On estimation of the L_r norm of a regression function

Lepski, O.
Humboldt University, SFB 373,
Spandauer Str. 1,
10178 Berlin Germany*

Nemirovski, A.
Technion–Israel Institute
of Technology,
Haifa 32000, Israel

Spokoiny, V.
Weierstrass Institute
Mohrenstr. 39,
10117 Berlin, Germany *

June 9, 2000

Keywords: non-smooth functionals, integral norm, rate of estimation.

1991 *Mathematics Subject Classification.* 62G07; Secondary 62G20

Running title: On estimation of the L_r norm.

The thoughtful comments of two anonymous referees, leading to significant improvement of this paper, are gratefully acknowledged. The authors thank A.Tsybakov and Yu.Golubev for important remarks and discussion.

* This research was partially supported by INTAS grant N93/1585.

Abstract

Let a Hölder continuousfunction f be observed with noise. In the present paper we study the problem of nonparametric estimation of certain nonsmooth functionals of f, specifically, L_r norms $\|f\|_r$ of f. Known from the literature results on functional estimation deal mostly with two extreme cases: estimating a smooth (differentiable in L_2) functional or estimating a singular functional like the value of f at certain point or the maximum of f. In the first case, the convergence rate typically is $n^{-1/2}$, n being the number of observations. In the second case, the rate of convergence coincides with the one of estimating the function f itself in the corresponding norm.

We show that the case of estimating $\|f\|_r$ is in some sense intermediate between the above extremes. The optimal rate of convergence is worse than $n^{-1/2}$ but is better than the rate of convergence of nonparametric estimates of f. The results depend on the value of r. For r even integer, the rate occurs to be $n^{-\beta/(2\beta+1-1/r)}$ where β is the degree of smoothness. If r is not an even integer, then the nonparametric rate $n^{-\beta/(2\beta+1)}$ can be improved, but only by a logarithmic in n factor.

1 Introduction

The problem of estimating a functional is one of the basic problems in statistical inference. Below we consider this problem in the nonparametric set-up. Let a function f be observed with noise, and our goal is to estimate a given real-valued functional $F(f)$. Clearly the quality of estimation heavily depends on smoothness properties of the functional F. The most developed theory here deals with linear functionals. The "hardest single-parameter subfamily" arguments yield both linear estimators with the smallest, as far as linear estimates are concerned, worst-case risk, and demonstrate that the resulting risk coincides, within factor $(1+o(1))$ as $n \to \infty$, with the minimax risk, see Levit (1974, 1975), Koshevnik and Levit (1976), Ibragimov and Khasminski (1979, 1987) and Donoho and Liu (1991).

Another well studied situation deals with the case of "smooth" functionals. Smoothness is usually understood as differentiability of F on L_2. It was shown in Levit (1978), Khasminski and Ibragimov (1979), Ibragimov, Nemirovski and Khasminski (1986) that if F is smooth and the underlying function f is also smooth enough then $F(f)$ can be estimated with the parametric rate of convergence $O(n^{-1/2})$, see also Ibragimov and Khasminski (1991), Birgé and Massart (1995). The problem of estimation of quadratic functionals is studied in details in Hall and Marron (1987), Bickel and Ritov (1988), Donoho and Nussbaum (1990), Fan (1991), Efroimovich and Low (1996), Laurent (1996) among others. Estimation of functionals of the type $\int f^3(t)dt$ is discussed in Kerkyacharian and Picard (1996).

The problem of estimation of non-smooth functionals is not well studied so far, and there are very few results of this sort in the literature. Ibragimov and Khasminski (1980) found the rate of convergence of estimating the maximum of f, Korostelev (1990) studied the problem of estimating the L_1 norm of f. Korostelev and Tsybakov (1994) considered some functional estimation problems in the image model, like estimating the area of an image.

In this paper we are focusing on estimating L_r norm $\|f\|_r$ with a given $r \geq 1$. It is worth to mention that at least three cases of this problem – those with $r = 1, 2$ and ∞ – have very natural interpretation. The case of $r = \infty$ corresponds to estimating the maximum of f. Ibragimov and Khasminski (1980) have shown that the convergence rate of estimating $F(f) = \|f\|_\infty$ coincides with the rate at which f itself can be recovered, the accuracy being measured in the uniform norm, and one may therefore use the plug-in estimator $\widehat{F} = \|\widehat{f}\|_\infty$ where $\widehat{f}$ is an optimal in order, with respect to the uniform norm of the error, estimate of f.

Korostelev (1990) announced similar result for estimating the L_1 norm $\|f\|_1 = \int |f(t)|dt$: the optimal rate of convergence is $O(n^{-\beta/(2\beta+1)})$, β being the order of smoothness of f, so that already a plug-in estimator $\int |\widehat{f}(t)|dt$ associated with an optimal in order, the accuracy being measured in the L_1 norm, non-parametric estimate $\widehat{f}$ of f, is optimal in order. However, the inspection of the proof shows a gap in establishing the lower bound, and a more detailed analysis demonstrates that the result itself is incorrect: when estimating the L_1 norm, a rate of convergence “slightly better” (namely, by a logarithmic in n factor) than $O(n^{-\beta/(2\beta+1)})$ is achievable.

Another interesting phenomenon occurs when estimating L_r norm for $r > 1$. It turns out that both the optimal rates of convergence and the underlying estimators heavily depend on whether r is or is not an even integer. When r is an even integer, the optimal rate of convergence is $n^{-\beta/(2\beta+1-1/r)}$, i.e., is “significantly” better than the standard non-parametric rate $n^{-\beta/(2\beta+1)}$ associated with the plug-in estimators; as about the remaining values of r, the optimal rate of convergence is only by a logarithmic in n factor better than the "plug-in" one.

It makes sense to compare the announced results with those related to a seemingly very close problem of nonparametric hypotheses testing associated with the case when the distance between the null hypothesis and the alternative set is measured in L_r norm, see Ingster (1982, 1993), Lepski and Spokoiny (1995) or Spokoiny (1996). A natural way to solve the testing problem is to estimate the L_r norm of the function in question and then use the estimate as a test statistics. This approach is known to work well for $r = 2$ and $r = \infty$. However, comparing the optimal convergence rates in the problem of nonparametric hypotheses testing and the one of estimating the L_r norms, one can see that the cases of $r = 2$ and $r = \infty$ are the only ones in which the outlined simple recipe works; for all other values of r, the convergence rates in the estimation and the testing problems differ from each other.

The rest of the paper is organized as follows. In Section 2 we state our main results, separately for r even integer and for the remaining cases. The estimation procedures for $r = 1$ and for even integer r are presented in Section 3. Section 4 contains the proofs.

2 Problem and main results

We start with formulating the problem. Consider the idealized "signal + white noise" model of observations as follows: the observed data $X(t)$, $t \in [0,1]$ is a trajectory of the stochastic differential equation

$$dX(t) = f(t)dt + n^{-1/2}dW(t) \tag{2.1}$$

where f is the unknown function, $W = (W(t), t \in [0,1])$ is the standard Wiener process, and the parameter n plays the role of the "volume of observations" (cf. more realistic models where we are given noisy observations of f at n equidistant or randomly generated points). Our a priori knowledge on f is that it possesses some smoothness, namely, belongs to Hölder class $\Sigma(\beta, L)$ with known parameters $\beta, L > 0$. Recall that the latter means that f is m times continuously differentiable on R^1, m being the largest integer which is less than β, and the m-th derivative $f^{(m)}$ of f is Hölder continuous with the exponentual $\beta - m$ and constant L:

$$|f^{(m)}(t) - f^{(m)}(s)| \le L|t-s|^{\beta-m}, \qquad t,s \in R^1.$$

By technical reasons, we assume also that f is bounded in the uniform norm by a constant $\varrho < 1$, so that

$$f \in \Sigma_\varrho(\beta, L) = \{f \in \Sigma(\beta, L) : \|f\|_\infty \le \varrho\}.$$

Our goal is to estimate the L_r norm of f

$$\|f\|_r = \left[\int_0^1 |f(t)|^r dt\right]^{1/r},$$

with a given $r \ge 1$.

We study our estimation problem in the standard asymptotic set-up, when the parameter n tends to infinity. For an estimate $\widehat{f_n}$ of $\|f\|_r$ via observation (2.1), let

$$\mathcal{R}(\widehat{f_n}) = \sup_{f \in \Sigma_\varrho(\beta,L)} \ell^{-1}\left(E\,\ell\left(\widehat{f_n} - \|f\|_r\right)\right)$$

be the worst, over f compatible with our a priori knowledge, risk of the estimate; here $\ell(\cdot)$ is a loss function. The results to follow are valid for every homogeneous loss function ℓ satisfying the standard conditions, see, e.g., Ibragimov and Khasminski (1979, Section 2.3). However, in order to simplify presentation, we prefer to restrict ourselves with the simplest case when $\ell(z) = |z|$, so that in what follows

$$\mathcal{R}(\widehat{f_n}) = \sup_{f \in \Sigma_\varrho(\beta,L)} E\left|\widehat{f_n} - \|f\|_r\right|.$$

Let also

$$\mathcal{R}^*(n) = \inf_{\widehat{f_n}} \sup_{f \in \Sigma_\varrho(\beta,L)} E\left|\widehat{f_n} - \|f\|_r,\right|$$

inf being taken over all estimates (i.e., measurable real-valued functions of observation X), be the associated minimax risk.

Our first result deals with the case of $r = 1$.

Theorem 2.1 *Let $r = 1$. There exist estimators $\widehat{f_n}$ and a positive $C > 0$ depending on β only such that for all large enough values of n one has*

$$\mathcal{R}(\widehat{f_n}) \le CL^{1/(2\beta+1)}(n \log n)^{-\beta/(2\beta+1)}. \tag{2.2}$$

This result shows that the L_1 norm can be estimated better than with the standard "nonparametric" convergence rate $O(n^{-\beta/(2\beta+1)})$, although the improvement is only by a logarithmic factor. The next result states that a more substantial improvement is impossible.

Theorem 2.2 *Assume that r is not an even integer. Then for n large enough it holds*

$$L^{-1/(2\beta+1)}(n\log n)^{\beta/(2\beta+1)}\mathcal{R}^*(n) \geq c/(\log n)^r$$

with some positive $c > 0$ depending only on β and r.

The situation with estimating L_r norm, r being an even integer, is as follows:

Theorem 2.3 *Let $r = 2k$ be an even integer. There exist positive constants c, C depending on β, r only such that for n large enough one has*

$$c \leq L^{-(1-1/r)/(2\beta+1-1/r)} n^{\beta/(2\beta+1-1/r)}\,\mathcal{R}^*(n) \leq C.$$

3 Estimation procedures

In this section we present two estimation procedures: one for estimating the L_1 norm, and the other one for estimating the L_r norm, r being an even integer.

