Optimal Recovery and Statistical Estimation in L_p Sobolev Classes

Jussi Klemelä*
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294, 69120 Heidelberg, Germany
Email: klemela@statlab.uni-heidelberg.de
Fax +49 6221 545331

Abstract

We present algorithms for the optimal recovery of a partial derivative of a function at a point when we have approximate measurements of the Riesz transform of the function, and this function belongs to a L_p Sobolev class. The algorithms are exactly optimal among linear algorithms for the cases $p = 1$ and $p = 2$ and we give tight bounds for the performance of the algorithms when $p > 1$, $p \neq 2$. Previously only the case $p = 2$ has been studied. Algorithms for the optimal recovery problem provide optimal estimators for several statistical problems, when we calibrate algorithms suitably. As examples we construct a nearly minimax estimator in the Gaussian white noise model and an asymptotically nearly minimax estimator for the problem of regression function estimation with i.i.d. data. We give also bounds for the asymptotic adaptive risk in the Gaussian white noise model.

Mathematics Subject Classifications: 62G07
Key Words: Exact constants in nonparametric smoothing, linear functionals, minimax risk, modulus of continuity, optimal kernels, Riesz transform.
Short title: Optimal Recovery and Statistical Estimation

*Writing of this article was financed by Deutsche Forschungsgemeinschaft under project MA1026/8-1.

1 Introduction

By the optimal recovery of an unknown function at a point we mean the estimation of the value of the function at this point when we have available a function from which we know that it is close to the unknown function in L_2 metric. More generally, we may also consider the inverse problem where we have available a transformation of the function and we know that this transformed function is close to the transformation of the original function in L_2 metric. For expositions of optimal recovery, see for example Micchelli and Rivlin (1977) and Arestov (1989).

Optimal recovery is connected to many statistical problems. We give in the following a list of statistical problems where the construction of an optimal or a nearly optimal statistical procedure can be reduced to the optimal recovery. Gaussian white noise model was studied in the most cases but some references address regression or density estimation models.

1. **Estimation of linear functionals.** Donoho and Liu (1991) and Donoho (1994*b*) show that minimax linear estimators of linear functionals may be obtained by a calibration of optimal recovery algorithms, and these linear estimators are nearly minimax optimal among all estimators. In particular, estimators constructed by Ibragimov and Hasminskii (1984) are a special case of such estimators. Donoho and Low (1992) show that the minimax rate of convergence of estimating linear functionals is determined by the modulus of continuity.

2. **Adaptive estimation of linear functionals.** Klemelä and Tsybakov (2001) construct an asymptotically sharp adaptive estimator by comparing linear estimators with different scales (Lepski method). The linear estimators were calibrations of optimal recovery algorithms. The adaptive estimators constructed in Lepski and Spokoiny (1997) and Tsybakov (1998) are either a special case of, or closely related to, such sharp adaptive estimators. Klemelä and Tsybakov (2004) construct a sharp adaptive estimation procedure when the Riesz transform is observed. Efromovich and Low (1994) show that adaptive rates of convergence may be expressed with the help of modulus of continuity.

3. **Estimation of a function with the supremum norm loss.** Donoho (1994*a*) shows that the asymptotically optimal minimax estimator (among all estimators) is a calibration of an optimal recovery algorithm when the sup-norm loss function is applied, over Hölder classes. The estimator of Korostelev (1993) is a special case of such an estimator. Tsybakov

(1998) considers L_2 Sobolev classes, and adaptive estimation, with the supremum norm loss.

4. **Large deviation optimality.** Korostelev (1996) and Puhalskii and Spokoiny (1998) show how minimax estimators in large deviation loss may be constructed from optimal recovery algorithms.

5. **Hypothesis testing.** Lepski and Tsybakov (2000) show how hypothesis testing problems in sup-norm and at a fixed point are connected to optimal recovery.

6. **Estimating the whole object in a sequence space.** Donoho, Johnstone, Kerkyacharian and Picard (1995) consider optimal recovery of a function in a sequence space model. They derive optimal rates of convergence for estimating a function under Besov or Triebel smoothness conditions, for Besov or Triebel loss functions.

7. **Estimation of quadratic functionals.** Donoho and Nussbaum (1990) show that minimax optimal quadratic estimators of quadratic functionals may be found by solving a related optimal recovery problem.

8. **Adaptive estimation of quadratic functionals.** Klemelä (2006) shows that a sharp adaptive estimator for certain quadratic functionals may be constructed by comparing minimax optimal quadratic estimators at different scales. Efromovich and Low (1996) apply the modulus of continuity to find the optimal adaptive rates of convergence.

In this article we consider approximation of partial derivatives of a function at a point. We assume that we have available a Riesz transform of a function which is close to the true function in L_2 metric. Based on this Riesz transformed function we construct a linear estimator for the unknown value of the partial derivative. We will assume that the unknown function satisfies a L_p Sobolev smoothness condition, in the sense of a moment condition for the Fourier transform of the function.

The results of this article are directly relevant to the statistical problems 1-5 in the above list. As examples we consider the item 1 and item 2 in the above list. We give bounds for the minimax risk both in the Gaussian white noise model and a regression estimation mode, and give bounds for the asymptotic adaptive risk in the Gaussian white noise model.

Previously only the case $p = 2$ of the Sobolev classes has been studied in the optimal recovery. Taikov (1969) gave an exact solution when $p = 2$, and Donoho and Low (1992), Klemelä and Tsybakov (2001), Klemelä and Tsybakov (2004) studied the problem in various degrees of generality but

only for the case $p = 2$ (see Remark 2 for a more precise description of the previous results).

We consider the cases $p \geq 1$ and give an exact solution for the cases $p = 1$ and $p = 2$, and give tight bounds for the cases $p > 1$, $p \neq 2$. For the case $p = 1$ we construct a kernel for the estimator which is of "Pinsker type", that is, its Fourier transform is of "Bartlett type". For the cases $p > 1$ we construct kernels which are of "Tikhonov type".

We show that the approximation in Sobolev L_p spaces is more feasible for low values of p than for high values of p, in high dimensional cases. Indeed, the ratio of the optimal approximation error correponding to $p = 2$, to the optimal approximation error correponding to $p = 1$, is increasing exponentially, as the dimension of the approximated function increases.

Typically, solutions to the optimal recovery are difficult to find. For Hölder spaces, for the estimation of the value of the function at one point, the solution has been found for smoothness indices $0 < s \leq 1$ and $s = 2$, see Fuller (1982), Gabushin (1968), Korostelev (1993), Zhao (1997), Leonov (1997), Leonov (1999). For Taylor spaces the solutions are given for smoothness indices $0 < s \leq 2$ in Klemelä and Tsybakov (2001). See also Legostaeva and Shiryayev (1971). For the case when we consider Sobolev classes of functions whose derivative of a given order has a bounded L_p norm (instead of posing moment conditions for the Fourier transform) Sz.-Nagy (1941) has given solutions (for $p > 2$), Gabushin (1967) has given the optimal rates of convergence, and Magaril-Il'yaev (1983) has characterized the optimal constants. In the connection of the estimation of quadratic functionals the exact solution for l_2-body is given by Donoho and Nussbaum (1990), and the exact solution for l_4-body and approximate solutions for l_p-bodies, $p > 2$, are given by Klemelä (2006).

We formulate the results in Section 2: Section 2.1 specifies the setting, Section 2.2 considers the case $p > 1$, Section 2.3 considers the case $p = 1$, Section 2.4 discusses the effect which parameter p and the dimension of the estimated function have on the optimal approximation error, and Section 2.5 gives three examples of statistical applications. Proofs are given in Section 3. Discussion of the results is given in Section 4.

2 Results

2.1 The setting

The functional. We study the estimation of the derivative $f^{(\alpha_0)}(0)$, where α_0 is a multi-index. For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d)$ and for a point

$\omega = (\omega_1, \ldots, \omega_d) \in \mathbf{R}^d$ we denote $|\alpha| = \alpha_1 + \cdots + \alpha_d$ and $\omega^\alpha = \omega_1^{\alpha_1} \cdots \omega_d^{\alpha_d}$. We assume that $|\alpha_0| = r$, where $r \geq 0$ is an integer. We may write

$$f^{(\alpha_0)}(x) = i^{|\alpha_0|} \int_{\mathbf{R}^d} \omega^{\alpha_0} \widehat{f}(\omega) \exp(ix^T\omega)\, d\omega$$

where $\widehat{f}(\omega)$ denotes the Fourier transform of f,

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

and i denotes the imaginary unit.

Smoothness condition. We assume that the function satisfies a Sobolev smoothness condition. Define Sobolev semi-norm by

$$\rho_{\beta,p}^p(f) = (2\pi)^d \int_{\mathbf{R}^d} \|\omega\|^q \left|\widehat{f}(\omega)\right|^p d\omega \tag{1}$$

where

$$q = p[\beta + d(1/2 - 1/p)], \tag{2}$$

$p \geq 1$. We discuss the smoothness condition more in detail in Section 2.4.1.

Riesz transform. We assume that we observe the Riesz transformation of the function. The operator $R_\gamma f$ is called the Riesz transform when for a function $f \in L_1(\mathbf{R}^d)$ and for $0 \leq \gamma < d$

$$(R_\gamma f)(x) = \begin{cases} \alpha_\gamma \int_{\mathbf{R}^d} f(y)\|x - y\|^{\gamma - d} dy & \text{if } 0 < \gamma < d, \\ f(x) & \text{if } \gamma = 0, \end{cases}$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbf{R}^d$ and

$$\alpha_\gamma = (2\pi)^{-d/2} \pi^{\gamma - d/2} \frac{\Gamma((d-\gamma)/2)}{\Gamma(\gamma/2)},$$

where Γ is gamma function, see Stein (1970). We have, by the Fourier convolution formula,

$$(R_\gamma f)^\wedge(\omega) = \widehat{f}(\omega)\|\omega\|^{-\gamma}, \ \ \omega \in \mathbf{R}^d, \tag{3}$$

see Stein (1970), page 73.

