AUSTRIAN JOURNAL OF STATISTICS
Volume 32 (2003), Number 1&2, 99–129

Adaptive Regression on the Real Line in Classes of Smooth Functions

L.M. Artiles and B.Y. Levit
Eurandom, Eindhoven, the Netherlands
Queen's University, Kingston, Canada

Abstract: Adaptive pointwise estimation of an unknown regression function $f(x)$, $x \in \mathbb{R}$ corrupted by additive Gaussian noise is considered in the equidistant design setting. The function f is assumed to belong to the class $\mathcal{A}(\alpha)$ of functions whose Fourier transform are rapidly decreasing in the weighted L^2-sense. The rate of decrease is described by a weight function that depends on the vector of parameters α which, in the adaptive setting, is typically unknown. For any of the classes $\mathcal{A}(\alpha)$, α fixed, we describe minimax estimators up to a constant as the bin-width goes to zero. Conditions under which an adaptive study is suitable are presented and a notion of adaptive asymptotic optimality is introduced based on distinguishing, among all possible functional scales, between the so-called non-parametric (NP) and pseudo-parametric (PP) scales. We propose adaptive estimators which 'tune up' point-wisely to the unknown smoothness of f. We prove them to be asymptotically adaptively minimax for large collections of NP functional scales, subject to being rate efficient for any of the PP functional scales.

Keywords: Non-parametric Statistics, Minimax Estimation, Adaptive Estimation, Fourier Transformations.

1 Introduction

During the last two decades adaptive estimation has become one of the most active areas of research in non-parametric statistics. The introduction of different models of adaptive estimation reflects the existing practical needs for more realistic models and flexible methods of estimation. Study of these models brought with it new challenging problems which required creation of new statistical methods and approaches.

In this paper we study non-parametric adaptive regression in a fixed design model in which an unknown regression function $f(x)$ can be observed on an equidistant grid of the whole real line. More precisely, for a given bin-width $h > 0$, we consider the additive model of observations given by

$$y_\ell = f(\ell h) + \xi_\ell, \qquad \ell = 0, \pm 1, \pm 2, \dots \tag{1}$$

where ξ_ℓ are independent centered Gaussian random variables $\mathcal{N}(0, \sigma^2)$, with a given variance $\sigma^2 > 0$. Often in the statistical literature more advanced results are obtained in the white noise model

$$dV(x) = f(x)\, dx + \epsilon\, dW(x), \quad -\infty < x < \infty, \tag{2}$$

which is just an approximation to the model (1), with $\epsilon = \sqrt{\sigma^2 h}$. Here V is the noisy observation of an unknown regression function f, ϵ is the resolving noise and $W(x)$ represents a standard Wiener process.

There exists a huge literature on the equivalence between these two models, cf. e.g. Brown and Low (1996) and Nussbaum (1996), but this does not cover our main problem here, namely adaptive non-parametric estimation. Our approach is greatly influenced by a recent paper, Lepski and Levit (1998), which was a milestone in adaptive estimation of infinitely differentiable functions, in the white noise model (2). Below we will explain main differences between our approach and that of Lepski and Levit (1998).

In non-parametric statistics, classes of functions are in general described by smoothness parameters. In this paper we shall study classes of functions defined in terms of positive parameters γ, β and r whose interpretation will be explained below. We will study estimation of f in (1), under the assumption that f belongs to the functional class $\mathcal{A}(\gamma, \beta, r)$ which is the collection of all continuous functions such that

$$\|f\|^2_{\gamma,\beta,r} := \int_{-\infty}^{\infty} \frac{\gamma}{\beta^2}\, e^{2|\gamma t|^r} |\mathcal{F}[f](t)|^2 dt \leq 1. \tag{3}$$

Here $\mathcal{F}[f]$ represents the Fourier transform of f. The collection of all such classes will be called *functional scale*. Note that when the parameters are assumed known, we are dealing with the problem of non-parametric estimation much studied recently, especially since the publications, Ibragimov and Has'miskii (1981, 1982, 1983), Stone (1982). The situation in which neither of these parameters is known *a priori* is much more realistic and complex. A real progress in this problem which is usually referred to as *adaptive* estimation, has been only achieved in the last decade, most notably since the publication of Lepski (1990, 1991, 1992a, 1992b). Further progress was achieved in Lepski and Levit (1998, 1999).

For all γ, β, r, the class $\mathcal{A}(\gamma, \beta, r)$ is a class of infinitely differentiable functions, and each of the parameters affects the smoothness – and the accuracy of the best non-parametric estimators – in its own way. The parameter γ is some kind of 'scale' parameter: one can verify that $f(\cdot) \in \mathcal{A}(1, \beta, r)$ if and only if $\frac{1}{\gamma} f(\frac{\cdot}{\gamma}) \in \mathcal{A}(\gamma, \beta, r)$. Therefore, of all parameters, it affects the smoothness of f most dramatically. The bigger is γ, the smoother are the functions of the class.

The parameter β can be interpreted as a 'size' parameter and represents the radius of the corresponding L^2-ellipsoid defined by (3). Note that $f(\cdot) \in \mathcal{A}(\gamma, 1, r)$ if and only if $\beta f(\cdot) \in \mathcal{A}(\gamma, \beta, r)$. Therefore the bigger is β, the less smooth are the functions of the class.

Finally, r can be best described as a parameter responsible for the 'type' of smoothness. It is well known that for $r = 1$ all functions in the class $\mathcal{A}(\gamma, \beta, r)$ admit bounded analytic continuation into the strip $\{z = x + iy : |y| < \gamma\}$ of the complex plane (Paley-Wiener theorem), and therefore for all $r > 1$ the functions in $\mathcal{A}(\gamma, \beta, r)$ are entire functions (i.e. functions admitting analytic continuation into the whole complex plane). For $r < 1$ these functions are 'only' infinitely differentiable, and their smoothness increases together with r.

In the Gaussian white noise model Lepski and Levit (1998) studied adaptive estimation for even broader classes of functions with rapidly vanishing Fourier transforms

$\mathcal{F}[f](t)$. However, their main conclusions are readily interpretable in the special example of functional classes $\mathcal{A}'(\gamma,\gamma,r)=\{f \text{ continuous}, |\mathcal{F}[f](t)|\le\gamma\exp-(\gamma t)^r\}$ which are quite similar to our classes $\mathcal{A}(\gamma,\gamma,r)$. Let us remind some of these conclusions here, as a starting point for outlining our main results. For simplicity, we will assume, after Lepski and Levit (1998), that $0<r_-<r<r_+<\infty$.

In the adaptive estimation, when the parameters such as γ,β,r are unknown, one is looking for statistical procedures which can 'adapt' to the largest possible scope of these parameters. As the smoothness of the underlying functions is most notably affected by the 'scale' parameter γ, we will mainly refer to the ensuing uncertainty in the value of this parameter. More specifically, the accuracy of the best methods of estimation will be determined by the 'effective noise' ϵ^2/γ, where ϵ is the average noise intensity in the observation model (1).

To realize the whole scope of the problem, it is useful to look at the extreme cases. On one hand, the situation could be so 'bad', that no consistent estimation of the unknown function would be possible at all, even if the parameter γ was completely known. On an intuitive level, it is quite clear that such a situation occurs when $\epsilon^2/\gamma\not\to 0$. We can exclude this case from consideration on the ground that "nothing can be done" in such an extreme situation. Thus one can restrict attention to the case $\gamma\gg\epsilon^2$. The situation deteriorates further in the adaptive setting, due to the uncertainty in parameter γ. According to Lepski and Levit (1998), adaptive methods can only work efficiently if $\gamma\gg\epsilon^{2-\tau}$, for some $0<\tau<2$. On the other hand, if γ becomes too big, the underlying functions become unrealistically smooth and can be estimated with accuracy $O(\epsilon)$, i.e. with the same accuracy which could be achieved if all underlying functions were either constant, or just included a few unknown parameters. According to Lepski and Levit (1998), such an offbeat situation occurs only when γ becomes of order $\log^{1/r}\epsilon^{-1}$. Therefore one can restrict attention to those γ for which $\epsilon^{2-\tau}\ll\gamma\ll\log^{1/r}\epsilon^{-1}$, which, in a sense, is the largest possible range for which adaptive procedure can exist. For all γ in this range, an efficient adaptive non-parametric procedure has been proposed in Lepski and Levit (1998). Note that this discussion led us, by the very nature of the statistical problem of adaptation, to a situation in which the unknown parameter of the scale γ belonged to a region $\Gamma=\Gamma_\epsilon$ depending on the index ϵ of the model. In other words, our adaptive setting leads us to a natural assumption that the unknown scale parameter γ may itself depend on the index ϵ.

Now, in the model we have just discussed the essential role was played by the noise intensity ϵ and the scale parameter γ. Our model of discrete regression is more realistic and also contains more parameters: σ,h,γ,β,r. Since the white noise model (2) is known to approximate the discrete regression model (1), one can expect some similarity between the ensuing results, namely that similar procedure could lead to an efficient adaptive method of estimation in the discrete regression. Without aiming at precise definitions, one could speak in this case of a "weak" equivalence between the white noise and discrete time adaptive regression schemes.

However, just as the relation between the two parameters involved played an important role in the above discussion, a more complicated relation between all involved parameters affects the quality of the optimal adaptive procedure in the discrete models. In fact, such relations become more complex in the discrete case, not only because of additional parameters, all of which may be unknown and, therefore vary together with ϵ, but also due to

the limitations to which the continuous time model (2) captures the underlying properties of the discrete model (1). In particular, the obvious naive recipe of just replacing ϵ in all the above restrictions by $\sqrt{\sigma^2 h}$ does not provide a correct answer.

We comment next that the classes similar to (3) are well known in statistics. Apparently they have been introduced first (for $r = 1$) in Ibragimov and Has'minskii (1983), where optimal rates of convergence were found in estimating an unknown density function $f \in \mathcal{A}(\gamma, \beta, 1)$. Later Golubev and Levit (1996) showed (again for $r = 1$) that these non-parametric classes are quite unique, in the sense that not only optimal rates, but exact asymptotically minimax estimators, even point-wisely, can be explicitly constructed for such classes. Asymptotically efficient non-parametric regression for the classes $\mathcal{A}(\gamma, \beta, 1)$ was studied in Golubev, Levit and Tsybakov (1996). Here we consider more general classes $\mathcal{A}(\gamma, \beta, r)$, use kernel-type estimators, different from Golubev, Levit and Tsybakov (1996) and, more significantly, consider the problem of adaptive estimation.

In the Gaussian white noise model Lepski and Levit (1998) considered still more general classes of infinitely differentiable functions, with rapidly vanishing Fourier transforms. However, the restriction on the Fourier transform of f in their paper was based on the L^∞-, rather than on the L^2-norm, as in our case. They have not only proposed asymptotically minimax estimators for all of the corresponding classes, but have also constructed asymptotically optimal adaptive estimators for the whole scale of such classes.

Since in most applications the information about an unknown function is typically conveyed by discrete measurements, our model can be viewed as a more realistic approximation, than the classical white noise model. Therefore our model contains an additional "discretization" parameter h – the bin-width.

Our goal is to study, to what degree the method of the adaptive procedure proposed in Lepski and Levit (1998) works in the discrete regression setting. More precisely, we are seeking to find natural conditions under which our equidistant regression model is weakly equivalent to the classical white noise model, in the sense that the asymptotically optimal adaptive estimators proposed for the later model, are still asymptotically optimal in the equidistant non-parametric regression models.

In the next section we introduce the model. In Section 3 we prove some auxiliary lemmas. In Section 4, the problem of asymptotic minimax regression is studied first under the assumption that the class of functions is completely determined by a fixed vector of parameters (γ, β, r), these parameters being independent of the index of the model h. At the end of this section we give the first steps towards the adaptive framework by allowing the parameters of the class depend on the index of the model. In Section 5 we consider the functional scales which are collections of functional classes, see (40). We define the optimality criteria based on the classification of the scales in pseudo-parametric (PP) and non-parametric (NP) scales. We then prove optimality of the adaptive procedure. We shall see that, compared to a given functional class $\mathcal{A}(\gamma, \beta, r)$, an additional logarithmic factor in the exact rate of convergence has to be paid as a price for the uncertainty about the actual class the regression function belongs to, see Theorem 3.

2 The Model

Let us formalize our model.

Definition 1 *Let $\gamma, \beta, r > 0$ be given. We denote by $\mathcal{A}(\gamma, \beta, r)$ the class of continuous functions $f : \mathbb{R} \to \mathbb{R}$, whose Fourier transforms $\mathcal{F}[f]$ satisfy*

$$\|f\|_{\gamma,\beta,r} := \int \frac{\gamma}{\beta^2}\, e^{2|\gamma t|^r} |\mathcal{F}[f](t)|^2 dt \le 1. \tag{4}$$

In this study we use the following definition of the Fourier transform,

$$\mathcal{F}[f](t) = \int e^{\mathbf{i}tx} f(x)\, dx. \tag{5}$$

Note that the Fourier inversion formula

$$f(x) = \frac{1}{2\pi} \int e^{-\mathbf{i}tx} \mathcal{F}[f](t)\, dt \tag{6}$$

certainly holds under assumption (4). It is easy to see that for all $\gamma, \beta, r > 0$, functions in $\mathcal{A}(\gamma, \beta, r)$ are infinitely differentiable.

