Density deconvolution with non–standard error distributions: Rates of convergence and adaptive estimation[*]

Alexander Goldenshluger and Taeho Kim

Department of Statistics, University of Haifa,
Haifa 3498838, Israel
e-mail: goldensh@stat.haifa.ac.il; ktaeho@campus.haifa.ac.il

Abstract: It is a standard assumption in the density deconvolution problem that the characteristic function of the measurement error distribution is non-zero on the real line. While this condition is assumed in the majority of existing works on the topic, there are many problem instances of interest where it is violated. In this paper we focus on non–standard settings where the characteristic function of the measurement errors has zeros, and study how zeros multiplicity affects the estimation accuracy. For a prototypical problem of this type we demonstrate that the best achievable estimation accuracy is determined by the multiplicity of zeros, the rate of decay of the error characteristic function, as well as by the smoothness and the tail behavior of the estimated density. We derive lower bounds on the minimax risk and develop optimal in the minimax sense estimators. In addition, we consider the problem of adaptive estimation and propose a data–driven estimator that automatically adapts to unknown smoothness and tail behavior of the density to be estimated.

Contents

---

1. Introduction

Statistical inverse problems are ubiquitous in science and engineering; they arise in such diverse areas as signal and image processing, physics, biology, operations research, to name but a few. In such problems transformation $y = \mathcal{A}f$ of an object of interest f (probability distribution/density, signal, image etc.) is observed with noise, and the goal is to reconstruct f from the noisy indirect observations. For discussion of various types of statistical inverse problems we refer to the review paper Cavalier [7] where further references to relevant literature can be found.

Deconvolution is an important class of linear inverse problems in which transformation $\mathcal{A}$ is a convolution integral operator with given kernel A, i.e. $y(x) = (\mathcal{A}f)(x) = \int A(x - t)f(t)\mathrm{d}t$. Nonparametric deconvolution problems naturally arise in models with measurement errors such as error–in–variables nonparametric regression (Fan and Truong [11]), density estimation from noisy samples (Carroll and Hall [6], Zhang [29], Fan [10]), and in signal and image deblurring (Johnstone and Raimondo [19], Johnstone et al. [20], Hall and Koch [17] and Bertero and Boccacci [3, Chapter 3]). In all these problems the Fourier transform techniques play a prominent role. Since in spectral domain the operation of convolution is reduced to multiplication, the standard technique for reconstruction of f is to estimate the Fourier transform of y from the available data, to divide it by the Fourier transform of the convolution kernel A and to apply to the ratio the inverse Fourier transform with some regularization. Such estimation strategies are applicable only if the Fourier transform of the convolution kernel A does not vanish, and nearly all publications in the area require this assumption. However, this assumption is rather restrictive: it does not hold in many cases of interest, e.g., when kernel A is compactly supported and/or discontinuous. In such settings other estimation strategies should be employed, and in this paper our goal is to develop such a strategy in the context of the density deconvolution problem.

Problem Formulation Density deconvolution is a problem of estimating a probability density from observations with additive measurement errors. Specifically, assume that we observe random sample $Y_1, Y_2, \ldots, Y_n$ generated by the model

$$Y_i = X_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where X_is are i.i.d. random variables with unknown density f with respect to the Lebesgue measure on $\mathbb{R}$, ϵ_i's are i.i.d. measurement errors with distribution function G, and X_is are independent of ϵ_i's. The objective is to estimate f on the basis of the sample $\mathcal{Y}_n := \{Y_1, Y_2, \ldots, Y_n\}$. Since Y_i is the sum of two independent random variables, X_i and ϵ_i, the density f_Y of Y_i is given by the convolution

$$f_Y(y) = (f \star \mathrm{d}G)(y) = \int_{-\infty}^{\infty} f(y - x)\mathrm{d}G(x). \tag{1.1}$$