Optimal recovery and the modulus of continuity. We will restrict ourselves to linear estimators and thus the optimal recovery problem is to find a kernel $K \in L_2$ which is close to the infimum in the minimax risk

$$R_\epsilon = \inf_{K \in L_2} \sup_{\{f:\rho_{\beta,p}(f) \le L\}} \sup_{\{g:\|R_\gamma(f-g)\|_2 \le \epsilon\}} \left| \int_{\mathbf{R}^d} K(R_\gamma g) - f^{(\alpha_0)}(0) \right|. \qquad (4)$$

That is, we try to estimate unknown value $f^{(\alpha_0)}(0)$ with the linear estimator (or linear algorithm) $\int_{\mathbf{R}^d} K(R_\gamma g)$, when we have available the Riesz transform $R_\gamma g$, which is close to the Riesz transform $R_\gamma f$ of the unknown function in L_2 metric. Note that R_γ is linear. We assume apriori knowledge that $\rho_{\beta,p}(f) \le L$. Define the modulus of continuity by

$$\omega(\epsilon) = \sup \left\{ |f^{(\alpha_0)}(0)| : \|R_\gamma f\|_2 \le \epsilon, \;\; \rho_{\beta,p}(f) \le L \right\}. \qquad (5)$$

Micchelli and Rivlin (1977), Theorem 6, show that under general assumptions, which are satisfied in our setting, the minimax risk is equal to the modulus of continuity:

$$R_\epsilon = \omega(\epsilon).$$

2.2 Solutions to the optimal recovery when $p > 1$

We give an upper and lower bound for the minimax risk (4) and we construct a kernel which achieves the upper bound. We consider the cases $p > 1$.

Constant for the upper bound. Denote

$$K_u(p) = \int_{S_d} |\xi^{\alpha_0}|^{p/(p-1)} d\mu(\xi), \qquad (6)$$

where $S_d = \{x \in \mathbf{R}^d : \|x\| = 1\}$ for $d = 2, 3, \ldots$, $S_1 = [-1, 1]$, μ is the Lebesgue measure on S_d so that

$$\mu(S_d) = 2\pi^{d/2}/\Gamma(d/2), \qquad (7)$$

$d = 1, 2, \ldots$, and Γ denotes the gamma-function. Denote with $B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt$, $a, b > 0$, the Beta function. Denote

$$\begin{aligned} I_1 \;\; = \;\; & \frac{(2\pi)^{d(1-p/(p-1))} K_u(p)}{q + 2\gamma} \\ & B\left(\frac{p(q-r)/(p-1) - q - d}{q + 2\gamma}, \frac{q + p(2\gamma + r)/(p-1) + d}{q + 2\gamma} \right) \end{aligned} \qquad (8)$$

and

$$I_2 = \frac{(2\pi)^{-d} K_u(2)}{q+2\gamma} B\left(\frac{2(q+\gamma-r)-d}{q+2\gamma}, \frac{2(\gamma+r)+d}{q+2\gamma}\right), \tag{9}$$

where q is defined in (2). Define the constant for the upper bound as

$$\begin{aligned} C_{\beta,p} &= \left(I_2^{1/2}\right)^{(\beta-r-d/2)/(\beta+\gamma)} \left(I_1^{(p-1)/p}\right)^{(\gamma+r+d/2)/(\beta+\gamma)} \\ &\quad \times \left(\frac{\gamma+r+d/2}{\beta-r-d/2}\right)^{(\beta-r-d/2)/(\beta+\gamma)} \frac{\beta+\gamma}{\gamma+r+d/2}. \end{aligned} \tag{10}$$

This constant is finite when

$$\beta > \max\{\iota_1(p,r,d), \iota_2(p,r,d,\gamma)\} \tag{11}$$

where $\iota_1(p,r,d) = r/p + d/2$, $\iota_2(p,r,d,\gamma) = (r-\gamma)/p + d(3/(2p) - 1/2)$.

Constant for the lower bound. Let

$$K_l(p) = \int_{S_d} |\xi^{\alpha_0}|^p d\mu(\xi), \tag{12}$$

where S_d and μ are as in the definition of K_u in (6). Define

$$I_3 = \frac{(2\pi)^{d(1-p/(p-1))} K_l(p)}{q+2\gamma} \tag{13}$$

$$\times B\left(\frac{p(q+\gamma)/(p-1) - q - p(\gamma+r) - d}{q+2\gamma}, \frac{q + p\gamma/(p-1) + p(\gamma+r) + d}{q+2\gamma}\right),$$

$$\begin{aligned} I_4 &= \frac{(2\pi)^{d(1-2/(p-1))} K_l(2)}{q+2\gamma} \\ &\quad B\left(\frac{2(q+\gamma)/(p-1) - 2r - d}{q+2\gamma}, \frac{2\gamma/(p-1) + 2r + d}{q+2\gamma}\right), \end{aligned} \tag{14}$$

and

$$\begin{aligned} I_5 &= \frac{(2\pi)^{-d/(p-1)} K_l(2)}{q+2\gamma} \\ &\quad B\left(\frac{[q+\gamma(2-p)]/(p-1) - 2r - d}{q+2\gamma}, \frac{p\gamma/(p-1) + 2r + d}{q+2\gamma}\right). \end{aligned} \tag{15}$$

Define the constant for the lower bound as

$$c_{\beta,p} = \left(I_4^{-1/2}\right)^{(\beta-r-d/2)/(\beta+\gamma)} \left(I_3^{-1/p}\right)^{(\gamma+r+d/2)/(\beta+\gamma)} I_5. \tag{16}$$

This constant is finite when

$$\beta > \max\{\iota_3(p,r,d,\gamma), \iota_4(p,r,d,\gamma), \iota_5(p,r,d,\gamma)\} \tag{17}$$

where $\iota_3(p,r,d,\gamma) = d/2 + (p-2)\gamma + (p-1)r$, $\iota_4(p,r,d,\gamma) = d(3/(2p) - 1/2) + r/p - \gamma/(p(p-1))$, $\iota_5(p,r,d,\gamma) = d/2 + 2r(p-1)/p + \gamma(p-2)/p$.

Kernel function. Denote

$$K_\beta(x) = b^{r+\gamma+d} L_\beta(bx), \qquad x \in \mathbf{R}^d, \tag{18}$$

where the Fourier transform of L_β is

$$\widehat{L}_\beta(\omega) = (2\pi)^{-d} i^r \omega^{\alpha_0} \|\omega\|^\gamma \left(1 + \|\omega\|^{q+2\gamma}\right)^{-1}, \qquad \omega \in \mathbf{R}^d, \tag{19}$$

$$b = \left(\frac{\beta - r - d/2}{\gamma + r + d/2} \frac{I_1^{(p-1)/p}}{I_2^{1/2}} \right)^{1/(\beta+\gamma)},$$

I_1 and I_2 are defined in (8) and (9), respectively. Denote

$$h = (\epsilon/L)^{1/(\beta+\gamma)} \tag{20}$$

and finally define the scaled kernel function of the estimator by

$$K_{\beta,h}(x) = h^{-\gamma-r-d} K_\beta(x/h). \tag{21}$$

The result. Bounds for the minimax risk are given by the following theorem. Define the exponent for the optimal rate of convergence as

$$\kappa = \frac{\beta - r - d/2}{\beta + \gamma}. \tag{22}$$

Theorem 1 *We have, for $\epsilon, L > 0$, $p > 1$, and β satisfying conditions (11) and (17), that*

$$c_{\beta,p} \le \epsilon^{-\kappa} L^{\kappa-1} R_\epsilon \le C_{\beta,p}$$

where $c_{\beta,p}$ is defined in (16), $C_{\beta,p}$ is defined in (10), and R_ϵ is defined in (4). The scaled kernel defined in (21) achieves the upper bound:

$$\sup_{\{f:\rho_{\beta,p}(f)\le L\}} \sup_{\{g:\|R_\gamma(f-g)\|_2 \le \epsilon\}} \left| \int_{\mathbf{R}^d} K_{\beta,h}(R_\gamma g) - f^{(\alpha_0)}(0) \right| \le \epsilon^\kappa L^{1-\kappa} C_{\beta,p}.$$

A proof of Theorem 1 is given in Section 3.1.

Remark 1. When $p = 2$, then the upper and lower bounds coincide and we have

$$C_{\beta,p} = c_{\beta,p} = I_1^{1/2} \left(\frac{\gamma + r + d/2}{\beta - r - d/2} \right)^{(\beta - r - d/2)/(2(\beta+\gamma))} \frac{\beta + \gamma}{\gamma + r + d/2}$$

where I_1 is defined in (8). Here we applied the property $B(a, b) = B(a - 1, b + 1)(a - 1)/b$ which implies $I_2 = I_1(\beta - r - d/2)/(\gamma + r + d/2)$. When $p \neq 2$, then we study tightness of the bounds only for the case $r = 0$. Figures 1 and 2 show that the upper bound and lower bound are close to each other. We plot the contour of $C_{\beta,p}/c_{\beta,p}$ as a function of (β, p). In Figure 1 we have $d = 1$ and in Figure 2 we have $d = 2$. We study the cases $\gamma = 0$, $\gamma = 0.5$, and $\gamma = 0.9$. In particular, from Figure 1 a) we have for the case $d = 1$, $r = 0$, $\gamma = 0$ that

$$\sup_{(\beta,p)\in[2,100]\times[1.1,2]} \frac{C_{\beta,p}}{c_{\beta,p}} \leq 1.13, \qquad \sup_{(\beta,p)\in[1,100]\times[2,100]} \frac{C_{\beta,p}}{c_{\beta,p}} \leq 1.039. \tag{23}$$

Figures 1 b.2) and 1 c.2) show that when $\gamma > 0$, smoothness parameter β is small, and p is large, then the bounds are not sharp. From Figure 2 a) we have for the case $d = 2$, $r = 0$, $\gamma = 0$ that

$$\sup_{(\beta,p)\in[2.5,100]\times[1.1,2]} \frac{C_{\beta,p}}{c_{\beta,p}} \leq 1.31, \qquad \sup_{(\beta,p)\in[1,100]\times[2.1,100]} \frac{C_{\beta,p}}{c_{\beta,p}} \leq 1.043.$$

Again, Figures 2 b.2) and 2 c.2) show that when $\gamma > 0$, smoothness parameter β is small, and p is large, then the bounds are not sharp.

