Adaptive Non-Parametric Estimation of Smooth Multivariate Functions

O.V. Lepski

Université de Provence

39, rue F. Joliot-Curie

13453 Marseille Cedex 13, France

B.Y. Levit

University of Utrecht

Budapestlaan 6

3584 CD Utrecht, Netherlands

March 4, 1999

Abstract

Adaptive pointwise estimation of smooth functions $f(\mathbf{x})$ in $\mathbb{R}^d$ is studied in the white Gaussian noise model of a given intensity $\varepsilon \to 0$. It is assumed that the Fourier transform of f belongs to a large class of rapidly vanishing functions but is otherwise unknown.

Optimal adaptation in higher dimensions presents several challenges. First, the number of essentially different estimates having a given variance $\varepsilon^2 S$ increases polynomially, as S^{d-1}. Second, the set of possible estimators, totally ordered when $d = 1$, becomes only partially ordered when $d > 1$. We demonstrate how these challenges can be met. The first one is to be matched by a meticulous choice of the estimators' net. The key to solving the second problem lies in a new method of *spectral majorants* introduced in this paper.

Extending our earlier approach used in [12], we restrict ourselves to a family of estimators, rate-efficient in an off-beat case of *partially parametric* functional classes. A proposed adaptive procedure is shown to be asymptotically minimax, simultaneously for any *ample regular non-parametric* family of underlying functions f.

---

[1] *AMS* 1991 *subject classifications.* Primary 62G05, 62G20; secondary 62C20.
Key words: nonparametric estimation, adaptive estimation, minimax estimation, Fourier transforms.

1 Introduction

The paper provides a multivariate generalization of the results of Lepski and Levit [12], dealing with the adaptive estimation for classes of infinitely differentiable functions. The setting is that of a function $f : \mathbb{R}^d \to \mathbb{R}^1$ corrupted by a white Gaussian noise of a given intensity ε:

$$dV(\mathbf{x}) = f(\mathbf{x})d\mathbf{x} + \varepsilon dW(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^d, \tag{1.1}$$

where we are interested in a pointwise estimation of $f(\mathbf{x})$ for small ε. Such a model is well suited for introducing new statistical methods, as it captures most of the essential features common to a variety of problems in functional estimation, in their pure form, while suppressing a number of technicalities, temporarily deemed less important.

For the last two decades functional estimation has been dominated by the assumption that the unknown function f belongs to some finite smoothness functional classes (Hölder, Sobolev, Besov). Although infinitely smooth functions already featured notably in non-parametrics in the mid-fifties – cf. e.g. Parzen's study of spectral density estimation for a stationary time series [13] – recent ideas about optimal (minimax, adaptive) non-parametric estimation did not yet exist then.

Rate minimax estimators for analytic functions in different $\mathbf{L}_p$-norms were studied by Ibragimov and Hasminskii [6],[8]. The possibility of asymptotically minimax pointwise estimation of functions analytic in a given vicinity of a segment has been demonstrated by Ibragimov and Hasminskii in [7] for the white noise model. Corresponding results for estimating a density or a distribution function, analytic in a strip around the real axis, were obtained by a different technique in Golubev and Levit [3].

Broader classes of infinitely differentiable functions introduced in Lepski and Levit [12] arguably present a viable alternative to the finite smoothness functional classes. By carrying out the whole program of adaptive minimax estimation for such functions they demonstrated that, essentially without losing anything, they stand to gain a great deal compared to the classes of finitely differentiable functions. In fact it was this effort to develop optimal adaptive methods for such classes that helped us understand their universal role. One can say that in the result, the notion of adaptivity has entirely changed our perception of the – hardly ever known – 'true order of smoothness'.

Our functional classes $\mathcal{A}_\alpha = \{f\}$ here are somewhat narrower than in Lepski and Levit [12], in that the assumption of a *rapidly vanishing* Fourier transform $\hat{f}$ is replaced by an *exponential* decline of $\hat{f}$. Even so the classes $\mathcal{A}_\alpha$ remain an interesting object for adaptive estimation. However, the main challenges ensue from the fact that the classes $\mathcal{A}_\alpha$ are now anisotropic, resulting in additional serious dimensionality problems. First, the number of essentially different estimators having a given variance $\varepsilon^2 S$ increases polynomially in S, as S^{d-1} (thus also exponentially in d). Second, the set of possible estimators, which is typically totally ordered for $d = 1$, becomes only partially ordered when $d > 1$. We will address both of these problems. The first one has to be tackled by a meticulous choice of the estimators' net. The key to solving the second problem lies in a new *method of spectral majorants* described below in Section 5. Handling these problems will bring us in this paper to a modification of the adaptive optimality concept proposed in Lepski and Levit [12] .

As in Lepski and Levit [12] we restrict ourselves exclusively to kernel estimators based on the so-called *sinc* function $k_s(x)$, with a bandwidth parameter s. Although such estimators are not quite suited for practical purposes, they have certain advantages in a theoretical analysis. This is manifested by the fact that, e.g. in the one-dimensional case, the corresponding estimator $\tilde{f}_s(x)$ can be represented as

$$\tilde{f}_s(x) \overset{\mathrm{w}}{=} f(x) + b_s(x) + \tilde{W}(s)$$

where $b_f(x)$ is the bias and $\tilde{W}(s)$, $s > 0$ is yet another standard Wiener process; cf. Lepski and Levit [12] . A generalization of this property easily obtainable when $d > 1$ considerably simplifies the analysis of such estimators. Note that in the *frequency domain* they represent a leading approximation term common to *all* estimators which are asymptotically minimax with respect to the corresponding classes $\mathcal{A}_\alpha$.

After starting with some notations in Section 2, we first derive in Section 3 asymptotically minimax estimators for fixed functional classes $\mathcal{A}_\alpha$. The availability of such estimators is a vivid manifestation of the difference between these functional classes and the more traditional – finite smoothness – classes. In Section 4 a modified concept of adaptive optimality is discussed with respect to multivariate functional scales. Adaptive estimators will be introduced and evaluated in Section 5, whereas the proof of their asymptotic optimality will be completed in Section 6.

This paper is dedicated to Johann Pfanzagl on the occasion of his 70th birthday and is a tribute to his pioneering scientific work which has inspired us for many years.

2 The model

The observation process. In this paper we study the problem of pointwise estimation of an unknown function $f : \mathbb{R}^d \to \mathbb{R}^1$, observed in white Gaussian noise. The corresponding model can be described by the stochastic differential equation (1.1), where V is the observation process, f is the unknown function, $\varepsilon > 0$ is a small given noise level and W is a standard Wiener process, or the so-called Brownian sheet, in $\mathbb{R}^d$. When $\mathbf{x}$ varies in the positive orthant, $W(\mathbf{x})$ can be defined as a centered continuous Gaussian process with zero boundary values, whose covariances are given by (cf. [4], [14])

$$\mathbf{E}W(\mathbf{x})W(\mathbf{y}) = \prod_{i=1}^{d} \min(x_i, y_i).$$

In other orthants one can use independent copies of $W(\mathbf{x})$. Such a process is easier to define first on bounded subsets of $\mathbb{R}^d$. The transition to the whole space can be accomplished following Ibragimov and Hasminskii [5], Appendix II.

Since we are going to use the underlying equation (1.1) only to a limited extent, its theory does not differ much from the one-dimensional case $d = 1$. In fact we will need only a few basic facts concerning (1.1) and the linear functionals of the form

$$\int k(\mathbf{x})\, dV(\mathbf{x}). \tag{2.1}$$

First, for any $k \in \mathbf{L}_2(\mathbb{R}^d)$ the stochastic integral $\int k(\mathbf{x})\, dW(\mathbf{x})$ is well defined and represents a Gaussian $\mathcal{N}(0, \sigma^2)$ random variable where, according to the so-called *Ito isometry formula*,

$$\sigma^2 = \|k\|_2.$$

Here $\|\cdot\|_2$ denotes the $\mathbf{L}_2(\mathbb{R}^d)$ norm.

To simplify things further and avoid technicalities, not important to our statistical problem, we will always assume that the underlying function $f(\mathbf{x})$ in (1.1) is continuous and also belongs to $\mathbf{L}_2(\mathbb{R}^d)$. A generalization is possible along the lines indicated in Lepski and Levit [12]. These assumptions allow us to exploit linear estimators (2.1), see e.g. Walsh (1986), which can be shown to be Borel measurable functions from $\mathbf{C}(\mathbb{R}^d)$ to $\mathbb{R}^1$ such that

$$\int k(\mathbf{x})\, dV(\mathbf{x}) = \int k(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} + \varepsilon \int k(\mathbf{x})\, dW(\mathbf{x}).$$

Finally, we will need the so-called Cameron-Martin-Girsanov formula for the solutions of (1.1). Denote by $\mathbf{P}_f^{(\varepsilon)}$ the distribution in $\mathbf{C}(\mathbb{R}^d)$, induced by the solutions of the equation (1.1). Then

$$\frac{d\mathbf{P}_f^{(\varepsilon)}}{d\mathbf{P}_0^{(\varepsilon)}}(V(\cdot)) = \exp\left\{\varepsilon^{-2}\int f(\mathbf{x})\, dV(\mathbf{x}) - \frac{\varepsilon^{-2}}{2}\int f^2(\mathbf{x})\, d\mathbf{x}\right\}. \tag{2.2}$$

This formula can be proved by the arguments similar to those used in [5], Appendix II.

Functional classes and functional scales. We will study estimation problems for functions f in $\mathbb{R}^d$ wich have rapidly vanishing Fourier transforms. Our motivation is similar to Lepski and Levit [12], but the approach here is somewhat different. On the one hand, we restrict our classes to functions with exponentially vanishing Fourier transforms. In this respect our classes are closer to those used e.g. by Parzen (1958) in the context of estimating the spectral density of (one-dimensional) time series. On the other hand, our classes are much more complex than in Lepski and Levit [12] or Parzen (1958), due to the multidimensionality of the problem.

Let us start with some notations. We will denote the ball with a radius $r > 0$ at the origin of $\mathbb{R}^d$ by $\mathbf{B}(r)$ and we denote a centrally symmetric cuboid in $\mathbb{R}^d$

$$\mathbf{C}(\mathbf{s}) = \{\mathbf{t} \in \mathbb{R}^d : |t_i| \le s_i\}, \quad \mathbf{s} = (s_1, \ldots, s_d).$$

Let $\mathbf{C} = \mathbf{C}(1, ..., 1)$ be the unit cube in $\mathbb{R}^d$. For a vector $\mathbf{x} \in \mathbb{R}^d$ denote by $\mathrm{diag}(\mathbf{x})$ the matrix $(x_i \delta_{ij})$ having elements of $\mathbf{x}$ on its diagonal. An inequality $\mathbf{x}_1 < \mathbf{x}_2,\ \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ will mean $x_{1i} < x_{2i},\ i = 1, \ldots, d$. This partial ordering gives meaning to vector notations such as '$\mathbf{x} \ge 0$' or '$\liminf \mathbf{x}_n > 0$'.

Below, B' will denote the transposition of a vector or a matrix, $B_{i\cdot}$, $B_{\cdot j}$ will denote the i-th row and the j-th column of the matrix B, respectively. For a square matrix B we denote by $|B| = |\det B\,|$, $\|B\| = \max_{i,j} |B_{ij}|$, $\mathrm{diag} B = (B_{ii}\delta_{ij})$. Let E be the unit matrix.

Denote further by $\mathfrak{A}_a$ the class of non-singular $d \times d$ matrices B (*angular matrices*) such that

$$\mathrm{diag}(B'B) = E,$$

and by $\mathfrak{A}_o$ the class of orthogonal matrices B:

$$B'B = E.$$

Obviously a non-singular matrix B can be represented as

$$B = B_a \mathrm{diag}(\mathbf{s}), \quad B_a \in \mathfrak{A}_a,\ \mathbf{s} > 0,$$

whereas a *rectangular matrix* B (having orthogonal column-vectors) can be represented as

$$B = B_o \mathrm{diag}(\mathbf{s}), \quad B_o \in \mathfrak{A}_o,\ \mathbf{s} > 0.$$

For a vector $\mathbf{r} = (R, r_1, \ldots, r_d) > 0$ consider the exponential function

$$\exp_{\mathbf{r}}(\mathbf{t}) = R \exp\left(-\sum_{i=1}^{d} |t_i|^{r_i}\right), \quad (\mathbf{t} \in \mathbb{R}^d). \tag{2.3}$$

Denote by $\mathcal{K}_0$ the class of functions $\alpha(\mathbf{t})$, $\mathbf{t} \in \mathbb{R}^d$ such that for some $\mathbf{r} > 0$ and a non-singular $d \times d$ matrix Γ

$$\alpha(\mathbf{t}) = |\Gamma| \exp_{\mathbf{r}}(\Gamma \mathbf{t}). \tag{2.4}$$

For any $\alpha \in \mathcal{K}_0$, let $\mathcal{A}_\alpha$ be the class of all continuous functions f in $\mathbf{L}_2(\mathbb{R}^d)$ with the Fourier transform $\hat{f}$ satisfying

$$\left|\hat{f}(\mathbf{t})\right| \leq \alpha(\mathbf{t}), \quad \mathbf{t} \in \mathbb{R}^d. \tag{2.5}$$

Finally, for any subset $\mathcal{K} \subset \mathcal{K}_0$ we define a functional scale $\mathcal{A}_\mathcal{K}$ as

$$\mathcal{A}_\mathcal{K} = \{\mathcal{A}_\alpha \,:\, \alpha \in \mathcal{K}\}. \tag{2.6}$$

Remark 2.1. Obviously α is defined by the parameters $\mathbf{r}$, Γ and, in turn, determines $\mathbf{r} = \mathbf{r}_\alpha$. The row-vectors of Γ, however, are defined except for their sign. To make the correspondence complete, assume, without loss of generality, that $\Gamma_{i1} \geq 0$, $i = 1, \ldots, d$. Then we can also write $\Gamma = \Gamma_\alpha$. Accordingly one can consider a scale $\mathcal{A}_\mathcal{K}$ to be parametrized either by the blend parameter α, or, separately, by the form/size ($\mathbf{r}$) and structural (Γ) parameters, the latter combining the band-scale and angular parameters. Different scale indexing will be introduced at the end of Section 3.