An estimator of the value of $f(x_0)$ is a measurable function of $\mathcal{Y}_n$, $\hat{f}(x_0) = \hat{f}(x_0; \mathcal{Y}_n)$, and the risk of $\hat{f}(x_0)$ is

$$\mathcal{R}_n[\hat{f}, f] := \left[\mathrm{E}_f|\hat{f}(x_0) - f(x_0)|^2\right]^{1/2},$$

where E_f stands for the expectation with respect to the probability measure P_f generated by the observation $\mathcal{Y}_n$ when the unknown density of X_is is f. For a particular functional class $\mathscr{F}$, accuracy of $\hat{f}(x_0)$ is measured by the *maximal risk*

$$\mathcal{R}_n[\hat{f}; \mathscr{F}] := \sup_{f \in \mathscr{F}} \mathcal{R}_n[\hat{f}, f],$$

and an estimator $\hat{f}_*(x_0)$ is called *rate–optimal* or *optimal in order* on $\mathscr{F}$ if

$$\mathcal{R}_n[\hat{f}_*; \mathscr{F}] \asymp \mathcal{R}_n^*[\mathscr{F}] := \inf_{\hat{f}} \mathcal{R}_n[\hat{f}; \mathscr{F}], \quad n \to \infty.$$

Here $\mathcal{R}_n^*[\mathscr{F}]$ is the *minimax risk*, and the infimum in its definition is taken over all possible estimators of $f(x_0)$. The objective in the density deconvolution problem is to construct an optimal in order estimator, and to study the rate at which the minimax risk $\mathcal{R}_n^*[\mathscr{F}]$ converges to zero as $n \to \infty$. In what follows we refer to the latter as *the minimax rate of convergence*.

The outlined problem is a subject of vast literature under various assumptions on the functional class $\mathscr{F}$ and distribution of measurement errors G; see, e.g., Carroll and Hall [6], Stefanski and Carroll [27], Zhang [29], Fan [10], Butucea and Tsybakov [4, 5], Meister [25], Lounici and Nickl [23] for representative publications, where further references can be found. Typically $\mathscr{F}$ is a class of functions satisfying smoothness conditions (e.g., Hölder or Sobolev functional class). As for assumptions on the measurement error distribution, they are usually put in terms of the characteristic function of G and read as follows.

Assumption (E0). Let $\phi_g(i\omega) := \mathcal{F}[\mathrm{d}G, \omega] := \int_{-\infty}^{\infty} e^{-i\omega x} \mathrm{d}G(x)$ be the characteristic function (the Fourier transform) of the measurement error distribution G. Then,

   I. $|\phi_g(i\omega)| \neq 0$ for all $\omega \in \mathbb{R}$.
  II. $|\phi_g(i\omega)|$ decreases at polynomial or exponential rate as $|\omega| \to \infty$:
      ordinary smooth errors: $|\phi_g(i\omega)| \asymp |\omega|^{-\gamma}$, $|\omega| \to \infty$ for some $\gamma > 0$, or
      super smooth errors: $|\phi_g(i\omega)| \asymp \exp\{-c|\omega|^{\gamma}\}$, $|\omega| \to \infty$ for some $c, \gamma > 0$.

Assumption (E0) is inarguably conventional and presumed in nearly all works dealing with density deconvolution problems. Under Assumption (E0) the accuracy in estimating f is determined by the rate at which ϕ_g tends to zero and by the smoothness of f as characterized in terms of functional class $\mathscr{F}$. In particular, it is well known that in the case of *the ordinary smooth errors* one has $\mathcal{R}_n^*[\mathscr{H}_\alpha(A)] \asymp (A^{1/\alpha} n^{-1})^{\alpha/(2\alpha + 2\gamma + 1)}$ as $n \to \infty$, where $\mathscr{H}_\alpha(A)$ denotes the Hölder functional class of regularity $\alpha > 0$ (the precise definition of $\mathscr{H}_\alpha(A)$ is given in Section 3.1). The parameter γ characterizes the rate of decay of $|\phi_g(i\omega)|$ as $|\omega| \to \infty$, and it is usually referred to as the degree of ill–posedness of the deconvolution problem. The indicated minimax rate of convergence is achieved by the standard kernel density deconvolution estimator of the form

$$\hat{f}(x_0) = \frac{1}{2\pi} \sum_{j=1}^{n} \int_{-\infty}^{\infty} \frac{\phi_K(i\omega h)}{\phi_g(-i\omega)} e^{i\omega(Y_j - x_0)} \mathrm{d}\omega, \qquad (1.2)$$

where K is a kernel satisfying standard properties, $\phi_K(i\cdot)$ is the Fourier transform of K, and $h > 0$ is a bandwidth that should be specified properly; see, e.g., [10].