Remark 2. For the case where $p = 2$, $d = 1$, $\gamma = 0$ (R_γ is equal to the identity operator), and $r \geq 0$ Taikov (1969) gave the solution to the optimal recovery. The case with $p = 2$, $d = 1$, $\gamma \geq 0$, and $r = 0$ was considered by Donoho and Low (1992), page 959. Klemelä and Tsybakov (2001) considered the case with $p = 2$, $d \geq 1$, $\gamma = 0$, and $r \geq 0$. Klemelä and Tsybakov (2004) considered the case $p = 2$, $d \geq 1$, $\gamma \geq 0$, and $r \geq 0$.

Remark 3. When $r = 0$ and $\gamma = 0$, then the condition on β posed in Theorem 1 is

$$\beta > \begin{cases} d/2, \text{ when} & p > 3/2 \\ d(3/(2p) - 1/2), \text{ when} & 1 < p \leq 3/2. \end{cases}$$

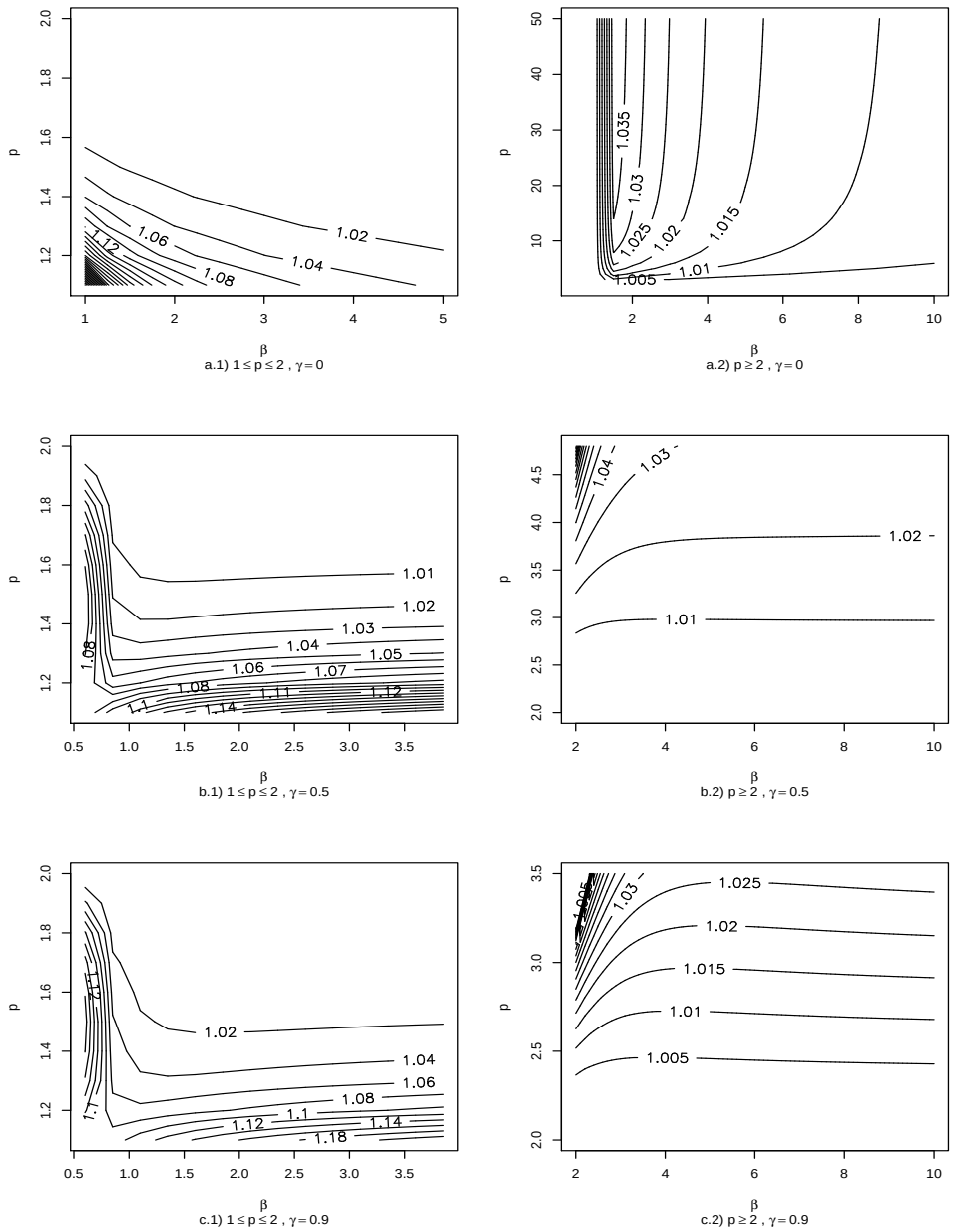

Figure 1: Contour of $C_{\beta,p}/c_{\beta,p}$ as a function of (β, p) for the case $d = 1$ and $r = 0$. In a) we have $\gamma = 0$, in b) $\gamma = 0.5$, and in c) $\gamma = 0.9$. On the left hand side $p \in [1, 2]$ and on the right hand side $p \geq 2$.

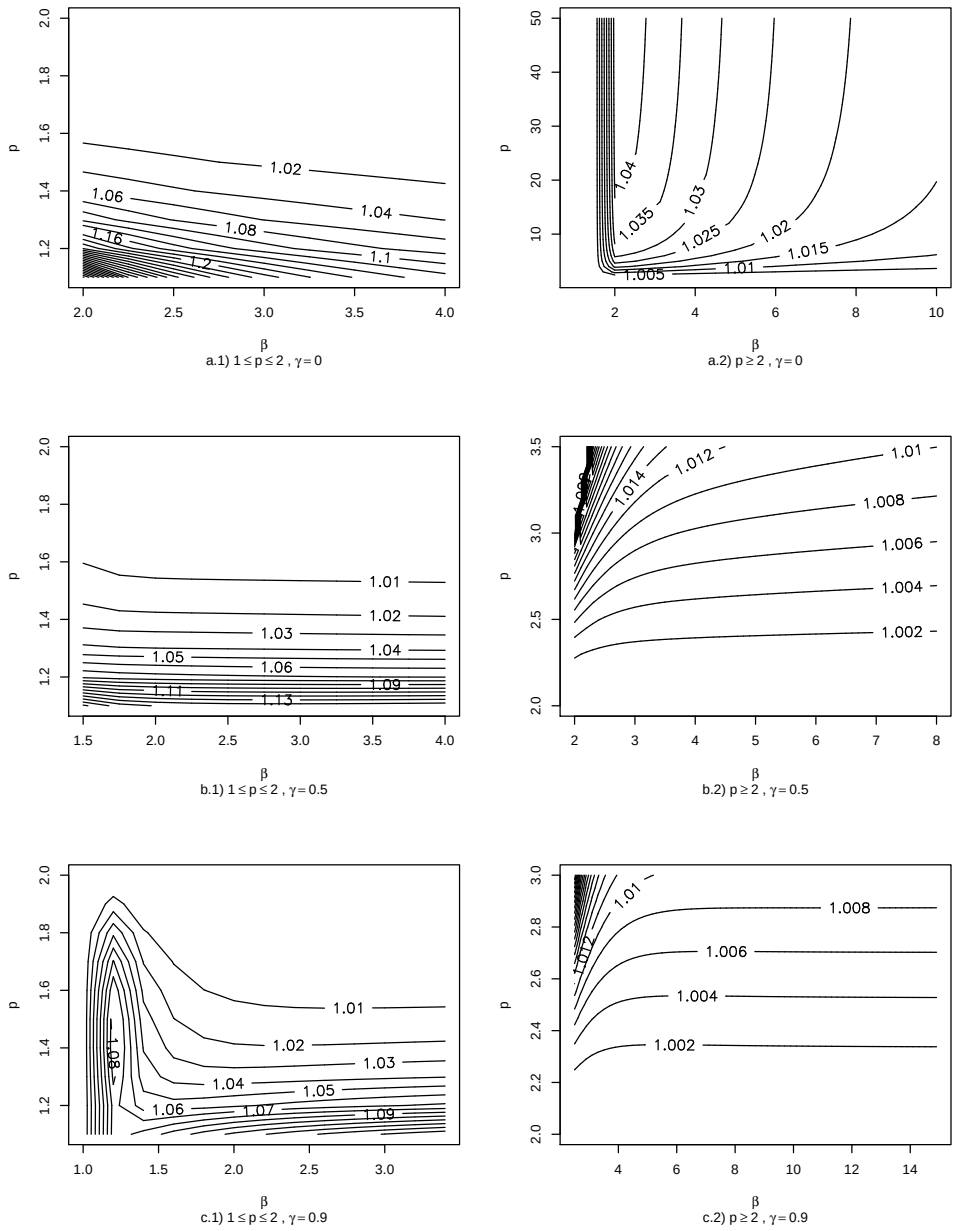

Figure 2: Contour of $C_{\beta,p}/c_{\beta,p}$ as a function of (β, p) for the case $d = 2$ and $r = 0$. In a) we have $\gamma = 0$, in b) $\gamma = 0.5$, and in c) $\gamma = 0.9$. On the left hand side $p \in [1, 2]$ and on the right hand side $p \geq 2$.

2.3 Solutions to the optimal recovery when $p = 1$

We consider the case when $p = 1$ and give an exact formula for the minimax risk (4). We also construct the optimal kernel. We consider the two cases

1. $r = 0$, $d \geq 1$,
2. $r \geq 0$ even, $d = 1$.

The exact constant. Let us denote

$$I_1 = (2\pi)^{-d} \tag{24}$$

and

$$I_2 = \frac{2\mu(S_d)(q-r)^2}{(2\pi)^d(2\gamma+2r+d)(q+2\gamma+r+d)(2q+2\gamma+d)} \tag{25}$$

where $q = \beta - d/2$ and $\mu(S_d)$ is defined in (7). The exact constant has the same form as (10), that is,

$$C_\beta = \left(I_2^{1/2}\right)^\kappa I_1^{1-\kappa} \left(\frac{\gamma+r+d/2}{\beta-r-d/2}\right)^\kappa \frac{\beta+\gamma}{\gamma+r+d/2} \tag{26}$$

where κ is defined in (22). This constant is finite when

$$\beta > r + d/2. \tag{27}$$

Kernel function. To define the scaled kernel function $K_{\beta,h}$ for the case $p = 1$, we take

$$\widehat{L}_\beta(\omega) = (2\pi)^{-d}\|\omega\|^\gamma \omega^r \left(1 - \|\omega\|^{q-r}\right)_+ \tag{28}$$

where $(a)_+ = \max\{a, 0\}$ and we denote $\omega^r = 1$ when $d \geq 1$ and $r = 0$. Define also

$$b = \left(\frac{\beta-r-d/2}{\gamma+r+d/2}\,\frac{I_1}{I_2^{1/2}}\right)^{1/(\beta+\gamma)},$$

and let I_1 be defined in (24), and I_2 be defined in (25). Then define $K_{\beta,h}$ by (21), (20), and (18).