We start with the case of $r = 1$. The idea behind the construction is as follows. The function $|t|$ is not smooth at the origin. However, it can be approximated on $[-1,1]$ by its truncated Fourier series:

$$|t| \approx \sum_{k=1}^{N} c_k \cos(\pi k t) \tag{3.1}$$

within accuracy of order of N^{-1}. Consequently, the functional $\int |f(t)|dt$ can be approximated by the finite sum

$$\sum_{k=1}^{N} c_k \int_0^1 \cos(\pi k f(t))dt \tag{3.2}$$

of smooth functionals which can be estimated with accuracy $O(n^{-1/2})$ each, e.g., by the method proposed in Ibragimov, Nemirovski and Khasminski (1986). Let $\widetilde{f}(t)$ be a proper nonparametric estimator of $f(t)$, e.g. a kernel estimator, with the variance λ. Then the estimator of $\int_0^1 \cos(\pi k f(t))dt$ can be taken in the form

$$\widehat{F}_k = E_\xi \int_0^1 \cos(\pi k(\widetilde{f}(t) + i\lambda\xi))dt = \int_0^1 \cos(\pi k \widetilde{f}(t)) \exp\{\pi^2 k^2 \lambda^2/2\} dt.$$

Here ξ is a $\mathcal{N}(0,1)$ random variable independent of our observation X and E_ξ is the expectation w.r.t. ξ. It remains to choose N in a way which balances the approximation error of (3.1) (which is the less the larger is N) and the "stochastic error" – the one of estimating the smooth functional (3.2) via noisy observations (the latter error is the larger the larger is N).

The outlined scheme can be implemented as follows. Let m be the largest integer which is smaller than β and let K be a compactly supported kernel of order m i.e., K is a continuous function satisfying the conditions

$(K.1)$ $\quad K(t) = 0$ for $|t| > 1$;

$(K.2)$ $\quad \int K(t)dt = 1$;

$(K.3)$ $\quad \int t^i K(t) = 0$ for $i = 1, \dots, m$.

We denote by $\|K\|$ the L_2 norm of K:

$$\|K\| = \sqrt{\int K^2(t)dt}. \tag{3.3}$$

Let $h \in (0,1)$ be a "bandwidth" (a parameter of the construction to be specified later), and let

$$\widetilde{f}_h(t) = \frac{1}{h}\int_0^1 K\left(\frac{t-u}{h}\right) dX(u) \tag{3.4}$$

be the standard kernel estimator of f associated with K, h. As always in the kernel estimation, the kernel K should be corrected near the endpoints $0, 1$: for $t \in [0, h]$ we should replace K in the right hand side of (3.4) by a kernel K_- vanishing outside $[0, 1]$, while for $t \in [1-h, 1]$, K should be replaced with a kernel K_+ vanishing outside $[-1, 0]$, the modified kernels satisfying the requirements (K.1) – (K.3). Without loss of generality we may assume that all three kernels $K, K_\pm$ have the same L_2 norm; with this assumption, in the constructions/proofs to follow we may use, with no risk of confusion, the same notation K for all three kernels, and we use this possibility in order to make the presentation more readable.

Due to (2.1), the estimate $\widetilde{f}_h(t)$ admits the usual decomposition into deterministic and stochastic components:

$$\widetilde{f}_h(t) = f_h(t) + \lambda_h \xi_h(t), \tag{3.5}$$

where

$$\begin{aligned}
f_h(t) &= \frac{1}{h}\int_0^1 K\left(\frac{t-u}{h}\right) f(u)du, \\
\lambda_h &= \sqrt{E\left\{\left(\frac{1}{h}\int_0^1 K\left(\frac{t-u}{h}\right) n^{-1/2} dW(u)\right)^2\right\}} = \frac{\|K\|}{\sqrt{nh}}, \\
\xi_h(t) &= \frac{1}{h\lambda_h}\int_0^1 K\left(\frac{t-u}{h}\right) n^{-1/2} dW(u) = \frac{1}{\|K\|\sqrt{h}}\int_0^1 K\left(\frac{t-u}{h}\right) dW(u).
\end{aligned}$$

$\xi_h(t)$ clearly is $\mathcal{N}(0,1)$ and hence

$$\begin{aligned}
E\,\widetilde{f}_h(t) &= f_h(t), \\
\mathrm{Var}\widetilde{f}_h(t) &\equiv E\left(\widetilde{f}_h(t) - f_h(t)\right)^2 = \lambda_h^2.
\end{aligned}$$

Let us now set

$$h = \left(L^2 n \log n\right)^{-1/(2\beta+1)}, \tag{3.6}$$

$$N = \lfloor \theta L^{-1/(2\beta+1)} (n \log n)^{\beta/(2\beta+1)} \rfloor, \tag{3.7}$$

where

$$\theta = \frac{1}{2\pi\|K\|\sqrt{2\beta+1}}.$$

For all $k = 1, 2, \dots, N$ and $\lambda > 0$, we define functions $\nu_{k,\lambda}(\cdot)$ as

$$\nu_{k,\lambda}(t) = \cos(\pi k t)\exp\{\pi^2 k^2 \lambda^2/2\} \tag{3.8}$$

and set

$$Q_{N,\lambda}(t) = c_0 + \sum_{k=1}^{N} c_k \nu_{k,\lambda}(t) \tag{3.9}$$

where c_k are the Fourier coefficients of the function $\mu(t) = |t|$:

$$c_k = 2\int_0^1 t\cos(\pi k t)dt = \begin{cases} 1, & k = 0 \\ 0, & k = 2, 4, 6, \dots, \\ 4(\pi k)^{-2}, & k = 1, 3, 5, \dots . \end{cases} \tag{3.10}$$

Finally, we define the estimator $\widehat{F}$ of $\|f\|_1$ as

$$\widehat{F}_n = \int_0^1 Q_{N,\lambda_h}(\widetilde{f}_h(t))dt = c_0 + \int_0^1 \sum_{k=1}^{N} c_k \nu_{k,\lambda_h}(\widetilde{f}_h(t))dt.$$

3.1 Estimating $\|f\|_r$ for an even integer r

The difference between this case and the previous one comes from the fact that for even integer r the function $|t|^r$ is analytic. This fact will be essentially used in the construction.

Let us first consider the functional $\Phi_r(f) = F_r^r(f)$:

$$\Phi_r(f) = \|f\|_r^r = \int_0^1 f^r(t)dt.$$

This functional is smooth and it can be estimated (under some mild conditions on f) from observations X with the convergence rate $n^{-1/2}$.

Let $\widetilde{f}_h(t)$ be the kernel estimator of f we have built. Applying the method from Ibragimov, Nemirovski and Khasminski (1986), we get the following estimator $\widehat{\Phi}_n$ of $\Phi_r(f)$:

$$\widehat{\Phi}_n = E_\xi \int_0^1 \left(\widetilde{f}_h(t) + i\lambda_h \xi\right)^r dt = \int_0^1 \sum_{j=0}^{r/2} b_{2j} \lambda_h^{2j} |\widetilde{f}_h(t)|^{r-2j} dt. \tag{3.11}$$

Here i is the imaginary unit, ξ is an $\mathcal{N}(0,1)$ random variable independent of observation X, and E_ξ is the expectation w.r.t. ξ, so that

$$b_{2j} = (-1)^j \binom{r}{2j} E_\xi \xi^{2j}. \tag{3.12}$$

Now we set (cf. (3.6))

$$h = (L^2 n)^{-\frac{1}{2\beta+1-1/r}} \tag{3.13}$$

and define the estimator $\widehat{F}_n$ of $\|f\|_r$ as

$$\widehat{F}_n = (\max\{0, \widehat{\Phi}_n\})^{1/r}.$$

Remark 3.1. Our estimate heavily exploits the fact that $|f|$ is known not to exceed a given quantity $\rho < 1$. Of course, applying the scaling $f \mapsto \text{const} f$, we can reduce to the case in question also the case when we have an a priori known upper bound of $|f|$.

4 Proofs

Below we present complete proofs of Theorems 2.1, 2.2 and 2.3. In what follows, κ (possibly, with sub- or superscripts) denote positive quantities (not necessary the same in independent proofs) depending on $r, \beta, K, K_\pm$ only.

4.1 Proof of the upper bound in Theorem 2.1

We start with several technical lemmas. Let $\nu_{k,\lambda}(t) = \cos(\pi k t)\exp\{\pi^2 k^2 \lambda^2/2\}$, $k \geq 1$, see (3.8).

Lemma 4.1 *Let* $z \in [-1,1]$, $\lambda > 0$ *and let* ξ *be* $\mathcal{N}(0,1)$ *random variable. Then for all* $k \geq 1$,

$$E\,\nu_{k,\lambda}(z+\lambda\xi) = \cos(\pi k z). \tag{4.1}$$

If $\sigma_{k,\lambda}(t)$ *is defined by*

$$\sigma^2_{k,\lambda}(t) \equiv \mathrm{Var}\nu_{k,\lambda} = E\left|\nu_{k,\lambda}(z+\lambda\xi) - \cos(\pi k z)\right|^2,$$

then

$$\sigma_{k,\lambda}(t) \leq \pi k \lambda \exp\{\pi^2 k^2 \lambda^2/2\}.$$

Proof. Let $\varphi(x) = (2\pi)^{-1/2}\exp\{-x^2/2\}$ be the standard Gaussian density. Then

$$\begin{aligned}
E\,\nu_{k,\lambda}(z+\lambda\xi) &= \int_{-\infty}^{\infty} \nu_{k,\lambda}(z+\lambda x)\varphi(x)dx \\
&= \exp\{\pi^2k^2\lambda^2/2\}\int_{\infty}^{\infty}\cos(\pi k(z+\lambda x))\varphi(x)dx \\
&= (2\pi)^{-1/2}\mathrm{Re}\left(\int_{\infty}^{\infty}\exp\{\pi^2k^2\lambda^2/2 + i\pi k(z+\lambda x) - x^2/2\}dx\right) \\
&= \mathrm{Re}\left(\exp\{i\pi k z\}\,(2\pi)^{-1/2}\int_{\infty}^{\infty}\exp\{-(x-i\pi k\lambda)^2/2\}dx\right) \\
&= \cos(\pi k z)
\end{aligned}$$

and (4.1) follows.

