Adaptive Minimax Estimation in Classes of Smooth Functions

Luis M. Artiles Martínez

Adaptive Minimax Estimation in Classes of Smooth Functions

Adaptieve Minimax Schatting in Klassen van Gladde Functies

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht, op gezag van de Rector
Magnificus, Prof. dr. W.H. Gispen, ingevolge het
besluit van het College voor Promoties
in het openbaar te verdedigen op
maandag 3 december 2001 des ochtends om 10:30 uur

door

Luis M. Artiles Martínez

geboren op 8 juni 1970, te Santa Clara, Cuba

Promotor : Prof. Dr. R.D. Gill
Mathematical Institute, Utrecht University, The Netherlands
Co-Promotor: Prof. Dr. B.Y. Levit
Department of Mathematics & Statistics, Queen's University, Canada

---

1991 Mathematics Subject Classification: 62G05, 62G07, 62G20

Artiles Martínez, Luis Manuel

Adaptive Minimax Estimation in Classes of Smooth Functions
Luis Manuel Artiles Martínez
Utrecht, Faculteit Wiskunde en Informatica
Proefscrift Universiteit Utrecht.– met samenvatting in het Nederlands
ISBN: 90-3932-903-6

Contents

1 Introduction **1**
- 1.1 Brief history of adaptive estimation . . . 2
- 1.2 Scope of the thesis . . . 4
  - 1.2.1 Adaptive regression on the real line . . . 4
  - 1.2.2 Adaptive regression on a bounded interval . . . 7
  - 1.2.3 Adaptive density estimation . . . 9

2 Adaptive regression on the real line **13**
- 2.1 The model . . . 14
- 2.2 Auxiliary results . . . 16
- 2.3 Minimax regression in $\mathcal{A}(\gamma, \beta, r)$ . . . 20
  - 2.3.1 Optimality in the case of fixed classes . . . 20
  - 2.3.2 An extension to non-fixed classes . . . 25
- 2.4 Adaptive minimax regression . . . 26
  - 2.4.1 Adaptive estimation in functional scales . . . 26
  - 2.4.2 The adaptive estimator: upper bound . . . 29
  - 2.4.3 Lower bound: optimality results . . . 36

3 Adaptive regression on a bounded interval **43**
- 3.1 The building blocks . . . 43
  - 3.1.1 The class $\mathcal{A}(\gamma, M)$ . . . 44
  - 3.1.2 Legendre polynomials . . . 44
  - 3.1.3 Chebyshev polynomials . . . 48
- 3.2 Minimax regression in $\mathcal{A}(\gamma, M)$ . . . 52
  - 3.2.1 The statistical setting . . . 52
  - 3.2.2 Estimation in the Legendre design . . . 52
  - 3.2.3 Estimation in the Chebyshev design . . . 60
  - 3.2.4 Estimation for non-fixed classes . . . 65
- 3.3 Adaptive minimax regression . . . 67
  - 3.3.1 Adaptive estimation in functional scales . . . 67
  - 3.3.2 Upper bound on the quality of adaptive estimators . . . 69
  - 3.3.3 Lower bound . . . 76

4 Adaptive density estimation **79**
4.1 The model 79
4.2 Auxiliary results 81
4.3 Minimax density estimation in $\mathcal{A}(\gamma, \beta, r)$ 84
4.4 Adaptive density estimation 95
4.4.1 The adaptive setting 95
4.4.2 The adaptive estimator and the upper bound 97
4.4.3 Lower bound: optimality results 107

Appendix **113**

Bibliography **115**

Samenvatting **119**

Gracias **121**

Curriculum Vitae **123**

Chapter 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.

The model of statistical estimation starts from the assumption that we have a sample from an unknown probability measure $\mathbf{P}$ on a given measurable space. The probability measure $\mathbf{P}$ ranges over a class $\mathcal{P}$. The goal of the statistician is then to find a method for approximating the unknown probability $\mathbf{P}$ using the given sample. Usually the class of probability distributions is modeled in the parameterized form $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$ and then the task is reduced to estimating θ or a functional of θ. In the non-parametric problems of the kind treated in this thesis, the parameter set Θ is usually infinite-dimensional.

How 'good' a method of estimation is, usually depends on the class Θ. In the classical non-adaptive framework this class is assumed to be known. In practice, however, it is rarely known to the statistician and thus more realistic methods are necessary. By *adaptive methods of estimation* we refer to the data-driven methods of estimation that in a sense adapt to the uncertainty about the actual class Θ, e.g. methods that choose an estimator $\hat{\theta}$ from a sequence of candidates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$ which are the optimal estimators corresponding to different classes $\Theta_1, \Theta_2, \ldots, \Theta_n$.

In this thesis we study adaptive methods of estimation for two particular types of statistical problems: regression and density estimation. For all these problems the classes of probabilities $\mathbf{P}$ are parameterized by real-valued functions θ. In each model, the underlying function θ is assumed to belong to some class Θ of smooth functions. In practice the 'true' smoothness of the function θ is unknown and so the actual class Θ is also unknown. Thus, finding adaptive methods of estimation becomes very important in such problems.

We study different regression problems with fixed discrete designs: regression on the real line and regression on a bounded interval. Formally, the distinction here lies just in the definition of the underlying functional classes Θ. The construction of optimal adaptive procedures however is quite different in these cases. This is underlined by the essential difference between these two models; namely, in the case of regression models on bounded

observation intervals, the presence of the boundary – the so called boundary effect – has to be incorporated in the study of optimal statistical procedures.

For each of the three problems: regression on the real line, regression on bounded intervals and density estimation, we introduce corresponding scales of functional classes Θ for which exact – up to constants – rates of convergence can be obtained, under the classical minimax non-parametric framework, i.e. in the case when the classes Θ are known. We proceed then by constructing adaptive estimators and prove them to be asymptotically optimal, for the corresponding functional scales.

How well one can do at the boundary of a bounded observation interval depends, in turn, on the chosen design. We analyze two classical designs on a bounded interval: the equidistant and the Chebyshev designs. In both cases, the quality of estimation near the boundary becomes worse than inside the interval. However, while in the case of the equidistant design this affects the rate of convergence, no severe boundary losses occur in the case of the Chebyshev design. Therefore, in studying adaptive estimation procedures on a bounded interval, we restrict ourselves to the case of Chebyshev designs.

1.1 Brief history of adaptive estimation

The modern theory of adaptive methods of estimation was started by Pinsker and Efromovich in [1984]. They proved adaptivity up to constants for the problem of regression in the white noise model, with respect to the L^2-optimality criterion. Efromovich [1985] provided similar results for density estimation.

Since then, several methods of adaptive estimation for different problems have been proposed, primarily dealing with the L^2-norms; see Golubev and Nussbaum [1992], Dohono and Johnstone [1994], Birgé and Massart [1997]. Further literature on adaptive approach and its applications include: Donoho, Johnstone, Kerkyacharian and Picard [1996], Golubev and Levit [1996], Goldenshluger and Nemirovski [1997], Tsybakov [1998], Golubev, Lepski and Levit [2001], Cavalier [2001], Compte [2001].

Lepski [1990], [1991], [1992a], [1992b], was the first to address the problem of finding adaptive rates of convergence in a broader context, including different norms. He showed that for point-wise estimation on Hölder and Sobolev classes a loss of logarithmic factor in the rate of convergence is unavoidable in the Gaussian white noise regression models. He proposed a method of adaptive estimation achieving the optimal rates in the adaptive setting which since then is commonly referred to as "Lepski's method".

Advancing these ideas still further, Lepski and Spokoiny [1997] elaborating on Yuditsky [1997], proposed a refinement of the Lepski's original idea, which for some Hölder classes, made it possible to achieve adaptive rates, including the exact constants. A comprehensive discussion of different approaches to non-parametric adaptive estimation and other related work can be found in Lepski, Mammen and Spokoiny [1997].

It turns out that much more complete solutions can be obtained when the underlying functional classes comprise entire, analytic or infinitely differentiable functions. Note that such functional classes are just as good for modeling the real world as classes of finitely

smooth functions, since, in both cases, such functional classes are everywhere dense among all continuous functions. Corresponding classes of analytic functions were first introduced in the statistical theory in Ibragimov and Has'minskii [1983] where optimal rates of convergence were found in estimating analytic density functions.

Later, Golubev and Levit [1996] showed that the class of analytic functions is 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 such classes was studied in Golubev, Levit and Tsybakov [1996].

Lepski and Levit [1998] introduced larger functional scales of infinitely differentiable functions in the white noise setting. They proposed asymptotically minimax estimators, for all of the corresponding functional classes, and, moreover, solved the problem of adaptive estimation for them, by properly modifying the adaptive procedure of Lepski and Spokoiny [1997], originally proposed for the finite smoothness functional classes. The adaptive estimation procedure proposed in Lepski and Levit [1998] was proved there to be asymptotically optimal, under most general assumptions. This paper was the primary motivation for the present work which, for functional scales close to the classes of infinitely differentiable functions they considered, carries their approach over in Chapters 2 and 3 to corresponding discrete Gaussian white noise models, and in Chapter 4 to the density estimation problem.

Both types of models we study – discretized regression models and the density estimation – can be approximated by a white noise model. This topic has been recently a subject of thorough investigation; see Nussbaum [1996], Brown and Low [1996]. The existing results, however, are primarily concerned with finite smoothness classes and, to the best of our knowledge, do not cover problems of adaptive estimation.

Our results can also be viewed from this perspective: although, strictly speaking, we do not establish equivalence relations between the above models, we do, in essence, investigate when procedures similar to those developed for the white noise model are also optimal for other models mentioned above. Such results can be interpreted as a "weak" equivalence between different models. They are useful from a practical point of view and serve as a good indication of which full type equivalence results for such models one might expect in the future.

Our study incorporates adaptive estimation for analytic functions observed on a bounded interval. For the first time such classes have been introduced in the white noise model in Ibragimov and Has'minskii [1984] where a point-wise asymptotically minimax estimator was proposed based on Legendre polynomials. Here we study the same type of functions as in Ibragimov and Has'minskii [1984] but in a more realistic discrete design model. This leads us to the interesting problem of choosing designs which can enhance the quality of estimation at the end-points of the observation interval. The problem of adaptive estimation for such designs is also studied.

In our last chapter we consider the problem of adaptive density estimation, for a scale of functional classes of infinitely differentiable densities. Adaptive estimation of density functions belonging to the scale of Sobolev classes was recently studied in Butucea [1999].

The number of publications in the area of adaptive estimation is growing very fast and there are many different approaches. Most of the references study just the optimal adaptive rates of convergence. The optimal constants for pointwise estimation has been made possible by introducing classes of infinitely differentiable functions, as in Lepski and Levit [1998] and [1999]. The goal of this thesis is to bring this idea further in some other classical statistical problems.

1.2 Scope of the thesis

1.2.1 Adaptive regression on the real line

In Chapter 2 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

$$y_\ell = f(\ell h) + \xi_\ell, \qquad \ell = 0, \pm 1, \pm 2, \dots \tag{1.1}$$

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

$$dV(x) = f(x)\, dx + \epsilon\, dW(x), \quad -\infty < x < \infty, \tag{1.2}$$

which is just an approximation to the model (1.1), 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 – 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 (1.2). Below we will explain main differences between our approach and that of Lepski and Levit [1998].

Classes of functions are in general described by smoothness parameters. In this chapter 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 (1.1), 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

$$\|f\|^2_{\gamma,\beta,r} := \int_{-\infty}^{\infty} \frac{\gamma}{\beta^2}\, e^{2|\gamma t|^r} |\mathcal{F}[f](t)|^2 dt \le 1. \tag{1.3}$$

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], [1984], Stone [1982], Ibragimov [2001]. 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 effects 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 (1.3). 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 $\mathcal{A}'(\gamma, \gamma, r) = \{f \text{ continuous}, |\mathcal{F}[f](t)| \leq \gamma \exp -(\gamma t)^r\}$ 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 (1.1).

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 off-beat 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: $\sigma, h, \gamma, \beta, r$. Since the white noise model (1.2) is known to approximate the discrete regression model (1.1), 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 effect 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 (1.2) captures the underlying properties of the discrete model (1.1). 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.

In Sect. 2.4 we study an adaptive procedure closely related to the one proposed in Lepski and Levit [1998]. Our aim is to describe precisely the conditions under which this procedure is asymptotically optimal. Using the above terminology, one could say that these assumptions describe the parameter limitations under which both regression models: the discrete model (1.1) and the continuous time model, are weakly equivalent.

In Section 2.3, 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 2.4 we consider the functional scales which are collections of the form

$$\Big\{ \mathcal{A}(\gamma, \beta, r) \;\Big|\; (\gamma, \beta, r) \in \mathcal{K} \subset \mathbb{R}^3_+ \Big\},$$

where $\mathcal{K}$ is the scale of the parameters. 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. Compared to a given 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 2.3.

1.2.2 Adaptive regression on a bounded interval

In Chapter 3 we study adaptive non-parametric regression models with a fixed design in the case when the unknown regression function f is analytic in a vicinity $\mathcal{V}$ of the observation interval $[-1, 1]$ of the complex plane. The smoothness of f is then characterized by the size of $\mathcal{V}$ and $\max_{\mathcal{V}} |f|$. A more concise description of such dependence – and more accurate results – become feasible when $\mathcal{V}$ is the region E_γ with boundary

$$\partial E_\gamma = \{z : \ z = \cosh\gamma \cos\phi + i \sinh\gamma \sin\phi, \ 0 \le \phi \le 2\pi\}.$$

This boundary set is the ellipse with foci at the end points of the interval $[-1, 1]$ and the sum of its semi-axes equal to $\exp\gamma$. The family of such elliptic areas is natural in the sense that $\cap E_\gamma = [-1, 1]$ and $\cup E_\gamma = \mathbb{C}$. Note, without loss of generality we have assumed that the regression interval is $[-1, 1]$, but an obvious generalization can be made to any real interval $[a, b]$.

We will denote by $\mathcal{A}(\gamma, M)$ the set of functions which are analytic and bounded in E_γ with $|f(z)| \le M$ in that region. For functions $f \in \mathcal{A}(\gamma, M)$ observed in the continuous-time Gaussian white noise on the interval $[-1, 1]$, Ibragimov and Has'minskii [1982] have demonstrated point-wise asymptotically minimax estimators based on Legendre polynomials.

We consider the problem of discrete regression in the model

$$y_k = f(x_k^n) + \xi_k, \quad k = 1, \dots, n, \tag{1.4}$$

where the points x_k^n form the design knots and the ξ_k are independent identically distributed Gaussian random variables, with zero mean and given variance σ^2. Given the observations $\mathbf{y} = (y_1, y_2, \dots, y_n)$, the function $f \in \mathcal{A}(\gamma, M)$ can be estimated by the projection-type estimators

$$\hat{f}_{n,N}(x, \mathbf{y}) = \sum_{r=0}^{N-1} \hat{c}_r Q_r(x), \qquad \hat{c}_r = \frac{1}{n} \sum_{k=1}^{n} y_k Q_r(x_k^n),$$

where Q_r are polynomials orthonormal over the design points x_k^n. This method is easier to implement and to study. For instance, if we consider the design of equally spaced knots

$$x_k^n = \frac{2k - n - 1}{n}, \quad k = 1, \dots, n, \tag{1.5}$$

one could use the so called Chebyshev discrete polynomials $C_r(x), r = 0, 1, ...,$ (cf. Bateman [1953], Sect. 10.23, p. 223). However, for this design we will find it more convenient to use a family $p_r(x), r = 0, 1, \dots,$ of normalized Legendre polynomials which are asymptotically equivalent to C_r (cf. Bateman, Sect. 10.23, eq. 7). In particular, the normalized Legendre polynomials $p_r(x)$ are asymptotically orthonormal over the design knots (1.5). Thus we

shall refer to these knots as the *Legendre knots* or *equidistant knots*, and to the set of these knots as the *Legendre design* or *equidistant design*.

As intuition suggests, when we use the equidistant design to estimate the unknown regression function at points which are close to the border of the interval, less information is gathered than when we are interested in estimation inside the interval. Although it might seem that the number of observations available at the end-points is just halved, we shall see that in fact the accuracy of estimation near the border becomes worse by a factor of order $\sqrt{\log n}$, compared to the accuracy obtained inside the interval.

This situation can be improved by using another – non-uniform – design which will balance the distribution of the design points, in favor of increasing the accuracy of the estimation at the end-points. A special classical design having this property is specified by the knots

$$x_k^n = \cos\frac{(2k-1)\pi}{2n}, \quad k = 1, \ldots, n, \tag{1.6}$$

which we will conveniently call the *Chebyshev knots* and the corresponding design the *Chebyshev design*. Remarkably, the classical orthornormal Chebyshev polynomials $t_r(x)$, $r = 0, 1, ...$, are also orthonormal over the Chebyshev design. As we shall see later, with this polynomials the same rate of convergence is achieved inside the interval, whilst at the end-points the rate is only a factor 2 slower. [1]

Given that for the equidistant design we use the Legendre polynomials and for the Chebyshev design we use the Chebyshev polynomials, one can question whether the difference in rates of convergence is due to the particular method of estimation we are studying or indeed is property of the design itself. To clarify this, we will demonstrate that in each of the corresponding designs, our estimators are asymptotically optimal among all possible estimators, at every point of the interval $[-1, 1]$. This leads us to the conclusion that the observed difference in the rates of convergence near the end-points is a direct consequence of the use of the equidistant design and that this problem does not present itself in the case of the Chebyshev design.

Several remarkable properties of the functions $p_r(x)$ and $t_r(x)$ make this approach attractive for practical purposes. The normalized Legendre polynomials $p_r(x)$ are asymptotically orthonormal over the equally spaced knots while the normalized Chebyshev polynomials $t_r(x)$ are orthonormal over the Chebyshev knots. This makes the evaluation of the projection polynomials straightforward. At the same time the orthonormality property allows an easy evaluation of the variance of the corresponding regression estimators in the statistical framework (1.4). In the case of known classes $\mathcal{A}(\gamma, M)$, the variance can be easily balanced against the systematic error, thus determining the optimal number N of polynomials in use.

A property that will play a major role in the application of Legendre or Chebyshev

[1] Motivated by this study B. Levit (2001) introduced a general theory of Optimal Designs in Nonparametric Regression.

design for estimation is the behavior of the functions

$$\frac{1}{N}\sum_{r=0}^{N-1} t_r^2(x) \quad \text{and} \quad \frac{1}{N}\sum_{r=0}^{N-1} p_r^2(x)$$

both inside the interval $[-1,1]$ and near the end-points (see Lemmas 3.1 and 3.2). These terms appear as variances of the corresponding estimators and to a great extent are important in shaping the results (see Theorems 3.1 and 3.2).

The structure of Chapter 3 is as follows. In Section 3.1 we introduce the functional classes $\mathcal{A}(\gamma, M)$ and discuss the Legendre and Chebyshev polynomials. In Section 3.2 we then describe the asymptotically minimax estimators $\hat{f}_n$, $n \to \infty$, are described in the case when the unknown regression function f belongs to a given fixed class $\mathcal{A}(\gamma, M)$, using Legendre and Chebyshev polynomials for corresponding designs. In both cases the polynomial estimates we consider are shown to be point-wise asymptotically efficient, for their corresponding designs. In Section 3.3 we introduce functional scales and construct asymptotically optimal adaptive estimators, under the assumption that the parameters γ and M are unknown.

1.2.3 Adaptive density estimation

In chapter 4 we study adaptive estimation of an unknown probability density function f, based on the sample of n independent identically distributed observations $X_1, \ldots, X_n$, with common density f. It is assumed that f belongs to the scale of functional classes $\mathcal{A}(\gamma, \beta, r)$ which was introduced before.

The study of asymptotically minimax estimators for the functional classes $\mathcal{A}(\gamma, \beta, r)$ is of recent origin. Golubev and Levit [1996] found exact asymptotics for the minimax point-wise estimation of an unknown density belonging to a class of functions analytic in the strip $\{z = x + iy : |y| < \gamma\}$ of the complex plane (this corresponds to our class $\mathcal{A}(\gamma, \beta, 1)$). They described asymptotically unbiased and asymptotically efficient estimators with a rate of convergence only slightly worse than $\sqrt{n}$.

Schipper [1996] studied asymptotically minimax estimators for this same class of densities, under the L^2-norm. Belitser [1997] was successful in studying more general classes $\mathcal{A}(\gamma, \beta, r)$, $0 < r \leq 1$, in a more difficult setting of density estimation under random censorship.

In general, the choice of the estimator depends on parameters γ, β, r regulating the degree of smoothness, which however are rarely known in practice. The use of adaptive methods of estimation at this stage appears to be an unavoidable evil.

The main goal of Chapter 4 is to generalize the approach proposed for the Gaussian white noise model in Lepski and Levit [1998], to estimating the unknown density function $f \in \mathcal{A}(\gamma, \beta, r)$ when the vector of parameters (γ, β, r) of the scale vary in a subset $\mathcal{K} \subset \mathbb{R}^3_+$. Adaptive estimation of a density function belonging to the scale of Sobolev classes has been recently studied in Butucea [1999] who used methods more closely related to

Tsybakov [1998]. Although there are some similarities between our study of the functional scales $\mathcal{A}(\gamma, \beta, r)$ and Butucea's work, the technical tools used are quite different.

In general, the problem of density estimation exhibits some similarities and, at the same time, a few important differences when compared to the Gaussian white noise regression model

$$dV(x) = f(x)\,dx + \frac{1}{\sqrt{n}} dW(x), \quad -\infty < x < \infty, \tag{1.7}$$

where V is the noisy observation of an unknown regression function f and $W(x)$ the Wiener process. These differences affect both the sample study of a chosen estimation procedure, and the proof of its optimality; that is, they affect both, deriving upper and lower bounds for the minimax risk. To demonstrate the difference of the first kind, let us assume that a kernel estimator is used to estimate $f(x)$ in both problems:

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} k(x - X_i)$$

in the density estimation problem, and

$$\bar{f}_n(x) = \int k(x - y)\,dV(y)$$

in the model (1.7). Assuming, as usual, that k is a δ-type kernel function, one obtains the familiar variance-bias decomposition for the mean square error

$$\mathbf{E}\left(f_n(x) - f(x)\right)^2 = \frac{1}{n}\,\mathbf{Var}\,k(x - X_1) + \left(\mathbf{E}k(x - X_1) - f(x)\right)^2 \approx$$

$$\frac{1}{n}\int k^2(x - y) f(y)\,dy + \left(\int k(x - y) f(y)\,dy - f(x)\right)^2 \approx$$

$$\frac{f(x)}{n}\int k^2(y)\,dy + \left(\int k(x - y) f(y)\,dy - f(x)\right)^2. \tag{1.8}$$

Similarly, using the well-known Itô isometry formula, one obtains

$$\mathbf{E}\left(\bar{f}_n(x) - f(x)\right)^2 = \mathbf{E}\left(\int k(x - y) f(y)\,dy + \frac{1}{\sqrt{n}}\int k(x - y)\,dW(y) - f(x)\right)^2 =$$

$$\frac{1}{n}\mathbf{E}\left(\int k(x - y)\,dW(y)\right)^2 + \left(\int k(x - y) f(y)\,dy - f(x)\right)^2 =$$

$$\frac{1}{n}\int k^2(y)\,dy + \left(\int k(x - y) f(y)\,dy - f(x)\right)^2. \tag{1.9}$$

Next, for our classes $\mathcal{A}(\gamma, \beta, r)$ the kernel k can be chosen in such a way, that the bias becomes negligible when compared to the variance:

$$\mathbf{E}\left(f_n(x) - f(x)\right)^2 \sim \frac{f(x)}{n}\int k^2(y)\,dy, \tag{1.10}$$

and

$$\mathbf{E}\left(\bar{f}_n(x) - f(x)\right)^2 \sim \frac{1}{n}\int k^2(y)\,dy. \tag{1.11}$$

Therefore, the mean-square errors in both problems are quite similar, and the effect of the additional factor $f(x)$ in (1.8) is not significant when the estimated density belongs to a *known* class $\mathcal{A}(\gamma, \beta, r)$, (cf. Section 4.1), since for such classes the optimal choice of the kernel k, minimizing (1.8), does not depend on the factor $f(x)$, up to a first order.

The situation changes dramatically in the adaptive setting. There, a choice has to be made between estimators, say f_{n1} and f_{n2}, related to different kernels k_1 and k_2 which correspond to different possible classes $\mathcal{A}(\gamma, \beta, r)$. Then we have similarly to (1.8)-(1.9):

$$\mathbf{E}\left(f_{n1}(x) - f_{n2}(x)\right)^2 \approx \frac{f(x)}{n}\int (k_1(y) - k_2(y))^2\,dy + \left(\int (k_1(x-y) - k_2(x-y))f(y)\,dy\right)^2,$$

$$\mathbf{E}\left(\bar{f}_{n1}(x) - \bar{f}_{n2}(x)\right)^2 \approx \frac{1}{n}\int (k_1(y) - k_2(y))^2\,dy + \left(\int (k_1(x-y) - k_2(x-y))f(y)\,dy\right)^2.$$

Here again, while the last two terms are exactly the same, in the density estimation problem and in the white noise model, an additional complication appears in the first term for the density case. However, in order to make an accurate choice between the two estimators f_{n1} and f_{n2}, a preliminary estimator of the factor $f(x)$ appearing in the density case, becomes important.

This complication can be solved in at least two different ways. One could estimate $\mathbf{E}\left(f_{n1}(x) - f_{n2}(x)\right)^2$ using the sample estimator

$$\frac{1}{n}\sum_{i=1}^{n}\left(k_1(x - X_i) - k_2(x - X_i)\right)^2. \tag{1.12}$$

Another possibility is to use, as an estimator of $f(x)$, one of the two estimators f_{n1} and f_{n2} themselves. It turns out that this second approach is easier to analyze although in some cases (1.12) may be better in practical implementations. In Chapter 4 we use this approach and develop the techniques necessary for analyzing the corresponding adaptive estimator. Similar approach, using a preliminary estimator, has been used in Butucea [1999]. Note that this complication, not present in Lepski and Levit's analysis of the white noise model, puts some additional restrictions on the density estimation model and on the final results.

Since the asymptotic variance in the density estimation problem depends on the unknown density $f(x)$, cf. (1.10), in such problems it is more accurate to use the concept of locally asymptotically minimax estimators. For comparison, since in the white noise model the asymptotic variance does not depend on $f(x)$, cf. (1.11), the use of globally asymptotically minimax estimators in such models is more justifiable; cf. Lepski and Levit. The method of proving locally asymptotically minimax lower bounds in non-parametric problems, was first proposed by B. Levit as early as 1974, Levit [1974]. The idea, which

since then was frequently used in non-parametric estimation, is to consider for an arbitrary density $f_0 \in \mathcal{A}(\gamma, \beta, r)$, an fixed x, a parametric sub-family

$$f_\theta(y) = f_0(y)\left(1 + \theta\frac{k(x-y) - \mathbf{E}_{f_0}\, k(x - X_1)}{\sqrt{\mathbf{Var}_{f_0}\, k(x - X_1)}}\right).$$

Here k is the same δ-type kernel function, as has been used in constructing the above estimators.

To obtain a lower bound for estimating $f(x)$, using this approach, one has to establish, first, that the resulting family is locally asymptotically normal, i.e. it behaves asymptotically similarly to a normal density with a shift parameter θ (cf. our Definition 4.1 below on p. 90), and second, that this parametric family belongs to the same class $\mathcal{A}(\gamma, \beta, r)$, at least for all sufficiently small θ. The first of these requirements is rather standard (cf. our Lemma 4.5 on p. 90).

The question of whether f_θ belongs to $\mathcal{A}(\gamma, \beta, r)$, however, is more complicated, since, in general, the properties $f_0 \in \mathcal{A}(\gamma, \beta, r)$ and $k \in \mathcal{A}(\gamma, \beta, r)$ do not imply that $f_0 k \in \mathcal{A}(\gamma, \beta, r)$, although as Belitser [1997] has demonstrated this is the case, under the additional assumption $r \leq 1$. In Section 4.3 we propose the following solution. Let $f \in \mathcal{A}(\gamma, \beta, r)$ and let $\mathcal{V}$ be an arbitrary vicinity of f in the class $\mathcal{A}(\gamma, \beta, r)$, described by a corresponding norm $\|f\|_{\gamma,\beta,r}$. First of all, we demonstrate that in any such vicinity $\mathcal{V}$, there exists a density f_0 which belongs to the class of the so called entire functions of an exponential type; equivalently, f_0 has a finitely supported Fourier transform $\mathcal{F}[f_0](t)$. Now, one can use the density function f_0 instead of f to construct the family f_θ. It can be proved then that the resulting family $f_\theta \in \mathcal{A}(\gamma, \beta, r)$ (see Lemma 4.5), leading to the required lower bound.

In Sections 4.1–4.3 we recall the model and collect the minimax results for known functional classes. In Section 4.4 we introduce the adaptive framework, construct the adaptive estimator, and prove its asymptotic optimality.

Chapter 2

Adaptive regression on the real line

In this chapter we discuss a non-parametric regression model in which the function $f(x)$ is observed under additive Gaussian perturbations, on an equidistant grid of the whole real line. This observation model is a discrete version of the white Gaussian noise model; see e.g. Lepski and Levit [1998].

The unknown function $f(x)$ is assumed to be infinitely differentiable. More precisely, $f(x)$ is assumed to belong to some non-parametric functional class $\mathcal{A}(\gamma, \beta, r)$, known *a priori* to belong to a family of such classes, parametrized by positive parameters γ, β, r. Such assumption provides a natural way of covering different types of smoothness, when one considers different values of the parameters involved. The functions of the class $\mathcal{A}(\gamma, \beta, r)$ are infinitely differentiable for all values of the parameter r. They admit, for $r = 1$, an analytic continuation into the strip of the complex plane of size 2γ symmetric around the real line, and are entire functions for $r > 1$. In practice the parameters of the class are unknown, thus leading to an interesting and important problem of finding adaptive methods of estimation, independent of the prior knowledge of the parameters γ, β, r.

Such functional classes were first introduced (for $r = 1$) in statistics 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 Section 2.1 we introduce the model. In Section 2.2 we prove some auxiliary lemmas. In Section 2.3 we study the problem of asymptotic minimax regression, both for fixed classes $\mathcal{A}(\gamma, \beta, r)$ and in the case of parameters which may depend on the level of the noise σ^2 or/and the discretization parameter h. Finally, in Section 2.4 we consider the adaptive case in which the unknown function f may belong to the whole functional scale of classes $\mathcal{A}(\gamma, \beta, r)$ and prove asymptotic optimality of the proposed adaptive estimator.

2.1 The model

For positive γ, β and r we study the problem of estimating a regression function in the class $\mathcal{A}(\gamma, \beta, r)$.

Definition 2.1 *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 transform $\mathcal{F}[f]$ satisfy*

$$\|f\|_{\gamma,\beta,r} := \int \frac{\gamma}{\beta^2}\, e^{2|\gamma t|^r} \left|\mathcal{F}[f](t)\right|^2 dt \le 1. \tag{2.1}$$

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

$$\mathcal{F}[f](t) = \int e^{\mathbf{i}tx} f(x)\, dx. \tag{2.2}$$

Note that the Fourier inversion formula

$$f(x) = \frac{1}{2\pi} \int e^{-\mathbf{i}tx} \mathcal{F}[f](t)\, dt \tag{2.3}$$

certainly holds under assumption (2.1). It is easy to see that for all $\gamma, \beta, r > 0$, functions in $\mathcal{A}(\gamma, \beta, r)$ are infinitely differentiable.

In the definition of the classes $\mathcal{A}(\gamma, \beta, r)$ each parameter plays a certain role. For instance, the parameter γ plays role of a scale parameter, since it is easy to see that

$$f(\cdot) \in \mathcal{A}(1, \beta, r) \quad \Longleftrightarrow \quad \tfrac{1}{\gamma} f\left(\tfrac{\cdot}{\gamma}\right) \in \mathcal{A}(\gamma, \beta, r).$$

Thus, γ affects the smoothness in a most stringent way. Note that the bigger is γ, the smoother are the functions of the class. The parameter β can be interpreted as the radius

of the corresponding L^2-ellipsoid defined by (2.1). Finally it is known that for $r < 1$ the functions in the class $\mathcal{A}(\gamma, \beta, r)$ are infinitely differentiable, while for $r = 1$ they admit an analytic continuation in the strip of the complex plane $\{z = x + iy : |y| < \gamma\}$. Consequently, for $r > 1$ the functions of this class are entire functions, i.e. they admit an analytic continuation into the whole complex plane. Similar classes of functions with $r = 1$ have been considered in Ibragimov and Has'minskii [1981] and Golubev and Levit [1996]. Another property, namely that the functions of $\mathcal{A}(\gamma, \beta, r)$ and their derivatives are globally bounded in terms of the parameters of the class can be found in Ch. 4, Lemma 4.1 below.

Now, let us consider the following observation model

$$y_\ell = f(\ell h) + \xi_\ell, \qquad \ell = 0, \pm 1, \pm 2, \ldots, \tag{2.4}$$

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

$$\hat{f}_{h,s}(x, \mathbf{y}) = h \sum_{\ell=-\infty}^{\infty} k_s(x - \ell h)\, y_\ell \tag{2.5}$$

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

$$k_s(x) = \frac{\sin sx}{\pi x}, \tag{2.6}$$

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

$$\mathcal{F}[k_s](t) = \mathbb{1}_{[-s,s]}(t) \tag{2.7}$$

and therefore, according to the convolution theorem,

$$\mathcal{F}[f * k_s](t) = \mathbb{1}_{[-s,s]}(t)\, \mathcal{F}[f](t), \tag{2.8}$$

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 Sect. 2.3, 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.

Our results in this chapter will refer to the following class of loss functions. 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

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

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*

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

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

2.2 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 defined by (2.1) and $k_s(x)$ is the kernel (2.6).

Lemma 2.1 *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*

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

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

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

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

(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* $\gamma, \beta, r, s_i \ge 0, i = 1, 2,$ *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, *for* $h \to 0$,

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

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

$$\sum_{\ell=-\infty}^{\infty} e^{2\pi\mathbf{i}\,\ell x} = \sum_{\ell=-\infty}^{\infty}\delta(x-\ell), \tag{2.9}$$

known in the theory of distributions (cf. e.g. Antonsik et al. [1973], Ch. 9.6). Using the Fourier inversion formula, the distributional formula (2.9) and with some algebra, one obtains

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

[2] Notice that if we take $s_1 = 0$ and $s_2 = s$ then $\Delta(x) = k_s(x)$.

$$
\begin{aligned}
&= \frac{h}{(2\pi)^2}\int\int e^{-\mathbf{i}tx}\,\mathcal{F}[g](t)\,e^{\mathbf{i}sy}\mathcal{F}[f](s)\sum_{\ell=-\infty}^{\infty} e^{-\mathbf{i}(s-t)\ell h}\,dt\,ds \\
&= \frac{h}{(2\pi)^2}\sum_{\ell=-\infty}^{\infty}\int\int e^{-\mathbf{i}tx}\mathcal{F}[g](t)\,e^{\mathbf{i}sy}\mathcal{F}[f](s)\,\delta\left(\frac{h(s-t)}{2\pi}-\ell\right)dt\,ds \\
&= \frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty}\int e^{-\mathbf{i}tx}\mathcal{F}[g](t)\int e^{\mathbf{i}sy}\mathcal{F}[f](s)\,\delta\left(s-t-\frac{2\pi\ell}{h}\right)ds\,dt \\
&= \frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty}\int e^{-\mathbf{i}tx}\mathcal{F}[g](t)\,e^{\mathbf{i}(t+\frac{2\pi\ell}{h})y}\mathcal{F}[f]\left(t+\frac{2\pi\ell}{h}\right)dt \\
&= \frac{1}{2\pi}\int e^{-\mathbf{i}t(x-y)}\,\mathcal{F}[g](t)\,\mathcal{F}[f](t)\,dt \\
&\qquad\qquad + \frac{1}{2\pi}\sum_{\ell\neq 0} e^{\mathbf{i}\frac{2\pi\ell}{h}y}\int e^{-\mathbf{i}t(x-y)}\,\mathcal{F}[g](t)\,\mathcal{F}[f]\left(t+\frac{2\pi\ell}{h}\right)dt \\
&= \int_{-\infty}^{\infty} g(x-z)f(z-y)\,dz \\
&\qquad\qquad + \frac{1}{2\pi}\sum_{\ell\neq 0} e^{\mathbf{i}\frac{2\pi\ell}{h}y}\int e^{-\mathbf{i}t(x-y)}\,\mathcal{F}[g](t)\,\mathcal{F}[f]\left(t+\frac{2\pi\ell}{h}\right)dt.
\end{aligned}
$$

(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 (2.1) and the Cauchy-Schwartz inequality. Thus we can apply the previous result in (a), using $g=\Delta$ and $y=0$. Notice 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

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

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

$$
\begin{aligned}
&\leq \frac{1}{2\pi}\sum_{\ell\neq 0}\left(\int \mathbb{1}_{(s_1,s_2]}(|t|)\,\frac{\beta^2}{\gamma}\,e^{2|\gamma t|^r}e^{-2|\frac{2\pi\ell\gamma}{h}|^r/c_r}dt\right)^{1/2}\\
&\leq \frac{1}{2\pi}\sum_{\ell\neq 0}e^{-|\frac{2\pi\ell\gamma}{h}|^r/c_r}\left(2\int \mathbb{1}_{(s_1,s_2]}(t)\,\frac{\beta^2}{\gamma}\,e^{2(\gamma t)^r}dt\right)^{1/2}\\
&= \frac{1}{\pi}\left(2\int_{s_1}^{s_2}\frac{\beta^2}{\gamma}\,e^{2(\gamma t)^r}dt\right)^{1/2}\sum_{\ell=1}^{\infty}e^{-(2\pi\ell\frac{\gamma}{h})^r/c_r}\\
&\leq \frac{1}{\pi}\left(2\int_{s_1}^{s_2}\frac{\beta^2}{\gamma}\,e^{2(\gamma t)^r}dt\right)^{1/2}\left(e^{-(2\pi\frac{\gamma}{h})^r/c_r}+\int_1^{\infty}e^{-(2\pi\frac{\gamma}{h}x)^r/c_r}dx\right)\\
&= O\left(e^{-(2\pi\frac{\gamma}{h})^r/c_r}\right)\left(\int_{s_1}^{s_2}\frac{\beta^2}{\gamma}\,e^{2(\gamma t)^r}\,dt\right)^{1/2}, \qquad (h\to 0),
\end{aligned}
$$

where the last asymptotic can be easily derived by partial integration (cf. Lemma 2.2, eq. (2.10)).