Condition (E0-I) ensures that the statistical model is identifiable (it is well known that if ϕ_g vanishes on a set that contains a non–empty open interval then f is not identifiable; see, e.g., Meister [25, Section 2.3]). It underlies applicability of standard Fourier–transform–based techniques for constructing estimators of f such as (1.2). Note however that (E0-I) does not hold if ϕ_g has isolated zeros which is the case in many interesting situations, e.g., for continuous distributions with compactly supported densities or for general discrete distributions. For example, if G is a uniform distribution on $[-1, 1]$, then $\phi_g(i\omega) = \sin\omega/\omega$ has zeros at $\omega = \pm\pi k$, $k \in \mathbb{N}$, and (E0-I) is not fulfilled. In this paper we focus on the settings where ϕ_g has isolated zeros on the imaginary axis so that standard Fourier–transform–based kernel density deconvolution estimator (1.2) is not directly applicable.

Related Literature The settings in which the error characteristic function ϕ_g may have isolated zeros have been studied to a considerably lesser extent; the available results in this area are fragmentary and disparate. Devroye [9] pointed out that density f can be estimated consistently in the $\mathbb{L}_1$–norm when the characteristic function ϕ_g of the error distribution is non-zero almost everywhere. Although the result is quite general, the convergence is not uniform, and the evaluation procedure is not based on the minimax criterion. Several

previous studies investigated the problem with the uniform error distribution. In particular, Groeneboom and Jongbloed [16] and Feuerverger, Kim and Sun [12] demonstrate that zeros of the characteristic function ϕ_g do not have influence on the minimax rate of convergence: it remains the same as under condition (E0-I) when the estimated density f is supported on the positive real line [16], or has bounded second moment [12]. Considering a more general class of so-called Fourier–oscillating error distributions, Delaigle and Meister [8] derives a similar result for densities f having finite left endpoint. In contrast to the aforementioned results, Hall and Meister [18] demonstrates that for the class of Fourier–oscillating error distributions zeros of the error characteristic function lead to a slower minimax convergence rate than the one under condition (E0-I). Hall and Meister [18] suggested a "ridge" modification of the kernel density deconvolution estimator in which characteristic function of the error distribution is regularized to avoid singularities due to the zeros. For another closely related work we also refer to Meister [24].

Recently a principled method for solving density deconvolution problems under general assumptions on the error characteristic function has been proposed in Belomestny and Goldenshluger [2]. This method uses the Laplace transform (the Fourier transform in complex domain) in conjunction with the linear functional strategy for constructing rate–optimal kernel deconvolution estimators. The results show that zeros of the error characteristic function have no influence on the achievable estimation accuracy when, in addition to usual smoothness conditions, the estimated density f has sufficiently light tails. On the other hand, if f is heavy tailed, then zeros of the error characteristic function *affect* the minimax rates of convergence that become slower. Belomestny and Goldenshluger [2] provide an explicit condition on the tail behavior of f and zeros geometry of ϕ_g under which the minimax rates of convergence are not influenced by the zeros of ϕ_g.

Main Contributions. In this paper we focus on the setting when ϕ_g has zeros, and f is heavy tailed relative to the multiplicity m of zeros of ϕ_g on the imaginary axis. The prototypical settings of this type arise when measurement error distribution is the binomial distribution with the parameters m and $p = 1/2$ or the m–fold convolution of uniform distributions on $[-\theta, \theta]$. Utilizing the methodology proposed in [2] we develop rate–optimal estimators of f and investigate their properties. It is shown that, in contrast to the well known results under Assumption (E0), in the considered regime the minimax rate of convergence is determined not only by the smoothness of f and the rate at which ϕ_g tends to zero, but also by the tail behavior of f and the zero multiplicity of ϕ_g. The derived lower bounds on the minimax risk demonstrate that dependence of the estimation accuracy on these factors is essential.