Similarly,

$$\begin{aligned}
\sigma^2_{k,\lambda}(t) &\equiv \int_{\infty}^{\infty} (\nu_k(z+\lambda x) - \cos(\pi k z))^2 \varphi(x) dx \\
&= \int_{\infty}^{\infty} \nu_k^2(z+\lambda x)\varphi(x)dx - \cos^2(\pi k z) \\
&= \exp\{\pi^2 k^2 \lambda^2\} \int_{\infty}^{\infty} 0.5\{1+\cos(2\pi k z + 2\pi k \lambda x)\}\varphi(x)dx - \cos^2(\pi k z) \\
&= 0.5\exp\{\pi^2 k^2 \lambda^2\} \left[1+\cos(2\pi k z)\exp\{-2\pi^2 k^2 \lambda^2\}\right] - 0.5\left[1+\cos(2\pi k z)\right] \\
&= 0.5\left[\exp\{\pi^2 k^2 \lambda^2\} - \cos(2\pi k z)\right] \cdot \left[1-\exp\{-\pi^2 k^2 \lambda^2\}\right] \\
&\le \pi^2 k^2 \lambda^2 \exp\{\pi^2 k^2 \lambda^2\}.
\end{aligned}$$

∎

Lemma 4.2 *Let $\lambda > 0$ be fixed and let $Q_{N,\lambda}$ be defined by (3.9). Then for every $z \in [-1,1]$*

$$\begin{aligned}
E\, Q_{N,\lambda}(z+\lambda\xi) &= c_0 + \sum_{k=1}^{N} c_k \cos(\pi k z), \\
\mathrm{Var} Q_{N,\lambda}(z+\lambda\xi) &\le \kappa_1^2 \lambda^2 \exp\{\pi^2 N^2 \lambda^2\} \log^2(N+1).
\end{aligned}$$

with $\kappa_1 \le 2/\pi$.

Proof. The first statement follows from the definition of $Q_{N,\lambda}$ by Lemma 4.1. Now,

$$[\mathrm{Var} Q_{N,\lambda}(z+\lambda\xi)]^{1/2} \le \sum_{k=1}^{N} c_k \left[\mathrm{Var}\nu_{k,\lambda}(z+\lambda\xi)\right]^{1/2},$$

so that by Lemma 4.1

$$\begin{aligned}
[\mathrm{Var} Q_{N,\lambda}(z+\lambda\xi)]^{1/2} &\le \sum_{k=1}^{N} c_k \pi k \lambda \exp\{\pi^2 k^2 \lambda^2/2\} \\
&\le \pi\lambda \exp\{\pi^2 N^2 \lambda^2/2\} \sum_{k=1}^{N} k c_k \\
&\le 2\lambda\pi^{-1} \exp\{\pi^2 N^2 \lambda^2/2\} \log(N+1),
\end{aligned}$$

as claimed. ∎

Lemma 4.3 *Let $c_k, k = 0, 1, \dots$ be given by (3.10). Then for every $N \geq 1$ and all $z \in [-1, 1]$ one has*

$$\left| |z| - c_0 - \sum_{k=1}^{N} c_k \cos(\pi k z) \right| \leq \kappa_2 N^{-1}$$

with $\kappa_2 = 2\pi^{-2}$.

Proof. By origin of c_k , we have for $|z| \leq 1$

$$|z| = c_0 + \sum_{k=1}^{\infty} c_k \cos(\pi k z)$$

and therefore

$$\left| |z| - c_0 - \sum_{k=1}^{N} c_k \cos(\pi k z) \right| \leq \sum_{k=N+1}^{\infty} c_k \leq \frac{1}{2} \sum_{k=N+1}^{\infty} \frac{4}{(\pi k)^2} \leq 2\pi^{-2} N^{-1}$$

as required. ∎

We are ready to prove the upper bound from Theorem 2.1. Consider decomposition (3.5) of the kernel estimate $\widetilde{f}_h(t)$ with h given by (3.6). Note first that the inclusion $f \in \Sigma(\beta, L)$ by standard reasons (see, e.g., Ibragimov and Khasminski (1979), Section 4.4, p. 317) implies that

$$|f_h(t) - f(t)| \leq \kappa_3 L h^{\beta} \tag{4.2}$$

with κ_3 depending on β and the kernels $K, K_{\pm}$ only. Since h is small for large n, from (4.2) combined with the fact that $\|f\|_\infty \leq \varrho < 1$ we conclude that for all large values of n for all $f \in \Sigma_\varrho(\beta, L)$ one has $|f_h(t)| \leq 1$. In what follows we assume that n is so large that the latter assumption is satisfied.

Let

$$\gamma_n(t) = Q_{N,\lambda_h}(\widetilde{f}_h(t))$$

so that $\widehat{F}_n = \int_0^1 \gamma_n(t) dt$. In view of decomposition (3.5) and by Lemma 4.2 as applied with $z = f_h(t)$ it holds

$$E\,\gamma_n(t) = c_0 + \sum_{k=1}^{N} c_k \cos(\pi k f_h(t)).$$

Applying Lemma 4.3 with $z = f_h(t)$ and $\lambda = \lambda_h$ and taking into account (4.2), we get

$$|E\,\gamma_n(t) - f(t)| \le |E\,\gamma_n(t) - f_h(t)| + |f_h(t) - f(t)| \le \kappa_2 N^{-1} + \kappa_3 L h^{\beta}$$

and hence

$$\left| E \int_0^1 \gamma_n(t)dt - \|f\|_1 \right| \le \int_0^1 |E\,\gamma_n(t) - f(t)| \le \kappa_2 N^{-1} + \kappa_3 L h^{\beta}.$$

Now let us bound the variance of the estimator $\widehat{F}_n$.

The definition of $\widetilde{f}_h(t)$ and the condition $(K.1)$ yield that $\widetilde{f}_h(t)$ and $\widetilde{f}_h(t')$ are independent random variables when $|t - t'| \ge 2h$. Let $\mathrm{Cov}\xi\xi'$ means the covariance $E(\xi - E\,\xi)(\xi' - E\xi')$ between two random variables ξ, ξ'. Using the Cauchy-Schwarz inequality, we get

$$\begin{aligned} \mathrm{Cov}(\gamma_n(t), \gamma_n(t')) &\le \left[\mathrm{Var}\gamma_n(t)\mathrm{Var}\gamma_n(t')\right]^{1/2} 1(|t - t'| \le 2h) \\ &\le 0.5\left(\mathrm{Var}\gamma_n(t) + \mathrm{Var}\gamma_n(t')\right) 1(|t - t'| \le 2h). \end{aligned}$$

This gives

$$\begin{aligned} \mathrm{Var}\widehat{F}_n &= \mathrm{Var}\left(\int_0^1 \gamma_n(t)dt\right) \\ &= \int_0^1 \int_0^1 \mathrm{Cov}(\gamma_n(t), \gamma_n(t'))dt\,dt' \\ &\le 0.5 \int_0^1 \int_0^1 \left(\mathrm{Var}\gamma_n(t) + \mathrm{Var}\gamma_n(t')\right) \mathbf{1}(|t - t'| \le 2h)\,dt\,dt' \\ &\le 4h \int_0^1 \mathrm{Var}\gamma_n(t)dt. \end{aligned} \tag{4.3}$$

Applying further Lemma 4.2 and recalling the origin of λ_h, we get

$$\mathrm{Var}\widehat{F}_n \le \kappa_1^2 4 \|K\|^2 n^{-1} \exp\{\pi^2 N^2 \|K\|^2/(nh)\} \log^2(N+1).$$

Now,

$$\begin{aligned} E\left|\widehat{F}_n - \|f\|_1\right| &\le E\left|E\,\widehat{F}_n - \|f\|_1\right| + E\left|\widehat{F}_n - E\,\widehat{F}_n\right| \\ &\le E\left|E\,\widehat{F}_n - \|f\|_1\right| + \left[\mathrm{Var}\widehat{F}_n\right]^{1/2} \\ &\le \kappa_2 N^{-1} + \kappa_3 L h^{\beta} \\ &\qquad + 2\kappa_1 \|K\| n^{-1/2} \log(N+1) \exp\left\{\frac{\pi^2 N^2 \|K\|^2}{2nh}\right\}. \end{aligned} \tag{4.4}$$

Substituting h, N given by (3.6), (3.7) respectively, we see that for all $n \geq \kappa_4$ it holds

$$\frac{\pi^2 N^2 \|K\|^2}{2nh} \leq \alpha \log n$$

with certain constant $\alpha < 1/(4\beta + 2)$. Therefore for all large enough values of n the exponent in (4.4) can be bounded as

$$\exp\{\pi^2 N^2 \|K\|^2/(2nh)\} \leq n^{1/(4\beta+2)};$$

with this bound, (4.4) implies (2.2).

4.2 Proof of the upper bound in Theorem 2.3

First we study the behavior of the estimator $\widehat{\Phi}_n$ of $\Phi_r(f)$, see (3.11).

Lemma 4.4 *Let $f_h(t)$ be given by (3.6). Then*

$$\begin{aligned} E\,\widehat{\Phi}_n &= \int_0^1 f_h^r(t)dt = \|f_h\|_r^r, \\ \mathrm{Var}\widehat{\Phi}_n &\leq \kappa_4 n^{-1} \max\{\lambda_h^{2r-2}, \|f_h\|_{2r-2}^{2r-2}\} \end{aligned}$$

where κ_4 depends only on r and the kernel K.

Proof. Observe first that for every two pair of independent $\mathcal{N}(0, \lambda^2)$ random variables ξ, ξ' and for every polynomial $p(\cdot)$ on the complex plane $\boldsymbol{C}$ one has

$$E(p(z + \xi + i\xi')) = p(z), \;\; z \in \boldsymbol{C}, \tag{4.5}$$

i being the imaginary unit. Indeed, our expectation can be rewritten as the mean value, over certain probability distribution on the ray $\{R \geq 0\}$, of the means $\frac{1}{2\pi}\int_0^{2\pi} p(z + R\exp\{i\phi\})\, d\phi$; all latter means are equal to $p(z)$ (the Cauchy Theorem on the integral representation of an analytic function).

Combining (4.5) and decomposition (3.5) of the kernel estimate $\widetilde{f}_t(t)$ we get

$$E\,\widehat{\Phi}_n = E\int_0^1 E_\xi\left(f_h(t) + \lambda_h \xi_h(t) + i\lambda_h \xi\right)^r dt = \int_0^1 f_h^r(t)dt, \tag{4.6}$$

which is the first assertion of Lemma.

Let

$$\gamma_n(t) = E_\xi(\widetilde{f}_h(t) + i\lambda_h \xi)^r;$$

by (4.6) we have $E\gamma_n(t) = |f_h(t)|^r$. Using (3.5) once more, we get

$$\begin{aligned}\gamma_n(t) - E\,\gamma_n(t) &= E_\xi\left(f_h(t) + \lambda_h\xi_h(t) + i\lambda_h\xi\right)^r - |f_h(t)|^r \\ &= \sum_{j=1}^{r}\binom{r}{j} f_h^{r-j}(t)\lambda_h^j E_\xi\left(\lambda_h\xi_h(t) + i\lambda_h\xi\right)^j,\end{aligned}$$

whence

$$\mathrm{Var}\gamma_n(t) \le \lambda_h^2 \sum_{j=1}^{r} a_j \lambda_h^{2j-2} |f_h(t)|^{2r-2j}$$

with some positive numbers a_j depending on r only (we have used the fact that for two independent $\mathcal{N}(0,1)$ random variables ξ', ξ'' one has $E\left[(\xi' + i\xi'')^j\,(\overline{\xi' + i\xi''})^k\right] = 0$ when $j \neq k$, $\overline{z}$ being the complex conjugate of z; to get this relation, it suffices to pass to integration in the polar coordinates, cf. (4.5)).