The result. The following theorem states the optimality of the kernel and the constant.

Theorem 2 *We have, for the two cases (1) $r = 0$, $d \geq 1$, and (2) $r \geq 0$ even, $d = 1$, and for $\epsilon, L > 0$, $p = 1$, and β satisfying the condition (27), that the minimax risk defined in (4) satisfies*

$$R_\epsilon = \epsilon^\kappa L^{1-\kappa} C_\beta$$

where C_β is defined in (26) and κ is defined in (22). Also, the scaled kernel $K_{\beta,h}$ defined with the help of (28) is optimal:

$$R_\epsilon = \sup_{\{f:\rho_{\beta,p}(f)\leq L\}} \sup_{\{g:\|R_\gamma(f-g)\|_2\leq\epsilon\}} \left| \int_{\mathbf{R}^d} K_{\beta,h}(R_\gamma g) - f^{(\alpha_0)}(0) \right|.$$

A proof of Theorem 2 is given in Section 3.2.

2.4 Discussion on the smoothness conditions

We discuss the role of parameter p in the smoothness condition and the effect of p on the largeness of the optimal approximation error.

2.4.1 Relations between L_p Sobolev classes

We may note that for $p = 2$, and when β is an integer, the semi-norm in (1) is the classical L_2 Sobolev semi-norm

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

By relating other L_p semi-norms to the L_2 semi-norm we may give a connection to the classical L_2 smoothness conditions. We give a lemma on embeddings for the balls

$$\mathcal{F}_{\beta,L,p} = \{f : \rho_{\beta,p}(f) \leq L\},$$

only for the case $p = 1$ and $p = 2$, since the other cases are analogous.

We state a lemma which says roughly that to get a L_1 ball inside a L_2 ball we need to choose the smoothness index of the L_1 ball essentially larger than the smoothness index of the L_2 ball. On the other hand the lemma says that to get a L_2 ball inside a L_1 ball we need to choose the smoothness index of the L_2 ball only slightly larger than the smoothness index of the L_1 ball. Thus the lemma shows that the parametrization of the smoothness classes defined by the choice of q in (2) is reasonable, since smoothness classes corresponding to different values of p but the same value of β are comparable. We make such comparison in Section 2.4.2.

Lemma 3 *Let* $\mathcal{G}_M = \{f : |\widehat{f}(\omega)| \leq M \text{ for all } \omega \in \mathbf{R}^d\}$. *The intersection of a* L_1 *ball with* $\mathcal{G}_M$ *is a subset of a* L_2 *ball:*

$$\mathcal{F}_{\beta_1,L',1} \cap \mathcal{G}_M \subset \mathcal{F}_{\beta_2,L'',2} \tag{29}$$

when $\beta_1 \geq 2\beta_2 + d/2$, *for suitable* L', L'', *where* $0 < M < \infty$. *An intersection of two* L_2 *balls is a subset of a* L_1 *ball:*

$$\mathcal{F}_{\beta_2,L'',2} \cap \mathcal{F}_{\beta_1-d/2,L''',2} \subset \mathcal{F}_{\beta_1,L',1} \tag{30}$$

when $\beta_2 > \beta_1 > d/2$, *for suitable* L', L'', L'''.

Proof. Denote $q_1 = \beta_1 - d/2$ and $q_2 = 2\beta_2$. Inclusion (29) follows from the facts that for $f \in \mathcal{F}_{\beta_1,L',1} \cap \mathcal{G}_M$,

$$\int_{\mathbf{R}^d} \|\omega\|^{q_2} \left|\widehat{f}(\omega)\right|^2 d\omega \leq M \int_{\mathbf{R}^d} \|\omega\|^{q_2} \left|\widehat{f}(\omega)\right| d\omega$$

and $q_1 \geq q_2 \Leftrightarrow \beta_1 \geq 2\beta_2 + d/2$. To prove (30) we apply the Cauchy inequality for $f \in \mathcal{F}_{\beta_2,L'',2} \cap \mathcal{F}_{\beta_1-d/2,L''',2}$:

$$\begin{aligned}
&\int_{\mathbf{R}^d} \|\omega\|^{q_1} \left|\widehat{f}(\omega)\right| d\omega \\
&\leq \left[\int_{\mathbf{R}^d} (1 + \|\omega\|^{2\beta_2'})^{-1} d\omega\right]^{1/2} \left[\int_{\mathbf{R}^d} (1 + \|\omega\|^{2\beta_2'}) \|\omega\|^{2q_1} \left|\widehat{f}(\omega)\right|^2 d\omega\right]^{1/2}
\end{aligned}$$

where $\beta_2' = \beta_2 - \beta_1 + d/2$. The first integral on the right hand side is finite since $\beta_2' > d/2$. The second integral on the right hand side is finite since $\rho_{\beta_1-d/2,2}(f) < \infty$, $\rho_{\beta_2,2}(f) < \infty$, and $2(\beta_2' + q_1) = q_2$. □

2.4.2 Curse of dimensionality and p

We have shown in Theorem 1 and Theorem 2 that the rate of convergence, as $\epsilon \to 0$, of the optimal pointwise approximation error is ϵ^κ, where κ is defined in (22). To guarantee that the rate of convergence is not unacceptably slow, the smoothness index β has to increase as dimension d increases. The rate of convergence does not depend on parameter p. We may however note that parameter p has a considerable influence on the largeness of the constants.

Figure 3 a) shows the magnitudes of the constants as a function of dimension d. We show with bullet "•" the constant $C_{\beta,p}$ for $p = 1$, with circle "○" the constant for $p = 2$, and with square "□" the constants for $p = 4$; for the case $p = 4$ we show both the constant for the lower bound and the

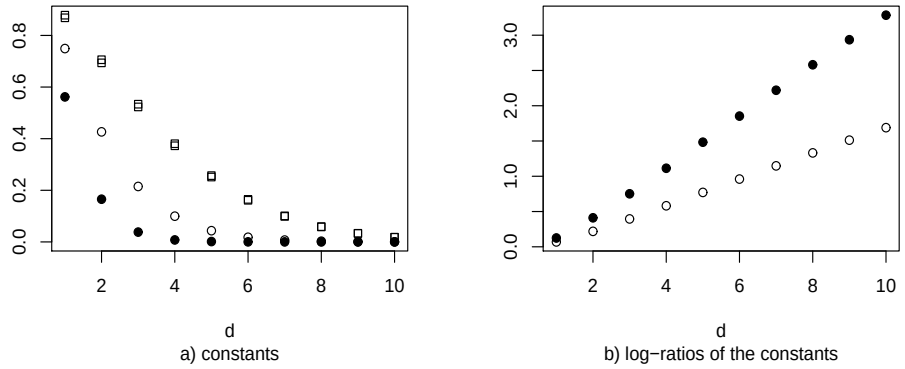

Figure 3: a) Magnitudes of the constants as a function of d, for $p = 1$, $p = 2$, and $p = 4$. b) Logarithms of the ratios of the constants.

constant for the upper bound. We take $d = 1, \ldots, 10$, $\beta = 2 + d/2$, $r = 0$, and $\gamma = 0$.

Figure 3 b) shows the logarithm of the ratio of constants as a function of dimension d. We show with bullet "•" the logarithm of the ratio of the constants corresponding to $p = 2$ and $p = 1$: $\log_{10}(C_{\beta,2}/C_{\beta,1})$. We show with circle "∘" the logarithm of the ratio of the constants corresponding to $p = 4$ and $p = 2$: $\log_{10}(C_{\beta,4}/C_{\beta,2})$. We take, as before, $d = 1, \ldots, 10$, $\beta = 2 + d/2$, $r = 0$, and $\gamma = 0$.

Figure 3 a) shows that the constants are in fact decreasing as the dimension grows, giving a compensation to the worsening rate of convergence as the dimension grows. Even more interestingly, Figure 3 b) highlights the fact that the constants are decreasing much faster for the lower values of p. In fact, the ratios of the constant corresponding to a larger value of p, to the constant corresponding to a smaller value of p, are increasing exponentially, as the dimension grows.

Previously, L_1 Sobolev smoothness conditions have been studied in high dimensional cases by Jones (1992), Barron (1993), Breiman (1993). They show that for this smoothness condition and for the integrated squared error loss the optimal approximation rate does not depend on the dimension d of the function. In fact, the rate is ψ_ϵ^2 where $\psi_\epsilon = (\epsilon^2 \cdot \log(\epsilon^{-1}))^{1/4}$. When we compare this to the classical nonparametric rate ϵ^κ, where $\kappa = (\beta - d/2)/\beta$, as defined in (22), for $r = 0$, $\gamma = 0$, then we note that

$$(\epsilon^2 \cdot \log(\epsilon^{-1}))^{1/4} < \epsilon^\kappa \Leftrightarrow d > \beta$$

when $0 < \epsilon < 1$, (and we have always $\beta > d/2$). Thus the rate ψ_ϵ is better

than the classical rate in the high dimensional cases: when $d > \beta$.

We have not considered the integrated squared error loss but the pointwise estimation, and for this case the classical rate ϵ^κ is optimal also for the L_1 Sobolev class. However, we have shown that the curse of dimensionality may be somewhat avoided also for the pointwise estimation, since the constant in the optimal approximation error is essentially smaller for the L_1 Sobolev class than for the L_2 Sobolev class, for the high dimensional cases.

2.5 Statistical applications

We study pointwise minimax estimation in Section 2.5.1 and Section 2.5.2. We formulate the results only for the case $p > 1$ but it is straightforward to formulate the results also for the case $p = 1$. We study adaptive estimation in Section 2.5.3, for the case $p > 1$.

In order to study regression estimation with i.i.d. data it is instructive to reduce regression estimation to Gaussian white noise model, and then in turn to reduce Gaussian white noise model to optimal recovery. We start with the Gaussian white noise model.