The construction of the proposed rate–optimal estimator of f depends on tuning parameters, and their specification requires prior information on smoothness and tail behavior of f. In practice such information is rarely available. To overcome this difficulty we propose and study an adaptive estimator of f that is based on the methodology developed in Goldenshluger and Lepski [14, 15].

An interesting feature of the proposed estimator is that it involves two tuning parameters, and the adaptation here is not only with respect to the unknown smoothness, but also with respect to the unknown tail behavior of f. We derive an oracle inequality for the developed adaptive estimator and show that it achieves the minimax rate of convergence up to a logarithmic factor which is unavoidable payment for adaptation in point-wise estimation.

Organization of the Paper The rest of the paper is organized as follows. In Section 2 we present the general idea for estimator construction and introduce our estimator. Section 3 deals with the minimax estimation of $f(x_0)$ with respect to proper functional classes. In Section 4 we introduce the corresponding adaptive procedure and investigate its properties. Lastly, Section 5 is reserved for discussion and concluding remarks. All the proofs are deferred to Appendix.

2. Estimator construction

2.1. Idea of construction

We start with presenting the key idea for estimator construction in our density deconvolution problem. The construction uses Laplace transform (Fourier transform in the complex domain) which allows us to handle the situation where the first condition of Assumption (E0) is not satisfied. Our goal is to deliver the main idea of construction; for further details we refer to Belomestny and Goldenshluger [2].

The following definitions will be utilized throughout the study. For a generic function w the bilateral Laplace transform of w is defined to be

$$\mathcal{L}[w; z] := \phi_w(z) = \int_{-\infty}^{\infty} w(x)e^{-zx}\mathrm{d}x. \tag{2.1}$$

The integral convergence region Σ_w (if exists) is a vertical strip in the complex plane, $\Sigma_w = \{z \in \mathbb{C} : Re(z) \in (\sigma_w^-, \sigma_w^+)\}$ for some $\sigma_w^-, \sigma_w^+ \in \mathbb{R}$, and $\phi_w(z)$ is analytic in Σ_w. The inverse Laplace transform is given by the formula

$$w(x) = \frac{1}{2\pi i} \int_{s-i\infty}^{s+i\infty} \phi_w(z)e^{zx}\mathrm{d}z = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_w(s+i\omega)e^{(s+i\omega)x}\mathrm{d}\omega, \quad s \in (\sigma_w^-, \sigma_w^+);$$

see Widder [28] for background on the properties of Laplace transform. For the error distribution function G we write $\phi_g(z) := \int_{-\infty}^{\infty} e^{-zx}\mathrm{d}G(x)$, and note that the integral convergence region necessarily includes the imaginary axis $\{z \in \mathbb{C} : Re(z) = 0\}$ with $\phi_g(i\omega)$ being the characteristic function of G. In what follows we assume that Σ_g is a vertical strip in the complex plane, $\Sigma_g := \{z \in \mathbb{C} : Re(z) \in (\sigma_g^-, \sigma_g^+)\}$ for some $\sigma_g^- < 0 < \sigma_g^+$.

Our estimator utilizes a kernel whose construction relies upon *the linear functional strategy* for the solution of ill–posed problems (see, e.g., [13]). Let

$K \in C^\infty(\mathbb{R})$ be a kernel on $[-1, 1]$ satisfying standard conditions such that for fixed $k \in \mathbb{Z}_+$

$$\int_{-1}^{1} K(t)dt = 1, \quad \int_{-1}^{1} t^j K(t)dt = 0, \ \ \forall j = 1, \ldots, k. \tag{2.2}$$

Note that $\phi_K(z)$ is an entire function, i.e. $\Sigma_K = \mathbb{C}$. We would like to find a function $L : \mathbb{R} \to \mathbb{R}$ such that for any given $x_0 \in \mathbb{R}$

$$\int_{-\infty}^{\infty} L(y - x_0)f_Y(y)\mathrm{d}y = \frac{1}{h} \int_{-\infty}^{\infty} K\Big(\frac{x - x_0}{h}\Big) f(x)\mathrm{d}x, \tag{2.3}$$

where we recall that f_Y and f are related to each other by the convolution integral (1.1). If function L satisfying (2.3) is found, then a reasonable estimator of $f(x_0)$ is given by the empirical estimator of the integral on the left hand side of (2.3) based on the sample $\mathcal{Y}_n$. In our deconvolution problem this strategy is realized as follows.