Now, let us consider the following observation model

$$y_\ell = f(\ell h) + \xi_\ell, \qquad \ell = 0, \pm 1, \pm 2, \ldots, \tag{7}$$

where ξ_ℓ are i.i.d. Gaussian random variables, $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. We assume that the function f belongs to the family $\mathcal{A}(\gamma, \beta, r)$, for some $\gamma, \beta, r > 0$.

Our purpose is to estimate the unknown function $f(x)$ based on the vector of observations $\mathbf{y} = (\ldots, y_{-2}, y_{-1}, y_0, y_1, y_2, \ldots)$. We will choose our optimal estimator from the family of kernel type estimators

$$\hat{f}_{h,s}(x, \mathbf{y}) = h \sum_{\ell=-\infty}^{\infty} k_s(x - \ell h)\, y_\ell \tag{8}$$

where $k_s, s \ge 0$, is the so-called *sinc*-function

$$k_s(x) = \frac{\sin sx}{\pi x}, \tag{9}$$

and $k_s(0) = \frac{s}{\pi}$. This kernel has the property

$$\mathcal{F}[k_s](t) = \mathbb{1}_{[-s,s]}(t) \tag{10}$$

and therefore, according to the convolution theorem,

$$\mathcal{F}[f * k_s](t) = \mathbb{1}_{[-s,s]}(t)\, \mathcal{F}[f](t), \tag{11}$$

where * represents the convolution operator.

The kernel k_s is just one of many possible, but its very tractable properties make it an attractive tool: it helps significantly in the search of the most general possible results and

clarifies the underlying ideas. For practical purposes some other kernels, such as *de la Vallée Poussin* kernel (cf. Nikol'skiĭ, 1975, p. 301), may be more relevant and typically would work better.

The parameter s is called the bandwidth. As we shall see in Section 4, for any fixed class there exists an optimum bandwidth s. The optimum bandwidth will depend on parameters γ, β, r, σ as well as the index of the model h, called the bin-width, which in our asymptotic study will tend to zero.

Denote by $\tilde{f}_h(x, \mathbf{y})$ an arbitrary estimator of $f(x)$ based on the observations $\mathbf{y}$. To shorten the notation we will often write $\tilde{f}_h(x)$ instead of $\tilde{f}_h(x, \mathbf{y})$. Let $\mathbf{P}_f$ be the distribution of the vector $\mathbf{y}$ and let $\mathbf{E}_f$ and $\mathbf{Var}_f$ denote the expectation and the variance with respect to this measure. When there is no possibility of confusion we will simply write $\mathbf{P}$, $\mathbf{E}$ and $\mathbf{Var}$ respectively.

Let $\mathcal{W}$ be the class of loss functions $w(x)$, $x \in \mathbb{R}$, such that

$$w(x) = w(-x),$$

$$w(x) \geq w(y) \quad \text{for } |x| \geq |y|, \quad x, y \in \mathbb{R},$$

and for some $0 < \eta < \frac{1}{2}$

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

With an appropriate normalizing factor σ_h to be defined shortly, and $w \in \mathcal{W}$, we will consider the maximum *risk*, over a fixed functional class $\mathcal{A}(\gamma, \beta, r)$, given by

$$\sup_{f \in \mathcal{A}(\gamma,\beta,r)} \mathbf{E}_f\, w\left(\sigma_h^{-1}(\tilde{f}_h(x, \mathbf{y}) - f(x))\right)$$

as a global measure of the error of the estimator $\tilde{f}_h$ over the whole class $\mathcal{A}(\gamma, \beta, r)$. When the classes $\mathcal{A}(\gamma, \beta, r)$ are considered fixed, our main goal is to find an estimator such that the corresponding maximum risk is as small as possible, i.e. achieves (asymptotically) the *minimax risk*

$$\inf_{\tilde{f}_h} \sup_{f \in \mathcal{A}(\gamma,\beta,r)} \mathbf{E}_f\, w\left(\sigma_h^{-1}(\tilde{f}_h(x, \mathbf{y}) - f(x))\right)$$

where $\tilde{f}_h$ is taken from the class of all possible estimators.

In the adaptive setting, we shall allow (γ, β, r) to vary freely inside large scales $\mathcal{K}$. Conditions under which an adaptive study is suitable are presented and a notion of adaptive asymptotic optimality is introduced based on distinguishing, among all possible functional scales, between the so-called non-parametric (NP) and pseudo-parametric (PP) scales.

3 Auxiliary Results

In this section we present, for the reader's convenience, two auxiliary results which will be used in the subsequent sections. The aim of the first lemma is to approximate summation formulas by integrals, with a good approximation error in the case of very smooth integrands. This result is a version of the celebrated *Poisson summation formula*. It

has been used in a similar situation in Golubev, Levit and Tsybakov (1996). Below $\mathcal{A}(\gamma,\beta,r),\gamma,\beta,r>0$ are the functional classes of infinitely differentiable functions previously defined and $k_s(x)$ is the kernel (9).

Lemma 1 *The following properties hold:*

(a) *Let f,g be continuous functions in $L^2(\mathbb{R})$ such that $\mathcal{F}[f],\mathcal{F}[g]\in L^1(\mathbb{R})$, then*

$$h\sum_{\ell=-\infty}^{\infty} g(x-\ell h)f(\ell h-y) \;=\; \tfrac{1}{2\pi}\int e^{-\mathbf{i}t(x-y)}\,\mathcal{F}[g](t)\,\mathcal{F}[f](t)\,dt \;+$$

$$\tfrac{1}{2\pi}\textstyle\sum_{\ell\neq 0} e^{\mathbf{i}\frac{2\pi\ell}{h}y}\int e^{-\mathbf{i}t(x-y)}\,\mathcal{F}[g](t)\,\mathcal{F}[f]\left(t+\tfrac{2\pi\ell}{h}\right)\,dt$$

$$=\; \textstyle\int_{-\infty}^{\infty} g(x-z)f(z-y)\,dz \;+$$

$$\tfrac{1}{2\pi}\textstyle\sum_{\ell\neq 0} e^{\mathbf{i}\frac{2\pi\ell}{h}y}\int e^{-\mathbf{i}t(x-y)}\,\mathcal{F}[g](t)\,\mathcal{F}[f](t+\tfrac{2\pi\ell}{h})\,dt.$$

(b) *For arbitrary numbers s_1,s_2 $(0\le s_1\le s_2)$ denote $\Delta(x)=k_{s_2}(x)-k_{s_1}(x)$.*[2] *Then, uniformly in $\gamma,\beta,r,s_i\ge 0, i=1,2$, $x\in\mathbb{R}$ and $f\in\mathcal{A}(\gamma,\beta,r)$ as $h\to 0$*

$$h\sum_{\ell=-\infty}^{\infty}\Delta(x-\ell h)f(\ell h)=\frac{1}{2\pi}\int e^{-\mathbf{i}tx}\,\mathcal{F}[\Delta](t)\,\mathcal{F}[f](t)\,dt+$$

$$O\left(e^{-\left(2\pi\frac{\gamma}{h}\right)^r/c_r}\right)\left(\int_{s_1}^{s_2}\frac{\beta^2}{\gamma}e^{2(\gamma t)^r}\,dt\right)^{1/2},$$

where $c_r=\max(1,2^{r-1})$.

(c) *Let s_1,s_2 and $\Delta(x)$ be as before. Then, uniformly in s_1,s_2 and $x\in\mathbb{R}$, for $h\to 0$,*

$$h\sum_{\ell=-\infty}^{\infty}\Delta^2(x-\ell h)=\frac{s_2-s_1}{\pi}\Big(1+O(1)\,h(s_2-s_1)\Big).$$

Proof. (a) The proof is based on the formula

$$\sum_{\ell=-\infty}^{\infty} e^{2\pi\mathbf{i}\,\ell x}=\sum_{\ell=-\infty}^{\infty}\delta(x-\ell), \tag{12}$$

known in the theory of distributions (cf. e.g. Antonsik et al., 1973, Ch. 9.6). Using the Fourier inversion formula, the distributional formula (12) and with some algebra, one obtains

[2] Note that $\Delta(x)=k_s(x)$ for $s_1=0$ and $s_2=s$.

$$h\sum_{\ell=-\infty}^{\infty} g(x-\ell h)f(\ell h-y) = \frac{h}{(2\pi)^2}\sum_{\ell=-\infty}^{\infty}\int e^{-\mathbf{i}t(x-\ell h)}\,\mathcal{F}[g](t)\,dt\int e^{-\mathbf{i}s(\ell h-y)}\mathcal{F}[f](s)\,ds$$

$$= \frac{h}{(2\pi)^2}\int\int e^{-\mathbf{i}tx}\,\mathcal{F}[g](t)\,e^{\mathbf{i}sy}\mathcal{F}[f](s)\sum_{\ell=-\infty}^{\infty} e^{-\mathbf{i}(s-t)\ell h}\,dt\,ds$$

$$= \frac{h}{(2\pi)^2}\sum_{\ell=-\infty}^{\infty}\int\int e^{-\mathbf{i}tx}\mathcal{F}[g](t)\,e^{\mathbf{i}sy}\mathcal{F}[f](s)\,\delta\left(\frac{h(s-t)}{2\pi}-\ell\right)dt\,ds$$

$$= \frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty}\int e^{-\mathbf{i}tx}\mathcal{F}[g](t)\int e^{\mathbf{i}sy}\mathcal{F}[f](s)\,\delta\left(s-t-\frac{2\pi\ell}{h}\right)ds\,dt$$

$$= \frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty}\int e^{-\mathbf{i}tx}\mathcal{F}[g](t)\,e^{\mathbf{i}(t+\frac{2\pi\ell}{h})y}\mathcal{F}[f]\left(t+\frac{2\pi\ell}{h}\right)dt$$

$$= \frac{1}{2\pi}\int e^{-\mathbf{i}t(x-y)}\,\mathcal{F}[g](t)\,\mathcal{F}[f](t)\,dt$$

$$+\frac{1}{2\pi}\sum_{\ell\neq 0} e^{\mathbf{i}\frac{2\pi\ell}{h}y}\int e^{-\mathbf{i}t(x-y)}\,\mathcal{F}[g](t)\,\mathcal{F}[f]\left(t+\frac{2\pi\ell}{h}\right)dt$$

$$= \int_{-\infty}^{\infty} g(x-z)f(z-y)\,dz$$

$$+\frac{1}{2\pi}\sum_{\ell\neq 0} e^{\mathbf{i}\frac{2\pi\ell}{h}y}\int e^{-\mathbf{i}t(x-y)}\,\mathcal{F}[g](t)\,\mathcal{F}[f]\left(t+\frac{2\pi\ell}{h}\right)dt.$$

(b) If $f\in\mathcal{A}(\gamma,\beta,r)$ then f belongs to $L^2(\mathbb{R})$ according to the Parseval's formula. Also, $\mathcal{F}[f]\in L^1(\mathbb{R})$ according to (4) and the Cauchy-Schwartz inequality. Thus we can apply the previous result in (a), using $g=\Delta$ and $y=0$. Note that $\mathcal{F}[\Delta](t)=\mathbb{1}_{(s_1,s_2]}(|t|)$. Applying the Fourier inversion formula, the Cauchy-Schwartz inequality and the c_r-inequality, we obtain after a few transformations

$$\left|h\sum_{\ell=-\infty}^{\infty}\Delta(x-\ell h)f(\ell h)-\frac{1}{2\pi}\int e^{-\mathbf{i}tx}\,\mathcal{F}[\Delta](t)\,\mathcal{F}[f](t)\,dt\right|$$

$$\leq \frac{1}{2\pi}\sum_{\ell\neq 0}\left|\int e^{-\mathbf{i}tx}\,\mathcal{F}[\Delta](t)\mathcal{F}[f](t+\frac{2\pi\ell}{h})\,dt\right|$$

$$\leq \frac{1}{2\pi}\left(\int\frac{\gamma}{\beta^2}e^{2|\gamma t|^r}\left|\mathcal{F}[f](t)\right|^2dt\right)^{1/2}\sum_{\ell\neq 0}\left(\int\left|\mathcal{F}[\Delta](t)\right|^2\frac{\beta^2}{\gamma}\,e^{-2|\gamma\left(t+\frac{2\pi\ell}{h}\right)|^r}dt\right)^{1/2}$$

$$\leq \frac{1}{2\pi}\sum_{\ell\neq 0}\left(\int \mathbb{1}_{\,(s_1,s_2]}(|t|)\,\frac{\beta^2}{\gamma}\,e^{2|\gamma t|^r}e^{-2|\frac{2\pi\ell\gamma}{h}|^r/c_r}dt\right)^{1/2}$$

$$\leq \frac{1}{2\pi}\sum_{\ell\neq 0}e^{-|\frac{2\pi\ell\gamma}{h}|^r/c_r}\left(2\int \mathbb{1}_{\,(s_1,s_2]}(t)\,\frac{\beta^2}{\gamma}\,e^{2(\gamma t)^r}dt\right)^{1/2}$$

$$= \frac{1}{\pi}\left(2\int_{s_1}^{s_2}\frac{\beta^2}{\gamma}\,e^{2(\gamma t)^r}dt\right)^{1/2}\sum_{\ell=1}^{\infty}e^{-(2\pi\ell\frac{\gamma}{h})^r/c_r}$$

$$\leq \frac{1}{\pi}\left(2\int_{s_1}^{s_2}\frac{\beta^2}{\gamma}\,e^{2(\gamma t)^r}dt\right)^{1/2}\left(e^{-(2\pi\frac{\gamma}{h})^r/c_r}+\int_1^{\infty}e^{-(2\pi\frac{\gamma}{h}x)^r/c_r}dx\right)$$

$$= O\left(e^{-(2\pi\frac{\gamma}{h})^r/c_r}\right)\left(\int_{s_1}^{s_2}\frac{\beta^2}{\gamma}\,e^{2(\gamma t)^r}\,dt\right)^{1/2}, \qquad (h\to 0),$$

where the last asymptotic can be easily derived by partial integration.

