\varepsilon 3} + \psi_\nu^{-1}|Z_{s'}|\right)^p I(\hat{s} = s')\right) \leq \left(1+\gamma_{\varepsilon 3} + \sqrt{\sigma_{s'}\psi_\nu^{-1}}\right)^p P_f(\hat{s} = s')\\
&+ E_f\left(\left(1+\gamma_{\varepsilon 3} + \psi_\nu^{-1}|Z_{s'}|\right)^p I\left(|Z_{s'}| \geq \sqrt{\sigma_{s'}\psi_\nu}\right)\right)\\
&\leq \left(1+\gamma_{\varepsilon 3} + \sqrt{\sigma_{s'}\psi_\nu^{-1}}\right)^p P_f(\hat{s} = s')\\
&+ C_5\left[P_f\left(|Z_{s'}| \geq \sqrt{\sigma_{s'}\psi_\nu}\right) + \psi_\nu^{-p}E_f\left(|Z_{s'}|^p I\left(|Z_{s'}| \geq \sqrt{\sigma_{s'}\psi_\nu}\right)\right)\right]
\end{aligned}
$$

$$\leq \left(1+\gamma_{\varepsilon 3}+\sqrt{\sigma_{s'}\psi_\nu^{-1}}\right)^p P_f\left(\hat{s}=s'\right)+C_6\left[1+\psi_\nu^{-p}\left((\sigma_{s'}\psi_\nu)^{p/2}+\sigma_{s'}^p\right)\right]\exp\left\{-\frac{\psi_\nu}{2\sigma_{s'}}\right\}$$

$$\leq \left(1+\gamma_{\varepsilon 3}+C_4^{1/2}\left(\log(1/\varepsilon)\right)^{-1/8}\right)^p P_f\left(\hat{s}=s'\right)+C_7\exp\left\{-\frac{\log^{1/4}(1/\varepsilon)}{2C_4}\right\}.$$

Since $s^+\in\mathcal{S}_1$, analogous bound holds for $E_f\left((1+\gamma_{\varepsilon 6}+\psi_\nu^{-1}\left|Z_{s^+}\right|)^p I\left(\hat{s}=s'\right)\right)$. We conclude therefore that

$$R^+_{\varepsilon,\nu}=\sup_{f\in\mathcal{F}_\nu}E_f\left(\psi_\nu^{-p}\left|T^*_\varepsilon-T(f)\right|^p I\left(\hat{s}\in\mathcal{S}_1\cup\mathcal{S}_2\right)\right)$$

$$\leq \left(1+\gamma_{\varepsilon 7}+\frac{C_4^{1/2}}{\log^{1/8}(1/\varepsilon)}\right)^p\sup_{f\in\mathcal{F}_\nu}P_f\left(\hat{s}\in\mathcal{S}_1\cup\mathcal{S}_2\right)+2C_7 m\exp\left\{-\frac{\log^{1/4}(1/\varepsilon)}{2C_4}\right\} \quad (57)$$

where $\gamma_{\varepsilon 7}=\max\{\gamma_{\varepsilon 3},\gamma_{\varepsilon 6}\}$. It remains to note that (37) follows from (57) and (48). □

Proof of the lower bound in Theorem 1.

Here we prove the bound (17). The proof consists in reducing the problem to getting a lower bound on the risk of two hypotheses $f=f_0$ and $f=f_1$, which are chosen to be distant enough. In fact, f_1 will be chosen on the "boundary" of our scale of classes. Let $L\in[L_*,L^*]$, $\nu'=(s_*,L)$ and $\nu''=(q,L)$. Consider the functions

$$f_0\equiv 0,\;\; f_1=(1-\delta)g_{s_*,L,\tilde{\varepsilon}(s_*)}$$

where $0<\delta<1/2$ and $\tilde{\varepsilon}(s_*)=\varepsilon d_\varepsilon(s_*,s^*)$. Obviously, $f_0\in\mathcal{F}_{q,L}$. Furthermore, $\rho_{s_*}(f_1)=(1-\delta)\rho_{s_*}(g_{s_*,L,\tilde{\varepsilon}(s_*)})\leq(1-\delta)L$ and thus $f_1\in\mathcal{F}_{s_*,L}$. From equation (15), $T(f_1)=(1-\delta)\psi_{\nu'}$. Also, $T(f_0)=0$. Thus, for any estimator T_ε,