From the general theory of Fourier transforms it follows that for any $f \in \mathcal{A}_{\mathcal{K}_0}$ the Fourier inversion formula

$$f(\mathbf{x}) = \frac{1}{(2\pi)^d} \int e^{-\mathrm{i}\mathbf{t}'\mathbf{x}} \hat{f}(\mathbf{t})\, d\mathbf{t} \tag{2.7}$$

holds. Here and in the sequel the integration is extended over the whole $\mathbb{R}^d$ unless stated otherwise.

Moreover, for any f, $g \in \mathbf{L}_2(\mathbb{R}^d)$ and $\mathbf{x} \in \mathbb{R}^d$

$$\int g(\mathbf{x} - \mathbf{y}) f(\mathbf{y})\, d\mathbf{y} = \frac{1}{(2\pi)^d} \int e^{-\mathrm{i}\mathbf{t}'\mathbf{x}} \hat{g}(\mathbf{t}) \hat{f}(\mathbf{t})\, d\mathbf{t} \tag{2.8}$$

and the Parseval formula

$$\int g(\mathbf{x})^2\, d\mathbf{x} = \frac{1}{(2\pi)^d} \int |\hat{g}(\mathbf{t})|^2\, d\mathbf{t}. \tag{2.9}$$

holds; see e.g. [2], Sect. 2.10.

3 Minimax estimation for the classes $\mathcal{A}_\alpha$

Band-limiting and its figures of merit. Consider the multivariate *sinc* function

$$k(\mathbf{x}) = \prod_{i=1}^{d} \frac{\sin x_i}{\pi x_i},$$

with a *cut-off* Fourier transform

$$\hat{k}(\mathbf{t}) = \int e^{\mathbf{i}\mathbf{t}'\mathbf{x}} k(\mathbf{x})\, d\mathbf{x} = \prod_{i=1}^{d} \mathbf{1}\left(|t_i| \le 1\right). \tag{3.1}$$

For a non-degenerate matrix B let

$$k_B(\mathbf{x}) = |B|\, k(B'\mathbf{x}). \tag{3.2}$$

It follows easily from (3.1) that

$$\hat{k}_B(\mathbf{t}) = \hat{k}(B^{-1}\mathbf{t}). \tag{3.3}$$

In this paper we are dealing exclusively with the kernel-type estimators of the form

$$\tilde{f}_B(\mathbf{x}) = \int k_B(\mathbf{x}-\mathbf{y}) dV(\mathbf{y}). \tag{3.4}$$

In the communication engineering literature such estimators are oftenly referred to as low-pass filters or band-limiting. One can regard the matrix B^{-1} as a generalized bandwidth.

Let us start with the sample distribution of the estimator $\tilde{f}_B(\mathbf{x})$ for a given B. Using (1.1), one arrives at the usual bias-variance decomposition:

$$\tilde{f}_B(\mathbf{x}) - f(\mathbf{x}) = \int k_B(\mathbf{x}-\mathbf{y}) f(\mathbf{y})\, d\mathbf{y} - f(\mathbf{x}) + \varepsilon \int k_B(\mathbf{x}-\mathbf{y})\, W(d\mathbf{y}) \overset{\mathrm{d}}{=} b_B + \xi_B. \tag{3.5}$$

From the stochastic calculus it is well known that

$$\tilde{f}_B(\mathbf{x}) - f(\mathbf{x}) \overset{\mathrm{w}}{\sim} \mathcal{N}(b_B, \sigma_B^2),$$

where by *Ito's isometry* and *Parseval's formula*

$$\sigma_B^2 = \varepsilon^2 \mathbf{E}\,\xi_B^2 = \varepsilon^2 \int k_B^2(\mathbf{x}-\mathbf{y})\,d\mathbf{y} = \frac{\varepsilon^2}{(2\pi)^d}\int |\hat{k}_B(\mathbf{t})|^2\,d\mathbf{t} = \frac{\varepsilon^2|B|}{\pi^d}. \tag{3.6}$$

Next by (2.7)-(2.8) and (3.3), for any $f \in \mathcal{A}_\alpha$

$$\begin{aligned} |b_B| &= \left| \frac{1}{(2\pi)^d} \int e^{-i\mathbf{t}'\mathbf{x}} \left(\hat{k}_B(\mathbf{t}) - 1\right) \hat{f}(\mathbf{t})\,d\mathbf{t} \right| \\ &\le \frac{1}{(2\pi)^d} \int \left|\hat{k}_B(\mathbf{t}) - 1\right| \alpha(\mathbf{t})\,d\mathbf{t} \\ &= \frac{|\Gamma|}{(2\pi)^d} \int \left|\hat{k}(B^{-1}\mathbf{t}) - 1\right| \exp_{\mathbf{r}}(\Gamma \mathbf{t})\,d\mathbf{t} \\ &= \frac{|\Gamma B|}{(2\pi)^d} \int \left|\hat{k}(\mathbf{t}) - 1\right| \exp_{\mathbf{r}}(\Gamma B \mathbf{t})\,d\mathbf{t} \stackrel{\text{def}}{=} b(\alpha, B). \end{aligned} \tag{3.7}$$

We will continue our study of these estimators shortly, but first let us note that relations (3.6)–(3.7) provide a basis for two different orderings in the set of estimators $\tilde{f}_B$.

Sinc **estimators: modes of comparison.** In the one-dimensional case $d = 1$ the collection of the kernels k_B (estimators, bandwidths respectively) is naturally ordered, say, in the direction of increasing B. Equations (3.6)-(3.7) do indeed show the variance increasing simultaneously with $|B|$, whereas the bias or, more precisely the upper bound $b(\alpha, B)$ in (3.7), is typically decreasing. Thus the estimators $\tilde{f}_B$ are *completely ordered* – a fact that significantly simplifies their comparison and makes it easier to balance their variances viz. bias.

When $d > 1$, however, the increase of the variance does not automatically lead to a decrease of the bias $b(\alpha, \cdot)$ with respect to which the estimators $\tilde{f}_B$ are only partially ordered. Accordingly we will introduce two types of orderings in the set of the kernels: a complete order associated with the variance σ_B^2 in (3.6), and a partial order related to the bound in (3.7). Without further mention we will in the sequel apply the same conventions to matrices B, kernels k_B and the estimators $\tilde{f}_B$.

We will say that

$$B_1 \preccurlyeq B_2, \quad \text{iff} \quad |B_1| \le |B_2|,$$

or, equivalently,

$$\text{meas}\left(\text{supp}\hat{k}_{B_1}\right) \le \text{meas}\left(\text{supp}\hat{k}_{B_2}\right).$$

Next we say that

$$B_1 \Subset B_2,$$

if

$$\mathrm{supp}\hat{k}_{B_1} \subseteq \mathrm{supp}\hat{k}_{B_2}.$$

Typically $B_1 \Subset B_2$ would imply that $\tilde{f}_{B_2}$ has a smaller bias than $\tilde{f}_{B_1}$; more precisely, $b(\alpha, B_2) \le b(\alpha, B_1)$. Note that the relation $B_1 \Subset B_2$ is stronger than, but obviously not equivalent to the relation $B_1 \preccurlyeq B_2$, except for $d = 1$ when the orderings coincide.

Minimum risk estimators. Up to Remark 3.1 below we will regard the classes $\mathcal{A}_\alpha$ in this section as being fixed and known. We will show how the matrix-valued bandwidth B can be adjusted to yield an asymptotically minimax estimator $\tilde{f}_B$. Later α, still treated as known, will be allowed to depend on ε, whereas in the following sections optimal adaptive estimation in the presence of an unknown α will be discussed.

Finding the optimal $B = B_\alpha$ becomes easier if one factorizes this matrix in the form suggested by (3.7):

$$B_\alpha = B_{1\alpha} B_{2\alpha}, \tag{3.8}$$

where

$$B_{1\alpha} = \Gamma^{-1}, \qquad B_{2\alpha} = \mathrm{diag}(\boldsymbol{\nu}), \qquad \boldsymbol{\nu} = (\nu_1, \ldots, \nu_d) > 0. \tag{3.9}$$

Thus only the optimal scalar bandwidth parameters $\boldsymbol{\nu} = \boldsymbol{\nu}_\alpha(\varepsilon^2)$ remain to be determined.

With the above choice of B_α, (3.6)–(3.7) simplify to

$$\sigma_B^2 = \frac{\varepsilon^2 \nu_1 \cdots \nu_d}{\pi^d |\Gamma|} \tag{3.10}$$

and

$$b(\alpha, B) = \frac{R}{\pi^d} \prod_{i=1}^{d} \int_{\nu_i}^{\infty} \exp(-t^{r_i})\, dt. \tag{3.11}$$

By (3.10)–(3.11),

$$\sigma_B^2 + b_B^2 \le \frac{\varepsilon^2}{\pi^d |\Gamma|} \prod_{i=1}^{d} \nu_i + \frac{R^2}{\pi^{2d}} \prod_{i=1}^{d} \left(\int_{\nu_i}^{\infty} \exp(-t^{r_i})\, dt \right)^2. \tag{3.12}$$

Minimization of the upper bound in (3.12) leads to the equations ($j = 1, \ldots, d$):

$$\frac{\varepsilon^2}{\pi^d |\Gamma| \nu_j} \prod_{i=1}^d \nu_i = \frac{2R^2}{\pi^{2d}} \exp\left(-\nu_j^{r_j}\right) \left(\int_{\nu_j}^\infty \exp(-t^{r_j})\, dt\right)^{-1} \prod_{i=1}^d \left(\int_{\nu_i}^\infty \exp(-t^{r_i})\, dt\right)^2. \tag{3.13}$$

We will use the asymptotic equivalence relation

$$\int_\nu^\infty \exp\left(-t^r\right)\, dt \sim r^{-1} \nu^{1-r} \exp\left(-\nu^r\right), \quad (\nu \to \infty), \tag{3.14}$$

along with the inequality (cf. e.g. Lepski and Levit [12]):

$$\int_\nu^\infty \exp\left(-t^r\right)\, dt \leq 2r^{-1} \nu^{1-r} \exp\left(-\nu^r\right), \quad (\nu \geq (2/r)^{1/r}). \tag{3.15}$$

Applying (3.14) to the right-hand side of (3.13) transforms it to an equivalent equation

$$\frac{\pi^d \varepsilon^2}{2R^2 |\Gamma|} = r_j \nu_j^{r_j} \prod_{i=1}^d r_i^{-2} \nu_i^{(1-2r_i)} \exp(-2\nu_i^{r_i}), \tag{3.16}$$

which is sufficient for our purposes. Obviously the solution of (3.16) satisfies, for some $c > 0$,

$$r_j \nu_j^{r_j} = c, \quad j = 1, \ldots, d. \tag{3.17}$$

Denote

$$\frac{1}{r} = \sum_{i=1}^d \frac{1}{r_i}, \quad r^* = \prod_{i=1}^d r_i^{1/r_i}.$$

Then from (3.16)

$$\frac{\pi^d r^* \varepsilon^2}{2R^2 |\Gamma|} = \exp(-2c/r) c^{1-2d+1/r}. \tag{3.18}$$

The equation (3.18) has a unique solution $c = c(|\Gamma|)$ satisfying, for $\varepsilon \to 0$,

$$c = \frac{r}{2} \log \frac{|\Gamma|}{\varepsilon^2} (1 + o(1)). \tag{3.19}$$

According to (3.17) the optimal bandwidths ν_j are given by

$$\nu_{j\alpha}(\varepsilon^2) = \left(\frac{c}{r_j}\right)^{1/r_j} = \left(\frac{r}{2r_j}\log\frac{|\Gamma|}{\varepsilon^2}\right)^{1/r_j}(1+o(1)). \tag{3.20}$$

Denote also by

$$\sigma^2_{\alpha}(\varepsilon^2) \stackrel{\text{def}}{=} \varepsilon^2 S_{\alpha}(\varepsilon^2) = \frac{\varepsilon^2 c^{1/r}(|\Gamma|)}{\pi^d r^* |\Gamma|}, \tag{3.21}$$

the variance in (3.6) with $B = B_\alpha$.