2.5.1 Gaussian white noise model.

Assume observation

$$dY_\epsilon(x) = (R_\gamma f)(x)dx + \epsilon\, dW(x), \qquad x \in \mathbf{R}^d, \tag{31}$$

where W is the d-dimensional Brownian sheet and $\epsilon > 0$. Given a realization of the process $Y_\epsilon(x)$, the problem is to estimate $f^{\alpha_0}(0)$ where $|\alpha_0| = r$. Denote minimax risk with squared error loss, for linear estimators, by

$$R_\epsilon^{(2)} = \inf_{K \in L_2} \sup_{\rho_{\beta,p}(f) \le L} E_f \left| \int_{\mathbf{R}^d} K dY_\epsilon - f^{\alpha_0}(0) \right|^2 .$$

Theorem 4 *Let $\epsilon, L > 0$, $p > 1$, and let β satisfy conditions (11) and (17). We have that*

$$c_{\beta,p}^2 \le \left[\left(\epsilon^\kappa L^{1-\kappa}\right)^2 \kappa^\kappa (1-\kappa)^{1-\kappa} \right]^{-1} R_\epsilon^{(2)} \le C_{\beta,p}^2$$

where $c_{\beta,p}$ is defined in (16), $C_{\beta,p}$ is defined in (10), and κ is defined in (22). Define

$$\tilde{b} = \left[\left(\frac{\gamma + r + d/2}{\beta - r - d/2} \right)^{1/2} \frac{I_2^{1/2}}{I_1^{(p-1)/p}} \right]^{1/(\beta+\gamma)}, \tag{32}$$

where I_1 and I_2 are defined in (8) and (9), respectively. Define scaled kernel $K_{\beta,h}$ with (18), (19), (20), and (21), but with b replaced by $\tilde{b}$. The kernel estimator with scaled kernel $K_{\beta,h}$ achieves the upper bound:

$$\sup_{\rho_{\beta,p}(f)\le L} E_f \left| \int_{\mathbf{R}^d} K_{\beta,h} dY_\epsilon - f^{\alpha_0}(0) \right|^2 \le \left(\epsilon^\kappa L^{1-\kappa} C_{\beta,p}\right)^2 \kappa^\kappa (1-\kappa)^{1-\kappa}.$$

A proof of Theorem 4 is given in Section 3.3.

Remark 4. To understand the difference between Theorems 1 and 4 note that the proofs show that for the optimal recovery we have the upper bound

$$A(h/b)^{\beta-d/2-r} + \epsilon B (h/b)^{-\gamma-r-d/2}$$

where $A = L I_1^{(p-1)/p}$ and $B = I_2^{1/2}$ and for the white noise model we have the upper bound

$$\left[A(h/\tilde{b})^{\beta-d/2-r}\right]^2 + \left[\epsilon B (h/\tilde{b})^{-\gamma-r-d/2}\right]^2.$$

2.5.2 Regression function estimation.

Consider estimation of regression function $f : \mathbf{R}^d \to \mathbf{R}$ and its derivatives at one point based on i.i.d. observations $(Y_1, X_1), \dots, (Y_n, X_n)$ where $Y_i \in \mathbf{R}$ and $X_i \in \mathbf{R}^d$. We assume that

$$Y_i = (R_\gamma f)(X_i) + \epsilon\, \xi_i \tag{33}$$

where ξ_i, $i = 1, \dots, n$, are i.i.d. random variables which are independent from X_i. We assume that $E(\xi_i) = 0$, and $\mathrm{Var}(\xi_i) = 1$. We assume (for the lower bound) that the Fisher information functional I_ξ is finite and positive, where we define

$$I_\xi = \int_{\{x\in\mathbf{R}: f_\xi(x)>0\}} \frac{f_\xi'(x)^2}{f_\xi(x)}\, dx$$

where $f_\xi : \mathbf{R} \to \mathbf{R}$ is the density of ξ_1. We assume that the first derivative f_ξ' is uniformly continuous. Variables X_i are i.i.d. and we denote the density of X_1 with $f_X : \mathbf{R}^d \to \mathbf{R}$, assume that this density is known, and $f_X(0) > 0$. Denote the kernel estimator with

$$\theta_n(K) = \frac{1}{n f_X(0)} \sum_{i=1}^n Y_i K(X_i)$$

where $K : \mathbf{R}^d \to \mathbf{R}$. Denote minimax risk with squared error loss, for linear estimators, by

$$R_n^{(3)} = \inf_{K \in L_2} \sup_{\rho_{\beta,p}(f) \le L} E_f \left| \theta_n(K) - f^{(\alpha_0)}(0) \right|^2 .$$

By the noise calibration

$$\tilde{\epsilon}_u = \frac{\epsilon}{\sqrt{n f_X(0)}} \tag{34}$$

we may reduce the calculation of the upper bound to the corresponding calculation in the Gaussian white noise model. By the noise calibration

$$\tilde{\epsilon}_l = \frac{\epsilon}{\sqrt{n f_X(0) I_\xi}}$$

we may reduce the calculation of the lower bound to the corresponding calculation in the Gaussian white noise model.

Theorem 5 *Let $\epsilon, L > 0$, $p > 1$, and let β satisfy conditions (11) and (17). We have that*

$$\begin{aligned} c_{\beta,p}^2 I_\xi^{-\kappa} &\le \frac{f_X^\kappa(0)}{\epsilon^{2\kappa} L^{2(1-\kappa)} \kappa^\kappa (1-\kappa)^{1-\kappa}} \liminf_{n\to\infty} n^\kappa R_n^{(3)} \\ &\le \frac{f_X^\kappa(0)}{\epsilon^{2\kappa} L^{2(1-\kappa)} \kappa^\kappa (1-\kappa)^{1-\kappa}} \limsup_{n\to\infty} n^\kappa R_n^{(3)} \le C_{\beta,p}^2 \end{aligned}$$

where $c_{\beta,p}$ is defined in (16), $C_{\beta,p}$ is defined in (10), and κ is defined in (22). Let $\tilde{b}$ be as in (32) and define scaled kernel $K_{\beta,h}$ with (18), (19), (20), and (21), but with b replaced with $\tilde{b}$ defined in (32) and ϵ replaced with $\tilde{\epsilon}_u$ defined in (34). The kernel estimator with scaled kernel $K_{\beta,h}$ achieves the upper bound:

$$\begin{aligned} &\limsup_{n\to\infty} n^\kappa \sup_{\rho_{\beta,p}(f) \le L} E_f \left| \theta_n(K_{\beta,h}) - f^{\alpha_0}(0) \right|^2 \\ &\quad \le \left(\frac{\epsilon^2}{f_X(0)} \right)^\kappa L^{2(1-\kappa)} \kappa^\kappa (1-\kappa)^{1-\kappa} C_{\beta,p}^2 . \end{aligned}$$

A proof of Theorem 5 is given in Section 3.4.

Remark 5. When the distribution of error ξ is the standard Gaussian distribution, then $I_\xi = 1$.

2.5.3 Adaptive pointwise estimation

Let us consider again the observation in the Gaussian white noise model defined in (31). Define

$$\tilde{\mathcal{F}}_{\beta,L} = \left\{ f \in L_1(\mathbf{R}^d) \,\middle|\, \rho^p_{\beta,p}(f) + (2\pi)^d \|(R_\gamma f)^\wedge\|_p^p \le L^p \right\}. \tag{35}$$

We want to consider adaptive estimation of $f^{(\alpha_0)}(0)$ over the scale of classes $\tilde{\mathcal{F}}_{\beta,L}$. We assume that we know that $f \in \tilde{\mathcal{F}}_{\beta,L}$ for some $\beta \in [\beta_*, \infty)$ and $L \in [L_*, L^*]$, where β_* satisfies

$$\max\{\iota_i : i = 1, \dots, 5\} < \beta_* < \infty,$$

ι_i are defined after displays (11) and (17), and $0 < L_* < L^* < \infty$. For the adaptive estimation the rate of convergence is slower than the rate in the non-adaptive case by a logarithmic factor. Define the "adaptive factor"

$$\alpha_{\beta,\epsilon} = \left[\log_e \left(\epsilon^{-1} \right) \frac{2\tau(r + \gamma + d/2)}{\beta + \gamma} \right]^{1/2}$$

where $\tau > 0$ will be the power of the loss function. The following theorem gives an upper and lower bound for the asymptotic adaptive risk. For a discussion and a construction of an estimator achieving the upper bound we refer to Klemelä and Tsybakov (2004).

Theorem 6 *Let $p > 1$ and let $C_{\beta,p}$ be the constant defined in (10) and $c_{\beta,p}$ be the constant defined in (16). Then,*

$$\limsup_{\epsilon \to 0} \inf_{T_\epsilon} \sup_{(\beta,L) \in B_\epsilon} [(\epsilon \alpha_{\beta,\epsilon})^\kappa L^{1-\kappa} C_{\beta,p}]^{-\tau} \sup_{f \in \tilde{\mathcal{F}}_{\beta,L}} E_f \left| T_\epsilon - f^{(\alpha_0)}(0) \right|^\tau \le 1$$

and

$$\liminf_{\epsilon \to 0} \inf_{T_\epsilon} \sup_{(\beta,L) \in B_\epsilon} [(\epsilon \alpha_{\beta,\epsilon})^\kappa L^{1-\kappa} c_{\beta,p}]^{-\tau} \sup_{f \in \tilde{\mathcal{F}}_{\beta,L}} E_f \left| T_\epsilon - f^{(\alpha_0)}(0) \right|^\tau \ge 1,$$

where $B_\epsilon = [\beta_*, \beta_\epsilon] \times [L_*, L^*]$, $\beta_\epsilon = (\log \log \epsilon^{-1})^\delta$, $0 < \delta < 1$, $\inf_{T_\epsilon}$ denotes the infimum over all estimators in the Gaussian model (31), and $\tau > 0$.

Proof. The proof is essentially given in Klemelä and Tsybakov (2004). The bias calculation in Lemma 1 of that paper has to modified, in the way shown in the proof of the upper bound of Theorem 1 of this paper. Otherwise the arguments are the same. □

3 Proofs

3.1 Proof of Theorem 1

We will denote for simplicity

$$J(a,b) = \int_0^\infty t^a \left(1+t^{q+2\gamma}\right)^{-b} dt = \frac{1}{q+2\gamma}\, B\left(b - \frac{a+1}{q+2\gamma}\,,\, \frac{a+1}{q+2\gamma}\right), \quad (36)$$

where $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$, $a,b>0$, is the Beta function and q is defined in (2).