$$|T_\varepsilon-T(f_i)|=\psi_{\nu'}\,D\left((1-\delta)^{-1}\psi_{\nu'}^{-1}T_\varepsilon,i\right)\quad i=0,1,$$

where $D(u,v)=(1-\delta)|u-v|$, $u,v\in\mathbf{R}$. Denoting $Q=\psi_{\nu'}/\psi_{\nu''}$ and $E_i=E_{f_i}$, we get

$$\begin{aligned}
&\inf_{T_\varepsilon}\sup_{\nu\in B_q}\sup_{f\in\mathcal{F}_\nu}E_f\left(\psi_\nu^{-p}\left|T_\varepsilon-T(f)\right|^p\right)\\
&\quad\geq \inf_{T_\varepsilon}\max\left\{E_0\left(\psi_{\nu''}^{-p}\left|T_\varepsilon-T(f_0)\right|^p\right),E_1\left(\psi_{\nu'}^{-p}\left|T_\varepsilon-T(f_1)\right|^p\right)\right\}\\
&\quad= \inf_{T_\varepsilon}\max\left\{Q^pE_0D^p(T_\varepsilon,0),E_1D^p(T_\varepsilon,1)\right\}.
\end{aligned}$$

Denote for brevity $P_i=P_{f_i}$, $i=0,1$. We apply now the following lemma, which is a special case of Theorem 6 (i) in Tsybakov (1998) adapted to the present notation.

Lemma 8 *Let $Q > 0$, $\tau > 0$, $0 < \delta < 1/2$, $0 < \alpha < 1$ be fixed and let $D : \mathbf{R} \times \mathbf{R} \to [0, \infty)$ be a distance such that*

$$D(0,1) \geq 1 - \delta.$$

Suppose that

$$P_1 \left(\frac{dP_0}{dP_1} \geq \tau \right) \geq 1 - \alpha.$$

Then, for $p > 0$,

$$\inf_{T_\varepsilon} \max \left\{ Q^p E_0 D^p(T_\varepsilon, 0), E_1 D^p(T_\varepsilon, 1) \right\} \geq \frac{(1-\alpha)(1-2\delta)^p \tau (Q\delta)^p}{(1-2\delta)^p + \tau (Q\delta)^p}$$

where the infimum is taken over all estimators.

Let us check the assumptions of Lemma 8. Clearly, the assumption $D(0,1) \geq 1-\delta$ is satisfied for our definition of $D(\cdot,\cdot)$. Next, by Lemma 2 (i), $\|f_1\|_2^2 = (1-\delta)^2 \tilde{\varepsilon}^2(s_*)$. Put

$$\tau = \exp\left\{ -\frac{1-\delta}{2}\, d_\varepsilon^2(s_*, s^*) \right\}.$$

Then

$$P_1 \left(\frac{dP_0}{dP_1} \geq \tau \right) = P\left(\exp\left\{ \varepsilon^{-1} \|f_1\|_2 \xi - \varepsilon^{-2} \|f_1\|_2^2 / 2 \right\} \geq \tau \right) = 1 - \Phi(l_\varepsilon)$$

where $\xi \sim \mathcal{N}(0,1)$, $\Phi(\cdot)$ is a standard normal c.d.f. and

$$l_\varepsilon = \frac{\varepsilon}{\|f_1\|_2} \left(\log \tau + \varepsilon^{-2} \|f_1\|_2^2 / 2 \right) = -\frac{\delta}{2}\, d_\varepsilon(s_*, s^*) \longrightarrow -\infty, \qquad \varepsilon \to 0.$$

Hence, we can use Lemma 8 with the choice $\alpha = \Phi(l_\varepsilon)$ which results in

$$\inf_{T_\varepsilon} \max \left\{ Q^p E_0 D^p(T_\varepsilon, 0), E_1 D^p(T_\varepsilon, 1) \right\} \geq \frac{(1-\Phi(l_\varepsilon))(1-2\delta)^p \tau (Q\delta)^p}{(1-2\delta)^p + \tau (Q\delta)^p}. \tag{58}$$