(c) Applying (a) and taking $f=g=\Delta$ and $x=y$, we see that

$$h\sum_{\ell=-\infty}^{\infty}\Delta^2(x-\ell h)=h\sum_{\ell=-\infty}^{\infty}\Delta(x-\ell h)\,\Delta(\ell h-x)$$

$$=\frac{1}{2\pi}\int\left(\mathcal{F}[\Delta](t)\right)^2dt+\frac{1}{2\pi}\sum_{\ell\neq 0}e^{i\frac{2\pi\ell}{h}x}\int\mathcal{F}[\Delta](t)\,\mathcal{F}[\Delta]\left(t+\frac{2\pi\ell}{h}\right)\,dt.$$

Therefore

$$\left|h\sum_{\ell=-\infty}^{\infty}\Delta^2(x-\ell h)-\frac{s_j-s_i}{\pi}\right|\leq\frac{1}{2\pi}\sum_{\ell\neq 0}\int\mathcal{F}[\Delta](t)\,\mathcal{F}[\Delta]\left(t+\frac{2\pi\ell}{h}\right)\,dt$$

$$\leq\frac{1}{\pi}\sum_{\ell=1}^{\infty}\int\mathbb{1}_{\,(s_1,s_2]}(|t|)\mathbb{1}_{\,(s_1,s_2]}\left(\left|t+\frac{2\pi\ell}{h}\right|\right)\,dt$$

$$\leq\frac{5h(s_2-s_1)^2}{2\pi^2}=O(1)\,h(s_2-s_1)^2,$$

which completes the proof of the lemma. □

The following elementary properties will be used below. They will help in bounding the bias and the approximation errors.

Lemma 2 *For any positive γ and r the following inequality holds*

$$\int_s^{\infty}e^{-2(\gamma t)^r}\,dt\leq\frac{s\,e^{-2(\gamma s)^r}}{r(\gamma s)^r} \qquad (13)$$

for all $s > t_0$ *where* t_0 *satisfies* $r(\gamma t_0)^r = 1$ *and*

$$\int_0^s e^{2(\gamma t)^r}\, dt = \frac{s\, e^{2(\gamma s)^r}}{2r(\gamma s)^r}(1 + o(1)) \tag{14}$$

uniformly in $r_- < r < r_+$ *for* $\gamma s \to \infty$, *where* $r_-, r_+ > 0$ *are arbitrary fixed numbers.*

For the first inequality see e.g. Lepski and Levit (1998), eqs. (2.8), (2.10). The second property can be easily proven by partial integration.

4 Minimax Regression in $\mathcal{A}(\gamma, \beta, r)$

4.1 Optimality in the Case of Fixed Classes

The first result we present in this section is obtained in the classical framework, i.e. in a situation where the function $f(x)$ although unknown belongs to a given class. In other words, the parameter $\alpha = (\gamma, \beta, r)$ of the class is known and fixed. Denote for shortness $\mathcal{A}(\alpha) = \mathcal{A}(\gamma, \beta, r)$. We will prove that asymptotically minimax estimators can be found among kernel estimators using a specified bandwidth and we will also calculate to a constant their maximal asymptotic risk, for a variety of loss functions.

Theorem 1 *Let* $\alpha > 0$ *and* $\omega \in \mathcal{W}$. *Then for any* $x \in \mathbb{R}$, *the kernel estimator* $\hat{f}_h = \hat{f}_{h,s_h}$, *in (8) with the bandwidth*

$$s_h = s_h(\alpha, \sigma^2) = \frac{1}{\gamma}\left(\frac{1}{2}\log\frac{\beta^2}{\pi\,\gamma\sigma^2 h}\right)^{1/r}, \tag{15}$$

satisfies

$$\lim_{h\to 0}\sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f\, w\left(\sqrt{\frac{\pi}{\sigma^2 h s_h}}(\hat{f}_h(x) - f(x))\right) =$$

$$\lim_{h\to 0}\inf_{\tilde{f}_h}\sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f\, w\left(\sqrt{\frac{\pi}{\sigma^2 h s_h}}(\tilde{f}_h(x) - f(x))\right) = \mathbf{E}\, w(\xi)$$

where $\tilde{f}_h$ *is taken from the class of all possible estimators of* f *and* $\xi \sim \mathcal{N}(0,1)$.

Proof. **Upper bound for the risk.** Let us first study the sample properties of the family of estimators we use. According to the model for the observations (7) and the formula for the estimator (8) one can split the error term as follows,

$$\hat{f}_{h,s}(x) - f(x) = \Big(h\sum_{\ell=-\infty}^{\infty} k_s(x - \ell h) f(\ell h) - f(x)\Big) + \Big(h\sum_{\ell=-\infty}^{\infty} k_s(x-\ell h)\xi_\ell\Big)$$

$$:= b(f, x, s, h) + v(\sigma, x, s, h).$$

For simplicity we shall write below $b_s = b(f, x, s, h)$, $v_s = v(\sigma, x, s, h)$. The mean square error can be decomposed as

$$\mathbf{E}\,(\hat{f}_{h,s}(x) - f(x))^2 \;=\; b_s^2 + \mathbf{Var}\, v_s, \tag{16}$$

where b_s is the bias and v_s is a normally distributed zero mean stochastic term.

First, let us consider the bias. In order to apply Lemma 1 we take $s_1 = 0$ and $s_2 = s$. In this case $\Delta = k_s$. Now, applying Lemma 1(b) and the Fourier inversion formula for $f(x)$ we see that uniformly in $f \in \mathcal{A}(\alpha)$

$$b_s = \frac{1}{2\pi}\int e^{-\mathbf{i}tx}(\mathcal{F}[k_s](t) - 1)\mathcal{F}[f](t)dt + O\left(e^{-(2\pi\frac{\gamma}{h})^r/c_r}\right)\left(\int_0^s \frac{\beta^2}{\gamma}e^{2(\gamma t)^r}\,dt\right)^{1/2},$$

for $h \to 0$. Furthermore, applying Cauchy-Schwartz inequality, property (10), and definition of the class $\mathcal{A}(\gamma, \beta, r)$ we get

$$\begin{aligned} b_s^2 &\le 2\left|\frac{1}{2\pi}\int e^{-\mathbf{i}tx}(\mathcal{F}[k_s](t) - 1)\mathcal{F}[f](t)\,dt\right|^2 + O\left(e^{-2(2\pi\frac{\gamma}{h})^r/c_r}\right)\int_0^s \frac{\beta^2}{\gamma}e^{2(\gamma t)^r}\,dt \\ &\le \frac{1}{2\pi^2}\int_{|t|>s}\frac{\beta^2}{\gamma}e^{-2|\gamma t|^r}dt + O\left(e^{-2(2\pi\frac{\gamma}{h})^r/c_r}\right)\int_0^s \frac{\beta^2}{\gamma}e^{2(\gamma t)^r}\,dt \\ &\le \frac{1}{\pi^2}\int_s^\infty \frac{\beta^2}{\gamma}e^{-2(\gamma t)^r}dt + O\left(e^{-2(2\pi\frac{\gamma}{h})^r/c_r}\right)\int_0^s \frac{\beta^2}{\gamma}e^{2(\gamma t)^r}\,dt. \end{aligned} \tag{17}$$

Second, let us consider the variance term. From Lemma 1(c), with $s_1 = 0$ and $s_2 = s$, we see that

$$\mathbf{Var}\, v_s = \sigma^2 h^2 \sum_{\ell=-\infty}^{\infty} k_s^2(x - \ell h) = \frac{\sigma^2 h\, s}{\pi}(1 + O(1)\, hs), \tag{18}$$

when $h \to 0$. For any s denote

$$\sigma_{h,s}^2 = \frac{\sigma^2 hs}{\pi} \tag{19}$$

and for the chosen bandwidth $s = s_h$ denote the resulting variance

$$\sigma_h^2 = \sigma_h^2(\alpha, \sigma^2) = \frac{\sigma^2 h s_h}{\pi}. \tag{20}$$

From equations (16)–(18) we see that the mean square error of the estimator $\hat{f}_{h,s}$ satisfies

$$\begin{aligned} \left|\mathbf{E}\,(\hat{f}_{h,s}(x) - f(x))^2 - \sigma_{h,s}^2\right| \le \sigma_{h,s}^2\Big(& O(hs) + (\pi\sigma_{h,s})^{-2}\int_s^\infty \frac{\beta^2}{\gamma}e^{-2(\gamma t)^r}dt \\ & + \sigma_{h,s}^{-2}O\left(e^{-2(2\pi\frac{\gamma}{h})^r/c_r}\right)\int_0^s \frac{\beta^2}{\gamma}e^{2(\gamma t)^r}\,dt\Big). \end{aligned} \tag{21}$$

Now we shall verify that, taking $s = s_h$ as defined in (15), the term of the right hand side of the previous equation is equal to $\sigma_h^2\, o(1)$. Before going into details, let us remark that

the bandwidth s_h is precisely the bandwidth that balances the main terms of the bias and the variance in the mean square error, i.e. it minimizes

$$\frac{\sigma^2 hs}{\pi} + \pi^{-2} \int_s^\infty \frac{\beta^2}{\gamma} e^{-2(\gamma t)^r} dt$$

(with respect to s), since by (15)

$$e^{2(\gamma s_h)^r} = \frac{\beta^2}{\pi \gamma \sigma^2 h}. \tag{22}$$

Let us return to equation (21). Note first that

$$h s_h \to 0, \quad \text{when} \quad h \to 0. \tag{23}$$

Second, applying the identity (22) and Lemma 2, we see that

$$\begin{aligned}(\pi\sigma_h)^{-2} \int_{s_h}^\infty \frac{\beta^2}{\gamma} e^{-2(\gamma t)^r} dt &= \frac{\beta^2}{\pi\,\gamma\sigma^2 h} \frac{\int_{s_h}^\infty e^{-2(\gamma t)^r} dt}{s_h} = \frac{\int_{s_h}^\infty e^{-2(\gamma t)^r} dt}{s_h e^{-2(\gamma s_h)^r}} \\ &\le \frac{1}{r(\gamma s_h)^r} = \left(\frac{r}{2} \log \frac{\beta^2}{\pi\gamma\,\sigma^2 h}\right)^{-1} = o(1),\end{aligned} \tag{24}$$

when $h \to 0$. Finally, applying the identity (22) and trivial inequality

$$\begin{aligned}\sigma_h^{-2} e^{-2(\frac{2\pi\gamma}{h})^r/c_r} \int_0^{s_h} \frac{\beta^2}{\gamma}\, e^{2(\gamma t)^r}\, dt &\le \pi \frac{\beta^2}{\gamma\,\sigma^2 h}\, e^{-2(\frac{2\pi\gamma}{h})^r/c_r + 2(\gamma s_h)^r} \\ &= \left(\frac{\beta^2}{\gamma\,\sigma^2 h}\right)^2 e^{-2\left(2\pi\frac{\gamma}{h}\right)^r/c_r} = o(1),\end{aligned} \tag{25}$$

when $h \to 0$. Thus, from (21) and (23)–(25) we have that

$$\mathbf{E}\,(\hat{f}_h(x) - f(x))^2 = \sigma_h^2\,(1 + o(1)), \qquad (h \to 0).$$

Note that when we normalize the error of our estimator by σ_h, the normalized error term $(\hat{f}_h(x) - f(x))/\sigma_h$ has a normal distribution, with mean of order $o(1)$ and variance equal to $1 + o(1)$ where the terms $o(1)$ are small uniformly in $f \in \mathcal{A}(\alpha)$ when h goes to zero. Because the loss function w has only countably many discontinuity points, applying the dominated convergence theorem

$$\lim_{h\to 0} \sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f\, w\left(\sigma_h^{-1}(\hat{f}_h(x) - f(x))\right) = \mathbf{E}\, w\,(\xi). \tag{26}$$

Lower bound for the risk. Consider the parametric family of functions

$$f_\theta(z) = \theta g(z), \quad g(z) = \frac{\pi}{s_h} k_{s_h}(z - x).$$

These functions satisfy $f_\theta(x) = \theta$, and if we assume that $|\theta| \le \theta(h)$ where

$$\theta^2(h) = \frac{s_h^2}{2\pi^2} \left(\int_0^{s_h} \frac{\gamma}{\beta^2} e^{2(\gamma t)^r} dt\right)^{-1} \tag{27}$$

then

$$\int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} |\mathcal{F}[f_\theta](t)|^2 dt \;=\; \theta^2 \frac{\pi^2}{s_h^2} \int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} |\mathcal{F}[k_{s_h}](t)|^2 \, dt$$

$$\leq \; \frac{\theta^2(h)\pi^2}{s_h^2} \int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} \mathbb{1}_{\,[-s_h, s_h]}(t) \, dt \;\leq\; 1.$$

Thus $f_\theta \in \mathcal{A}(\alpha)$ for all θ such that $|\theta| \leq \theta(h)$.