In addition to the analyticity of ϕ_g in Σ_g we suppose that $\phi_g(z)$ does not vanish on the set $\{z : Re(z) \in (\varkappa_g^-, \varkappa_g^+) \setminus \{0\}\}$ for some $\varkappa_g^-, \varkappa_g^+$ such that $\sigma_g^- \leq \varkappa_g^- < 0 < \varkappa_g^+ \leq \sigma_g^+$. Note that ϕ_g may have zeros on the imaginary axis $\{z : Re(z) = 0\}$, so that the conventional Fourier transform technique would not work in this situation. Let $S_g := \{z : Re(z) \in (-\varkappa_g^+, -\varkappa_g^-) \setminus \{0\}\}$; in fact, S_g is the union of two open vertical strips in the complex plane having the imaginary axis as the boundary. Note that $\phi_g(-z) \neq 0$ on S_g, and for $h > 0$ define

$$\phi_L(z) := \frac{\phi_K(zh)}{\phi_g(-z)}, \quad z \in S_g.$$

Obviously, ϕ_L is analytic on S_g, and we define a kernel L_h^s through the following inverse Laplace transform of ϕ_L:

$$L_h^s(x) := \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\phi_K((s + i\omega)h)}{\phi_g(-s - i\omega)} e^{(s+i\omega)x}d\omega, \quad s \in (-\varkappa_g^+, -\varkappa_g^-) \setminus \{0\}. \tag{2.4}$$

Depending on the sign of parameter s formula (2.4) defines two different kernels which in the sequel are denoted $L_h^+(\cdot)$ for $s > 0$ and $L_h^-(\cdot)$ for $s < 0$. If the integral on the right hand side of (2.4) is absolutely convergent and

$$\int_{-\infty}^{\infty} |L_h^s(y - x_0)| f_Y(y)dy < \infty,$$

then by Lemma 1 in [2] kernels L_h^s and K are related to each other via (2.3). Therefore we define the resulting density deconvolution estimator by

$$\hat{f}_h^s(x_0) = \frac{1}{n} \sum_{j=1}^{n} L_h^s(Y_j - x_0), \quad s \in (-\varkappa_g^+, -\varkappa_g^-) \setminus \{0\}.$$

While a general form of the kernel L_h^s is given in (2.4), it would be beneficial to specialize it for particular error distributions. We handle this in the next subsection in relation to the error characteristic functions ϕ_g having zeros on the imaginary axis.

2.2. Measurement error distributions

The following assumption on characteristic function of measurement errors has been introduced in [2].

Assumption (E1). ϕ_g is analytic in $\Sigma_g := \{z : Re(z) \in (\sigma_g^-, \sigma_g^+)\}$ with $\sigma_g^- < 0 < \sigma_g^+$ and admits the following representation

$$\phi_g(z) = \frac{1}{\psi(z)} \prod_{k=1}^{q} \left(1 - e^{a_k z - i b_k}\right)^{m_k}, \tag{2.5}$$

where $\{a_k\}_{k=1}^q$ and $\{b_k\}_{k=1}^q$ are real numbers, $a_k > 0$ and $b_k \in [0, 2\pi)$, such that their pairs (a_k, b_k)'s are distinct throughout $k = 1, 2, \ldots, q$, and $\{m_k\}_{k=1}^q$ are non-negative integer numbers. The function $\psi(z)$ is represented as

$$\psi(z) = \psi_0(z) \prod_{k:b_k=0} (-a_k z)^{m_k} \prod_{k:b_k \neq 0} (1 - e^{-ib_k})^{m_k},$$

where $\psi_0(z)$ is an analytic function, it has no zeros in a vertical strip Σ_ψ, $\{z : Re(z) = 0\} \subset \Sigma_\psi \subseteq \Sigma_g$, and $\psi_0(0) = 1$. This particular form of $\psi(z)$ follows from (2.5) and the fact that ϕ_g is a characteristic function.