Here, in view of (14) and (33),

$$\begin{aligned} \tau Q^p &= \exp\left\{ -\frac{1-\delta}{2}\, d_\varepsilon^2(s_*, s^*) \right\} \left(\frac{\psi_{\nu'}}{\psi_{\nu''}} \right)^p \\ &\geq \exp\left\{ \frac{2p(2r+d)}{(2s_*+d)(2s^*+d)} \left[q - s^* + \delta(s^* - s_*) \right] \log \varepsilon^{-1} \right\} \\ &\quad \exp\left\{ -\frac{p(2r+d)}{(2s_*+d)(2q+d)}\, \log\log \varepsilon^{-1} \right\} \left(\frac{c_{\nu'}}{c_{\nu''}} \right)^p. \end{aligned}$$

Choose $\delta = (1-\delta_1)\frac{s^*-q}{s^*-s_*} + \delta_1$, where $0 < \delta_1 < 1/2$, and consider only q that is close enough to s^*, so that $\delta < 1/2$. Then $q - s^* + \delta(s^* - s_*) > 0$, and $\tau Q^p \to \infty$, as $\varepsilon \to 0$. Since also $l_\varepsilon \to -\infty$, we conclude that the RHS of (58) tends to $(1-2\delta)^p$, as $\varepsilon \to 0$. Taking the limit of $(1-2\delta)^p$, as $q \to s^*$, and using the fact that δ_1 can be chosen arbitrarily small, we get

$$\lim_{q\to s^*} \liminf_{\varepsilon\to 0} \inf_{T_\varepsilon} \sup_{\nu\in B_q} \mathcal{R}_{\varepsilon,\nu}(T_\varepsilon, \psi_\nu) \geq 1.$$

□

Proof of Theorem 3. We start with the proof of Theorem 3 (i). Note that the function

$$F(s) \stackrel{\text{def}}{=} \left(\max\{\lambda(s), \lambda(\frac{s^*+q}{2})\}\varepsilon^2 \log \varepsilon^{-1}\right)^{1/(2s+d)}$$

is a continuous function of s on the interval $[r', s^*]$, and $F(r') = h(r', \tilde{\varepsilon}(r')) \leq h_i < h_{\max} = \varepsilon^{2/(2s^*+d)} < F(s^*)$ for any $i \in \{1, \ldots, m\}$ and ε small enough. Hence for every $i \in \{1, \ldots, m\}$ there exists at least one $s_i \in [r', s^*]$ such that

$$h_i = \left(\max\{\lambda(s_i), \lambda(\frac{s^*+q}{2})\}\varepsilon^2 \log \varepsilon^{-1}\right)^{1/(2s_i+d)}. \tag{59}$$

Fix a sequence $S = \{s_1, s_2, \ldots, s_m\}$ where s_i is a solution of (59). Using (18), (19) it is easy to check that s_i defined by (59) satisfies (10) with $\gamma_1 = \gamma = 1 + \gamma_0$, and

$$h_i = h(s_i, \tilde{\varepsilon}(s_i)), \quad \forall s_i \leq (s^*+q)/2.$$

Therefore, we can apply the argument as in the proof of Theorem 1 for this particular grid S. Some modifications of the proof are needed here since, unlike the case of Theorem 1, the kernel and the threshold are defined with $s(h_i)$ in place of s_i. These modifications are easy to establish if one notes that

$$|s_i - s(h_i)| \leq \frac{C \log\log(1/\varepsilon)}{\sqrt{\log(1/\varepsilon)}}, \quad \forall s_i \leq (s^*+q)/2. \tag{60}$$

In fact, Assumption 3 and (60) entail

$$\left|\frac{\eta_{h_i}}{\eta(s_i)} - 1\right| \leq \gamma_{\varepsilon 8}, \quad \forall s_i \leq (s^*+q)/2, \tag{61}$$

and, in view of (60) and Assumption 4,

$$b_{s,s(h_i)} \le b_{s,s_i}(1+\gamma_{\varepsilon 9}) \le b_{\max}(r', s^*)(1+\gamma_{\varepsilon 9}), \quad \forall\ s, s_i \in [r', q],\ s_i \le s. \qquad (62)$$