Next we will demonstrate that, with the thus defined bandwidth matrix B_α, the bias of the estimator $\tilde{f}_B$ becomes negligibly small. More precisely, for any $B \Supset B_\alpha$ and all sufficiently small ε, according to (3.7), (3.11), (3.15) and (3.18),

$$\begin{aligned}
b^2(\alpha, B) &= \frac{R^2}{\pi^{2d}}\prod_{i=1}^{d}\left(\int_{\nu_i}^{\infty}\exp(-t^{r_i})\,dt\right)^2 \\
&\le \frac{2^{2d}R^2}{\pi^{2d}}\prod_{i=1}^{d}\left(\frac{\nu_i^{1-r_i}}{r_i}\right)^2\exp\left(-2\nu_i^{r_i}\right) \\
&= \frac{\varepsilon^2 2^{2d-1}}{c\pi^d|\Gamma|}\prod_{i=1}^{d}\nu_i = \frac{2^{2d-1}}{c}\sigma^2_{\alpha}(\varepsilon) \\
&= o(\sigma^2_{\alpha}(\varepsilon)).
\end{aligned} \tag{3.22}$$

Remark 3.1. Following the approach discussed at length in [12], Section 3.2, we can substantially broaden our original framework by allowing the functional classes $\mathcal{A}_\alpha$ and scales $\mathcal{A}_\mathcal{K}$ also to depend on ε. In other words both the form/size parameters $\mathbf{r}$ and the structural parameters Γ need not be regarded as fixed but rather may vary with ε. It is easy to see that even in this broader framework the asymptotic relations derived above hold uniformly with respect to any functional scale $\mathcal{A}_\mathcal{K}$ such that

$$0 < \liminf_{\varepsilon\to 0}\inf_{\mathcal{K}}\mathbf{r}_\alpha \le \limsup_{\varepsilon\to 0}\sup_{\mathcal{K}}\mathbf{r}_\alpha < \infty \tag{3.23}$$

and

$$\lim_{\varepsilon\to 0}\sup_{\mathcal{K}}\frac{\varepsilon^2}{|\Gamma|} = 0. \tag{3.24}$$

Following Lepski and Levit [12] we will call any such scale a *uniform scale.*

Asymptotically minimax estimators. To highlight the scope of the results presented below we will invoke in the rest of the paper a broader framework, following Lepski and Levit [12] , in which the functional classes $\mathcal{A}_\alpha = \mathcal{A}_{\alpha_\varepsilon}$ and functional scales $\mathcal{A}_\mathcal{K} = \mathcal{A}_{\mathcal{K}_\varepsilon}$ may themselves depend on ε. On the one hand, such an extension follows almost immediately from the present analysis, cf. Remark 3.1; but perhaps a more convincing argument in favor of such an extension comes naturally from the perspective of adaptive estimation.

In such a setting one would like to estimate an unknown function f under the broadest possible assumptions, i.e. within the limits determined by the consistency requirement. As (3.24) itself clearly demonstrates, such limits depend on ε, although this condition alone may be not sufficient for the existence of *consistent adaptive* estimators; cf. Lepski and Levit [12] , Example 6.1.

Consider an arbitrary estimator $\tilde{f}_\varepsilon(\mathbf{x})$ of $f(\mathbf{x})$, based on the observation process V in (1.1). Denote by $\mathcal{W}$ the class of loss functions $w(x) \geq 0$, $x \in \mathbb{R}$ such that

$$w(x) = w(-x), \quad w(x) \geq w(y), \quad |x| \geq |y| \,,$$

and for some $0 < \kappa < 1/2$

$$\int e^{-\kappa x^2} w(x)\, dx < \infty \,.$$

Let ξ be a standard Gaussian random variable.

Theorem 3.1. *For any $\alpha \in \mathcal{K}_0$ and $w \in \mathcal{W}$, uniformly in $\mathbf{x} \in \mathbb{R}^d$,*

$$\lim_{\varepsilon \to 0} \inf_{\tilde{f}_\varepsilon} \sup_{f \in \mathcal{A}_\alpha} \mathbf{E}_f^{(\varepsilon)} w\left(\sigma_\alpha^{-1}(\varepsilon^2)(\tilde{f}_\varepsilon(\mathbf{x}) - f(\mathbf{x})) \right) = \mathbf{E}\, w(\xi) \,. \tag{3.25}$$

Moreover (3.25) holds uniformly with respect to any uniform scale $\{\mathcal{A}_{\alpha_\varepsilon}\}$.

The proof is standard; cf. Lepski and Levit [12] , Theorems 3.1, 3.2. The upper bound in (3.25) follows easily from the above relations. To obtain the corresponding lower bound, consider, following [12], a subfamily of functions

$$f_\theta(\mathbf{y}) = \theta\varepsilon \|k_B\|^{-1} k_B(\mathbf{y} - \mathbf{x}),$$

where $B = B_\alpha$ is chosen according to (3.8), (3.18) and (3.20). Note first that by (2.7), (2.9) and (3.6)

$$f_\theta(\mathbf{x}) = \theta\varepsilon\|k_B\|^{-1}k_B(0) = \theta\varepsilon\|k_B\| = \theta\sigma_\alpha(\varepsilon).$$

Thus, restricted to the family $f_\theta(\cdot)$, the loss function in (3.25) becomes

$$w\left(\sigma_\alpha^{-1}(\varepsilon)\left(\tilde{f}_\varepsilon(\mathbf{x}) - f(\mathbf{x})\right)\right) = w\left(\sigma_\alpha^{-1}(\varepsilon)\left(\tilde{f}_\varepsilon(\mathbf{x}) - \theta\sigma_\alpha(\varepsilon)\right)\right) = w(\tilde{\theta} - \theta).$$

Next by (3.3) and (3.8) the inequality

$$|\hat{f}_\theta(\mathbf{t})| \leq \alpha(\mathbf{t}) \tag{3.26}$$

is equivalent to

$$\frac{\pi^d\theta^2\varepsilon^2|\Gamma|}{\prod_{i=1}^d \nu_i}\,\hat{k}(B_{2\alpha}^{-1}\Gamma\mathbf{t}) \leq \alpha^2(\mathbf{t}) = R^2|\Gamma|^2\exp_{\mathbf{r}}^2(\Gamma\mathbf{t}). \tag{3.27}$$

Therefore $f_\theta(\cdot) \in \mathcal{A}_\alpha$ for all θ such that

$$\frac{\pi^d\theta^2\varepsilon^2}{|\Gamma|\prod_{i=1}^d \nu_i} \leq R^2\exp_{\mathbf{r}}^2(\boldsymbol{\nu}), \tag{3.28}$$

with $\nu_i = \nu_{i\alpha}(\varepsilon^2)$ defined by (3.20). According to (3.16)

$$\frac{\pi^d\varepsilon^2}{|\Gamma|\prod_{i=1}^d \nu_i} \leq 2R^2\exp_{\mathbf{r}}^2(\boldsymbol{\nu})c^{1-2d},$$

and therefore $f_\theta(\cdot) \in \mathcal{A}_\alpha$ for all θ such that

$$\theta^2 \leq c^2(\varepsilon) \stackrel{\text{def}}{=} \frac{c^{2d-1}}{2}. \tag{3.29}$$

The proof can now be concluded as in [12] using the well-known LAN methods and the Cameron-Martin-Girsanov formula (2.2).

A convenient reparametrization. Up till now we could think of the index-set $\mathcal{K}$ of an arbitrary scale $\mathcal{A}_{\mathcal{K}}$ as being specified by the parameters $(\mathbf{r}, \Gamma)$; cf. (2.3)-(2.4). Below we introduce a different parametrization of uniform scales that satisfy (3.24) which will turn out to be more convenient in the rest of the paper.

Without loss of generality one can normalize the structural matrix Γ as follows

$$\Gamma = \operatorname{diag}(\boldsymbol{\gamma})\, \Gamma_a \tag{3.30}$$

where $\Gamma_a^{-1} \in \mathfrak{A}_a$ represents the angular parameters, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d) > 0$ are the band-scale parameters. The above representation (3.8)–(3.9) then becomes

$$B_\alpha \overset{\text{def}}{=} B_{a\alpha} B_{b\alpha} = \Gamma_a^{-1} \operatorname{diag}(\mathbf{s}), \tag{3.31}$$

where

$$s_i = s_{i\alpha}(\varepsilon^2) = \frac{\nu_i}{\gamma_i} = \frac{1}{\gamma_i} \left(\frac{c}{r_i} \right)^{1/r_i}, \tag{3.32}$$

with c defined by (3.18). The relevance of such a representation is obvious from the fact that both the bandwidth-matrix B_α above and the asymptotic variance

$$S_\alpha(\varepsilon^2) = \frac{c^{1/r}(|\Gamma|)}{\pi^d r^* |\Gamma|} = \frac{1}{\pi^d |\Gamma_a|} \prod_{i=1}^d s_i, \tag{3.33}$$

cf. (3.10), (3.32) can be easily expressed in terms of the parameters

$$\alpha = \alpha(\mathbf{r}, \Gamma_a, \mathbf{s}). \tag{3.34}$$

Thus in the sequel we can – and will – think of the uniform scales $\mathcal{A}_{\mathcal{K}}$ as being indexed by (3.34). Note that s_i can be interpreted as the *difficulty* of estimating functions f belonging to $\mathcal{A}_\alpha$, in the direction determined by the i-th column-vector of the angular matrix Γ_α^{-1}.

Let us demonstrate that these parameters uniquely define the class $\mathcal{A}_\alpha$, for all sufficiently small $\varepsilon^2/|\Gamma|$, say $\varepsilon^2/|\Gamma| < \varepsilon_0$. Indeed by (3.18) it is sufficient to demonstrate that these parameters determine Γ, from which one can successively recover $c(|\Gamma|)$, $\boldsymbol{\nu}$ and $\boldsymbol{\gamma}$. By (3.33) we need only show that $v(|\Gamma|) = c^{1/r}(|\Gamma|)/|\Gamma|$ is an invertible function, for $\varepsilon^2/|\Gamma| < \varepsilon_0$. Using (3.18) it is easy to show that, under the assumption (3.24),

$$v'(|\Gamma|) = -\frac{v(|\Gamma|)}{|\Gamma|}(1 + o(1)) < 0.$$

4 Adaptive estimation for functional scales

Essentially non-parametric scales. In the adaptive setting our attention shifts from fixed functional classes $\mathcal{A}_\alpha$ to functional scales $\mathcal{A}_\mathcal{K}$. To motivate our discussion of such scales let us concentrate for a moment on the bandwidth matrix B_α in (3.31). Now the smoothness of f is determined essentially by Γ_a and the band-scale parameters $\mathbf{s}$. To avoid a possible degeneration of our functional classes we will consider only scales $\mathcal{K} = \mathcal{K}_\varepsilon$ such that

$$\liminf_{\varepsilon \to 0} \inf_{\alpha \in \mathcal{K}} |B_{a\alpha}| > 0. \tag{4.1}$$

However, the consistency condition (3.24) leaves us a great deal of freedom in determining the parameters $\mathbf{s}$. To obtain a better insight into different situations that may transpire, let us try to recognize those $\mathbf{s}$ which occur 'typically' in non-parametric problems.

One of the characteristic features that one might expect from any non-parametric problem, on the basis of past experience and intuition, is that the bandwidths would tend to 0 as $\varepsilon \to 0$. This means in our case that the parameters s_i in (3.31) are all becoming large in a typical non-parametric problem, whereas s_i remaining bounded may be regarded as non-typical, unless the smoothness of a function allows a *parametric* rate of estimation 'in some direction'. Such a situation can actually occur only when the smoothness of a function is increasing in some direction when $\varepsilon \to 0$. Let us make this discussion more concise by introducing the following definition.

Definition 4.1. A scale $\mathcal{A}_\mathcal{K}$ is called *essentially non-parametric* or an *NP*-scale, if

$$\lim_{\varepsilon \to 0} \inf_{\alpha \in \mathcal{K}} \min_i s_i = \infty\,. \tag{4.2}$$

Otherwise $\mathcal{A}_{\mathcal{K}_\varepsilon}$ is called *partially parametric* or a *PP*-scale.

Aggregated classes. To formulate our results describing asymptotic performance of an optimal adaptive procedure with regard to the classes $\mathcal{A}_\alpha$ we will need to pool together some of these classes. For a given functional scale $\mathcal{A}_\mathcal{K}$ and arbitrary $S = S(\varepsilon^2) > 0$ denote

$$\mathcal{A}_{\mathcal{K}(S)} = \bigcup_{\alpha \in \mathcal{K}} \left\{ \mathcal{A}_\alpha : S_\alpha(\varepsilon^2) = S \right\},$$

where $S_\alpha(\varepsilon^2)$ was introduced earlier by (3.21). Obviously

$$\bigcup_{\alpha \in \mathcal{K}} \mathcal{A}_\alpha = \bigcup_{S > 0} \mathcal{A}_{\mathcal{K}(S)}.$$

Thus if a scale $\mathcal{A}_{\mathcal{K}}$ is essentially non-parametric then for any S_0 and all sufficiently small ε

$$\bigcup_{S<S_0} \mathcal{A}_{\mathcal{K}(S)} = \emptyset,$$

whereas a scale $\mathcal{A}_{\mathcal{K}}$ is partially parametric, if for some S_0 and all sufficiently small ε

$$\bigcup_{S>S_0} \mathcal{A}_{\mathcal{K}(S)} = \emptyset.$$

For any uniform scale $\mathcal{A}_{\mathcal{K}}$, the mean square error of the minimax estimators $\tilde{f}_{B_\alpha}$ is asymptotically bounded by $\varepsilon^2 S$ uniformly over $\mathcal{A}_\alpha \subset \mathcal{A}_{\mathcal{K}}(S)$, cf. Theorem 3.1. Recent ventures into the adaptive non-parametric problems have made it increasingly clear that in an adaptive setting a $\log S$ factor is an unavoidable price for not knowing the functional classes $\mathcal{A}_\alpha$ exactly. Lepski (1991,1992a) was the first to formulate corresponding general results. Another demonstration of this fact is exhibited by our Theorem 6.1 below.