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

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

Therefore

$$
\begin{aligned}
\left|h\sum_{\ell=-\infty}^{\infty}\Delta^2(x-\ell h)-\frac{s_j-s_i}{\pi}\right| &\leq \frac{1}{2\pi}\sum_{\ell\neq 0}\int\mathcal{F}[\Delta](t)\,\mathcal{F}[\Delta]\left(t+\frac{2\pi\ell}{h}\right)dt\\
&\leq \frac{1}{\pi}\sum_{\ell=1}^{\infty}\int\mathbb{1}_{(s_1,s_2]}(|t|)\,\mathbb{1}_{(s_1,s_2]}\left(\left|t+\frac{2\pi\ell}{h}\right|\right)dt\\
&\leq \frac{5h(s_2-s_1)^2}{2\pi^2}=O_h(1)\,h(s_2-s_1)^2,
\end{aligned}
$$

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 2.2 *For any positive γ and r the following inequality holds*

$$\int_s^\infty e^{-2(\gamma t)^r}\,dt \le \frac{s\,e^{-2(\gamma s)^r}}{r(\gamma s)^r} \tag{2.10}$$

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

$$\int_0^s e^{2(\gamma t)^r}\,dt = \frac{s\,e^{2(\gamma s)^r}}{2r(\gamma s)^r}(1+o(1)) \tag{2.11}$$

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.

2.3 Minimax regression in $\mathcal{A}(\gamma, \beta, r)$

2.3.1 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 2.1 *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 (2.5) with the bandwidth*

$$s_h = s_h(\alpha) = \frac{1}{\gamma}\left(\frac{1}{2}\log\frac{\beta^2}{\pi\,\gamma\sigma^2 h}\right)^{1/r}, \tag{2.12}$$

satisfies

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

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

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 (2.4) and the formula for the estimator (2.5) one can split the error term as follows,

$$\begin{aligned}\hat{f}_{h,s}(x) - f(x) &= \Big(h\sum_{\ell=-\infty}^{\infty} k_s(x-\ell h)f(\ell h) - f(x)\Big) + \Big(h\sum_{\ell=-\infty}^{\infty} k_s(x-\ell h)\xi_\ell\Big) \\ &= b(f,x,s,h) + v(\sigma,x,s,h).\end{aligned}$$

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

$$\mathbf{E}\left(\hat{f}_{h,s}(x) - f(x)\right)^2 = b_s^2 + \mathbf{Var}\, v_s, \tag{2.13}$$

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 2.1 we take $s_1 = 0$ and $s_2 = s$. In this case $\Delta = k_s$. Now, applying Lemma 2.1(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 (2.7), and definition of the class $\mathcal{A}(\gamma,\beta,r)$ we get

$$\begin{aligned}b_s^2 &\le 2\left|\frac{1}{2\pi}\int e^{-\mathbf{i}tx}(\mathcal{F}[k_s](t) - 1)\mathcal{F}[f](t)\,dt\right|^2 + O\left(e^{-2(2\pi\frac{\gamma}{h})^r/c_r}\right)\int_0^s \frac{\beta^2}{\gamma}e^{2(\gamma t)^r}\,dt \\ &\le \frac{1}{2\pi^2}\int_{|t|>s}\frac{\beta^2}{\gamma}e^{-2|\gamma t|^r}dt + O\left(e^{-2(2\pi\frac{\gamma}{h})^r/c_r}\right)\int_0^s \frac{\beta^2}{\gamma}e^{2(\gamma t)^r}\,dt \\ &\le \frac{1}{\pi^2}\int_s^\infty \frac{\beta^2}{\gamma}e^{-2(\gamma t)^r}dt + O\left(e^{-2(2\pi\frac{\gamma}{h})^r/c_r}\right)\int_0^s \frac{\beta^2}{\gamma}e^{2(\gamma t)^r}\,dt. \end{aligned}\tag{2.14}$$

Second, let us consider the variance term. From Lemma 2.1(c), with $s_1 = 0$ and $s_2 = s$, we see that

$$\mathbf{Var}\, v_s = \sigma^2 h^2\sum_{\ell=-\infty}^{\infty} k_s^2(x-\ell h) = \frac{\sigma^2 h\, s}{\pi}\big(1 + O_h(1)\, hs\big), \tag{2.15}$$

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

$$\sigma_{h,s}^2 = \frac{\sigma^2 hs}{\pi}, \tag{2.16}$$

so that, with the chosen bandwidth $s = s_h$, the resulting variance becomes

$$\sigma_h^2 = \sigma_h^2(\alpha) = \frac{\sigma^2 h s_h}{\pi}. \tag{2.17}$$

From equations (2.13)–(2.15) we see that the mean square error of the estimator $\hat{f}_{h,s}$ satisfies

$$\begin{aligned}\left|\mathbf{E}\left(\hat{f}_{h,s}(x)-f(x)\right)^2-\sigma_{h,s}^2\right| &\leq \sigma_{h,s}^2\Big(O(hs)+(\pi\sigma_{h,s})^{-2}\int_s^\infty \frac{\beta^2}{\gamma}e^{-2(\gamma t)^r}dt \\ &\quad +\sigma_{h,s}^{-2}O\left(e^{-2(2\pi\frac{\gamma}{h})^r/c_r}\right)\int_0^s \frac{\beta^2}{\gamma}e^{2(\gamma t)^r}\,dt\Big). \qquad (2.18)\end{aligned}$$

Now we shall verify that, taking $s=s_h$ as defined in (2.12), 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 finds a compromise between 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), this is because by (2.12)

$$e^{2(\gamma s_h)^r}=\frac{\beta^2}{\pi\gamma\sigma^2 h}. \qquad (2.19)$$

Let us return to equation (2.18). First notice that

$$hs_h\to 0, \quad \text{when} \quad h\to 0. \qquad (2.20)$$

Second, applying the identity (2.19) and Lemma 2.2, we see that

$$\begin{aligned}(\pi\sigma_h)^{-2}\int_{s_h}^\infty \frac{\beta^2}{\gamma}e^{-2(\gamma t)^r}dt &= \frac{\beta^2}{\pi\,\gamma\sigma^2 h}\frac{\int_{s_h}^\infty e^{-2(\gamma t)^r}dt}{s_h} = \frac{\int_{s_h}^\infty e^{-2(\gamma t)^r}dt}{s_h e^{-2(\gamma s^h)^r}} \\ &\leq \frac{1}{r(\gamma s_h)^r} = \left(\frac{r}{2}\log\frac{\beta^2}{\pi\gamma\,\sigma^2 h}\right)^{-1} = o(1), \qquad (2.21)\end{aligned}$$

when $h\to 0$. Finally, applying the identity (2.19) and the maximum inequality

$$\begin{aligned}\sigma_h^{-2}e^{-2(\frac{2\pi\gamma}{h})^r/c_r}\int_0^{s_h}\frac{\beta^2}{\gamma}\,e^{2(\gamma t)^r}\,dt &\leq \pi\frac{\beta^2}{\gamma\,\sigma^2 h}\,e^{-2(\frac{2\pi\gamma}{h})^r/c_r+2(\gamma s_h)^r} \\ &= \left(\frac{\beta^2}{\gamma\,\sigma^2 h}\right)^2 e^{-2\left(2\pi\frac{\gamma}{h}\right)^r/c_r}=o(1), \qquad (2.22)\end{aligned}$$

when $h\to 0$. Thus, from (2.18) and (2.20)–(2.22) we have that

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

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. Applying dominated convergence

$$\lim_{h\to 0} \sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f\, w\left(\sigma_h^{-1}\big(\hat{f}_h(x) - f(x)\big)\right) = \mathbf{E}\, w\,(\xi). \tag{2.23}$$

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

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

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

$$\theta^2(h) = \frac{s_h^2}{2\pi^2}\left(\int_0^{s_h} \frac{\gamma}{\beta^2} e^{2(\gamma t)^r} dt\right)^{-1} \tag{2.24}$$

then

$$\begin{aligned}\int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} \big|\mathcal{F}[f_\theta](t)\big|^2 dt \;&=\; \theta^2 \frac{\pi^2}{s_h^2} \int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} \big|\mathcal{F}[k_{s_h}](t)\big|^2 dt \\ &\le\; \frac{\theta^2(h)\pi^2}{s_h^2} \int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} \mathbb{1}_{[-s_h, s_h]}(t)\, dt \;\le\; 1.\end{aligned}$$

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

Now, we can apply Kakutani's theorem using the fact that $\sum_{\ell=-\infty}^{\infty} g^2(\ell h) < \infty$ according to Lemma 4.2(c), and see that

$$\frac{d\mathbf{P}_\theta^{(h)}}{d\mathbf{P}_0^{(h)}}(\mathbf{y}) \;=\; \exp\left\{\frac{1}{2\sigma^2}\sum_{\ell=-\infty}^{\infty}\big(2\theta\, y_\ell\, g(\ell h) - \theta^2 g(\ell h)\big)\right\}, \tag{2.25}$$

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

$$T = \frac{\sum_{\ell=-\infty}^{\infty} y_\ell\, g(\ell h)}{\sum_{\ell=-\infty}^{\infty} g^2(\ell h)} \tag{2.26}$$

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

$$T \sim \mathcal{N}\left(\theta\,, \frac{\sigma^2}{\sum_{\ell=-\infty}^{\infty} g^2(\ell h)}\right), \tag{2.27}$$

and applying Lemma 2.1(c), with $s_1 = 0$ and $s_2 = s_h$, we see that

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

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

$$T = \theta + \varphi\,\xi \quad \text{where} \quad \xi \sim \mathcal{N}(0,1) \tag{2.28}$$

and, according to the previous arguments,

$$\varphi^2 = \frac{\sigma^2}{\sum_{\ell=-\infty}^{\infty} g^2(\ell h)} = \sigma_h^2\big(1+o(1)\big). \tag{2.29}$$

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

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

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

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

Here the last equation follows from Levit [1980]. According to (2.29), $\frac{\varphi}{\sigma_h} = 1+o(1), (h\to 0)$, while applying identity (2.19) and Lemma 2.2 we see that

$$\frac{\sigma_h^2}{\theta^2(h)} = 2\,\frac{\pi\gamma\sigma^2 h}{\beta^2}\,\frac{\int_0^{s_h}\gamma e^{2(\gamma t)^r}dt}{\gamma s_h} = \frac{2\int_0^{\gamma s_h}\gamma e^{2t^r}dt}{\gamma s_h e^{2(\gamma s_h)^r}} \le \frac{1}{r(\gamma s_h)^r}\to 0, \tag{2.30}$$

when $h\to 0$. Thus we have that

$$\begin{aligned}
&\liminf_{h\to 0}\sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f\, w\left(\sqrt{\frac{\pi}{\sigma^2 h s_h}}\big(\hat f_h(x)-f(x)\big)\right) \ge\\
&\qquad \liminf_{h\to 0}\inf_{\tilde f_h}\sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f\, w\left(\sqrt{\frac{\pi}{\sigma^2 h s_h}}\big(\tilde f_h(x)-f(x)\big)\right) \ge \mathbf{E}\, w(\xi).
\end{aligned} \tag{2.31}$$

Combining the relations (2.23) and (2.31) proves the theorem. □

2.3.2 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 2.1. 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 2.4 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 2.4. 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 2.1. Note that now the optimum bandwidth s_h depends on h also through the parameters γ, β and r. Nevertheless the statement of Theorem 2.1 still holds, as we shall see, under corresponding assumptions.

Theorem 2.2 *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*

$$0 < \liminf_{h\to 0} r \le \limsup_{h\to 0} r < \infty, \tag{2.32}$$

$$\liminf_{h\to 0} \frac{\beta^2}{\gamma\, \sigma^2 h} = \infty, \tag{2.33}$$

$$\limsup_{h\to 0} \frac{h}{\gamma} \left(\log \frac{\beta^2}{\gamma\, \sigma^2 h} \right)^{1/r} = 0. \tag{2.34}$$

Then

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

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

where s_h, $\tilde{f}_h$ and $\hat{f}_h$ are the same as in Theorem 2.1.

Remark 2.1 *Note that the conditions (2.32) and (2.34) 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; on other words they apply even when there is no consistency! There is no contradiction in that! This situation is similar to having a minimax estimator of the normal means whose variance may not be small. In other words, the asymptotic optimality (minimaxity) of the proposed estimator does not hinge on its consistency, although we might prefer to have both!*

Proof: We prove this theorem following the same proof of Theorem 2.1. It is sufficient to see that relations (2.20)–(2.22) and (2.30) still hold for the class $\mathcal{A}(\gamma_h, \beta_h, r_h)$. The limit (2.20) follows from (2.32) and (2.34), the limits (2.21) and (2.30) follow from (2.32) and (2.33). Finally (2.22) follows from the identity

$$\frac{\beta^2}{\gamma\,\sigma^2 h}\, e^{-\left(2\pi\frac{\gamma}{h}\right)^r/c_r} = \exp\Big\{ - c_r^{-1}\Big(2\pi\frac{\gamma}{h}\Big)^r\Big(1 - \frac{c_r}{(2\pi)^r}\Big(\frac{h}{\gamma}\Big(\log\frac{\beta^2}{\gamma\,\sigma^2 h}\Big)^{1/r}\Big)^r\Big)\Big\}. \tag{2.35}$$

and conditions (2.32)–(2.34). Notice that $h/\gamma \to 0$, by (2.33) and (2.34). 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

$$\sigma_h^2(\alpha) = \frac{\sigma^2 h}{\pi\gamma}\left(\frac{1}{2}\log\frac{\beta^2}{\pi\gamma\,\sigma^2 h}\right)^{\frac{1}{r}}, \tag{2.36}$$

can be of any order, with respect to any of 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.

2.4 Adaptive minimax regression

2.4.1 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*

$$\mathcal{A}_{\mathcal{K}} = \Big\{\mathcal{A}(\alpha) \;\Big|\; \alpha \in \mathcal{K}\Big\},$$

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

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

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 2.2 *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 $\tilde{f}_h \in \mathcal{F}$*

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

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 2.2 hold uniformly on the scale $\mathcal{K}$; below we formulate these assumptions more explicitly.

Definition 2.3 *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.*

$$0 < \liminf_{h\to 0} \inf_{\alpha\in\mathcal{K}_h} r \le \limsup_{h\to 0} \sup_{\alpha\in\mathcal{K}_h} r < \infty, \tag{2.37}$$

$$\liminf_{h\to 0} \inf_{\alpha\in\mathcal{K}_h} \frac{\beta^2}{\gamma\,\sigma^2 h} = \infty, \tag{2.38}$$

$$\limsup_{h\to 0} \sup_{\alpha\in\mathcal{K}_h} \frac{h^{1-\delta}}{\gamma} \left(\log \frac{\beta^2}{\gamma\,\sigma^2 h} \right)^{1/r} = 0. \tag{2.39}$$

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 2.2. 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 2.4 *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

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

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

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

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.

2.4.2 The adaptive estimator: upper bound

Section 2.4.1 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 2.4.3 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 2.2.

Remark 2.2 *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

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

where $s_h(\alpha)$ and $\sigma_h^2(\alpha)$ were defined in (2.12) and (2.17).

Theorem 2.3 *For any $p > 0$ there exists an adaptive estimator $\hat{f}_h$ such that for any $x \in \mathbb{R}$ and for any RNP-functions scale $\mathcal{A}_{\mathcal{K}_h}$, $\hat{f}_h \in \mathcal{F}_p$*

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

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 $\mathcal{S}_h = \big\{s_0, s_1, \ldots s_{I_h}\big\}$ where

$$I_h = \arg\max_i \big\{h s_i \le \log^{-1} 1/h\big\}, \tag{2.40}$$

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

Now, let us denote

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

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

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

and define the thresholds

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

Finally we define

$$\hat{i} = \min \left\{ 1 \le i \le I_h \ : \ \left| \hat{f}_j(x) - \hat{f}_i(x) \right| \le \lambda_j \, \hat{\sigma}_{i,j} \quad \forall j \ (i \le j \le I_h) \right\}. \tag{2.41}$$

We will prove below that the estimator

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

satisfies both the statements contained in Theorem 2.3.

Let us get first some insight into the algorithm. The sequence of bandwidths s_I 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. (2.39), are such that $h s_h(\alpha) \le h^\delta$ 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 $\mathcal{S}_h$, 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

$$\left| \hat{f}_j(x) - \hat{f}_i(x) \right| > \lambda_j \, \mathbf{Var}^{1/2}\big(\hat{f}_j(x) - \hat{f}_i(x) \big), \tag{2.42}$$

is very small since, except for an event of a small probability, this can only occur if the bias $(b_j - b_i) \gg \mathbf{Var}^{1/2}\big(\hat{f}_j(x) - \hat{f}_i(x) \big)$ 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 (2.42) 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 the our procedure warrants that, cf. (2.41),

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

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 2.1 and 2.6.

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 2.3 *For $h \to 0$, uniformly with respect to $1 \le i, j \le I_h$ and with respect to α varying in an R-scale,*

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

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

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

Proof. (a) Using the bound for the bias given in (2.14), equation (2.19) and Lemma 2.2 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 $s_j \ge s_h(\alpha)$ and using conditions (2.40) in the definition of the sequence of bandwidths $\mathcal{S}_h$ and conditions (2.37)–(2.39) 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 ($i(\alpha) \le j \le I_h$) and with respect to α in $\mathcal{K}_h$. .

(b) This just a reformulation of the asymptotic relation (2.15) using the fact that $hs_j \le \log^{-1}(1/h)$ according to (2.40).

(c) Applying Lemma 2.1(b) taking $s_1 = s_i$ and $s_2 = s_j$, and arguing as in (2.14) and in the proof (a), we see that

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

(d) It follows directly from Lemma 2.1(c), taking $s_1 = s_i$ and $s_2 = s_j$. Here, as in (2.15), we can verify

$$
\begin{aligned}
\sigma_{i,j}^2 &= \sigma^2 h^2\sum_{\ell=-\infty}^{\infty}\big(k_{s_j}(x-\ell h) - k_{s_i}(x-\ell h)\big)^2 \\
&= \frac{\sigma^2 h\,(s_j-s_i)}{\pi}\big(1+O_h(1)\,h(s_j-s_i)\big). \qquad (2.43)
\end{aligned}
$$

and thus, using (2.40), this completes the proof of the lemma. □

We now proceed with proving Theorem 2.3. For arbitrary f in any R-functional scale $\mathcal{A}_{\mathcal{K}_h}$,

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

where

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

and

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

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

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

therefore

$$\begin{aligned}R_h^-(f) &\le \mathbf{E}\Big\{ \mathbb{1}_{\{\hat{i}\le i(\alpha)\}} \Big(\big|\, \hat{f}_{\hat{i}}(x) - \hat{f}_{i(\alpha)}(x) \,\big| + \big|\, \hat{f}_{i(\alpha)}(x) - f(x) \,\big| \Big)^p \Big\} \\ &\le \mathbf{E}\Big(\hat{\sigma}_{i(\alpha)}\lambda_{i(\alpha)} + \big|\, \hat{f}_{i(\alpha)}(x) - f(x) \,\big| \Big)^p \\ &\le \mathbf{E}\Big(\hat{\sigma}_{i(\alpha)}\lambda_{i(\alpha)} + |b_{i(\alpha)}| + \sigma_{i(\alpha)}|\xi| \Big)^p\end{aligned}$$

where $\xi \sim \mathcal{N}(0,1)$. Now according to Lemma 2.3, (a) and (b), uniformly with respect to α in any R-scale

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

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

$$R_h^-(f) = O(h^{p/2}), \tag{2.44}$$

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

$$R_h^-(f) \le \psi_h^p(\alpha)(1+o(1)). \tag{2.45}$$

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

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

Applying Hölder's inequality we obtain

$$\begin{aligned}R_h^+(f) &= \mathbf{E}\left\{ \mathbb{1}_{\{\hat{i}>i(\alpha)\}} \big|\hat{f}_{\hat{i}}(x) - f(x)\big|^p \right\} = \sum_{i=i(\alpha)+1}^{I_h} \mathbf{E}\left\{ \mathbb{1}_{\{\hat{i}=i\}} \big|\hat{f}_i(x) - f(x)\big|^p \right\} \\ &= \sum_{i=i(\alpha)+1}^{I_h} \mathbf{E}\left\{ \big|\hat{f}_i(x) - f(x)\big|^p \left(\mathbb{1}_{\{\hat{i}=i\}\cap A_i} + \mathbb{1}_{\{\hat{i}=i\}\cap A_i^c} \right) \right\} \\ &= R_{h,1}^+(f) + R_{h,2}^+(f),\end{aligned}$$

where

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

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

$$\begin{aligned}\mathbf{P}(\hat{i} = i) \;&\le\; \mathbf{P}(\hat{i} \ge i) \\ &\le\; \sum_{j=i+1}^{\infty} \mathbf{P}\left(|\hat{f}_{j-1}(x) - \hat{f}_{i-1}(x)| > \hat{\sigma}_{i-1,j-1}\, \lambda_{j-1} \right). \qquad (2.46)\end{aligned}$$

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 2.3(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

$$\mathbf{P}\Big(|\hat{f}_j(x) - \hat{f}_i(x)| > \hat{\sigma}_{i,j}\lambda_j \Big) \;\le\; \mathbf{P}\Big(|\xi| > \frac{\hat{\sigma}_{i,j}}{\sigma_{i,j}}\lambda_j - \frac{|b_j - b_i|}{\sigma_{i,j}} \Big) \qquad (2.47)$$

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

$$\le\; \exp\left\{ -\frac{1}{2}\lambda_j^2 + C\frac{\hat{\sigma}_{i,j}}{\sigma_{i,j}}\lambda_j + \frac{1}{2}\left(1 - \frac{\hat{\sigma}^2_{i,j}}{\sigma^2_{ij}}\right)\lambda_j^2 \right\}. \qquad (2.48)$$

Since by Lemma 2.3(c) and (2.40)

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

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

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

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

Returning to (2.46) we obtain that

$$\begin{aligned}
\mathbf{P}(\hat{i} \geq i) &\leq C_1 \sum_{j=i+1}^{\infty} \exp\left\{-\frac{1}{2}\lambda_{j-1}^2 + 2C\lambda_{j-1}\right\} = C_1 \sum_{j=i}^{\infty} \exp\left\{-\frac{1}{2}\lambda_j^2 + 2C\lambda_j\right\} \\
&= C_1 \sum_{j=i}^{\infty} \exp\left\{-\frac{pj^l + p_1 j^{l_1}}{2} + 2C\sqrt{pj^l + p_1 j^{l_1}}\right\} \\
&\leq C_1 \sum_{j=i}^{\infty} \exp\Big\{-\frac{pj^l}{2} - \frac{p_1 j^{l_1}}{3}\Big\} \sim C_1 \frac{2}{pl} i^{1-l} \exp\Big\{-\frac{pi^l}{2} - \frac{p_1 i^{l_1}}{3}\Big\} \\
&= C_1 \frac{2}{pl} i^{1-l} s_i^{-p/2} \exp\Big\{-\frac{p_1 i^{l_1}}{3}\Big\} \leq C_2 s_i^{-p/2} \exp\Big\{-\frac{p_1 i^{l_1}}{4}\Big\} \qquad (2.49)
\end{aligned}$$

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

$$R_1^+(f) = O(h^{p/2}) \sum_{i=1}^{\infty} i^{pl/2} \exp\Big\{-p_1 i^{l_1}/4\Big\} = O(h^{p/2}), \qquad (h \to 0).$$

In order to obtain a bound on $R_2^+(f)$ we write again $\hat{f}_i - f(x) = b_i + \sigma_i \xi$, $\xi \sim \mathcal{N}(0,1)$. Applying Lemma 2.3, (a) and (b), in the same way as before, we have

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

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

$$\begin{aligned}
R_{h,2}^+(f) &= \sum_{i=i(\alpha)+1}^{I_h} \mathbf{E}^{1/2}\big|\hat{f}_i(x) - f(x)\big|^{2p}\, \mathbf{P}^{1/2}(A_i^c) \\
&\leq \sum_{i=i(\alpha)+1}^{I_h} \hat{\sigma}_i^p\, \mathbf{E}^{1/2}\big|\, o(1) + (1+o(1))\xi\,\big|^{2p}\, \mathbf{P}^{1/2}(A_i^c) \\
&\leq O(1)\Big(\frac{\sigma^2 h}{\pi}\Big)^{p/2} \sum_{i=1}^{\infty} \exp\Big\{-p_1 i^{r_1}/4\Big\} \\
&= O\big(h^{p/2}\big), \qquad (h \to 0), \qquad (2.50)
\end{aligned}$$

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

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

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

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

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

when $h\to 0$. □

2.4.3 Lower bound: optimality results

In Section 2.4.2 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 2.2.

Theorem 2.4 *Let $p>0$. Let $\mathcal{A}_{\mathcal{K}_h}$ be an arbitrary RNP-scale. Assume that $\tilde{s}_h=\tilde{s}_h(\alpha)$, $\tilde{s}_h\le s_h(\alpha)$, and $\tilde{\phi}_h(\alpha)$ are defined in such a way that for all sufficiently small h and $\alpha\in\mathcal{K}_h$*

$$\tilde{\phi}_h^2=\tilde{\phi}_h^2(\alpha)\le\min(\,p\log\tilde{s}_h\,,r(\gamma\tilde{s}_h)^r/2\,)\tag{2.51}$$

and

$$\lim_{h\to 0}\inf_{\alpha\in\mathcal{K}_h}\tilde{\phi}_h=\infty.\tag{2.52}$$

Denote

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

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

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

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

$$\begin{aligned}f_0(z)&\equiv 0,\\ f_1(z)&=\theta\,\tilde{g}(z),\quad \tilde{g}(z)=\sqrt{\frac{\sigma^2h\,\pi}{\tilde{s}_h}}k_{\tilde{s}_h}(x-z).\end{aligned}\tag{2.53}$$

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 (2.7), definition (2.12) of s_h, and Lemma 2.2, we get

$$\begin{aligned}
\int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} \big|\,\mathcal{F}[f_1](t)\big|^2\,dt \;&=\; \theta^2 \frac{\sigma^2 h\,\pi}{\tilde{s}_h} \int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} \big|\mathcal{F}[k_{\tilde{s}_h}](t)\big|^2\,dt \\
&=\; 2\,\theta^2 \frac{\gamma\sigma^2 h\,\pi}{\beta^2} \frac{\int_0^{\tilde{s}_h} \gamma e^{2(\gamma t)^r}\,dt}{\gamma\tilde{s}_h} \;=\; 2\,\theta^2\, e^{-2(\gamma s_h)^r}\, \frac{\int_0^{\tilde{s}_h} \gamma e^{2(\gamma t)^r}\,dt}{\gamma\tilde{s}_h} \\
&=\; \frac{\theta^2}{r(\gamma\tilde{s}_h)^r} e^{2(\gamma\tilde{s}_h)^r - 2(\gamma s_h)^r}(1+o(1)) \;\le\; \frac{\tilde{\phi}_h^2}{r(\gamma\tilde{s}_h)^r} e^{2(\gamma\tilde{s}_h)^r - 2(\gamma s_h)^r}(1+o(1)) \qquad (2.54) \\
&\le\; \frac{1}{2}(1+o(1)) \le 1,
\end{aligned}$$

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

$$\psi_h^{-1}\big(\tilde{f}_h(x) - f_1(x)\big) = f_h^* - \psi_h^{-1} f_1(x) = f_h^* - \tilde{\phi}_h^{-1}\theta = f_h^* - L \qquad (2.55)$$

whereas

$$\begin{aligned}
\sqrt{\frac{\pi}{\sigma^2 h}}\big(\tilde{f}_h(x) - f_0(x)\big) \;&=\; \sqrt{\frac{\pi}{\sigma^2 h}}\,\tilde{f}_h(x) = \sqrt{\frac{\pi}{\sigma^2 h}}\,\tilde{\psi}_h\, f_h^*(x) \\
&=\; \sqrt{\frac{\pi}{\sigma^2 h}}\,\sqrt{\frac{\sigma^2 h\,\tilde{s}_h}{\pi}}\,\tilde{\phi}_h\, f_h^*(x) = \tilde{s}_h^{1/2}\,\tilde{\phi}_h\, f_h^*(x) \\
&=\; f_h^* \exp\Big\{\frac{\log\tilde{s}_h}{2} + \log\tilde{\phi}_h\Big\}. \qquad (2.56)
\end{aligned}$$

Denote $q = \exp\big\{-\tilde{\phi}_h\big\}$ so that by (2.52), $q \to 0$ 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

$$\begin{aligned}
\mathcal{R}^{(h)} \;&:=\; \sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}_f\left(\tilde{\psi}_h^{-1}\big|\tilde{f}_h(x) - f(x)\big|\right) \ge \mathbf{E}_1\left(\tilde{\psi}_h^{-1}\big|\tilde{f}_h(x) - f_1(x)\big|\right) \\
&\ge\; q\,\mathbf{E}_0\left(\sqrt{\frac{\pi}{\sigma^2 h}}\big|\tilde{f}_h(x) - f_0(x)\big|\right) + (1-q)\,\mathbf{E}_1\left(\tilde{\psi}_h^{-1}\big|\tilde{f}_h(x) - f_1(x)\big|\right) + O(q). \qquad (2.57)
\end{aligned}$$

According to (2.51) and (2.55)–(2.57),

$$\begin{aligned}
\mathcal{R}^{(h)} &\geq q\,\exp\left\{\frac{\tilde{\phi}_h}{2}+p\log\tilde{\phi}_h\right\}\mathbf{E}_0\left|f_h^*(x)\right|^p+(1-q)\,\mathbf{E}_1\left|f_h^*(x)-L\right|^p+O(q)\\
&\geq (1-q)\,\mathbf{E}_1\left(Z\left|f_h^*(x)\right|^p+\left|f_h^*(x)-L\right|^p\right)+O(q)\\
&\geq (1-q)\,\mathbf{E}_1\inf_x\left(Z|x|^p+|x-L|^p\right)+O(q) \qquad (2.58)
\end{aligned}$$

where

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

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],

$$\min_x g(x)=\begin{cases}\min(Z,1)L^p & \text{if } p\leq 1,\\ \left(1+Z^{-\frac{1}{p-1}}\right)^{-(p-1)}L^p & \text{if } p>1.\end{cases} \qquad (2.59)$$

Thus for any $p>0$ we can write

$$\min_x g(x)=\chi L^p, \qquad (2.60)$$

where χ is defined by (2.59) and satisfies $0<\chi\leq 1$.

Now, let us consider the likelihood corresponding to f_0 and f_1. Using the same arguments that we used in (2.25)–(2.29) we can see that

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

where $\xi\sim\mathcal{N}(0,1)$ with respect to $\mathbf{P}_1$. Using the definition of θ and (2.51) and definition (2.53) we can see that

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

Note that by (2.52)

$$Z = \big(1+o(1)\big)\exp\left\{-\tilde\phi_h + \frac{\tilde\phi_h^2}{2} + p\log\tilde\phi_h - (\,\tilde\phi_h - \sqrt{\tilde\phi_h})\xi - \frac{1}{2}(\,\tilde\phi_h - \sqrt{\tilde\phi_h})^2\right\} \stackrel{\mathbf{P}_1}{\to} \infty.$$

hence $\chi \stackrel{\mathbf{P}_1}{\to} 1$. Also $L = 1 + o(1)$, according to its definition. Therefore according to equations (2.58)–(2.60), uniformly in $\alpha \in \mathcal{K}_h$,

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

□

Corollary 2.1 *Let $\mathcal{A}_{\mathcal{K}_h}$ be an arbitrary RNP-scale such that*

$$\liminf_{h\to 0}\inf_{\alpha\in\mathcal{K}_h}\frac{r(\gamma s_h)^r}{\log s_h} = \infty \tag{2.61}$$

where s_h is the optimum bandwidth defined in (2.12). Then for any $p > 0$ and $x \in \mathbb{R}$, the estimator $\hat f_h$ of Theorem 2.3 is $(p, \mathcal{K}_h, \mathcal{F}_p(x))$-adaptively minimax at x.

Proof. This is a consequence of Theorem 2.3 and 2.4. In order to prove the lower bound take $s_h = \tilde s_h$ and apply the previous theorem. □

Now, we prove a version of Theorem 2.4 under a weaker condition. It will be used below to provide an easily verifiable conditions for adaptive optimality of the estimator proposed in Section 2.4.2.

Theorem 2.5 *Let $\mathcal{A}_{\mathcal{K}_h}$ be an arbitrary RNP-scale such that*

$$\liminf_{h\to 0}\inf_{\alpha\in\mathcal{K}_h}\frac{r(\gamma s_h)^r}{\log\log s_h} = \infty \tag{2.62}$$

where the optimum bandwidth s_h was defined in (2.12). Then for any estimator $\tilde f_h \in \mathcal{F}_p^0(x)$,

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

where

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

Proof: We prove this theorem in the same way as Theorem 2.4 by choosing $\tilde\phi_h^2 = p\log\tilde s_h$ and subsequently defining $\tilde s_h$ in such a way that

$$\frac{2p\log\tilde s_h}{r(\gamma\tilde s_h)^r}\,e^{2(\gamma\tilde s_h)^r - 2(\gamma s_h)^r} \leq 1 \tag{2.63}$$

for h small enough. The point here is that condition (2.63) was only needed in proving (2.54), which now becomes (2.63). 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

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

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 (2.62). Thus the auxiliary bandwidth $\bar{s}_h$ is asymptotically equivalent to s_h. It also satisfies (2.62), 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

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

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,

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

and we see that

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

for h small enough. The rest of the proof is the same as for Theorem 2.4. □

Finally, we prove that the estimator we constructed in Theorem 2.3 is adaptively minimax, for any RNP-scale satisfying a condition just a little stronger than condition (2.36) used in the definition of a regular scale.

Theorem 2.6 *Let $\mathcal{K}_h$ be a RNP-scale such that*

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

for some τ ($0 < \tau < 1$) and $C > 0$. Then for any $p > 0$ and $x \in \mathbb{R}$, the estimator $\hat{f}_h$ of Theorem 2.3 is $(p, \mathcal{K}_h, \mathcal{F}_p(x))$-adaptively minimax at x.

Proof. The upper bound result was proved in Theorem 2.3. To prove the lower bound we notice that

$$r(\gamma s_h)^r = \frac{r}{2}\log\frac{\beta^2}{\pi\gamma\sigma^2 h} \geq \frac{r\tau}{2}\log Ch^{-1}$$

while according to conditions 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 2.5. □

Chapter 3

Adaptive regression on a bounded interval

In the previous chapter we discussed a regression model in which the unknown function $f(x)$ was assumed to be analytic on the whole real line. This model required that the observations of $f(x)$ were made on the the whole line. In practice however observations of the regression function $f(x)$ are often available only on a bounded interval. This is exactly the case which will be discussed in the current chapter. To begin with, we will introduce relative classes $\mathcal{A}(\gamma, M)$ of functions analytic in a vicinity of a given interval (Section 3.1.1).

Next, an important issue of the observation design will be highlighted. A seemingly natural, but somewhat naive approach is to use the simplest possible uniform, or equidistant, design. We will see however that such a design loses substantially in accuracy, near the end-points of the observation interval. We will explain that this is not a drawback of any specific method of estimation, but rather an in-built defect of the equidistant design itself.

A much more satisfactory design is the Chebyshev design. In Section 3.2 we describe, pointwisely, the quality of the best attainable accuracy of estimation for both designs. Finally, in Section 3.3 we will present our main results about adaptive estimation of functions $f \in \mathcal{A}(\gamma, M)$. Here we restrict ourselves only to the Chebyshev designs, in view of their greater efficiency. Without any loss of generality, we will assume throughout the chapter that our observation interval is $[-1, 1]$; a generalization to an arbitrary bounded interval $[a, b]$ is straightforward; cf. e.g. Timan [1963], Sect. 3.7.

3.1 The building blocks

The purpose of this section is to introduce classes $\mathcal{A}(\gamma, M)$ of analytic functions, as well as the Legendre and Chebyshev polynomials. We discuss their properties and the relation between them. Classes $\mathcal{A}(\gamma, M)$ will serve as the underlying functional classes in the regression problems that we will study, while Legendre and Chebyshev polynomials will be

used, in corresponding designs, for constructing the estimators.

3.1.1 The class $\mathcal{A}(\gamma, M)$

For $\gamma > 0$ let E_γ be the open ellipse in the complex plane, with its boundary defined by

$$\partial E_\gamma = \{z \in \mathbb{C} \ : \ z = \cosh\gamma\cos\phi + i\sinh\gamma\sin\phi, \ 0 \le \phi \le 2\pi\}.$$

The ellipses E_γ represent a convenient family of vicinities of the interval $[-1, 1]$, expanding from $[-1, 1]$ to $\mathbb{C}$, as γ increases from 0 to ∞. One can verify by simple algebra that the elliptic boundary ∂E_γ has its foci at the end-points of the interval $[-1, 1]$, thus

$$E_\gamma = \{z \in \mathbb{C} : |z-1| + |z+1| < e^\gamma + e^{-\gamma}\}.$$

Definition 3.1 *We denote by $\mathcal{A}(\gamma, M)$ the class of functions analytic inside E_γ such that $|f(z)| \le M$, for all $z \in E_\gamma$.*

Denote by ρ_γ the distance from the interval $[-1, 1]$ to the boundary ∂E_γ. From the integral Cauchy formula for the *mth* derivative of analytic functions we know that for any $\epsilon > 0$ and any ball $B_{\rho_\gamma - \epsilon}$ of radii $\rho_\gamma - \epsilon$ centered at $x \in [-1, 1]$,

$$f^{(m)}(x) = \frac{m!}{2\pi\mathbf{i}} \int_{B_{\rho_\gamma-\epsilon}} \frac{f(z)}{(z-x)^{m+1}}\, dz, \qquad m = 1, 2, \ldots.$$

Thus, since ϵ is arbitrary, one obtains for the derivatives of the functions $f \in \mathcal{A}(\gamma, M)$ the following bounds:

$$|f^{(m)}(x)| \le M m!/\rho_\gamma^m \tag{3.1}$$

for all $x \in [-1, 1]$. An elementary calculation shows that

$$\rho_\gamma = \cosh\gamma - 1. \tag{3.2}$$

Equations (3.1) and (3.2) will be used later in Section 3.2, in obtaining some discrete-type approximations to analytic functions.