Using (60) - (62), the proof of (36) is almost the same as in Theorem 1. For the proof of (37) we mention only the modifications in the key relations (50) and (54). Instead of (50) we now obtain (with the notation $s' = s_i$):

$$|E_f T_\varepsilon(h_i) - T(f)| \le L h_i^{s-r} b_{s,s(h_i)} = L h^{s-r}(s_i, \tilde{\varepsilon}(s_i)) b_{s,s(h_i)}$$

where, by (51), $|s_i - s| \le C \log\log\log(1/\varepsilon)/\log(1/\varepsilon)$. This and (60), together with Assumption 4 (ii), yield $b_{s,s(h_i)} \le b_{s,s}(1+\gamma_{\varepsilon 10})$. Other elements of (50) remain as in the proof of Theorem 1. Turning to (54), we have to evaluate now η_{h^+} in place of $\eta(s^+)$, where $h^+ = h(s^+, \tilde{\varepsilon}(s^+))$. By virtue of (61), the only difference from the case of Theorem 1 appears in the inclusion of the extra factor $(1+\gamma_{\varepsilon 8})$.

Consider now the proof of Theorem 3 (ii). We have $F(r') = h(r', \tilde{\varepsilon}(r')) \le h_i < h_{\max} = 1 = \lim_{s\to\infty} F(s)$. Hence, the solutions s_i exist, as above, but we get an additional set of gridpoints that extends to the right beyond s^*:

$$\mathcal{S}_3 = \{s_i : s_i > (s^* + q)/2\}\,.$$

We can apply the argument as in the proof of Theorem 3 (i) with a modification as to address the set S_3 and the choice $h_{\max} = 1$. The latter is equivalent to putting $s^* = \infty$, and all the calculations in the proof of Theorems 1 and Theorem 3 (i) remain valid with this modification if ψ_ν is replaced by $\tilde{\psi}_\nu$. Inclusion of the set $\mathcal{S}_3$ leads to a modification only in the proof of (37). In fact,

$$R^+_{\varepsilon,\nu} = \sup_{f\in\mathcal{F}_\nu} E_f\left(\psi_\nu^{-p}\,|T^*_\varepsilon - T(f)|^p\, I\left(\hat{s} \in \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3\right)\right)$$

and the inclusion of $\mathcal{S}_3$ results in the consideration of the third component of $R^+_{\varepsilon,\nu}$, namely

$$R^{++}_{\varepsilon,\nu} = \sup_{f\in\mathcal{F}_\nu} E_f\left(\psi_\nu^{-p}\,|T^*_\varepsilon - T(f)|^p\, I\left(\hat{s} \in \mathcal{S}_3\right)\right).$$

We treat this component similarily to (56) (we have the same expression with $\mathcal{S}_3$ instead of $\mathcal{S}_2$). Hence

$$R^{++}_{\varepsilon,\nu} \le \sup_{f\in\mathcal{F}_\nu} \sum_{s'\in\mathcal{S}_3} E_f\left(\left(1+\gamma_{\varepsilon 8} + \psi_\nu^{-1}\,|Z_{s^+}|\right)^p I\left(\hat{s} = s'\right)\right),$$

and acting as in the calculation following the formula (56), we get instead of (57)

$$R^+_{\varepsilon,\nu} \le \left(1+\gamma_{\varepsilon 7}+\frac{C_4^{1/2}}{\log^{1/8}(1/\varepsilon)}\right)^p \sup_{f\in\mathcal{F}_\nu} P_f\left(\hat{s}\in\mathcal{S}_1\cup\mathcal{S}_2\cup\mathcal{S}_3\right)+2C_6 m\exp\left\{-\frac{\log^{1/4}(1/\varepsilon)}{2C_4}\right\}$$

and we conclude the proof by noting that $m = \mathrm{Card}(S) \le C\left(\log(1/\varepsilon)\right)^{1+\gamma_0}$. □

Proof of Proposition 1. Note that

$$\begin{aligned} \|\tilde{K}_s\|_2^2 &= (2\pi)^d\|\hat{\tilde{K}}_s\|_2^2 = (2\pi)^{-d}\int_{\mathbf{R}^d}\frac{\omega^{2\alpha_0}}{(1+\|\omega\|^{2\beta})^2}\,d\omega \\ &= \frac{1}{2\beta}B\left(\frac{2r+d}{2\beta},2-\frac{2r+d}{2\beta}\right)(2\pi)^{-d}\int_{S_d}\xi^{2\alpha_0}d\mu(\xi), \end{aligned}$$