Assumption (E1) postulates that the characteristic function $\phi_g(z)$ is analytic in a vertical strip and can be factorized in a product of two functions: the first function $1/\psi(z)$ does not vanish in a strip around the imaginary axis while the second function $\prod_{k=1}^q (1 - e^{a_k z - i b_k})^{m_k}$ has zeros on the imaginary axis. Therefore, the condition (2.5) ensures that the zeros of $\phi_g(z)$ are $z_{k,j} = i(b_k + 2\pi j)/a_k$, $j = 0, \pm1, \pm2, \ldots$, $z_{k,j} \neq 0$, and the multiplicity of $z_{k,j}$ is equal to m_k for any j. Note that Assumption (E1) is rather general. It holds for a wide class of discrete and continuous distributions; for specific examples we refer to [2, Section 3.2]. Since the main focus of this study is to investigate the effect of zeros multiplicity of $\phi_g(z)$ on the estimation accuracy, we will concentrate on the following prototypical examples:

(a) [m–convolution of $U(-\theta, \theta)$ distribution] Let G be the distribution function of m–fold convolution of the uniform distribution on $[-\theta, \theta]$, $\theta > 0$. Then,

$$\phi_g(z) = \left[\frac{\sinh(\theta z)}{\theta z}\right]^m = e^{-m\theta z}(-2\theta z)^{-m}(1 - e^{2\theta z})^m, \tag{2.6}$$

so that Assumption (E1) holds with $q = 1$, $a_1 = 2\theta$, $b_1 = 0$, $m_1 = m$ and $\psi(z) = (-2\theta z)^m e^{m\theta z}$.

(b) [binomial distribution with $(m, 1/2)$] Let G be the distribution function of the binomial random variable with parameters m and $p = 1/2$. Then,

$$\phi_g(z) = 2^{-m}(1 + e^z)^m, \tag{2.7}$$

so that Assumption (E1) holds with $q = 1$, $a_1 = 1$, $b_1 = \pi$ and $\psi(z) = 2^m$.

It is worth noting that in setting (a) parameter m characterizes both the multiplicity of zeros and the rate of decay of $|\phi_g(i\omega)|$ as $|\omega| \to \infty$. Thus, parameter m also plays the role of the degree of ill–posedness, similar to the parameter γ of Assumption (E0) in the case of the ordinary smooth errors.

2.3. Estimator and zero multiplicity

Under Assumption (E1) the kernel in (2.4) takes the following particular form:

$$L_h^s(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\phi_K((s+i\omega)h)\psi(-s-i\omega)}{\prod_{k=1}^{q}(1 - e^{-a_k(s+i\omega)-ib_k})^{m_k}} e^{(s+i\omega)t} d\omega, \quad s+i\omega \in S_g. \quad (2.8)$$

While the denominator does not vanish for $s \in (-\varkappa_g^+, -\varkappa_g^-) \setminus \{0\}$, the kernel representation is either L_h^+ or L_h^-, depending on the sign of s. For examples (a) and (b) discussed above we can substitute expressions of $\phi_g(z)$ given by (2.6) and (2.7) for (2.4). Then expanding formally the integrand in series (see, [2, Section 4.1] for details), we can obtain the following infinite series representation for the kernels:

(a) m–convolution of $U(-\theta, \theta)$ distribution:

$$L_h^{\pm}(t) = \frac{(\pm 2\theta)^m}{h^{m+1}} \sum_{j=0}^{\infty} C_{j,m} K^{(m)} \left(\frac{t \mp \theta(2j+m)}{h} \right);$$