Upper bound. Let f and g be such that

$$\rho_{\beta,p}(f) \le L, \qquad \|R_\gamma(f-g)\|_2 \le \epsilon,$$

and let $K_{\beta,h}$ be the scaled kernel defined in (21). We have

$$\begin{aligned}
&\left|\int_{\mathbf{R}^d} K_{\beta,h}(R_\gamma g) - f^{(\alpha_0)}(0)\right| \\
&\quad \le \left|\int_{\mathbf{R}^d} K_{\beta,h}(R_\gamma f) - f^{(\alpha_0)}(0)\right| + \left|\int_{\mathbf{R}^d} K_{\beta,h}(R_\gamma g - R_\gamma f)\right|. \qquad (37)
\end{aligned}$$

We have that

$$\widehat{K}_{\beta,h}(\omega) = (2\pi)^{-d} i^r \omega^{\alpha_0} \|\omega\|^\gamma \left(1+\|h\omega/b\|^{2(\beta+\gamma)}\right)^{-1}$$

where $\widehat{K}_{\beta,h}$ is defined in (21). Hence, using the formula for the Fourier transform of $R_\gamma f$ given in (3), by the Hölder inequality,

$$\begin{aligned}
&\left|\int_{\mathbf{R}^d} K_{\beta,h}(R_\gamma f) - f^{(\alpha_0)}(0)\right| \\
&= \left|\int_{\mathbf{R}^d} \widehat{f}(\omega)\left((2\pi)^d \widehat{K}_{\beta,h}(\omega)\|\omega\|^{-\gamma} - i^r\omega^{\alpha_0}\right) d\omega\right| \\
&= \left|\int_{\mathbf{R}^d} \widehat{f}(\omega)\left(\frac{i^r\omega^{\alpha_0}}{1+\|h\omega/b\|^{2(q+2\gamma)}} - i^r\omega^{\alpha_0}\right) d\omega\right| \\
&= \left|\int_{\mathbf{R}^d} \widehat{f}(\omega) i^r\omega^{\alpha_0}\, \frac{\|h\omega/b\|^{q+2\gamma}}{1+\|h\omega/b\|^{q+2\gamma}}\, d\omega\right| \\
&\le (h/b)^{q/p-r-d/p'} \left[(2\pi)^d \int_{\mathbf{R}^d} \|\omega\|^q \left|\widehat{f}(\omega)\right|^p d\omega\right]^{1/p} \\
&\qquad \times \left[\int_{\mathbf{R}^d} \left((2\pi)^{-d/p}\, \frac{\omega^{\alpha_0}\|\omega\|^{2\gamma+q(1-1/p)}}{1+\|\omega\|^{q+2\gamma}}\right)^{p'} d\omega\right]^{1/p'} \\
&\le (h/b)^{q/p-r-d/p'} L I_1^{1/p'} \qquad (38)
\end{aligned}$$

where $p' = p/(p-1)$ and I_1 is the integral defined in (8), which can be written as

$$\begin{aligned} I_1 &= (2\pi)^d \int_{\mathbf{R}^d} \|\omega\|^q \left(\|\omega\|^\gamma \left| \widehat{L}_\beta(\omega) \right| \right)^{p/(p-1)} d\omega \\ &= (2\pi)^{d(1-p/(p-1))} K_u(p)\, J\left(q + \frac{p}{p-1}(2\gamma + r) + d - 1,\, \frac{p}{p-1} \right) \end{aligned}$$

where $\widehat{L}_\beta$ is defined in (19), $K_u(p)$ is defined in (6), and J is defined in (36). Also,

$$\left| \int_{\mathbf{R}^d} K_{\beta,h}(R_\gamma f - R_\gamma g) \right| \le \|K_{\beta,h}\|_2 \|R_\gamma(f-g)\|_2 \le (h/b)^{-\gamma-r-d/2} I_2^{1/2}\, \epsilon \quad (39)$$

where I_2 is the integral defined in (9), which can be written as

$$I_2 = \|L_\beta\|_2^2 = (2\pi)^d \|\widehat{L}_\beta\|_2^2 = (2\pi)^{-d} K_u(2)\, J(2(\gamma + r) + d - 1, 2).$$

Thus, by (37), (38), and (39),

$$\left| \int_{\mathbf{R}^d} K_{\beta,h}(R_\gamma g) - f^{(\alpha_0)}(0) \right| \le L(h/b)^{q/p-r-d/p'} I_1^{1/p'} + \epsilon (h/b)^{-\gamma-r-d/2} I_2^{1/2}.$$

The upper bound follows by the choice of h and b.

Lower bound. Let

$$g_\beta(x) = a_1 b_1^{r+\gamma+d} h_\beta(b_1 x)$$

where the Fourier transformation of h_β is

$$\widehat{h}_\beta(\omega) = \|\omega\|^\gamma i^r \omega^{\alpha_0} \left[(2\pi)^{-d} \|\omega\|^\gamma \left(1 + \|\omega\|^{q+2\gamma}\right)^{-1} \right]^{1/(p-1)},$$

and a_1 and b_1 are such that

$$\rho_{\beta,p}(g_\beta) = L, \qquad \|R_\gamma g_\beta\|_2 = \epsilon. \quad (40)$$

We have that,

$$\widehat{g}_\beta(\omega) = a_1 b_1^{r+\gamma} \widehat{h}_\beta(\omega/b_1),$$

and thus,

$$\begin{aligned} \rho_{\beta,p}^p(g_\beta) &= (2\pi)^d \int_{\mathbf{R}^d} \|\omega\|^q \, |\widehat{g}_\beta(\omega)|^p \, d\omega \\ &= a_1^p b_1^{p(r+\gamma)} (2\pi)^d \int_{\mathbf{R}^d} \|\omega\|^q \left| \widehat{h}_\beta(\omega/b_1) \right|^p d\omega \\ &= a_1^p b_1^{p(r+\gamma)+d+q} I_3, \end{aligned} \quad (41)$$

where I_3 is the integral defined in (13), which can also be written as

$$\begin{aligned} I_3 &= (2\pi)^d \int_{\mathbf{R}^d} \|\omega\|^q \left|\widehat{h}_\beta(\omega)\right|^p d\omega \qquad (42) \\ &= (2\pi)^{d(1-p/(p-1))} K_l(p)\, J\left(q + \frac{p\gamma}{p-1} + p(\gamma+r) + d - 1,\, \frac{p}{p-1}\right), \end{aligned}$$

where $K_l(p)$ is defined in (12), and J is defined in (36). Also, using Equation (3) for the Fourier transform of the Riesz transform,

$$\begin{aligned} \|R_\gamma g_\beta\|_2^2 &= (2\pi)^d \int_{\mathbf{R}^d} |(R_\gamma g_\beta)^\wedge|^2 \\ &= (2\pi)^d \int_{\mathbf{R}^d} \|\omega\|^{-2\gamma} \left|\widehat{g}_\beta(\omega)\right|^2 d\omega \\ &= (2\pi)^d a_1^2 b_1^{2(r+\gamma)} \int_{\mathbf{R}^d} \|\omega\|^{-2\gamma} \left|\widehat{h}_\beta(\omega/b_1)\right|^2 d\omega \\ &= a_1^2 b_1^{2r+d} I_4 \qquad (43) \end{aligned}$$

where I_4 is the integral defined in (14), which can also be written as

$$\begin{aligned} I_4 &= (2\pi)^d \int_{\mathbf{R}^d} \|\omega\|^{-2\gamma} \left|\widehat{h}_\beta(\omega)\right|^2 d\omega \qquad (44) \\ &= (2\pi)^{d(1-2/(p-1))} K_l(2)\, J\left(\frac{2\gamma}{p-1} + 2r + d - 1,\, \frac{2}{p-1}\right). \end{aligned}$$

We have from (40), (43), and (41), that $a_1 = \epsilon I_4^{-1/2} b_1^{-r-d/2}$, and

$$b_1 = \left(\frac{L I_4^{1/2}}{\epsilon I_3^{1/p}}\right)^{1/(\beta+\gamma)}.$$

We have that

$$\begin{aligned} g_\beta^{\alpha_0}(0) &= i^{|\alpha_0|} \int_{\mathbf{R}^d} \omega^{\alpha_0} \widehat{g}_\beta(\omega) d\omega \\ &= a_1 b_1^{r+\gamma} i^r \int_{\mathbf{R}^d} \omega^{\alpha_0} \widehat{h}_\beta(\omega/b_1) d\omega \\ &= a_1 b_1^{2r+\gamma+d} I_5, \end{aligned}$$

where I_5 is the integral defined in (15), which can also be written as

$$\begin{aligned} I_5 &= \int_{\mathbf{R}^d} \omega^{\alpha_0} i^r \widehat{h}_\beta(\omega) d\omega \qquad (45) \\ &= (2\pi)^{-d/(p-1)} K_l(2)\, J\left(\frac{p\gamma}{p-1} + 2r + d - 1,\, \frac{1}{p-1}\right). \end{aligned}$$

By the formulas for a_1 and b_1,

$$g_\beta^{\alpha_0}(0) = \left(\epsilon I_4^{-1/2}\right)^{(\beta-r-d/2)/(\beta+\gamma)} \left(L I_3^{-1/p}\right)^{(\gamma+r+d/2)/(\beta+\gamma)} I_5. \tag{46}$$

For a kernel $K \in L_2$, for $f = g_\beta$, and $g \equiv 0$,

$$\left| \int_{\mathbf{R}^d} K(R_\gamma g) - f^{(\alpha_0)}(0) \right| = g_\beta^{\alpha_0}(0). \tag{47}$$

The lower bound follows from (46) and from (47).