and

$$\begin{aligned} \rho_s^2(\tilde{K}_s) &= (2\pi)^d\int_{\mathbf{R}^d}\|\omega\|^{2\beta}\left|\hat{\tilde{K}}_s(\omega)\right|^2 d\omega = (2\pi)^{-d}\int_{\mathbf{R}^d}\frac{\omega^{2\alpha_0}\|\omega\|^{2\beta}}{(1+\|\omega\|^{2\beta})^2}\,d\omega \\ &= \frac{1}{2\beta}B\left(1+\frac{2r+d}{2\beta},1-\frac{2r+d}{2\beta}\right)(2\pi)^{-d}\int_{S_d}\xi^{2\alpha_0}d\mu(\xi) = C_*^2. \end{aligned}$$

Since the beta-function satisfies $B(c,d) = B(c-1,d+1)(c-1)/d$, $\forall\, c>1, d>0$, it follows that

$$\|\tilde{K}_s\|_2^2 = \frac{2(s-r)}{2r+d}\,\rho_s^2(\tilde{K}_s),$$

and thus

$$b = \left(\frac{\|\tilde{K}_s\|_2}{\rho_s(\tilde{K}_s)}\right)^{2/(2s+d)}. \tag{63}$$

Now,

$$(-1)^r T(\tilde{K}_s) = (-i)^r\int\omega^{\alpha_0}\hat{\tilde{K}}_s(\omega)d\omega = (2\pi)^{-d}\int_{\mathbf{R}^d}\frac{\omega^{2\alpha_0}}{1+\|\omega\|^{2\beta}}\,d\omega = \|\tilde{K}_s\|_2^2+\rho_s^2(\tilde{K}_s)$$

and hence, using (63), we get

$$\begin{aligned} T(g_{s,1,1}(\cdot)) &= (-1)^r aT(\tilde{K}_s(b\,\cdot)) = (-1)^r ab^r T(\tilde{K}_s) \\ &= \left(\frac{\rho_s(\tilde{K}_s)}{\|\tilde{K}_s\|_2}\right)^{2(s-r)/(2s+d)}\rho_s(\tilde{K}_s)^{-1}\left(\|\tilde{K}_s\|_2^2+\rho_s^2(\tilde{K}_s)\right) \\ &= C_*\left(\frac{2r+d}{2(s-r)}\right)^{(s-r)/(2s+d)}\frac{2s+d}{2r+d}. \end{aligned} \tag{64}$$

Now,

$$\|K_s\|_2 = b^{r+d/2}\|\tilde{K}_s\|_2. \tag{65}$$

To evaluate the bias constant we use the following lemma.

Lemma 9 *Let the assumptions of Proposition 1 be satisfied, and let* $r < r' < s^*$, $r' \le s, s' \le s^*$, $\beta = s + d/2,\ \beta' = s' + d/2,\ b' = \left(\frac{2(s'-r)}{2r+d}\right)^{1/(2s'+d)}$. *Then*

$$b_{s,s'} \le (b')^{r-s} \left[(2\pi)^{-d} \int_{\mathbf{R}^d} \frac{\omega^{2\alpha_0} \|\omega\|^{4\beta'-2\beta}}{(1+\|\omega\|^{2\beta'})^2} d\omega \right]^{1/2}$$

provided the last integral is finite.

Proof. We have $\hat{K}_{s'}(\omega) = (2\pi)^{-d} i^r \omega^{\alpha_0} (1 + \|\omega/b'\|^{2\beta'})^{-1}$. Hence, by the Cauchy inequality,

$$\begin{aligned} b_{s,s'} &= \sup_{\rho_s(f)\le 1} \left| \int K_{s'} f - T(f) \right| = \sup_{\rho_s(f)\le 1} \left| \int_{\mathbf{R}^d} \hat{f}(\omega) \left((2\pi)^d \hat{K}_{s'}(\omega) - i^r \omega^{\alpha_0} \right) d\omega \right| \\ &= \sup_{\rho_s(f)\le 1} \left| \int_{\mathbf{R}^d} \widehat{f}(\omega) i^r \omega^{\alpha_0} \left(\frac{\|\omega/b'\|^{2\beta'}}{1+\|\omega/b'\|^{2\beta'}} \right) d\omega \right| \\ &\le (b')^{r-s} \sup_{\rho_s(f)\le 1} \left[(2\pi)^d \int_{\mathbf{R}^d} \|\omega\|^{2\beta} \left| \hat{f}(\omega) \right|^2 d\omega \right]^{1/2} \left[(2\pi)^{-d} \int_{\mathbf{R}^d} \frac{\omega^{2\alpha_0} \|\omega\|^{4\beta'-2\beta}}{(1+\|\omega\|^{2\beta'})^2} d\omega \right]^{1/2}. \end{aligned}$$