3.1.2 Legendre polynomials

Legendre polynomials form a complete system of orthogonal polynomials in $L^2([-1, 1])$. Their explicit definition is (cf. Szegö [1975], p. 68)

$$P_r(x) = 2^{-r} \sum_{\nu=0}^{r} \binom{r}{r-\nu} \binom{r}{\nu} (x-1)^\nu (x+1)^{r-\nu}, \tag{3.3}$$

and their recurrent form is (cf. Szegö, p. 71)

$$P_0 \equiv 1,$$

$$P_1(x) = x,$$
$$rP_r(x) = (2r-1)xP_{r-1}(x) - (r-1)P_{r-2}(x), \quad r \geq 2.$$

In particular, from the definition (3.3), it holds

$$P_r(1) = 1, \quad P_r(-1) = (-1)^r. \tag{3.4}$$

An important bound for the derivatives of Legendre polynomials can be obtained by combining the A.A. Markov inequality (cf. Timan [1963], Sect. 4.8.8)

$$|P_r^{(m)}(x)| \leq r^{2m} \max_{-1\leq x\leq 1} |P_r(x)|, \qquad m = 1, 2, \ldots; \tag{3.5}$$

with the fact that the maximum of $|P_r(x)|$ is attained at the end points of the interval (cf. Szegö, Sect. 7.21),

$$\max_{-1\leq x\leq 1} |P_r(x)| = |P_r(\pm 1)| = 1. \tag{3.6}$$

The normalized Legendre polynomials, given by

$$p_r(x) = (2r+1)^{1/2} P_r(x), \qquad r = 0, 1, \ldots, \tag{3.7}$$

satisfy, from (3.5)–(3.7),

$$\max_{-1\leq x\leq 1} |p_r^{(m)}(x)| \leq (2r+1)^{1/2} r^{2m} \qquad m = 1, 2, \ldots. \tag{3.8}$$

The defined normalized Legendre polynomials form an orthonormal basis in the space $L^2([-1,1])$ corresponding to the inner product

$$\langle f \,|\, g \rangle := \frac{1}{2} \int_{-1}^{1} f(x) g(x)\, dx.$$

Besides that, they are asymptotically orthonormal with respect to a "discrete" inner product defined below which is a discrete version of the "continuous" inner product just mentioned. For a given design, x_k^n, $k = 1, 2, \ldots, n$, we define the corresponding discrete inner product of the functions f and g to be

$$(f \,|\, g) := \frac{1}{n} \sum_{k=1}^{n} f(x_k^n) g(x_k^n).$$

In this subsection, we consider the discrete inner product with respect to the Legendre design, for which x_k^n represent the equidistant knots

$$x_k^n = \frac{2k - n - 1}{n}, \quad k = 1, \ldots, n. \tag{3.9}$$

Let us denote the *kernel* corresponding to the Legendre family p_r by

$$K_N(x, y) := \sum_{r=0}^{N-1} p_r(x) p_r(y). \tag{3.10}$$

Underlying the quality of our estimators will be remarkable properties of the following type.

Lemma 3.1 *Let $N \in \mathbb{N}$. The normalized Legendre polynomials p_r satisfy*

(a) *Uniformly for $0 \le r_1, r_2 \le N$,*

$$(\, p_{r_1} | \, p_{r_2}) \;=\; \frac{1}{n}\sum_{k=1}^{n} p_{r_1}(x_k^n) p_{r_2}(x_k^n) \;=\; \delta_{r_1\, r_2} + O\left(\frac{N^6}{n^2}\right), \qquad (n \to \infty). \quad (3.11)$$

(b) *If*

$$\alpha_N^2(x) := \; \frac{1}{N} K_N(x,x) \;=\; \frac{1}{N}\sum_{r=0}^{N-1} p_r^2(x), \quad (3.12)$$

then

$$\alpha_N^2(x) = \frac{2}{\pi\sqrt{1-x^2}}(1+o(1)), \qquad (N \to \infty), \quad (3.13)$$

uniformly on any interval $[a,b] \subset (-1,1)$, and $\alpha_N^2(\pm 1) = N$.

Remark 3.1 *Note the different behavior of α_N inside the interval and at the end-points. This will explain why the results presented below hold uniformly only on the compact subsets of $(-1,1)$ while at the extremes of the interval the accuracy of estimation, based on the equidistant design, will deteriorate, even to the extent of being of a different order!*

The property (b) is illustrated by Figure 3.1.

Proof. (a) The numerical integration method for approximating $\int_a^b g(x)dx$, in which the interval is divided in n equally spaced sub-intervals and the function is evaluated at the middle points of the sub-intervals, has the accuracy bounded by

$$\frac{(b-a)^2}{24n^2} \max_{a\le x\le b} \left| \frac{d^2}{dx^2} f(x) \right| \quad (3.14)$$

when the function $f \in C^2[a,b]$ (cf. e.g. Stoer and Bulirsch). Thus, we have

$$\begin{aligned} |\,(\, p_{r_1} | \, p_{r_2}) - \langle p_{r_1} \,|\, p_{r_2}\rangle\,| \;&=\; \left| \frac{1}{n}\sum_{k=1}^{n} p_{r_1}(x_k^n) p_{r_2}(x_k^n) - \frac{1}{2}\int_{-1}^{1} p_{r_1}(x) p_{r_2}(x) dx \right| \\ &\le\; \frac{1}{3n^2} \max_{-1\le x\le 1} \frac{d^2}{dx^2}\left(p_{r_1}(x) p_{r_2}(x)\right). \end{aligned} \quad (3.15)$$

Applying L^2-orthonormality and bounds (3.8) for the derivatives of $p_r(x)$ we get

$$|\,(\, p_{r_1} | \, p_{r_2}) - \delta_{r_1,r_2}\,| \;\le\; \frac{1}{3n^2}(2r_1+1)(2r_2+1)(r_1^2+r_2^2)^2 = O\left(\frac{N^6}{n^2}\right)$$

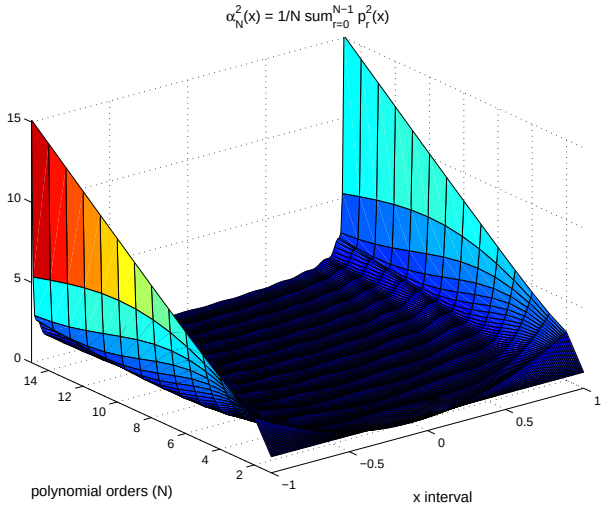

Figure 3.1: Sums of squared Legendre polynomials

as $n \to \infty$.

(b) Using the asymptotic formula of Laplace (cf. Szegö, p. 194)

$$p_r(x) \sim \frac{2}{\sqrt{\pi}(1-x^2)^{1/2}} \cos((r+1/2)\sqrt{1-x^2} - \frac{\pi}{4}) + O(r^{-1}), \qquad r \to \infty, |x| < 1 \tag{3.16}$$

and formula (cf. e.g. Gradshtein and Ryzhik, f. 1.341(1), p. 29)

$$\sum_{r=0}^{N-1} \sin(r\theta_1 + \theta_2) = \sin(\frac{N-1}{2}\theta_1 + \theta_2) \sin\frac{N\theta_1}{2} \csc\frac{\theta_1}{2}$$

we obtain, with some algebra,

$$\begin{aligned} \frac{1}{N}\sum_{r=0}^{N-1} p_r^2(x) &= \frac{2}{\pi\sqrt{1-x^2}}\Big(1 - \frac{1}{N}\sum_{r=0}^{N-1} \sin((2r+1)\theta) + O(N^{-1}\log N)\Big) \\ &= \frac{2}{\pi\sqrt{1-x^2}}(1+o(1)), \qquad (N \to \infty), \end{aligned}$$

uniformly on compacts in $(-1, 1)$. At the end-points

$$\frac{1}{N}\sum_{r=0}^{N-1} p_r^2(\pm 1) = \frac{1}{N}\sum_{r=0}^{N-1}(2r+1) = N.$$

□

Finally, let us mention the following bound on the growth of the Legendre polynomials outside the interval $[-1,1]$. According to Timan, Theorem 2.9.11, for any polynomial P_r of order r and any $z \in \mathbb{C}$

$$|P_r(z)| \le |T_r(z)| \max_{-1\le x\le 1} |P_r(x)|.$$

Here $T_r(x)$ are the Chebyshev polynomials which will be discussed in the next section. In particular we will see that $|T_r(z)| \le e^{\gamma r}, z \in E_\gamma$. Therefore according to (3.8),

$$|p_r(z)| \le (2r+1)^{1/2} e^{\gamma r} \tag{3.17}$$

for every $z \in E_\gamma$.

3.1.3 Chebyshev polynomials

Chebyshev polynomials appeared for the first time in the problem of finding polynomials $T_r(x) = x^r + a_1 x^{r-1} + \cdots + a_r$ least deviating from zero, in the uniform norm on the interval $[-1,1]$; Chebyshev [1859]. Normed by $T_r(1) = 1$, they can be represented as

$$T_r(x) = \cos r \arccos x, \quad r = 0, 1, \dots , \tag{3.18}$$

or in the recurrent form

$$T_0(x) = 1,$$

$$T_1(x) = x,$$

$$T_{r+1}(x) = 2xT_r(x) - T_{r-1}(x), \quad r = 1, 2, \dots .$$

The Chebyshev polynomials are extensively used as an appropriate Fourier basis for approximating non-periodic functions. Consider the normalized family

$$t_r(x) = \begin{cases} T_0(x), & r = 0 \\ \sqrt{2}\, T_r(x) & r \ne 0. \end{cases}$$

These polynomials constitute an orthonormal system in the weighted L^2-space with the scalar product

$$\langle f \,|\, g \rangle := \frac{1}{\pi} \int_{-1}^{1} \frac{f(x)g(x)}{\sqrt{1-x^2}}\, dx, \tag{3.19}$$

i.e. they satisfy $< t_{r_1} \,|\, t_{r_2} > = \delta_{r_1, r_2}$ for all integers $r_1, r_2 \ge 0$.

Denote by

$$K_N(x,y) := \sum_{r=0}^{N-1} t_r(x) t_r(y)$$

the kernel associated with the polynomials $t_r(x)$. For a given function f, the corresponding Chebyshev-Fourier series is given by

$$\sum_{r=0}^{\infty} \langle f \,|\, t_r \rangle t_r(x). \tag{3.20}$$

This expansion becomes just the classical trigonometric series if the change of variables $x = \cos\theta$ is made. The partial sum

$$f_N(x) = \sum_{r=0}^{N-1} \langle f \,|\, t_r \rangle t_r(x) \;=\; \langle f \,|\, K_N(x,\cdot) \rangle \tag{3.21}$$

provides the best approximation to a function f, with respect to the weighted L^2-norm corresponding to (3.19), among all polynomials of degree less than N. The class $\mathcal{A}(\gamma, M)$ has the important property that the coefficients of the Chebyshev-Fourier series (3.20) decrease very fast (cf. Timan, Sect. 3.7.3). For all $r = 0, 1, \ldots$, the inequality

$$\sup_{f \in \mathcal{A}(\gamma,M)} |\langle f \,|\, t_r \rangle| \;\leq\; \sqrt{\pi} M e^{-\gamma r} \tag{3.22}$$

holds. From (3.21),(3.22) and the bound $|t_r(x)| \leq \sqrt{2}$ it follows that for every $f \in \mathcal{A}(\gamma, M)$

$$\max_{x \in [-1,1]} |f_N(x) - f(x)| \leq \sum_{r=N}^{\infty} |\langle f \,|\, t_r \rangle| \; |t_r(x)| \leq \frac{\sqrt{2\pi} M}{1 - e^{-\gamma}} e^{-\gamma N}, \tag{3.23}$$

(cf. Timan, Sect. 3.7.3 and 5.4.1).

The function $f_N(x)$ is the polynomial of the best approximation in the weighted L^2-space. Remarkably, for analytic functions of the classes $\mathcal{A}(\gamma, M)$, the approximation $f_N(x)$ based on Chebyshev polynomials is asymptotically also the polynomial of the best uniform approximation on $[-1, 1]$. More precisely,

$$\sup_{f \in \mathcal{A}(\gamma,M)} \limsup_{N \to \infty} \left(\inf_{p \in Q_N} \|f - p\|_\infty \right)^{1/N} = \sup_{f \in \mathcal{A}(\gamma,M)} \limsup_{N \to \infty} \left(\|f - f_N\|_\infty \right)^{1/N},$$

where Q_N is the class of all the polynomials of the form $p = \sum_{k=0}^{N-1} a_k x^k$, (cf. Timan, Sect. 6.5.2).

According to their definition, the Chebyshev polynomials satisfy $|t_r(x)| \leq \sqrt{2}$ for all $x \in [-1, 1]$. Now we shall exhibit an interesting bound that can be obtained in the whole region E_γ. From the identity

$$2 \cos rt = (\cos t + i \sin t)^r + (\cos t - i \sin t)^r$$

it follows that

$$T_r(x) = \frac{1}{2} \Big(\big(x + \sqrt{x^2 - 1} \big)^r + \big(x - \sqrt{x^2 - 1} \big)^r \Big)$$

$$= \frac{1}{2}(\omega^r + \omega^{-r}),$$

where $x = \frac{1}{2}(\omega + \omega^{-1})$. Further, the transformation $z = \frac{1}{2}(\omega + \omega^{-1})$ maps the ring

$$\left\{\omega \in \mathbb{C} : e^{-\gamma} < |\omega| < e^{\gamma}\right\}$$

into E_γ and therefore $T_r(z) = \frac{1}{2}(\omega^r + \omega^{-r})$. Thus the normalized Chebyshev polynomials are bounded in E_γ by

$$|t_r(z)| = \sqrt{2}\,|T_r(z)| \le \sqrt{2}\,e^{\gamma r}. \tag{3.24}$$

Denote the discrete inner product by

$$(\, f \,|\, g \,) := \frac{1}{n}\sum_{k=1}^{n} f(x_k^n) g(x_k^n) \tag{3.25}$$

where the points x_k^n correspond to the Chebyshev design[3]

$$x_k^n = \cos\frac{(2k-1)\pi}{2n}, \quad k = 1, \ldots, n. \tag{3.26}$$

We can state next a lemma which is similar to Lemma 3.1. The first of the properties is usually referred to as 'double-orthogonality' (cf. e.g. Fox and Parker, Sect. 2.7) and is closely related to the corresponding property of the classical trigonometric polynomials. The second property follows from a standard calculation.

Lemma 3.2 *The normalized Chebyshev polynomials t_r satisfy*

(a) *For any $r_1, r_2 = 0, 1, ...$*

$$(\, t_{r_1} |\, t_{r_2}) = \frac{1}{n}\sum_{k=1}^{n} t_{r_1}(x_k^n) t_{r_2}(x_k^n) \;=\; \delta_{r_1\, r_2}, \tag{3.27}$$

(b) *If*

$$\beta_N^2(x) := \;\frac{1}{N} K_N(x,x) \;=\; \frac{1}{N}\sum_{r=0}^{N-1} t_r^2(x) \tag{3.28}$$

and we denote $x = \cos\theta$ then, for $N \to \infty$,

$$\begin{aligned} \beta_N^2(x) &= 1 + \frac{1}{N}\frac{\cos(N\theta)\;\sin((N-1)\theta)}{\sin\theta} \\ &= 1 + \frac{O(1)}{N}, \end{aligned} \tag{3.29}$$

uniformly on any $[a,b] \subset (-1,1)$, and $\beta_N^2(x) = 2$ for $x = \pm 1$.

[3]Given the parallel between our work with Legendre and Chebyshev polynomials we decided to duplicate some of the notations, e.g. x_k^n, the inner products, the projection operator K_N, etc. The reader must just keep in mind whether we are working under the Chebyshev or the Legendre setting.

Remark 3.2 *Note the slightly different behavior at the end-points when compared with the inner points. Compare this with Lemma 3.1.*

The second property is illustrated in Figure 3.2.

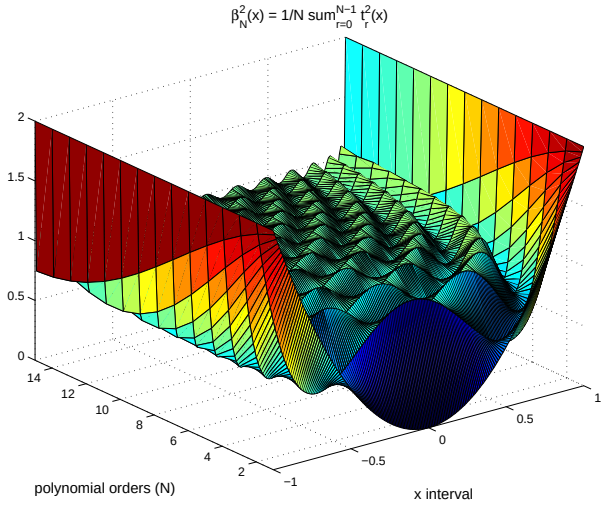

Figure 3.2: Averaged sum of squared Chebyshev polynomials

Proof. (a) This is a consequence of the double orthogonality property of the trigonometric Fourier basis (cf. e.g. Gradshtein and Ryzhik, f. 1.351(1), p. 30).

(b) This is a classical identity (cf. e.g. Gradshtein and Ryzhik, f. 1.351(2), p. 31); compare with the proof of Lemma 3.1.

□

In the following section we will discuss the use of the Legendre and Chebyshev polynomials in constructing pointwise asymptotically minimax estimators for analytic functions, in the non-adaptive (known γ, M) setting.

We shall see, in particular, that the best achievable rate of convergence at the end-points using the Chebyshev design is faster than that in the case of the Legendre design. Here we have only considered and compared two most important designs: one which is often appears to be the natural choice – the equidistant design, and one which is actually more preferable – the Chebyshev design. There are of course many others designs; their importance and a more comprehensive study has only started recently, partly as a result of the study presented in this chapter.

In Section 3.3 we shall restrict our study to Chebyshev designs, in constructing minimax estimator in the adaptive (unknown γ, M) setting. Statistical estimation using the uniform

norm as the quality criterion of estimators requires a different approach (cf. Golubev, Lepski and Levit [2001]).

3.2 Minimax regression in $\mathcal{A}(\gamma, M)$

3.2.1 The statistical setting

Our observation model in this chapter is given by

$$y_k = f(x_k^n) + \xi_k, \quad k = 1, \dots, n, \tag{3.30}$$

where the random variables ξ_k are independent identically distributed $\mathcal{N}(0, \sigma^2)$ and the design x_k^n is either Legendre and Chebyshev design. Throughout this chapter the unknown regression function f belongs to $\mathcal{A}(\gamma, M)$. In this section we assume that the parameters γ and M which determine the class are fixed and known to the statistician. We prove that it is possible, asymptotically, to have as good minimax risk using projection-type estimators based on the Legendre-Fourier and Chebyshev-Fourier series, for their respective designs, as with any other estimator.

Let $\mathcal{W}$ be the class of loss functions $w : \mathbb{R} \to \mathbb{R}^+$ such that

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

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

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

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

Let $\tilde{f}_n(x) = \tilde{f}_n(x, \mathbf{y})$ be an arbitrary estimator of $f(x)$ based on the observation vector $\mathbf{y} = (y_1, \dots, y_n)$, and denote by $\mathbf{P}_f$, $\mathbf{E}_f$ and $\mathbf{Var}_f$ the distribution, the expectation and the variance corresponding to f. Sometimes the sub-index f will be dropped, when there is no possibility of confusion.

Our main interest will be in the asymptotic behavior of the minimax risk

$$\inf_{\tilde{f}_n} \sup_{f \in \mathcal{A}(\gamma, M)} \mathbf{E}_f\, w\Big(\sigma_n^{-1}\big(\tilde{f}_n(x) - f(x)\big)\Big)$$

where $w \in \mathcal{W}$. The parameter σ_n defining the minimax rate of convergence, for each of the corresponding designs, Legendre or Chebyshev, will be specified later in Theorems 3.1 and 3.2.

3.2.2 Estimation in the Legendre design

Given the observations $\mathbf{y}$ taken at the Legendre knots (3.9), and following the notation introduced in Section 3.1.2, define the estimator

$$\hat{f}_{n,N}(x) \;=\; \frac{1}{n}\sum_{k=1}^{n} y_k K_N(x, x_k^n) \;=\; \sum_{r=0}^{N-1} \left(\frac{1}{n}\sum_{k=1}^{n} y_k p_r(x_k^n)\right) p_r(x). \tag{3.31}$$

With a slight abuse of the notation, we will write

$$\hat{f}_{n,N}(x) \;=\; (\mathbf{y} \,|\, K_N(x,\cdot)) \;=\; \sum_{r=0}^{N-1} (\mathbf{y}|p_r)\, p_r(x). \tag{3.32}$$

Now consider two auxiliary functions:

$$f_N(x) \;=\; \langle f \,|\, K_N(x,\cdot)\rangle \;=\; \sum_{r=0}^{N-1} \langle f \,|\, p_r\rangle\, p_r(x), \tag{3.33}$$

and

$$f_{n,N}(x) \;=\; (f \,|\, K_N(x,\cdot)) \;=\; \sum_{r=0}^{N-1} (f|p_r)\, p_r(x). \tag{3.34}$$

Notice that the projection-type estimator $\hat{f}_{n,N}(x)$ is an unbiased estimator of the finite expansion term $f_{n,N}(x)$ which, in turn, approximates the sum f_N of the first N terms of the Legendre-Fourier series.

The following theorem holds.

Theorem 3.1 *For any $w \in \mathcal{W}$ and every $x \in [-1,1]$*

$$\lim_{n\to\infty} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\, w\left(\alpha_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\,\big(\hat{f}_n(x) - f(x)\big)\right)$$

$$= \lim_{n\to\infty} \inf_{\tilde{f}_n} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\, w\left(\alpha_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\,\big(\tilde{f}_n(x) - f(x)\big)\right) \;=\; \mathbf{E}\, w(\xi)$$

where $\alpha_N(x)$ is defined in (3.12), $\tilde{f}_n$ is an arbitrary estimator of f, $\hat{f}_n = \hat{f}_{n,N}$ is the projection estimator (3.32) with

$$N = N_n := \left\lfloor \frac{1}{2\gamma} \log n \right\rfloor \qquad \textit{and} \qquad \xi \sim \mathcal{N}(0,1). \tag{3.35}$$

Proof: the upper bound. Let N be given by (3.35). As usual we decompose the mean square error as

$$\mathbf{E}\big(\hat{f}_{n,N}(x) - f(x)\big)^2 = \mathbf{Var}\, v_N^2(x) + b_N^2(x) \tag{3.36}$$

where, according to (3.32) and (3.34),

$$v_N(x) \;=\; \hat{f}_{n,N}(x) - f_{n,N}(x) \;=\; \frac{1}{n}\sum_{k=1}^{n} \xi_k K_N(x, x_k^n) \tag{3.37}$$

is a zero-mean stochastic term and

$$b_N(x) \;=\; \big(f_{n,N}(x) - f_N(x)\big) \;+\; \big(f_N(x) - f(x)\big) \tag{3.38}$$

is the bias.

Let us first analyze the variance of $v_N(x)$. Applying Lemma 3.1(a) we get

$$\begin{aligned}
\mathbf{Var}\, v_N(x) &= \frac{\sigma^2}{n^2}\sum_{k=1}^{n} K_N^2(x,x_k^n) \;=\; \frac{\sigma^2}{n^2}\sum_{k=1}^{n}\left(\sum_{r=0}^{N-1} p_r(x)p_r(x_k^n)\right)^2 \\
&= \frac{\sigma^2}{n}\sum_{r_1=0}^{N-1}\sum_{r_2=0}^{N-1} p_{r_1}(x)p_{r_2}(x)\,\frac{1}{n}\sum_{k=1}^{n} p_{r_1}(x_k^n)p_{r_2}(x_k^n) \\
&= \frac{\sigma^2}{n}\sum_{r_1=0}^{N-1}\sum_{r_2=0}^{N-1} p_{r_1}(x)p_{r_2}(x)\left(\delta r_1 r_2 + O\left(\frac{N^6}{n^3}\right)\right) \\
&= \frac{\sigma^2}{n}\sum_{r=0}^{N-1} p_r^2(x) + O\left(\frac{N^6}{n^3}\right)\sum_{r_1=0}^{N-1}\sum_{r_2=0}^{N-1} p_{r_1}(x)p_{r_2}(x). \qquad (3.39)
\end{aligned}$$

Now, applying the Cauchy-Schwartz inequality we see that

$$\begin{aligned}
\left|\sum_{r_1=0}^{N-1}\sum_{r_2=0}^{N-1} p_{r_1}(x)p_{r_2}(x)\right| &= \left(\sum_{r=0}^{N-1} p_r(x)\right)^2 \;\le\; N\sum_{r=0}^{N-1} p_r^2(x) \\
&= N\,K_N(x,x) \;=\; N^2\,\alpha_N^2(x). \qquad (3.40)
\end{aligned}$$

Thus, according to the last two equations and (3.35),

$$\mathbf{Var}\, v_N(x) \;=\; \alpha_N^2(x)\,\frac{\sigma^2 N}{n}\,(1+o(1)) \qquad (3.41)$$

for any $x \in [-1,1]$, as n goes to infinity.

Now let us consider the bias. First, we have

$$f_{n,N}(x) - f_N(x) = \sum_{r=0}^{N-1}\big((f\,|\,p_r) - \langle f\,|\,p_r\rangle\big)\,p_r(x). \qquad (3.42)$$

By definition

$$\big|\,(f\,|\,p_r) - \langle f\,|\,p_r\rangle\big| \;=\; \left|\frac{1}{2}\int_{-1}^{1} f(x)p_r(x)\,dx - \frac{1}{n}\sum_{k=1}^{n} f(x_k^n)p_r(x_k^n)\right|. \qquad (3.43)$$

Next, applying (3.14), this difference can be bounded by

$$\frac{1}{3n^2}\max_{x\in[-1,1]}\left|\frac{d^2}{dx^2}f(x)p_r(x)\right|. \qquad (3.44)$$

Thus, applying the bounds for the derivatives $|f^{(m)}(x)| \le Mm!/\rho_\gamma^m$ (cf. Sect. 3.1.1) and $|p_r^{(m)}(x)| \le (2r+1)^{1/2} r^{2m}$ (cf. eq. (3.8)), it follows that

$$\begin{aligned} \big|\,(f\,|\,p_r) - \langle f\,|\,p_r\rangle\big| &\le \frac{M}{3n^2}\,(\,(2\rho_\gamma)^{-2} + 2\rho_\gamma^{-1}(2r+1)^{1/2}r^2 + (2r+1)^{1/2}r^4) \\ &= O\Big(\frac{r^5}{n^2}\Big), \qquad (n\to\infty). \end{aligned} \tag{3.45}$$

Combining Cauchy-Schwartz inequality with the previous bound and using the fact that N is of order $O(\log n)$, cf. eq. (3.35), we find

$$\begin{aligned} (\,f_{n,N}(x) - f_N(x)\,)^2 &\le \sum_{r=0}^{N-1}\big((f\,|\,p_r) - \langle f\,|\,p_r\rangle\big)^2 \sum_{r=0}^{N-1} p_r^2(x) \\ &= \alpha_N^2(x)\,N\sum_{r=0}^{N-1}\big((f\,|\,p_r) - \langle f\,|\,p_r\rangle\big)^2 \;=\; \alpha_N^2(x)\,O\Big(\frac{N^{12}}{n^4}\Big) \\ &= \alpha_N^2(x)\frac{\sigma^2 N}{n}\,O\Big(\frac{N^{11}}{n^3}\Big) \;=\; o(1)\,\mathbf{Var}v_N(x). \end{aligned} \tag{3.46}$$

As demonstrated in Ibragimov and Has'minskii [1981], for functions $f \in \mathcal{A}(\gamma, M)$

$$|\langle f\,|\,p_r\rangle| \le C_1\,e^{-\gamma r}$$

for some constant $C_1 > 0$. According to the Laplace formula (3.16) the polynomials $p_r(x)$ are uniformly bounded, on any interval $[a,b] \subset (-1,1)$. Thus, from previous inequality, for some $C_2 > 0$,

$$\begin{aligned} \big(f_N(x) - f(x)\big)^2 &\le \Big(\sum_{r=N}^{\infty}|\langle f\,|\,p_r\rangle|\,|p_r(x)|\Big)^2 \\ &\le C_2\,e^{-2\gamma N} \;\sim\; C_2 n^{-1} \;=\; o(1)\,\mathbf{Var}\,v_N(x). \end{aligned} \tag{3.47}$$

At the end-points of the interval we have $|p_r(\pm 1)| = (2r+1)^{1/2}$, see eqs. (3.4) and (3.7), thus for $x = \pm 1$

$$\begin{aligned} |f_N(x) - f(x)| &\le C_1\sum_{r=N}^{\infty}(2r+1)^{1/2}e^{-\gamma r} \;\le\; C_3\sum_{r=N+1}^{\infty} r^{1/2}e^{-\gamma r} \\ &\le C_3\,e^{\gamma}\int_{N+1}^{\infty} r^{1/2}e^{-\gamma r}dr \;=\; C_3\,N^{1/2}e^{-\gamma N}(\,1+o(1)\,) \end{aligned}$$

as $N\to\infty$. Therefore for some $C_4 > 0$ and N large enough

$$\big(f_N(x) - f(x)\big)^2 \le C_4\,Ne^{-2\gamma N} \;\sim\; C_4\frac{N}{n} \;=\; o(1)\,\mathbf{Var}\,v_N(x). \tag{3.48}$$

From (3.36), (3.41), (3.46) and (3.47) or (3.48) we can conclude that

$$\mathbf{E}\big(\hat{f}_{n,N}(x) - f(x)\big)^2 = \alpha_N^2(x)\,\frac{\sigma^2 N}{n}(\,1 + o(1)\,),$$

uniformly on $[-1,1]$. It follows that

$$\alpha_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\big(\hat{f}_{n,N_n}(x) - f(x)\big)$$

is normally distributed with mean of order $o(1)$ and variance equal to $1 + o(1)$, when n goes to infinity, uniformly with respect to $f \in \mathcal{A}(\gamma, M)$. Therefore using the dominated convergence theorem we obtain the following upper bound:

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\, w\left(\alpha_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\,\big(\hat{f}_n(x) - f(x)\big)\right) \;=\; \mathbf{E}\, w(\xi). \tag{3.49}$$

Proof of the lower bound for the risk. For fixed $x \in [-1,1]$ and any $z \in \mathbb{C}$ consider the following parametric sub-family of functions

$$f_\theta(z) = \theta\,\sqrt{\frac{\sigma^2}{n}}\,\frac{K_{\bar{N}}(x,z)}{\sqrt{K_{\bar{N}}(x,x)}}, \qquad |\theta| < \theta_n = \bar{N}^{1/2}, \tag{3.50}$$

where we will use

$$\bar{N} = \bar{N}_n = \lfloor N_n - 3\log N_n \rfloor\,, \tag{3.51}$$

see (3.35). Note that $\bar{N}$ is asymptotically equivalent to $N = N_n$ when $N \to \infty$. This implies, according to Lemma 3.1(b), that

$$\frac{\alpha_{\bar{N}}^2(x)}{\alpha_N^2(x)} \to 1, \tag{3.52}$$

uniformly in $[-1,1]$, when $n \to \infty$.

We need the following lemma.

Lemma 3.3 *For a given $x \in [-1,1]$ and any $z \in E_\gamma$, let $f_\theta(z)$ be defined by (3.50). Then*

(a) $f_\theta(x) = \theta\,\alpha_{\bar{N}}(x)\,\sqrt{\frac{\sigma^2\bar{N}}{n}}$.

(b) $f_\theta \in \mathcal{A}(\gamma, M)$, $\quad |\theta| < \theta_n$, *for all n big enough.*

(c) *The statistic*

$$T = \frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n} y_k\,\frac{K_{\bar{N}}(x, x_k^n)}{\sqrt{K_{\bar{N}}(x,x)}}$$

has a normal distribution $\mathcal{N}(\theta\mathcal{I}_n, \mathcal{I}_n)$ under f_θ, where $\mathcal{I}_n = 1 + o(1)$.

(d) *The statistic T is sufficient and the log-likelihood $\Lambda := \log \frac{d\mathbf{P}_\theta}{d\mathbf{P}_0}(\mathbf{y})$ satisfies*

$$\Lambda \;=\; \theta T - \frac{\theta^2}{2}\mathcal{I}_n$$

where $\mathbf{P}_\theta$ and $\mathbf{P}_0$ denote the probabilities associated with f_θ and f_0 respectively.

Proof of lemma.

(a) This follows directly from the definitions of f_θ and $\alpha_{\bar N}(x)$.

(b) Obviously $f_\theta(z)$ is analytic in the whole complex plane, thus also in E_γ. Using (3.17), applying the Cauchy-Schwartz inequality and recalling the definition of $\bar N = \bar N_n$, we obtain

$$|f_\theta(z)| \;\le\; \theta_{\bar N}\sqrt{\frac{\sigma^2}{n}}\left(\frac{K^2_{\bar N}(x,z)}{K_{\bar N}(x,x)}\right)^{1/2} \le \sqrt{\frac{\sigma^2\bar N}{n}}K^{1/2}_{\bar N}(z,z) \;=\; \sqrt{\frac{\sigma^2\bar N}{n}}\left(\sum_{r=0}^{\bar N-1} p_r^2(z)\right)^{1/2}$$

$$\le \sqrt{\frac{\sigma^2\bar N}{n}}\left(\sum_{r=0}^{\bar N-1}(2r+1)e^{2\gamma r}\right)^{1/2} \;=\; O(1)\frac{\bar N}{\sqrt{n}}e^{\gamma\bar N} = O(\bar N^{-1/2}) \;\le\; M,$$

in E_γ for all n large enough.