Now, we can apply Kakutani's theorem using the fact that $\sum_{\ell=-\infty}^{\infty} g^2(\ell h) < \infty$ according to Lemma 1(c), and see that

$$\frac{d\mathbf{P}_\theta^{(h)}}{d\mathbf{P}_0^{(h)}}(\mathbf{y}) \;=\; \exp\left\{ \frac{1}{2\sigma^2} \sum_{\ell=-\infty}^{\infty} (2\theta \, y_\ell \, g(\ell h) - \theta^2 g^2(\ell h)) \right\}, \tag{28}$$

where $\mathbf{P}_\theta = \mathbf{P}_{f_\theta}$ (cf. e.g. Hui-Hsiung, 1975, Sect. II.2). The statistic

$$T = \frac{\sum_{\ell=-\infty}^{\infty} y_\ell \, g(\ell h)}{\sum_{\ell=-\infty}^{\infty} g^2(\ell h)} \tag{29}$$

is sufficient for the parameter θ of the family of distributions $\mathbf{P}_\theta$. Obviously T is normally distributed. Given $f_\theta(\ell h) = \theta g(\ell h)$, we can easily verify that

$$T \sim \mathcal{N}\left(\theta \, , \frac{\sigma^2}{\sum_{\ell=-\infty}^{\infty} g^2(\ell h)} \right), \tag{30}$$

and applying Lemma 1(c), with $s_1 = 0$ and $s_2 = s_h$, we see that

$$\frac{1}{\sigma^2} \sum_{\ell=-\infty}^{\infty} g^2(\ell h) \;=\; \frac{\pi^2}{\sigma^2 h s_h^2} \Big(h \sum_{\ell=-\infty}^{\infty} k_{s_h}^2(x - \ell h) \Big) \;=\; \frac{\pi}{\sigma^2 h s_h} \, (1 + O(1) \, h s_h),$$

when h goes to zero. Thus, T can be represented as

$$T = \theta + \varphi \, \xi \quad \text{where} \quad \xi \sim \mathcal{N}(0, 1) \tag{31}$$

and, according to the previous arguments,

$$\varphi^2 = \frac{\sigma^2}{\sum_{\ell=-\infty}^{\infty} g^2(\ell h)} = \sigma_h^2 (1 + o(1)). \tag{32}$$

To derive the required lower bound, let us assume the unknown parameter θ has a prior density $\lambda(\theta)$; a convenient choice is

$$\lambda(\theta) = \frac{1}{\theta(h)} \cos^2 \frac{\pi\theta}{2\theta(h)}, \quad |\theta| \leq \theta(h).$$

We obtain then, due to the sufficiency of the statistic T,

$$\begin{aligned}
&\inf_{\tilde{f}_h}\sup_{f\in\mathcal{A}(\alpha)}\mathbf{E}_f\ \ w\left(\sqrt{\frac{\pi}{\sigma^2 h s_h}}(\tilde{f}_h(x)-f(x))\right)\\
&\qquad\geq\ \inf_{\tilde{f}_h}\sup_{|\theta|<\theta(h)}\mathbf{E}_f\, w\left(\sqrt{\frac{\pi}{\sigma^2 h s_h}}(\tilde{f}_h(x)-f_\theta(x))\right)\\
&\qquad\geq\ \inf_{\hat{\theta}}\sup_{|\theta|<\theta(h)}\mathbf{E}_\theta w\ \left(\sqrt{\frac{\pi}{\sigma^2 h s_h}}(\hat{\theta}-\theta)\right)\\
&\qquad\geq\ \inf_{\hat{\theta}}\int_{-\theta(h)}^{\theta(h)}\mathbf{E}_\theta w\ \left(\sqrt{\frac{\pi}{\sigma^2 h s_h}}(\hat{\theta}-\theta)\right)\ \lambda(\theta)d\theta\\
&\qquad=\ \inf_{\hat{\theta}(T)}\int_{-\theta(h)}^{\theta(h)}\mathbf{E}_\theta w\ \left(\sqrt{\frac{\pi}{\sigma^2 h s_h}}(\hat{\theta}(T)-\theta)\right)\ \lambda(\theta)d\theta\\
&\qquad=\ \mathbf{E}\, w\left(\frac{\varphi}{\sigma_h}\xi\right)-\frac{\varphi^2}{\theta^2(h)}\frac{1}{\sqrt{2\pi}}\int(x^2-1)w(x)e^{-\frac{x^2}{2}}dx\ (1+o(1)).
\end{aligned}$$

Here the last equation follows from Levit (1980). According to (32), $\frac{\varphi}{\sigma_h}=1+o(1)$, $(h\to 0)$, while applying identity (22) and Lemma 2 we see that

$$\frac{\sigma_h^2}{\theta^2(h)}=2\,\frac{\pi\gamma\sigma^2 h}{\beta^2}\,\frac{\int_0^{s_h}\gamma e^{2(\gamma t)^r}dt}{\gamma s_h}\ =\ \frac{2\int_0^{\gamma s_h}\gamma e^{2t^r}dt}{\gamma s_h e^{2(\gamma s_h)^r}}\ \leq\ \frac{1}{r(\gamma s_h)^r}\to 0, \qquad (33)$$

when $h\to 0$. Thus we have that, according to the dominated convergence theorem,

$$\begin{aligned}
&\liminf_{h\to 0}\sup_{f\in\mathcal{A}(\alpha)}\mathbf{E}_f\, w\left(\sqrt{\frac{\pi}{\sigma^2 h s_h}}(\hat{f}_h(x)-f(x))\right)\ \geq\\
&\qquad\liminf_{h\to 0}\ \inf_{\tilde{f}_h}\sup_{f\in\mathcal{A}(\alpha)}\mathbf{E}_f\, w\left(\sqrt{\frac{\pi}{\sigma^2 h s_h}}(\tilde{f}_h(x)-f(x))\right)\ \geq\ \mathbf{E}\, w(\xi). \qquad (34)
\end{aligned}$$

Together the relations (26) and (34) prove the theorem. □

4.2 An Extension to Non-fixed Classes

Up till now we assumed that the classes $\mathcal{A}(\alpha)$ were fixed, i.e. not depending on the parameter h, though the function we wanted to estimate could vary freely within the given class $\mathcal{A}(\alpha)$ and, in particular, could depend on h. The possible dependency of f on h implies that the estimated function could be as 'bad' as our model allowed it to be which justified the minimax approach of Theorem 1. To summarize, the assumption that our functional class $\mathcal{A}(\alpha)$ is fixed implies that the smoothness properties of the elements of the class are fixed. However, we might want to further relax this restriction by allowing the class itself depend on h. Indeed, there is neither practical justification, nor a logical

requirement, that the smoothness of the underlying function remains the same while the level of noise decreases and consequently the resolution of the available statistical procedures increases. This will become even more natural in the adaptive setting of Section 5 where the smoothness of the underlying function is not known beforehand.

Thus, as a first step towards introducing the adaptive framework, we let the parameters of the model γ, β and r depend on h. Even so, they still be assumed to be known to the statistician – this assumption will be abolished later in the adaptive framework of Section 5. This approach will allow us to explore the 'limits' of the model where its parameters are allowed to change freely. Let s_h be as defined in Theorem 1. Note that now the optimum bandwidth s_h depends on h also through the parameters γ, β and r. Nevertheless the statement of Theorem 1 still holds, as we shall see, under corresponding assumptions.

Theorem 2 *Let $w \in \mathcal{W}$, and let the parameters $\beta = \beta_h$, $r = r_h$, $\gamma = \gamma_h$ and $\sigma = \sigma_h$ be all positive and such that*

$$0 < \liminf_{h\to 0} r \le \limsup_{h\to 0} r < \infty, \tag{35}$$

$$\liminf_{h\to 0} \frac{\beta^2}{\gamma\,\sigma^2 h} = \infty, \tag{36}$$

$$\limsup_{h\to 0} \frac{h}{\gamma}\left(\log \frac{\beta^2}{\gamma\,\sigma^2 h}\right)^{1/r} = 0. \tag{37}$$

Then

$$\lim_{h\to 0} \sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f\, w\left(\sqrt{\tfrac{\pi}{\sigma^2 h s_h}}(\hat f_h(x) - f(x))\right) =$$

$$\lim_{h\to 0} \inf_{\tilde f_h} \sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f\, w\left(\sqrt{\tfrac{\pi}{\sigma^2 h s_h}}(\tilde f_h(x) - f(x))\right) = \mathbf{E}\, w(\xi)$$

where s_h, $\tilde f_h$ and $\hat f_h$ are the same as in Theorem 1.

Remark 1 *Note that the conditions (35) and (37) imply $hs_h \to 0$ when $h \to 0$. As a direct consequence of this, we obtain consistency, provided σ^2 is bounded, since then $\frac{\sigma^2 h s_h}{\pi} \to 0$. However, our asymptotic optimality result doesn't require σ^2 to be bounded; in other words they apply even when there is no consistency!*

Proof. We prove this theorem following the same proof of Theorem 1. It is sufficient to see that relations (23)–(25) and (33) still hold for the class $\mathcal{A}(\gamma_h, \beta_h, r_h)$. The limit (23) follows from (35) and (37), the limits (24) and (33) follow from (35) and (36). Finally (25) follows from the identity

$$\frac{\beta^2}{\gamma\,\sigma^2 h}\, e^{-\left(2\pi\frac{\gamma}{h}\right)^r/c_r} = \exp\Big\{ -c_r^{-1}\Big(2\pi\frac{\gamma}{h}\Big)^r\Big(1 - \frac{c_r}{(2\pi)^r}\Big(\frac{h}{\gamma}\Big(\log\frac{\beta^2}{\gamma\,\sigma^2 h}\Big)^{1/r}\Big)^r\Big)\Big\} \tag{38}$$

and conditions (35)–(37). Note that $h/\gamma \to 0$, by (36) and (37). The rest of the proof remains the same. □

The important conclusion which can be drawn from the last result is that in order to prove asymptotic optimality of our estimation procedure, we do not have to invoke the assumption – not always realistic – that the smoothness of the estimated function remains the same, even when the level of noise decreases and, as a consequence, the resolution of available statistical methods increases. Note that in this more general situation the corresponding optimal rate of convergence

$$\sigma_h^2(\alpha, \sigma^2) = \frac{\sigma^2 h}{\pi\gamma} \left(\frac{1}{2} \log \frac{\beta^2}{\pi\gamma\, \sigma^2 h} \right)^{\frac{1}{r}}, \tag{39}$$

can be of any order, with respect to any of the parameters, h or $\sigma^2 h$, varying from extremely fast, parametric rates, to extremely slow, non-parametric ones, and even all the way down to no consistency at all. The problem which we will face in next section, is that in practice we often do not know the real class at all.