(b) binomial distribution with $(m, 1/2)$:

$$L_h^{\pm}(t) = \frac{(\pm 2)^m}{h} \sum_{j=0}^{\infty} C_{j,m} K \left(\frac{t \mp j}{h} \right),$$

where

$$C_{j,m} := \binom{j+m-1}{m-1}$$

is the number of weak compositions of j into m parts (see, e.g., [26]). Note that the derived kernels $L_h^{\pm}$ are not integrable and, in general, condition (2.3) is fulfilled only if f has sufficiently light tails. Thus we truncate the infinite series in the estimator construction by a cut–off parameter N, so that the resulting kernels for examples (a) and (b) are of the following forms, respectively:

$$L_{h,N}^{\pm}(t) := \frac{(\pm 2\theta)^m}{h^{m+1}} \sum_{j=0}^{N} C_{j,m} K^{(m)} \left(\frac{t \mp \theta(2j+m)}{h} \right); \quad (2.9)$$

$$L_{h,N}^{\pm}(t) := \frac{(\pm 2)^m}{h} \sum_{j=0}^{N} C_{j,m} K \left(\frac{t \mp j}{h} \right). \quad (2.10)$$

The multiplicity of zeros clearly manifests itself in construction of kernel $L_{h,N}^{\pm}$: in setting (a) multiplicity m determines the degree of ill–posedness of

the deconvolution problem, and in both settings coefficients $C_{j,m}$ in (2.9) and (2.10) grow with m affecting the variance of the corresponding estimators in the case of heavy tailed densities f. Intuitively, the larger multiplicity m, the flatter the characteristic function $\phi_g(z)$ in the vicinity of zeros, and the harder the deconvolution problem.

Based on the derived kernels we define the estimators of $f(x_0)$ in examples (a) and (b) by

$$(a) \quad \hat{f}^{\pm}_{h,N}(x_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{(\pm 2\theta)^m}{h^{m+1}} \sum_{j=0}^{N} C_{j,m} K^{(m)} \left(\frac{Y_i - x_0 \mp \theta(2j+m)}{h} \right); \quad (2.11)$$

$$(b) \quad \hat{f}^{\pm}_{h,N}(x_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{(\pm 2)^m}{h} \sum_{j=0}^{N} C_{j,m} K \left(\frac{Y_i - x_0 \mp j}{h} \right), \quad (2.12)$$

where h and N are tuning parameters that should be specified.

3. Minimax results

In this section we derive upper bounds on the risk of the estimators constructed in the previous section, and show that they are rate optimal over functional classes characterized by smoothness and tail conditions. The analysis of the risk for both estimators in cases (a) and (b) coincides in almost every detail. Therefore in the sequel we concentrate on the example (a); the corresponding results for binomial error distribution are discussed in Section 5.

3.1. Functional classes

The following assumption introduces the functional class over which the accuracy of $\hat{f}^{\pm}_{h,N}(x_0)$ will be assessed.

Assumption (F). Let A and B be a positive real numbers.

(I) For $\alpha > 0$, a probability density f belongs to the functional class $\mathscr{H}_\alpha(A)$ if f is $\lfloor \alpha \rfloor := \max\{n \in \mathbb{N} \cup \{0\} : n < \alpha\}$ times continuously differentiable, and

$$\left| f^{(\lfloor \alpha \rfloor)}(t) - f^{(\lfloor \alpha \rfloor)}(t') \right| \leq A|t - t'|^{\alpha - \lfloor \alpha \rfloor}, \quad \forall t, \, t' \in \mathbb{R}. \quad (3.1)$$

(II) Let q be a positive real number. We say that a probability density f belongs to the functional class $\mathscr{N}_q(B)$ if

$$f(t) \leq B|t|^{-q}, \quad \forall t \in \mathbb{R}. \quad (3.2)$$

Combining the two conditions in Assumption (F), we define the following functional class:

$$\mathscr{W}_{\alpha,q}(A, B) := \mathscr{H}_\alpha(A) \cap \mathscr{N}_q(B).$$

Remark. While first assumption defines the traditional Hölder class, the second condition imposes a uniform upper bound on the decay of the tails of the density. Note that this tail condition is comparable to the moment condition in [2, Definition 3].