3.2 Proof of Theorem 2

Upper bound. Let f and g be such that

$$\rho_{\beta,p}(f) \le L, \qquad \|R_\gamma(f-g)\|_2 \le \epsilon,$$

and let $K_{\beta,h}$ be the scaled kernel defined with the help of (28). Then,

$$\widehat{K}_{\beta,h}(\omega) = (2\pi)^{-d} \|\omega\|^\gamma \omega^r \left(1 - \|h\omega/b\|^{q-r}\right)_+,$$

where we denote $\omega^r = 1$ when $d \ge 1$ and $r = 0$. Then,

$$\begin{aligned}
&\left| \int_{\mathbf{R}^d} K_{\beta,h}(R_\gamma f) - f^{(r)}(0) \right| \\
&= \left| \int_{\mathbf{R}^d} \widehat{f}(\omega) \left((2\pi)^d \widehat{K}_{\beta,h}(\omega) \|\omega\|^{-\gamma} - \omega^r \right) d\omega \right| \\
&\le \int_{\|h\omega/b\| \le 1} \omega^r \|h\omega/b\|^{q-r} \left| \widehat{f}(\omega) \right| d\omega + \int_{\|h\omega/b\|>1} |\omega|^r \left| \widehat{f}(\omega) \right| d\omega \\
&\le (h/b)^{q-r} \int_{\mathbf{R}^d} \|\omega\|^q \left| \widehat{f}(\omega) \right| d\omega \\
&\le (h/b)^{q-r} \rho_{\beta,p}(f) (2\pi)^{-d} \\
&\le (h/b)^{q-r} L I_1
\end{aligned} \tag{48}$$

since $\|h\omega/b\| > 1$ is equivalent to $(h/b)^{q-r}\|\omega\|^q > \omega^r$, and where we denote $I_1 = (2\pi)^{-d}$. We have for the integral I_2 defined in (25) that

$$I_2 = \|L_\beta\|^2 = (2\pi)^d \|\widehat{L}_\beta\|_2^2.$$

The rest of the derivation of the upper bound is similar to the proof of the upper bound for Theorem 1.

Lower bound. The proof of the lower bound is otherwise similar to the proof of the lower bound of Theorem 1 but now we take

$$\widehat{h}_\beta(\omega) = (2\pi)^{-d}\|\omega\|^{2\gamma+r}\left(1 - \|\omega\|^{q-r}\right)_+ .$$

Now the integrals (42), (44), and (45) have different formulas:

$$I_3 = \frac{\mu(S_d)(q-r)}{(q+2\gamma+r+d)(2q+2\gamma+d)},$$

$I_4 = I_2$ where I_2 is defined in (25), and

$$I_5 = \frac{\mu(S_d)(q-r)}{(2\pi)^d(2\gamma+2r+d)(q+2\gamma+r+d)} .$$

We have a similar lower bound as in (46), and by direct calculation one sees that the lower bound is equal to the upper bound. Theorem 2 is proved.

3.3 Proof of Theorem 4

Upper bound. We have for the expectation and variance of the estimator that

$$E_f \int_{\mathbf{R}^d} K_{\beta,h} dY_\epsilon = \int_{\mathbf{R}^d} K_{\beta,h}(R_\gamma f), \qquad \mathrm{Var}_f\left(\int_{\mathbf{R}^d} K_{\beta,h} dY_\epsilon\right) = \epsilon^2 \int_{\mathbf{R}^d} K_{\beta,h}^2.$$

Thus, by (38),

$$\begin{aligned} E_f \left| \int_{\mathbf{R}^d} K_{\beta,h} dY_\epsilon - f^{\alpha_0}(0) \right|^2 &= \left| \int_{\mathbf{R}^d} K_{\beta,h}(R_\gamma f) - f^{(\alpha_0)}(0) \right|^2 + \epsilon^2 \int_{\mathbf{R}^d} K_{\beta,h}^2 \\ &\le (h/\tilde{b})^{2\beta-d-2r} \left(L I_1^{(p-1)/p} \right)^2 + \epsilon^2 (h/\tilde{b})^{-2\gamma-2r-d} I_2 \end{aligned}$$

where I_2 is defined by (9) (and satisfies $I_2 = \|L_\beta\|_2^2$). The upper bound follows by the choice of h and $\tilde{b}$.

Lower bound. The lower bound follows from the proof of Theorem 2 of Donoho and Liu (1991). In fact, that theorem gives the formula

$$R_\epsilon^{(2)} = \sup_\alpha \left(\frac{\tilde{\omega}(\alpha)}{\alpha} \right)^2 \rho_A(\alpha/2, \epsilon)$$

where $\rho_A(\alpha/2,\epsilon) = (\alpha/2)^2\epsilon^2/[(\alpha/2)^2 + \epsilon^2]$ and

$$\tilde{\omega}(\epsilon) = \sup\left\{ |f^{(\alpha_0)}(0) - g^{(\alpha_0)}(0)| : \|R_\gamma f - R_\gamma g\|_2 \le \epsilon,\ \ \rho_{\beta,p}(f), \rho_{\beta,p}(g) \le L \right\}.$$

Note that when we define the modulus of continuity ω as in (5), then $\omega(\epsilon) = \tilde{\omega}(2\epsilon)/2$.

3.4 Proof of Theorem 5

Upper bound. The expectation of the estimator is

$$\begin{aligned} E_f(\theta_n(K_{\beta,h})) &= \frac{1}{f_X(0)} \int_{\mathbf{R}^d} (R_\gamma f)(x) K_{\beta,h}(x) f_X(x)\, dx \\ &\sim \int_{\mathbf{R}^d} (R_\gamma f)(x) K_{\beta,h}(x)\, dx \end{aligned}$$

and the variance is

$$\begin{aligned} \mathrm{Var}_f(\theta_n(K_{\beta,h})) &= \frac{\epsilon^2}{n f_X^2(0)} \mathrm{Var}_f\left(K_{\beta,h}(X_1)\right) \le \frac{\epsilon^2}{n f_X^2(0)} \int_{\mathbf{R}^d} K_{\beta,h}^2(x) f_X(x)\, dx \\ &\sim \frac{\epsilon^2}{n f_X(0)} \int_{\mathbf{R}^d} K_{\beta,h}^2. \end{aligned}$$

Thus the the proof is otherwise similar to the proof of the upper bound of Theorem 1 but now we replace ϵ in Theorem 1 by $\tilde{\epsilon}_u = \epsilon/\sqrt{n f_X(0)}$.

Lower bound. We construct the proof by a renormalization argument as in Low (1992) and Klemelä (2003). That is, we construct a local nonparametric experiment which converges to the Gaussian white noise experiment. An alternative approach would be to construct a one dimensional subexperiment and consider its convergence as in Donoho and Liu (1991). These papers considered density estimation. Lower bounds for regression estimation were considered also by Golubev (1991).

For $g : \mathbf{R}^d \to \mathbf{R}$, let

$$(T_n g)(x) = f_X(0)^{-1/2} \epsilon I_\xi^{-1/2} n^{-1/2} \kappa_n^{d/2} g(\kappa_n x), \qquad x \in \mathbf{R}^d,$$

where

$$\kappa_n = n^{1/(2\beta)}.$$

Transformation T_n is in a certain sense L_2-invariant which will lead to local asymptotic normality. Define

$$L' = L f_X(0)^{1/2} \epsilon^{-1} I_\xi^{1/2} \tag{49}$$

and

$$\mathcal{G}_c = \left\{ g : \mathbf{R}^d \to \mathbf{R} \,|\, \rho_{\beta,p}(g) \le L',\ \ \|g\|_\infty \le c \right\},$$

where $0 < c < \infty$. Now, when $\rho_{\beta,p}(g) \le L'$, then

$$\begin{aligned} \rho_{\beta,p}(T_n g) &= f_X(0)^{-1/2} \epsilon I_\xi^{-1/2} n^{-1/2} \kappa_n^\beta \rho_{\beta,p}(g) \\ &= f_X(0)^{-1/2} \epsilon I_\xi^{-1/2} \rho_{\beta,p}(g) \le L. \end{aligned}$$

Thus

$$\{T_n g : g \in \mathcal{G}_c\} \subset \{f : \rho_{\beta,p}(f) \leq L\}. \tag{50}$$

Let E_n be the regression experiment with observations (33), but in place of regression function f we take $T_n g$. Let the parameter space of the experiment be $\mathcal{G}_c$. Denote

$$E_n = (P_{g,n} : g \in \mathcal{G}_c).$$

Let E be the Gaussian white noise experiment with the observation

$$(R_\gamma g)(t)dt + dW(t), \qquad t \in \mathbf{R}^d.$$

Let the parameter space of the experiment be also $\mathcal{G}_c$. Denote

$$E = (P_g : g \in \mathcal{G}_c).$$

Now it holds that the local experiment E_n converges to the Gaussian white noise experiment E.

Lemma 7 *The experiment E_n convergences weakly, in Le Cam's sense, to the experiment E.*

Proof. We apply the proof of Theorem 3.1 of Klemelä (2003). We need to check 4 facts. (i) Sequence E_n is contiguous, that is, for all $g_1, g_2 \in \mathcal{G}_c$, for all sequences A_n of measurable sets, if $P_{g_1,n}(A_n) \to 0$, then $P_{g_2,n}(A_n) \to 0$. Denote

$$Z_{gni} = \frac{f_\xi([Y_i - (R_\gamma T_n g)(X_i)]/\epsilon)}{f_\xi(Y_i/\epsilon)} - 1.$$

(ii) For all $g \in \mathcal{G}_c$, $EZ_{gn1} = o(1)$. (iii) For all $g_1, g_2 \in \mathcal{G}_c$, $nEZ_{g_1n1}Z_{g_2n1} = \int g_1 g_2 + o(1)$. (iv) Finally, for all $\alpha > 0$, for all $g \in \mathcal{G}_c$,

$$n \int_{(|Z_{gn1}| > \alpha)} Z_{gn1}^2 dP = o(1).$$

□

We need to state that weak convergence of experiments implies a lower bound for the minimax risk.

Lemma 8 *Let $\mathcal{T}_n$ be the set of linear real valued estimators in regression experiment E_n and $\mathcal{T}$ the set of linear real valued estimators in Gaussian white noise experiment E. Then*

$$\begin{aligned} &\liminf_{n\to\infty} n^\kappa \inf_{\hat{\theta}_n \in \mathcal{T}_n} \sup_{g \in \mathcal{G}_c} E_{P_{g,n}} \left| \hat{\theta}_n - (T_n g)^{(\alpha_0)}(0) \right|^2 \\ &\geq f_X(0)^{-1} \epsilon^2 I_\xi^{-1} \inf_{\hat{\theta} \in \mathcal{T}} \sup_{g \in \mathcal{G}_c} E_{P_g} \left| \hat{\theta} - g^{(\alpha_0)}(0) \right|^2 \end{aligned} \tag{51}$$

where κ is the exponent in the optimal rate of convergence defined in (22).