(c) Denote

$$\mathcal{I}_n \;=\; \frac{1}{n}\sum_{k=1}^{n}\frac{K^2_{\bar N}(x,x_k^n)}{K_{\bar N}(x,x)}.$$

We can see that T is normally distributed,

$$\mathbf{E}\,T \;=\; \frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n} f_\theta(x_k^n)\,\frac{K_{\bar N}(x,x_k^n)}{\sqrt{K_{\bar N}(x,x)}} \;=\; \theta\,\frac{1}{n}\sum_{k=1}^{n}\frac{K^2_{\bar N}(x,x_k^n)}{K_{\bar N}(x,x)} \;=\; \theta\mathcal{I}_n, \qquad \text{and}$$

$$\mathbf{Var}\,T \;=\; \frac{1}{n}\sum_{k=1}^{n}\frac{K^2_{\bar N}(x,x_k^n)}{K_{\bar N}(x,x)} \;=\; \mathcal{I}_n.$$

Thus $T \sim \mathcal{N}(\theta\mathcal{I}_n, \mathcal{I}_n)$. Now let us show that $\mathcal{I}_n \to 1$ when $n \to \infty$. Using Lemma 3.1(a)

and the Cauchy-Schwartz inequality, we find obtains

$$\begin{aligned}
\mathcal{I}_n &= \frac{1}{n} K_{\bar{N}}^{-1}(x,x) \sum_{k=1}^{n} K_{\bar{N}}^2(x,x_k^n) \;=\; \frac{1}{n} K_{\bar{N}}^{-1}(x,x) \sum_{k=1}^{n} \left(\sum_{r=0}^{\bar{N}-1} p_r(x) p_r(x_k^n) \right)^2 \\
&= K_{\bar{N}}^{-1}(x,x) \sum_{r_1=0}^{\bar{N}-1} \sum_{r_2=0}^{\bar{N}-1} \left(p_{r_1}(x) p_{r_2}(x) \frac{1}{n} \sum_{k=1}^{n} p_{r_1}(x_k^n) p_{r_2}(x_k^n) \right) \\
&= K_{\bar{N}}^{-1}(x,x) \sum_{r_1=0}^{\bar{N}-1} \sum_{r_2=0}^{\bar{N}-1} \left(p_{r_1}(x) p_{r_2}(x) \left(\delta_{r_1 r_2} + O\left(\frac{N^6}{n^2} \right) \right) \right) \\
&= K_{\bar{N}}^{-1}(x,x) \sum_{r=0}^{\bar{N}-1} p_r^2(x) + O\left(\frac{\bar{N}^6}{n^2} \right) K_{\bar{N}}^{-1}(x,x) \sum_{r_1=0}^{\bar{N}-1} \sum_{r_2=0}^{\bar{N}-1} p_{r_1}(x) p_{r_2}(x) \\
&= 1 + O\left(\frac{\bar{N}^6}{n^2} \right) K_{\bar{N}}^{-1}(x,x) \left(\sum_{r=0}^{\bar{N}-1} p_r(x) \right)^2 \\
&= 1 + o(1), \qquad\qquad (n \to \infty). \qquad\qquad (3.53)
\end{aligned}$$

(d) It is easy to see that the log-likelihood

$$\begin{aligned}
\Lambda &= \log \prod_{k=0}^{n-1} \exp\left\{ -\frac{1}{2\sigma^2} \big(y_k - f_\theta(x_k^n) \big)^2 + \frac{1}{2\sigma^2} y_k^2 \right\} \\
&= -\frac{1}{2\sigma^2} \sum_{k=1}^{n} (y_k - f_\theta(x_k^n))^2 + \frac{1}{2\sigma^2} \sum_{k=1}^{n} y_k^2 \\
&= \theta \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} y_k \frac{K_{\bar{N}}(x,x_k^n)}{\sqrt{K_{\bar{N}}(x,x)}} - \frac{\theta^2}{2n} \sum_{k=1}^{n} \frac{K_{\bar{N}}^2(x,x_k^n)}{K_{\bar{N}}(x,x)} \\
&= \theta\, T - \frac{\theta^2}{2} \mathcal{I}_n.
\end{aligned}$$

This completes the proof of the lemma. □

Now we can continue the proof of the theorem. Given $\alpha^2_{\bar{N}}(x) \sim \alpha^2_N(x)$, see eq. (3.52),

$$\mathcal{R} := \inf_{\tilde{f}_n} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\, w\left(\alpha^{-1}_N(x)\sqrt{\frac{n}{\sigma^2 N}}\big(\tilde{f}_n(x) - f(x)\big)\right) \tag{3.54}$$

$$= \inf_{\tilde{f}_n} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\, w\left(\alpha^{-1}_{\bar{N}}(x)\sqrt{\frac{n}{\sigma^2 \bar{N}}}\big(\tilde{f}_n(x) - f(x)\big)(1+o(1))\right) \tag{3.55}$$

$$\geq \inf_{\tilde{f}_n} \sup_{f_\theta} \mathbf{E}_{f_\theta}\, w\left((1+o(1))\alpha^{-1}_{\bar{N}}(x)\sqrt{\frac{n}{\sigma^2 \bar{N}}}\big(\tilde{f}_n(x) - f_\theta(x)\big)\right), \quad (\bar{N}\to\infty).$$

Denote $\tilde{\theta} = \alpha^{-1}_{\bar{N}}(x)\sqrt{\frac{n}{\sigma^2 N}}\tilde{f}_n(x)$. Then applying Lemma 3.3(a)

$$\mathcal{R} \geq \inf_{\tilde{\theta}} \sup_{|\theta|\leq\theta_n} \mathbf{E}_\theta\, w\Big((\tilde{\theta}-\theta)(1+o(1))\Big), \qquad (n\to\infty).$$

Since $|\theta| \leq \theta_n$, we can restrict ourselves exclusively to estimators such that $|\tilde{\theta}| \leq \theta_n$; otherwise trimming $\tilde{\theta}$, at an appropriate level, will produce a smaller risk. For such estimators $|\tilde{\theta}-\theta| \leq 2\theta_n$. Now, from equations (3.54) and (3.55), applying Lemma 3.1(b) and definition (3.51) of $\bar{N}$ we can verify that the term $o(1)$ in the previous equation is of order $(\log N)/N$. Thus $\theta_n o(1) \to 0$ and therefore the previously mentioned estimators satisfy $|\tilde{\theta}-\theta|o(1) \to 0$. Hence

$$\mathcal{R} \geq \inf_{\tilde{\theta}} \sup_{|\theta|\leq\theta_n} \mathbf{E}_\theta\, w\Big((\tilde{\theta}-\theta)+o(1)\Big), \qquad (n\to\infty).$$

We can approximate any loss function $w \in \mathcal{W}$, by a sequence of bounded uniformly continuous functions $w_\delta \in \mathcal{W}$ such that $w_\delta \nearrow w$ when $\delta \to 0$ and see that for any δ

$$\mathcal{R} \geq \inf_{\tilde{\theta}} \sup_{|\theta|\leq\theta_n} \mathbf{E}_\theta\, w_\delta\, \Big((\tilde{\theta}-\theta)+o(1)\Big) = \inf_{\tilde{\theta}} \sup_{|\theta|\leq\theta_n} \mathbf{E}_\theta\, w_\delta\,(\tilde{\theta}-\theta) + o(1).$$

Now let us fix an arbitrary prior density λ on $(-\theta_n, \theta_n)$ with a finite Fisher information $I(\lambda)$. Then

$$\inf_{\tilde{\theta}} \sup_{|\theta|\leq\theta_n} \mathbf{E}_\theta\, w_\delta\,(\tilde{\theta}-\theta) \geq \inf_{\tilde{\theta}} \int_{-\theta_n}^{\theta_n} \mathbf{E}_\theta\, w_\delta\,(\tilde{\theta}-\theta)\lambda(\theta)d\theta$$

$$= \inf_{\tilde{\theta}(T)} \int_{-\theta_n}^{\theta_n} \mathbf{E}_\theta\, w_\delta\,(\tilde{\theta}(T)-\theta)\lambda(\theta)d\theta$$

given that T is sufficient for θ, according to Lemma 3.3(c). Applying results presented in Levit [1980], we get that

$$\inf_{\tilde{\theta}} \sup_{|\theta|\leq\theta_n} \mathbf{E}_\theta\, w_\delta\,(\tilde{\theta}-\theta) \geq \mathbf{E}\; w_\delta(\xi) + O(\theta_n^{-2}), \qquad (n\to\infty),$$

where $\xi \sim \mathcal{N}(0,1)$. Thus $\liminf_{n\to\infty} \mathcal{R} \ge \mathbf{E}\, w_\delta(\xi)$. Applying the dominate convergence theorem for $\delta \to 0$ we get

$$\liminf_{n\to\infty} \inf_{\tilde f_n} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\ w\left(\alpha_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\big(\tilde f_n(x) - f(x)\big)\right) \ge \mathbf{E}\, w(\xi). \tag{3.56}$$

Finally, from (3.49) and (3.56) the theorem is proved. □

Corollary 3.1 *For any $[a,b] \subset (-1,1)$, uniformly in $x \in [a,b]$,*

$$\lim_{n\to\infty} \inf_{\tilde f_n} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}\ w\left(\sqrt{(1-x^2)^{1/2}\frac{\pi n}{\sigma^2 N_n}}\ \big(\tilde f_n(x) - f(x)\big)\right)$$

$$= \lim_{n\to\infty} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}\ w\left(\sqrt{(1-x^2)^{1/2}\frac{\pi n}{\sigma^2 N_n}}\big(\hat f_n(x) - f(x)\big)\right) = \mathbf{E}\, w(\xi)$$

where $\tilde f_n$ and $\hat f_n$ are as in the Theorem 3.1. For $x = \pm 1$,

$$\lim_{n\to\infty} \inf_{\tilde f_n} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\ w\left(\sqrt{\frac{n}{\sigma^2 N_n^2}}\big(\tilde f_n(x) - f(x)\big)\right)$$

$$= \lim_{n\to\infty} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\ w\left(\sqrt{\frac{n}{\sigma^2 N_n^2}}\big(\hat f_n(x) - f(x)\big)\right) = \mathbf{E}\, w(\xi).$$

3.2.3 Estimation in the Chebyshev design

Consider now the design given by the Chebyshev knots (3.26). Following the notation of Section 3.1.3 define the estimator

$$\hat f_{n,N}(x) \ = \ \frac{1}{n}\sum_{k=1}^{n} y_k K_N(x, x_k^n) \ = \ \sum_{r=0}^{N-1}\left(\frac{1}{n}\sum_{k=1}^{n} y_k t_r(x_k^n)\right) t_r(x). \tag{3.57}$$

As before, we will write, with a slight abuse of the notation

$$\hat f_{n,N}(x) \ = \ (\mathbf{y}\,|\,K_N(x,\cdot)) \ = \ \sum_{r=0}^{N-1} (\mathbf{y}|t_r)\, t_r(x), \tag{3.58}$$

and consider the two functions

$$f_N(x) \ = \ \langle f\,|\,K_N(x,\cdot)\rangle \ = \ \sum_{r=0}^{N-1} \langle f\,|\,t_r\rangle\, t_r(x), \tag{3.59}$$

and

$$f_{n,N}(x) \ = \ (f\,|\,K_N(x,\cdot)) \ = \ \sum_{r=0}^{N-1} (f|t_r)\, t_r(x); \tag{3.60}$$

see the footnote on page 50 with regards to these notations. Then the following result holds.

Theorem 3.2 *For any $w \in \mathcal{W}$ and every $x \in [-1, 1]$*

$$\lim_{n\to\infty} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\, w\left(\beta_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\,\big(\hat{f}_n(x) - f(x)\big)\right) \;=$$

$$\lim_{n\to\infty} \inf_{\tilde{f}_n} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\, w\left(\beta_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\,\big(\tilde{f}_n(x) - f(x)\big)\right) \;=\; \mathbf{E}\, w(\xi)$$

where $\tilde{f}_n$ is an arbitrary estimator of f, $\hat{f}_n = \hat{f}_{n,N}$ is the projection estimator (3.58) with

$$N = N_n := \left\lfloor \frac{1}{2\gamma} \log\big(M^2\,\gamma\,(1 - e^{-\gamma})^{-2} n\big) \right\rfloor, \tag{3.61}$$

$\beta_N^2(x)$ is defined by (3.28) and $\xi \sim \mathcal{N}(0, 1)$.

Remark 3.3 *Note that $\beta_N^2(x)$ plays the same role in the present context of estimation using Chebyshev design as played by $\alpha_N^2(x)$ in the previous Legendre case.*

Proof: the upper bound. The proof of this theorem is similar to the proof of the equivalent result for Legendre polynomials, Theorem 3.1. However, notice that in the case of Chebyshev polynomials we have exact orthogonality, and not just asymptotic orthogonality, as for the Legendre polynomials; compare the Lemmas 3.1(a) and 3.2(a). This will make some computations more straightforward. Some steps in this proof will be presented somewhat differently; we will keep track of the dependency in the variance and the bias on the parameters of the class, γ and M. This will be used in the next section for adaptive estimation.

Let $N \in \mathbb{N}$. Applying the same decomposition as in Theorem 3.1, cf. (3.37) and (3.38), we have

$$\mathbf{E}(\hat{f}_{n,N}(x) - f(x))^2 = \mathbf{Var}\, v_N^2(x) + b_N^2(x). \tag{3.62}$$

Let us first analyze the variance of $v_N(x)$. As before (cf. eq. (3.39)), applying Lemma 3.2(a) we obtain

$$\mathbf{Var}\, v_N(x) \;=\; \frac{\sigma^2}{n} \sum_{r_1=0}^{N-1} \sum_{r_2=0}^{N-1} t_{r_1}(x) t_{r_2}(x) \delta r_1 r_2 \;=\; \beta_N^2(x)\,\frac{\sigma^2 N}{n} \tag{3.63}$$

for any $x \in [-1, 1]$.

Now let us consider the bias

$$b_N(x) \;=\; \big(f_{n,N}(x) - f_N(x)\big) \;+\; \big(f_N(x) - f(x)\big). \tag{3.64}$$

Using Cauchy-Schwartz inequality we see that

$$\begin{aligned} \big(f_{n,N}(x) - f_N(x)\big)^2 \;&\le\; \sum_{r=0}^{N-1} \big((f \,|\, t_r) - \langle f \,|\, t_r\rangle\big)^2 \sum_{r=0}^{N-1} t_r^2(x) \\ &=\; N\,\beta_N^2(x) \sum_{r=0}^{N-1} \big((f \,|\, t_r) - \langle f \,|\, t_r\rangle\big)^2. \end{aligned} \tag{3.65}$$

If we rewrite the inner products as

$$(f \,|\, t_r) = \frac{1}{\pi} \sum_{k=1}^{n} f\left(\cos(k-1/2)\frac{\pi}{n}\right) \cos\left(r(k-1/2)\frac{\pi}{n}\right) \frac{\pi}{n}$$

and

$$\langle f \,|\, t_r \rangle = \frac{1}{\pi} \int_0^{\pi} f(\cos\zeta)\cos(r\zeta) d\zeta$$

(cf. eqs. (3.19) and (3.25)), we can apply the same arguments that we used in (3.43)–(3.45). Using the bounds for the derivatives of f given in eq. (3.1) we find that

$$\begin{aligned} \left|(f \,|\, t_r) - \langle f \,|\, t_r \rangle\right| &\leq \frac{\pi}{24}\left(\frac{\pi}{n}\right)^2 \max_{\zeta} \left|\frac{d^2}{d\zeta^2} f(\cos\zeta)\cos(r\zeta)\right| \\ &\leq \frac{\pi^3}{24\, n^2} M \left(r^2 + \frac{(2r+1)}{\rho_\gamma} + \frac{2}{\rho_\gamma^2}\right) \\ &\leq \frac{\pi^3 (r+1)^2}{6\, n^2}\, M\, \max(1, \rho_\gamma^{-1}, \rho_\gamma^{-2}) \;=\; M C_\gamma \frac{(r+1)^2}{n^2} \end{aligned} \tag{3.66}$$

where, using (3.2), one can verify that

$$C_\gamma = O(1 - e^{-\gamma})^{-4}, \tag{3.67}$$

both at $\gamma = 0$ and $\gamma = \infty$ and it is bounded when γ is varying in compact subsets of $(0, \infty)$. Thus, both for $\gamma \to 0$ and for $\gamma \to \infty$, uniformly in N

$$\left(f_{n,N}(x) - f_N(x)\right)^2 \;=\; \beta_N^2(x)\, O\left(M^2 \left(1 - e^{-\gamma}\right)^{-8} \frac{N^6}{n^4}\right). \tag{3.68}$$

If we choose $N = N_n$

$$\left(f_{n,N}(x) - f_N(x)\right)^2 \;=\; o(1)\, \mathbf{Var} v_N(x), \qquad (n \to \infty). \tag{3.69}$$

In the previous section we saw that in order to bound the truncation error term $f_N(x) - f(x)$ it was necessary to consider separately two cases: $|x| < 1$ and $|x| = 1$ (cf. eqs. (3.47) and (3.48)). Now, one can see that both cases can be considered simultaneously. From (3.23) one can see that for any x and $N = N_n$

$$\left(f_N(x) - f(x)\right)^2 \;\leq\; 2\pi M^2 (1 - e^{-\gamma})^{-2}\, e^{-2\gamma N} \;=\; O\Big(\frac{1}{\gamma n}\Big) \tag{3.70}$$

$$= \beta_N^2(x) \frac{\sigma^2 N}{n} O\Big(\frac{1}{\gamma N}\Big) \;=\; o(1)\, \mathbf{Var}\, v_N(x), \tag{3.71}$$

when $n \to \infty$. From (3.62)–(3.64), (3.69) and (3.71) we have proved that

$$\mathbf{E}\big(\hat{f}_{n,N}(x) - f(x)\big)^2 = \beta_N^2(x)\,\frac{\sigma^2 N}{n}(1 + o(1)), \quad (n \to \infty),$$

which holds uniformly on $[-1, 1]$. It follows that

$$\beta_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\big(\hat{f}_{n,N}(x) - f(x)\big)$$

is normally distributed with mean of order $o(1)$ and variance equal $1 + o(1)$, $n \to \infty$, uniformly with respect to $f \in \mathcal{A}(\gamma, M)$. Therefore using the dominated convergence theorem we obtain the upper bound:

$$\lim_{n\to\infty} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\, w\left(\beta_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\,\big(\hat{f}_n(x) - f(x)\big)\right) \;=\; \mathbf{E}\, w(\xi). \tag{3.72}$$

Proof of the lower bound for the risk. We can follow the same proof of the lower bound we did in Theorem 3.1. For fixed $x \in [-1, 1]$ and any $z \in \mathbb{C}$ consider again the parametric sub-family of functions

$$f_\theta(z) = \theta\sqrt{\frac{\sigma^2}{n}}\,\frac{K_{\bar{N}}(x, z)}{\sqrt{K_{\bar{N}}(x, x)}} \qquad |\theta| < \theta_n = \bar{N}^{1/2} \tag{3.73}$$

where $K_{\bar{N}}$ is now defined in terms of the Chebyshev polynomials and

$$\bar{N} = \bar{N}_n = \lfloor N_n - 3\log N_n \rfloor \tag{3.74}$$

(cf. definition of N_n in eq. (3.61)).

Lemma 3.4 *The following properties are satisfied for any $x \in [-1, 1]$:*

(a) $f_\theta(x) = \theta\,\beta_{\bar{N}}(x)\,\sqrt{\frac{\sigma^2\bar{N}}{n}}$.

(b) $f_\theta \in \mathcal{A}(\gamma, M)$, $\quad |\theta| < \theta_n$, *for n big enough.*

(c) *The statistic*

$$T = \frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n} y_k\,\frac{K_{\bar{N}}(x, x_k^n)}{\sqrt{K_{\bar{N}}(x, x)}}$$

has the normal distribution $\mathcal{N}(\theta, 1)$ under f_θ, i.e. it can be represented as

$$T = \theta + \xi \tag{3.75}$$

where $\xi \sim \mathcal{N}(0, 1)$.

(d) *The statistic T is sufficient and the log-likelihood satisfies*

$$\Lambda \;:=\; \log\frac{d\mathbf{P}_\theta}{d\mathbf{P}_0} \;=\; \theta T - \frac{\theta^2}{2}. \tag{3.76}$$

where $\mathbf{P}_\theta$ and $\mathbf{P}_0$ denote the probabilities associated with f_θ and f_0 respectively.

Proof of the lemma. The proof is the same as that of Lemma 3.3. Nevertheless, a couple of remarks can be made. First, the bound (3.17) for Legendre polynomials is also a bound for the Chebyshev polynomials, thus the proof of (b) remaines the same. Second, in the present case, $\mathcal{I}_n = 1$ given exact orthogonality of Chebyshev polynomials (cf. eq. (3.53)). The rest of the proofs of the lemma and the theorem remain the same and we get

$$\liminf_{n\to\infty}\ \inf_{\tilde f_n}\ \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\ w\left(\beta_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\big(\tilde f_n(x) - f(x)\big)\right) \ge \mathbf{E}\, w(\xi). \tag{3.77}$$

The theorem follows from (3.72) and (3.77). □

Corollary 3.2 *For any $[a,b] \subset (-1,1)$ uniformly in $x \in [a,b]$*

$$\lim_{n\to\infty}\ \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\ w\left(\sqrt{\frac{n}{\sigma^2 N_n}}\big(\hat f_n(x) - f(x)\big)\right) \;=$$

$$\lim_{n\to\infty}\ \inf_{\tilde f_n}\ \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\ w\left(\sqrt{\frac{n}{\sigma^2 N_n}}\big(\tilde f_n(x) - f(x)\big)\right) \;= \mathbf{E}\, w(\xi)$$

where $\tilde f_n$ and $\hat f_n$ are as in the previous Thorem. For $x = \pm 1$,

$$\lim_{n\to\infty}\ \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\ w\left(\sqrt{\frac{n}{2\sigma^2 N_n}}\big(\hat f_n(x) - f(x)\big)\right) \;=$$

$$\lim_{n\to\infty}\ \inf_{\tilde f_n}\ \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\ w\left(\sqrt{\frac{n}{2\sigma^2 N_n}}\big(\tilde f_n(x) - f(x)\big)\right) \;= \mathbf{E}\, w(\xi)$$

Till now we have proved, first, that the polynomial estimators we proposed, with the order of polynomials adequately chosen, are asymptotically minimax for fixed classes $\mathcal{A}(\gamma, M)$. Secondly, we have seen that the optimal rate of convergence may be affected by the chosen design; in particular the rate of convergence at the end-points of the interval is worse for the Legendre design as compared to the Chebyshev design. For that reason, we will restrict ourselves to the study of the regression problem on a bounded interval under the Chebyshev design. In the next subsection we will make necessary steps towards the adaptive framework.

3.2.4 Estimation for non-fixed classes

In order to create an adaptive framework we follow the same procedure as in the previous chapter. This procedure is based on the ideas introduced in Lepski and Levit [1998]. The basic underlying idea is to allow the parameters of the model – in our case γ and M – take values from the broadest possible set, pushed to its 'limits'. Such 'limits' can be taken to be the extreme values for which either there is no consistency or, on the other hand, a parametric rate $O(n^{-1})$ is possible. Since in both cases these extreme values are not some fixed values $(\gamma^{extr}, M^{extr})$, but rather should be thought as some sequences $(\gamma_n^{extr}, M_n^{extr})$, our first step towards the adaptive framework will be to look for corresponding results in the situation where the parameters of the model, though known, are allowed to depend on n.

Thus we will assume in this subsection that although the parameters $\gamma = \gamma_n > 0$ and $M = M_n > 0$ are still known, they may depend on the number of observations n. As we saw in the previous chapter, this is not yet a proper adaptive framework. However it will allow us to explore the 'limits' of the model if the parameters have more freedom. Let N_n be as it was defined in Theorem 3.2. The dependence of N_n on n comes also from the parameters γ, M in the present situation. Nevertheless, the statement of Theorem 3.2 will still hold provided the appropriate assumptions are fulfilled.

Theorem 3.3 *Let $w \in \mathcal{W}$, $\gamma = \gamma_n$, $M = M_n$ and let $N = N_n$ be as defined in (3.61). If the following conditions are satisfied*

$$\lim_{n\to\infty} \gamma N = \infty, \tag{3.78}$$

$$\lim_{n\to\infty} M^2(1-e^{-\gamma})^{-8} N^5\, n^{-3} = 0, \tag{3.79}$$

$$\lim_{n\to\infty} N = \infty, \tag{3.80}$$

then

$$\lim_{n\to\infty} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\, w\left(\beta_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\,\left(\hat{f}_n(x) - f(x)\right)\right) \;=$$

$$\lim_{n\to\infty} \inf_{\tilde{f}_n} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\, w\left(\beta_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\,\left(\tilde{f}_n(x) - f(x)\right)\right) \;=\; \mathbf{E}\, w(\xi),$$

for all $x \in [-1, 1]$. Here $\tilde{f}_n$ is an arbitrary estimator of f and $\hat{f}_n = \hat{f}_{n,N}$ is the projection estimator (3.58).

Proof. Note that the previous conditions were automatically fulfilled in the case of fixed classes. The proof in the general case is similar to the proof of Theorem 3.2, and consists in checking that the conditions (3.78) and (3.79) guarantee asymptotic unbiasness of the optimal estimator (cf. eqs. (3.68) and (3.71)), while (3.80) allow us to prove the lower

bound result. The rest of the proof is the same. □

Though conditions (3.78)–(3.80) are sufficient to prove optimality results in non-fixed classes, it may be more convenient to express them explicitly in terms of the parameters γ and M, as is done in the following theorem.

Theorem 3.4 *Let $w \in \mathcal{W}$ and the parameters $\gamma = \gamma_n$ and $M = M_n$ be such that*

$$\limsup_{n\to\infty} \frac{M^2}{\log n} = 0, \tag{3.81}$$

$$\liminf_{n\to\infty} M^2 \log n = \infty, \tag{3.82}$$

$$\limsup_{n\to\infty} \frac{\gamma}{\log\log n} = 0, \tag{3.83}$$

$$\liminf_{n\to\infty} \gamma \log n = \infty, \tag{3.84}$$

then, with $N = N_n$ defined by (3.61),

$$\lim_{n\to\infty} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\, w\left(\beta_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\,\left(\hat{f}_n(x) - f(x)\right)\right) \;=$$

$$\lim_{n\to\infty} \inf_{\tilde{f}_n} \sup_{f\in\mathcal{A}(\gamma,M)} \mathbf{E}_f\, w\left(\beta_N^{-1}(x)\sqrt{\frac{n}{\sigma^2 N}}\,\left(\tilde{f}_n(x) - f(x)\right)\right) \;=\; \mathbf{E}\, w(\xi),$$

for all $x \in [-1, 1]$. Here $\tilde{f}_n$ is an arbitrary estimator of f and $\hat{f}_n = \hat{f}_{n,N}$ is the projection estimator (3.58).

Proof. In order to prove the theorem, we only need to verify that hypothesis of the Theorem 3.3 are satisfied, i.e. we just need to assure that the limits (3.78)–(3.80) are still valid (cf. eqs. (3.68) and (3.71)). If γ and M are bounded then trivially (3.78)–(3.80) hold. Let us consider the two extreme cases $\gamma \to 0$ and $\gamma \to \infty$. Remember that

$$N = N_n = \left\lfloor \frac{1}{2\gamma} \log\left(M^2\,\gamma(1 - e^{-\gamma})^{-2} n\right) \right\rfloor.$$

Case $\gamma \to 0$: Applying some asymptotics and conditions (3.82) and (3.83), we see that for n large enough

$$M^2\gamma(1 - e^{-\gamma})^{-2} n \;\sim\; M^2\gamma^{-1} n \;\geq\; \gamma^{-1}\log n \;\to\; \infty.$$

Thus γN and N go to infinity. Using (3.81) and (3.84)

$$\begin{aligned} M^2(1 - e^{-\gamma})^{-8} N^5 n^{-3} &= O\left(M^2\gamma^{-13} n^{-3} \log^5(M^2\gamma^{-1} n)\right) \\ &= O\left(n^{-3}\log^{14} n \log^5(n\log^2 n)\right) \;=\; o(1). \end{aligned}$$

Case $\gamma \to \infty$: Applying (3.82) and (3.83)

$$N \geq \frac{\log M^2 n}{2\gamma} \;=\; O\left(\frac{\log n}{\log\log n}\right) \to \infty, \quad (n \to \infty),$$

thus N and γN go to infinity. From (3.81) and (3.84)

$$\begin{aligned} M^2(1-e^{-\gamma})^{-8}N^5 n^{-3} &= O\left(M^2\gamma^{-5}n^{-3}\log^5(M^2\gamma n)\right) \\ &= O\left(n^{-3}\log n \log^5(n\log n)\right) \;=\; o(1), \quad (n\to\infty). \end{aligned}$$

Thus the theorem is proved. □

3.3 Adaptive minimax regression

3.3.1 Adaptive estimation in functional scales

In the previous section we described asymptotically minimax estimators for classes $\mathcal{A}(\gamma, M)$ where the parameters γ and M were known. However, in practice we do not usually know to which class the unknown function belongs, in other words we do not know the smoothness parameters. A data-dependent method for choosing an estimator in the presence of the unknown smoothness parameters is then necessary. We follow here the same procedure that we used in the previous chapter in order to create the adaptive framework in a situation where γ and M are unknown.

Let $\upsilon = (\gamma, M)$ where υ belongs to the region $\Gamma_n \subset \mathbb{R}^2_+$. Let $\mathcal{A}(\upsilon) = \mathcal{A}(\gamma, M)$ and define the functional scale $\mathcal{A}_{\Gamma_n}$,

$$\mathcal{A}_{\Gamma_n} := \Big\{ \mathcal{A}(\upsilon) \;\Big|\; \upsilon \in \Gamma_n \Big\},$$

corresponding to the *parameter class* Γ_n. As our scales $\mathcal{A}_{\Gamma_n}$ can be identified with corresponding subsets Γ_n, we will speak sometimes about a scale Γ_n, instead of $\mathcal{A}_{\Gamma_n}$, when there is no risk it could lead to a confusion.

From now on we will restrict ourselves to the loss functions $w(x) = |x|^p$, $p > 0$. Let $\mathcal{A}_{\Gamma_n}$ be a functional scale, and $\mathcal{F}$ a class of estimators $\tilde{f}_n$, both possibly depending on n.

Definition 3.2 *An estimator $\hat{f}_n \in \mathcal{F}$ is called $(p, \Gamma_n, \mathcal{F})$-adaptively minimax, at a point $x \in \mathbb{R}$, if for any other estimator $\tilde{f}_n \in \mathcal{F}$*

$$\limsup_{n\to\infty} \sup_{\upsilon\in\Gamma_n} \frac{\sup_{f\in\mathcal{A}(\upsilon)} \mathbf{E}_f |\hat{f}_n(x) - f(x)|^p}{\sup_{f\in\mathcal{A}(\upsilon)} \mathbf{E}_f |\tilde{f}_n(x) - f(x)|^p} \leq 1.$$

As it was discussed in the previous chapter, this property depends crucially on which classes Γ_n and $\mathcal{F}$ are considered. The rate of convergence in estimating $f(x)$ over the whole scale $\mathcal{A}_{\mathbb{R}^2_+}$ can be of any order; it can vary from extremely fast parametric rates to

extremely slow non-parametric ones, even to no consistency at all. We thus define a type of scales, so-called *regular-pseudo-parametric* scales, for which the parametric rate $n^{-1/2}$ can be achieved, consider estimators which are rate efficient on these scales and build an adaptive minimax estimator in *regular-non-parametric* ones.

Definition 3.3 *A functional scale $\mathcal{A}_{\Gamma_n}$ (or the corresponding scale Γ_n) is called a regular, or an R-scale if the condition*

$$\lim_{n\to\infty} \sup_{\upsilon\in\Gamma_n} M^2(1-e^{-\gamma})^{-8} N_n^5(\upsilon)\, n^{-3} = 0, \tag{3.85}$$

where $N_n(\upsilon)$ was defined in (3.61), is satisfied.

The previous condition is aimed to guarantee that the approximation arguments which were used in (3.68) and (3.69) are still applicable. Let us remark that in this condition the powers of the terms are not so relevant as far as we have $N_n(\upsilon)$ of orden $\log n$ at most.

We shall restric our study to regular scales. Two special cases of regular scales are:

Definition 3.4 *A functional scale $\mathcal{A}_{\Gamma_n}$ (a scale Γ_n) is called a regular-pseudo-parametric, or RPP-functional scale (regular-pseudo-parametric, or RPP-scale) if there exit finite constants M_+ and C_+ such that for all $(\gamma, M) \in \Gamma_n$ uniformly*

$$\limsup_{n\to\infty} \sup_{\upsilon\in\Gamma_n} M \le M_+, \quad \textit{and} \tag{3.86}$$

$$\limsup_{n\to\infty} \sup_{\upsilon\in\Gamma_n} \gamma^{-1} \log n \le C_+. \tag{3.87}$$

Regular-pseudo-parametric scales are regular, in the sense of Definition 3.3, and uniformly on them, we have parametric rates, i.e. the rate $n^{-1/2}$ is achieved given

$$\limsup_{n\to\infty} \sup_{\upsilon\in\Gamma_n} N_n(\upsilon) < \infty. \tag{3.88}$$

Definition 3.5 *A functional scale $\mathcal{A}_{\Gamma_n}$ (a scale Γ_n) is called a regular-non-parametric, or RNP-functional scale (regular-non-parametric, or RPP-scale) if*

$$\limsup_{n\to\infty} \sup_{\upsilon\in\Gamma_n} \frac{M^2}{\log n} = 0, \tag{3.89}$$

$$\liminf_{n\to\infty} \inf_{\upsilon\in\Gamma_n} M^2 \log n = \infty, \tag{3.90}$$

$$\liminf_{n\to\infty} \sup_{\upsilon\in\Gamma_n} \frac{\gamma}{\log\log n} = 0, \tag{3.91}$$

$$\liminf_{n\to\infty} \inf_{\upsilon\in\Gamma_n} \gamma \log n = \infty. \tag{3.92}$$

Note that conditions for regular-non-parametric scales require that the assumptions of Theorem 3.4 hold uniformly on RNP-scales. Thus, according to the proof of Theorem 3.4, the conditions of Theorem 3.3 also hold uniformly in RNP-scales; in particular

$$\liminf_{n\to\infty} \inf_{\upsilon\in\Gamma_n} N_n(\upsilon) = \infty. \tag{3.93}$$

Also notice that regular-non-parametric scales are regular, in the sense of Definition 3.3.

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

$$\limsup_{n\to\infty} \sup_{\upsilon\in\Gamma_n} \sup_{f\in\mathcal{A}(\upsilon)} \mathbf{E}_f \Big| n^{1/2} \big(\tilde{f}_n(x) - f(x) \big) \Big|^p < \infty$$

for any RPP-functional scales $\mathcal{A}_{\Gamma_n}$ and let $\mathcal{F}_p^0 = \mathcal{F}_p^0(x)$ be the class of all estimators such that

$$\limsup_{n\to\infty} \mathbf{E}_0 \Big| n^{1/2}\, \tilde{f}_n(x) \Big|^p < \infty.$$

One can see that $\mathcal{F}_p \subset \mathcal{F}_p^0$, since $f \equiv 0$ belongs to any of the classes $\mathcal{A}(\gamma, M)$. Below we present an adaptive estimator $\hat{f}_n \in \mathcal{F}_p$ and prove an upper bound on the quality of the estimator in RNP-functional scales. Then we prove a lower bound with the same rate for any estimator in $\mathcal{F}_p^0$. Finally we shall conclude that our adaptive estimator is $(p, \Gamma_n, \mathcal{F}_p)$-adaptive minimax for RNP-functional scales.

3.3.2 Upper bound on the quality of adaptive estimators

Theorem 3.5 *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 $\mathcal{A}_{\Gamma_n}$, $\hat{f}_n \in \mathcal{F}_p$ and*

$$\limsup_{n\to\infty} \sup_{\upsilon\in\Gamma_n} \sup_{f\in\mathcal{A}(\upsilon)} \mathbf{E}_f \left| \psi_n^{-1}(\upsilon) \big(\hat{f}_n(x) - f(x) \big) \right|^p \le 1.$$

Here

$$\psi_n^2(\upsilon) = p\,(\log N_n) \cdot \beta_{N_n}^2(x) \frac{\sigma^2 N_n}{n}$$

where N_n was defined in (3.61) for any $\upsilon \in \Gamma_n$.

The estimator. Let us first describe our adaptive estimator. Fix the parameters, $1/2 < l < 1$, $1/2 < \delta < 1$, $p_1 > 0$, $l_1 = \delta l$ and consider the sequence of truncation orders $N_0 = 0$, $N_i = \lfloor \exp(i^l) \rfloor$ for $i = 1, 2, \ldots$. Two consecutive elements of this sequence satisfy

$$N_{i+1} - N_i \sim l (\log N_i)^{1-\frac{1}{l}} N_i \to \infty \qquad (i \to \infty) \tag{3.94}$$

but, at the same time, they are close enough so that they are asymptotically equivalent,

$$\frac{N_{i+1}}{N_i} \sim e^{l i^{l-1}} \sim 1 \qquad (i \to \infty). \tag{3.95}$$

For each n we will consider the subsequence $\mathcal{S}_n = \left\{N_0, N_1, \ldots, N_{I_n}\right\}$, where

$$I_n = \arg\max_i \left\{N_i \le n^{1/2}\right\}. \tag{3.96}$$

Since for any δ, $(0 < \delta < 1/2)$ and for n large enough, $N_n(\upsilon) \le n^{1/2-\delta}$ for all υ in any RPP scale as well as any RNP scales, one can always find $i(\upsilon) \le I_n$ such that

$$N_{i(\upsilon)-1} < N_n(\upsilon) \le N_{i(\upsilon)}. \tag{3.97}$$

Let us denote

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

$$\sigma_i^2 = \mathbf{Var}_f\, \hat{f}_i(x), \qquad \hat{\sigma}_i^2 = \beta_{N_i}^2(x)\frac{\sigma^2 N_i}{n},$$

$$\sigma_{i,j}^2 = \mathbf{Var}_f\left(\hat{f}_j(x) - \hat{f}_i(x)\right), \qquad \hat{\sigma}_{i,j}^2 = \hat{\sigma}_j^2 - \hat{\sigma}_i^2,$$

and define the sequence of thresholds

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

Adaptive procedure. Define

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

We will prove that the estimator

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

satisfies Theorem 3.5. First, however, we derive some inequalities which are necessary for the proof.

Lemma 3.5 *Using the previous notation, uniformly with respect to υ in any RPP or RNP-scale, and uniformly with respect to $1 \le i, j \le I_n$, as $n \to \infty$,*

(a) $b_j^2 = o(1)\,\hat{\sigma}_j^2$ *for all j such that $i(\upsilon) \le j \le I_n$;*

(b) $\sigma_j^2 = \hat{\sigma}_j^2$ *for all j;*

(c) $(b_j - b_i)^2 = O(1)\,\hat{\sigma}_{i,j}^2$ *for all i, j such that $i(\upsilon) \le i \le j \le I_n$;*

(d) $\sigma_{i,j}^2 = \hat{\sigma}_{i,j}^2$ *for all i, j.*

Proof of lemma. (a) As we saw before

$$b_j^2 \;\le\; 2\,\big(f_{n,N_j}(x) - f_{N_j}(x)\big)^2 + 2\,\big(f_{N_j}(x) - f(x)\big)^2.$$

From equations (3.68), (3.96), and conditions for RPP scales, or as well, conditions for RNP-scales (cf. Definitions 3.4 and 3.5), we have

$$\begin{aligned}\big(f_{n,N_j}(x) - f_{N_j}(x)\big)^2 \;&\le\; \beta_{N_j}^2(x)\frac{\sigma^2 N_j}{n}\,O\left(M^2(1-e^{-\gamma})^{-8}N_j^5 n^{-3}\right)\\ &\le\; \beta_{N_j}^2(x)\frac{\sigma^2 N_j}{n}\,O\left(M^2(1-e^{-\gamma})^{-8}n^{-1/2}\right) \;=\; o(1)\,\hat\sigma_j^2.\end{aligned}$$

From (3.71),

$$\begin{aligned}\big(f_{N_j}(x) - f(x)\big)^2 \;&\le\; 2\pi M^2(1-e^{-\gamma})^{-2}\,e^{-2\gamma N_j} \;\le\; 2\pi M^2(1-e^{-\gamma})^{-2}\,e^{-\gamma N_n}\\ &=\; O\Big(\frac{1}{\gamma n}\Big) \;=\; O\Big(\frac{1}{\gamma N_j}\Big)\,\hat\sigma_j^2.\end{aligned}$$

In RPP-scales γ goes to infinity uniformly, thus γN_j goes to infinity uniformly for all $N_j \ge N_1$. In RNP-scales $\gamma N_j \ge \gamma N_n \to \infty$. Thus

$$\big(f_{N_j}(x) - f(x)\big)^2 \;=\; o(1)\,\hat\sigma_j^2,$$

as $n \to \infty$. Thus from previous equations we have that $b_j^2 = o(1)\,\hat\sigma_j^2$ for all $j \ge i(\upsilon)$, uniformly in RPP- as well as RNP-functional scales.