The main aim of this paper is to evaluate the asymptotically exact risk of the optimal adaptive minimax estimators, with regard to the aggregated classes $\mathcal{A}_{\mathcal{K}(S)}$. Such results for ample non-parametric scales $\mathcal{A}_{\mathcal{K}}$ will be presented in the following sections. However, it is apparently impossible to obtain similar results for the *PP*-scales, within the same classes of procedures; cf. Lepski and Levit [12] . Our viewpoint here is similar to that of Lepski and Levit [12] , namely provided the rate $(\varepsilon^2 S \log S)^{1/2}$ can be guaranteed for such scales, they still receive privileged risk treatment compared to some *NP*-scales, even without pursuit of the optimal constants.

To help the formulation of these results, let $\mathfrak{F}_p = \mathfrak{F}_p(\mathbf{x})$ be the class of all estimators $\tilde{f}_\varepsilon$ satisfying the relation

$$\limsup_{\varepsilon\to 0} \sup_{S>0} \sup_{f\in\mathcal{A}_{\mathcal{K}(S)}} \mathbf{E}_f^{(\varepsilon)} \left|(\varepsilon^2 S \log S)^{-1/2}(\tilde{f}_\varepsilon(\mathbf{x}) - f(\mathbf{x}))\right|^p < \infty \tag{4.3}$$

for any partially-parametric scale $\mathcal{A}_{\mathcal{K}}$.

Below we will develop asymptotically optimal adaptive estimation procedures for arbitrary *NP*-scales, under the restriction that such procedures are rate efficient with respect to arbitrary *PP*-scales. Accordingly, the role of generic classes $\mathfrak{F}$ in a formal definition below will be played by $\mathfrak{F}_p(\mathbf{x})$. However, the particular dichotomy between *NP*- and *PP*-scales will be needed only to fix a reference point. The careful reader will note that there are many different ways of doing the same thing. For example, in defining *NP*- and *PP*-scales, one could postulate an arbitrary asymptotic bound $g(\varepsilon)$ on the band-scale parameters in (4.2) (by replacing s_i by $g(\varepsilon)s_i$) and draw a distinction there. We choose $g = 1$ here simply to illustrate the whole idea. Besides, such a choice can be interpreted easily in terms of the parametric–non-parametric slogan.

Let $\mathfrak{F}$ be a class of (sequences of) estimators $\tilde{f}_\varepsilon(\mathbf{x})$ and $\mathcal{A}_\mathcal{K}$, $\mathcal{K} = \mathcal{K}_\varepsilon \subset \mathcal{K}_0$, a functional scale. The following convention is used below in (4.4): if some of the classes $\mathcal{A}_{\mathcal{K}(S)}$ are empty the corresponding *sup* is set to be 1.

Definition 4.2. An estimator $\hat{f}_\varepsilon \in \mathfrak{F}$ is called a) $(p, \mathcal{K}, \mathfrak{F})$-*adaptively minimax*, at a point $\mathbf{x} \in \mathbf{R}$, if for any other estimator $\tilde{f}_\varepsilon \in \mathfrak{F}$

$$\limsup_{\varepsilon\to 0} \sup_{S>0} \frac{\sup_{f\in\mathcal{A}_{\mathcal{K}(S)}} \mathbf{E}_f^{(\varepsilon)} |\hat{f}_\varepsilon(\mathbf{x}) - f(\mathbf{x})|^p}{\sup_{f\in\mathcal{A}_{\mathcal{K}(S)}} \mathbf{E}_f^{(\varepsilon)} |\tilde{f}_\varepsilon(\mathbf{x}) - f(\mathbf{x})|^p} \le 1\,; \tag{4.4}$$

b) *adaptively rate minimax* at $\mathbf{x}$, if the above limit is finite.

5 Adaptive estimation procedure

Below we will use following notations:

$$\phi^2(S) = (p + 2(d-1)) \log S, \tag{5.1}$$

$$\psi^2(S) = \varepsilon^2 S \phi^2(S). \tag{5.2}$$

Theorem 5.1. *There exists a sequence of estimators $\hat{f}_\varepsilon$ such that for any $\mathbf{x} \in \mathbb{R}^d$ and any NP-scale $\mathcal{A}_\mathcal{K}$ $\hat{f}_\varepsilon \in \mathfrak{F}_p(\mathbf{x})$ and*

$$\limsup_{\varepsilon\to 0} \sup_{S>0} \sup_{f\in\mathcal{A}_{\mathcal{K}(S)}} \mathbf{E}_f^{(\varepsilon)} \left|\psi^{-1}(S)\left(\hat{f}_\varepsilon(\mathbf{x}) - f(\mathbf{x})\right)\right|^p \le 1. \tag{5.3}$$

An estimator with the above properties will be described shortly. In Section 6 we will show that under a mild additional assumption such an estimator is adaptively minimax, in the sense of Definition 4.2. The parameters $\mathbf{x}$, p will be fixed in the rest of the paper.

Estimators' net: facing the dimension. In this section we will be interested primarily in the following question: how 'big' does a discrete subset $\tilde{f}_\mathbf{i} = \tilde{f}_{B_i}$ of estimators have to be to provide a satisfactory approximation to all possible minimax estimators $\tilde{f}_{B_\alpha}$, in an arbitrary non-parametric scale $\mathcal{A}_\mathcal{K}$? First we will describe a representative net of estimators $\tilde{f}_\mathbf{i}$ and evaluate its cardinality. In the subsequent sections we will present a selection procedure for comparing and choosing between such estimators. The optimality results are deferred to Section 6.

In accordance with (3.31) we consider estimators $\bar{f}_{\mathbf{i}} = \tilde{f}_{B_{\mathbf{i}}}$ such that

$$\mathrm{B}_{\mathbf{i}} = B_{a\mathbf{i}}\, B_{b\mathbf{i}}$$

where

$$B_{a\mathbf{i}} \in \mathfrak{A}_a, \qquad B_{b\mathbf{i}} = \mathrm{diag}(\mathbf{s}_{\mathbf{i}}), \qquad \mathbf{s}_{\mathbf{i}} > 0. \tag{5.4}$$

We will refer to $B_{\alpha\mathbf{i}}$ as angular parameters of the estimators $\bar{f}_{\mathbf{i}}$ and $s_{\mathbf{i}}$ as the band-scale parameters.

Let $0 < \varrho < 1$ be fixed and

$$s_i = \exp(i^{\varrho}), \qquad i = 0, 1, \ldots . \tag{5.5}$$

The diagonal elements of $B_{b\mathbf{i}}$ in (5.4) will comprise all possible collections of s_i. More precisely for any $(i_1, ..., i_d)$ we set $\mathbf{s}_{\mathbf{i}} = (s_{i_1}, \ldots, s_{i_d})$.

Next assume that for any $(i_1, ..., i_d)$ a set $\mathcal{B}_{\mathbf{i}} = \{B_q,\ q = 1, \ldots, Q(\mathbf{i})\} \subset \mathfrak{A}_a$ has been fixed in such a way that for some $m_1, m_2, M_1, M_2, M_3 > 0$

$$Q(\mathbf{i}) \le M_1 (s_{i_1} \cdots s_{i_d})^{d-1} \log^{m_1}(s_{i_1} \cdots s_{i_d}), \tag{5.6}$$

$$\inf_{B \in \mathcal{B}_{\mathbf{i}}} |B| \ge M_2 \log^{-m_2}(s_{i_1} \cdots s_{i_d}), \tag{5.7}$$

$$\sup_{\|v\|=1} \inf_{B \in \mathcal{B}_{\mathbf{i}}} \|v - B_{\cdot j}\| \le M_3 s_{i_j}^{-1}, \qquad j = 1, \ldots, d. \tag{5.8}$$

By definition, the bandwidth matrix $B_{\mathbf{i}}$, supplied with the multi-index

$$\mathbf{i} = (q, i_1, \ldots, i_d), \qquad i_j = 0, 1, ..., \qquad q = 1, \ldots, Q(\mathbf{i}),$$

is

$$B_{\mathbf{i}} = B_q \mathrm{diag}(\mathbf{s}_{\mathbf{i}}).$$

Denote also

$$\bar{f}_{\mathbf{i}} = \tilde{f}_{B_{\mathbf{i}}}, \qquad\qquad S_{\mathbf{i}} = |B_{\mathbf{i}}|,$$

$$N(S) = \#\{\mathbf{i} : S_{\mathbf{i}} \le S\}, \qquad N(q_0; S) = \#\{\mathbf{i} : S_{\mathbf{i}} \le S,\ q = q_0\}.$$

Proposition 5.1. *There exist $m_4, M_4 > 0$ such that for any q_0 and $S \geq 2$*

$$N(S) \leq M_4 S^{d-1} \log^{m_4} S,$$

$$N(q_0; S) \leq M_4 \log^{m_4} S.$$

Proof. By (5.5)–(5.7) there exist $m_5, M_5 > 0$ such that for all $\mathbf{i}$ with $S_{\mathbf{i}} \leq S$

$$\exp(i_1^\varrho + \cdots + i_d^\varrho) \leq M \stackrel{\text{def}}{=} M_5 S \log^{m_5} S.$$

There are at most $(\log^{1/\varrho} M + 1)^d$ different values of the multi-index $\mathbf{i}$ which satisfy this inequality. Therefore, according to (5.6) for some $m_4, M_4 > 0$ and all $S \geq 2$

$$N(S) \leq M_1 S^{d-1} (\log^{1/\varrho} M + 1)^d \log^{m_1} S \leq M_4 S^{d-1} \log^{m_4} S.$$

The second inequality follows in a similar way. $\square$

Note that $N(S)$ grows much more slowly when the angular parameters are fixed. Our proof of Theorem 5.1 below hinges heavily on this idea. The following result demonstrates that the thus defined estimators $\bar{f}_{\mathbf{i}}$ provide a sufficiently good approximation to the minimax estimators $\tilde{f}_{B_\alpha}$, $\alpha \in \mathcal{K}_0$. Recall that we assume that the condition (4.1) holds.

Proposition 5.2. *For any $\alpha \in \mathcal{K}_0$ there exists $\mathbf{i}(\alpha) = (q(\alpha), i_1(\alpha), , \ldots, i_d(\alpha))$ such that*

$$B_\alpha \Subset B_{\mathbf{i}(\alpha)}, \tag{5.9}$$

whereas uniformly in any NP-scale

$$B_\alpha \preccurlyeq B_{\mathbf{i}(\alpha)} \preccurlyeq (1 + o(1)) B_\alpha. \tag{5.10}$$

Proof. Let us determine first the band-scale parameters of the matrix $B_{\mathbf{i}(\alpha)}$. For any sufficiently small ε^2 there exist $i_j = i_j(\alpha)$ such that

$$s_{i_j - 2} < s_{j\alpha}(\varepsilon^2) \leq s_{i_j - 1}. \tag{5.11}$$

By (4.2), (5.5) uniformly in any *NP*-scale $\lim_{\varepsilon \to 0} i_j = \infty$, $j = 1, \ldots, d$, and

$$\frac{s_{i_j}}{s_{j\alpha}(\varepsilon^2)} = 1 + \delta_j, \tag{5.12}$$

where $\boldsymbol{\delta} = (\delta_1, \dots, \delta_d)$,

$$\delta_j = (1+o(1))\varrho {i_j}^{\varrho-1} = (1+o(1))\varrho \log^{(\varrho-1)/\varrho} s_{j\alpha}(\varepsilon^2). \tag{5.13}$$

Now by (5.8) there exists a matrix $B_{a\mathbf{i}(\alpha)} \in \mathcal{B}_{\mathbf{i}(\alpha)}$ such that for some $M > 0$

$$\|(B_{a\alpha} - B_{a\mathbf{i}(\alpha)})B_{b\mathbf{i}(\alpha)}\| \le M.$$

Note that (5.9) is equivalent to the following inclusion:

$$\mathbf{C} \subseteq B_\alpha^{-1} B_{\mathbf{i}(\alpha)} \mathbf{C} = B_{b\alpha}^{-1} B_{a\alpha}^{-1} B_{a\mathbf{i}(\alpha)} B_{b\mathbf{i}(\alpha)} \mathbf{C}.$$

Denoting $\Delta = B_{a\mathbf{i}(\alpha)} - B_{a\alpha}$, this becomes

$$\mathbf{C} \subseteq B_{b\alpha}^{-1} B_{b\mathbf{i}(\alpha)} \mathbf{C} + B_{b\alpha}^{-1} B_{a\alpha}^{-1} \Delta B_{b\mathbf{i}(\alpha)} \mathbf{C} =$$

$$\mathrm{diag}(1+\boldsymbol{\delta})\mathbf{C} + O(1) B_{b\alpha}^{-1}\mathbf{C},$$

and (5.9) follows now by (5.12)-(5.13). The relation (5.10) follows directly from (5.12). □

Below we will assume that the matrices $B_{\mathbf{i}(\alpha)}$ have been fixed. However these will have no effect on the construction of estimators and will be used only as a strategy in proving the theorem.