By exactly the same reasons which led us to (4.3) we have

$$\mathrm{Var}\widehat{\Phi}_n \le 4h \int_0^1 \mathrm{Var}\gamma_n(t)dt,$$

whence

$$\begin{aligned}\mathrm{Var}\widehat{\Phi}_n &\le 4h\lambda_h^2 \sum_{j=1}^{r} \int_0^1 a_j \lambda_h^{2j-2} |f_h(t)|^{2r-2j} dt \\ &\le 4\|K\|^2 n^{-1} \sum_{j=1}^{r} a_j \lambda_h^{2j-2} \|f_h\|_{2r-2}^{2r-2j},\end{aligned}$$

which clearly implies the second assertion of Lemma.

Lemma 4.5 *There exists a constant κ_5 depending only on r and on the kernels $K, K_\pm$ such that*

$$\|f_h\|_{2r-2}^{2r-2} \le \kappa_5 h^{-1+1/r} \|f\|_r^{r-1} \|f_h\|_r^{r-1}.$$

Proof. Applying the Minkovski inequality, we get

$$\begin{aligned}|f_h(t)|^{r-1} &= \left|\int_0^1 f(u)h^{-1}K\left(\frac{t-u}{h}\right)du\right|^{r-1} \\ &\le \left[\left(\int_0^1 |f(u)|^r du\right)^{\frac{1}{r}} \left(h^{-\frac{r}{r-1}} \int_0^1 |K\left(h^{-1}(t-u)\right)|^{\frac{r}{r-1}} du\right)^{\frac{r-1}{r}}\right]^{r-1} \\ &= \kappa_5 h^{-1+1/r} \|f\|_r^{r-1}\end{aligned}$$

where κ_5 depends on r and the kernels $K, K_\pm$ only. Consequently,

$$\begin{aligned} \|f_h\|_{2r-2}^{2r-2} &= \int_0^1 |f_h(t)|^{2r-2} dt \\ &\le \kappa_5 h^{-1+1/r} \|f\|_r^{r-1} \int_0^1 |f_h(t)|^{r-1} dt \\ &\le \kappa_5 h^{-1+1/r} \|f\|_r^{r-1} \|f_h\|_r^{r-1}, \end{aligned}$$

the concluding $\le$ being given by the Jensen inequality. ∎

Now we are ready to complete the proof of the theorem. Denote

$$\varrho_n = L^{\frac{1-1/r}{2\beta+1-1/r}} n^{-\frac{\beta}{2\beta+1-1/r}}. \tag{4.7}$$

Then ϱ_n is exactly the convergence rate mentioned in the theorem; note that by (3.13) one has $\varrho_n = Lh^\beta$.

Recall that the Hölder smoothness constraint implies the bound

$$\|f - f_h\|_r \le \kappa_3 L h^\beta = \kappa_3 \varrho_n, \tag{4.8}$$

whence $\|f_h\|_r \le \|f\|_r + \kappa_3 \varrho_n$. Consider separately the cases of $\|f\|_r \le 2\kappa_3 \varrho_n$ and of $\|f\|_r > 2\kappa_3 \varrho_n$. If $\|f\|_r \le 2\kappa_3 \varrho_n$, then

$$\begin{aligned} E|\widehat{F}_n - \|f\|_r| &\le E|\widehat{F}_n| + 2\kappa_3 \varrho_n \\ &\le (E\,\widehat{\Phi}_n^2)^{1/(2r)} + 2\kappa_3 \varrho_n \\ &\le [\mathrm{Var}\widehat{\Phi}_n + (E\,\widehat{\Phi}_n)^2]^{1/(2r)} + 2\kappa_3 \varrho_n \\ &\le (\mathrm{Var}\widehat{\Phi}_n)^{1/(2r)} + (E\,\widehat{\Phi}_n)^{1/r} + 2\kappa_3 \varrho_n. \end{aligned}$$

It is easily seen that for $n \ge \kappa_4$ one has $\varrho_n < \lambda_h^2 = \|K\|^2/(nh)$. Using Lemma 4.4, we come to the bound

$$\begin{aligned} E|\widehat{F}_n - \|f\|_r| &\le (\kappa_5 n^{-1} \lambda_h^{2r-2})^{1/(2r)} + \|f_h\|_r + 2\kappa_3 \varrho_n \\ &\le (\kappa_5 n^{-1} \lambda_h^{2r-2})^{1/(2r)} + [\|f\|_r + \kappa_3 \varrho_n] + 2\kappa_3 \varrho_n \\ &\le (\kappa_5 n^{-1} \lambda_h^{2r-2})^{1/(2r)} + \kappa_6 \varrho_n. \end{aligned}$$

Substituting $\lambda_h = (nh)^{-1/2}$ in the expression for h from (3.13), and using the bound (4.8), we get the desired risk bound.

It remains to consider the case of $\|f\|_r > 2\kappa_3\varrho_n$. In this case from (4.8) it follows that $\|f_h\|_r \geq \|f\|_r - \kappa_3\varrho_n \geq \kappa_3\varrho_n$ whence

$$\begin{aligned}
E|\widehat{F}_n - \|f\|_r| &\leq E|\widehat{F}_n - \|f_h\|_r| + \kappa_3\varrho_n \\
&\leq \frac{E|\widehat{F}_n^r - \|f_h\|_r^r|}{\|f_h\|_r^{r-1}} + \kappa_3\varrho_n \\
&\leq \frac{E|\widehat{\Phi}_n - E\,\widehat{\Phi}_n|}{\|f_h\|_r^{r-1}} + \kappa_3\varrho_n \\
&\leq \frac{(\mathrm{Var}\widehat{\Phi}_n)^{1/2}}{\|f_h\|_r^{r-1}} + \kappa_3\varrho_n.
\end{aligned}$$

Combining Lemma 4.4 and (4.8), we get

$$(\mathrm{Var}\widehat{\Phi}_n)^{1/2} \leq \kappa_6 n^{-1/2}(\lambda_h^{r-1} + h^{-(r-1)/(2r)}\|f_h\|_r^{r-1})$$

and we end up with

$$E|\widehat{F}_n - \|f\|_r| \leq \kappa_6 n^{-1/2}(\lambda_h^{r-1}\varrho_n^{-r+1} + h^{-(r-1)/(2r)}) + \kappa_3\varrho_n.$$

Recalling that $\lambda_h = \|K\|^2/(nh)$ and substituting the expression for h, we come to the desired risk bound.

4.3 Proof of the lower bound in Theorem 2.3

The problem under consideration is rather special, and the standard techniques for establishing lower bounds in the problems of estimating the value of a functional (e.g., the one of "the hardest single-parametric subfamily") seemingly do not work. The reason is that the functional $\|f\|_r$, r being an even integer, is "nearly smooth" – it looses smoothness at the unique point $f = 0$. Note that the value of an "actually smooth" functional can be estimated with the parametric convergence rate $O(n^{-1/2})$, while our goal is to establish a kind of nonparametric lower bound. To this end we intend to build a pair of "high-dimensional" distributions concentrated each near its own small "r-sphere" $\{f \in \Sigma_\varrho(\beta, L) \mid \|f\|_r = R\}$, $R = R_1, R_2$, in such a way that the Kullback distance between the distributions is small, so that they cannot be distinguished reliably from the observations. Ensuring this property, we can use the standard arguments to demonstrate that the minimax risk in our problem of estimating $\|\cdot\|_r$ is (at least) $O(|R_1 - R_2|)$.

Our first step is to replace the nonparametric set $\Sigma_\varrho(\beta, L)$ with its properly chosen parametric subset where the aforementioned distributions will be concentrated. Let us fix a function $g \in \Sigma(\beta, 1)$ vanishing outside the interval $[0,1]$ and such that $\|g\|^2(t)dt = \int g^2 > 0$. Note that by evident reasons all functions of the form $Lb^{-\beta}g(a+bt)$ with $b \geq 1$ belong to $\Sigma_\varrho(\beta, L)$, provided that b is greater than a constant depending on ϱ only.

Let us set

$$\begin{aligned} N &= \lfloor (L^2 n)^{\frac{1}{2\beta+1-1/r}} \rfloor, \\ h &= N^{-1}; \end{aligned} \tag{4.9}$$

note that our new values of N, h differ from those used in the construction of the estimators $\widehat{F}_n$.

Now let $\mathcal{I} = \{I_i,\, i = 1, \dots, N\}$ be the partition of the interval $[0,1]$ into $N = h^{-1}$ subintervals $I_1, ..., I_N$ of length h each, and let t_i be the left endpoint of subinterval I_i. With a point $\theta = (\theta_1, \dots, \theta_N)$ from the N-dimensional cube $B_N = [-1,1]^N$ we associate the function

$$f_\theta(t) = L \sum_{i=1}^{N} \theta_i h^\beta g((t - t_i)/h)$$

Assuming n large enough, for all $\theta \in B_N$ we have $f_\theta \in \Sigma_\varrho(\beta, L)$ and

$$\|f_\theta\|_r^r \;=\; L^r h^{\beta r} \sum_{i=1}^{N} |\theta_i|^r \int_{I_i} \left| g\left(\frac{t - t_i}{h}\right) \right|^r dt = \left(L \|g\|_r h^\beta F_r(\theta) \right)^r \tag{4.10}$$

where

$$F_r(\theta) = \left(\frac{1}{N} \sum_{i=1}^{N} |\theta_i|^r \right)^{1/r}. \tag{4.11}$$

For $i = 1, \dots, N$ let

$$Y_i = Y_i^\theta = \frac{\sqrt{n}}{\|g\|\sqrt{h}} \int_{I_i} g\left(\frac{t - t_i}{h}\right) dX^\theta(t),$$

where X^θ is observation (2.1) associated with $f = f_\theta$. We clearly have

$$Y_i = \alpha(N)\theta_i + \xi_i, \qquad i = 1, \dots, N, \tag{4.12}$$

where

$$\begin{aligned} \alpha(N) &= L\|g\| n^{1/2} h^{\beta+1/2} = L\|g\| n^{1/2} N^{-\beta-1/2}, \\ \xi_i &= \frac{1}{\|g\|\sqrt{h}} \int_{I_i} g\left(\frac{t - t_i}{h}\right) dW(t). \end{aligned} \tag{4.13}$$