(b) From (3.63), taking $N = N_j$, we obtain

$$\sigma_j^2 \;=\; \mathbf{Var}\,\hat f_j(x) \;=\; \beta_{N_j}^2(x)\,\frac{\sigma^2 N_j}{n} \;=\; \hat\sigma_j^2.$$

(c) We have

$$\begin{aligned}\big(\,b_j - b_i\,\big)^2 \;&=\; \big(f_{n,N_j}(x) - f_{n,N_i}(x)\big)^2\\ &\le\; 2\big((f_{n,N_j}(x) - f_{N_j}(x)) - (f_{n,N_i}(x) - f_{N_i}(x))\big)^2\\ &\qquad\qquad + 2\big(f_{N_j}(x) \,-\, f_{N_i}(x)\big)^2\\ &:=\; 2\,b_1^2(x) + 2\,b_2^2(x).\end{aligned}$$

Now,

$$b_1 = \big(f_{n,N_j}(x) - f_{N_j}(x)\big) - \big(f_{n,N_i}(x) - f_{N_i}(x)\big) = \sum_{r=N_i}^{N_j-1}\big((f|t_r) - \langle f\,|\,t_r\rangle\big)\,t_r(x).$$

Applying the Cauchy-Schwartz inequality, (3.66) and (3.67) we see that, as we did in (a),

$$\begin{aligned} b_1^2 &= O\Big(M^2(1-e^{-\gamma})^{-8}N_j^5 n^{-4}\Big)\Big(\sum_{r=0}^{N_j-1} t_r^2(x) - \sum_{r=0}^{N_i-1} t_r^2(x)\Big) \\ &= O\Big(M^2(1-e^{-\gamma})^{-8}n^{-1/2}\Big)\ \Big(\ \beta_{N_j}^2(x)\frac{\sigma^2 N_j}{n} - \beta_{N_i}^2(x)\frac{\sigma^2 N_i}{n}\ \Big) \\ &= o(1)\,(\hat\sigma_j^2 - \hat\sigma_i^2), \quad (n\to\infty). \end{aligned}$$

Also, applying the Cauchy-Schwartz inequality,

$$b_2^2 \ \le\ \Big(\sum_{r=N_i}^{N_j-1} |\langle f\,|\,t_r\rangle|\ |t_r(x)|\Big)^2 \ \le\ \sum_{r=N_n}^{\infty} |\langle f\,|\,t_r\rangle|^2 \sum_{r=N_i}^{N_j-1} t_r^2(x),$$

where using (3.22), the definition (3.61) of N_n and condition (3.92) one can verify that

$$\sum_{r=N_n}^{\infty} |\langle f\,|\,t_r\rangle|^2 \ =\ O\Big(M^2\frac{e^{-2\gamma N_n}}{1-e^{-2\gamma}}\Big)\ =\ O\Big(\frac{(1-e^{-\gamma})^2}{\gamma(1-e^{-2\gamma})}\Big)\frac{1}{n}\ =\ O(1)\,\frac{1}{n}.$$

Now,

$$\begin{aligned} b_2^2 &= O(1)\,\frac{1}{n}\Big(\ \sum_{r=0}^{N_j-1} t_r^2(x) - \sum_{r=0}^{N_i-1} t_r^2(x)\Big) \\ &= O(1)\Big(\ \beta_{N_j}^2(x)\frac{\sigma^2 N_j}{n} - \beta_{N_i}^2(x)\frac{\sigma^2 N_i}{n}\ \Big) \\ &= O(1)\,(\hat\sigma_j^2 - \hat\sigma_i^2), \qquad (n\to\infty). \end{aligned}$$

Thus

$$\big(\,b_j - b_i\,\big)^2 = O(1)\,(\hat\sigma_j^2 - \hat\sigma_i^2)$$

for any $x\in[-1,1]$, when $n\to\infty$.

(d) Applying again the Cauchy-Schwartz inequality together with Lemma 3.2(a) we see that

$$\begin{aligned} \mathbf{Var}\big(\hat f_j(x) - \ \hat f_i(x)\big) &= \frac{\sigma^2}{n^2}\sum_{k=1}^{n}\big(K_{N_j}(x,x_k^n) - K_{N_i}(x,x_k^n)\big)^2 \\ &= \frac{\sigma^2}{n}\sum_{r_1=N_i}^{N_j-1}\sum_{r_2=N_i}^{N_j-1}\Big(t_{r_1}(x)t_{r_2}(x)\,\frac{1}{n}\sum_{k=1}^{n} t_{r_1}(x_k^n)t_{r_2}(x_k^n)\Big) \end{aligned}$$

$$= \frac{\sigma^2}{n} \sum_{r_1=N_i}^{N_j-1} \sum_{r_2=N_i}^{N_j-1} t_{r_1}(x) t_{r_2}(x) \delta r_1 r_2 = \frac{\sigma^2}{n} \sum_{r=N_i}^{N_j-1} t_r^2(x)$$

$$= \hat{\sigma}_j^2 - \hat{\sigma}_i^2.$$

□

Proof of the theorem. For arbitrary scales of parameters Γ_n and for any $f \in \mathcal{A}(\upsilon)$ for some $\upsilon \in \Gamma_n$,

$$\begin{aligned} R_n(f) &= \mathbf{E} \left| \hat{f}_{\hat{i}}(x) - f(x) \right|^p \\ &= \mathbf{E} \left\{ \mathbb{1}_{\{\hat{i} \le i(\upsilon)\}} \left| \hat{f}_{\hat{i}}(x) - f(x) \right|^p \right\} + \mathbf{E} \left\{ \mathbb{1}_{\{\hat{i} > i(\upsilon)\}} \left| \hat{f}_{\hat{i}}(x) - f(x) \right|^p \right\} \\ &:= R_n^-(f) + R_n^+(f). \end{aligned}$$

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

$$\begin{aligned} \left\{ \hat{i} \le i(\upsilon) \right\} &\subset \left\{ \left| \hat{f}_{\hat{i}}(x) - \hat{f}_{i(\upsilon)}(x) \right| \le \hat{\sigma}_{\hat{i},i(\upsilon)} \lambda_{i(\upsilon)} \right\} \\ &\subset \left\{ \left| \hat{f}_{\hat{i}}(x) - \hat{f}_{i(\upsilon)}(x) \right| \le \hat{\sigma}_{i(\upsilon)} \lambda_{i(\upsilon)} \right\}, \end{aligned}$$

given the definition of $\hat{i}$ and the property $\hat{\sigma}_{i,j}^2 = \hat{\sigma}_j^2 - \hat{\sigma}_i^2$. Therefore

$$\begin{aligned} R_n^-(f) &\le \mathbf{E} \left\{ \mathbb{1}_{\{\hat{i} \le i(\upsilon)\}} \Big(\left| \hat{f}_{\hat{i}}(x) - \hat{f}_{i(\upsilon)}(x) \right| + \left| \hat{f}_{i(\upsilon)}(x) - f(x) \right| \Big)^p \right\} \\ &\le \mathbf{E} \left(\hat{\sigma}_{i(\upsilon)} \lambda_{i(\upsilon)} + \left| \hat{f}_{i(\upsilon)}(x) - f(x) \right| \right)^p \\ &\le \mathbf{E} \left(\hat{\sigma}_{i(\upsilon)} \lambda_{i(\upsilon)} + |b_{i(\upsilon)}| + \sigma_{i(\upsilon)} |\xi| \right)^p \end{aligned} \tag{3.98}$$

where $\xi \sim \mathcal{N}(0,1)$.

In RPP-scales, the family of $N_n(\upsilon)$, the optimum bandwidths, is uniformly bounded with respect to υ. Thus, the families of $N_{i(\upsilon)}$ and $\lambda_{i(\upsilon)}$ are also uniformly bounded in Γ_n, and we can see that the variance satisfies

$$\sigma_{i(\upsilon)}^2 = \frac{\sigma^2}{n} \sum_{r=0}^{N_{i(\upsilon)}-1} t_r^2(x) \le 2 \frac{\sigma^2 N_{i(\upsilon)}}{n} = O(n^{-1}), \tag{3.99}$$

uniformly in such scales, when $n \to \infty$. From Lemma 3.5 we know that $b_{i(\upsilon)}^2 = o(1)\, \hat{\sigma}_{i(\upsilon)}^2$, thus $b_{i(\upsilon)}^2 = o(n^{-1})$. Using the above in (3.98) we have that for any RPP-scale, uniformly,

$$\sup_{f \in \mathcal{A}(\upsilon)} R_n^-(f) = O(n^{-p/2}), \qquad (n \to \infty). \tag{3.100}$$

From (3.98), applying Lemma 3.5, the dominated convergence theorem and asymptotic (3.95), uniformly in any RNP-scale

$$\sup_{f\in\mathcal{A}(\upsilon)} R_n^-(f) \le \psi_n^p(\upsilon)\big(1+o(1)\big), \qquad (n\to\infty). \tag{3.101}$$

Now let us examine $R_n^+(f)$. Consider the auxiliary event

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

Applying the Hölder and Cauchy-Schwartz inequalities we obtain

$$\begin{aligned}
R_n^+(f) &= \mathbf{E}\Big\{\mathbb{1}_{\{\hat i>i(\upsilon)\}}\big|\hat f_{\hat i}(x)-f(x)\big|^p\Big\} = \sum_{i=i(\upsilon)+1}^{I_n}\mathbf{E}\Big\{\mathbb{1}_{\{\hat i=i\}}\big|\hat f_i(x)-f(x)\big|^p\Big\}\\
&= \sum_{i=i(\upsilon)+1}^{I_n}\mathbf{E}\Big\{\big|\hat f_i(x)-f(x)\big|^p\big(\mathbb{1}_{\{\hat i=i\}\cap A_i}+\mathbb{1}_{\{\hat i=i\}\cap A_i^c}\big)\Big\}\\
&\le \sum_{i=i(\upsilon)+1}^{I_n}\mathbf{E}\Big\{\big|\hat f_i(x)-f(x)\big|^p\mathbb{1}_{\{\hat i=i\}\cap A_i}\Big\}+\sum_{i=i(\upsilon)+1}^{I_n}\mathbf{E}\Big\{\big|\hat f_i(x)-f(x)\big|^p\mathbb{1}_{A_i^c}\Big\}\\
&:= R_{n,1}^+(f)+R_{n,2}^+(f).
\end{aligned}$$

where

$$R_{n,1}^+(f) = \sum_{i=i(\upsilon)+1}^{I_n}(2\hat\sigma_i^2\lambda_i^2)^{p/2}\,\mathbf{P}(\hat i=i)$$

and

$$R_{n,2}^+(f) = \sum_{i=i(\upsilon)+1}^{I_n}\mathbf{E}^{1/2}\big|\hat f_i(x)-f(x)\big|^{2p}\,\mathbf{P}^{1/2}(A_i^c).$$

We have that

$$\begin{aligned}
\mathbf{P}(\hat i=i) &\le \mathbf{P}(\hat i\ge i)\\
&\le \sum_{j=i+1}^{\infty}\mathbf{P}\Big(\big|\hat f_{j-1}(x)-\hat f_{i-1}(x)\big|>\hat\sigma_{i-1,j-1}\,\lambda_{j-1}\Big),
\end{aligned}\tag{3.102}$$

but $\hat f_j(x)-\hat f_i(x)=\sigma_{i,j}\xi+b_j-b_i$, where $\xi\sim\mathcal{N}(0,1)$. Therefore applying Lemma 3.5, (c) and (d), and a well known bound for the tails of the normal distribution (cf. Feller [1968], Lemma 2) we find that

$$\begin{aligned}
\mathbf{P}(\,|\,\hat f_j(x)-\hat f_i(x)|>\hat\sigma_{i,j}\lambda_j\,) &\le \mathbf{P}\left(|\xi|>\lambda_j-\frac{|b_j-b_i|}{\hat\sigma_{i,j}}\right)\\
&\le \exp\left\{-\frac{1}{2}\,(\lambda_j-C_1)^2\right\} \le \exp\left\{-\frac{1}{2}\lambda_j^2+C_1\lambda_j\right\},
\end{aligned}$$

for some $C_1 > 0$ and n large enough. Returning to (3.102) we obtain that

$$\begin{aligned}
\mathbf{P}(\hat{i} \geq i) &\leq \sum_{j=i+1}^{\infty} \exp\left\{-\frac{1}{2}\lambda_{j-1}^2 + C_1\lambda_{j-1}\right\} = \sum_{j=i}^{\infty} \exp\left\{-\frac{1}{2}\lambda_j^2 + C_1\lambda_j\right\} \\
&= \sum_{j=i}^{\infty} \exp\left\{-\frac{pj^l + p_1 j^{l_1}}{2} + C_1\sqrt{pj^l + p_1 j^{l_1}}\right\} \\
&\leq \sum_{j=i}^{\infty} \exp\Big\{-\frac{pj^l}{2} - \frac{p_1 j^{l_1}}{3}\Big\} \sim \frac{2}{pl} i^{1-l} \exp\Big\{-\frac{pi^l}{2} - \frac{p_1 i^{l_1}}{3}\Big\} \\
&= \frac{2}{pl} i^{1-l} N_i^{-p/2} \exp\Big\{-\frac{p_1 i^{l_1}}{3}\Big\} \leq C_2 N_i^{-p/2} \exp\Big\{-\frac{p_1 i^{l_1}}{4}\Big\} \qquad (3.103)
\end{aligned}$$

for some $C_2 > 0$ and all $i \geq i(\upsilon)$, when n is sufficiently large. Therefore uniformly in Γ_n

$$\sup_{f \in \mathcal{A}(\upsilon)} R_1^+(f) = O(n^{-p/2}) \sum_{i=1}^{\infty} i^{pl/2} \exp\Big\{-p_1 i^{l_1}/4\Big\} = O(n^{-p/2}), \qquad (3.104)$$

when $n \to \infty$. In order to bound $R_2^+(f)$ note that $\hat{f}_i - f(x) = b_i + \sigma_i\xi$, $\xi \sim \mathcal{N}(0,1)$. Then applying Lemma 3.5, (a) and (b), in the same way as before, we have

$$\begin{aligned}
\mathbf{P}(A_i^c) &\leq \mathbf{P}\Big(|\xi| > \sqrt{2}\,\lambda_i - \frac{|b_i|}{\sigma_i}\Big) \leq \mathbf{P}\Big(|\xi| > \sqrt{2}\,\lambda_i - \sqrt{2}\Big) \\
&\leq \exp\left\{-\frac{1}{2}\Big(\sqrt{2}\,\lambda_i - \sqrt{2}\Big)^2\right\} \leq \exp\Big\{-\lambda_i^2 + 2\,\lambda_i\Big\} \\
&\leq \exp\Big\{-pi^l - p_1 i^{l_1}/2\Big\} \sim N_i^{-p} \exp\Big\{-p_1 i^{l_1}/2\Big\},
\end{aligned}$$

for all $i \geq i(\upsilon)$, n large enough. Thus, applying again Lemma 3.5, (a) and (b), and previous bound

$$\begin{aligned}
R_{n,2}^+(f) &= \sum_{i=i(\upsilon)+1}^{I_n} \mathbf{E}^{1/2}\Big|\hat{f}_i(x) - f(x)\Big|^{2p} \mathbf{P}^{1/2}(A_i^c) \\
&\leq \sum_{i=i(\upsilon)+1}^{I_n} \hat{\sigma}_i^p\, \mathbf{E}^{1/2}\Big|\, o(1) + \xi\,\Big|^{2p} \mathbf{P}^{1/2}(A_i^c) \\
&= O\Big(\beta_{N_i}^2 \frac{\sigma^2}{n}\Big)^{p/2} \sum_{i=1}^{\infty} \exp\Big\{-p_1 i^{r_1}/4\Big\}
\end{aligned}$$

and finally

$$\sup_{f \in \mathcal{A}(\upsilon)} R_{n,2}^+(f) = O\big(n^{-p/2}\big). \qquad (3.105)$$

Finally we can conclude from (3.100), (3.101), (3.104) and (3.105) that $\hat{f}_n \in \mathcal{F}_p(x)$ and

$$\limsup_{n\to\infty} \sup_{\upsilon\in\Gamma_n} \sup_{f\in\mathcal{A}(\upsilon)} \mathbf{E}\left|\psi_n^{-1}(\upsilon)\big(f_n(x)-f(x)\big)\right|^p \le 1,$$

in RNP-scales, thus ending the proof of the theorem.

3.3.3 Lower bound

Theorem 3.6 *Let $p>0$. Let $\mathcal{A}_{\Gamma_n}$ be an arbitrary RNP-functional scale. For each $\upsilon\in\Gamma_n$, define*

$$\psi_n(\upsilon) \;= \sigma_n(\upsilon)\phi_n(\upsilon)$$

where

$$\sigma_n^2(\upsilon) \;= \beta_{N_n}^2(x)\frac{\sigma^2 N_n}{n}, \qquad\qquad \phi_n^2(\upsilon) \;=\; p\log N_n,$$

and N_n is the same as in Theorem 3.5. Then, for any estimator $\tilde{f}_n \in \mathcal{F}_p^0(x)$

$$\liminf_{n\to 0} \inf_{\upsilon\in\Gamma_n} \sup_{f\in\mathcal{A}(\upsilon)} \mathbf{E}\left|\psi_n^{-1}(\upsilon)\left(\tilde{f}_n(x)-f(x)\right)\right|^p \ge 1.$$

Proof. This proof is similar to the proof of Theorem 2.5 in Ch. 2. Denote for shortness $\psi_\upsilon=\psi_n(\upsilon)$, $\phi_\upsilon=\phi_n(\upsilon)$ and $\sigma_\upsilon=\sigma_n(\upsilon)$. Choose $\bar{N}$ as it was defined in (3.74), and define $\bar{\psi}_\upsilon \;= \bar{\sigma}_\upsilon\bar{\phi}_\upsilon$ where

$$\bar{\sigma}_\upsilon^2 \;= \beta_{\bar{N}}^2(x)\frac{\sigma^2\bar{N}}{n} \qquad \text{and} \qquad \bar{\phi}_\upsilon^2 \;=\; p\log\bar{N}.$$

Define $f_0\equiv 0$ and $f_1=f_\theta$ for $\theta=\bar{\phi}_\upsilon-\bar{\phi}_\upsilon^{\,1/2}$, where f_θ belongs to the parametric family defined in (3.73). Notice that $|\theta|<\bar{N}^{1/2}$ for all n big enough. According to Lemma 3.4, $f_1\in\mathcal{A}(\upsilon)$ and

$$f_1(x)=\theta\,\beta_{\bar{N}}(x)\sqrt{\frac{\sigma^2\bar{N}}{n}}.$$

For an arbitrary estimator $\tilde{f}_n\in\mathcal{F}_p^0(x)$ denote $f_n^*=\bar{\psi}_\upsilon^{-1}\tilde{f}_n(x)$ and $L=\bar{\phi}_\upsilon^{-1}\theta$. Then

$$\bar{\psi}_\upsilon^{-1}\big(\tilde{f}_n(x)-f_1(x)\big)=f_n^*-\bar{\psi}_\upsilon^{-1}f_1(x)=f_n^*-\bar{\phi}_\upsilon^{-1}\theta=f_n^*-L \tag{3.106}$$

whereas

$$\begin{aligned}\frac{\sqrt{n}}{\sigma}\big(\tilde{f}_n(x)-f_0(x)\big) \;&=\; \frac{\sqrt{n}}{\sigma}\,\bar{\psi}_\upsilon\,f_n^*(x)=\sqrt{\bar{N}}\,\bar{\phi}_\upsilon\,f_n^*(x)\\ &=\; f_n^*\exp\left\{\frac{\log\bar{N}}{2}+\log\bar{\phi}_\upsilon\right\}.\end{aligned} \tag{3.107}$$

Denote $\mathbf{P}_0$ and $\mathbf{P}_1$ the probabilities associated with f_0 and f_1 respectively. From equations (3.75) and (3.76),

$$\frac{d\mathbf{P}_0}{d\mathbf{P}_1}(y)=\exp\left\{-\frac{\theta^2}{2}-\theta\xi\right\} \tag{3.108}$$

with respect to $\mathbf{P}_1$, where $\xi \overset{\mathbf{P}_1}{\sim} \mathcal{N}(0,1)$. Denote $q = \exp\{-\bar{\phi}_\upsilon\}$ so that $q \to 0$ since $\bar{N} \to \infty$ ($n \to \infty$) in NP-scales. Now, given $f_1 \in \mathcal{A}(\upsilon)$, for any $\tilde{f}_n \in \mathcal{F}_p^0(x)$, uniformly in $\upsilon \in \Gamma_n$ as n goes to infinity, we have

$$\begin{aligned}\bar{\mathcal{R}} &:= \sup_{f\in\mathcal{A}(\upsilon)} \mathbf{E}^{(n)} \left| \bar{\psi}_\upsilon^{-1}\big(\tilde{f}_n(x) - f(x)\big)\right|^p \geq \mathbf{E}_1 \left| \bar{\psi}_\upsilon^{-1}\big(\tilde{f}_n(x) - f_1(x)\big)\right|^p \\ &\geq q\,\mathbf{E}_0 \left| \frac{\sqrt{n}}{\sigma}\big(\tilde{f}_n(x) - f_0(x)\big)\right|^p + \\ &\qquad (1-q)\,\mathbf{E}_1 \left| \bar{\psi}_\upsilon^{-1}\big(\tilde{f}_n(x) - f_1(x)\big)\right|^p + O(q). \end{aligned} \tag{3.109}$$

According to (3.106)–(3.109),

$$\begin{aligned}\bar{\mathcal{R}} &\geq q\,\exp\Big\{\frac{\bar{\phi}_\upsilon}{2} + p\log\bar{\phi}_\upsilon\Big\}\mathbf{E}_0\left|f_n^*(x)\right|^p + (1-q)\,\mathbf{E}_1\left|f_n^*(x) - L\right|^p + O(q) \\ &\geq (1-q)\,\mathbf{E}_1\,(\,Z\,|\,f_n^*(x)\,|^p + |\,f_n^*(x) - L\,|^p\,) + O(q) \\ &\geq (1-q)\,\mathbf{E}_1 \inf_x(\,Z\,|x|^p + |\,x - L\,|^p\,) + O(q) \end{aligned} \tag{3.110}$$

where

$$Z = q\,\exp\Big\{\frac{\bar{\phi}_\upsilon}{2} + p\log\bar{\phi}_\upsilon\Big\}\frac{d\mathbf{P}_0}{d\mathbf{P}_1}.$$

From (3.108) and definition of θ we have

$$Z = \exp\Big\{-\bar{\phi}_\upsilon + \frac{\bar{\phi}_\upsilon^2}{2} + p\log\bar{\phi}_\upsilon - \big(\bar{\phi}_\upsilon - \bar{\phi}_\upsilon^{1/2}\big)\,\xi - \frac{1}{2}\big(\bar{\phi}_\upsilon - \bar{\phi}_\upsilon^{1/2}\big)^2\Big\} \overset{\mathbf{P}_1}{\to} \infty$$

given $\bar{\phi}_\upsilon \to \infty$. Now consider the same optimization problem as before:

$$\min_x \{g(x) := Z|x|^p + |L - x|^p\}.$$

We saw in the previous chapter that

$$g(x_{min}) = \chi\,L^p \tag{3.111}$$

where $\chi \overset{\mathbf{P}_1}{\to} 1$. Therefore according to equations (3.110) and (3.111), uniformly in $\upsilon \in \Gamma_n$,

$$\bar{\mathcal{R}} \geq (1-q)L^p\mathbf{E}_1\chi + O(q) = 1 + o(1).$$

Finally, uniformly in Γ_n

$$\begin{aligned}\sup_{f\in\mathcal{A}(\upsilon)} \mathbf{E}^{(n)}\left|\psi_\upsilon^{-1}\big(\tilde{f}_n(x) - f(x)\big)\right|^p &= \sup_{f\in\mathcal{A}(\upsilon)} \mathbf{E}^{(n)}\left|\bar{\psi}_\upsilon^{-1}\big(\tilde{f}_n(x) - f(x)\big)\right|^p(1+o(1)) \\ &\geq 1 + o(1).\end{aligned}$$

This completes the proof of the theorem. □

Corollary 3.3 *Let $\mathcal{A}_{\Gamma_n}$ be an arbitrary RNP-scale. Then for any $p > 0$ and $x \in \mathbb{R}$, the estimator $\hat{f}_n$ of Theorem 3.5 is $(p, \Gamma_n, \mathcal{F}_p(x))$-adaptively minimax at x.*

Proof. This is a consequence of Theorems 3.5 and 3.6. □

Chapter 4

Adaptive density estimation

Until now we have been concerned with the problem of adaptive estimation in regression models. The problem of adaptive estimation of a probability density function is no less important. The problem of density estimation can be seen as one of the basic problems of statistics. Here, as before, adaptivity appears given the uncertainty about the actual class of functions. In this study we assume that the unknown density function belongs to some of the classes $\mathcal{A}(\gamma, \beta, r)$ that we considered in Chapter 2.

As we shall see all through this chapter, a few important differences make the solution to this problem technically different when compared with the regression problems previously considered. For example, in this case the optimum bandwidth depends on $f(x)$, so that an asymptotic equivalent bandwidth has to be used. The optimal rate of convergence will also depend on $f(x)$, this will require the results be presented in restricted classes where the function $f(x)$ will be bounded away from zero from below. In the studied regression models, we had normality of the estimators; in this case, we have local asymptotic normality. In the density estimation problem, under the non-adaptive framework, we can prove local minimax results which is better than the typically global minimax results. In other words we can proof more precise results. The fact that the optimal rates depend on $f(x)$ will also make the proof of the adaptive procedure more complicated.

4.1 The model

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having common density function f. We assume f belongs to the class $\mathcal{A}(\alpha) = \mathcal{A}(\gamma, \beta, r)$ which is the family of continuous density functions whose Fourier transform satisfy

$$\|f\|_\alpha^2 := \int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} \left|\mathcal{F}[f](t)\right|^2 dt \leq 1. \tag{4.1}$$

In Section 4.3 we shall prove local asymptotic optimality in neighborhoods of the topology generated by the norm $\|\cdot\|_\alpha$. Recall the definition for the Fourier transform and its inverse, in equations (2.2) and (2.3). We saw, in Chapter 2, the relation between the parameters

of the class $\mathcal{A}(\alpha)$ and the smoothness of its elements. Lemma 4.1 in this section also illustrates the relation between the derivatives of the elements of $\mathcal{A}(\alpha)$ and the parameters γ, β and r.

For $x \in \mathbb{R}$, our goal is to estimate $f(x)$ based on the observation $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. For the purpose of this study we use the family of kernel type density estimators

$$\hat{f}_{n,s}(x) = \hat{f}_{n,s}(x, \mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} k_s(x - X_i) \tag{4.2}$$

where k_s is the already known *sinc*-function

$$k_s(x) = \frac{\sin sx}{\pi x}. \tag{4.3}$$

for $x \neq 0$ and $k_s(0) = \frac{s}{\pi}$. Let us recall the properties

$$\mathcal{F}[k_s](t) = \mathbb{1}_{[-s,s]}(t) \tag{4.4}$$

and

$$\mathcal{F}[f * k_s](t) = \mathbb{1}_{[-s,s]}(t)\mathcal{F}[f](t) \tag{4.5}$$

where * represents the convolution operator.

Let $\overline{\mathcal{W}}$ be the class of loss functions $w : \mathbb{R} \to \mathbb{R}$, such that for $x, y \in \mathbb{R}$

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

$$w(x) \geq w(y) \qquad \text{for } |x| > |y|,$$

and for some $q > 0$ and δ $(0 < \delta < 2)$

$$w(x) \leq q\, e^{|x|^{2-\delta}}.$$

Denote by $\tilde{f}_n(x, \mathbf{X})$ an arbitrary estimator of $f(x)$ based on $\mathbf{X}$. Let $\mathbf{P}_f$ be the distribution of the vector $\mathbf{X}$ and let $\mathbf{E}_f$ and $\mathbf{Var}_f$ denote expectation and variance with respect to this measure. To shorten the notation we write $\tilde{f}_n(x)$ instead of $\tilde{f}_n(x, \mathbf{X})$ and we shall not make explicit reference, in the notation of $\mathbf{P}$, $\mathbf{E}$ and $\mathbf{Var}$, to the dependence on f and n, unless it could lead to misunderstandings.

For each n and fixed α we can define a bandwidth

$$s_n = s_n(\alpha) = \frac{1}{\gamma} \left(\frac{1}{2} \log \frac{\beta^2 n}{2\pi\gamma} \right)^{1/r} \tag{4.6}$$

which helps to describe both the efficient estimator and its asymptotic variance. Using this bandwidth we define

$$\rho_n(\alpha) = \min \left\{ s_n,\, r(\gamma s_n)^r \right\}$$

and, for fixed x, consider the restricted class of functions such that for some fixed $0 < \nu < 1$ and $C > 0$

$$\mathcal{A}_\alpha^{(n)} = \mathcal{A}_\alpha^{(n)}(x) = \left\{ f \in \mathcal{A}(\alpha) \ : \ f(x) \geq C\rho_n^{\nu-1}(\alpha) \right\}. \tag{4.7}$$

Note that $\rho_n(\alpha) \sim \tilde{C} \log^{\min(1,1/r)} n$, where $\tilde{C}$ is is a constant, and thus $\rho_n^{\nu-1}(\alpha) \to 0$ when $n \to \infty$. Using the bandwidth s_n for the estimator (4.2) we prove the asymptotic optimum rate

$$\sigma_n^2 = \frac{f(x)s_n}{\pi n}$$

when optimizing the minimax risk

$$\inf_{\tilde{f}_n} \sup_{f \in \mathcal{A}_\alpha^{(n)}} \mathbf{E}\, w\left(\sigma_n^{-1}\big(\tilde{f}_n(x) - f(x)\big)\right)$$

in the class of all possible estimators, for a fixed $w \in \overline{\mathcal{W}}$, cf. Theorem 4.1.

Before stating any estimation results we shall prove a couple of useful lemmas. Some of these results will be presented in a form that will be useful in Sections 4.3 and 4.4; particularly to evaluate the accuracy of the estimates in the Section 4.3 and to compare estimates pairwise in Section 4.4.

4.2 Auxiliary results

The following lemma gives us a bound for any function $f \in \mathcal{A}(\alpha)$ and its derivatives.

Lemma 4.1 *Let $f \in \mathcal{A}(\alpha)$ and denote by $f^{(m)}$ the mth-derivative of f $(m = 0, 1, \ldots)$. Then there exists a constant $C_{r,m}$ which does not depend on γ and β such that*

$$\left|f^{(m)}(x)\right| \le C_{r,m}\, \frac{\beta}{\gamma^{m+1}}.$$

Proof. Applying the Fourier Inversion formula (2.3), properties of the Fourier transform of derivatives of f, Cauchy-Schwartz inequality and (4.1),

$$\begin{aligned}
|f^{(m)}(x)| &= \frac{1}{2\pi}\left|\int (-\mathbf{i}t)^m e^{-\mathbf{i}tx} \mathcal{F}[f](t)dt\right| \\
&\le \frac{1}{2\pi}\left(\int |t|^{2m} \frac{\beta^2}{\gamma} e^{-2|\gamma t|^r} dt\right)^{1/2} \left(\int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} |\mathcal{F}[f](t)|^2 dt\right)^{1/2} \\
&= \frac{1}{2\pi}\left(\int |t|^{2m} \frac{\beta^2}{\gamma} e^{-2|\gamma t|^r} dt\right)^{1/2} \\
&= \frac{\beta}{2\pi\gamma^{m+1}}\left(\int |t|^{2m} e^{-2|t|^r} dt\right)^{1/2} = C_{r,m}\, \frac{\beta}{\gamma^{m+1}}. \qquad (4.8)
\end{aligned}$$

The following lemma gives some properties of the kernel $k_s(x)$ as well as some properties of the random variable $k_s(x - X)$. Those properties will prove to be very useful in characterizing the kernel estimator (4.2).

Lemma 4.2 *Let $f \in \mathcal{A}(\alpha)$ and let X be a random variable with distribution density f. For $x \in \mathbb{R}$ and arbitrary bandwidths s_1, s_2 $(0 \le s_1 \le s_2)$ denote*

$$\Delta(x) = k_{s_2}(x) - k_{s_1}(x),^4 \tag{4.9}$$

Then the following properties hold:

(a)

$$\big|\Delta(x)\big| \le \frac{s_2 - s_1}{\pi};$$

(b)

$$\mathbf{E}^2\,\Delta(x - X) \le \frac{1}{2\pi^2}\int_{s_1}^{s_2} \frac{\beta^2}{\gamma}\, e^{-2(\gamma t)^r}\, dt;$$

(c)

$$\big(\mathbf{E}\, k_{s_1}(x - X) - f(x)\big)^2 \le \frac{1}{2\pi^2}\int_{s_1}^{\infty} \frac{\beta^2}{\gamma}\, e^{-2(\gamma t)^r}\, dt;$$

(d) *there exists $C_\alpha > 0$ such that*

$$\left|\mathbf{E}\,\Delta^2(x - X) - \frac{f(x)(s_2 - s_1)}{\pi}\right| \le C_\alpha\,.$$

Proof. (a) From the definition of k_s we easily see that

$$\begin{aligned}
\big|\Delta(x)\big| &= \big|k_{s_2}(x) - k_{s_1}(x)\big| \\
&= \left|\frac{\sin s_2 x - \sin s_1 x}{\pi x}\right| = \left|\frac{\cos\big((s_2 + s_1)x/2\big)\sin\big((s_2 - s_1)x/2\big)}{\pi x/2}\right| \\
&\le \left|\frac{\sin\big((s_2 - s_1)x/2\big)}{\pi x/2}\right| \le \frac{s_2 - s_1}{\pi}.
\end{aligned}$$

[4] Note that if we take $s_1 = 0$ and $s_2 = s$ then $\Delta(x) = k_s(x)$.

(b) Applying the Fourier inversion formula (2.3), properties (4.4) and (4.5), and Cauchy-Schwartz inequality, we see that

$$\begin{aligned}
\mathbf{E}^2\,\Delta(x-X) &= \left|\int \big(k_{s_2}(x-y)-k_{s_1}(x-y)\big)f(y)dy\right|^2 \\
&= \left(\frac{1}{2\pi}\left|\int e^{-\mathbf{i}tx}\big(\mathcal{F}[k_{s_2}](t)-\mathcal{F}[k_{s_1}](t)\big)\mathcal{F}[f](t)dt\right|\right)^2 \\
&\le \frac{1}{4\pi^2}\left(\int_{s_1\le|t|\le s_2}\big|\mathcal{F}[f](t)\big|dt\right)^2 \\
&\le \frac{1}{2\pi^2}\int\frac{\gamma}{\beta^2}e^{2|\gamma t|^r}\big|\mathcal{F}[f](t)\big|^2dt\int_{s_1}^{s_2}\frac{\beta^2}{\gamma}e^{-2(\gamma t)^r}dt \\
&\le \frac{1}{2\pi^2}\int_{s_1}^{s_2}\frac{\beta^2}{\gamma}e^{-2(\gamma t)^r}dt.
\end{aligned}$$

(c) This proposition is proven in the same way as the previous one.

$$\begin{aligned}
\big(\mathbf{E}\,k_{s_1}(x-X)-f(x)\big)^2 &= \left(\frac{1}{2\pi}\left|\int e^{-\mathbf{i}tx}\big(\mathcal{F}[k_{s_1}](t)-1\big)\mathcal{F}[f](t)dt\right|\right)^2 \\
&\le \frac{1}{2\pi^2}\left(\int_{s_1}^{\infty}\big|\mathcal{F}[f](t)\big|dt\right)^2 \qquad (4.10)\\
&\le \frac{1}{2\pi^2}\int_{s_1}^{\infty}\frac{\beta^2}{\gamma}e^{-2(\gamma t)^r}dt.
\end{aligned}$$

(d) Applying the symmetry of k_s, Parseval's formula, property (4.4) and the infinite differentiability of f

$$\begin{aligned}
\left|\mathbf{E}\,\Delta^2(x-X)-\frac{f(x)(s_2-s_1)}{\pi}\right| &= \left|\int\Delta^2(x-y)f(y)\,dy-f(x)\int\Delta^2(y)\,dy\right| \\
&= \left|\int\Delta^2(u)(f(x+u)-f(x))\,du\right| \\
&\le \left|\int_{|u|\le1}u^2\Delta^2(u)\;h(u)\,du\right|+\left|\int_{|u|>1}\Delta^2(u)(f(x+u)-f(x))\,du\right|
\end{aligned}$$

where $h(u)=u^{-2}(f(x+u)-f(x)-f'(x)u)$. Using the Fourier representation of h in terms of the Fourier transform of f, Cauchy-Schwartz inequality, and (4.1) one can bound $h(u)$

as follows:

$$\begin{aligned} |h(u)| &= \frac{1}{2\pi}\left|\int \frac{e^{-\mathbf{i}t(x+u)} - e^{-\mathbf{i}tx} + \mathbf{i}tue^{-\mathbf{i}tx}}{u^2}\mathcal{F}[f](t)dt\right| \\ &\le \frac{1}{2\pi}\int\left|\frac{e^{-\mathbf{i}tu} - 1 + \mathbf{i}tu}{u^2}\mathcal{F}[f](t)\right|dt \\ &\le \frac{1}{4\pi}\int t^2|\mathcal{F}[f](t)|dt \le \frac{1}{4\pi}\left(\int \frac{\beta^2}{\gamma}\, t^2 e^{-2(\gamma t)^r}dt\right)^{1/2} = C_{r,1}\frac{\beta}{2\gamma^2}. \end{aligned}$$

(cf. Lemma 4.1 for definition of $C_{r,m}$). Thus

$$\left|\mathbf{E}\,\Delta^2(x-X) - \frac{f(x)(s_2-s_1)}{\pi}\right| \le$$

$$C_{r,1}\frac{\beta}{2\gamma^2}\int_{|u|\le 1}\Delta^2(u)u^2\,du \;+\; 2\max_x|f(x)|\int_{|u|>1}\Delta^2(u)\,du$$

$$= C_{r,1}\frac{\beta}{2\gamma^2}\int_{|u|\le 1}\left(\frac{2}{\pi}\sin\big((s_2-s_1)u/2\big)\right)^2 du + C_{r,0}\frac{2\beta}{\gamma}\int_{|u|>1}\left(\frac{\sin\big((s_2-s_1)u/2\big)}{\pi u/2}\right)^2 du$$

$$\le \frac{4}{\pi^2}\frac{\beta}{\gamma}\left(\frac{C_{r,1}}{\gamma} + 4C_{r,0}\right) = C_\alpha. \tag{4.11}$$

□

4.3 Minimax density estimation in $\mathcal{A}(\gamma,\beta,r)$

Theorem 4.1 *Let $w \in \overline{\mathcal{W}}$. For any $x \in \mathbb{R}$, the following global risk bound holds*

$$\limsup_{n\to\infty}\sup_{f\in\mathcal{A}_\alpha^{(n)}}\mathbf{E}_f\, w\left(\frac{\sqrt{\pi n}\,\left|\hat f_{n,s_n}(x) - f(x)\right|}{\sqrt{f(x)s_n}}\right) \le \mathbf{E}\, w(\xi). \tag{4.12}$$

where $\hat f_{n,s_n}$ is the projection estimator (4.2) with the bandwidth s_n, defined in (4.6), and $\xi \sim \mathcal{N}(0,1)$.