How to choose between incomparable estimators? A selection procedure described below is a generalization of the method of constructing optimal adaptive estimators used in [12]. The method in [12] was based on pairwise comparisons of an essentially one-dimensional family of estimators $\bar{f}_{\mathbf{i}}$; see further references therein regarding the origin of this method. Let us recall briefly the idea of [12] procedure and explain why a radical modification is needed in the multivariate case.

One could easily carry out a comparison between any two competing estimators $\bar{f}_{\mathbf{i}}(\mathbf{x})$, $\bar{f}_{\mathbf{j}}(\mathbf{x})$ exactly as in [12] in the special case of ordered estimators:

$$\bar{f}_{\mathbf{i}}(\mathbf{x}) \Subset \bar{f}_{\mathbf{j}}(\mathbf{x}). \tag{5.14}$$

As in [12] the differences $\bar{f}_{\mathbf{j}}(\mathbf{x}) - \bar{f}_{\mathbf{i}}(\mathbf{x})$ are then Gaussian random variables, with known variances $\bar{\sigma}_{\mathbf{ij}}^2$ and expectations equal to the bias differences $\bar{b}_{\mathbf{j}} - \bar{b}_{\mathbf{i}}$. Now (5.14) implies that

$$\bar\sigma_{\mathbf{i}}^2 \le \bar\sigma_{\mathbf{j}}^2$$

and

$$\bar\sigma_{\mathbf{ij}}^2 \le \bar\sigma_{\mathbf{j}}^2.$$

Suppose that one of these estimators, say $\bar f_{\mathbf{i}}(\mathbf{x})$, coincided with $\bar f_{\mathbf{i}(\alpha)}$, which for all our purposes can be thought as the 'true' estimator. Then, as we have seen in Section 3, cf. (3.7), (3.22), both biases are small:

$$\bar b_{\mathbf{i}}, \bar b_{\mathbf{j}} \ll \bar\sigma_{\mathbf{i}} \le \bar\sigma_{\mathbf{j}}.$$

Therefore except for negligible non-random terms, the random variables

$$\frac{\bar f_{\mathbf{j}}(\mathbf{x}) - \bar f_{\mathbf{i}}(\mathbf{x})}{\bar\sigma_{\mathbf{j}}}$$

are zero-mean Gaussian with variances not exceeding 1. Now if a sequence of thresholds $\lambda_{\mathbf{j}}$ was properly chosen, none of these variables would exceed corresponding thresholds $\lambda_{\mathbf{j}}$, with a high probability. Thus the fulfillment of events

$$\left|\frac{\bar f_{\mathbf{j}}(\mathbf{x}) - \bar f_{\mathbf{i}}(\mathbf{x})}{\bar\sigma_{\mathbf{j}}}\right| \le \lambda_{\mathbf{j}}, \quad \forall\, \mathbf{j} \Supset \mathbf{i} \tag{5.15}$$

would circumstantially suggest that $\bar f_{\mathbf{i}}$ can be a good candidate for the 'true' estimator. Among all such candidates one could choose e.g. the one with the smallest variance $\bar\sigma_{\hat{\mathbf{i}}}^2$.

This procedure would guarantee that the 'true' estimator $\bar f_{\mathbf{i}(\alpha)}$ is among the good candidates. Moreover, had we chosen a different estimator $\bar f_{\hat{\mathbf{i}}} \preccurlyeq \bar f_{\mathbf{i}(\alpha)}$, the inequality (5.15) would provide sufficient control of the deviation of $\bar f_{\hat{\mathbf{i}}}$ from the 'true' estimator $\bar f_{\mathbf{i}(\alpha)}$. The exact choice of the thresholds then becomes a balancing act based on large deviation calculations: $\lambda_{\mathbf{j}}$ should be big enough to guarantee a sufficiently high probability of (5.15), and at the same time small enough to provide good control of $\bar f_{\hat{\mathbf{i}}} - \bar f_{\mathbf{i}(\alpha)}$. For the moment we will leave this at this merely intuitive level, in order not to obscure the main idea.

Obviously everything in the reasoning just sketched hinged on the fact that the estimators $\bar f_{\mathbf{i}}, \bar f_{\mathbf{j}}$ were comparable in the sense of (5.14). However, since the set of estimators $\bar f_{\mathbf{i}}$ is only partially ordered – unless $d = 1$ – the relation (5.14) does not need to hold in general. Thus a rather radical modification of the above procedure is needed!

The main idea of the adaptive procedure proposed in this paper can be summarized as follows. In order to preserve the pivotal order relation (5.14), the estimator $\bar f_{\mathbf{j}}$ appearing in (5.14) is replaced by the smallest majorant $\bar f_{\mathbf{i},\mathbf{j}} = \bar f_{\mathbf{i}} \vee \bar f_{\mathbf{j}}$ with respect to the order relation $\Subset$. Obviously when (5.14) holds (in particular, for $d = 1$) the method remains essentially the same. In general, however, it leads to a substantial revision of the approach used in

[12] as well as to a modified version of the optimality results as presented by Theorems 5.1 and 6.1.

Upon recalling (3.2), (3.4)–(3.6), (3.10) let

$$\bar{k}_{\mathbf{i}}(\mathbf{x}) = k_{B_{\mathbf{i}}}(\mathbf{x}), \qquad \hat{k}_{\mathbf{i}}(\mathbf{t}) = \hat{k}_{B_{\mathbf{i}}}(\mathbf{t}),$$

$$\bar{\sigma}_{\mathbf{i}}^2 = \sigma_{B_{\mathbf{i}}}^2 = \varepsilon^2 S_{\mathbf{i}}, \qquad \bar{b}_{\mathbf{i}} = b_{B_{\mathbf{i}}},$$

and

$$\bar{b}_{\mathbf{i}}(\alpha) = b(\alpha, B_{\mathbf{i}}),$$

so that

$$\bar{f}_{\mathbf{i}}(\mathbf{x}) - f(\mathbf{x}) = \bar{\sigma}_{\mathbf{i}}\xi_{\mathbf{i}} + \bar{b}_{\mathbf{i}}, \tag{5.16}$$

where $\xi_{\mathbf{i}} \sim \mathcal{N}(0,1)$. Denote also

$$\bar{\psi}_{\mathbf{i}} = \psi(S_{\mathbf{i}}),$$

$$\bar{v}_{\mathbf{i}} = \frac{\bar{b}_{\mathbf{i}}}{\bar{\sigma}_{\mathbf{i}}}.$$

For a pair of kernels $\bar{k}_{\mathbf{i}}, \bar{k}_{\mathbf{j}}$ the smallest majorant with respect to the order relation $\Subset$ is obviously given by

$$\bar{k}_{\mathbf{i},\mathbf{j}}(\mathbf{x}) = \bar{k}_{\mathbf{i}}(\mathbf{x}) + \bar{k}_{\mathbf{j}}(\mathbf{x}) - \bar{k}_{\mathbf{i}}(\mathbf{x}) * \bar{k}_{\mathbf{j}}(\mathbf{x}), \tag{5.17}$$

i.e. by the smallest *spectral majorant*

$$\hat{k}_{\mathbf{i},\mathbf{j}}(\mathbf{t}) = \max(\hat{k}_{\mathbf{i}}(\mathbf{t}), \hat{k}_{\mathbf{j}}(\mathbf{t})). \tag{5.18}$$

Consider also the corresponding estimator

$$\bar{f}_{\mathbf{i},\mathbf{j}}(\mathbf{x}) = \int \bar{k}_{\mathbf{i},\mathbf{j}}(\mathbf{x}-\mathbf{y})\, dV(\mathbf{y}). \tag{5.19}$$

Following the above discussion, our selection procedure will be based on a comparison between the differences $\bar{f}_{\mathbf{i},\mathbf{j}}(\mathbf{x}) - \bar{f}_{\mathbf{i}}(\mathbf{x})$ and the corresponding thresholds $\lambda_{\mathbf{j}}$. One has

$$\bar{f}_{\mathbf{i},\mathbf{j}}(\mathbf{x}) - \bar{f}_{\mathbf{i}}(\mathbf{x}) = \bar{\sigma}_{\mathbf{i},\mathbf{j}}\xi_{\mathbf{i},\mathbf{j}} + \bar{b}_{\mathbf{i},\mathbf{j}}, \tag{5.20}$$

where $\xi_{\mathbf{i},\mathbf{j}} \sim \mathcal{N}(0,1)$,

$$\bar{b}_{\mathbf{i},\mathbf{j}} \stackrel{\text{def}}{=} \bar{v}_{\mathbf{i},\mathbf{j}}\bar{\sigma}_{\mathbf{i},\mathbf{j}} = \mathbf{E}\left(\bar{f}_{\mathbf{i},\mathbf{j}}(\mathbf{x}) - \bar{f}_{\mathbf{i}}(\mathbf{x})\right),$$

$$\bar{\sigma}^2_{\mathbf{i},\mathbf{j}} = \mathbf{Var}(\bar{f}_{\mathbf{i},\mathbf{j}} - \bar{f}_{\mathbf{i}}) \stackrel{\text{def}}{=} \varepsilon^2 S_{\mathbf{i},\mathbf{j}} = \varepsilon^2 \|\bar{k}_{\mathbf{j}} - \bar{k}_{\mathbf{i}} * \bar{k}_{\mathbf{j}}\|^2.$$

The next proposition follows easily from the above definitions and (3.22).

Proposition 5.3. *For any pair of kernels $\bar{k}_{\mathbf{i}}, \bar{k}_{\mathbf{j}}$ the following holds:*

(a)

$$\bar{\sigma}^2_{\mathbf{i},\mathbf{j}} \le \bar{\sigma}^2_{\mathbf{j}}; \tag{5.21}$$

(b) for any $f \in \mathcal{A}_\alpha$

$$\left|\bar{b}_{\mathbf{i},\mathbf{j}}\right| \le \bar{b}_{\mathbf{i}}(\alpha); \tag{5.22}$$

(c) for f varying in a uniform scale, $\mathbf{i} \Supset \mathbf{i}(\alpha)$ and all sufficiently small ε

$$|\bar{v}_{\mathbf{i}}| \le 1, \quad |\bar{v}_{\mathbf{i},\mathbf{j}}| \le 1. \tag{5.23}$$

Selection procedure. Let us introduce a sequence of thresholds

$$\lambda_{\mathbf{j}} = \sqrt{p_1 \log S_{\mathbf{j}} + p_2 \log^\delta S_{\mathbf{j}}}$$

where $p_1 = p + 2(d-1)$, and $1/2 < \delta < 1$ and $p_2 > 0$ are arbitrary constants. According to [12] a selection procedure could be based on the following inequality:

$$\left|\bar{f}_{\mathbf{i},\mathbf{j}}(\mathbf{x}) - \bar{f}_{\mathbf{i}}(\mathbf{x})\right| \le \bar{\sigma}_{\mathbf{i},\mathbf{j}}\lambda_{\mathbf{j}}. \tag{5.24}$$

Such an approach can be realized with a slightly modified version of $S_{\mathbf{i}}$. However, to avoid any further complications we will replace $\bar{\sigma}_{\mathbf{i},\mathbf{j}}$ in (5.24) by $\bar{\sigma}_{\mathbf{j}}$; cf. (5.15). From our numerical experience this is also likely to make the procedure computationally more robust.

By definition, $\mathbf{i} \in \mathcal{I} = \mathcal{I}(\mathbf{x})$ if for all $\mathbf{j} \succcurlyeq \mathbf{i}$

$$\left|\bar{f}_{\mathbf{i},\mathbf{j}}(\mathbf{x}) - \bar{f}_{\mathbf{i}}(\mathbf{x})\right| \le \bar{\sigma}_{\mathbf{j}}\lambda_{\mathbf{j}}, \tag{5.25}$$

whereas

$$\hat{\mathbf{i}} = \arg\min_{\mathbf{i}\in\mathcal{I}} S_{\mathbf{i}}.$$

If the above minimum is not unique, the index $\hat{\mathbf{i}}$ should be determined by independent randomization. Finally we define our adaptive estimator pointwise as

$$\hat{f}_{\varepsilon}(\mathbf{x}) = \bar{f}_{\hat{\mathbf{i}}}(\mathbf{x}).$$

In the following section we will demonstrate that thus defined estimator satisfies (5.3). In the proof the following elementary result will be used.