Clearly $\xi = (\xi_1, \dots, \xi_N)$ is a collection of independent $\mathcal{N}(0,1)$ random variables. It is also straightforward to see that the set of statistics $Y_i, i = 1, \dots, n$ is sufficient for the parametric submodel (with $f \in \Sigma^N = \{f_\theta, \theta \in B_N\}$). Therefore, when restricting f to belong to Σ^N and setting $s_i = \alpha(N)\theta_i$, $i = 1, \dots, N$, the original "signal + white noise" model (2.1) becomes the "sequence space" model

$$Y_i = s_i + \xi_i, \qquad i = 1, \dots, N, \tag{4.14}$$

with $s = (s_1, \dots, s_N)$ from the cube $S_N = B_N^{\alpha(N)} = [-\alpha(N), \alpha(N)]^N$. With this transformation, the original estimation problem (reduced to Σ^N) becomes the problem of estimating the quantity

$$F_r(s) = \left(\frac{1}{N} \sum_{i=1}^N |s_i|^r \right)^{1/r}$$

(cf. (4.11)) via observations (4.14). Let $\mathcal{R}_s(N)$ be the corresponding minimax risk:

$$\mathcal{R}_s(N) = \inf_{\widehat{F}} \sup_{s \in S_N} E_s|\widehat{F} - F_r(s)|,$$

the infimum being taken over all Borel functions $\widehat{F} = \widehat{F}(y)$ on R^N and E_s being the expectation over the observations (4.14) associated with a given s. Comparing (4.11) and the definition of $F_r(s)$ and taking into account (4.10), we get

$$\mathcal{R}^*(n) \geq L\|g\|_r h^\beta \alpha^{-1}(N) \mathcal{R}_s(N) = \kappa_g \sqrt{N/n}\, \mathcal{R}_s(N) \tag{4.15}$$

where $\kappa_g = \|g\|_r / \|g\|$.

Now we are going to establish the following

Proposition 4.1 *For all large enough values of N one has*

$$\mathcal{R}_s(N) \geq \kappa_7 \alpha(N) \tag{4.16}$$

with κ_7 depending on r, β only.

Note that the statement of Theorem 2.3 is an immediate consequence of Proposition 4.1. Indeed, combining (4.16), (4.9), (4.15) and (4.13), we get

$$\mathcal{R}^*(n) \geq \kappa_7 \kappa_g \sqrt{N/n}\alpha(N) = \kappa_7 \kappa_g L \|g\| N^{-\beta} = \kappa_8 L^{\frac{1-1/r}{2\beta+1-1/r}} n^{-\frac{\beta}{2\beta+1-1/r}}$$

with κ_8 depending on r, β only, as claimed in Theorem 2.3.

Proof of Proposition 4.1 is based on the following idea. We introduce two prior measures $\mu_{N,0}$ and $\mu_{N,1}$ on the parameter set S_N and denote by $P_{N,0}$ and $P_{N,1}$ the corresponding marginal measures on R^N,

$$P_{N,j} = \mu_{N,j} * \mathcal{L}, \qquad j = 0, 1;$$

here $\mathcal{L}$ is the distribution of the observation noises ξ in (4.14). Let also $\mathcal{K}(P_{N,0}, P_{N,1})$ be the Kullback distance between $P_{N,0}$ and $P_{N,1}$

$$\mathcal{K}(P_{N,0}, P_{N,1}) = \int \log\left(\frac{dP_{N,1}}{dP_{N,0}}\right) dP_{N,1}.$$

We will bound the minimax risk from below by the maximum of two Bayesian risks corresponding to the distributions $\mu_{N,0}$ and $\mu_{N,1}$ on the space S_N of "signals" s. To this end we need the following statement (which can be obtained from the Fano inequality; we, however, prefer to present a direct proof).

Lemma 4.6 *Let prior measures $\mu_{N,0}$ and $\mu_{N,1}$ be such that the Kullback distance $\mathcal{K}(P_{N,0}, P_{N,1})$ satisfies the condition*

$$\mathcal{K}(P_{N,0}, P_{N,1}) \le \Omega \tag{4.17}$$

with some positive Ω. Let Φ be a function on the parametric set S_N, and let

$$\begin{aligned} v_{N,j} &= \int \Phi(s)\mu_{N,j}(ds), && (4.18)\\ d^2_{N,j} &= \int (\Phi(s) - v_{N,j})^2 \mu_{N,j}(ds), && (4.19) \end{aligned}$$

for $j = 0, 1$. One has

$$R(N) \equiv \inf_{\widehat{\Phi}} \sup_{s \in S_N} E_s|\widehat{\Phi} - \Phi(s)| \ge 0.25|v_{N,0} - v_{N,1}|e^{-\Omega} - \max\{d_{N,0}, d_{N,1}\}, \tag{4.20}$$

the infimum being taken over all estimators of $\Phi(s)$ via observations (4.14).

Proof. First note that for an arbitrary prior measure μ and every estimator $\widehat{\Phi}$ of $\Phi(s)$ via observations (4.14) one has

$$\begin{aligned} \sup_{s \in S_N} E_s|\widehat{\Phi} - \Phi(s)| &\ge E_{N,\mu}|\widehat{\Phi} - \Phi(s)| \\ &\ge E_{N,\mu}|\widehat{\Phi} - E_{N,\mu}\Phi(s)| - E_{N,\mu}|\Phi(s) - E_{N,\mu}\Phi(s)| \\ &\ge E_{N,\mu}|\widehat{\Phi} - E_{N,\mu}\Phi(s)| - d_{N,\mu}. \end{aligned}$$

It follows that

$$\begin{aligned} R(N) &\geq 0.5 \inf_{\widehat{\Phi}} \Big\{ E_{N,0}|\widehat{\Phi} - v_{N,0}| - d_{N,0} + E_{N,1}|\widehat{\Phi} - v_{N,1}| - d_{N,1} \Big\} \\ &\geq 0.5 \inf_{\widehat{\Phi}} \Big\{ E_{N,0}|\widehat{\Phi} - v_{N,0}| + E_{N,1}|\widehat{\Phi} - v_{N,1}| \Big\} - \max\{d_{N,0}, d_{N,1}\}. \qquad (4.21) \end{aligned}$$

Now let us use the well known fact (see e.g. Borovkov (1984, Theorem 2.1, Chapter 3)) that the maximum likelihood test $\widehat{T}_N = \mathbf{1}(dP_{N,1}/dP_{N,0} > 1)$ is optimal for testing the hypothesis $H_0 : P = P_{N,0}$ versus the alternative $H_1 : P = P_{N,1}$ (P is the distribution of observations (4.14)) in the sense that it minimizes the sum of probabilities of errors: for an arbitrary test T_N ,

$$P_{N,0}(T_N = 1) + P_{N,1}(T_N = 0) \geq P_{N,0}(\widehat{T}_N = 1) + P_{N,1}(\widehat{T}_N = 0). \qquad (4.22)$$

Let $Z_N = dP_{N,0}/dP_{N,1}$. Then $\widehat{T}_N = \mathbf{1}(Z_N \leq 1)$ and, since the function $\log(z)$ is concave, using Jensen's inequality we get

$$\begin{aligned} &\log\Big(P_{N,0}(\widehat{T}_N = 1) + P_{N,1}(\widehat{T}_N = 0)\Big) \\ &\geq \log P_{N,0}(Z_N \leq 1) \\ &= \log \int Z_N \mathbf{1}(Z_N \leq 1) dP_{N,1} \\ &\geq \int \log(Z_N)\mathbf{1}(\log(Z_N) \leq 0) dP_{N,1} \\ &\geq -\mathcal{K}(P_{N,0}, dP_{N,1}) \geq -\Omega. \qquad (4.23) \end{aligned}$$

Let now $\widehat{\Phi}$ be an estimator of $\Phi(s)$. Consider the following test

$$T_N = \mathbf{1}(\widehat{\Phi} - v_{\mu,0} > \Delta_N)$$

where

$$\Delta_N = (v_{N,1} - v_{N,0})/2$$

(we assume that $v_{N,1} > v_{N,0}$). Applying (4.22) and (4.23), we get

$$P_{N,0}(T_N = 1) + P_{N,1}(T_N = 0) \geq e^{-\Omega}$$

or

$$P_{N,0}(\widehat{\Phi} - v_{N,0} > \Delta_N) + P_{N,1}(\widehat{\Phi} - v_{N,1} < -\Delta_N) \geq e^{-\Omega}.$$

Since

$$\begin{aligned}
&E_{N,0}|\widehat{\Phi} - v_{N,0}| + E_{N,1}|\widehat{\Phi} - v_{N,1}| \\
&\geq \ \Big(P_{N,0}(\widehat{\Phi} - v_{N,0} > \Delta_N) + P_{N,1}(\widehat{\Phi} - v_{N,1} < -\Delta_N)\Big)\,|\Delta_N| \\
&\geq \ 0.5|v_{N,1} - v_{N,0}|e^{-\Omega},
\end{aligned}$$

(4.21) implies (4.20). ∎

We shall apply Lemma 4.6 to the function $\Phi(s) = N^{-1}(s_1^r + \ldots + s_N^r)$ and a pair of prior measures $\mu_{N,0}$ and $\mu_{N,1}$ with the product structure:

$$\begin{aligned}
\mu_{N,0} &= \mu_0^N, \\
\mu_{N,1} &= \mu_1^N.
\end{aligned}$$

We shall build the measures μ_0, μ_1 on $[-\alpha(N), \alpha(N)]$ in such a way that (4.17) holds with some fixed Ω, while and the difference $|v_{N,1} - v_{N,0}|$ is “large”.

First we note that, for $j = 0, 1$,

$$v_{N,j} = \frac{1}{N}\int \sum_{i=1}^{N} |s_i|^r \mu_{N,j}(ds) = \int |s|^r \mu_j(ds) = v_j \tag{4.24}$$

and similarly

$$d_{N,j}^2 \ = \ \frac{1}{N^2}\int \sum_{i=1}^{N}(|s_i|^{2r} - v_j^2)\mu_{N,j}(ds) = N^{-1}\int (|s|^{2r} - v_j^2)\mu_j(ds) = N^{-1}d_j^2$$

where

$$\begin{aligned}
v_j &= \int |s|^r \mu_j(ds) \leq \alpha^r(N) \\
d_j^2 &= \int |s|^{2r}\mu_j(ds) - v_j^2 \leq \alpha^{2r}(N).
\end{aligned} \tag{4.25}$$

To bound the Kullback distance between the marginal measures $P_{N,0}$ and $P_{N,1}$, note that the product structure of model (4.14) and of the priors $\mu_{N,0}, \mu_{N,1}$ altogether imply that

$$\mathcal{K}(P_{N,0},P_{N,1}) = N\int \log(p_{\mu_0}(y)/p_{\mu_1}(y))p_{\mu_0}(y)dy \tag{4.26}$$

where, for a finitely supported measure μ on the axis,

$$p_\mu(y) = \int \varphi(y-t)\mu(dt),$$

$$\varphi(y) = (2\pi)^{-1}\exp\{-y^2/2\}$$

being the standard Gaussian density on the axis.