5 Adaptive Minimax Regression

5.1 Adaptive Estimation in Functional Scales

As a transition from the classical minimax setting, studied in the previous sections, to the adaptive setting we introduce *functional scales*

$$\mathcal{A}_{\mathcal{K}} = \Big\{ \mathcal{A}(\alpha) \mid \alpha \in \mathcal{K} \Big\}, \tag{40}$$

corresponding to a subset $\mathcal{K} \subset \mathbb{R}^3_+$ in the underlying parameter space. As our scales $\mathcal{A}_{\mathcal{K}}$ can be identified with corresponding subsets $\mathcal{K}$, we will speak sometimes about a scale $\mathcal{K}$, instead of $\mathcal{A}_{\mathcal{K}}$, when there is no risk that could lead to a confusion. Sometimes we can think of the scale $\mathcal{A}_{\mathcal{K}}$ as the collection of functions

$$\Big\{ f \in \mathcal{A}(\alpha) \mid \alpha \in \mathcal{K} \Big\}.$$

We will say that some limit exists uniformly in $\mathcal{A}_{\mathcal{K}}$ to express that it exists uniformly in $f \in \mathcal{A}(\alpha)$ for every α and they converge uniformly in $\alpha \in \mathcal{K}$.

Our goal is to estimate a function which belongs to $\mathcal{A}(\alpha)$ for some $\alpha \in \mathcal{K}$. So, we must find an estimator, which does not depend on α and such that it performs "optimally" well over the whole scale $\mathcal{K}$. For this new setting a new definition of optimality is necessary. We use the following definition which was used in Lepski and Levit (1998). From now on we will restrict ourselves to the loss functions $w(x) = |x|^p$, $p > 0$. Let $\mathcal{A}_{\mathcal{K}}$ be a functional scale and $\mathcal{F}$ a class of estimators $\tilde{f}_h$.

Definition 2 *An estimator $\hat{f}_h \in \mathcal{F}$ is called $(p, \mathcal{K}, \mathcal{F})$-adaptively minimax, at a point $x \in \mathbb{R}$, if for any other estimator $\tilde{f}_h \in \mathcal{F}$*

$$\limsup_{h \to 0} \sup_{\alpha \in \mathcal{K}} \frac{\sup_{f \in \mathcal{A}(\alpha)} \mathbf{E}_f |\hat{f}_h(x) - f(x)|^p}{\sup_{f \in \mathcal{A}(\alpha)} \mathbf{E}_f |\tilde{f}_h(x) - f(x)|^p} \leq 1.$$

The simplest example of a scale $\mathcal{A}_{\mathcal{K}}$ can be obtained when $\mathcal{K}$ is a fixed compact subset of $\mathbb{R}^3_+$. Our results below cover a much broader setting in which the set $\mathcal{K}$ itself can depend on the parameter h. In our approach, such results serve two goals. First of all, they allow a better understanding of the true scope of adaptivity of statistical procedures, since they describe the 'extreme' situation in which an adaptation is still possible. In fact all what is needed below is that the assumptions of our 'non-adaptive' Theorem 2 hold uniformly on the scale $\mathcal{K}$; below we formulate these assumptions more explicitly.

Definition 3 *A functional scale $\mathcal{A}_{\mathcal{K}_h}$ (or the corresponding scale $\mathcal{K}_h$) is called a regular, or an R-scale if the following conditions are satisfied:*

$$0 < \liminf_{h\to 0} \inf_{\alpha\in\mathcal{K}_h} r \le \limsup_{h\to 0} \sup_{\alpha\in\mathcal{K}_h} r < \infty, \tag{41}$$

$$\liminf_{h\to 0} \inf_{\alpha\in\mathcal{K}_h} \frac{\beta^2}{\gamma\,\sigma^2 h} = \infty, \tag{42}$$

and

$$\limsup_{h\to 0} \sup_{\alpha\in\mathcal{K}_h} \frac{h^{1-\delta}}{\gamma} \left(\log \frac{\beta^2}{\gamma\,\sigma^2 h}\right)^{1/r} = 0 \tag{43}$$

for some $0 < \delta < 1$.

The second goal that can be achieved by considering more general scales $\mathcal{K}_h$ is to introduce the notion of optimality in adaptive estimation, by specifying a natural set of estimators $\mathcal{F}$ in the above Definition 2. Note that within a large scale $\mathcal{A}_{\mathcal{K}_h}$, unknown functions f can vary from extremely smooth ones, allowing parametric rate $\sigma^2 O(h^2)$, to much less smooth functions, allowing slower rates $\sigma^2 O(h^{2\delta}),\ \delta < 1$, or even extremely slow rates $\sigma^2 O(\log^{-1}(1/h))$. The first possibility is not typical in non-parametric estimation and only can happen in some extreme cases. These ideas are made more precise by introducing the following terminology classifying functional scales $\mathcal{A}_{\mathcal{K}_h}$ into *pseudo-parametric* (PP) and *non-parametric* (NP) scales depending of their global rates of convergence.

Definition 4 *A functional scale $\mathcal{A}_{\mathcal{K}_h}$ (or the corresponding parameter scale $\mathcal{K}_h$) is called*

(a) *pseudo-parametric, or a PP scale if*

$$\limsup_{h\to 0} \sup_{\alpha\in\mathcal{K}_h} s_h(\alpha) < \infty,$$

(b) *non-parametric, or an NP-scale if*

$$\lim_{h\to 0} \inf_{\alpha\in\mathcal{K}_h} s_h(\alpha) = \infty.$$

We shall call regular pseudo-parametric and regular non-parametric scales respectively RPP and RNP scales.

Since pseudo-parametric scales are not typical, in non-parametric estimation and can only happen in some extreme cases, we will only require our statistical procedure to

achieve the optimal rate $\sigma^2 O(h^2)$ for such scales; cf. the Definition of the corresponding classes $\mathcal{F}_p$ below. Note that even with such procedures, a better rate will be achieved, in estimating functions in any pseudo-parametric scale than in any of the non-parametric scales. Further a strong evidence suggests that there is hardly much more one can do than require rate optimality, for any of the pseudo-parametric scales. On the other hand, such an approach allows to develop natural optimality criteria, for any adaptive procedure in the classes $\mathcal{F}$ in the case of non-parametric scales.

Let $\mathcal{F}_p = \mathcal{F}_p(x)$ be the class of all estimators $\tilde{f}_h$ that satisfy

$$\limsup_{h\to 0} \sup_{\alpha\in\mathcal{K}_h} \sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f \left| (\sigma^2 h)^{-1/2} (\tilde{f}_h(x) - f(x)) \right|^p < \infty$$

for arbitrary RPP functional scales $\mathcal{A}_{\mathcal{K}_h}$. Let $\mathcal{F}_p^0 = \mathcal{F}_p^0(x)$ denote the class of estimators such that

$$\limsup_{h\to 0} \mathbf{E}_0 |(\sigma^2 h)^{-1/2}\, \tilde{f}_h(x)|^p < \infty.$$

It is easy to notice that $\mathcal{F}_p \subset \mathcal{F}_p^0$. In the next subsection we present an adaptive estimator $\hat{f}_h \in \mathcal{F}_p$ and prove it to be $(p, \mathcal{K}, \mathcal{F}_p)$-adaptively minimax for arbitrary RNP functional scales.

5.2 The Adaptive Estimator: Upper Bound

Section 5.1 outlined the general adaptive setting, introduced a notion of optimal adaptive estimation and described regular non-parametric scales of infinitely differentiable functions. Our first result describes accuracy which can be achieved for such scales. Its proof starts with the construction of an adaptive estimator achieving this accuracy. In this, the Lepski's method will be used, with the recent modification of Lepski and Levit (1998). Note that the accuracy of our procedure loses a logarithmic factor compared to the non-adaptive case where the parameters of the underlying classes are known. In Section 5.3 we will see that this is an unavoidable pay for not knowing the smoothness *a priori* and we will prove optimality of the proposed procedure in the sense of Definition 2.

Remark 2 *In principle, one could also study adaptation to the unknown parameter σ^2. This however leads to entirely different problems, and is not considered in this thesis. Therefore we always assume that σ^2 is known, although it can vary with h.*

Denote

$$\psi_h^2 = \psi_h^2(\alpha) = p(\log s_h(\alpha))\, \sigma_h^2(\alpha)$$

where $s_h(\alpha)$ and $\sigma_h^2(\alpha)$ were defined in (15) and (20).

Theorem 3 *For any $p > 0$ there exists an adaptive estimator $\hat{f}_h$ such that for any $x \in \mathbb{R}$ and for any RNP functional scale $\mathcal{A}_{\mathcal{K}_h}$, $\hat{f}_h \in \mathcal{F}_p$*

$$\limsup_{h\to 0} \sup_{\alpha\in\mathcal{K}_h} \sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f \left| \psi_h^{-1} (\hat{f}_h(x) - f(x)) \right|^p \le 1.$$

The adaptive estimator. First, let us choose parameters, $1/2 < l < 1$, $1/2 < \delta < 1$, $p_1 > 0$, $l_1 = \delta l$, and define the sequence of bandwidths $s_0 = 0$, $s_i = \exp(i^l)$ for $i = 1, \ldots$. For each h, we take a subsequence $\mathcal{S}_h = \{s_0, s_1, \ldots s_{I_h}\}$ where

$$I_h = \arg\max_i \{hs_i \le \log^{-1} 1/h\}, \tag{44}$$

$h < 1$. Our asymptotic study considers $h \to 0$, thus, without loss of generality, we define I_h just for $h < 1$.

Now, let us denote

$$\hat{f}_i(x) = \hat{f}_{h,s_i}(x), \qquad b_i = \mathbf{E}_f \hat{f}_i(x) - f(x),$$

$$\sigma_i^2 = \mathbf{Var}\, \hat{f}_i(x), \qquad \hat{\sigma}_i^2 = \frac{\sigma^2 h\, s_i}{\pi},$$

$$\sigma_{i,j}^2 = \mathbf{Var}\, (\hat{f}_j(x) - \hat{f}_i(x)), \qquad \hat{\sigma}_{i,j}^2 = \frac{\sigma^2 h\, (s_j - s_i)}{\pi},$$

and define the thresholds

$$\lambda_j^2 = p \log s_j + p_1 \log^\delta s_j.$$

Finally we define

$$\hat{i} = \min \left\{ 1 \le i \le I_h \ : \ |\hat{f}_j(x) - \hat{f}_i(x)| \le \lambda_j\, \hat{\sigma}_{i,j} \quad \forall j \ (i \le j \le I_h) \right\}. \tag{45}$$

We will prove below that the estimator

$$\hat{f}_h(x) = \hat{f}_{\hat{i}}(x)$$

satisfies both the statements contained in Theorem 3.

Let us get first some insight into the algorithm. The sequence $\mathcal{S}_h$ of bandwidths has several important properties. First, it is increasing, thus the variance of the corresponding estimators is also increasing.

Second, according to the definition of R-scales the bandwidths $s_h(\alpha)$, see eq. (43), are such that $hs_h(\alpha) \le h^\delta$ uniformly in $\mathcal{K}_h$ for some $\delta < 1$, and h small enough. Thus, s_{I_h} is large enough for h small enough, so that for each α, the optimum bandwidth $s_h(\alpha)$ corresponding to $\mathcal{A}(\alpha)$, can be sandwiched between two consecutive elements of the sequence $\mathcal{S}_h$, i.e. there exists $i(\alpha) = i(\alpha, h)$ such that

$$s_{i(\alpha)-1} < s_h(\alpha) \le s_{i(\alpha)}.$$

The sequence is also dense enough so that

$$\lim_{i \to \infty} \frac{s_{i+1}}{s_i} = 1.$$

This guarantees that $s_h(\alpha)$ and $s_{i(\alpha)}$ are asymptotically equivalent since $s_h(\alpha) \to \infty$ for $h \to 0$ in NP scales.

The sequence of thresholds λ_j has been chosen in such a way that, for large i, j $(i(\alpha) \le i \le j)$, the probability of the event

$$|\hat{f}_j(x) - \hat{f}_i(x)| > \lambda_j \, \mathbf{Var}^{1/2}(\hat{f}_j(x) - \hat{f}_i(x)), \tag{46}$$

is very small since, except for an event of a small probability, this can only occur if the bias $(b_j - b_i) \gg \mathbf{Var}^{1/2}(\hat{f}_j(x) - \hat{f}_i(x))$ which is not the case for bandwidths greater than $s_h(\alpha)$ as we will see. Therefore, for any given i and $j > i$ we reject s_i in favor of the subsequent elements of the sequence $\mathcal{S}_h$, if the event (46) occurs. This pairwise comparison is performed for every i, and from all the accepted s_i we select the smallest, i.e. we choose the estimator with the smallest variance. Note that according to the previous argument no bandwidth s_i, $i \ge i(\alpha)$ will be rejected, with high probability. However it is possible that a bandwidth s_i, $i < i(\alpha)$ is chosen. In that case our procedure warrants that, cf. (45),

$$|\hat{f}_{\hat{i}}(x) - \hat{f}_{i(\alpha)}(x)| \le \lambda_{i(\alpha)} \, \mathbf{Var}^{1/2}(\hat{f}_{\hat{i}}(x) - \hat{f}_{i(\alpha)}(x))(1 + o(1))$$

Thus in the worst case the accuracy of $\hat{f}_h$ decreases by a factor $1 + \lambda_{i(\alpha)}$ which is of order $\log s_h(\alpha)$ asymptotically as $h \to 0$. In the next subsection we prove that the accuracy of this algorithm is asymptotically optimal in the adaptive setting, for all RNP scales subject to certain mild additional assumptions; see Theorems 1 and 6.