Lemma 5.1. *For any* $\mathbf{i} \preccurlyeq \mathbf{j}$ *the random variables* $\xi_{\mathbf{j},\mathbf{i}}$ *and* $\chi\{\hat{\mathbf{i}} = \mathbf{i}\}$ *are independent.*

Proof. Let us fix an index $\mathbf{i}$. For an arbitrary measurable subset $U \subset \mathbb{R}^d$ consider functions $k_U(\mathbf{x})$, $k_U^c(\mathbf{x})$ such that

$$\hat{k}_U(\mathbf{x}) = \hat{k}_{\mathbf{i}}(\mathbf{t})\chi_U(\mathbf{t}),$$

$$\hat{k}_U^c(\mathbf{x}) = (1 - \hat{k}_{\mathbf{i}}(\mathbf{t}))\chi_U(\mathbf{t}),$$

and let

$$\xi_U = \int k_U(\mathbf{x})\, dW(\mathbf{x}),\ \xi_U^c = \int k_U^c(\mathbf{x})\, dW(\mathbf{x}).$$

Obviously the σ-algebras $\mathfrak{B}_{\mathbf{i}}$, $\mathfrak{B}_{\mathbf{i}}^c$ generated by the Gaussian random variables $\{\xi_U\}$, $\{\xi_U^c\}$ respectively are independent. All we need to note now is that, by definition, the random variables $\xi_{\mathbf{j},\mathbf{i}}$ are $\mathfrak{B}_{\mathbf{i}}$-measurable, whereas $\chi\{\hat{\mathbf{i}} = \mathbf{i}\}$ are determined by the σ-algebra $\mathfrak{B}_{\mathbf{i}}^c$ and an independent randomization device.

Proof of the Theorem 5.1. Obviously for arbitrary f

$$\mathbf{E}_f^{(\varepsilon)} \left|\hat{f}_\varepsilon(\mathbf{x}) - f(\mathbf{x})\right|^p = \mathbf{E}_f^{(\varepsilon)} \left|\bar{f}_{\hat{\mathbf{i}}}(\mathbf{x}) - f(\mathbf{x})\right|^p \chi\left\{\hat{\mathbf{i}} \preccurlyeq \mathbf{i}(\alpha)\right\} +$$

$$\mathbf{E}_f^{(\varepsilon)} \left|\bar{f}_{\hat{\mathbf{i}}}(\mathbf{x}) - f(\mathbf{x})\right|^p \chi\left\{\hat{\mathbf{i}} \succ \mathbf{i}(\alpha)\right\} \stackrel{\text{def}}{=} \mathbf{R}_\varepsilon^-(f) + \mathbf{R}_\varepsilon^+(f). \tag{5.26}$$

Let us examine $\mathbf{R}_\varepsilon^-(f)$ first. By (5.25)

$$\left\{\hat{\mathbf{i}} \preccurlyeq \mathbf{i}(\alpha)\right\} \subset \left\{\left|\bar{f}_{\hat{\mathbf{i}},\mathbf{i}(\alpha)}(\mathbf{x}) - \bar{f}_{\hat{\mathbf{i}}}(\mathbf{x})\right| \leq \bar{\sigma}_{\mathbf{i}(\alpha)}\lambda_{\mathbf{i}(\alpha)}\right\}.$$

Therefore by (5.16), (5.20) and Proposition 5.3

$$\mathbf{R}_\varepsilon^-(f) \leq$$

$$\sum_{\mathbf{i}\preccurlyeq\mathbf{i}(\alpha)} \mathbf{E}_f^{(\varepsilon)} \left(\left|\bar{f}_{\mathbf{i}}(\mathbf{x}) - \bar{f}_{\mathbf{i},\mathbf{i}(\alpha)}(\mathbf{x})\right| + \left|\bar{f}_{\mathbf{i},\mathbf{i}(\alpha)}(\mathbf{x}) - \bar{f}_{\mathbf{i}(\alpha)}(\mathbf{x})\right| + \left|\bar{f}_{\mathbf{i}(\alpha)}(\mathbf{x}) - f(\mathbf{x})\right|\right)^p \chi\left\{\hat{\mathbf{i}} = \mathbf{i}\right\} \leq$$

$$\sum_{\mathbf{i}\preccurlyeq\mathbf{i}(\alpha)} \mathbf{E}_f^{(\varepsilon)} \left(\bar{\sigma}_{\mathbf{i}(\alpha)}\lambda_{\mathbf{i}(\alpha)} + \left|\bar{f}_{\mathbf{i}(\alpha),\mathbf{i}}(\mathbf{x}) - \bar{f}_{\mathbf{i}(\alpha)}(\mathbf{x})\right| + \left|\bar{f}_{\mathbf{i}(\alpha)}(\mathbf{x}) - f(\mathbf{x})\right|\right)^p \chi\left\{\hat{\mathbf{i}} = \mathbf{i}\right\} \leq$$

$$(\bar{\sigma}_{\mathbf{i}(\alpha)}\lambda_{\mathbf{i}(\alpha)})^p \sum_{\mathbf{i}\preccurlyeq\mathbf{i}(\alpha)} \mathbf{E}_f^{(\varepsilon)} \left(1 + \frac{\left|\sigma_{\mathbf{i}(\alpha),\mathbf{i}}\xi_{\mathbf{i}(\alpha),\mathbf{i}} + b_{\mathbf{i}(\alpha),\mathbf{i}}\right| + \left|\sigma_{\mathbf{i}(\alpha)}\xi_{\mathbf{i}(\alpha)} + b_{\mathbf{i}(\alpha)}\right|}{\bar{\sigma}_{\mathbf{i}(\alpha)}\lambda_{\mathbf{i}(\alpha)}}\right)^p \chi\left\{\hat{\mathbf{i}} = \mathbf{i}\right\} \leq$$

$$(\bar{\sigma}_{\mathbf{i}(\alpha)}\lambda_{\mathbf{i}(\alpha)})^p \sum_{\mathbf{i}\preccurlyeq\mathbf{i}(\alpha)} \mathbf{E}_f^{(\varepsilon)} \left(1 + \frac{1}{\lambda_{\mathbf{i}(\alpha)}}\left(|\xi_{\mathbf{i}(\alpha),\mathbf{i}}| + |\xi_{\mathbf{i}(\alpha)}| + 2|\bar{v}_{\mathbf{i}(\alpha)}|\right)\right)^p \chi\left\{\hat{\mathbf{i}} = \mathbf{i}\right\}. \tag{5.27}$$

If $p \le 1$, using the c_r-inequality, (5.23) and Lemma 5.1 we can bound (5.27) further as follows

$$\mathbf{R}_\varepsilon^-(f) \le (\bar{\sigma}_{\mathbf{i}(\alpha)}\lambda_{\mathbf{i}(\alpha)})^p \sum_{\mathbf{i}\preccurlyeq\mathbf{i}(\alpha)} \mathbf{E}_f^{(\varepsilon)} \left(1 + \frac{1}{\lambda_{\mathbf{i}(\alpha)}^p}\left(|\xi_{\mathbf{i}(\alpha),\mathbf{i}}| + |\xi_{\mathbf{i}(\alpha)}| + 2\right)^p\right) \chi\left\{\hat{\mathbf{i}} = \mathbf{i}\right\}$$

$$\le (\bar{\sigma}_{\mathbf{i}(\alpha)}\lambda_{\mathbf{i}(\alpha)})^p \sum_{\mathbf{i}\preccurlyeq\mathbf{i}(\alpha)} \mathbf{E}_f^{(\varepsilon)} \left(1 + \frac{1}{\lambda_{\mathbf{i}(\alpha)}^p}\left(|\xi_{\mathbf{i}(\alpha),\mathbf{i}}|^p + |\xi_{\mathbf{i}(\alpha)}|^p + 2^p\right)\right) \chi\left\{\hat{\mathbf{i}} = \mathbf{i}\right\}$$

$$\le (\bar{\sigma}_{\mathbf{i}(\alpha)}\lambda_{\mathbf{i}(\alpha)})^p \left(1 + \frac{1}{\lambda_{\mathbf{i}(\alpha)}^p}\left(\mathbf{E}_f^{(\varepsilon)}|\xi_{\mathbf{i}(\alpha)}|^p + 2^p + \sum_{\mathbf{i}\preccurlyeq\mathbf{i}(\alpha)} \mathbf{E}_f^{(\varepsilon)}|\xi_{\mathbf{i}(\alpha),\mathbf{i}}|^p \chi\left\{\hat{\mathbf{i}} = \mathbf{i}\right\}\right)\right)$$

$$= (\bar{\sigma}_{\mathbf{i}(\alpha)}\lambda_{\mathbf{i}(\alpha)})^p \left(1 + \frac{1}{\lambda_{\mathbf{i}(\alpha)}^p}\left(O(1) + \sum_{\mathbf{i}\preccurlyeq\mathbf{i}(\alpha)} \mathbf{E}_f^{(\varepsilon)}|\xi_{\mathbf{i}(\alpha),\mathbf{i}}|^p \mathbf{E}_f^{(\varepsilon)} \chi\left\{\hat{\mathbf{i}} = \mathbf{i}\right\}\right)\right)$$

$$= (\bar{\sigma}_{\mathbf{i}(\alpha)}\lambda_{\mathbf{i}(\alpha)})^p \left(1 + \frac{O(1)}{\lambda_{\mathbf{i}(\alpha)}^p}\right). \tag{5.28}$$

If $p > 1$, we first use a convexity inequality in (5.27): for any $0 < \eta < 1$

$$\mathbf{R}_\varepsilon^-(f) \le$$

$$(\bar{\sigma}_{\mathbf{i}(\alpha)}\lambda_{\mathbf{i}(\alpha)})^p \left(\frac{1}{(1-\eta)^{p-1}} + \frac{1}{\eta^{p-1}} \sum_{\mathbf{i}\preccurlyeq\mathbf{i}(\alpha)} \mathbf{E}_f^{(\varepsilon)} \left(\frac{1}{\lambda_{\mathbf{i}(\alpha)}}\left(|\xi_{\mathbf{i}(\alpha),\mathbf{i}}| + |\xi_{\mathbf{i}(\alpha)}| + 2\right)\right)^p \chi\left\{\hat{\mathbf{i}} = \mathbf{i}\right\}\right) \le$$

$$(\bar{\sigma}_{\mathbf{i}(\alpha)}\lambda_{\mathbf{i}(\alpha)})^p \left(\frac{1}{(1-\eta)^{p-1}} + \frac{1}{\eta^{p-1}} \sum_{\mathbf{i}\preccurlyeq\mathbf{i}(\alpha)} \mathbf{E}_f^{(\varepsilon)} \left(\frac{3^{p-1}}{\lambda_{\mathbf{i}(\alpha)}^p}\left(|\xi_{\mathbf{i}(\alpha),\mathbf{i}}|^p + |\xi_{\mathbf{i}(\alpha)}|^p + 2^p\right)\right) \chi\left\{\hat{\mathbf{i}} = \mathbf{i}\right\}\right),$$

and then, proceeding as above and choosing a sufficiently small η, we again arrive at (5.28).

According to (5.28) uniformly with respect to any PP-scale $\mathcal{A}_{\mathcal{K}}$

$$\sup_{f\in\mathcal{A}_{\mathcal{K}(S)}} \mathbf{R}_\varepsilon^-(f) = O(\varepsilon^2 S\log S)^{p/2}$$

while uniformly with respect to any NP-scale $\mathcal{A}_{\mathcal{K}}$

$$\sup_{f\in\mathcal{A}_{\mathcal{K}(S)}} \mathbf{R}_\varepsilon^-(f) \le \psi^p(S)(1+o(1)).$$

Now let us turn to $\mathbf{R}_\varepsilon^+(f)$ in (5.26). For any $\mathbf{i}=(q,i_1,\dots,i_d)\succ\mathbf{i}(\alpha)$ let $\mathbf{i}',\mathbf{i}''$ be defined by

$$\mathbf{i}'=(q(\alpha),i_1(\alpha)+l,\dots,i_d(\alpha)+l)\prec\mathbf{i}\preccurlyeq(q(\alpha),i_1(\alpha)+l+1,\dots,i_d(\alpha)+l+1)=\mathbf{i}'',$$

where $l\ge 0$ is an integer. Note that

$$\mathbf{i}'\Supset\mathbf{i}(\alpha) \tag{5.29}$$

and for some $C>0$

$$S_{\mathbf{i}''}\le CS_{\mathbf{i}}\le C^2S_{\mathbf{i}'}. \tag{5.30}$$

Consider the following auxiliary events

$$A_{\mathbf{i}}=\left\{\omega:\ \left|\bar f_{\mathbf{i}}(\mathbf{x})-f(\mathbf{x})\right|\le\bar\sigma_{\mathbf{i}}\sqrt{2p_1\log S_{\mathbf{i}}}\right\}.$$