Assuming that the priors $\mu_{N,0} = \mu_0^N$, $\mu_{N,1} = \mu_1^N$ and an $\Omega > 0$ satisfy (4.17) and applying Lemma 4.6, we get the following lower bound on the risk of an arbitrary estimate $\widehat{\Phi}$ of $\Phi(s)$:

$$\sup_{s\in S_N} E_s|\widehat{\Phi} - \Phi(s)| \geq 0.25|v_1 - v_0|e^{-\Omega} - \alpha^r(N)N^{-1/2} \tag{4.27}$$

(see 4.25).

Now let us derive from the latter bound a lower bound for the risk $\mathcal{R}_s(N)$ of estimating $F_r(s)$. Let $\widehat{F}$ be an estimate of $F_r(s)$, $s \in S_N$. When bounding from below the risk of $\widehat{F}$ on S_N, we may assume without loss of generality that $|\widehat{F}(\cdot)| \leq \alpha(N)$. Indeed, since $|F_r(s)| \leq \alpha(N)$ for $s \in S_N$, we only decrease the risk of $\widehat{F}$ at $s \in S_N$ when passing from $\widehat{F}$ to the "projected" estimate $\psi(\widehat{F}(\cdot))$, where

$$\psi(t) = \begin{cases} -\alpha(N), & t \leq -\alpha(N), \\ t, & -\alpha(N) \leq t \leq \alpha(N), \\ \alpha(N), & t \geq \alpha(N). \end{cases}$$

Let $\widehat{\Phi} = \widehat{F}^r$ be the estimate of $\Phi(s) = F_r^r(s)$ induced by $\widehat{F}$. Since $|\widehat{F}| \leq \alpha(N)$, we have

$$E_s|\widehat{\Phi} - \Phi(s)| = E_s|\widehat{F}^r - F_r^r(s)| \leq r\alpha^{r-1}(N)E_s|\widehat{F} - F_r(s)|.$$

Applying (4.27), we get

$$\begin{aligned} \mathcal{R}_s(N) &\geq (r\alpha^{r-1}(N))^{-1}(0.25|v_1 - v_0|e^{-\Omega} - \alpha^r(N)N^{-1/2}) \\ &= r^{-1}\alpha(N)(0.25\alpha^{-r}(N)|v_1 - v_0|e^{-\Omega} - N^{-1/2}). \end{aligned} \tag{4.28}$$

It is time now to specify our choice of the measures μ_0, μ_1. Let δ be the distance (in the uniform norm on $[-1,1]$) from the function t^r to the space L_{r-2} of polynomials of degree $\le r-2$. We claim that there exists a measure μ on $[-1,1]$ of variation 2 such that $\int t^l \mu(dt) = 0$ for $l = 0,1,...,r-2$, while $\int t^r \mu(dt) = 2\delta$. The justification of our claim is quite standard. Consider the space $C(-1,1)$ of continuous real-valued functions on $[-1,1]$ (equipped with the uniform norm) along with its finite-dimensional subspace L spanned by L_{r-2} and the polynomial t^r. L is a finite-dimensional linear space equipped with the norm $\|\cdot\|$ inherited from $C(-1,1)$, and L_{r-2} is a linear subspace in L of codimension 1. Let the linear functional $\psi(\cdot)$ on L be defined by the requirements that ψ vanishes on L_{r-2} and is equal to δ at t^r. Observe that the norm of our functional is 1:

$$\|\psi\|_* \equiv \max\{\psi(q(\cdot)) \mid q(\cdot) \in L, \|q\| \le 1\} = 1.$$

Indeed, if $q(\cdot)$ is the closest to t^r element of L_{r-2}, then $\psi(t^r - q(\cdot)) = \delta = \|t^r - q(\cdot)\|$, so that $\|\psi\|_* \ge 1$. On the other hand, assuming that $\|\psi\|_* > 1$, we are able to find $d \in L$ with $\|d\| = 1$ and $\psi(d) = \|\psi_*\| > 1$; the vector $t^r - (\delta/\|\psi\|_*)d \in L$ belongs to L_{r-2} (since the value of ψ at this vector is 0) and is at a smaller than δ $\|\cdot\|$-distance from t^r, which is impossible.

By the Hahn-Banach Theorem, we can extend the linear functional ψ from L on the entire $C(-1,1)$ not increasing the norm of the functional, and by the Riesz Theorem, the resulting linear functional $\widehat{\psi}(g)$ on $C(-1,1)$ can be represented as

$$\widehat{\psi}(g) = \int_{-1}^{1} g(t) d\nu(t)$$

for a Borel (not necessarily nonnegative) measure ν with variation equal to the norm of $\widehat{\psi}$, i.e., to 1.

Setting $\mu = 2\nu$, we get a measure on $[-1,1]$ of variation 2 such that

$$\int_{-1}^{1} t^l \mu(dt) = 0, \qquad l = 0,1,...,r-2, \qquad\qquad \int_{-1}^{1} t^r \mu(dt) = 2\delta.$$

Note that if μ possesses the indicated properties, so is the "reflected" measure μ^* ($\mu^*(A) = \mu(-A)$) and hence the measure $(\mu + \mu^*)/2$; therefore μ may be assumed to be symmetric. Let $\mu_+, -\mu_-$ be the positive and the negative components of μ, respectively. Since μ

is symmetric with variation 2 and $\int_{-1}^{1}\mu(dt) = \int_{-1}^{1} t^0\mu(dt) = 0$, both μ_+ and μ_- are symmetric probability distributions on $[-1,1]$ such that

$$\begin{aligned} \int_{-1}^{1} t^l \mu_+(dt) &= \int_{-1}^{1} t^l \mu_-(dt), \ l = 0,1,...,r-2; \\ \int_{-1}^{1} t^r \mu_+(dt) &= \int_{-1}^{1} t^r \mu_-(dt) + 2\delta. \end{aligned} \tag{4.29}$$

Let μ_0, μ_1 be obtained from $\mu_\pm$ by "expanding" associated with the similarity transformation which maps $[-1,1]$ onto $[-\alpha(N), \alpha(N)]$: $\mu_0(A) = \mu_+(\alpha^{-1}(N)A)$, $\mu_1(A) = \mu_-(\alpha^{-1}(N)A)$, $A \subset [-\alpha(N), \alpha(N)]$. The quantities v_0, v_1 associated with our μ_0, μ_1 (see (4.24)) clearly satisfy the relation

$$v_0 - v_1 = \alpha^r(N)\int_{-1}^{1} |t|^r \mu(dt) = 2\delta\alpha^r(N)$$

and the associated bound (4.28) is

$$\mathcal{R}_s(N) \geq r^{-1}\alpha(N)(\delta e^{-\Omega} - N^{-1/2}), \tag{4.30}$$

Ω being the Kullback distance between the marginal distributions $P_{N,0}, P_{N,1}$ given by the priors μ_0^N, μ_1^N. All we need is to evaluate Ω.

Let us associate with a symmetric probability distribution ν on $[-1,1]$ and a real α the distribution F_ν^α on the axis with the density

$$p_\nu(\alpha, y) = \int_{-1}^{1} \varphi(y - \alpha t)\nu(dt) = \varphi(y)\int_{-1}^{1} \mathrm{ch}(\alpha t y)\exp\{-\alpha^2 t^2/2\}\nu(dt), \tag{4.31}$$

so that

$$p_{\mu_0}(y) = p_{\mu_+}(\alpha(N), y), \qquad p_{\mu_1}(y) = p_{\mu_-}(\alpha(N), y). \tag{4.32}$$

Note that (4.31) defines function $p_\nu(\alpha, y)$ for an arbitrary (not necessarily nonnegative) symmetric measure ν on $[-1,1]$.

Let

$$\mathcal{K}(\alpha) = \int_{-\infty}^{\infty} \log(p_{\mu_+}(\alpha, y)/p_{\mu_-}(\alpha, y))p_{\mu_+}(\alpha, y)dy \tag{4.33}$$

be the Kullback distance from $p_{\mu_+}(\alpha, \cdot)$ to $p_{\mu_-}(\alpha, \cdot)$. Note that by (4.26) and (4.32) it holds

$$\Omega = \mathcal{K}(P_{N,0}, P_{N,1}) = N\mathcal{K}(p_{\mu_0}, p_{\mu_1}) = N\mathcal{K}(\alpha(N)). \tag{4.34}$$

Lemma 4.7 *The function $\mathcal{K}(\alpha)$ is C^∞ smooth and it has a zero of order at least $2r$ at the point $\alpha = 0$.*

Proof. It is clearly seen that one may differentiate $\mathcal{K}(\alpha)$ arbitrarily many times and that

$$\mathcal{K}^{(l)}(\alpha) = \int_{-\infty}^{\infty} \frac{\partial^l}{\partial \alpha^l} \left[\log \left(\frac{p_{\mu_+}(\alpha, y)}{p_{\mu_-}(\alpha, y)} \right) p_{\mu_+}(\alpha, y) \right] dy$$

for all l. Note that

$$p_{\mu_+}(\alpha, y) = p_{\mu_-}(\alpha, y) + p_\mu(\alpha, y).$$

Let us first demonstrate that for all x

$$\left. \frac{\partial^l p_\mu(\alpha, y)}{\partial \alpha^l} \right|_{\alpha=0} = 0, \qquad l = 0, 1, ..., r-1. \tag{4.35}$$

Indeed, one has

$$\begin{aligned} & \left. \frac{\partial^l p_\mu(\alpha, y)}{\partial \alpha^l} \right|_{\alpha=0} \\ = \quad & \varphi(x) \int_{-1}^{1} \left[\sum_{i=0}^{l} \binom{l}{i} \left(\frac{\partial^i \exp\{-\alpha^2 t^2/2\}}{\partial \alpha^i} \right) \left(\frac{\partial^{l-i} \mathrm{ch}(\alpha t y)}{\partial \alpha^{l-i}} \right) \right] \mu(dt) \Bigg|_{\alpha=0} \\ = \quad & \int_{-1}^{1} t^l (a_0 + a_1 y + \ldots + a_l y^l) \mu(t) = 0 \end{aligned}$$

(we have used (4.29)), as required in (4.35).