For any estimator $\tilde f_n$ and any non-empty vicinity $V \subset \mathcal{A}(\alpha)$

$$\liminf_{n\to\infty}\inf_{\tilde f_n}\sup_{f\in V\cap\mathcal{A}_\alpha^{(n)}}\mathbf{E}_f\, w\left(\frac{\sqrt{\pi n}\,\left|\tilde f_n(x) - f(x)\right|}{\sqrt{f(x)s_n}}\right) \ge \mathbf{E}\, w(\xi). \tag{4.13}$$

The vicinity V corresponds to the topology generated by the norm $\|\cdot\|_\alpha$ in $\mathcal{A}(\alpha)$.

From (4.12) and (4.13) we can conclude that the proposed estimator $\hat{f}_{n,s_n}$ is locally asymptotic minimax.

$$\lim_{n\to\infty} \sup_{f\in V\cap\mathcal{A}_\alpha^{(n)}} \mathbf{E}_f w\left(\frac{\sqrt{\pi n}\,\left|\hat{f}_{n,s_n}(x)-f(x)\right|}{\sqrt{f(x)s_n}}\right)$$
$$\lim_{n\to\infty} \inf_{\tilde{f}_n} \sup_{f\in V\cap\mathcal{A}_\alpha^{(n)}} \mathbf{E}_f w\left(\frac{\sqrt{\pi n}\,\left|\tilde{f}_n(x)-f(x)\right|}{\sqrt{f(x)s_n}}\right) \;=\; \mathbf{E}\, w(\xi).$$

Proof of the upper bound. As usual we decompose the mean square error as the sum of the bias and the variance terms. Using the estimator defined on (4.2) the mean square error can be decomposed in the following way,

$$\mathbf{E}\big(\hat{f}_{n,s}(x)-f(x)\big)^2 \;=\; \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n k_s(x-X_i)-f(x)\right)^2 \;=\; \frac{1}{n}\sigma_s^2+b_s^2 \qquad (4.14)$$

where

$$\sigma_s^2 = \mathbf{Var}\, k_s(x-X) \qquad (4.15)$$

and

$$b_s = \mathbf{E}k_s(x-X)-f(x). \qquad (4.16)$$

Here X represents any of the random variables $X_1,\ldots,X_n$ and thus has density function f.

Let us sketch the proof: First we show that using the bandwidth s_n the estimator $\hat{f}_{n,s_n}$ is asymptotically unbiased to the variance uniformly in $\mathcal{A}_\alpha^{(n)}$. More precisely we prove that

$$\sigma_{s_n}^2 = \frac{f(x)s_n}{\pi}(1+o(1)) \qquad \text{and} \qquad b_{s_n}^2 = \frac{1}{n}\sigma_{s_n}^2 o(1).$$

Secondly we prove that there exists $\tilde{\delta}$ $(0<\tilde{\delta}<1)$ and $C_{w,\tilde{\delta}}>0$ such that

$$\mathbf{E}\, w^{1+\tilde{\delta}}(\eta_n) < C_{w,\tilde{\delta}}$$

and we prove that uniformly with respect to $f\in\mathcal{A}_\alpha^{(n)}$

$$\eta_n = \sqrt{\frac{\pi n}{f(x)s_n}}\big(\hat{f}_{n,s_n}(x)-f(x)\big) \xrightarrow{d} \xi.$$

where $\xi\sim\mathcal{N}(0,1)$. From the last two statements and taking a sequence of uniformly continuous functions w_δ that approximate w when $\delta\to 0$ we show that uniformly in $\mathcal{A}_\alpha^{(n)}$, $\mathbf{E}_f\, w_\delta(\eta_n)\to\mathbf{E}\, w_\delta(\xi)$ and thus the desired upper bound when $\delta\to 0$.

Now let us continue the proof. First let us consider the variance. Applying Lemma 4.2(d) taking $s_1=0$ and $s_2=s_n$ we have

$$\left|\sigma_{s_n}^2-\frac{f(x)s_n}{\pi}\right| \;\le\; \mathbf{E}^2 k_{s_n}(x-X)+\left|\mathbf{E}k_{s_n}^2(x-X)-\frac{f(x)s_n}{\pi}\right| \;\le\; \mathbf{E}^2 k_{s_n}(x-X)+C_\alpha.$$

Now, from Lemma 4.2(b), taking $s_1 = 0$ and $s_2 = s_n$ we can see that there is a constant $\tilde{C}_\alpha = C_{r,0}\frac{\beta}{\gamma}$ such that $\mathbf{E}k_{s_n}(x - X) \le \tilde{C}_\alpha$ for any $f \in \mathcal{A}_\alpha^{(n)}$. Thus uniformly in $\mathcal{A}_\alpha^{(n)}$

$$\sigma_{s_n}^2 \;=\; \frac{f(x)s_n}{\pi}\big(1 + o(1)\big), \quad (n \to \infty). \tag{4.17}$$

Note that this asymptotic property is also valid for any bandwidth $\tilde{s}_n$ such that $f(x)\tilde{s}_n \to \infty$ uniformly in $\mathcal{A}_\alpha^{(n)}$.

Now let us consider the bias. We know from Lemma 4.2(c), taking $s_1 = s_n$, that

$$b_{s_n}^2 \le \frac{1}{2\pi^2}\int_{s_n}^{\infty} \frac{\beta^2}{\gamma} e^{-2(\gamma t)^r} dt. \tag{4.18}$$

Using the Lemma 2.2, definition of s_n, and condition (4.7) for functions on the class $\mathcal{A}_\alpha^{(n)}$, we get that uniformly in $\mathcal{A}_\alpha^{(n)}$

$$b_{s_n}^2 \;\le\; \frac{f(x)s_n}{\pi n}\frac{1}{f(x)r(\gamma s_n)^r} \;=\; \frac{f(x)s_n}{\pi n}o(1), \qquad (n \to \infty). \tag{4.19}$$

Using the same tools one can verify that this property also holds for any bandwidth $\tilde{s}_n \ge s_n$.

Lemma 4.3 *Let $f \in \mathcal{A}_\alpha^{(n)}$ and let s_n be defined as before by (4.6). Let $w \in \overline{\mathcal{W}}$ be a loss function, then there exist $\tilde{\delta} > 0$ and $C > 0$ such that*

$$\mathbf{E}w^{1+\tilde{\delta}}\left(\sqrt{\frac{\pi n}{f(x)s_n}}\big(\hat{f}_{n,s_n}(x) - f(x)\big)\right) < C. \tag{4.20}$$

Proof. For any $w \in \overline{\mathcal{W}}$ there exists $\tilde{\delta}$ such that $w^{1+\tilde{\delta}} \in \overline{\mathcal{W}}$. So, without loss of generality, we just prove that for any $w \in \overline{\mathcal{W}}$ there exists a constant C such that

$$\mathbf{E}w\left(\sqrt{\frac{\pi n}{f(x)s_n}}\big(\hat{f}_{n,s_n}(x) - f(x)\big)\right) < C. \tag{4.21}$$

To prove that, it is sufficient to show that for any λ and any δ $(0 < \delta < 2)$, there is a $C_{\lambda,\delta} > 0$ such that for all sufficiently large n and all y

$$\mathbf{P}\left(\sqrt{\frac{\pi n}{f(x)s_n}}\big|\hat{f}_{n,s_n}(x) - f(x)\big| > y\right) \le C_{\lambda,\delta}\, e^{-\lambda y^{2-\delta}}. \tag{4.22}$$

According to Lemma 4.1 the functions $f \in \mathcal{A}(\alpha)$ are uniformly bounded. Thus, for all sufficiently large n, say $n > n_0$, given s_n goes to infinity then $\sup_{f\in\mathcal{A}(\alpha)}\sup_x |f(x)| \le s_n$. Since $|k_s| \le s$, according to Lemma 4.2(a), it follows that $|\hat{f}_{n,s_n}(x)| \le s_n$ and thus $\hat{f}_{n,s_n}(x) \equiv$

$\bar{f}_n(x)$. From the asymptotic unbiasness of $\hat{f}_{n,s_n}$ to the variance which was shown in (4.19) follows that for n big enough

$$\mathbf{P}\left(\sqrt{\frac{\pi n}{f(x)s_n}}\big|\hat{f}_{n,s_n}(x) - f(x)\big| > y\right) \leq \mathbf{P}\left(\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n} k_{s_n}(x - X_i) - \mathbf{E}k_{s_n}(x - X_i)\right| > \frac{y}{2}\sqrt{\frac{f(x)s_n}{\pi}}\right).$$

According to Bernstein's inequality(see Appendix), Lemma 4.2(a) and asymptotic (4.17) this does not exceed

$$2\exp\left\{-\frac{y^2\pi^{-1}s_n f(x)}{8\left(\sigma^2_{s_n} + \frac{1}{3}\frac{2s_n}{\pi}\sqrt{\frac{f(x)s_n}{\pi n}}\frac{y}{2}\right)}\right\} = 2\exp\left\{-\frac{1}{8}\frac{y^2}{\left((1+o(1)) + \frac{1}{3}\sqrt{\frac{s_n}{\pi f(x)n}}\,y\right)}\right\}$$

$$\leq \begin{cases} 2e^{-\frac{y^2}{24}} \leq C_{\lambda,\delta}\, e^{-\lambda y^{2-\delta}} & \text{if } y \leq \frac{3}{2}\sqrt{\frac{\pi f(x)n}{s_n}} \\ 2e^{-\frac{1}{8}\sqrt{\frac{\pi n f(x)}{s_n}}y} \leq 2e^{-\lambda y^{2-\delta}} & \text{if } \frac{3}{2}\sqrt{\frac{\pi f(x)n}{s_n}} \leq y \leq 2\sqrt{\frac{\pi n s_n}{f(x)}} \end{cases}$$

since for any δ $(0 < \delta < 2)$ and any $\lambda > 0$, for n large enough

$$\frac{1}{8}\sqrt{\frac{\pi f(x)n}{s_n}}y \geq \frac{3}{16}\pi n = \frac{3\,(2^\delta)}{64}\left(\frac{f(x)^{1-\delta/2}n^{\delta/2}}{s_n^{1-\delta/2}}\right)\left(2\sqrt{\frac{\pi n s_n}{f(x)}}\right)^{2-\delta} \geq \lambda y^{2-\delta}.$$

From (4.22), taking $\lambda = 2$, and conditions for $w \in \overline{\mathcal{W}}$, we see

$$\int w(y)d\mathbf{P}\left(\sqrt{\frac{\pi n}{f(x)s_n}}\big|\bar{f}_n(x) - f(x)\big| \leq y\right) \leq \int qe^{y^{2-\delta}}d\mathbf{P}\left(\sqrt{\frac{\pi n}{f(x)s_n}}\,\big|\bar{f}_n(x) - f(x)\big| \leq y\right)$$

$$= \int\left(1 - \mathbf{P}\left(\sqrt{\frac{\pi n}{f(x)s_n}}\big|\bar{f}_n(x) - f(x)\big| \leq y\right)\right) q(2-\delta)y^{1-\delta}e^{y^{2-\delta}}dy$$

$$= \int \mathbf{P}\left(\sqrt{\frac{\pi n}{f(x)s_n}}\big|\bar{f}_n(x) - f(x)\big| > y\right) q(2-\delta)y^{1-\delta}e^{y^{2-\delta}}dy$$

$$= q\,(2-\delta)C_\lambda\int_0^\infty y^{1-\delta}e^{-y^{2-\delta}}dy \leq C.$$

Thus, the lemma is proved. □

For the purpose of this section we need the following result just for the bandwidth s_n. However, this property will be used for a wider collection of bandwidths along this section and the next, so we consider a more general bandwidth $\tilde{s}$.

Lemma 4.4 *Let $f \in \mathcal{A}_\alpha^{(n)}$ and let $\tilde{s} = \tilde{s}_n$ be a bandwidth such that $\tilde{s} \to \infty$ and $\frac{\tilde{s}}{f(x)n} \to 0$ as $n \to \infty$ uniformly in $\mathcal{A}_\alpha^{(n)}$. Let us define*

$$\xi_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i}{\sigma_{\tilde{s}}}$$

where

$$Y_i = k_{\tilde{s}}(x - X_i) - \mathbf{E}_f\, k_{\tilde{s}}(x - X),$$

and $\sigma_{\tilde{s}}^2 = \mathbf{Var}\, k_{\tilde{s}}$. Then uniformly w.r.t. $f \in \mathcal{A}_\alpha^{(n)}$ (cf. Ibragimov and Has'minskii [1981], Def. on p. 365)

$$\xi_n \xrightarrow{d} \xi$$

where $\xi \sim \mathcal{N}(0, 1)$.

Proof. We have that $Y_i,\ i = 1, \ldots, n$ are i.i.d. random variables. Suppose they distribute as a random variable Y. We see that $\mathbf{E}Y = 0$. We prove that

$$\mathbf{E}\left(\frac{Y^2}{\sigma_{\tilde{s}}^2}\ \mathbb{1}\left\{\frac{|Y|}{\sigma_{\tilde{s}}} > \epsilon\sqrt{n}\right\}\right) \overset{n\to\infty}{\longrightarrow} 0 \tag{4.23}$$

thus, as a consequence of Lindenberg's theorem (cf. Ibragimov and Has'minskii [1981]), the lemma will be proved. Applying Lemma 4.2(a) and asymptotic (4.17) we have

$$\begin{aligned}
\mathbf{E}\left(\frac{Y^2}{\sigma_{\tilde{s}}^2}\ \mathbb{1}\left\{\frac{|Y|}{\sigma_{\tilde{s}}} > \epsilon\sqrt{n}\right\}\right) &= \mathbf{E}\left(\frac{Y^2}{\sigma_{\tilde{s}}^2}\ \mathbb{1}\left\{\frac{|k_{\tilde{s}}(x - X_i) - \mathbf{E}\, k_{\tilde{s}}(x - X)|}{\sigma_{\tilde{s}}} > \epsilon\sqrt{n}\right\}\right) \\
&\le \mathbf{E}\left(\frac{Y^2}{\sigma_{\tilde{s}}^2}\ \mathbb{1}\left\{\frac{2\max|k_{\tilde{s}}(x)|}{\sigma_{\tilde{s}}} > \epsilon\sqrt{n}\right\}\right) = \mathbf{E}\left(\frac{Y^2}{\sigma_{\tilde{s}}^2}\ \mathbb{1}\left\{\frac{2\tilde{s}}{\sigma_{\tilde{s}}\pi} > \epsilon\sqrt{n}\right\}\right) \\
&= \mathbf{E}\left(\frac{Y^2}{\sigma_{\tilde{s}}^2}\ \mathbb{1}\left\{\frac{4\tilde{s}^2}{\sigma_{\tilde{s}}^2\pi^2} > \epsilon^2 n\right\}\right) = \mathbf{E}\left(\frac{Y^2}{\sigma_{\tilde{s}}^2}\ \mathbb{1}\left\{\frac{4}{\pi}(1 + o(1))\frac{\tilde{s}}{f(x)n} > \epsilon^2\right\}\right)
\end{aligned}$$

when $n \to \infty$. By hypothesis $\frac{\tilde{s}}{f(x)n} \to 0$ as n goes to infinity thus

$$\mathbf{E}\left(\frac{Y^2}{\sigma_{\tilde{s}}^2}\ \mathbb{1}\left\{\frac{|Y|}{\sigma_{\tilde{s}}} > \epsilon\sqrt{n}\right\}\right) = 0 \quad \forall n \text{ big enough.}$$

Thus the lemma is proved. □

Now, let us go back to the proof of the theorem. Define

$$\eta_n = \sqrt{\frac{\pi n}{f(x)s_n}}\big(\hat{f}_{n,s_n}(x) - f(x)\big)$$

then

$$\begin{aligned}
\eta_n &= \sqrt{\frac{\pi n}{f(x)s_n}}\left(\frac{1}{n}\sum_{i=1}^{n} k_{s_n}(x - X_i) - \mathbf{E}\, k_{s_n}(x) + b_{s_n}(x)\right) \\
&= \sqrt{\frac{\pi n}{f(x)s_n}}\left(\frac{\sigma_{s_n}}{\sqrt{n}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{Y_i}{\sigma_{s_n}}\right) + b_{s_n}(x)\right) \\
&= \left(\frac{\pi\sigma^2_{s_n}}{f(x)s_n}\right)^{1/2}\xi_n + \left(\frac{\pi n b^2_{s_n}(x)}{f(x)s_n}\right)^{1/2} \\
&= a_n\,\xi_n + b_n.
\end{aligned}$$

From (4.17) and (4.18) we have that $a_n \overset{n\to\infty}{\longrightarrow} 1$ and $b_n \overset{n\to\infty}{\longrightarrow} 0$, uniformly in $\mathcal{A}^{(n)}_\alpha$, while from Lemma 4.4, $\xi_n \overset{d}{\to} \xi$ as n goes to infinity. Thus, applying Slutsky's Theorem (cf. Bickel and Doksum [1977], p. 461), we conclude that $\eta_n \overset{d}{\to} \xi$ where $\xi \sim \mathcal{N}(0,1)$. Given $\eta_n \to \xi$ in distribution then $w(\eta_n) \to w(\xi)$ for any function w continuous almost everywhere. The loss functions $w \in \overline{\mathcal{W}}$ are monotone in the real semiaxes then they are continuous almost everywhere, since they can have only countable number of jumps. Converge in distribution implies convergence in probability thus $w(\eta_n) \to w(\xi)$ also in probability. We also proved in Lemma 4.3 that there exist $\tilde{\delta}$ and $C > 0$ such that $\mathbf{E}\, w^{1+\tilde{\delta}}(\eta_n) < C$. Now, applying the last two arguments we get, as a consequence of the L_r-convergence Theorem (cf. Loève [1977], Corollary 2, p. 166), that $\mathbf{E}\, w(\eta_n) \to \mathbf{E}\, w(\xi)$, and thus

$$\lim_{n\to\infty} \sup_{f\in\mathcal{A}^{(n)}_\alpha} \mathbf{E}\, w\left(\sqrt{\frac{\pi n}{f(x)s_n}}\big(\hat{f}_{n,s_n}(x) - f(x)\big)\right) = \mathbf{E}\, w(\xi). \tag{4.24}$$

Proof of the lower bound for the risk. In order to prove the lower bound let us define a couple of auxiliary bandwidths which are smaller than s_n but asymptotically equivalent to it. So, let us denote

$$\bar{s}_n = \gamma^{-1}\left((\gamma s_n)^r - vr\log(\gamma s_n)\right)^{1/r}, \quad \text{and} \tag{4.25}$$

$$\tilde{s}_n = \bar{s}_n - \log\bar{s}_n \tag{4.26}$$

where v satisfies $r(v+1) > 1$.

Our proof of the lower bound will be based on exhibiting a special parametric subfamily in any given neighborhood $V \subset \mathcal{A}(\alpha)$ defined by the topology corresponding to the norm $\|\cdot\|_\alpha$. We will choose such a family of the form

$$f_\theta(y) = f_0(y)\left(1 + \theta\sqrt{\frac{\pi}{f_0(x)\tilde{s}_n n}}\left(k_{\tilde{s}_n}(x-y) - \bar{k}_{\tilde{s}_n}(x)\right)\right), \quad |\theta| \le \theta_n = \tilde{s}_n^{1/4} \tag{4.27}$$

where

$$\bar{k}_{\tilde{s}_n}(x) = \mathbf{E}_{f_0} k_{\tilde{s}_n}(x - X).$$

Definition 4.1 *Let $\mathbf{P}_\theta^{(n)}$ be the joint distribution of the sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, given by n i.i.d observations corresponding to the marginal common density f_θ, $(\theta \in \Theta_n)$. Assume that $\mathbf{P}_\theta^{(n)}$ is absolutely continuous w.r.t. $\mathbf{P}_0^{(n)}$ (for a more general case cf. Ibragimov and Has'minskii [1981]). Let the log-likelihood*

$$L(\theta) = \log \frac{d\mathbf{P}_\theta^{(n)}}{d\mathbf{P}_0^{(n)}}(\mathbf{X}).$$

We say $\mathbf{P}_\theta^{(n)}$ is locally asymptotically normal (LAN) at $\theta = 0$ as $n \to \infty$ if the log-likelihood function $L(\theta)$ satisfies

$$L(\theta) = \theta\Delta_n - \frac{1}{2}\theta^2 + r_{n,\theta}$$

where $\Delta_n \xrightarrow{d} \mathcal{N}(0,1)$ and $r_{n,\theta} \xrightarrow{p} 0$ with respect to $\mathbf{P}_0^{(n)}$ as n goes to infinity.

Lemma 4.5 *Let $x \in \mathbb{R}$, then for any neighborhood V in $\mathcal{A}(\alpha)$ there exist a function f_0 such that $f_0(x) > 0$ and a parametric family of functions $\{f_\theta\}$, as it was defined in (4.27) such that*

(a) $f_\theta(x) = f_0(x) + \theta\sqrt{\frac{f_0(x)\tilde{s}_n}{\pi n}}(1 + o(1))$ *as* $n \to \infty$.

(b) $f_\theta(y) = f_0(y)(1 + o(1))$ *as* $n \to \infty$.

(c) $f_\theta \in V$, *for all* $|\theta| \le \theta_n$ *and n sufficiently large.*

(d) *The family $\mathbf{P}_\theta^{(n)}$ corresponding to the densities f_θ is LAN at $\theta = 0$.*

Proof. Let us sketch the proof. There exists a density function $f \in V$, $\|f\|_\alpha < 1$ such that $f(x) > 0$. Rescaling f we build a function $f_1 \in V$, which is close enough to f, $f_1(x) > 0$ and

$$\int \frac{\gamma}{\beta^2}|t|^2 e^{2|\gamma t|^r}\left|\mathcal{F}[f_1](t)\right|^2 dt < \infty. \tag{4.28}$$

Then we build a function $f_0 \in V$ such that $f_0(x) > 0$ and $\mathcal{F}[f_0]$ has finite support. We will use this chosen f_0 in (4.27) and prove (a)–(d).

To do so, consider the density function $f_1(x) = \frac{1}{a}f(\frac{x}{a})$. First, it is continuous. Second, we know that $\mathcal{F}[f_1](t) = \mathcal{F}[f](at)$. For a given $a > 1$

$$\begin{aligned}\int \frac{\gamma}{\beta^2}e^{2|\gamma t|^r}\left|\mathcal{F}[f_1](t)\right|^2 dt &\le \int \frac{\gamma}{\beta^2}e^{2|a\gamma t|^r}|\mathcal{F}[f_1](t)|^2 dt \\ &= \frac{1}{a}\int \frac{\gamma}{\beta^2}e^{2|\gamma a t|^r}\left|\mathcal{F}[f](at)\right|^2 d(at) \le \frac{1}{a} < 1. \end{aligned} \tag{4.29}$$

Thus $f_1 \in \mathcal{A}(\alpha)$. The function f is continuous thus $\mathcal{F}[f](at) \to \mathcal{F}[f](t)$ for all t when $a \to 1$. From this, (4.1) and (4.29), by dominated convergence theorem we see that

$$\|f_1 - f\|_\alpha^2 = \int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} \left|\mathcal{F}[f](at) - \mathcal{F}[f](t)\right|^2 dt \to 0, \qquad (a \to 1).$$

Thus, the parameter a can be taken close enough to 1, such that $f_1 \in V$ and $f_1(x) > 0$. We know that for every $a > 1$ there exist a C_a such that

$$|t|^2 e^{2(\gamma t)^r} \leq C_a e^{2(a\gamma t)^r},$$

thus

$$\begin{aligned} C_1 \;&:=\; \int \frac{\gamma}{\beta^2} |t|^2 e^{2|\gamma t|^r} \left|\mathcal{F}[f_1](t)\right|^2 dt \;\leq\; C_a \int \frac{\gamma}{\beta^2} e^{2|a\gamma t|^r} \left|\mathcal{F}[f_1](t)\right|^2 dt \\ &=\; \frac{C_a}{a} \int \frac{\gamma}{\beta^2} e^{2|a\gamma t|^r} \left|\mathcal{F}[f](at)\right|^2 a\, dt \;\leq\; \frac{C_a}{a}. \end{aligned}$$

Now, for a given $M > 0$ let us define $f_0 = f_1 * g_M$ where g_M is the density function

$$g_M(x) = \frac{1 - \cos^2(Mx)}{\pi M x^2}.$$

The Fourier transform of g_M is known to be

$$\mathcal{F}[g_M](t) = \left(1 - M^{-1}|t|\right) \mathbb{1}_{[-M,M]}(t).$$

Thus the Fourier transform of the function f_0, which satisfies $\mathcal{F}[f_0] = \mathcal{F}[f_1]\mathcal{F}[g_M]$, cf. (4.5), has finite support in the interval $[-M, M]$. Note that $\mathcal{F}[f_0] \leq \mathcal{F}[f_1]$ thus $\|f_0\|_\alpha \leq \|f_1\|_\alpha \leq 1$, therefore $f_0 \in \mathcal{A}(\alpha)$. Let us show that $f_0 \in V$ and $f_0(x) > 0$, at least if M is big enough. See that

$$\mathcal{F}[f_1](t) - \mathcal{F}[f_0](t) = \mathcal{F}[f_1](t) - \left(1 - M^{-1}|t|\right)\mathcal{F}[f_1](t) = M^{-1}|t|\mathcal{F}[f_1](t)$$

for all $|t| < M$ then, using (4.30),

$$\begin{aligned} \| f_1 - f_0 \|_\alpha^2 \;&=\; \frac{1}{M^2} \int_{|t|<M} \frac{\gamma}{\beta^2} |t|^2 e^{2|\gamma t|^r} \left|\mathcal{F}[f_1](t)\right|^2 dt + \int_{|t|>M} \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} |\mathcal{F}[f_1](t)|^2 dt \\ &\leq\; \frac{C_1}{M^2} + \int_{|t|>M} \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} |\mathcal{F}[f_0](t)|^2 dt \;\to\; 0, \quad (M \to \infty). \end{aligned}$$

So we can fix an M such that $f_0 \in V$. Moreover, given continuity and boundness of f_1 we know that

$$f_0(x) = \int f_1(x - y) g_M(y) dy \to f_1(x), \quad (M \to \infty),$$

thus, we can take M big enough such that $f_0(x) \geq \frac{f_1(x)}{2} > 0$.

We are now ready to prove the lemma for the function f_θ defined in (4.27).

(a) We know that $k_{\tilde{s}_n}(0) = \frac{\tilde{s}_n}{\pi}$ and $\bar{k}_{\tilde{s}_n} \to f_0(x)$ when $\tilde{s}_n$ goes to infinity. From equations (4.25) and (4.26) one can see that $\tilde{s}_n$ goes to infinity when n goes to infinity, hence

$$k_{\tilde{s}_n}(0) - \bar{k}_{\tilde{s}_n}(x) = \frac{\tilde{s}_n}{\pi}(1 + o(1)), \quad (n \to \infty). \tag{4.30}$$

Now, if we substitute this in the definition (4.27), taking $y = x$, we get the desired result.

(b) Using Lemma 4.2(a) and the value of θ_n, given in (4.27), we can see that

$$\begin{aligned} \left| \theta \sqrt{\frac{\pi}{f_0(x)\tilde{s}_n n}} \big(k_{\tilde{s}_n}(x-y) - \bar{k}_{\tilde{s}_n}(x)\big) \right| &\leq 2\,\theta_n \sqrt{\frac{\pi \tilde{s}_n}{f_0(x) n}}(1 + o(1)) \\ &\leq 2 \sqrt{\frac{\pi \tilde{s}_n^{3/2}}{f_0(x) n}}(1 + o(1)) = o(1) \end{aligned} \tag{4.31}$$

when n goes to infinity. Thus $f_\theta(y) = f_0(y)(1 + o(1))$.

(c) From the previous result the function $f_\theta(y)$ is non-negative for all y for n sufficiently large. Obviously

$$\int f_\theta(y)dy = \int f_0(y)dy = 1.$$

We then conclude that f_θ is a density function for all θ, $|\theta| \leq \theta_n$, for n big enough. As we saw before f_0 is continuous so f_θ is continuous too.

Let us first evaluate the α-norm of the function $\psi(y) = f_0(y)\, k_{\tilde{s}_n}(x - y)$. Note that

$$\mathcal{F}[\psi](t) = \frac{1}{2\pi} \int e^{\mathbf{i}xu} \mathcal{F}[f_0](t + u)\, \mathcal{F}[k_{\tilde{s}_n}](u)\, du$$

thus

$$\|\psi\|_\alpha \;=\; \frac{1}{2\pi} \left(\int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} \left| \int e^{\mathbf{i}xu} \mathcal{F}[f_0](t + u) \mathcal{F}[k_{\tilde{s}_n}](u)\, du \right|^2 dt \right)^{1/2}.$$

Given $\mathcal{F}[f_0](t+u) = 0$ when $|t+u| > M$ and $\mathcal{F}[k_{\tilde{s}_n}](u) = 0$ for $u > \tilde{s}_n$ one can see that $\mathcal{F}[f_0](t+u)\mathcal{F}[k_{\tilde{s}_n}](u) = 0$ for $|t| > \tilde{s}_n + M$ then

$$\begin{aligned}
\|\psi\|_\alpha &\leq \frac{1}{2\pi}\left(\int_{|t|\leq \tilde{s}_n+M} \frac{\gamma}{\beta^2} e^{2|\gamma t|^r}\left(\int \left|\mathcal{F}[f_0](t+u)\mathcal{F}[k_{\tilde{s}_n}](u)\right| du\right)^2 dt\right)^{1/2} \\
&\leq \frac{1}{2\pi}\left(\int_{|t|\leq \tilde{s}_n+M} \frac{\gamma}{\beta^2} e^{2|\gamma t|^r}\left(\int \left|\mathcal{F}[f_0](t+u)\right| du\right)^2 dt\right)^{1/2} \\
&\leq \frac{C_2}{2\pi}\left(\int_{|t|<\tilde{s}_n+M} \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} dt\right)^{1/2} \leq \frac{C_2}{2\pi}\left(\frac{2\gamma}{\beta^2}\int_0^{\tilde{s}_n+M} e^{2(\gamma t)^r} dt\right)^{1/2}.
\end{aligned}$$

Now we see that $\tilde{s}_n + M \leq \tilde{s}_n + \log \bar{s}_n$ for n large enough, given $\bar{s}_n \to \infty$ when n goes to infinity (cf. eqs. (4.25) and (4.26)). Thus, applying Lemma 2.2 we get

$$\begin{aligned}
\|\psi\|_\alpha &\leq \frac{C_2}{2\pi}\left(\frac{2\gamma}{\beta^2}\int_0^{\tilde{s}_n+\log \bar{s}_n} e^{2(\gamma t)^r} dt\right)^{1/2} = \frac{C_2}{2\pi}\left(\frac{2\gamma}{\beta^2}\int_0^{\bar{s}_n} e^{2(\gamma t)^r} dt\right)^{1/2} \\
&\leq \frac{C_2}{2\pi}\left((\gamma \bar{s}_n)^{-r(v+1)+1}\frac{n}{2\pi\gamma r}\right)^{1/2} = C_3 n^{1/2}
\end{aligned}$$

given $r(v+1) > 1$. Now

$$\begin{aligned}
\| f_\theta - f_0 \|_\alpha &\leq \sqrt{\frac{\pi}{f_0(x)\tilde{s}_n n}}\theta_n \left(\|\psi\|_\alpha + \|f_0\|_\alpha \bar{k}_{\tilde{s}_n}(x)\right) \\
&\leq \sqrt{\frac{\pi}{f_0(x)\tilde{s}_n^{1/2} n}}\left(O(n^{1/2}) + O(\tilde{s}_n)\right) = o(1), \quad (n \to \infty).
\end{aligned}$$

(d) Now we prove the last part. Denote

$$Y_i = k_{\tilde{s}_n}(x - X_i) - \bar{k}_{\tilde{s}_n}(x). \tag{4.32}$$

Using the Taylor expansion, the log-likelihood w.r.t. f_0 can be expressed as

$$\begin{aligned}
L(\theta) &= \sum_{i=1}^n \log\left(1 + \theta\sqrt{\frac{\pi}{f_0(x) n \tilde{s}_n}} Y_i\right) \\
&= \theta\sqrt{\frac{\pi}{f_0(x) n \tilde{s}_n}}\sum_{i=1}^n Y_i - \frac{\theta^2}{2}\frac{\pi}{f_0(x)\tilde{s}_n}\frac{1}{n}\sum_{i=1}^n Y_i^2 + O(1)\frac{\theta^3}{f_0^{3/2}(x) n^{3/2} s_n^{1/2}(n)}\sum_{i=1}^n Y_i^3 \\
&= \theta\Delta_n - \frac{\theta^2}{2} + r_n
\end{aligned} \tag{4.33}$$

where

$$\Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i}{\sqrt{\frac{f_0(x)\tilde{s}_n}{\pi}}},$$

and

$$r_n = \frac{\theta^2}{2} \left(\frac{\pi}{f_0(x)\tilde{s}_n} \frac{1}{n} \sum_{i=1}^{n} Y_i^2 - f_0(x) \right) + O(1) \frac{\theta^3}{f_0^{3/2}(x) n^{3/2} \tilde{s}_n^{3/2}} \sum_{i=1}^{n} Y_i^3.$$

Using (4.17), Lemma 4.4 and Slutsky's theorem we have that $\Delta_n \to \mathcal{N}(0,1)$ in distribution with respect to f_0. By the Law of Large Number

$$\frac{\pi}{f_0(x)\tilde{s}_n} \frac{1}{n} \sum_{i=1}^{n} Y_i^2 - 1 \xrightarrow{\mathbf{P}_0^{(n)}} 0$$

where $\mathbf{P}_0^{(n)}$ is the distribution of $X_1, X_2, \ldots, X_n$ corresponding to f_0 and

$$r_n \le o_{P_0^{(n)}}(1) + \frac{O(1)\theta^3}{n^{\frac{3}{2}} \tilde{s}_n^{\frac{3}{2}}} n \frac{\tilde{s}_n^3}{\pi^3} = o_{P_0^{(n)}}(1) + o(1) \frac{\tilde{s}_n^{3/2}}{n^{1/2}} = o_{P_0^{(n)}}(1), \qquad (n \to \infty).$$

Thus the family $\{f_\theta\}$ is locally asymptotically normal and the lemma is proved. □

Once we have the Local Asymptotic Normality we are ready to apply the following result, (see Ibragimov and Has'minskii [1981], Sect. II.2). Let a family of density f_θ, $|\theta| \le \theta_n$ which satisfies LAN property at $\theta = 0$, such that for any $K > 0$, $\theta_n > K$ for all n sufficiently large, then for any quasi-convex loss function $W(u)$, $u \in \mathbb{R}$

$$\lim_{K \to \infty} \liminf_{n \to \infty} \inf_{T_n} \sup_{|\theta| < K} \mathbf{E}_\theta W\left(T_n - \theta\right) \ge \mathbf{E}\, W(\xi) \tag{4.34}$$

where T_n is an arbitrary estimator of θ and $\xi \sim \mathcal{N}(0,1)$. We can approximate any loss function by a sequence of uniformly continuous functions $w_\delta \in \overline{\mathcal{W}}$ such that $w_\delta \nearrow w$ when $\delta \to 0$. Recall that s_n is asymptotically equivalent to $\tilde{s}_n$. Now for any element w_α of the mentioned sequence we can see that

$$\begin{aligned} \mathcal{R}_\delta(V) &= \liminf_{n \to \infty} \inf_{\tilde{f}_n} \sup_{f \in V \cap \mathcal{A}_\alpha^{(n)}} \mathbf{E} w_\delta \left(\sqrt{\frac{\pi n}{f(x) s_n}} \big(\tilde{f}_n(x) - f(x) \big) \right) \\ &= \liminf_{n \to \infty} \inf_{\tilde{f}_n} \sup_{f \in V \cap \mathcal{A}_\alpha^{(n)}} \mathbf{E} w_\delta \left(\sqrt{\frac{\pi n}{f(x) \tilde{s}_n}} \big(\tilde{f}_n(x) - f(x) \big) \big(1 + o(1) \big) \right) \end{aligned}$$

but from the previous lemma we know that the family $\{f_\theta\} \subset V$ for n big enough. Also as $f_\theta(x) > \frac{1}{2} f_0(x) > 0$, $f_\theta(x) > \rho_n^{\nu-1}(\alpha) \to 0$ and so $f_\theta \in \mathcal{A}_\alpha^{(n)}$ for all $|\theta| < \theta_n$ for large n.