Using first the c_r-inequality and then Hölder inequality, one obtains by (5.25) for some $C_1,C_2>0$ and all sufficiently small ε:

$$\begin{aligned}
\mathbf{R}_\varepsilon^+(f) &= \sum_{\mathbf{i}\succ\mathbf{i}(\alpha)}\mathbf{E}_f^{(\varepsilon)}\left|\bar f_{\mathbf{i}}(\mathbf{x})-f(\mathbf{x})\right|^p\chi\{\hat{\mathbf{i}}=\mathbf{i}\}\\
&\le C_1\sum_{\mathbf{i}\succ\mathbf{i}(\alpha)}\mathbf{E}_f^{(\varepsilon)}\left|\bar f_{\mathbf{i}}(\mathbf{x})-f_{\mathbf{i}''}(\mathbf{x})\right|^p\chi\{\hat{\mathbf{i}}=\mathbf{i}\}\\
&+ C_1\sum_{\mathbf{i}\succ\mathbf{i}(\alpha)}\mathbf{E}_f^{(\varepsilon)}\left|\bar f_{\mathbf{i}''}(\mathbf{x})-f(\mathbf{x})\right|^p\chi\{\hat{\mathbf{i}}=\mathbf{i}\}
\end{aligned}$$

$$
\begin{aligned}
&\le\ C_1 \sum_{\mathbf{i}\succ \mathbf{i}(\alpha)} \left(\bar\sigma_{\mathbf{i}''}\sqrt{p_1\log S_{\mathbf{i}''}+p_2\log^\delta S_{\mathbf{i}''}}\right)^p \mathbf{P}_f^{(\varepsilon)}\{\hat{\mathbf{i}}=\mathbf{i}\}\\
&+\ C_1 \sum_{\mathbf{i}\succ \mathbf{i}(\alpha)} \mathbf{E}_f^{(\varepsilon)}\left|\bar f_{\mathbf{i}''}(\mathbf{x})-f(\mathbf{x})\right|^p \ \left(\chi\left\{\hat{\mathbf{i}}=\mathbf{i},A_{\mathbf{i}''}\right\}+\chi\left\{\hat{\mathbf{i}}=\mathbf{i},A^c_{\mathbf{i}''}\right\}\right)\\
&\le\ C_1 \sum_{\mathbf{i}'\succcurlyeq \mathbf{i}(\alpha)} \left(C\bar\sigma_{\mathbf{i}'}\sqrt{p_1\log C^2S_{\mathbf{i}'}+p_2\log^\delta C^2S_{\mathbf{i}'}}\right)^p \mathbf{P}_f^{(\varepsilon)}\{\hat{\mathbf{i}}'=\mathbf{i}'\}\\
&+\ C_1 \sum_{\mathbf{i}'\succcurlyeq \mathbf{i}(\alpha)} \left(C\bar\sigma_{\mathbf{i}'}\sqrt{p_1\log C^2S_{\mathbf{i}'}}\right)^p \mathbf{P}_f^{(\varepsilon)}\{\hat{\mathbf{i}}'=\mathbf{i}'\}\\
&+\ C_1 \sum_{\mathbf{i}''\succ \mathbf{i}(\alpha)} \mathbf{E}_f^{(\varepsilon)}\left|\bar f_{\mathbf{i}''}(\mathbf{x})-f(\mathbf{x})\right|^p \chi\left\{A^c_{\mathbf{i}''}\right\}\\
&\le\ C_2 \sum_{\mathbf{i}'\succcurlyeq \mathbf{i}(\alpha)} \left(\bar\sigma_{\mathbf{i}'}\sqrt{\log S_{\mathbf{i}'}}\right)^p \mathbf{P}_f^{(\varepsilon)}\{\hat{\mathbf{i}}'=\mathbf{i}'\}\\
&+\ C_1 \sum_{\mathbf{i}''\succ \mathbf{i}(\alpha)} \bar\sigma^p_{\mathbf{i}''}\ \left(\mathbf{E}_f^{(\varepsilon)}\left|\bar\sigma^{-1}_{\mathbf{i}''}(\bar f_{\mathbf{i}''}(\mathbf{x})-f(\mathbf{x}))\right|^{2p}\mathbf{P}_f^{(\varepsilon)}\left\{A^c_{\mathbf{i}''}\right\}\right)^{1/2}\\
&\le\ C_2 \sum_{\mathbf{i}'\succcurlyeq \mathbf{i}(\alpha)} \left(\bar\sigma_{\mathbf{i}'}\sqrt{\log S_{\mathbf{i}'}}\right)^p \mathbf{P}_f^{(\varepsilon)}\{\hat{\mathbf{i}}'=\mathbf{i}'\}\\
&+\ C_2 \sum_{\mathbf{i}''\succ \mathbf{i}(\alpha)} \bar\sigma^p_{\mathbf{i}''}\ \left(\mathbf{P}_f^{(\varepsilon)}\left(A^c_{\mathbf{i}''}\right)\right)^{1/2}\\
&\stackrel{\text{def}}{=}\ \mathbf{R}_1^+(f)+\mathbf{R}_2^+(f). \qquad (5.31)
\end{aligned}
$$

Note that

$$\mathbf{P}_f^{(\varepsilon)}\{\hat{\mathbf{i}}'=\mathbf{i}'\}\ \le\ \sum_{\mathbf{j}\succcurlyeq\mathbf{i}'}\mathbf{P}_f^{(\varepsilon)}\{|\bar f_{\mathbf{i}',\mathbf{j}}(\mathbf{x})-\bar f_{\mathbf{i}'}(\mathbf{x})|>\bar\sigma_{\mathbf{j}}\lambda_{\mathbf{j}}\}\,.$$

For a generic standard normal ξ we will use the inequality $\mathbf{P}(|\xi| \geq x) \leq \exp(-x^2/2)$ for $x \geq \sqrt{2/\pi}$. According to (5.20)-(5.21), (5.23), (5.29)-(5.30) and Proposition 5.1, for a large enough I, some $C_3 > 0$ and all $\mathbf{i}' \succcurlyeq \mathbf{i}(\alpha)$ such that $S_{\mathbf{i}'} \geq I$

$$\begin{aligned}
\mathbf{P}_f^{(\varepsilon)}\{\hat{\mathbf{i}}' = \mathbf{i}'\} &\leq \sum_{\mathbf{j} \succcurlyeq \mathbf{i}'} \mathbf{P}\left\{|\xi| > \sqrt{p_1 \log S_{\mathbf{j}} + p_2 \log^\delta S_{\mathbf{j}}} - 1\right\} \\
&\leq \sum_{\mathbf{j} \succcurlyeq \mathbf{i}'} \exp\left(-\frac{p_1 \log S_{\mathbf{j}}}{2} - \frac{p_2 \log^\delta S_{\mathbf{j}}}{3}\right) \\
&\leq \sum_{k=\lfloor \log S_{\mathbf{i}'} \rfloor}^{\infty} \exp\left(-\frac{p_1 k}{2} - \frac{p_2 \log^\delta k}{3}\right) N(e^k) \\
&\leq \sum_{k=\lfloor \log S_{\mathbf{i}'} \rfloor}^{\infty} \exp\left(-\frac{pk}{2} - \frac{p_2 \log^\delta k}{3}\right) \cdot M_4 k^{m_4} \\
&\leq C_3 S_{\mathbf{i}'}^{-p/2} \log^{m_4} S_{\mathbf{i}'} \exp\left(-\frac{p_2}{3} \log^\delta S_{\mathbf{i}'}\right).
\end{aligned}$$

Hence this inequality holds for all $\mathbf{i}' \succcurlyeq \mathbf{i}(\alpha)$ and some $C_3 > 0$. Therefore, again according to Proposition 5.1, uniformly in $f \in \mathcal{A}_{\mathcal{K}_0}$

$$\begin{aligned}
\mathbf{R}_1^+(f) &= O(\varepsilon^p) \sum_{\mathbf{i}' \succcurlyeq \mathbf{i}(\alpha)} \log^{m_4} S_{\mathbf{i}'} \exp\left(-\frac{p_2 \log^\delta S_{\mathbf{i}'}}{3}\right) \\
&= O(\varepsilon^p) \sum_{k=\lfloor \log S_{\mathbf{i}(\alpha)} \rfloor}^{\infty} \log^{m_4} k \exp\left(-\frac{p_2}{3} k^\delta\right) N(q(\alpha); e^k) \\
&= O(\varepsilon^p) \sum_{k=\lfloor \log S_{\mathbf{i}(\alpha)} \rfloor}^{\infty} \log^{2m_4} k \exp\left(-\frac{p_2}{3} k^\delta\right) = O(\varepsilon^p).
\end{aligned}$$

For the second term in (5.31), a usual normal tail bound suffices to evaluate the probabilities $\mathbf{P}_f^{(\varepsilon)}(A_{\mathbf{i}''}^c)$ in a similar way. $\square$

6 Adaptive optimality

Our next goal is to demonstrate that the upper bound presented by Theorem 5.1 is exact for 'typical' non-parametric functional scales. By 'typical' we mean sufficiently large scales admitting minimax estimators converging at a non-parametric rate ε^{τ}, $0 < \tau < 1$.

In the previous section we dealt with *uniform* scales described by (3.23)–(3.24). Note that although these assumptions provided uniform consistency for such scales, cf. Theorem 3.1, they do not yet guarantee consistency of any *adaptive* procedure. Indeed if e.g. $|\Gamma| \sim \varepsilon^2 \log^{\tau}(1/\varepsilon^2)$ in (3.24) for some $\tau > 0$ (slow non-parametric scale, in the terminology of [12]), the estimator described in Theorem 5.1 may become inconsistent. Moreover, no consistent adaptive procedures exist in $\mathfrak{F}_p$ when $\tau \leq 1$. This case can be dealt with in the same way as in [12]. Besides, due to an extremely slow logarithmic rate, it can be regarded as '*untypical*'.

Here we are interested primarily in the *regular* scales for which there exists $0 < \tau < 1$ such that

$$\liminf_{\varepsilon \to 0} \inf_{\mathcal{A}_{\mathcal{K}}} \frac{|\Gamma|}{\varepsilon^{2(1-\tau)}} > 0, \tag{6.1}$$

or equivalently, for some $0 < \tau < 1$,

$$\limsup_{\varepsilon \to 0} \sup_{\mathcal{A}_{\mathcal{K}}} \varepsilon^{2(1-\tau)} S_{\alpha}(\varepsilon^2) < \infty. \tag{6.2}$$

It turns out that this is also the case where the dimension plays a substantial role. Underlying this difference we will need an additional *ampleness* assumption with regard to the functional scales $\mathcal{A}_{\mathcal{K}}$; this assumption is redundant in the one-dimensional case. In the following definition we again exploit the representation (3.31)-(3.32).

Definition 6.1. $\mathcal{A}_{\mathcal{K}}$ is called an *ample* functional scale if there exists a slowly increasing function $h(x) \to \infty$ $(x \to \infty)$, with $g(x) = x/h^{d-1}(x)$ such that, together with any $\alpha = \alpha(\mathbf{r}, \Gamma_a, \mathbf{s})$ and an arbitrary orthogonal matrix Γ_o, $\mathcal{K}$ contains also $\alpha' = \alpha(\mathbf{r}, \Gamma_o \Gamma_a, \mathbf{s}')$ where

$$s'_1 = g(s_1 \times \cdots \times s_d),$$

$$s'_i = h(s_1 \times \cdots \times s_d), \qquad i = 2, \ldots, d.$$

The meaning of the thus introduced notion is obvious: a scale has to sustain arbitrary rotations of the underlying functional classes as well as certain volume-preserving pull-and-squeeze transformations which do not affect the precision of the estimator $\tilde{f}_{B_\alpha}$ ($S_\alpha(\varepsilon^2) =$

$(\prod s_i)/(\pi^d|\Gamma_a|) = \text{const}$). Recall that the functions $\phi(S), \psi(S)$ were defined by (5.1)-(5.2) above and that $\liminf_{\varepsilon\to 0}\inf_{\alpha\in\mathcal{K}} S_\alpha(\varepsilon^2) = \infty$ for any *NP*-scale $\mathcal{A}_\mathcal{K}$; cf. (4.1)-(4.2).

Theorem 6.1. *Assume (6.1). For any ample NP-scale $\mathcal{A}_\mathcal{K}$, arbitrary estimator $\tilde{f}_\varepsilon \in \mathfrak{F}_p(\mathbf{x})$ and all $\mathbf{x}\in\mathbb{R}^d$, $p>0$*

$$\liminf_{\varepsilon\to 0}\inf_{S>0}\sup_{f\in\mathcal{A}_{\mathcal{K}(S)}} \mathbf{E}_f^{(\varepsilon)}\left|\psi^{-1}(S)\left(\tilde{f}_\varepsilon(\mathbf{x}) - f(\mathbf{x})\right)\right|^p \geq 1.$$

Proof. To make the arguments below more accessible we will assume additionally that $\mathcal{K}(S)$ contains $\alpha(\mathbf{r},\Gamma_o,\mathbf{s})$, for some $\mathbf{r},\mathbf{s}>0, \Gamma_o\in\mathfrak{A}_o$. The general case can be dealt with similarly with some minor changes. Further we can restrict ourselves to the case $d>1$. Assuming S is sufficiently large, let

$$\theta = \phi(S) - \sqrt{\phi(S)},$$

$$f_0(\mathbf{y}) \equiv 0.$$

To introduce an alternative family of functions $f_j \in \mathcal{A}_{\mathcal{K}(S)}$, consider first a prolate cuboid

$$\mathbf{C}_0 = \{\mathbf{t} : |t_i| \leq s_i',\ i = 1,\ldots,d\},$$

having one large symmetry axis and $d-1$ small axes and $\mathrm{vol}(\mathbf{C}_0) = (2\pi)^d S$. Let

$$\delta = S^{-1}h^{d+1}(S).$$

There exist

$$J = O(\delta^{-(d-1)}) = O\left(\frac{S^{d-1}}{h^{d^2-1}(S)}\right)$$

centrally symmetric cuboids $\mathbf{C}_j^*$, $j=1,\ldots,J$ congruent to $\mathbf{C}_0$, such that the angles between their larger axes are no smaller than δ. The intersections of these cuboids belong to a ball $\mathbf{B}(r)$ with a radius no larger than

$$r = O\left(\frac{S}{h^d(S)}\right),$$

so that

$$\mathrm{vol}\left(\mathbf{C}_j^*\cap\mathbf{B}(r)\right) = O\left(\frac{S}{h(S)}\right) \stackrel{\text{def}}{=} \kappa\,\mathrm{vol}\mathbf{C}_0,$$

where

$$\kappa = o(1), \quad S \to \infty .$$

Let $\mathbf{C}_j = \mathbf{C}_j^* \backslash \mathbf{B}(r),\ j = 1, \ldots, J$ be the same cuboids minus their intersections with $\mathbf{B}(r)$:

$$\mathbf{C}_i \cap \mathbf{C}_j = \emptyset, \quad i \neq j, \quad \mathrm{vol}(\mathbf{C}_j) = (2\pi)^d S(1-\kappa).$$