Now, let us turn to the proof of the theorem. We start with an auxiliary result needed in the proof where we use the same notations as those used in describing the estimation procedure.

Lemma 3 *For $h \to 0$, uniformly with respect to i, j $(1 \le i, j \le I_h)$ and with respect to α varying in a regular scale,*

(a) $b_j^2 = o(1)\hat{\sigma}_j^2$ *for all j such that $i(\alpha) \le j \le I_h$.*

(b) $\sigma_j^2 = \hat{\sigma}_j^2(1 + O(\log^{-1}(1/h)))$.

(c) $(b_j - b_i)^2 \le (1 + o(1))\hat{\sigma}_{i,j}^2$ *for all i, j such that $i(\alpha) \le i \le j \le I_h$.*

(d) $\sigma_{i,j}^2 = \hat{\sigma}_{i,j}^2(1 + O(\log^{-1}(1/h)))$.

Proof. (a) Using the bound for the bias given in (17), equation (22), and Lemma 2 we see, with some algebra, that

$$\begin{aligned} b_j^2 &\le \frac{1}{\pi^2} \int_{s_j}^{\infty} \frac{\beta^2}{\gamma} e^{-2(\gamma t)^r} dt + O\left(e^{-2(2\pi\frac{\gamma}{h})^r/c_r}\right) \int_0^{s_j} \beta\gamma e^{2(\gamma t)^r} \, dt \\ &\le \frac{\sigma^2 h s_j}{\pi} \frac{\beta^2}{\pi\gamma\,\sigma^2 h} \frac{e^{-2(\gamma s_j)^r}}{r(\gamma s_j)^r} + O\left(e^{-2(2\pi\frac{\gamma}{h})^r/c_r}\right) \frac{\sigma^2 h s_j}{\pi} \frac{\beta^2}{\pi\gamma\,\sigma^2 h} e^{2(\gamma s_j)^r} \\ &= \hat{\sigma}_j^2 \left(\frac{e^{2(\gamma s_h)^r - 2(\gamma s_j)^r}}{r(\gamma s_h)^r} + O\left(e^{-(2\pi\frac{\gamma}{h})^r/c_r + 2(\gamma s_h)^r} e^{-(2\pi\frac{\gamma}{h})^r/c_r + 2(\gamma s_j)^r}\right) \right). \end{aligned}$$

Now, given $s_j \geq s_h(\alpha)$ and using conditions (44) in the definition of the sequence of bandwidths $\mathcal{S}_h$ and conditions (41)–(43) in the definition of R scales, we obtain $b_j^2 = o(1)\hat{\sigma}_j^2$ when $h \to 0$, uniformly with respect to j ($i(\alpha) \leq j \leq I_h$) and with respect to α in $\mathcal{K}_h$.

(b) This is just a reformulation of the asymptotic relation (18) using the fact that, according to (44), $hs_j \leq \log^{-1}(1/h)$.

(c) Applying Lemma 1(b) taking $s_1 = s_i$ and $s_2 = s_j$, and arguing as in (17) and in the proof (a), we see that

$$
\begin{aligned}
(b_j - b_i)^2 &\leq 2 \left| \frac{1}{2\pi} \int e^{-\mathbf{i}tx} \mathcal{F}[\Delta_{i,j}](t) \mathcal{F}[f](t)\, dt \right|^2 + \\
&\qquad\qquad O\left(e^{-2(2\pi\frac{\gamma}{h})^r/c_r}\right) \int_{s_i}^{s_j} \frac{\beta^2}{\gamma} e^{2(\gamma t)^r}\, dt \\
&\leq \frac{1}{\pi^2} \int_{s_i}^{s_j} \frac{\beta^2}{\gamma} e^{-2(\gamma t)^r} dt + O\left(e^{-2\left(2\pi\frac{\gamma}{h}\right)^r/c_r}\right) \int_{s_i}^{s_j} \frac{\beta^2}{\gamma} e^{2(\gamma t)^r} dt \\
&\leq \frac{\sigma^2 h(s_j - s_i)}{\pi} \frac{\beta^2}{\pi\gamma\,\sigma^2 h} e^{-2(\gamma s_i)^r} + \\
&\qquad\qquad O\left(e^{-2(2\pi\frac{\gamma}{h})^r/c_r}\right) \frac{\sigma^2 h(s_j - s_i)}{\pi} \frac{\beta^2}{\pi\gamma\,\sigma^2 h} e^{2(\gamma s_j)^r} \\
&= \hat{\sigma}_{i,j}^2 \left(e^{2(\gamma s_h)^r - 2(\gamma s_i)^r} + O\left(e^{-(2\pi\frac{\gamma}{h})^r/c_r + 2(\gamma s_h)^r} e^{-(2\pi\frac{\gamma}{h})^r/c_r + 2(\gamma s_j)^r}\right)\right) \\
&= \hat{\sigma}_{i,j}^2 (1 + o(1)), \qquad\qquad (h \to 0).
\end{aligned}
$$

(d) It follows directly from Lemma 1(c), taking $s_1 = s_i$ and $s_2 = s_j$. Here, as in (18), we can verify that

$$
\begin{aligned}
\sigma_{i,j}^2 &= \sigma^2 h^2 \textstyle\sum_{\ell=-\infty}^{\infty} \left(k_{s_j}(x - \ell h) - k_{s_i}(x - \ell h)\right)^2 \\
&= \tfrac{\sigma^2 h\,(s_j - s_i)}{\pi} (1 + O(1)\, h(s_j - s_i)). \qquad (47)
\end{aligned}
$$

and thus, using (44), this completes the proof of the lemma. □

We now proceed with proving Theorem 3. For arbitrary f in any R-functional scale $\mathcal{A}_{\mathcal{K}_h}$,

$$R_h(f) := \mathbf{E}\,|\hat{f}_{\hat{i}}(x) - f(x)|^p \;=\; R_h^-(f) + R_h^+(f)$$

where

$$R_h^-(f) = \mathbf{E}\left\{ \mathbb{1}_{\{\hat{i} \leq i(\alpha)\}} |\hat{f}_{\hat{i}}(x) - f(x)|^p \right\}$$

and

$$R_h^+(f) = \mathbf{E}\left\{ \mathbb{1}_{\{\hat{i} > i(\alpha)\}} |\hat{f}_{\hat{i}}(x) - f(x)|^p \right\}.$$

Let us examine $R_h^-(f)$ first. We have

$$\Big\{\hat{i} \le i(\alpha)\Big\} \subset \Big\{\, |\hat{f}_{\hat{i}}(x) - \hat{f}_{i(\alpha)}(x)| \le \lambda_{i(\alpha)}\hat{\sigma}_{\hat{i},i(\alpha)} \,\Big\}$$

$$\subset \Big\{\, |\hat{f}_{\hat{i}}(x) - \hat{f}_{i(\alpha)}(x)| \le \lambda_{i(\alpha)}\hat{\sigma}_{i(\alpha)} \,\Big\},$$

therefore

$$R_h^-(f) \le \mathbf{E}\Big(\mathbb{1}_{\{\hat{i}\le i(\alpha)\}} \Big(\,|\, \hat{f}_{\hat{i}}(x) - \hat{f}_{i(\alpha)}(x) \,|\, + |\, \hat{f}_{i(\alpha)}(x) - f(x) \,|\, \Big)^p \Big)$$

$$\le \mathbf{E}\,\Big(\lambda_{i(\alpha)}\hat{\sigma}_{i(\alpha)} + |\, \hat{f}_{i(\alpha)}(x) - f(x) \,|\, \Big)^p$$

$$\le \mathbf{E}\Big(\lambda_{i(\alpha)}\hat{\sigma}_{i(\alpha)} + |b_{i(\alpha)}| + \sigma_{i(\alpha)}|\xi| \,\Big)^p$$

where $\xi \sim \mathcal{N}(0,1)$. Now according to Lemma 3, (a) and (b), uniformly with respect to α in any regular scale

$$\sigma_{i(\alpha)} = \hat{\sigma}_{i(\alpha)}(1 + o(1)) \quad \text{and} \quad |b_{i(\alpha)}| = o(1)\hat{\sigma}_{i(\alpha)}, \qquad\qquad (h \to 0).$$

It follows that for $h \to 0$ uniformly with respect to any RPP scale

$$R_h^-(f) = O(h^{p/2}), \tag{48}$$

while by the dominated convergence theorem, uniformly in any RNP scale

$$R_h^-(f) \le \psi_h^p(\alpha)(1 + o(1)). \tag{49}$$

Now let us examine $R_h^+(f)$. Consider the auxiliary events

$$A_i = \Big\{ \omega : |\, \hat{f}_i(x) - f(x) \,|\, \le \sqrt{2}\,\lambda_i\hat{\sigma}_i \Big\}.$$

Applying Hölder's inequality we obtain

$$R_h^+(f) = \mathbf{E}\left(\mathbb{1}_{\{\hat{i} > i(\alpha)\}} \,|\hat{f}_{\hat{i}}(x) - f(x)|^p \right) \;=\; \sum_{i=i(\alpha)+1}^{I_h} \mathbf{E}\left(\mathbb{1}_{\{\hat{i}=i\}} \,|\hat{f}_i(x) - f(x)|^p \right)$$

$$= \sum_{i=i(\alpha)+1}^{I_h} \mathbf{E}\left(|\hat{f}_i(x) - f(x)|^p \,(\, \mathbb{1}_{\{\hat{i}=i\}\cap A_i} + \mathbb{1}_{\{\hat{i}=i\}\cap A_i^c} \,) \right)$$

$$\le R_{h,1}^+(f) + R_{h,2}^+(f),$$

where

$$R_{h,1}^+(f) \;=\; \sum_{i=i(\alpha)+1}^{I_h} (2\lambda_i^2\hat{\sigma}_i^2)^{p/2}\, \mathbf{P}(\hat{i} = i)$$

and

$$R^{+}_{h,2}(f) \;=\; \sum_{i=i(\alpha)+1}^{I_h} \mathbf{E}^{1/2} \left|\hat{f}_i(x) - f(x)\right|^{2p} \mathbf{P}^{1/2}(A_i^c).$$

We have

$$\mathbf{P}(\hat{i}=i) \le \sum_{j=i+1}^{\infty} \mathbf{P}\left(|\hat{f}_{j-1}(x) - \hat{f}_{i-1}(x)| > \hat{\sigma}_{i-1,j-1}\,\lambda_{j-1} \right). \tag{50}$$

By writing $\hat{f}_j(x) - \hat{f}_i(x) = \sigma_{i,j}\xi + b_j - b_i$, where $\xi \sim \mathcal{N}(0,1)$, applying Lemma 3(d), and using the well known bound on the tails of the normal distribution (cf. Feller, 1968, Lemma 2), we find for some $C>0$ and all h small enough

$$\mathbf{P}\Big(|\hat{f}_j(x) - \hat{f}_i(x)| > \lambda_j\hat{\sigma}_{i,j} \Big) \;\le\; \mathbf{P}\Big(|\xi| > \lambda_j\frac{\hat{\sigma}_{i,j}}{\sigma_{i,j}} - \frac{|b_j-b_i|}{\sigma_{i,j}} \Big)$$

$$\le\; \exp\Big\{ -\frac{1}{2}\Big(\lambda_j\frac{\hat{\sigma}_{i,j}}{\sigma_{i,j}} - C\Big)^2 \Big\} \;\le\; \exp\Big\{ -\frac{1}{2}\lambda_j^2\frac{\hat{\sigma}^2_{i,j}}{\sigma^2_{ij}} + C\lambda_j\frac{\hat{\sigma}_{i,j}}{\sigma_{i,j}} \Big\}$$

$$\le\; \exp\left\{ -\frac{1}{2}\lambda_j^2 + C\lambda_j\frac{\hat{\sigma}_{i,j}}{\sigma_{i,j}} + \frac{1}{2}\lambda_j^2\left(1 - \frac{\hat{\sigma}^2_{i,j}}{\sigma^2_{ij}}\right) \right\}.$$

Since by Lemma 3(c) and (44)

$$\lambda_j^2\frac{\sigma^2_{i,j} - \hat{\sigma}^2_{i,j}}{\sigma^2_{i,j}} = \lambda_j^2\, O(\log^{-1}(1/h)) \;=\; o(1), \qquad\qquad (h\to 0),$$

it follows from the last inequality that for some $C_1>0$

$$\mathbf{P}\Big(|\hat{f}_j(x) - \hat{f}_i(x)| > \lambda_j\hat{\sigma}_{i,j} \Big) \;\le\; C_1 \exp\left\{ -\frac{1}{2}\lambda_j^2 + 2C\lambda_j \right\}$$

for all $\alpha,\ j \ge i \ge i(\alpha)$ and all sufficiently small h.