Thus applying Lemma 4.5(a)

$$\begin{aligned}\mathcal{R}_\delta(V) &\geq \liminf_{n\to\infty} \inf_{\tilde{f}_n} \sup_{f_\theta} \mathbf{E}_{f_\theta} w_\delta\left(\sqrt{\frac{\pi n}{f_\theta(x)\tilde{s}_n}}\big(\tilde{f}_n(x) - f_\theta(x)\big)\big(1+o(1)\big)\right) \\ &= \liminf_{n\to\infty} \inf_{T_n} \sup_{|\theta|\leq\theta_n} \mathbf{E}_\theta w_\delta\Big(\big(T_n-\theta\big)\big(1+o(1)\big)\Big).\end{aligned}$$

Given $\theta_n \to \infty$ for $n \to \infty$ we can see that

$$\begin{aligned}\mathcal{R}_\delta(V) &\geq \liminf_{n\to\infty} \inf_{T_n} \sup_{|\theta|\leq K} \mathbf{E}_\theta w_\delta\Big(\big(T_n-\theta\big)\big(1+o(1)\big)\Big) \\ &\geq \liminf_{n\to\infty} \inf_{\tilde{T}_n} \sup_{|\theta|\leq K} \mathbf{E}_\theta w_\delta\Big(\big(\tilde{T}_n-\theta\big)\big(1+o(1)\big)\Big).\end{aligned}$$

for any K and taking the trimmed estimator $\tilde{T}_n = T_n \mathbb{1}(T_n \leq K)$. Now we can apply uniform continuity and (4.34) for $W(u) = w_\delta(u)$. Finally,

$$\begin{aligned}\liminf_{n\to\infty} \inf_{\tilde{f}_n} \sup_{f\in V\cap\mathcal{A}_\alpha^{(n)}} \mathbf{E} w\left(\sqrt{\frac{\pi n}{f(x)s_n}}\big(\tilde{f}_n(x)-f(x)\big)\right) &\geq \lim_{\delta\to 0}\mathcal{R}_\delta(V) \\ &\geq \lim_{\delta\to 0}\mathbf{E}\, w_\delta(\xi) = \mathbf{E}\, w(\xi). \qquad (4.35)\end{aligned}$$

Finally, from (4.24) and (4.35), the Theorem 4.1 is proved. □

4.4 Adaptive density estimation

4.4.1 The adaptive setting

Now we follow the same procedure as in the previous chapters in order to create the adaptive framework. We define first the scale of functions corresponding to a class of parameters $\mathcal{K}$, then we classify the functional scales with respect to the behavior of the optimum bandwidths $s_n(\alpha)$, corresponding to the different parameters α of the scale $\mathcal{K}$. Finally we introduce the adaptive estimator and prove optimality results.

In previous chapters we considered functional scales of the type

$$\mathcal{A}_\mathcal{K} = \Big\{ A(\alpha) \;\Big|\; \alpha \in \mathcal{K} \Big\}. \qquad (4.36)$$

However, we saw in the previous section of this chapter that in the case of density estimation a new problem arises. This is the problem of performing estimation in points where the density is very small. We thus consider ϵ-restricted functional scales. For $\epsilon > 0$, define the ϵ-restricted class

$$\mathcal{A}_\alpha^\epsilon = \mathcal{A}_\alpha^\epsilon(x) = \Big\{ f \in \mathcal{A}(\alpha) \;\Big|\; f(x) \geq \epsilon \Big\}. \qquad (4.37)$$

Note that $\mathcal{A}_\alpha^\epsilon \subset \mathcal{A}_\alpha^{(n)}$ for n sufficiently large. The parameter ϵ can be taken as small as necessary. Note, from Lemma 4.1, that ϵ at least must satisfy $\epsilon < C_{r,0}\frac{\beta}{\gamma}$.

Let $\mathcal{K}$ be any subset of parameters in $\mathcal{R}_+^3$ and consider the corresponding ϵ-restricted functional scale

$$\mathcal{A}_\mathcal{K}^\epsilon = \mathcal{A}_\mathcal{K}^\epsilon(x) = \Big\{ A_\alpha^\epsilon(x) \;\Big|\; \alpha \in \mathcal{K} \Big\}. \tag{4.38}$$

We shall study adaptive estimation in ϵ-restricted functional scales.

As in the previous chapters we shall restrict our adaptive study to the loss functions $\omega(x) = |x|^p$, $p > 0$ and consider the optimality criteria given in the following definition. Let $\mathcal{K}$, be a parameter class, possibly depending on n, and let $\mathcal{F}$ be a class of estimators $\tilde{f}_n$.

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

$$\limsup_{n\to\infty} \sup_{\alpha\in\mathcal{K}} \frac{\sup_{f\in\mathcal{A}_\alpha^\epsilon} \mathbf{E}_f \big|\hat{f}_n(x) - f(x)\big|^p}{\sup_{f\in\mathcal{A}_\alpha^\epsilon} \mathbf{E}_f \big|\tilde{f}_n(x) - f(x)\big|^p} \le 1.$$

As we have seen before this property depends on the classes $\mathcal{K}$ and $\mathcal{F}$ involved. The rate of convergence in estimating $f(x)$ over the whole scale $\mathcal{A}_{\mathbb{R}_3^+}$ can be of any order, thus certain restrictions to the class of parameters $\mathcal{K}$ are necessary.

Definition 4.3 *A functional scale $\mathcal{A}_{\mathcal{K}_n}^\epsilon$ (a scale $\mathcal{K}_n$) is called a regular, R-scale if*

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

$$0 < \liminf_{n\to\infty} \inf_{\alpha\in\mathcal{K}_n} \frac{\beta}{\gamma} \le \limsup_{n\to\infty} \sup_{\alpha\in\mathcal{K}_h} \frac{\beta}{\gamma} < \infty,$$

A regular scale is called a

(a) *regular-pseudo-parametric or RPP-scale if*

$$\limsup_{n\to\infty} \sup_{\alpha\in\mathcal{K}_h} \frac{1}{\gamma} \left(\log \frac{\beta^2 n}{\gamma} \right)^{1/r} < \infty,$$

(b) *regular-non-parametric or RNP-scale if*

$$0 < \liminf_{n\to\infty} \inf_{\alpha\in\mathcal{K}_n} \gamma \le \limsup_{n\to\infty} \sup_{\alpha\in\mathcal{K}_h} \gamma < \infty.$$

One can see that any RPP-scale satisfy

$$\limsup_{n\to\infty} \sup_{\alpha\in\mathcal{K}_n} s_n(\alpha) < \infty \tag{4.39}$$

while RNP-scales are non-parametric in the sense that

$$\lim_{n\to\infty} \inf_{\alpha\in\mathcal{K}_n} s_n(\alpha) = \infty \tag{4.40}$$

and the corresponding rates of convergence are slower than parametric rates. Note that $s_n(\alpha)$ goes uniformly to infinity at a logarithmic speed in RNP-scales when $n \to \infty$ (cf. eq. (4.6)).

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

$$\limsup_{n\to\infty} \sup_{\alpha\in\mathcal{K}_n} \sup_{f\in\mathcal{A}^\epsilon_\alpha} \mathbf{E}_f \left| n^{1/2}\big(\tilde{f}_n(x) - f(x)\big)\right|^p < \infty \tag{4.41}$$

for RPP-functional scales $\mathcal{A}_{\mathcal{K}_n}$.

In the next theorem we present an adaptive estimator $\hat{f}_n \in \mathcal{F}_p$ and prove it to be $(p, \epsilon, \mathcal{K}_n, \mathcal{F}_p)$-adaptively minimax for RNP-scales of functions.

4.4.2 The adaptive estimator and the upper bound

Theorem 4.2 *Let $\epsilon > 0$ and $\mathcal{K}_n$ be any RNP-scale. Consider the corresponding functional scale $\mathcal{A}^\epsilon_{\mathcal{K}_n}$. Then for all $p > 0$ there exists an adaptive estimator $\hat{f}_n$, no depending on α, such that for any $x \in \mathbb{R}$, $\hat{f}_n \in \mathcal{F}_p(x)$ and*

$$\limsup_{n\to\infty} \sup_{\alpha\in\mathcal{K}_n} \sup_{f\in\mathcal{A}^\epsilon_\alpha} \mathbf{E}_f \left| \psi_n^{-1}(\alpha)\big(\hat{f}_n(x) - f(x)\big)\right|^p \le 1$$

where

$$\psi_n^2(\alpha) = p\,\big(\log s_n(\alpha)\big)\frac{f(x)s_n(\alpha)}{\pi n}.$$

The estimator. Let us fix some parameters, $1/2 < l < 1$, $1/2 < \delta < 1$, $p_1 > 0$, $l_1 = \delta l$ and consider the sequence of bandwidths $s_0 = 0$, $s_i = \exp(i^l)$. For fixed n let us take the subsequence of bandwidth $\mathcal{S}_n = \big\{s_0, s_1, \ldots, s_{I_n}\big\}$ where

$$I_n = \arg\max_i \big\{s_i < n^{1/2}\big\}. \tag{4.42}$$

One can verify that for both, RPP and RNP-scales, $s_n(\alpha) \ll n^{1/2}$, thus for each α we can define $i(\alpha)$ such that

$$s_{i(\alpha)-1} < s_n(\alpha) \le s_{i(\alpha)}. \tag{4.43}$$

Consider the sequence of estimators $\hat{f}_i = \hat{f}_{n,s_i}$ defined in (4.2). As we have seen before, in Lepski's algorithm the selection of the adaptive estimator from the sequence of candidates $\hat{f}_i$ is based upon the comparison of $\hat{f}_j - \hat{f}_i$ with $\tau_j \mathbf{Var}^{1/2}(\hat{f}_j - \hat{f}_i)$ where $\tau_j = O(\log s_j)$. Now,

$$\mathbf{Var}(\hat{f}_j - \hat{f}_i) = \frac{1}{n}\mathbf{Var}\big(k_{s_j}(x - X) - k_{s_i}(x - X)\big).$$

Two natural estimators of this quantity come to mind immediately. One is

$$\hat{\sigma}^2_{i,j} = \frac{1}{n^2}\sum_{\ell=1}^{n}\big(k_{s_j}(x-X_\ell) - k_{s_i}(x-X_\ell)\big)^2 .$$

The other estimator follows from Lemma 4.2(d) and the application of the estimator $\hat{f}_j(x)$ of $f(x)$,

$$\tilde{\sigma}^2_{i,j} = \frac{(s_j - s_i)}{\pi n}\big|\hat{f}_j(x)\big|. \tag{4.44}$$

It turns out that $\tilde{\sigma}_{i,j}$ results in more theoretical results. It appears however that $\hat{\sigma}_{i,j}$ may be better in some practical implementations. Now, let us define the sequence of thresholds

$$\lambda_{i,j} = \tau_j \tilde{\sigma}_{i,j} \tag{4.45}$$

where

$$\tau_j^2 = p\log s_j + p_1 \log^\delta s_j$$

and

$$\hat{i} = \min\Big\{1 \le i \le I_n \ : \ \big|\hat{f}_j(x) - \hat{f}_i(x)\big| \le \lambda_{i,j} \quad \forall j \ (i \le j \le I_n)\Big\}. \tag{4.46}$$

We will prove that the estimator

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

satisfies Theorem 4.2.

Before proceeding with the proof let us define some new notation that will allow us to write more succinctly

$$k_i = k_{s_i}(x-X) \qquad k_i^\ell = k_{s_i}(x-X_\ell) \qquad \sigma^2_{k_i} = \mathbf{Var}_f\, k_i$$

$$\Delta_{i,j} = k_j - k_i \qquad \Delta^\ell_{i,j} = k^\ell_j - k^\ell_i \qquad \sigma^2_{\Delta_{i,j}} = \mathbf{Var}_f\, \Delta_{i,j}$$

$$b_i = \mathbf{E}_f k_i - f(x).$$

Auxiliary lemmas. In order to prove the theorem we need the following lemmas.

Lemma 4.6 *Let $\epsilon > 0$ and let $\mathcal{K}_n$ be any R-scale, then uniformly with respect to $f \in \mathcal{A}^\epsilon_\alpha$ and $\alpha \in \mathcal{K}_n$, and uniformly with respect to i, j when $n \to \infty$*

(a) $b_j^2 = o(1)\dfrac{f(x)s_j}{\pi n}$ *for all j such that $i(\alpha) \le j \le I_n$.*

(b) $(b_j - b_i)^2 = O(1)\dfrac{f(x)(s_j - s_i)}{\pi n}$, *for all j such that $i(\alpha) \le i \le j \le I_n$.*

(c) $\sigma^2_{k_j} = \dfrac{f(x)s_j}{\pi}\big(1 + o(1)\big)$ *when $s_j \to \infty$.*

(d) $\sigma^2_{\Delta_{i,j}} = \dfrac{f(x)(s_j - s_i)}{\pi}\big(1 + o(1)\big)$ *when* $(s_j - s_i) \to \infty$.

Proof.

(a) From Lemma 4.2(c) taking $s_1 = s_j$ we know that

$$b^2_{s_j} \le \frac{1}{2\pi^2}\int_{s_j}^{\infty} \frac{\beta^2}{\gamma} e^{-2(\gamma t)^r} dt,$$

thus, given $s_j \ge s_n$, using the Lemma 2.2, definition of s_n, and restriction $f(x) > \epsilon$, we get that uniformly

$$b^2_{s_j} \;\le\; \frac{f(x)s_j}{\pi n}\frac{e^{2(\gamma s_n)^r - 2(\gamma s_j)^r}}{f(x)r(\gamma s_n)^r} \;=\; \frac{f(x)s_j}{\pi n}o(1), \qquad (n \to \infty).$$

(b) Applying Lemma 4.2, (b) and (d), taking $s_1 = s_i$ and $s_2 = s_j$, the definition of s_n, condition $f(x) > \epsilon$, and the maximum inequality we get

$$\begin{aligned}(b_j - b_i)^2 \;&\le\; \frac{1}{2\pi^2}\int_{s_i}^{s_j}\frac{\beta^2}{\gamma}e^{-2(\gamma t)^r}dt \;\le\; \frac{(s_j - s_i)}{\pi n}\frac{\beta^2 n}{2\pi\gamma}e^{-2(\gamma s_i)^r} \\ &\le\; \epsilon^{-1}\frac{f(x)(s_j - s_i)}{\pi n}e^{2(\gamma s_n)^r - 2(\gamma s_i)^r} \;=\; O(1)\frac{f(x)(s_j - s_i)}{\pi n}.\end{aligned}$$

(c) It follows from Lemma 4.2, (c) and (d), taking $s_1 = 0$ and $s_2 = s_j$, condition $f(x) > \epsilon$, and conditions for RNP-scales. The procedure is the same as in (4.17). Note that the constants C_α and $\tilde{C}_\alpha$ which were used there are bounded in R-scales.

(d) It follows from Lemma 4.2, (b) and (d), condition $f(x) > \epsilon$, and conditions for R-scales in the same way as in (c).

□

Lemma 4.7 *Let $\mathcal{K}_n$ be any R-scale and consider the corresponding functional scale. Let $p > 0$, and $i \le i(\alpha)$ then, uniformly in any RPP-scale,*

$$\mathbf{E}\left|\tilde{\sigma}_{i,i(\alpha)}\right|^p = O\left(n^{-p/2}\right).$$

Let $\epsilon > 0$, then uniformly in any ϵ-restricted RNP-scale $\mathcal{A}^{\epsilon}_{\mathcal{K}_n}$

$$\mathbf{E}\left|\tilde{\sigma}_{i,i(\alpha)}\right|^p \le \left(\frac{f(x)(s_{i(\alpha)} - s_i)}{\pi n}\right)^{p/2}(1 + o(1)).$$

Proof. First let us remind that according to definition (4.44)

$$\mathbf{E}\left|\tilde{\sigma}_{i,i(\alpha)}\right|^p = \left(\frac{s_{i(\alpha)}-s_i}{\pi n}\right)^{p/2}\mathbf{E}\left|\frac{1}{n}\sum_{\ell=1}^{n}k_{i(\alpha)}^{\ell}\right|^{p/2}.$$

We know that $s_{i(\alpha)}$ is uniformly bounded in RPP-scales $\mathcal{K}_n$, cf. (4.39). It follows by Lemma 4.2(a) and the previous argument that $k_{i(\alpha)}$ is also uniformly bounded there. Thus, uniformly in RPP-scales

$$\mathbf{E}\left|\tilde{\sigma}_{i,i(\alpha)}\right|^p = O\left(n^{-p/2}\right).$$

Now, let us consider the RNP-scales. For the case $0 < p < 2$: Applying c_r-inequality ($p/2 < 1$) and some transformations

$$\begin{aligned}\mathbf{E}\left|\tilde{\sigma}_{i,i(\alpha)}\right|^p \le{}& \left(\frac{f(x)(s_{i(\alpha)}-s_i)}{\pi n}\right)^{p/2}\left(\frac{\pi}{f(x)n}\right)^{p/2}\mathbf{E}\left|\sum_{\ell=1}^{n}\left(k_{i(\alpha)}^{\ell}+\mathbf{E}\,k_{i(\alpha)}\right)\right|^{p/2}\\ &+\left(\frac{s_{i(\alpha)}-s_i}{\pi n}\mathbf{E}\,k_{i(\alpha)}\right)^{p/2}. \qquad (4.47)\end{aligned}$$

Denote $Z_{i(\alpha)}^{\ell}=k_{i(\alpha)}^{\ell}-\mathbf{E}k_{i(\alpha)}$. Let us first analize the first term of the previous equation. For some fixed p' ($2<p'<4$), applying the inequality for the moments and Bretagnolle-Huber inequality ($0 \le p/2 \le 1$), see Appendix,

$$\begin{aligned}\big(f(x)n\big)^{-p/2}\mathbf{E}\left|\sum_{\ell=1}^{n}Z_{i(\alpha)}^{\ell}\right|^{p/2} &\le \big(f(x)n\big)^{-p/2}\left(\mathbf{E}^{-p'/2}\left|\sum_{\ell=1}^{n}Z_{i(\alpha)}^{\ell}\right|^{p'/2}\right)^{p/2}\\ &\le \big(f(x)n\big)^{-p/2}\left(\mathcal{C}_{p'}^{2/p'}\sigma_{k_{i(\alpha)}}\,n^{1/2}\right)^{p/2}\\ &\le \mathcal{C}_p\big(f(x)n\big)^{-p/2}\sigma_{k_{i(\alpha)}}^{p/2}\,n^{p/4}.\end{aligned}$$

Now, applying Lemma 4.6(c) and condition $f(x) > \epsilon$

$$\begin{aligned}\big(f(x)n\big)^{-p/2}\mathbf{E}\left|\sum_{\ell=1}^{n}Z_{i(\alpha)}^{\ell}\right|^{p/2} &\le C_p\big(f(x)n\big)^{-p/2}\left(\frac{f(x)s_{i(\alpha)}}{\pi}\right)^{p/4}\big(1+o(1)\big)\,n^{p/4}\\ &= O\left(\frac{s_{i(\alpha)}}{n}\right)^{p/4}\big(1+o(1)\big) \;=\; o(1).\end{aligned}$$

On the second term of equation (4.47), $\mathbf{E}\,k_{i(\alpha)} = f(x)(1+o(1))$ when $s_{i(\alpha)} \to \infty$, thus

$$\mathbf{E}\left|\tilde{\sigma}_{i,i(\alpha)}\right|^p \le \left(\frac{f(x)(s_{i(\alpha)}-s_i)}{\pi n}\right)^{p/2}(1+o(1)).$$

When $p > 2$ we can apply λ-inequality (cf. Appendix) and see that

$$\mathbf{E}\left|\tilde{\sigma}_{i,i(\alpha)}\right|^p \;\le\; \frac{1}{\lambda^{p/2-1}}\left(\frac{f(x)(s_{i(\alpha)}-s_i)}{\pi n}\right)^{p/2}\left(\frac{\pi}{f(x)n}\right)^{p/2}\mathbf{E}\Big|\sum_{\ell=1}^{n} Z^{\ell}_{i(\alpha)}\Big|^{p/2} + \frac{1}{(1-\lambda)^{p/2-1}}\left(\frac{s_{i(\alpha)}-s_i}{\pi n}\mathbf{E}\,k_{i(\alpha)}\right)^{p/2}.$$

The value of λ is taken small enough at our convenience. Using again the fact that $\mathbf{E}\,k_{i(\alpha)} = f(x)(1+o(1))$ when $s_{i(\alpha)} \to \infty$

$$\mathbf{E}\left|\tilde{\sigma}_{i,i(\alpha)}\right|^p \;\le\; \frac{1}{\lambda^{p/2-1}}\left(\frac{f(x)(s_{i(\alpha)}-s_i)}{\pi n}\right)^{p/2}\left(\frac{\pi}{f(x)n}\right)^{p/2}\mathbf{E}\Big|\sum_{\ell=1}^{n} Z^{\ell}_{i(\alpha)}\Big|^{p/2} + \frac{1}{(1-\lambda)^{p/2-1}}\left(\frac{f(x)(s_{i(\alpha)}-s_i)}{\pi n}\right)^{p/2}\big(1+o(1)\big).$$

Now we do as before but we must make first a distinction between the following two subcases.

The case $2 < p \le 4$ is in principle similar to the previous one, except that it was necessary to apply the λ-inequality. We take $\lambda \to 0$ as slow as required in order to get the desired result. The rest is the same.

For $p > 4$ we see that applying Bretagnolle-Huber inequality ($p/2 \ge 2$) using the bound given in Lemma 4.2(a), Lemma 4.6(c) and condition $f(x) > \epsilon$

$$\begin{aligned}\big(f(x)n\big)^{-p/2}\mathbf{E}\Big|\sum_{\ell=1}^{n} Z^{\ell}_{i(\alpha)}\Big| \;&\le\; \big(f(x)n\big)^{-p/2}\Big(\mathcal{C}_{p,1}\,f(x)ns^{p/2-1}_{i(\alpha)} + \mathcal{C}_{p,2}\,\sigma^{p/2}_{k_{i(\alpha)}}\,n^{p/4}\Big)\\ &\le\; C_{p,1}\left(\frac{s_{i(\alpha)}}{f(x)n}\right)^{p/2-1} + o(1) \;=\; o(1).\end{aligned}$$

Taking an appropriate λ such that it goes to zero slow enough we conclude that

$$\mathbf{E}\left|\tilde{\sigma}_{i,i(\alpha)}\right|^p \le \left(\frac{f(x)(s_{i(\alpha)}-s_i)}{\pi n}\right)^{p/2}(1+o(1)).$$

□

Proof of the theorem. The idea of the proof is the same as the proofs of the equivalent theorems in the previous chapters. For fixed $\alpha \in \mathcal{K}_n$, we consider two major mutually exclusive events, $\hat{i} \le i(\alpha)$ and $\hat{i} > i(\alpha)$. For the case $\hat{i} \le i(\alpha)$ we can see that the difference $\hat{f}_{i(\alpha)} - \hat{f}_{\hat{i}}$ is bounded by $\lambda_{\hat{i},i(\alpha)}$. If f belongs to a RPP -scale this bound is of order $O(n^{-1/2})$, while for f in a RNP-functional scale it is of order $O(\frac{s_n}{n}\log s_n)^{1/2}$. For the case $\hat{i} > i(\alpha)$ we prove that no bandwidth s_i, $i > i(\alpha)$ will be rejected with high probability cause in

those cases the bias is of small order with respect to the variance. Thus the risk in that case, will be proved to be of order $O(n^{-p/2})$, uniformly in RPP as well as RNP-scales.

For any f in a functional scale $\mathcal{A}_{\mathcal{K}_n}$,

$$\begin{aligned}
\mathbf{R}^n(f) &:= \mathbf{E}_f \left|\hat{f}_{\hat{i}} - f(x)\right|^p \\
&= \mathbf{E}_f \left\{\left|\hat{f}_{\hat{i}} - f(x)\right|^p \mathbb{1}_{\{\hat{i}\le i(\alpha)\}}\right\} + \mathbf{E}_f \left\{\left|\hat{f}_{\hat{i}} - f(x)\right|^p \mathbb{1}_{\{\hat{i}> i(\alpha)\}}\right\} \\
&= \mathbf{R}^n_-(f) + \mathbf{R}^n_+(f).
\end{aligned}$$

Let us examine $\mathbf{R}^n_-(f)$ first. For the case $p \ge 1$: Applying the λ-inequality for $p \ge 1$

$$\begin{aligned}
\mathbf{R}^n_-(f) &\le \mathbf{E}\left\{\left[\frac{1}{(1-\lambda)^{p-1}}\left|\hat{f}_{\hat{i}}(x) - \hat{f}_{i(\alpha)}(x)\right|^p + \frac{1}{\lambda^{p-1}}\left|\hat{f}_{i(\alpha)}(x) - f(x)\right|^p\right] \mathbb{1}_{\{\hat{i}\le i(\alpha)\}}\right\} \\
&\le \frac{1}{(1-\lambda)^{p-1}}\mathbf{E}\left|\hat{f}_{\hat{i}}(x) - \hat{f}_{i(\alpha)}(x)\right|^p \mathbb{1}_{\{\hat{i}\le i(\alpha)\}} + \frac{1}{\lambda^{p-1}}\mathbf{E}\left|\hat{f}_{i(\alpha)}(x) - f(x)\right|^p \\
&\le \frac{1}{(1-\lambda)^{p-1}}\mathbf{E}\left|\lambda_{i(\alpha)}\tilde{\sigma}_{\hat{i},i(\alpha)}\right|^p + \frac{1}{\lambda^{p-1}}\mathbf{E}\left|\hat{f}_{i(\alpha)}(x) - f(x)\right|^p.
\end{aligned}$$

On one hand, from Lemma 4.7 we know that for RPP-scales

$$\mathbf{E}\left|\tilde{\sigma}_{i,i(\alpha)}\right|^p = O\left(n^{-p/2}\right).$$

On the other hand, applying Lemma 4.2(a) we see that $\sigma^2_{k_{i(\alpha)}}$ is uniformly bounded in RPP-scales, while according to Lemma 4.6(a) we have that $b^2_{i(\alpha)} = o(1)\sigma^2_{k_{i(\alpha)}}$ so

$$\mathbf{E}\left|\hat{f}_{i(\alpha)}(x) - f(x)\right|^p = O\left(n^{-p/2}\right)$$

uniformly in RPP-scales and thus

$$\mathbf{R}^n_-(f) = O\left(n^{-p/2}\right), \quad (n \to \infty).$$

Now, from Lemma 4.7 we know that uniformly in RNP-scales

$$\mathbf{E}\left|\tilde{\sigma}_{\hat{i},i(\alpha)}\right|^p < \left(\frac{f(x)s_{i(\alpha)}}{\pi n}\right)^{p/2}(1+o(1))$$

and from Lemma 4.6(a) we see that in RNP-scales there exists C_p such that

$$\mathbf{E}\left|\sqrt{\frac{\pi n}{f(x)s_{i(\alpha)}}}\left(\hat{f}_{i(\alpha)}(x) - f(x)\right)\right|^p \le C_p.$$

Thus, if we apply the dominated convergence theorem for an appropriate λ going to zero, and the asymptotic equivalence $s_{i(\alpha)} \sim s_n(\alpha)$, we get that uniformly in RNP-scales

$$\mathbf{R}^n_-(f) \le \left(p\big(\log s_n(\alpha)\big)\ \frac{f(x) s_n(\alpha)}{\pi n} \right)^{p/2} \big(1 + o(1)\big), \qquad (n \to \infty). \tag{4.48}$$

The case $0 < p < 1$ is proved, in the same way, using the c_r-inequality.

In order to bound $\mathbf{R}^n_+(f)$ let us define the following event:

$$B_i = \Big\{ \omega : \big|\hat{f}_i(x) - f(x)\big| \ < \ \tau_i \sqrt{2n^{-1}\,\mathbf{E}\, k^2_{s_i}} \Big\} \quad \forall i, \tag{4.49}$$

One can see after applying (4.49), Cauchy-Schwartz inequality and Lemma 4.2(d) that

$$\begin{aligned}
\mathbf{R}^n_+(f) \ &= \ \mathbf{E}\Big\{ \big|\hat{f}_{\hat{i}}(x) - f(x)\big|^p \, \mathbb{1}_{\{\hat{i} > i(\alpha)\}} \Big\} \ = \ \sum_{i=i(\alpha)+1}^{I_n} \mathbf{E}\Big\{ \big|\hat{f}_i(x) - f(x)\big|^p \, \mathbb{1}_{\{\hat{i}=i\}} \Big\} \\
&= \ \sum_{i=i(\alpha)+1}^{I_n} \mathbf{E}\Big\{ \big|\hat{f}_i(x) - f(x)\big|^p \, \big(\mathbb{1}_{\{\hat{i}=i\}\cap B_i} \ + \mathbb{1}_{\{\hat{i}=i\}\cap B_i^c} \big) \Big\} \\
&\le \ \sum_{i=i(\alpha)+1}^{I_n} \Big(\tau_i \sqrt{2\, n^{-1} \mathbf{E}\, k^2_{s_i}} \Big)^p \, \mathbf{P}(\hat{i} = i) \\
&\qquad\qquad + \sum_{i=i(\alpha)+1}^{I_n} \mathbf{E}^{1/2} \Big| \hat{f}_i(x) - f(x) \Big|^{2p} \, \mathbf{P}^{1/2}(B_i^c).
\end{aligned}$$

Applying Lemma 4.6, after few transformations, we get that

$$\begin{aligned}
\mathbf{R}^n_+(f) \ &\le \ O\left(n^{-p/2}\right) \sum_{i=i(\alpha)+1}^{I_n} \left(\frac{\mathbf{E}\, k^2_{s_i}}{s_i} \right)^{p/2} (\tau_i^2 s_i)^{p/2} \mathbf{P}(\hat{i} = i) \\
&\qquad + \sum_{i=i(\alpha)+1}^{I_n} \left(\frac{f(x) s_i}{\pi n} \right)^{p/2} \mathbf{E}^{1/2} \left| \sqrt{\frac{\pi n}{f(x) s_i}} \big(\hat{f}_i(x) - f(x) \big) \right|^{2p} \mathbf{P}^{1/2}(B_i^c) \\
&\le \ O\left(n^{-p/2}\right) \left(\sum_{i=i(\alpha)+1}^{I_n} (s_i \log s_i)^{p/2} \, \mathbf{P}(\hat{i} = i) + \sum_{i=i(\alpha)+1}^{I_n} s_i^{p/2} \, \mathbf{P}^{1/2}(B_i^c) \right).
\end{aligned}$$

The rest of the proof is just to show that both sums are of order $O(1)$ when n goes to infinity. We shall show that for $i > i(\alpha)$ the probabilities of the events $\{\hat{i} = i\}$ and B_i^c are exponentially small.

From definition of $\hat{i}$ (cf. eq. (4.46)) one can see that

$$\mathbf{P}(\hat{i} = i) \ \le \ \sum_{j=i+1}^{I_n} \mathbf{P}\Big(\big|\hat{f}_{j-1}(x) - \hat{f}_{i-1}(x)\big| > \lambda_{i-1,j-1} \Big).$$

Applying Lemmas 4.6, (b) and (d), we can see that for some $C > 0$

$$\mathbf{P}\Big(\big|\hat{f}_j(x) - \hat{f}_i(x)\big| > \tau_j\, \tilde{\sigma}_{i,j} \Big) \;\le\; \mathbf{P}\left(\Big|\frac{1}{n}\sum_{\ell=1}^{n} \Delta_{i,j}^{\ell}\Big| > \tau_j \sqrt{\frac{s_j - s_i}{\pi n^2}\Big|\sum_{\ell=1}^{n} k_j^{\ell}\Big|} \right)$$

$$\le\; \mathbf{P}\left(\Big|\frac{1}{\sqrt{n}\,\sigma_{\Delta_{i,j}}}\sum_{\ell=1}^{n}(\Delta_{i,j}^{\ell} - \mathbf{E}\Delta_{i,j})\Big| > \tau_j\,\big(1+o(1)\big)\sqrt{f(x)^{-1}\,\frac{1}{n}\Big|\sum_{\ell=1}^{n} k_j^{\ell}\Big| - \frac{\sqrt{n}|b_j - b_i|}{\sigma_{\Delta_{i,j}}}} \right)$$

$$\le\; \mathbf{P}\left(\Big|\frac{1}{\sqrt{n}\,\sigma_{\Delta_{i,j}}}\sum_{\ell=1}^{n}(\Delta_{i,j}^{\ell} - \mathbf{E}\Delta_{i,j})\Big| > \tau_j\,\big(1+o(1)\big)\sqrt{f(x)^{-1}\,\frac{1}{n}\Big|\sum_{\ell=1}^{n} k_j^{\ell}\Big| - C} \right) \qquad (4.50)$$

For $j > i(\alpha)$ define the event

$$C_j \;:=\; \left\{ \omega : \Big| \frac{1}{n}\sum_{\ell=1}^{n} k_j^{\ell} - f(x) \Big| \le \epsilon_n\, f(x) \right\}, \qquad (4.51)$$

where the sequence $\epsilon_n = o(1)$, $(n \to \infty)$. Considering the conditional probability on the event C_j

$$\mathbf{P}\Big(\big|\hat{f}_j(x) - \hat{f}_i(x)\big| > \lambda_{i,j}\Big) \;\le\; \mathbf{P}\Big(\big|\hat{f}_j(x) - \hat{f}_i(x)\big| > \lambda_{i,j} \;\Big|\; C_j\Big) + P(C_j^c).$$

When the event C_j is true, $\frac{1}{n}\big|\sum_{\ell=1}^{n} k_j^{\ell}\big| \ge (1-\epsilon_n) f(x)$. Thus, if we substitute this in eq. (4.50),

$$\mathbf{P}\Big(\big|\hat{f}_j(x) - \hat{f}_i(x)\big| > \lambda_{i,j} \;\Big|\; C_j\Big) \le$$

$$\le\; \mathbf{P}\left(\Big|\frac{1}{\sqrt{n}\,\sigma_{\Delta_{i,j}}}\sum_{\ell=1}^{n}(\Delta_{i,j}^{\ell} - \mathbf{E}\Delta_{i,j})\Big| > \tau_j\,\big(1+o(1)\big)(1-\epsilon_n) - C \right)$$

$$=\; \mathbf{P}\left(\Big|\frac{1}{\sqrt{n}\,\sigma_{\Delta_{i,j}}}\sum_{\ell=1}^{n}(\Delta_{i,j}^{\ell} - \mathbf{E}\Delta_{i,j})\Big| > \tau_j\,\big(1+o(1)\big) - C \right).$$

Now, applying Bernstein's inequality,

$$\mathbf{P}\left(\Big|\frac{1}{\sqrt{n}\sigma_{\Delta_{i,j}}}\sum_{\ell=1}^{n}(\Delta_{i,j}^{\ell} - \mathbf{E}\Delta_{i,j})\Big| > \tau_j\,\big(1+o(1)\big) - C \right) \;\le\; \exp\left\{ -\frac{(\tau_j\big(1+o(1)\big) - C)^2}{2(1+F_n)} \right\}$$

where using $B_{\Delta_{i,j}} = \frac{2(s_j - s_i)}{\pi}$ as a bound of $\Delta_{i,j} - \mathbf{E}\Delta_{i,j}$ (cf. Lemma 4.2(a)), the asymptotic $\sigma^2_{\Delta_{i,j}} = \frac{f(x)(s_j - s_i)}{\pi}\big(1+o(1)\big)$ (cf. Lemma 4.2(d)), condition $f(x) > \epsilon$ and definition of τ_j, we have

$$F_n = \frac{B_{\Delta_{i,j}}(\tau_j(1+\tilde{\epsilon}_n) - 1)}{3\sigma_{\Delta_{i,j}}\sqrt{n}} = O\Big(\frac{(s_j - s_i)\log s_j}{n}\Big)^{1/2} = o(1), \qquad (n \to \infty).$$