Using the definition of ρ_s for the Sobolev classes, we get the lemma. □

It follows from Lemma 9 with $s = s'$ that $b_{s,s} \le \rho_s(\tilde{K}_s) b^{r-s}$. Moreover,

$$b_{s,s} = \rho_s(\tilde{K}_s) b^{r-s}. \tag{66}$$

Indeed, the function f_* with the Fourier transform

$$\widehat{f}_*(\omega) = (2\pi)^{-d} \rho_s(\tilde{K}_s)^{-1} b^{-r-s-d} \frac{i^r \omega^{\alpha_0}}{1+\|\omega/b\|^{2\beta}}$$

satisfies $\rho_s(f_*) = 1$ and

$$\left| \int K_s f_* - T(f_*) \right| = \left| \int_{\mathbf{R}^d} \widehat{f}_*(\omega) i^r \omega^{\alpha_0} \left(\frac{\|\omega/b\|^{2\beta}}{1+\|\omega/b\|^{2\beta}} \right) d\omega \right| = \rho_s(\tilde{K}_s) b^{r-s}.$$

Combining (64) - (66), and using (63), we get (6). This proves the proposition. □

Proof of Proposition 2. Assumptions 1 - 3 are straightforward to verify. We check only Assumption 4. For $s, s' \in [r', s^*]$ and $\beta = s + d/2,\ \beta' = s' + d/2$ using the inequality $r' > (s^* - r - d)/2$ assumed in Proposition 2, we obtain $\beta' \ge r' + d/2 > (s^* - r)/2 \ge (s - r)/2 = (\beta - r)/2 - d/4$. Thus

$$\int_{\mathbf{R}^d} \frac{\omega^{2\alpha_0} \|\omega\|^{4\beta'-2\beta}}{(1+\|\omega\|^{2\beta'})^2} d\omega \le \int_{\mathbf{R}^d} \frac{\|\omega\|^{4\beta'-2\beta+2r}}{(1+\|\omega\|^{2\beta'})^2} d\omega = \mu(S_d) \int_0^\infty \frac{t^{4\beta'-2\beta+2r+d-1}}{(1+t^{2\beta'})^2} dt$$

$$\leq \mu(S_d)\left(\int_0^1 t^{4\beta'-2\beta+2r+d-1}dt + \int_1^\infty t^{d-1-2\beta+2r}dt\right)$$
$$\leq \mu(S_d)\left[\frac{1}{4r'-2(s^*-r-d)} + \frac{1}{2(r'-r)}\right]$$

where for the last inequality we used that $\beta, \beta' > r' + d/2$. This and Lemma 9 yield Assumption 4(i). Next, Assumption 4(ii) follows from (66), the continuity of the beta-function and the fact that (by Lemma 9)

$$b_{s,s'} \leq (b')^{r-s}\left[\frac{1}{2\beta'}B\left(1+\frac{2r+d+2(\beta'-\beta)}{2\beta'}, 1-\frac{2r+d+2(\beta'-\beta)}{2\beta'}\right)\right]^{1/2}$$
$$\times\left[(2\pi)^{-d}\int_{S_d}\xi^{2\alpha_0}d\mu(\xi)\right]^{1/2}.$$

□

Proof of Propositions 3 and 4. Note that, for $\tilde{K}_s$ defined in (23),

$$\int \|x\|^s \tilde{K}_s(x)dx + \|\tilde{K}_s\|_2^2 = \int \tilde{K}_s.$$

Now, $g_{s,1,1}(\cdot) = a\tilde{K}_s(b\,\cdot)$, where $b = \|\tilde{K}_s\|_2^{2/(2s+d)}$. Let

$$K_s(x) = \left[\int g_{s,1,1}\right]^{-1} g_{s,1,1}(x) = \left[\int \tilde{K}_s\right]^{-1} b^d \tilde{K}_s(bx).$$