Proof. We apply the proof of Theorem 4.1 in Klemelä (2003). We need the fact (i) that $E_n \to E$ weakly, stated in Lemma 7, and the fact (ii) that

$$(T_n g)^{(\alpha_0)}(0) = n^{-\kappa} f_X(0)^{-1/2} \epsilon I_\xi^{-1/2} \left[g^{(\alpha_0)}(0) + \delta_n(g) \right], \tag{52}$$

where $\lim_{n\to\infty} \sup_{g\in\mathcal{G}_c} |\delta_n(g)| = 0$. Due to (52) we may apply a renormalization argument for the loss function on the left hand side of (51). To prove (52) note that

$$(T_n g)^\wedge(\omega) \sim f_X(0)^{-1/2} \epsilon I_\xi^{-1/2} n^{-1/2} \kappa_n^{-d/2} \widehat{g}(\omega/\kappa_n), \ \ \omega \in \mathbf{R}^d,$$

and then

$$(T_n g)^{(\alpha_0)}(0) \sim f_X(0)^{-1/2} \epsilon I_\xi^{-1/2} n^{-1/2} \kappa_n^{r+d/2} g^{(\alpha_0)}(0),$$

uniformly in $g \in \mathcal{G}_c$. We have $n^{-1/2}\kappa_n^{r+d/2} = n^{-\kappa}$, and (52) follows. □

Let us collect the results. By Equation (50) we have that

$$\begin{aligned} &\liminf_{n\to\infty} n^\kappa \inf_{\hat\theta_n \in \mathcal{T}_n} \sup_{\rho_{\beta,p}(f)\le L} E_f \left| \hat\theta_n - f^{(\alpha_0)}(0) \right|^2 \\ &\quad \ge \liminf_{n\to\infty} n^\kappa \inf_{\hat\theta_n \in \mathcal{T}_n} \sup_{g\in\mathcal{G}_c} E_{P_{g,n}} \left| \hat\theta_n - (T_n g)^{(\alpha_0)}(0) \right|^2. \end{aligned} \tag{53}$$

By an application of Lemma 8 and by Theorem 4, and because $c > 0$ was chosen arbitrarily, we get a lower bound for the lower bound in (53):

$$\begin{aligned} &f_X(0)^{-1} \epsilon^2 I_\xi^{-1} \inf_{\hat\theta\in\mathcal{T}} \sup_{g\in\mathcal{G}_\infty} E_{P_g} \left| \hat\theta - g^{(\alpha_0)}(0) \right|^2 \\ &\quad \ge f_X(0)^{-1} \epsilon^2 I_\xi^{-1} c_{\beta,p}^2 (L')^{2-2\kappa} \kappa^\kappa (1-\kappa)^{1-\kappa} \end{aligned}$$

where L' is defined in (49). We have proved the lower bound of Theorem 5.

4 Discussion

We have constructed optimal approximations among the set of linear algorithms for the cases $p = 1$ and $p = 2$. The kernels of the algorithms are different: for the case $p = 1$ the kernel is of "Pinsker type" and for the case $p = 2$ the kernel is of "Tikhonov type". When $p > 1$, $p \neq 2$, then the algorithms are not proved to be optimal but we give tight bounds for the performance of the algorithms. For example, when $d = 1$, $r = 0$, and $\gamma = 0$, then the upper bound for the minimax risk is only 1.13 times larger than

the lower bound, over a large range of smoothness parameter values β and values of $p > 1$, see equation (23).

A motivation to study L_p Sobolev classes for small values of p, in particular for $p = 1$, comes from the fact that these classes are large nonparametric classes, but at the same time the approximation and statistical estimation is more feasible in these classes in high dimensional cases, than for the high values of p. The rate of the approximation error is independent of p, but we have studied the effect of p on the constants in Section 2.4.2, and shown that the constant in the optimal approximation error is much smaller for $p = 1$ than for $p = 2$, in high dimensional cases.

Statistical inference in discrete statistical models may asymptotically be reduced to the inference in the Gaussian white noise model, and inference in the Gaussian white noise model may be reduced in turn to the optimal recovery.

We have considered regression estimation with i.i.d. observations. To make the reduction to the Gaussian white noise model we have utilized the weak convergence of statistical experiments. The strong convergence of experiments has been proved to hold in the univariate case by Brown and Low (1996) and Nussbaum (1996) for the smooth functions, and we conjecture that the strong convergence holds in the multivariate case only for smoothness indices $\beta > d$. Thus we may get bounds under weaker conditions by utilizing weak convergence, see Remark 3.

References

Arestov, V. V. (1989), 'Optimal recovery of operators and related problems', *Proc. Steklov Inst. Math.* **189**, 3–20.

Barron, A. (1993), 'Universal approximation bounds for superpositions of a sigmoidal function', *IEEE Trans. Information Theory* **39**, 930–945.

Breiman, L. (1993), 'Hinging hyperplanes for regression, classification, and function approximation', *IEEE Trans. Inform. Theory* **39**(3), 999–1013.

Brown, L. and Low, M. (1996), 'Asymptotic equivalence of nonparametric regression and white noise', *Ann. Statist.* **24**, 2384–2398.

Donoho, D. L. (1994*a*), 'Asymptotic minimax risk for sup-norm loss: Solution via optimal recovery', *Probab. Theory Relat. Fields* **99**, 145–170.

Donoho, D. L. (1994*b*), 'Statistical estimation and optimal recovery', *Ann. Statist.* **22**, 238–270.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995), 'Wavelet shrinkage: asymptotia? (with discussion)', *J. Roy. Statist. Soc. B* **57**, 301–369.

Donoho, D. L. and Liu, R. C. (1991), 'Geometrizing rates of convergence III', *Ann. Statist.* **19**, 668–701.

Donoho, D. L. and Low, M. (1992), 'Renormalization exponents and optimal pointwise rates of convergence', *Ann. Statist.* **20**, 944–970.

Donoho, D. L. and Nussbaum, M. (1990), 'Minimax quadratic estimation of a quadratic functional', *J. Complexity* **6**, 290–323.

Efromovich, S. Y. and Low, M. (1994), 'Adaptive estimates of linear functionals', *Probab. Theory Relat. Fields* **98**, 261–275.

Efromovich, S. Y. and Low, M. (1996), 'On optimal adaptive estimation of a quadratic functional', *Ann. Statist.* **24**, 1106–1125.

Fuller, A. T. (1982), 'Optimization of stochastic relay control systems by means of dimensional analysis', *Int. J. Control* **35**, 575–604.

Gabushin, V. N. (1967), 'Inequalities for norms of functions and their derivatives in the L_p metric', *Math. Notes* **1**, 291–298.

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

Golubev, G. K. (1991), 'LAN in problems of nonparametric estimation of functions and lower bounds for quadratic risks', *Theory Probab. Appl.* **63**, 152–157.

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.

Jones, L. K. (1992), 'A simple lemma on greedy approximation in Hilbert space and convergence rate for projection pursuit regression', *Ann. Statist.* **20**(1), 608–613.

Klemelä, J. (2003), 'Lower bounds for the asymptotic minimax risk with spherical data', *J. Statist. Plann. Inference* **113**, 113–136.

Klemelä, J. (2006), 'Sharp adaptive estimation of quadratic functionals', *Probab. Theory Relat. Fields* **134**(4), 539 – 564.

Klemelä, J. and Tsybakov, A. B. (2001), 'Sharp adaptive estimation of linear functionals', *Ann. Statist.* **29**, 1567–1600.

Klemelä, J. and Tsybakov, A. B. (2004), 'Exact constants for pointwise adaptive estimation under the Riesz transform', *Probab. Theory Relat. Fields* **129**(3), 441–467.

Korostelev, A. P. (1993), 'Asymptotically minimax regression estimator in the uniform norm up to exact constant', *Theory Probab. Appl.* **38**, 775–782.

Korostelev, A. P. (1996), 'A minimaxity criterion in nonparametric regression based on large-deviations probabilities', *Ann. Statist.* **24**, 1075–1083.

Legostaeva, I. L. and Shiryayev, A. N. (1971), 'Minimax weigths in a trend detection problem of a random process', *Theory Probab. Appl.* **16**, 344–349.

Leonov, S. L. (1997), 'On the solution of an optimal recovery problem and its applications in nonparametric regression', *Math. Methods Statist.* **6**, 476–490.

Leonov, S. L. (1999), 'Remarks on extremal problems in nonparametric curve estimation', *Statist. Probab. Letters.* **43**, 169–178.

Lepski, O. and Spokoiny, V. (1997), 'Optimal pointwise adaptive methods in nonparametric estimation', *Ann. Statist.* **25**, 2512–2546.

Lepski, O. and Tsybakov, A. B. (2000), 'Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point', *Probab. Theory Relat. Fields* **117**, 17–48.

Low, M. (1992), 'Renormalization and white noise approximation for nonparametric functional estimation problems', *Ann. Statist.* **20**, 545–554.

Magaril-Il'yaev, G. G. (1983), 'Inequalities for derivatives and duality', *Proc. Steklov Inst. Math.* **161**, 199–212.

Micchelli, C. A. and Rivlin, T. J. (1977), A survey of optimal recovery, *in* C. A. Micchelli and T. J. Rivlin, eds, 'Optimal Estimation in Approximation Theory', Plenum, New York, pp. 1–54.

Nussbaum, M. (1996), 'Asymptotic equivalence of density estimation and Gaussian white noise', *Ann. Statist.* **24**, 2399–2430.

Puhalskii, A. and Spokoiny, V. (1998), 'On large-deviation efficiency in statistical inference', *Bernoulli* **4**, 203–272.

Stein, E. M. (1970), *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, New Jersey.

Sz.-Nagy, B. (1941), 'Über Integralgleichungen zwischen einer Funktion und ihrer Ableitung', *Acta Sci. Math.* **10**, 64–74.

Taikov, L. V. (1969), 'Kolmogorov type inequalities and the best formulas of numerical differentiation', *Math. Notes* **4**, 233–238.

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

Zhao, L. H. (1997), 'Minimax linear estimation in a white noise problem', *Ann. Statist.* **25**, 745–755.

Jussi Klemelä
Universität Heidelberg
Institut für Angewandte Mathematik
Im Neuenheimer Feld 294
69120 Heidelberg, Germany
Email: klemela@statlab.uni-heidelberg.de