Returning to (50) we obtain that

$$\mathbf{P}(\hat{i}=i) \le C_1 \sum_{j=i+1}^{\infty} \exp\left\{ -\frac{1}{2}\lambda_{j-1}^2 + 2C\lambda_{j-1} \right\} \;=\; C_1 \sum_{j=i}^{\infty} \exp\left\{ -\frac{1}{2}\lambda_j^2 + 2C\lambda_j \right\}$$

$$= C_1 \sum_{j=i}^{\infty} \exp\left\{ -\frac{pj^l + p_1 j^{l_1}}{2} + 2C\sqrt{pj^l + p_1 j^{l_1}} \right\}$$

$$\le C_1 \sum_{j=i}^{\infty} \exp\Big\{ -\frac{pj^l}{2} - \frac{p_1 j^{l_1}}{3} \Big\} \;\sim\; C_1\frac{2}{pl} i^{1-l} \exp\Big\{ -\frac{pi^l}{2} - \frac{p_1 i^{l_1}}{3} \Big\}$$

$$= C_1\frac{2}{pl} i^{1-l} s_i^{-p/2} \exp\Big\{ -\frac{p_1 i^{l_1}}{3} \Big\} \;\le\; C_2 s_i^{-p/2} \exp\Big\{ -\frac{p_1 i^{l_1}}{4} \Big\} \tag{51}$$

for some $C_2 > 0$ and all $i \geq i(\alpha)$, when h is sufficiently small. Therefore uniformly in $\mathcal{A}_{\mathcal{K}_h}$

$$R^+_{h,1}(f) = O(h^{p/2}) \sum_{i=1}^{\infty} i^{pl/2} \exp\Big\{ - p_1 i^{l_1}/4 \Big\} \; = \; O(h^{p/2}), \qquad (h \to 0).$$

In order to obtain a bound on $R^+_{h,2}(f)$ we write again $\hat{f}_i - f(x) = b_i + \sigma_i \xi$, $\xi \sim \mathcal{N}(0,1)$. Applying Lemma 3, (a) and (b), in the same way as before, we have

$$\begin{aligned}
\mathbf{P}(A_i^c) &\leq \mathbf{P}\Big(|\xi| > \sqrt{2}\,\lambda_i \frac{\hat{\sigma}_i}{\sigma_i} - \frac{|b_i|}{\sigma_i}\Big) \; \leq \; \mathbf{P}\left(|\xi| > \sqrt{2}\,\lambda_i \frac{\hat{\sigma}_i}{\sigma_i} - \sqrt{2}\right) \\
&\leq \exp\left\{ -\frac{1}{2}\left(\sqrt{2}\,\lambda_i \frac{\hat{\sigma}_i}{\sigma_i} - \sqrt{2} \right)^2 \right\} \; \leq \; C_3 \exp\Big\{ - \lambda_i^2 + 2\,\lambda_i \Big\} \\
&\leq C_3 \exp\Big\{ - p i^l - p_1 i^{l_1}/2 \Big\} \; = \; C_3 s_i^{-p} \exp\Big\{ - p_1 i^{l_1}/2 \Big\},
\end{aligned}$$

for some C_3, all $i \geq i(\alpha)$ and all α provided h is small enough. Thus,

$$\begin{aligned}
R^+_{h,2}(f) &= \sum_{i=i(\alpha)+1}^{I_h} \mathbf{E}^{1/2} |\hat{f}_i(x) - f(x)|^{2p}\, \mathbf{P}^{1/2}(A_i^c) \\
&\leq \sum_{i=i(\alpha)+1}^{I_h} \hat{\sigma}_i^p\, \mathbf{E}^{1/2} |\, o(1) + (1 + o(1))\xi\,|^{2p}\, \mathbf{P}^{1/2}(A_i^c) \\
&= O\Big(\frac{\sigma^2 h}{\pi}\Big)^{p/2} \sum_{i=1}^{\infty} \exp\Big\{ - p_1 i^{r_1}/4 \Big\} \\
&= O(h^{p/2}), \qquad (h \to 0), \qquad (52)
\end{aligned}$$

uniformly in $\mathcal{A}_{\mathcal{K}_h}$.

We can thus conclude that, uniformly in any RPP scale $\mathcal{K}_h$, our estimator satisfies

$$\sup_{\alpha \in \mathcal{K}_h} \sup_{f \in \mathcal{A}(\alpha)} \mathbf{E} \Big| h^{-1/2} (\hat{f}_h(x) - f(x)) \Big|^p = O(1),$$

while for any RNP scale $\mathcal{K}_h$

$$\sup_{\alpha \in \mathcal{K}_h} \sup_{f \in \mathcal{A}(\alpha)} \mathbf{E} \Big| \psi_h^{-1}(\alpha) (\hat{f}_h(x) - f(x)) \Big|^p \leq 1 + o(1),$$

when $h \to 0$. □

5.3 Lower Bound: Optimality Results

In Section 5.2 we have established an upper bound for the risk of adaptive procedures, by evaluating the quality of a proposed adaptive estimator. In this section we will establish a lower bound for arbitrary such estimator, which will allow us to establish optimality of the proposed procedure in the sense of Definition 2.

Theorem 4 *Let $p > 0$. Let $\mathcal{A}_{\mathcal{K}_h}$ be an arbitrary RNP scale such that quantities $\tilde{s}_h = \tilde{s}_h(\alpha)$, $\tilde{s}_h \le s_h(\alpha)$, and $\tilde{\phi}_h(\alpha)$ can be defined in such a way that for all sufficiently small h and $\alpha \in \mathcal{K}_h$*

$$\tilde{\phi}_h^2 = \tilde{\phi}_h^2(\alpha) \le \min(\, p \log \tilde{s}_h \,, r(\gamma \tilde{s}_h)^r/2 \,) \tag{53}$$

and

$$\lim_{h\to 0} \inf_{\alpha \in \mathcal{K}_h} \tilde{\phi}_h = \infty. \tag{54}$$

Denote

$$\tilde{\psi}_h^2 = \tilde{\psi}_h^2(\alpha) = \frac{\sigma^2 h\, \tilde{s}_h}{\pi} \tilde{\phi}_h^2.$$

Then for any estimator $\tilde{f}_h \in \mathcal{F}_p^0(x)$

$$\liminf_{h\to 0} \inf_{\alpha \in \mathcal{K}_h} \sup_{f \in \mathcal{A}(\alpha)} \mathbf{E}_f \left| \tilde{\psi}_h^{-1} (\tilde{f}_h(x) - f(x)) \right|^p \ge 1.$$

Proof. Letting $\theta = \tilde{\phi}_h - \sqrt{\tilde{\phi}_h}$ consider the following pair of functions:

$$\begin{aligned} &f_0(z) \equiv 0, \\ &f_1(z) = \theta\, \tilde{g}(z), \quad \tilde{g}(z) = \sqrt{\frac{\sigma^2 h\, \pi}{\tilde{s}_h}} k_{\tilde{s}_h}(x - z). \end{aligned} \tag{55}$$

Note that f_1 satisfies

$$f_1(x) = \theta \sqrt{\frac{\sigma^2 h\, \tilde{s}_h}{\pi}}.$$

Obviously f_1 is a continuous function and using (10), definition (15) of s_h, and Lemma 2, we get

$$\begin{aligned} \int \tfrac{\gamma}{\beta^2} e^{2|\gamma t|^r} |\, \mathcal{F}[f_1](t)|^2\, dt \;&=\; \theta^2 \tfrac{\sigma^2 h\, \pi}{\tilde{s}_h} \int \tfrac{\gamma}{\beta^2} e^{2|\gamma t|^r} |\mathcal{F}[k_{\tilde{s}_h}](t)|^2\, dt \\ &=\; 2\,\theta^2 \tfrac{\gamma \sigma^2 h\, \pi}{\beta^2} \tfrac{\int_0^{\tilde{s}_h} \gamma e^{2(\gamma t)^r}\, dt}{\gamma \tilde{s}_h} \;=\; 2\,\theta^2\, e^{-2(\gamma s_h)^r}\, \tfrac{\int_0^{\tilde{s}_h} \gamma e^{2(\gamma t)^r}\, dt}{\gamma \tilde{s}_h} \\ &=\; \tfrac{\theta^2}{r(\gamma \tilde{s}_h)^r} e^{2(\gamma \tilde{s}_h)^r - 2(\gamma s_h)^r} (1 + o(1)) \;\le\; \tfrac{\tilde{\phi}_h^2}{r(\gamma \tilde{s}_h)^r} e^{2(\gamma \tilde{s}_h)^r - 2(\gamma s_h)^r} (1 + o(1)) \\ &\le\; \tfrac{1}{2}(1 + o(1)) \le 1, \end{aligned} \tag{56}$$

uniformly in $\mathcal{K}_h$ for h small enough. Thus $f_1 \in \mathcal{A}(\alpha)$ for all sufficiently small h and every $\alpha \in \mathcal{K}_h$.

Let $\tilde{f}_h \in \mathcal{F}_p^0(x)$ be an arbitrary estimator and denote $f_h^* = \tilde{\psi}_h^{-1}\tilde{f}_h(x)$ and $L = \tilde{\phi}_h^{-1}\theta$; then

$$\psi_h^{-1}(\tilde{f}_h(x) - f_1(x)) = f_h^* - \psi_h^{-1}f_1(x) = f_h^* - \tilde{\phi}_h^{-1}\theta = f_h^* - L \tag{57}$$

whereas

$$\begin{aligned}\sqrt{\frac{\pi}{\sigma^2 h}}(\tilde{f}_h(x) - f_0(x)) &= \sqrt{\frac{\pi}{\sigma^2 h}}\,\tilde{f}_h(x) = \sqrt{\frac{\pi}{\sigma^2 h}}\,\tilde{\psi}_h\, f_h^*(x)\\ &= \sqrt{\frac{\pi}{\sigma^2 h}}\,\sqrt{\frac{\sigma^2 h\,\tilde{s}_h}{\pi}}\,\tilde{\phi}_h\, f_h^*(x) = \tilde{s}_h^{1/2}\,\tilde{\phi}_h\, f_h^*(x)\\ &= f_h^* \exp\Big\{\frac{\log \tilde{s}_h}{2} + \log\tilde{\phi}_h\Big\}.\end{aligned} \tag{58}$$

Denote $q = \exp\{-\tilde{\phi}_h\}$ so that by (54), $q \to 0$ uniformly with respect to α for $h \to 0$. Now, with the thus defined $f_1 \in \mathcal{A}(\alpha)$, for any $\tilde{f}_h \in \mathcal{F}_p^0(x)$, uniformly in $\alpha \in \mathcal{K}_h$ as $h \to 0$, we have

$$\begin{aligned}\mathcal{R}_h &:= \sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f\left(\tilde{\psi}_h^{-1}|\tilde{f}_h(x) - f(x)|\right)^p \geq \mathbf{E}_1\left(\tilde{\psi}_h^{-1}|\tilde{f}_h(x) - f_1(x)|\right)^p\\ &\geq q\,\mathbf{E}_0\left(\sqrt{\frac{\pi}{\sigma^2 h}}|\tilde{f}_h(x) - f_0(x)|\right)^p + (1-q)\,\mathbf{E}_1\left(\tilde{\psi}_h^{-1}|\tilde{f}_h(x) - f_1(x)|\right)^p + O(q).\end{aligned} \tag{59}$$

According to (53) and (57)–(59),

$$\begin{aligned}\mathcal{R}_h &\geq q\,\exp\Big\{\frac{\tilde{\phi}_h}{2} + p\log\tilde{\phi}_h\Big\}\,\mathbf{E}_0\,|f_h^*(x)\,|^p + (1-q)\,\mathbf{E}_1\,|f_h^*(x) - L|^p + O(q)\\ &\geq (1-q)\,\mathbf{E}_1\left(Z|f_h^*(x)|^p + |f_h^*(x) - L|^p\right) + O(q)\\ &\geq (1-q)\,\mathbf{E}_1\,\inf_x\left(Z|x|^p + |x - L|^p\right) + O(q)\end{aligned} \tag{60}$$

where

$$Z = q\,\exp\Big\{\frac{\tilde{\phi}_h}{2} + p\log\tilde{\phi}_h\Big\}\,\frac{d\mathbf{P}_0^{(h)}}{d\mathbf{P}_1^{(h)}}(\mathbf{y}).$$

For each value of Z consider the optimization problem of minimizing the function:

$$g(x) = Z|x|^p + |L - x|^p.$$

As was shown in Lepski and Levit (1998),

$$\min_x g(x) = \begin{cases}\min(Z,1)L^p & \text{if } p \leq 1,\\ \left(1 + Z^{-\frac{1}{p-1}}\right)^{-(p-1)}L^p & \text{if } p > 1.\end{cases} \tag{61}$$

Thus for any $p > 0$ we can write

$$\min_x g(x) = \chi L^p, \tag{62}$$

where χ is defined by (61) and satisfies $0 < \chi \le 1$.