Then

$$\begin{aligned}\int \|x\|^s K_s(x)dx + \|K_s\|_2 &= \left(b^{-s}\int \|x\|^s\tilde{K}_s(x)dx + b^{d/2}\|\tilde{K}_s\|_2\right)/\int \tilde{K}_s \\ &= \|\tilde{K}_s\|_2^{-2s/(2s+d)}\left(\int \|x\|^s\tilde{K}_s(x)dx + \|\tilde{K}_s\|_2^2\right)/\int\tilde{K}_s \\ &= \|\tilde{K}_s\|_2^{-2s/(2s+d)}.\end{aligned}$$

For $f \in \{g : \|g\|_2 \leq 1,\ \rho_s(g) \leq 1\}$ we get

$$\begin{aligned}|f(0)| &\leq \left|\int K_s f - f(0)\right| + \left|\int K_s f\right| \leq \left|\int (f(x)-f(0))K_s(x)dx\right| + \|f\|_2\|K_s\|_2 \\ &\leq \int \|x\|^s K_s(x)dx + \|K_s\|_2 = \|\tilde{K}_s\|_2^{-2s/(2s+d)}.\end{aligned}$$

On the other hand, this upper bound is achieved by $g_{s,1,1}$ because

$$g_{s,1,1}(0) = a\tilde{K}_s(0) = a = \|\tilde{K}_s\|_2^{-2s/(2s+d)}.$$

Propositions 3 and 4 follow from these remarks and the equations

$$\int \tilde{K}_s = \frac{\mu(S_d)s}{(s+d)d}, \quad \|\tilde{K}_s\|_2^2 = \frac{2s^2\mu(S_d)}{(2s+d)(s+d)d} \ .$$

□

Proof of Proposition 5. Assumptions 1 - 3 are straightforward. To check Assumption 4, it suffices to remark that in this case $b_{s,s'} \leq \int \|x\|^s K_{s'}(x)dx$ and $b_{s,s} = \int \|x\|^s K_s(x)dx$. □

References

[1] Arestov, V. V. (1989). Optimal recovery of operators and related problems. *Proceedings of Steklov Mathematical Institute* **189** 3-20.

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

[3] Brown, L. and Low, M. (1996a). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384-2398.

[4] Brown, L. and Low, M. (1996b). A constrained risk inequality with applications to nonparametric estimation. *Ann. Statist.* **24** 2524-2535.

[5] Butucea, C. (2000). Exact adaptive pointwise estimation on Sobolev classes of densities. Preprint n.621, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6, Paris 7 (http://www.proba.jussieu.fr).

[6] Butucea, C. (2001). Numerical results concerning a sharp adaptive density estimator. *Computational Statistics* (to appear).

[7] Cavalier, L. and Tsybakov, A.B. (2000) Sharp adaptation for inverse problems with random noise. Preprint N.559, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 - Paris 7. http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2000. *Probability Theory and Related Fields* (to appear).

[8] Donoho, D. (1994a). Asymptotic minimax risk for sup-norm loss: solution via optimal recovery. *Probab. Theory and Related Fields* **99** 145-170.

[9] Donoho, D. (1994b). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238-270.

[10] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *J. Royal Statist. Soc. Ser. B* **57** 301-369.

[11] Donoho, D. and Liu, R. (1991). Geometrizing rates of convergence, III. *Ann. Statist.* **19** 668-701.

[12] Donoho, D. and Low, M. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20** 944-970.

[13] Efroimovich, S. Yu. and Pinsker, M. S. (1984). Learning algorithm for nonparametric filtering. *Automation and Remote Control* **11** 1434-1440.

[14] Erfomovich, S. (2000) Sharp adaptive estimation of multivariate curves. *Math. Meth. Statist.* **9** 117-139.

[15] Efromovich, S. Y. and Low, M. (1994). Adaptive estimates of linear functionals. *Probab. Theory and Related Fields* **98** 261-275.

[16] Fuller, A. T. (1960). Optimization of relay systems via various performance indices. In: *Proc. 1st International IFAC Congress, Moscow* v. 2, 584-607.

[17] Gabushin, V. N. (1968). Exact constants in inequalities between norms of the derivatives of a function. *Math. Notes* **4** 221-232.

[18] Gabushin, V. N. (1970). Best approximation of functionals on some sets. *Math. Notes* **8** 551-562.