According to (4.35), $p_\mu(\alpha, y)$ can be represented in the form

$$p_\mu(\alpha, y) = \alpha^r w(\alpha, y)$$

with smooth function $w(\cdot, \cdot)$ (which, as it is easily seen, is a summable function of y). Since $\int_{-\infty}^{\infty} p_\mu(\alpha, y) dy = 0$ for all α, it also is the case for $w(\alpha, y)$:

$$\int_{-\infty}^{\infty} w(\alpha, y) dy = 0, \qquad \forall \alpha.$$

Now we have

$$\log \left(\frac{p_{\mu_-}(\alpha, y)}{p_{\mu_+}(\alpha, y)} \right) \quad = \quad \log \left(1 - \frac{\alpha^r w(\alpha, y)}{p_{\mu_+}(\alpha, y)} \right) = -\frac{\alpha^r w(\alpha, y)}{p_{\mu_+}(\alpha, y)} - \alpha^{2r} v(\alpha, y),$$

v being a smooth function of y, α. Hence

$$\begin{aligned} \mathcal{K}(\alpha) \quad &= \quad -\int_{-\infty}^{\infty} \log \left(\frac{p_{\mu_-}(\alpha, y)}{p_{\mu_+}(\alpha, y)} \right) p_{\mu_+}(\alpha, y) dy \\ &= \quad \alpha^r \int_{-\infty}^{\infty} w(\alpha, y) dy + \alpha^{2r} \int_{-\infty}^{\infty} v(\alpha, y) p_{\mu_+}(\alpha, y) dy \\ &= \quad \alpha^{2r} \int_{-\infty}^{\infty} v(\alpha, y) p_{\mu_+}(\alpha, y) dy \end{aligned}$$

and the assertion of Lemma follows.

The result of Lemma 4.7 says that for small positive α one has

$$\mathcal{K}(\alpha) \leq \kappa_{10}\alpha^{2r}. \tag{4.36}$$

In particular, for all large enough values of n (and thus – of N) we have

$$\begin{array}{rcll} \Omega & = & N\mathcal{K}(\alpha(N)) & [\text{by } (4.34)] \\ & \leq & \kappa_{10}N\alpha^{2r}(N) & [\text{by } (4.36)] \\ & \leq & \kappa_{11}N(Ln^{1/2}N^{-\beta-1/2})^{2r} & [\text{see } (4.13)] \\ & \leq & \kappa_{12} & [\text{see } (4.9)] \end{array}$$

Applying (4.30), we see that for n large enough it holds

$$\mathcal{R}_s(N) \geq \kappa_{13}\alpha(N),$$

as required in Proposition 4.1.

4.4 Proof of the lower bound in Theorem 2.2

Here we establish the lower bound from Theorem 2.2 for the case when r is not an even integer. We follow the line of the proof of the lower bound from Theorem 2.3; the only difference is in construction of the priors μ_0 and μ_1.

We start with translating the problem into the "sequence space" model in exactly the same manner as in Section 4.3, with the only difference that now we set

$$N = \lfloor (200L\|g\|)^{2/(2\beta+1)}(n\log n)^{1/(2\beta+1)} \rfloor. \tag{4.37}$$

Note that with this setup for all large enough values of n one has (see (4.13))

$$\alpha(N) \equiv L\|g\|\sqrt{n}N^{-\beta-1/2} \leq \frac{0.01}{\sqrt{\log N}}. \tag{4.38}$$

Relation (4.15) for $\mathcal{R}^*(n)$ remains valid for our new setup as well, and the required result is obtained from this relation and a lower bound on the worst case, over $s \in S_N$, risk of recovering the functional $F_r(s) = (N^{-1}(s_1^r + \ldots + s_N^r))^{1/r}$ via observations (4.14). The latter bound is given by the following statement (which now plays the role of Proposition 4.1):

Proposition 4.2 *For all large enough values of N one has*

$$\mathcal{R}_s(N) \equiv \inf_{\widehat{F}} \sup_{s\in S_N} E_s|\widehat{F} - F_r(s)| \geq \kappa_9(\log N)^{-r}\alpha(N) \tag{4.39}$$

where $\kappa_9 > 0$ depends on r and β only.

Postponing for the moment proof of Proposition, let us derive from this statement Theorem 2.2. Indeed, we have

$$\begin{aligned}
\mathcal{R}^*(n) &\geq \kappa_g\sqrt{N/n}\,\mathcal{R}_s(N) && \text{[by (4.15)]}\\
&\geq \kappa_9\kappa_g\sqrt{N/n}(\log N)^{-r}\alpha(N) && \text{[by (4.39)]}\\
&\geq \kappa_{10}L\|g\|N^{-\beta}(\log N)^{-r} && \text{[by (4.13)]}\\
&\geq \kappa_{11}L^{1/(2\beta-1)}(n\log n)^{-\beta/(2\beta+1)}(\log n)^{-r} && \text{[by (4.37)]}
\end{aligned}$$

with κ_{11} depending on β, r only, as required in Theorem 2.2.

Proof of Proposition 4.2 differs from the one of Proposition 4.1 only in how we define the measures $\mu_\pm$. Let $\mathcal{P}_k$ be the space of polynomials of degree $\leq k$, and let $\delta(k)$ be the distance (in the uniform norm on $[-1,1]$) from the function $|t|^r$ to the space $\mathcal{P}_{2k}$. It is known (see, e.g., Timan A.F., *Theory of approximation of functions of real variable*, Moscow, 1960, p.430) that if k is a nonnegative integer, then

$$\delta(k) \geq \kappa_{10}k^{-r},$$

with $\kappa_{10} > 0$ depending on r only. Let us set

$$k(N) = \lfloor \log N \rfloor,$$

with N given by (4.37); we assume n to be so large that $N \geq 3$. Same as in the proof of Proposition 4.1, for our N there exists a symmetric measure μ_N on $[-1,1]$ with variation 2 such that

$$\begin{aligned}
\int_{-1}^{1} t^l\mu_N(dt) &= 0,\ l = 0,1,...,2k(N), \qquad (4.40)\\
\int_{-1}^{1} |t|^r\mu_N(dt) &= 2\delta(k(N)) \geq 2\kappa_{10}k^{-r}(N),
\end{aligned}$$

and the positive and the negative components, μ_+, μ_- ($\mu_N = \mu_+ - \mu_-$) are symmetric probability distributions on $[-1,1]$.

Same as in Section 4.3, we define the measures μ_0 and μ_1 on $[-\alpha(N), \alpha(N)]$ "expanding" the measures $\mu_\pm$, thus coming to a pair of symmetric probability distributions μ_0, μ_1 on $[-\alpha(N), \alpha(N)]$ satisfying the relations

$$\begin{aligned}\int_{-\alpha(N)}^{\alpha(N)} t^l \mu_0(dt) &= \int_{-\alpha(N)}^{\alpha(N)} t^l \mu_1(dt), \qquad l = 0,1,...,2k(N);\\ \int_{-\alpha(N)}^{\alpha(N)} |t|^r \mu_0(dt) &\geq \int_{-\alpha(N)}^{\alpha(N)} |t|^r \mu_1(dt) + 2\delta(k(N))\alpha^r(N). \end{aligned} \tag{4.41}$$

Setting $\mu_{N,0} = \mu_0^N$, $\mu_{N,1} = \mu_1^N$ and denoting by $P_{N,0}, P_{N,1}$ the marginal distributions of observations (4.14) associated with the priors $\mu_{N,0}, \mu_{N,1}$, we, same as in the proof of the lower bound in Theorem 2.3, come to the inequality (cf. (4.30))

$$\mathcal{R}_s(N) \geq r^{-1}\alpha(N)\left(0.25\delta(k(N))e^{-\Omega} - N^{-1/2}\right), \tag{4.42}$$

where Ω is the Kullback distance between the distributions $P_{N,0}, P_{N,1}$:

$$\begin{aligned}\Omega &= \mathcal{K}(P_{N,0}, P_{N,1}) = N\mathcal{K}(\alpha(N)),\\ \mathcal{K}(\alpha) &= \int_{-\infty}^{\infty} \log(p_{\mu_+}(\alpha,y)/p_{\mu_-}(\alpha,y))p_{\mu_+}(\alpha,y)dy,\end{aligned} \tag{4.43}$$

with $p_\nu(\alpha,\cdot)$ given by (4.31).

For $T > 0$, let us set

$$\mathcal{K}_T(\alpha) = \int_{|y|\leq T} \log(p_{\mu_+}(\alpha,y)/p_{\mu_-}(\alpha,y))p_{\mu_+}(\alpha,y)dy. \tag{4.44}$$

Lemma 4.8 *For every $T > 0$*

$$\left.\frac{d^l\mathcal{K}_T(\alpha)}{d\alpha^l}\right|_{\alpha=0} = 0,\ l = 0,...,2k(N).$$

Proof. We have

$$\mathcal{K}_T(\alpha) = \int_{-T}^{T} \log\left(1 + \frac{p_\mu(\alpha,y)}{p_{\mu_-}(\alpha,y)}\right) p_{\mu_-}(\alpha,y)dy,$$

and the result is readily given by (4.35) (in view of the first relation in (4.41), equality (4.35) is now valid for $l = 0,1,...,2k(N)$, see the proof of (4.35)).

The remaining part of the required information on $\mathcal{K}_T(\cdot)$ is given by

Lemma 4.9 *For every $T \geq 20$ and all $\alpha \in [-1,1]$, one has*

$$\mathcal{K}(\alpha) \leq \exp\{-(T-1)^2/2\} + \mathcal{K}_T(\alpha). \tag{4.45}$$

The function $\mathcal{K}_T(\alpha)$ can be extended analytically into the circle $|\alpha| \leq (10T)^{-1}$, and in this circle

$$|\mathcal{K}_T(\alpha)| \leq 2/3.$$

Proof. We clearly have

$$\begin{array}{rcl} \mathcal{K}(\alpha) & = & \mathcal{K}_T(\alpha) + R_T(\mu_+, \mu_-), \\ R_T(\nu, \nu') & = & \displaystyle\int_{|y|>T} \log(p_\nu(\alpha, y)/p_{\nu'}(\alpha, y)) p_\nu(\alpha, y) dy, \end{array}$$

ν, ν' being probability distributions on $[-1,1]$. Now, $R_T(\nu, \nu')$ is a convex functional of probability distributions ν, ν'; therefore its supremum, over all pairs (even non-symmetric) probability distributions on $[-1,1]$ is the same as its supremum over the set P_s^2 of pairs of distributions on the same segment with singleton supports. Indeed, every probability distribution ν on $[-1,1]$ can be approximated by a sequence $\{\nu_i\}$ of discrete distributions with finite supports in the sense that $\int g(x)\nu_i(dx) \to \int g(x)\nu(dx)$ for every continuous on $[-1,1]$ function g. From this observation and the fact that R_T, as it is easily seen, is lower semicontinuous (in fact even continuous) with respect to the weak topology on the set P^2 of pairs of probability distributions on $[-1,1]$ we conclude that $\sup_{(\nu,\nu')\in P^2} R_T(\nu, \nu') = \sup_{(\nu,\nu')\in P_d^2} R_T(\nu, \nu')$, P_d^2 being the set of pairs of discrete probability distributions on $[-1,1]$ with finite supports. Finally, every pair $(\nu, \nu') \in P_d^2$ is a convex combination of pairs from P_s^2; since R_T is convex, its supremum over P_d^2 is the same as the one over P_s^2, whence $\sup_{(\nu,\nu')\in P^2} R_T(\nu, \nu') = \sup_{(\nu,\nu')\in P_s^2} R_T(\nu, \nu')$, as claimed.