Thus one can see that after few transformations

$$\mathbf{P}\left(\left|\frac{1}{\sqrt{n}\,\sigma_{\Delta_{i,j}}}\sum_{\ell=1}^{n}(\Delta_{i,j}^{\ell}-\mathbf{E}\Delta_{i,j})\right|>\tau_j\left(1+o(1)\right)-C\right)\le\ 2\,\exp\left\{-\frac{1+o(1)}{2}(\,\tau_j-C_1\,)^2\right\}$$

$$\le\ 2\,\exp\left\{-\frac{1+o(1)}{2}(\,\sqrt{pj^l+p_1j^{l_1}}-C_1\,)^2\right\}$$

$$\le\ 2\,\exp\left\{\left(1+o(1)\right)\left(-\frac{pj^l+p_1j^{l_1}}{2}+\sqrt{pj^l+p_1j^{l_1}}\right)\right\}$$

$$\le\ 2\,\exp\left\{\left(1+o(1)\right)\left(-\frac{pj^l}{2}-\frac{p_1j^{l_1}}{3}\right)\right\}$$

for large enough $\bar{i}$, and all $i>\max(\bar{i},i(\alpha))$. Now,

$$\mathbf{P}(\hat{i}=i)\ \le\ 2\sum_{j=i}^{\infty}\exp\left\{\left(1+o(1)\right)\left(-\frac{pj^l}{2}-\frac{p_1j^{l_1}}{3}\right)\right\}+\sum_{j=i}^{I_n}\mathbf{P}(C_j^c),$$

but for some $C_2>0$ this is

$$\mathbf{P}(\hat{i}=i)\le\ C_2i^{1-l}\exp\left\{\left(-\frac{pi^l}{2}-\frac{p_1i^{l_1}}{3}\right)\right\}+\sum_{j=i}^{I_n}\mathbf{P}(C_j^c)$$

$$\le\ C_2i^{1-l}s_i^{-p/2}\exp\{-p_1i^{l_1}/3\}+\sum_{j=i}^{I_n}\mathbf{P}(C_j^c)$$

$$\le\ C_3s_i^{-p/2}\,\exp\{-p_1i^{l_1}/4\}+\sum_{j=i}^{I_n}\mathbf{P}(C_j^c).$$

On the other hand doing some transformations and applying Lemma 4.6(c) we obtain

$$\mathbf{P}(C_j^c)\ =\ \mathbf{P}\left(\left|\frac{1}{n}\sum_{\ell=1}^{n}k_j^{\ell}-f(x)\right|>\epsilon_nf(x)\right)$$

$$\le\ \mathbf{P}\left(\frac{1}{n}\left|\sum_{\ell=1}^{n}k_j^{\ell}-\mathbf{E}\,k_j\right|>\epsilon_n\,f(x)-|b_j|\right)$$

$$=\ \mathbf{P}\left(\frac{1}{\sqrt{n}\sigma_{k_j}}\left|\sum_{\ell=1}^{n}k_j^{\ell}-\mathbf{E}\,k_j\right|>\frac{\epsilon_nf(x)\sqrt{n}}{\sigma_{k_j}}-\left(\frac{n\,b_j^2}{\sigma_{k_j}^2}\right)^{1/2}\right)$$

$$=\ \mathbf{P}\left(\frac{1}{\sqrt{n}\sigma_{k_j}}\left|\sum_{\ell=1}^{n}k_j^{\ell}-\mathbf{E}\,k_j\right|>\frac{\epsilon_nf(x)\sqrt{n}}{\sigma_{k_j}}-1\right)$$

where, using $B_{k_j} = \frac{2s_j}{\pi}$ as a bound of centered k_j, the asymptotic $\sigma^2_{k_j} = \frac{f(x)s_j}{\pi}\big(1+o(1)\big)$, condition $f(x) > \epsilon$ and $s_j \le n^{1/2}$ we have

$$\mathbf{P}(C_j^c) \le 2\,\exp\left\{-\frac{\epsilon_n^2 f(x) n}{C s_j}\right\} \le 2\exp\left\{-\frac{\epsilon_n^2 \epsilon n}{C n^{1/2}}\right\}.$$

Thus, taking $\epsilon_n = O(\log^{-1} n)$

$$\mathbf{P}(C_j^c) \le 2\,\exp\left\{-n^{1/3}\right\}$$

for n large enough. Finally we see that

$$\begin{aligned}
\sum_{i=i(\alpha)+1}^{I_n} (s_i \log s_i)^{p/2}\,\mathbf{P}(\hat{i}=i) &\le \sum_{i=i(\alpha)+1}^{I_n} \log^{p/2} s_i\,\exp\{-p_1 i^{l_1}/4\} \\
&\quad + \sum_{i=i(\alpha)+1}^{I_n} (s_i \log s_i)^{p/2} \sum_{j=i}^{I_n} \mathbf{P}(C_j^c) \\
&\le \sum_{i=1}^{\infty} i^{lp/2}\,\exp\{-p_1 i^{l_1}/4\} + \sum_{i=i(\alpha)+1}^{I_n} 2 I_n\, s_i^p\,\exp\{-n^{1/3}\} \\
&\le C_{p,p_1} + 2 I_n^2\, n^p\,\exp\{-n^{1/3}\} \;=\; O(1)
\end{aligned}$$

for $n \to \infty$.

$$\begin{aligned}
\mathbf{P}(B_i^c) &= \mathbf{P}\left(\left|\hat{f}_i(x) - f(x)\right| \ge \tau_i \sqrt{2\,\frac{1}{n}\mathbf{E}\,k_i^2}\right) \\
&= \mathbf{P}\left(\left|\frac{1}{n}\sum_{\ell=1}^{n} k_i^\ell - f(x)\right| \ge \tau_i \sqrt{2\,\frac{1}{n}\mathbf{E}\,k_i^2}\right) \\
&\le \mathbf{P}\left(\left|\frac{1}{n}\sum_{\ell=1}^{n} (k_i^\ell - \mathbf{E}k_i)\right| \ge \tau_i \sqrt{2\,\frac{1}{n}\sigma_{k_i}^2} - |b_i|\right) \\
&\le \mathbf{P}\left(\frac{1}{\sqrt{n}\,\sigma_{k_i}}\left|\sum_{\ell=1}^{n} (k_i^\ell - \mathbf{E}k_i)\right| \ge \sqrt{2\tau_i^2} - \left(\frac{n b_i^2}{\sigma_{k_i}^2}\right)^{1/2}\right)
\end{aligned}$$

but applying Lemma 4.6(c) and Bernstein's inequality we see that

$$\begin{aligned}
\mathbf{P}(B_i^c) &\le \mathbf{P}\left(\frac{1}{\sqrt{n}\,\sigma_{k_{s_i}}}\Big|\sum_{\ell=1}^{n} (k_i^\ell - \mathbf{E}k_i)\Big| \ge \sqrt{2\tau_i^2} - 2\right) \\
&\le 2\,\exp\left\{-\frac{(\sqrt{2\,\tau_i^2} - 2)^2}{2(1+F_n)}\right\}
\end{aligned}$$

where, using $B_{k_i} = \frac{2s_i}{\pi}$ as a bound of centered k_i, the asymptotic $\sigma^2_{k_i} = \frac{f(x)s_i}{\pi}\big(1+o(1)\big)$, and condition $f(x) > \epsilon$ we then have

$$F_n \;=\; O\left(\frac{B_i\tau_i}{n^{1/2}\sigma_{k_i}}\right) = O\left(\frac{s_i \log s_i}{f(x)n}\right)^{1/2} = O\left(\frac{s_i \log s_i}{\epsilon n}\right)^{1/2} = o(1).$$

Applying the continuity of the exponential

$$\begin{aligned}
\mathbf{P}(B_i^c) &\leq 2\,\exp\left\{-\frac{1+o(1)}{2}\left(\sqrt{2\,(pi^l + p_1 i^{l_1})} - 2\right)^2\right\} \\
&\leq 2\,\exp\left\{\big(1+o(1)\big)\left(-pi^l - p_1 i^{l_1} + 2\sqrt{2\,(pi^l + p_1 i^{l_1})}\right)\right\} \\
&\leq 2\,\exp\left\{\big(1+o(1)\big)\left(-pi^l - p_1 i^{l_1}/2\right)\right\} \\
&= C_4\, s_i^{-p}\,\exp\left\{-p_1 i^{l_1}/2\right\}.
\end{aligned}$$

for some C_4. Thus

$$\sum_{i=i(\alpha)+1}^{I_n} s_i^{p/2}\mathbf{P}^{1/2}(B_i^c) \;\leq\; O(1)\sum_{i=i(\alpha)+1}^{\infty}\exp\left\{-p_1 i^{l_1}/4\right\} \;=\; O(1).$$

□.

4.4.3 Lower bound: optimality results

Theorem 4.3 *Let $\epsilon > 0$ and $\mathcal{K}_n$ be any RNP-scale. For each $\alpha \in \mathcal{K}_n$ and $f \in \mathcal{A}^\epsilon_\alpha$ consider $s_n = s_n(\alpha)$, as defined in (4.6), and denote*

$$\psi_n^2(\alpha) = p\big(\log s_n(\alpha)\big)\frac{f(x)s_n(\alpha)}{\pi n}. \tag{4.52}$$

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

$$\lim_{n\to\infty}\inf_{\alpha\in\mathcal{K}_n}\sup_{f\in\mathcal{A}^\epsilon_\alpha}\mathbf{E}\left|\psi_n^{-1}(\alpha)\big(\tilde{f}_n(x) - f(x)\big)\right|^p \geq 1. \tag{4.53}$$

Proof. First we show that it is possible to find a function $f_0 \in \mathcal{A}^\epsilon_\alpha(x)$, for every $\alpha \in \mathcal{K}_n$, for which there exists an estimator with a parametric rate of convergence. Let us take the density function

$$f_a(y) = \frac{1-\cos^2 y}{\pi y^2}$$

and consider the rescaling of f_a

$$f_0(y) = \frac{1}{m}f_a\left(\frac{y-x}{m}\right).$$

The Fourier transform of the function f_0 satisfies

$$\mathcal{F}[f_0](t) = 1 - |mt| \quad \text{for} \quad |t| \le \frac{1}{m}$$

and is zero otherwise. Thus for any bandwidth $s_0 > \frac{1}{m}$ the kernel estimator (4.2) is unbiased, see equation (4.10). Our goal now is to prove that there exists $m = m(\alpha)$ such that $\frac{1}{m}$ is uniformly bounded in $\mathcal{K}_n$ and $f_0 \in \mathcal{A}_\alpha^\epsilon$. First we can easily make the norm $\|f_0\|_\alpha$ less than 1. According to the definition of RNP-scales there exist finite positive constants $\gamma_0, \gamma_1, \beta_0, \beta_1, r_0, r_1$ such that for n big enough $\mathcal{K}_n \in [\gamma_0, \gamma_1] \times [\beta_0, \beta_1] \times [r_0, r_1]$. Now,

$$\begin{aligned} \|f_0\|_\alpha &= \int \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} |\mathcal{F}[f_0](t)|^2 dt \le \int_{-1/m}^{1/m} \frac{\gamma}{\beta^2} e^{2|\gamma t|^r} dt \\ &\le \frac{2}{\beta^2} \int_0^{\gamma/m} e^{2t^r} dt \le \frac{2}{\beta_0^2} \int_0^{\gamma_1/m} e^{2t^{r_1}} dt. \end{aligned}$$

Denote $c_0 = \frac{\gamma_1}{m}$ and take it small enough such that

$$\frac{2}{\beta_0^2} \int_0^{c_0} e^{2t^{r_1}} dt < 1.$$

Moreover by definition,

$$f_0(x) = \frac{1}{m} f_a(0) = \frac{1}{2\pi m} = \frac{c_0}{2\pi} \min\left(1, \frac{1}{\gamma}\right).$$

Therefore both the conditions $f_0(x) > \epsilon$ and $\|f_0\|_\alpha < 1$ are satisfied if $2\pi\epsilon \le \frac{1}{m} \le \frac{c_0}{\gamma_1}$. Hence, with this choice of m, $f_0 \in \mathcal{A}_\alpha^\epsilon$.

Now, let us take $\tilde{s} = \tilde{s}_n(\alpha)$ as it was defined in (4.26) and define $\tilde{\phi}_\alpha = p \log \tilde{s}$. For $\theta = \tilde{\phi}_\alpha - \sqrt{\tilde{\phi}_\alpha}$, let us consider the function f_1 given by

$$f_1(y) = f_0(y) \left(1 + \theta \sqrt{\frac{\pi}{f_0(x) n \tilde{s}}} \left(k_{\tilde{s}}(x - y) - \mathbf{E}_0 k_{\tilde{s}}(x - X)\right)\right),$$

where $\mathbf{E}_0$ denotes the expectation with respect to f_0. Note that f_1 belongs to the family defined in (4.27). From (4.52) one can see that $\theta \le \tilde{s}^{1/4}$ for n big enough. Thus, as we saw in Lemma 4.5(a)

$$f_1(y) = f_0(y)\big(1 + o(1)\big), \qquad (n \to \infty), \tag{4.54}$$

and $f_1 \in \mathcal{A}_\alpha^\epsilon(x)$ for all n big enough. Next we can derive a representation for the log-likelihood

$$L(\theta) = \log \frac{d\mathbf{P}_1^{(n)}}{d\mathbf{P}_0^{(n)}}$$

just as we did in Lemma 4.5(c). If we denote

$$\bar{Y}_i = k_{\tilde{s}}(x - X_i) - \mathbf{E}_1 k_{\tilde{s}}(x - X),$$

and

$$Y_i = k_{\tilde{s}}(x - X_i) - \mathbf{E}_0 k_{\tilde{s}}(x - X),$$

we can verify that

$$\begin{aligned} Y_i &= \bar{Y}_i + \big(\mathbf{E}_1 k_{\tilde{s}}(x - X) - \mathbf{E}_0 k_{\tilde{s}}(x - X)\big) = \bar{Y}_i + \int k_{\tilde{s}}(x - y)\big(f_1(y) - f_0(y)\big)\, dy \\ &= \bar{Y}_i + \theta\sqrt{\frac{\pi}{f_0(x) n \tilde{s}}}\, \mathbf{Var}_0\, k_{\tilde{s}}(x - X) = \bar{Y}_i + \theta\sqrt{\frac{f_0(x)\tilde{s}}{\pi n}}\,\big(1 + o(1)\big), \quad (n \to \infty). \end{aligned}$$

As in Lemma 4.5(c), cf. (4.33), we find

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{n} \log\left(1 + \theta\sqrt{\frac{\pi}{f_0(x) n \tilde{s}_n}} Y_i\right) \\ &= \sum_{i=1}^{n} \log\left(1 + \theta\sqrt{\frac{\pi}{f_0(x) n \tilde{s}}}\Big(\bar{Y}_i + \theta\sqrt{\frac{f_0(x)\tilde{s}}{\pi n}}\,\big(1 + o(1)\big)\Big)\right) \\ &= \sum_{i=1}^{n} \log\left(1 + \theta\sqrt{\frac{\pi}{f_0(x) n \tilde{s}}}\bar{Y}_i + \frac{\theta^2}{n}\big(1 + o(1)\big)\right) \\ &= \theta\sqrt{\frac{\pi}{f_0(x) n \tilde{s}}} \sum_{i=1}^{n} \bar{Y}_i + \theta^2\big(1 + o(1)\big) - \frac{\theta^2}{2}\frac{\pi}{f_0(x)\tilde{s}}\frac{1}{n}\sum_{i=1}^{n} \bar{Y}_i^2 + r_n \\ &= \theta\sqrt{\frac{\pi}{f_0(x) n \tilde{s}}} \sum_{i=1}^{n} \bar{Y}_i + \frac{\theta^2}{2} + \frac{\theta^2}{2}\left(1 - \frac{\pi}{f_0(x)\tilde{s}}\frac{1}{n}\sum_{i=1}^{n} \bar{Y}_i^2\right) + r_n + o(1) \\ &= \theta\bar{\Delta}_n + \frac{\theta^2}{2} + \tilde{r}_n, \end{aligned} \tag{4.55}$$

where, applying (4.17), Lemma 4.4, and Slutsky's theorem,

$$\bar{\Delta}_n = \sqrt{\frac{f_1(x)}{f_0(x)}}\,\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\bar{Y}_i}{\sqrt{\frac{f_1(x)\tilde{s}}{\pi}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

with respect to f_1 and $\tilde{r}_n = o_{P_1}(1)$. Here we use the same arguments as in (4.33) but this time with respect to f_1. Note here that Lindenberg theorem which was applied in Lemma 4.4 can also be applied in this situation and the bound for r_n here remains the same as before.

$$\tilde{\psi}_{\alpha,1} = \tilde{\phi}_\alpha \sqrt{\frac{\tilde{s} f_1(x)}{\pi n}}.$$

Let $\tilde{f}_n \in \mathcal{F}_p^0(x)$ be an arbitrary estimator. If we denote

$$f_n^* = \tilde{\psi}_{\alpha,1}^{-1}(\tilde{f}_n(x) - f_0(x))$$

and

$$L = \tilde{\psi}_{\alpha,1}^{-1}(f_1(x) - f_0(x))$$

then

$$\begin{aligned} \tilde{\psi}_{\alpha,1}^{-1}(\tilde{f}_n(x) - f_1(x)) &= \tilde{\psi}_{\alpha,1}^{-1}(\tilde{f}_n(x) - f_0(x)) - \tilde{\psi}_{\alpha,1}^{-1}(f_1(x) - f_0(x)) \\ &= f_n^* - L, \end{aligned} \tag{4.56}$$

whereas

$$\begin{aligned} \sqrt{\frac{\pi n}{f_0(x)}}\,(\tilde{f}_n(x) - f_0(x)) &= \tilde{s}^{1/2}\tilde{\phi}_\alpha \sqrt{\frac{f_1(x)}{f_0(x)}}\, f_n^* \\ &= \exp\left\{\frac{\log \tilde{s}}{2} + \log \tilde{\phi}_\alpha\right\} \sqrt{\frac{f_1(x)}{f_0(x)}}\, f_n^*. \end{aligned} \tag{4.57}$$

Denote

$$q = \exp(-\tilde{\phi}_\alpha)$$

so that $q \to 0 \quad (n \to \infty)$, according to (4.52). Now

$$\begin{aligned} \tilde{\mathcal{R}} &:= \sup_{f \in \mathcal{A}_\alpha^{(n)}} \mathbf{E}_f \left(\tilde{\phi}_\alpha^{-1} \sqrt{\frac{\pi n}{f(x)\tilde{s}}}\, |\, \tilde{f}_n(x) - f(x)\, | \right)^p \geq \mathbf{E}_1 \left(\tilde{\psi}_{\alpha,1}^{-1}\, |\, \tilde{f}_n(x) - f_1(x)\, | \right)^p \\ &\geq q\, \mathbf{E}_0 \left(\sqrt{\frac{\pi n}{f_0(x)}}\, |\, \tilde{f}_n(x) - f_0(x)\, | \right)^p + (1-q)\, \mathbf{E}_1 \left(\tilde{\psi}_{\alpha,1}^{-1}\, |\, \tilde{f}_n(x) - f_1(x)\, | \right)^p + O(q) \\ &\geq q \exp\left\{\frac{\tilde{\phi}_\alpha^2}{2} + p \log \tilde{\phi}_\alpha\right\} \mathbf{E}_0 \left(\left(\frac{f_1(x)}{f_0(x)}\right)^{p/2} |\, f_n^*(x)\, |^p \right) + (1-q)\ \mathbf{E}_1 |\, f_n^*(x) - L\, |^p + O(q) \end{aligned}$$

due to $\tilde{f}_n \in \mathcal{F}_p(x)$ and equations (4.56) and (4.57). Therefore

$$\begin{aligned} \tilde{\mathcal{R}} &\geq \exp\left\{-\tilde{\phi}_\alpha + \frac{\tilde{\phi}_\alpha^2}{2} + p \log \tilde{\phi}_\alpha\right\} \mathbf{E}_1 \Big(\exp\{-L(\theta)\} \Big(\frac{f_1(x)}{f_0(x)}\Big)^{p/2} |\, f_n^*(x)\, |^p \Big) \\ &\qquad + (1-q)\, \mathbf{E}_1 |\, f_n^*(x) - L\, |^p + O(q) \\ &\geq (1-q)\, \mathbf{E}_1 \Big(Z\, |\, f_n^*(x)|^p + |\, f_n^*(x) - L\, |^p \Big) + O(q) \end{aligned}$$

where $L(\theta)$ is the log-likelihood and

$$Z = \exp\left\{-\tilde{\phi}_\alpha + \frac{\tilde{\phi}_\alpha^2}{2} + p\log\tilde{\phi}_\alpha - L(\theta)\right\}\left(\frac{f_1(x)}{f_0(x)}\right)^{p/2}.$$

From the definition of θ and (4.55) we can see that

$$\exp\left\{-\tilde{\phi}_\alpha + \frac{\tilde{\phi}_\alpha^2}{2} + p\log\tilde{\phi}_\alpha - (\tilde{\phi}_\alpha - \sqrt{\tilde{\phi}_\alpha})\,\bar{\Delta}_n - \frac{1}{2}(\tilde{\phi}_\alpha - \sqrt{\tilde{\phi}_\alpha})^2 - \tilde{r}_n\right\} \xrightarrow{P_1} \infty.$$

As we saw in the previous chapters

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

Thus for any $p > 0$

$$\min_x g(x) = \chi L^p,$$

where $0 < \chi \le 1$, $\chi \xrightarrow{\mathbf{P}_1} 1$. As a last step, note that

$$\begin{aligned} L &= \tilde{\phi}_\alpha^{-1}\sqrt{\frac{\pi n}{\tilde{s}f_1(x)}}(f_1(x) - f_0(x)) \\ &= \tilde{\phi}_\alpha^{-1}\sqrt{\frac{\pi n}{\tilde{s}f_1(x)}}f_0(x)\,\theta\,\sqrt{\frac{\pi}{n\tilde{s}f_0(x)}}\Big(k_{\tilde{s}}(0) - \mathbf{E}_0 k_{\tilde{s}}(x)\Big) \\ &= \tilde{\phi}_\alpha^{-1}(\tilde{\phi}_\alpha - \sqrt{\tilde{\phi}_\alpha})\sqrt{\frac{f_0(x)}{f_1(x)}}\,\frac{\pi}{\tilde{s}}\left(\frac{\tilde{s}}{\pi} - f_0(x)\big(1+o(1)\big)\right) \;=\; 1 + o(1). \end{aligned}$$

Therefore according to the previous arguments, uniformly in $\alpha \in \mathcal{K}_n$,

$$\begin{aligned} \sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}^{(n)}\left|\psi_\alpha^{-1}\big(\,\tilde{f}_n(x) - f(x)\,\big)\right|^p &= \sup_{f\in\mathcal{A}(\alpha)} \mathbf{E}^{(n)}\left|\tilde{\psi}_\alpha^{-1}\big(\,\tilde{f}_n(x) - f(x)\,\big)\right|^p(1+o(1)) \\ &\ge (1-q)L^p\mathbf{E}_1\chi + O(q) \;=\; 1 + o(1), \quad (n\to\infty). \end{aligned}$$

Thus the theorem is proved. □

Corollary 4.1 *Let $\mathcal{A}_{\mathcal{K}_n}$ be an arbitrary RNP-scale. Then for any $p > 0$, $\epsilon > 0$ and $x \in \mathbb{R}$, the estimator $\hat{f}_n$ of Theorem 4.2 is $(p, \epsilon, \mathcal{K}_n, \mathcal{F}_p(x))$-adaptively minimax at x.*

Proof. This is a consequence of Theorems 4.2 and 4.3. □

Appendix

A.1 Fundamental inequalities

Bernstein's inequality, Pollard [1984], p. 193

Let X_1 , ..., X_n be independent identically distributed random variables with $\mathbf{E}X_i = 0$, $\mathbf{E}X_i^2 = \sigma^2$ and bounded ranges: $|X_i| \leq B$. Then for each $c > 0$,

$$\mathbf{P}\left(\frac{1}{\sigma\sqrt{n}}\left|\sum_{i=1}^{n} X_i\right| \geq c\right) \leq 2\,\exp\left\{-\frac{c^2}{2\left(1+\frac{Bc}{3\sigma\sqrt{n}}\right)}\right\} \quad ,$$

or, equivalently,

$$\mathbf{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq c\right) \leq 2\,\exp\left\{-\frac{c^2}{2\left(\sigma^2 n+\frac{Bc}{3}\right)}\right\} \quad .$$

or

$$\mathbf{P}\left(\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n} X_i\right| \geq c\right) \leq 2\,\exp\left\{-\frac{c^2}{2\left(\sigma^2+\frac{Bc}{3\sqrt{n}}\right)}\right\} \quad ,$$

and for $0 < c < 3\delta\sqrt{n}/B, \delta > 0$,

$$\mathbf{P}\left(\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n} X_i\right| \geq c\right) \leq 2\,\exp\left\{-\frac{c^2}{2\left(\sigma^2+\delta\right)}\right\} \quad .$$

Bretagnolle-Huber's inequality, Bretagnolle and Huber [1979]

Let X_1 , ..., X_n be independent identically distributed random variables with $\mathbf{E}X_i = 0$, $\mathbf{E}X_i^2 = \sigma^2$ and bounded ranges: $|X_i| \leq B$. There exist constants C_p, such that

$$\mathbf{E}\Big|\sum_{i=1}^{n} X_i\Big|^p \leq \mathcal{C}_p\, n^{p/2}\sigma^p, \qquad 1 \leq p \leq 2.$$

$$\mathbf{E}\Big|\sum_{i=1}^{n} X_i\Big|^p \leq \mathcal{C}_p\left(n\sigma^2 B^{p-2} + n^{p/2}\sigma^p\right), \qquad p \geq 2$$

A.2 Elementary inequalities

$\mathbf{c_r}$-inequality, Loève [1977], p. 157

For $r > 0$, and

$$|a+b|^r \leq c_r|a|^r + c_r|b|^r,$$

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

λ-inequality

Let $a, b \in \mathbb{R}$. For $p \geq 1$, and $0 < \lambda < 1$

$$|a+b|^p \leq \frac{1}{(1-\lambda)^{p-1}}|a|^p + \frac{1}{\lambda^{p-1}}|b|^p.$$

This inequality follows from the convexity inequality by taking $f(x) = |x|^p$ and setting $x = a/(1-\lambda)$, $y = b/\lambda$.

Bibliography

Antonsik, P., Mikusiński, J. & Sikorski, R. [1973] *Theory of Distribution. The sequential approach* (Elsevier, Amsterdam).

Bateman, H. [1953] *Higher Transcendental Functions*, volume II (McGraw-Hill).

Belitser, E.N. *Minimax Estimation in Regression and Random Censorship Models* PhD thesis, Utrecht University, Department of Mathematics, [1997].

Bickel, P.J. & Doksum, K.A. [1977] *Mathematical Statistics, Basic ideas and selected topics* (Holden-Day Inc.).

Birgé, L. & Massart, P. [1997] *From model selection to adaptive estimation*, pp 55–87 Springer.

Bretagnolle, J. & Huber, C. [1979] "Estimation des densités: risque minimax," *Z. Wahrsch. Verw. Gebiete* **47**, 119–137.

Brown, L.D. & Low, M.G. [1996] "Asymptotic equivalence of nonparametric regression and white noise," *Ann. Statist.* **24(6)**, 2384–2398.

Butucea, C. *Estimation Non Paramétrique Adaptative de la Densité de Probabilité; Vitesses de Convergence, Constante Exact et Résultats Numeriques* PhD thesis, These de Doctorat de la Université Paris VI, [1999].

Cavalier, L. [2001] "On the problem of adaptive estimation in tomography," *Bernoulli* **7 (1)**, 63–78.

Chebyshev, P.L. [1859] "Problems on the least values connected with the approximate representation of functions. (voprosy naimen'shikh velichinakh, svyazannye s priblizhennym predstavleniyem funktsii)," *Sochineniya* **II**, 151–235.

Compte, F. [2001] "Adaptive estimation of the spectrum of a stationary gaussian sequence," *Bernoulli* **7(2)**, 267–298.

Dohono, D.L. & Johnstone, I. M. [1994] "Ideal spatial adaptation via wavelet shrinkage," *Biometrica* **81**, 425–455.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. & Picard, D. [1996] "Density estimation by wavelet thresholding," *Ann. Statist.* **24**, 508–539.

Efromovich, S.Y. [1985] "Nonparametric estimation of a density with unknown smoothness," *Theory Probab. Appl.* **30**, 557–568.

Efromovich, S.Y. & Pinsker, M.S. [1984] "An adaptive algorithm of nonparametric filtering," *Automat. Remote. Control* **11**, 1434–1440.

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

Fox, L. & Parker, I.B. [1968] *Chebyshev Polynomials in Numerical Analysis* (Oxford University Press, London).

Goldenshluger, A. & Nemirovski, A. [1997] "On spatially adaptive estimation of nonparametric regression," *Math. Meth. Statist.* **6**, 135–170.

Golubev, G.K., Lepski, O.V. & Levit, B.Y. [2001] "On adaptive estimation using the *sup*-norm losses," *Math. Meth. Statist.* **10**, 23–37.

Golubev, G.K. & Levit, B.Y. [1996] "Asymptotically efficient estimation for analytic distributions," *Math. Meth. Statist.* **5**, 357–368.

Golubev, G.K., Levit, B.Y. & Tsybakov, A.B. [1996] "Asymptotically efficient estimation of analytic functions in gaussian noise," *Bernoulli* **2**, 167–181.

Golubev, G.K. & Nussbaum, M. [1992] "Adaptive spline estimates for nonparametric regression models," *Theory Probab. Appl.* **37**, 521–529.

Gradshtein, I.S. & Ryzhik, I.M. [1965] *Table of Integrals, Series, and Products* (Academic Press, New York).

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

Ibragimov, I.A. "Estimation of analytic functions," In *Conf. State of the art in probability and statistics (Leiden, 1999)*, number 36 in IMS Lecture Notes Monogr., pp 359–383, Inst. Math. Statist., Beachwood, OH, [2001].

Ibragimov, I.A. & Has'minskii, R.I. [1981] *Statistical Estimation, Asymptotic Theory* (Springer, New York).

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

Ibragimov, I.A. & Has'minskii, R.I. [1983] "Estimation of distribution density," *Journ. Sov. Math.* , 40–57.

Ibragimov, I.A. & Has'minskii, R.I. [1984] "On nonparametric estimation of the value of a functional in gaussian white noise," *Theor. Probab. Appl.* **29(1)**, 18–32.

Lepski, O.V. [1990] "On a problem of adaptive estimation in gaussian noise," *Theory Probab. Appl.* **35**, 454–466.

Lepski, O.V. [1991] "Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates," *Theory Probab. Appl.* **36**, 682–697.

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

Lepski, O.V. [1992]b "On problems of adaptive estimation in white gaussian noise," *Adv. Soc. Math.* **12**, 87–106.

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

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

Lepski, O.V., Mammen, E. & Spokoiny, V.G. [1997] "Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection," *Ann. Statist.* **25(3)**, 929–947.

Lepski, O.V. & Spokoiny, V.G. [1997] "Optimal pointwise adaptive methods in nonparametric estimation," *Ann. Statist.* **25(6)**, 2512–2546.

Levit, B.Y. "On the optimality of some statistical estimates," In Hájek, J, editor, *Proc.,Prague Symp. Asympt.,Statist.*, volume 2, pp 215–238, [1974].

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

Loève, M. [1977] *Probability Theory* (Springer, New York), 4th edition.

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

Nussbaum, M. [1996] "Asymptotic equivalence of density estimation and gaussian white noise," *Ann. Statist.* **24(6)**, 2399–2430.

Pollard, D. [1984] *Convergence of Stochastic Processes* (Springer-Verlag, New York, Heidelberg, Berlin).

Schipper, M. [1996] "Optimal rate and constants in l_2-minimax estimation of probability density functions," *Math. Methods Statist.* **5(3)**, 253–274.

Stoer, J. & Bulirsch, R. [1992] *Introduction to Numerical Analysis* (Springer-Verlag), 2nd edition.

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

Szegö, G. [1975] *Orthogonal Polynomials*, volume XXIII (American Math. Soc.), 4th edition.

Timan, A.F. [1963] *Theory of Approximation of Functions of a Real Variable* (Pergamon Press).

Tsybakov, A.B. [1998] "Pointwise and sup-norm adaptive signal estimation on the Sobolev classes," *Ann. of Statist.* **26**, 2420–2469.

Yuditsky, A. [1997] "Wavelet estimators: Adapting to unknown smoothness," *Math. Meth. Statist.* **6**, 1–25.

Samenvatting

De afgelopen twintig jaar heeft de adaptieve schatting zich ontwikkeld tot één van de meest actieve onderzoeksgebieden in de niet-parametrische statistiek. De vraag uit de praktijk naar realistischere modellen en flexibere schattingsmethoden leidde tot de introductie van verschillende modellen. Bij de bestudering van deze modellen ontstonden uitdagende vraagstukken, die om nieuwe statistische modellen en benaderingen vroegen.

Statistische schatting begint met de aanname dat we een steekproef hebben volgens een onbekende kansmaat $\mathbf{P}$ op een gegeven maatruimte. De kansmaat $\mathbf{P}$ behoort tot een zekere klasse $\mathcal{P}$. Het doel van de statisticus is nu een methode te vinden om de onbekende kansmaat $\mathbf{P}$ te schatten, daarbij gebruikmakend van de gegeven steekproef. Gewoonlijk modelleren we de klasse kansverdelingen in de geparametriseerde vorm $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$ en de taak van de statisticus komt dan neer op het schatten van θ of een functie van θ. Voor het soort niet-parametrische problemen dat we in dit proefschrift behandelen, is de parameter-verzameling Θ meestal oneindig-dimensionaal.

Hoe 'goed' een schattingsmethode is, hangt meestal van de klasse Θ af. In het klassieke geval wordt aangenomen dat deze klasse bekend is. In de praktijk is dit echter zelden het geval, wat realistischere methoden noodzakelijk maakt. Met *adaptieve schattingsmethoden* bedoelen we schattingsmethoden die gebaseerd zijn op data en zich in zekere zin aanpassen aan de onzekerheid over de echte klasse Θ. Een voorbeeld is een methode die een schatting $\hat{\theta}$ uit een rijtje kandidaten $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$ kiest, die de optimale schattingen zijn behorend bij de verschillende klassen $\Theta_1, \Theta_2, \ldots, \Theta_n$.

In dit proefschrift bestuderen we adaptieve schattingsmethoden voor twee speciale soorten statistische problemen: regressie en dichtheidsschatting. Voor deze problemen wordt de klasse van kansmaten $\mathbf{P}$ geparametriseerd door reëelwaardige functies θ. In ieder model nemen we aan dat de onderliggende functie θ tot een bepaalde klasse Θ van gladde functies behoort. In de praktijk is de 'echte' gladheid van de functie θ onbekend, evenals de klasse Θ. Dus bij dit soort problemen is het ontwikkelen van adaptieve schattingsmethoden erg belangrijk.

We bestuderen verschillende regressie-problemen met een vaste discrete opzet, namelijk regressie op de reële rechte en regressie op een begrensd interval. Hoewel het onderscheid tussen deze problemen slechts ligt in de definitie van de onderliggende functieklassen Θ, is er een groot verschil in de constructie van optimale adaptieve procedures in deze twee gevallen. Dit wordt veroorzaakt door het essentiële verschil tussen deze twee modellen: in het geval van regressie-modellen op begrensde observatie-intervallen moet namelijk de aanwezigheid

van de rand – het zogenaamde rand-effect – meegenomen worden in de bestudering van optimale statistische procedures.

Voor elk van de drie problemen (regressie op de reële rechte, regressie op begrensde intervallen en dichtheidsschatting) introduceren we overeenkomstige "schalen" van klassen Θ, van analytische functies. Voor deze Θ's kunnen we de convergentiesnelheid, op een constante na, bepalen in het klassieke niet-parametrische minimax raamwerk, waarin de klassen Θ bekend zijn. We vervolgen dan met het construeren van adaptieve schattingen en bewijzen dat deze asymptotisch optimaal zijn.

Hoe goed men een schatting kan geven op de rand van een begrensd observatie-interval, hangt af van de gekozen opzet. We analyseren twee klassieke methoden op een begrensd interval: de equidistante en de Chebyshev opzet. In beide gevallen is de kwaliteit van de schatting dicht bij de rand slechter dan in het interval. Terwijl dit in het geval van de equidistante opzet de convergentiesnelheid beïnvloedt, blijft bij de Chebyshev opzet het verlies van convergentiesnelheid aan de rand beperkt. Bij het bestuderen van adaptieve schattingsprocedures beperken we ons daarom tot de Chebyshev opzet.

Gracias

I would like to express my gratitude to the University of Utrecht for giving me the opportunity to carry out my PhD research in a very pleasant environment, as well as for giving me the financial support necessary to participate in several conferences and other events.

I am profoundly indebted to my supervisor Boris Levit for his always supportive and patient guidance of my research, for all his advice and for the stimulating and very enjoyable conversations. I also want to thank Richard Gill for being my promotor and always being kind and helpful.

I want to thank the Institute Henri Poincaré (IHP) – Paris, France – for hosting me while attending part of the semester "2001, l'Odysseé de la Statistique". I would like to thank the organizers, for providing me with a grant to attend, and in particular to Alexandre Tsybakov, for his invitation to this event.

I want to thank the Statistical Department at Queen's University, Kingston, Canada, for hosting me during a three month visit to my supervisor Boris Levit.

I want to thank Lorna Booth for reading the manuscript and suggesting possible improvements. I also want to thank Damien White, Silvia Caserta and Corrie Quant in this regard.

Not all I have done in The Netherlands is mathematics. I have also been lucky and have good friends and colleagues, with whom I spent a great time.

I want to thank the family van Heeswijk, including David, and specially Tineke. They were the first to introduce me in the Dutch culture with a gezellige Christmas dinner and later there was more – nice talks, skiing holidays, a chance to try diving, etc.

I had a great time in the Mathematical department of Utrecht University. I enjoyed very much the time with the people of the "7de verdieping" – Menno Verbeek, Lennaert van Veen, Martijn van Manen, Theo Tuwankotta, Barbara van den Berg, Mischja van Bossum and Bob Rink. To Menno and Martijn thanks for their help and contribution to the computer freaky side of the moon.

To other friends who were always close to me, in a way or another – Marisela Mainegra, Alejandro León, Ernesto Reinaldo, Stephanie Godey, Eric Porras, Adán Simón, and many others – thanks.

My parents and my brother have been fara way from here, but I have always felt as if they were very close. For their love, thanks.

Lastly and very specially I want to thank Alina for her lovely smile and company during all these years.

Curriculum Vitae

Luis M. Artiles Martínez was born in Santa Clara, Cuba, on 8 June 1970. During his secondary and high-school studies he took part in National Mathematics Olympiads where he obtained some prizes, including a first and a second position. He also took part in the 28th International Mathematics Olympiad as a member of the Cuban team.

Between September 1988 and July 1993 he followed studies of Mathematics at Havana University where he graduated with honors.

From September 1993 he worked for three years as an Instructor at the Faculty of Physics, Mathematics and Computer Science at the Central University of Las Villas, Santa Clara, Cuba.

During the academic year 1996/97 he attended the Master Class program "Stochastics and Operational Research" organized by the Mathematical Research Institute (MRI) in the Netherlands.

After the Master Class he received a fellowship from Utrecht University to conduct the PhD research that resulted in the present thesis.