Finally denote by $k_j(\mathbf{x}), k_j^*(\mathbf{x})$ real-valued functions whose Fourier transforms are the indicators of $\mathbf{C}_j, \mathbf{C}_j^*$ respectively. Define

$$f_j(\mathbf{y}) = \theta\varepsilon \|k_j\|^{-1} k_j(\mathbf{y} - \mathbf{x}), \quad j = 1, \ldots, J.$$

Obviously for $i, j = 1, \ldots, J$

$$\int_{\mathbb{R}^d} f_i(\mathbf{y}) f_j(\mathbf{y})\, d\mathbf{y} = \theta^2 \varepsilon^2 \delta_{ij},$$

and

$$f_i(\mathbf{x}) = \theta\varepsilon \sqrt{S(1-\kappa)} .$$

We will demonstrate next that

$$f_j \in \mathcal{A}_{\mathcal{K}(S)}, \quad j = 1, \ldots, J.$$

Note that for some suitable B_j

$$|\hat{f}_j(\mathbf{t})| \ \leq \ \theta\varepsilon \|k_j\|^{-1} \hat{k}_j(\mathbf{t}) \ \leq$$

$$\frac{\theta}{1-\kappa} \varepsilon \|k_j^*\|^{-1} \hat{k}_j^*(\mathbf{t}) = \frac{\theta}{1-\kappa} \varepsilon \|k_{B_j}\|^{-1} \hat{k}(B_j^{-1}\mathbf{t}).$$

In fact the range of the parameter θ for which $f_j \in \mathcal{A}_{\mathcal{K}(S)}$ has already been evaluated in the proof of Theorem 3.1. By (3.19), (6.1)-(6.2) $f_j \in \mathcal{A}_{\mathcal{K}(S)}$ for all j, provided

$$\frac{\theta^2}{1-\kappa} \leq \frac{c^{2d-1}}{2} = O\left(\log \frac{1}{\varepsilon^2}\right)^{2d-1},$$

which obviously holds under our assumption (6.1), for all sufficiently small ε and $d > 1$.

Denoting $f_\varepsilon^* = \psi^{-1}(S)\tilde{f}_\varepsilon(\mathbf{x})$, $L = \sqrt{1-\kappa}\,\phi^{-1}(S)\,\theta$ one obtains for $j = 1, \ldots, J$

$$\psi^{-1}(S)(\tilde{f}_\varepsilon(\mathbf{x}) - f_j(\mathbf{x})) = f_\varepsilon^* - L$$

and

$$\frac{\pi^{d/2}}{\varepsilon}(\tilde{f}_\varepsilon(\mathbf{x}) - f_0(\mathbf{x})) = \frac{\pi^{d/2}}{\varepsilon}\sqrt{S}\,\phi(S)\,f_\varepsilon^* = f^* \exp\left(\frac{\log S}{2} + \log\phi(S)\right).$$

Let $q = \exp(-\phi(S))$. It is easy to see that

$$\begin{aligned}
R \quad &\stackrel{\text{def}}{=} \quad \sup_{\mathcal{A}_{\mathcal{K}(S)}} \mathbf{E}_f^{(\varepsilon)} \left|\psi^{-1}(S)(\tilde{f}_\varepsilon(\mathbf{x}) - f(\mathbf{x}))\right|^p \\
&\geq \quad \frac{1}{J}\sum_{j=1}^{J} \mathbf{E}_j^{(\varepsilon)} \left|\psi^{-1}(S)(\tilde{f}_\varepsilon(\mathbf{x}) - f_j(\mathbf{x}))\right|^p
\end{aligned}$$

$$\geq \quad q\mathbf{E}_0^{(\varepsilon)} \left|\frac{\pi^{d/2}}{\varepsilon}(\tilde{f}_\varepsilon(\mathbf{x}) - f_0(\mathbf{x}))\right|^p + \frac{1-q}{J}\sum_{j=1}^{J} \mathbf{E}_j^{(\varepsilon)} \left|\psi^{-1}(S)(\tilde{f}_\varepsilon(\mathbf{x}) - f_j(\mathbf{x}))\right|^p + O(q)$$

$$\geq \quad q\exp\left(\frac{p\log S}{2} + p\log\phi(S)\right)\mathbf{E}_0|f_\varepsilon^*|^p + \frac{1-q}{J}\sum_{j=1}^{J}\mathbf{E}_0^{(\varepsilon)}\eta_j\,|f_\varepsilon^* - L|^p + O(q)$$

$$\begin{aligned}
&\geq \quad (1-q)\mathbf{E}_0^{(\varepsilon)}\left(M|f_\varepsilon^*|^p + Z\,|f_\varepsilon^* - L|^p\right) + O(q) \\
&\geq \quad (1-q)\mathbf{E}_0^{(\varepsilon)}\min_x\left(M|x|^p + Z\,|x - L|^p\right) + O(q), \qquad (6.3)
\end{aligned}$$

where

$$M = q\exp\left(\frac{p\log S}{2} + p\log\phi(S)\right), \qquad Z = \frac{1}{J}\sum_{j=1}^{J}\eta_j.$$

Here

$$\eta_j = \frac{d\mathbf{P}_j^{(\varepsilon)}}{d\mathbf{P}_0^{(\varepsilon)}}(V(\cdot)) = \exp\left\{\varepsilon^{-2}\int f_j(\mathbf{y})\,dV(\mathbf{y}) - \frac{\varepsilon^{-2}}{2}\int f_j^2(\mathbf{y})\,d\mathbf{y}\right\} = \exp\left\{\theta T_j - \frac{\theta^2}{2}\right\},$$

where

$$T_j = \theta\varepsilon^{-1}\|k_j\|^{-1}\int k_j(\mathbf{y}-\mathbf{x})\,dV(\mathbf{y}) \overset{\mathbf{P}_1^{(\varepsilon)}}{=} \begin{cases} \theta+\xi_j, & \text{if } j=1, \\ \xi_j, & \text{if } j\neq 1, \end{cases}$$

and $\xi_j = \|k_j\|^{-1}\int k_j(\mathbf{y}-\mathbf{x})\,dW(\mathbf{y})$ are independent standard normal random variables. Thus

$$\eta_j \overset{\mathbf{P}_1^{(\varepsilon)}}{=} \begin{cases} \exp\left(\theta\xi_1+\theta^2/2\right), & \text{if } j=1, \\ \exp\left(\theta\xi_1-\theta^2/2\right), & \text{if } j\neq 1. \end{cases}$$

Note in particular that for $j \neq 1$

$$\mathbf{E}_1^{(\varepsilon)}\eta_j = 1, \quad \mathbf{E}_1^{(\varepsilon)}\eta_j^{1+\nu} = \exp\Big(\nu(\nu+1)\theta^2/2\Big), \quad (\nu > 0). \tag{6.4}$$

Let $p \leq 1$. Then by (6.3), cf. Lepski (1992b),

$$\begin{aligned} R &\geq (1-q)\mathbf{E}_0^{(\varepsilon)}\min(M,Z)L^p + O(q) \qquad (6.5)\\ &\geq (1-q)L^p\mathbf{E}_0^{(\varepsilon)}Z\,\mathbf{1}(Z\leq M) + O(q) \\ &= \frac{(1-q)L^p}{J}\sum_{j=1}^{J}\mathbf{P}_j^{(\varepsilon)}(Z\leq M) + O(q) \\ &= (1-q)L^p\mathbf{P}_1^{(\varepsilon)}(Z\leq M) + O(q). \end{aligned}$$

Now $M \to \infty$ uniformly in $\mathcal{A}_{\mathcal{K}}$ and

$$\frac{Z}{M} = \frac{J-1}{JM} + \frac{J-1}{J}\frac{1}{(J-1)M}\sum_{j=2}^{J}(\eta_j - 1) + \frac{1}{JM}\eta_1,$$

where

$$\frac{\eta_1}{JM} \overset{\mathbf{P}_1}{=} \exp\left(\theta\xi_1 + \frac{\theta^2}{2} - \frac{p\log S}{2} - p\log\phi(S) - \log q - (d-1)\log S + O(1)\right) =$$

$$\exp\left((\phi(S)-\sqrt{\phi(S)})\xi_1 + \frac{(\phi(S)-\sqrt{\phi(S)})^2}{2} - \frac{p+2(d-1)}{2}\log S - 2\log\phi(S) + O(1)\right) \overset{\mathbf{P}_1^{(\varepsilon)}}{\to} 0.$$

Next, according to the Bahr-Esseen inequality [1], by (6.4) for any $0 \le \nu \le 1$ and $\varkappa > 0$

$$\mathbf{P}_1^{(\varepsilon)}\left(\left|\frac{1}{(J-1)M}\sum_{j=2}^{J}(\eta_j - 1)\right| \ge \varkappa\right) \le \frac{2\mathbf{E}_1^{(\varepsilon)}|\eta_J - 1|^{1+\nu}}{\varkappa^{1+\nu}M^{1+\nu}(J-1)^{\nu}} \le$$

$$\frac{16\mathbf{E}_1^{(\varepsilon)}\eta_J^{1+\nu}}{\varkappa^{1+\nu}M^{1+\nu}(J-1)^{\nu}} = O(1)M^{-(1+\nu)}J^{-\nu}\exp\left(\nu(\nu+1)\theta^2/2\right) =$$

$$O(1)\exp\left(\frac{1}{2}\left(\nu(\nu+1)(p+2(d-1)) - p(1+\nu) - 2(d-1)\nu + o(1)\right)\log S\right) = o(1),$$

provided ν has been chosen small enough, namely $\nu^2(p+2(d-1)) < p$.

For $p > 1$ elementary algebra gives

$$\begin{aligned}
R &\ge (1-q)L^p\mathbf{E}_0^{(\varepsilon)}Z\left(1 + (Z/M)^{1/(p-1)}\right)^{-(p-1)} + O(q) \\
&\overset{\text{def}}{=} (1-q)L^p\mathbf{E}_0^{(\varepsilon)}Z\zeta + O(q) \\
&= \frac{(1-q)L^p}{J}\sum_{j=1}^{J}\mathbf{E}_j^{(\varepsilon)}\zeta + O(q) \\
&= (1-q)L^p\mathbf{E}_1^{(\varepsilon)}\zeta + O(q) \\
&= 1 + o(1),
\end{aligned}$$

by the relation $\zeta \overset{\mathbf{P}_1^{(\varepsilon)}}{\to} 1$ which has been proved while we treated the case $p \le 1$. $\square$

References

[1] B. van Bahr and C.-G. Esseen, *Inequalities for the rth absolute moment of a sum of random variables,* $1 \leq r \leq 2$, Ann. Math. Statist., 36 (1965), 299-303.

[2] H. Dym and H.P. McKean, *Fourier Series and Integrals,* Academic Press, New York-London, 1972.

[3] Yu.K. Golubev and B.Y. Levit, *Asymptotically efficient estimation for analytic distributions,* Math. Meth. Statistics, 5 (1996), 357-368.

[4] H. Holden, B. Øksendal, J. Ubøe and T. Zhang, *Stochastic Partial Differential Equations.* Birkhäuser, Boston-Basel-Berlin, 1996.

[5] I.A. Ibragimov and R.Z. Hasminskii, *Statistical Estimation. Asymptotic Theory,* Springer, New York, 1981.

[6] I.A. Ibragimov and R.Z. Hasminskii, *Estimation of distribution density,* J. Sov. Math., 25 (1983), 40-57.

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

[8] I.A. Ibragimov and R.Z. Hasminskii, *On density estimation in the view of Kolmogorov's ideas in approximation theory,* Ann. Statis., 18 (1990), 999-1010.

[9] O.V. Lepski, *Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates,* Theory Probab. Appl, **36** (1991), 682-697.

[10] O.V. Lepski, *Asymptotically minimax adaptive estimation. II: Schemes without optimal adaptation. Adaptive estimators,* Theory Probab. Appl., **37** (1992a), 433-448.

[11] O.V. Lepski, *On problems of adaptive estimation in white Gaussian noise*, Adv. Sov. Math., **12** (1992b), 87-106.

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

[13] E. Parzen, *On asymptotically efficient consistent estimates of the spectral density function of a stationary time series.* J. Roy. Stat. Soc., Ser.B, 20 (1958), 303-322.

[14] J.B. Walsh, *An introduction to stochastic partial differential equations.* Lect. Notes Math., 1180 (1986), 265-439.