Now consider a pair of distributions $(\nu_+, \nu_-) \in P_s^2$; let ν_+ be concentrated at a point

t and ν_- be concentrated at a point τ $(t, \tau \in [-1,1])$. In this case we have

$$\begin{aligned}
R_T(\nu,\nu') &= \int_{|y|>T} \left[-\frac{(y-\alpha t)^2}{2} + \frac{(y-\alpha\tau)^2}{2} \right] \exp\{-\frac{(y-\alpha t)^2}{2}\} \frac{1}{\sqrt{2\pi}} dy \\
&= \int_{\{y \le -T-\alpha t\} \cup \{y \ge T-\alpha t\}} \left[\alpha(t-\tau)y + \alpha^2(t-\tau)^2/2 \right] \varphi(y) dy \\
&= \alpha(t-\tau)(2\pi)^{-1/2} \left[\exp\{-(T-\alpha t)^2/2\} - \exp\{-(T+\alpha t)^2/2\} \right] \\
&\quad + 2(2\pi)^{-1/2}\alpha^2(t-\tau)^2(T-1)^{-1} \exp\{-(T-1)^2/2\} \\
&\le (2\pi)^{-1/2}(2 + 8(T-1)^{-1}) \exp\{-(T-1)^2/2\} \\
&\le \exp\{-(T-1)^2/2\}
\end{aligned}$$

(we have taken into account that $T \ge 20$). Thus,

$$\sup_{P^2} R_T(\nu,\nu') = \sup_{P_s^2} R_T(\nu,\nu') \le \exp\{-(T-1)^2/2\},$$

and (4.45) follows.

Let us now look at the function $\mathcal{K}_T$. Let y be a real with $|y| < T$, and let t be a real with $|t| \le 1$. The absolute value of the derivative of the function $g(\alpha) = \exp\{-\alpha^2 t^2/2\} \mathrm{ch}(\alpha t y)$ in the circle $|\alpha| \le z \le 1$ clearly does not exceed $(T+1)\exp\{zT + z^2/2\}$, and therefore $|g(\alpha) - 1| = |g(\alpha) - g(0)| \le (zT + z)\exp\{zT + z^2/2\}$ in this circle. It follows that in the circle $|\alpha| \le z \equiv (10T)^{-1}$ we have

$$\begin{aligned}
&\left| \int_{-1}^{1} \exp\{-\alpha^2 t^2/2\} \mathrm{ch}(\alpha t y) \nu(dt) - 1 \right| \\
&\qquad \le (zT + z)\exp\{zT + z^2/2\} \le 1/5 \exp\{0.105\} \le 1/4,
\end{aligned}$$

both for $\nu = \mu_+$ and for $\nu = \mu_-$. Consequently, for the indicated z and $|\alpha| \le z$ we have

$$\left| \frac{p_{\mu_+}(\alpha, y)}{p_{\mu_-}(\alpha, y)} - 1 \right| \le 1/3.$$

We see that if y is real and $|y| \le T$, then the function $\log(p_{\mu_+}(\alpha,y)/p_{\mu_-}(\alpha,y))$, regarded as a function of α, can be extended analytically from the segment $|\alpha| \le d_T = (10T)^{-1}$ of the real axis onto the circle $|\alpha| \le d_T$ in the complex plane, and the absolute value of the extended function in this circle does not exceed the quantity

$$\sum_{m=1}^{\infty} \frac{1}{m} \left(\frac{1}{3} \right)^m = \log(3/2).$$

By the same reasons, for real y with $|y| \le T$ and every α from the circle $|\alpha| \le d_T$ we have $|p_{\mu_+}(\alpha, y)| \le 5/4\varphi(y)$, and we see that $\mathcal{K}_T$ is an analytic function in the circle $|\alpha| \le d_T$ with absolute value in the circle not exceeding $5/4 \log 3/2 \le 2/3$.

According to Lemma 4.9, $\mathcal{K}_T(\alpha)$ is an analytic function of α in the circle $|\alpha| \le d_T = (10T)^{-1}$ which is bounded in absolute value in this circle by $2/3$; according to Lemma 4.7, $\mathcal{K}_T(\alpha)$ has zero of order at least $2k(N) + 1$ at the origin, and since the function is even, the order of this zero is at least $2k(N) + 2$. Consequently, the function $d_T^{2k(N)+2} \mathcal{K}_T(\alpha)\alpha^{-2k(N)-2}$ is analytic in the circle $|\alpha| \le d_T$ and therefore the maximum of its absolute value in the circle is equal to the one on the boundary of the circle, i.e., it does not exceed $2/3$. We conclude that

$$\mathcal{K}_T(\alpha) \le \frac{2}{3} \frac{\alpha^{2k(N)+2}}{d_T^{2k(N)+2}}, \qquad -d_T \le \alpha \le d_T. \tag{4.46}$$

Now let us set

$$T = T(N) = 1 + \sqrt{2 \log N}$$

and let us look what (4.46) with this T implies for $\alpha = \alpha(N)$. In view of (4.38) for large enough values of n we have

$$\frac{\alpha(N)}{d_{T(N)}} = 10T(N)\alpha(N) \le 0.2 < \exp\{-1\},$$

so that (4.46) indeed is applicable to $\alpha = \alpha(N)$ and results in

$$\mathcal{K}_{T(N)}(\alpha(N)) \le \exp\{-2k(N) - 2\} \le N^{-2}$$

(see (4.4)). Applying (4.45) with $\alpha = \alpha(N)$, $T = T(N)$, we therefore get

$$\mathcal{K}(\alpha(N)) \le N^{-2} + \exp\{-(T(N) - 1)^2/2\} \le N^{-2} + N^{-1},$$

so that (see (4.43))

$$\Omega = N\mathcal{K}(\alpha(N)) \le 1 + N^{-1}.$$

The latter relation, in view of (4.42), (4.4) and the lower bound for $\delta(k(N))$ from (4.41), implies (4.39). Proposition 4.2 is proved.

References

[1] Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23** 11–29.

[2] Bickel, P.J. and Ritov, Y. (1988) Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhya*, **50** 381–393.

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

[4] Donoho, D.L. and Liu, R.C. (1991). Geometrizing rate of convergence, III. *Ann. Statist.* **19** 668–701.

[5] Efroimovich, S. and Low, M. (1996). On Bickel and Ritov's conjecture about adaptive estimation of the integral of the square of density derivative. *Ann. Statist.* **24** 682–686.

[6] Fan, J. (1991). On the estimation of quadratic functionals. *Ann. Statist.* **19** 1273–1294.

[7] Hall, P. and Marron, J.S. (1987). Estimation of integrated squared density derivatives. *Statist. and Prob. Letters*, **6**, 109–115.

[8] Khasminski, R. and Ibragimov, I. (1979). On the nonparametric estimation of functionals. In *Proceedings of the Second Prague Symp. on Asymptotic Statistics* (P.Mandl and M.Huskova eds.) North-Holland, Amsterdam, 41–51.

[9] Ibragimov, I. and Khasminski, R. (1979). *Asymptotic Estimation Theory.* (in Russian). Nauka. Moscow. English translation: *Statistical Estimation: Asymptotic Theory.* (1981) Springer. Berlin, Heidelberg, New York.

[10] Khasminski, R. and Ibragimov, I. (1980). Some estimation problems for stochastic differential equations. *Lecture Notes Control Inform. Sci.* Springer, New York. **25** 1–12.

[11] Ibragimov, I.A. and Khasminski, R. (1980). On the estimation of distribution density. *Zap. Nauch. Sem. LOMI*, **98** 61–85.

[12] Ibragimov, I. and Nemirovski, A. and Khasminski, R. (1986). Some problems on nonparametric estimation in Gaussian white noise. *Theory Probab. Appl.*, **31** 391–406.

[13] Ibragimov, I. and Khasminski, R. (1987). Estimation of linear functionals in Gaussian noise. *Theory Probab. Appl.*, **32** 30–39.

[14] Ibragimov, I. and Khasminski, R. (1991). Asymptotic normal families of distributions and effective estimation. *Ann. Statist.*, **19** 1681–1724.

[15] Ingster, Yu.I. (1982). Minimax nonparametric detection of signals in white Gaussian noise. *Problems Inform. Transmission*, **18** 130 – 140.

[16] Ingster, Yu.I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I–III. *Math. Methods of Statist.* **2** (1993) 85 – 114, **3** (1993) 171 – 189, **4** (1993) 249 – 268.

[17] Kerkyacharian, J. and Picard, D. (1996). Estimating nonquadratic functionals of a density using Haar basis. *Ann: Statist.*, **24** 485–508.

[18] Korostelev, A.P. (1990). On the accuracy of estimation of non-smooth functionals of regression. *Theory Probab. Appl.*, **35** 768–770.

[19] Korostelev, A.P. and Tsybakov, A.B. (1994). *Minimax Theory of Image Reconstruction.* Lecture Notes in Statist. Springer, New York.

[20] Koshevnik, Yu. and Levit, B. Ya. (1976). On a nonparametric analogue of the information matrix. *Theory Probab. Appl.*, **21** 738–753.

[21] Laurent, B. (1996) Efficient estimation of integral functionals of a density. *Ann: Statist.*, **24** 659–682.

[22] Lehmann, E.L. (1959) *Testing Statistical Hypothesis* Wiley, New York.

[23] Lepski, O. and Spokoiny, V. (1995). Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*, to appear.

[24] Levit, B.Ya. (1974). On optimality of some statistical estimates. In *Proceedings of the Prague Symp. on Asymptotic Statistics* (J. Hajek, ed.) Univ. Karlova, Prague. **2** 215–238.

[25] Levit, B.Ya. (1975). Efficiency of a class of nonparametric estimates. *Theory Probab. Appl.*, **20** 738–754.

[26] Levit, B.Ya. (1978). Asymptotically efficient estimation of nonlinear functionals. *Problems info. Transmission*, **14** 65–72.

[27] Spokoiny, V. (1996). Adaptive hypothesis testing using wavelets. *Annals of Statistics*, **26** 2477–2498.