Now, let us consider the likelihood corresponding to f_0 and f_1. Using the same arguments that we used in (28)–(32) we can see that

$$\begin{aligned}\frac{d\mathbf{P}_0^{(h)}}{d\mathbf{P}_1^{(h)}}(\mathbf{y}) &= \exp\left\{-\frac{1}{2\sigma^2}\sum_{\ell=-\infty}^{\infty}(\theta^2\tilde{g}^2(\ell h) + 2\theta\, y_\ell\, \tilde{g}(\ell h))\right\},\\ &= \exp\left\{\Big(-\theta\xi - \frac{\theta^2}{2}\Big)\Big(\frac{\pi}{\tilde{s}_h}\, h \sum_{\ell=-\infty}^{\infty} k^2_{\tilde{s}_h}(x-\ell h)\Big)\right\}\\ &= \exp\left\{\Big(-\theta\xi - \frac{\theta^2}{2}\Big)\Big(1 + O(1)h\tilde{s}_h\Big)\right\}\end{aligned}$$

where $\xi \sim \mathcal{N}(0,1)$ with respect to $\mathbf{P}_1$. Using the definition of θ, condition (53) and definition (55) we can see that

$$\frac{d\mathbf{P}_0^{(h)}}{d\mathbf{P}_1^{(h)}}(\mathbf{y}) = (1+o(1))\,\exp\left\{-\frac{\theta^2}{2} - \theta\xi\right\}, \quad (h \to 0).$$

Note that by (54)

$$Z = (1+o(1))\exp\left\{-\tilde{\phi}_h + \frac{\tilde{\phi}_h^2}{2} + p\log\tilde{\phi}_h - (\tilde{\phi}_h - \sqrt{\tilde{\phi}_h})\xi - \frac{1}{2}(\tilde{\phi}_h - \sqrt{\tilde{\phi}_h})^2\right\} \overset{\mathbf{P}_1}{\to} \infty$$

when $h \to 0$, hence $\chi \overset{\mathbf{P}_1}{\to} 1$. Also $L = 1 + o(1)$, according to its definition. Therefore according to equations (60)–(62), uniformly in $\alpha \in \mathcal{K}_h$,

$$\mathcal{R}_h \ge (1-q)L^p\mathbf{E}_1\chi + O(q) = 1 + o(1), \qquad (h \to 0).$$

□

Corollary 1 *Let $\mathcal{A}_{\mathcal{K}_h}$ be an arbitrary RNP scale such that*

$$\liminf_{h\to 0}\ \inf_{\alpha\in\mathcal{K}_h} \frac{r(\gamma s_h)^r}{\log s_h} = \infty \tag{63}$$

where s_h is the optimum bandwidth defined in (15). Then for any $p > 0$ and $x \in \mathbb{R}$, the estimator $\hat{f}_h$ of Theorem 3 is $(p, \mathcal{K}_h, \mathcal{F}_p(x))$-adaptively minimax at x.

Proof. This is a consequence of Theorems 3 and 4. In order to prove the lower bound use the previous theorem taking s_h in place of $\tilde{s}_h$. □

Now, we prove a version of Theorem 4 under a weaker condition. It will be used below to provide an easily verifiable conditions for adaptive optimality of the estimator proposed in Section 5.2.

Theorem 5 *Let $\mathcal{A}_{\mathcal{K}_h}$ be an arbitrary RNP scale such that*

$$\liminf_{h\to 0} \inf_{\alpha\in\mathcal{K}_h} \frac{r(\gamma s_h)^r}{\log\log s_h} = \infty \tag{64}$$

where the optimum bandwidth s_h was defined in (15). Then for any estimator $\tilde{f}_h \in \mathcal{F}_p^0(x)$,

$$\liminf_{h\to 0} \inf_{\alpha\in\mathcal{K}_h} \sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f \left|\psi_h^{-1}(\tilde{f}_h(x) - f(x))\right|^p \geq 1,$$

where

$$\psi_h^2 = \psi_h^2(\alpha) = p\,(\log s_h)\frac{\sigma^2 h s_h}{\pi}.$$

Proof. We prove this theorem in the same way as Theorem 4 by choosing $\tilde{\phi}_h^2 = p\log\tilde{s}_h$ and subsequently defining $\tilde{s}_h$ in such a way that

$$\frac{2p\log\tilde{s}_h}{r(\gamma\tilde{s}_h)^r}\, e^{2(\gamma\tilde{s}_h)^r - 2(\gamma s_h)^r} \leq 1 \tag{65}$$

for h small enough. The point here is that condition (65) was only needed in proving (56), which now becomes (65). We construct an appropriate $\tilde{s}_h$ asymptotically equivalent to s_h that satisfies the previous inequality for h small enough. Let us first, for fixed α, define the auxiliary bandwidth $\bar{s}_h$ as the solution of the equation

$$2(\gamma s_h)^r = 2(\gamma\bar{s}_h)^r + \log r(\gamma\bar{s}_h)^r. \tag{66}$$

We know that γs_h goes to infinity as h goes to zero uniformly in regular scales. Thus from the previous equation, $\gamma\bar{s}_h$ goes to infinity too and we can see that

$$\left(\frac{s_h}{\bar{s}_h}\right)^r = 1 + \frac{\log r(\gamma\bar{s}_h)}{2(\gamma\bar{s}_h)^r} = 1 + o(1),$$

uniformly in $\mathcal{K}_h$ according to (64). Thus the auxiliary bandwidth $\bar{s}_h$ is asymptotically equivalent to s_h. It also satisfies (64), see that

$$\liminf_{h\to 0} \inf_{\alpha\in\mathcal{K}_h} \frac{r(\gamma\bar{s}_h)^r}{\log\log\bar{s}_h} = \liminf_{h\to 0} \inf_{\alpha\in\mathcal{K}_h} \frac{r(\gamma s_h)^r}{\log\log\bar{s}_h}(1+o(1)) \geq \liminf_{h\to 0} \inf_{\alpha\in\mathcal{K}_h} \frac{r(\gamma s_h)^r}{\log\log s_h} = \infty.$$

Now, let us define $\tilde{s}_h = \vartheta\bar{s}_h$ where ϑ ($0 < \vartheta < 1$) is the closest solution to 1 of the equation

$$\frac{2r(\gamma\bar{s}_h)^r}{\log\log\bar{s}_h}\,\vartheta^r\log\vartheta^{-1} = 1.$$

We can see that $\vartheta \to 1$ as $h \to 0$ thus implying that $\tilde{s}_h$ is asymptotically equivalent to $\bar{s}_h$ and s_h. Now, after few transformations,

$$-2(\gamma\tilde{s}_h)^r = -2(\gamma\bar{s}_h)^r + 2\int_{\tilde{s}_h}^{\bar{s}_h} r(\gamma t)^r t^{-1}\,dt$$

$$= -2(\gamma s_h)^r \log r(\gamma \bar{s}_h)^r + 2\int_{\tilde{s}_h}^{\bar{s}_h} r(\gamma t)^r t^{-1}\, dt$$

$$\geq -2(\gamma s_h)^r + \log r(\gamma \bar{s}_h)^r + 2r(\gamma \tilde{s}_h)^r \int_{\tilde{s}_h}^{\bar{s}_h} t^{-1}\, dt$$

$$= -2(\gamma s_h)^r + \log r(\gamma \bar{s}_h)^r + 2r(\gamma \bar{s}_h)^r \vartheta^r \log \vartheta^{-1}$$

$$= -2(\gamma s_h)^r + \log r(\gamma \bar{s}_h)^r + \log\log \bar{s}_h$$

and we see that

$$e^{-2(\gamma \tilde{s}_h)^r} \geq e^{-2(\gamma s_h)^r} r(\gamma \bar{s}_h)^r \log \bar{s}_h = e^{-2(\gamma s_h)^r \frac{2p \log \tilde{s}_h}{r(\gamma \tilde{s}_h)}} \vartheta^r r^2 (\gamma \bar{s}_h)^{2r}/(2p)$$

$$\geq e^{-2(\gamma s_h)^r \frac{2p \log \tilde{s}_h}{r(\gamma \tilde{s}_h)}}$$

for h small enough. The rest of the proof is the same as for Theorem 4. Finally, given $\tilde{\psi}_h := p \log \tilde{s}_h \frac{\sigma^2 h \tilde{s}_h}{\pi}$ is asymptotically equivalent to ψ_h we have the proof of the lemma. □

Finally, we prove that the estimator we constructed in Theorem 3 is adaptively minimax, for any RNP scale satisfying a condition just a little stronger than condition (42) used in the definition of a regular scale.

Theorem 6 *Let $\mathcal{K}_h$ be a RNP scale such that*

$$\liminf_{h\to 0} \inf_{\alpha \in \mathcal{K}_h} \frac{\beta^2}{\gamma \sigma^2 h^{1-\delta}} \geq C$$

for some δ ($0 < \delta < 1$) and $C > 0$. Then for any $p > 0$ and $x \in \mathbb{R}$, the estimator $\hat{f}_h$ of Theorem 3 is $(p, \mathcal{K}_h, \mathcal{F}_p(x))$-adaptively minimax at x.

Proof. The upper bound result was proved in Theorem 3. To prove the lower bound we notice that

$$r(\gamma s_h)^r = \frac{r}{2} \log \frac{\beta^2}{\pi \gamma \sigma^2 h} \geq \frac{r}{2} \log C h^{-\delta}$$

while according to conditions (41)–(43) for R scales

$$\log\log s_h = \log\log \frac{1}{\gamma} \left(\frac{1}{2} \log \frac{\beta^2}{\pi \gamma \sigma^2 h} \right)^{1/r} < \log\log h^{-1}$$

thus $\frac{r(\gamma s_h)^r}{\log\log s_h}$ goes to infinity when $h \to 0$, uniformly with respect to the scale $\mathcal{K}_h$. The desired lower bound follows now from Theorem 5. □

References

P. Antonsik, J. Mikusiński, and R. Sikorski. *Theory of Distribution. The Sequential Approach*. Elsevier, Amsterdam, 1973.

L.D. Brown and M.G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24:2384–2398, 1996.

W. Feller. *An Introduction to Probability Theory and its Applications*, volume I. Wiley, New York, 3rd edition, 1968.

G.K. Golubev and B.Y. Levit. Asymptotically efficient estimation for analytic distributions. *Math. Meth. Statist.*, 5:357–368, 1996.

G.K. Golubev, B.Y. Levit, and A.B. Tsybakov. Asymptotically efficient estimation of analytic functions in Gaussian noise. *Bernoulli*, 2:167–181, 1996.

Kuo Hui-Hsiung. *Gaussian Measures in Banach Spaces*. Number 463 in Lect. Notes Math. Springer-Verlag, Berlin-Heidelberg-New York, 1975.

I.A. Ibragimov and R.I. Has'minskii. *Statistical Estimation, Asymptotic Theory*. Springer, New York, 1981.

I.A. Ibragimov and R.I. Has'minskii. Bounds for the risks of non-parametric regression estimates. *Theor. Probab. Appl.*, 27:84–99, 1982.

I.A. Ibragimov and R.I. Has'minskii. Estimation of distribution density. *Journ. Sov. Math.*, 25:40–57, 1983.

O.V. Lepski. On a problem of adaptive estimation in Gaussian noise. *Theory Probab. Appl.*, 35:454–466, 1990.

O.V. Lepski. Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.*, 36:682–697, 1991.

O.V. Lepski. Asymptotically minimax adaptive estimation. II: Schemes without optimal adaptation. Adaptive estimators. *Theory Probab. Appl.*, 7:433–448, 1992a.

O.V. Lepski. On problems of adaptive estimation in white Gaussian noise. *Adv. Soc. Math.*, 12:87–106, 1992b.

O.V. Lepski and B.Y. Levit. Adaptive minimax estimation of infinitely differentiable functions. *Math. Meth. Statist.*, 7:123–156, 1998.

O.V. Lepski and B.Y. Levit. Adaptive non-parametric estimation of smooth multivariate functions. *Math. Meth. Statist.*, 8:344–370, 1999.

B.Y. Levit. On the asymptotic minimax estimates of the second order. *Theory Prob. Appl.*, 25:552–568, 1980.

S. Nikol'skiĭ. *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer-Verlag, Berlin Heidelberg New York, 1975.

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

C.J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10:1040–1053, 1982.

Authors' addresses:

L.M. Artiles
Eurandom
P.O. Box 513
5600 MB Eindhoven
The Netherlands

E-mail: artiles@mfc.uclv.edu.cu

B.Y. Levit
Department of Mathematics & Statistics
Queen's University
Kingston, ON, K7L 3N6
Canada

E-mail: blevit@mast.queensu.ca