[19] Goldenshluger, A. and Nemirovskii, A. (1997). On spatial adaptive estimation of nonparametric regression. *Math. Meth. Statist.* **6** 135-170.

[20] Golubev, G. K. and Nussbaum, M. (1990). Adaptive spline estimates for nonparametric regression models. *Theory Probab. Appl.* **37** 521-529.

[21] Ibragimov, I. A. and Hasminskii, R. Z. (1981). *Statistical Estimation. Asymptotic Theory*, Springer-Verlag. (Originally published in Russian in 1979).

[22] Ibragimov, I. A. and Hasminskii, R. Z. (1984). On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.* **29** 18-32.

[23] Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, Approximation and Statistical Applications.* Lecture Notes in Statistics, vol. 129. Springer, N.Y. e.a.

[24] Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics* **11** 337-381.

[25] Korostelev, A. P. (1993). Asymptotically minimax regression estimator in the uniform norm up to exact constant. *Theory Probab. Appl.* **38** 737-743.

[26] Legostaeva, I. L. and Shiryayev, A. N. (1971). Minimax weigths in a trend detection problem of a random process. *Theory Prob. Appl.* **16** 344-349.

[27] Leonov, S. L. (1997). On the solution of an optimal recovery problem and its applications in nonparametric regression. *Math. Meth. Statist.* **6** 476-490.

[28] Leonov, S. L. (1999). Remarks on extremal problems in nonparametric curve estimation. *Statist. Prob. Letters.* **43** 169-178.

[29] Lepski, O. V. (1990). One problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454-466.

[30] Lepski, O. V. (1992a). Asymptotically minimax adaptive estimation II: Statistical models without optimal adaptation. Adaptive estimators. *Theory Probab. Appl.* **37** 433-468.

[31] Lepski, O. V. (1992b). On problems of adaptive estimation in white Gaussian noise. In *Topics in Nonparametric Estimation*. Advances in Soviet Math., vol. 12 (Khasminskii R. Z., ed.), p. 87-106, Amer. Math. Soc., Providence, R.I.

[32] Lepski, O. V. and Levit, B. Ya. (1998). Adaptive minimax estimation of infinitely differentiable functions. *Math. Meth. Statist.* **7** 123-156.

[33] Lepski, O. V. and Levit, B. Ya. (1999). Adaptive nonparametric estimation of smooth multivariate functions. *Math. Meth. Statist.* **8** 344-370.

[34] Lepski, O. V., Mammen, E., and Spokoiny, V. G. (1997). Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection. *Ann. Statist.* **25** 929-947.

[35] Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25** 2512-2546.

[36] Lepski, O. V. and Tsybakov, A. B. (2000). Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. *Probab. Theory and Related Fields*, **117**, 17-48.

[37] Micchelli, C. A. and Rivlin, T. J. (1977). A survey of optimal recovery. In: *Optimal Estimation in Approximation Theory*, p.1-54, Plenum Press, N.Y. e.a.

[38] Nemirovski, A. (2000). Topics in Non-Parametric Statistics. In *Lectures on Probability and Statistics, Ecole d'Été de Probabilités de Saint-Flour XXVIII - 1998.* Lecture Notes in Mathematics v.1738 (P.Bernard, ed.), p.85-277, Springer, Berlin e.a.

[39] Nussbaum, M. (1996). Asymptotic equivalence of density estimation and gaussian white noise. *Ann. Statist.* **24** 2399-2430.

[40] Sacks, J. and Ylvisaker, N. D. (1981). Asymptotically optimum kernels for density estimation at a point. *Ann. Statist.* **9** 334-346.

[41] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann.Statist.* **8** 1348-1360.

[42] Taikov, L. V. (1968). Kolmogorov type inequalities and the best formulas of numerical differentiation. *Math. Notes* **4** 233-238.

[43] Tsybakov, A. B. (1998). Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.* **26** 2420-2469.

[44] Zhao, L.H. (1997). Minimax linear estimation in a white noise problem. *Ann. Statist.* **25** 745-755.

Jussi Klemelä
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg, Germany

Alexandre Tsybakov
Laboratoire de Probabilités et Modèles Aléatoires
UMR CNRS 7599
Université Paris 6
4, pl.Jussieu, BP 188, 75252 Paris, France
email: tsybakov@ccr.jussieu.fr