3.2. Rates of convergence

Now we are in a position to establish upper bounds on the maximal risk of the estimator $\hat{f}_{h,N}^{\pm}(x_0)$ defined in (2.11). Let

$$
\hat{f}_{h,N}(x_0) := \left\{ \begin{array}{ll} \hat{f}_{h,N}^{+}(x_0), & x_0 \geq 0; \\ \hat{f}_{h,N}^{-}(x_0), & x_0 < 0, \end{array} \right. \tag{3.3}
$$

$$
r := \left\{ \begin{array}{ll} (\alpha/q)(2m-1-q), & q < 2m-1; \\ 0, & q \geq 2m-1, \end{array} \right. \quad \text{and} \quad \nu := \frac{\alpha}{2\alpha + 2m + 1 + r}; \tag{3.4}
$$

and define

$$
\varphi(n) := \left\{ \begin{array}{ll} \left(B^{1/\alpha}A^{\frac{2m+1}{\alpha}}\right)^{\nu}n^{-\nu}, & q > 2m-1; \\[2mm] \left(B^{1/\alpha}A^{\frac{2m+1}{\alpha}}\right)^{\nu}\left(\frac{\log n}{n}\right)^{\nu}, & q = 2m-1; \\[2mm] \left(B^{\frac{2m-1}{\alpha q}}A^{\frac{2m+1}{\alpha}}\right)^{\nu}n^{-\nu}, & q < 2m-1. \end{array} \right. \tag{3.5}
$$

Theorem 1. *Let $\phi_g(z) = [\sinh(\theta z)/(\theta z)]^m$, $m \in \mathbb{N}$. Assume that $f \in \mathscr{W}_{\alpha,q}(A, B)$ with $q > 0$, and let $\hat{f}_{h,N}(x_0)$ be the estimator defined in (3.3) and (2.11) and associated with kernel K satisfying condition (2.2) with parameter $k \geq \alpha + 1$. Then with $h = h_*$ and $N = N_*$ defined in (A.6)–(A.8) in the proof of the theorem one has*

$$
\limsup_{n \to \infty} \left\{ [\varphi(n)]^{-1} \mathcal{R}_n[\hat{f}_{h_*,N_*}; \mathscr{W}_{\alpha,q}(A, B)] \right\} \leq C_1,
$$

where C_1 is a constant independent of A and B.

Remark.

(a) The result of Theorem 1 shows how the tail behavior of f and zeros multiplicity m affect the estimation accuracy. If the tail of f is sufficiently light, i.e., $q > 2m - 1$, then the risk of $\hat{f}_{h_*,N_*}(x_0)$ converges to zero at the rate $n^{-\alpha/(2\alpha+2m+1)}$ which is obtained in the ordinary smooth case with $\gamma = m$ and non–vanishing characteristic function ϕ_g [see Assumption (E0)]. On the other hand, for heavy tailed densities f with $q < 2m - 1$ the maximal risk of $\hat{f}_{h_*,N_*}(x_0)$ converges at a slower rate, and parameter r in (3.4) characterizes the deterioration in the convergence rate.

(b) The existence of different regimes depending on the tail behavior of f and zeros multiplicity m has been noticed in [2]; however, the case of heavy tailed densities has not been studied there.

Next theorem provides a lower bound on the minimax risk of estimation over the functional class $\mathscr{W}_{\alpha,q}(A, B)$.

Theorem 2. *Let* $\phi_g(z) = [\sinh(\theta z)/(\theta z)]^m$, $m \in \mathbb{N}$. *Then for any fixed* $\alpha > 0$, $q > 1$, $A > 0$ *and* $B > 0$, *one has*

$$\liminf_{n \to \infty} \left\{ \left(A^{-(2m+1)/\alpha} \, n \right)^{\nu} \mathcal{R}_n^* \left[\mathscr{W}_{\alpha,q}(A, B) \right] \right\} \geq C_2,$$

where ν *is defined in (3.4), and* C_2 *is a positive constant independent of* A.

Remark.

(a) Theorems 1 and 2 show that there are two regimes in behavior of the minimax risk. These regimes are characterized by the tail behavior of the estimated density f and the multiplicity of zeros of the error characteristic function ϕ_g. In the *light tail regime*, $q > 2m - 1$, zeros of ϕ_g have no influence on the minimax rate of convergence: it is fully determined by the tail behavior of ϕ_g. On the other hand, if $q < 2m - 1$ (the *heavy tail regime*) then zeros of ϕ_g have significant influence on the minimax rate, it becomes much slower than in the case of non–vanishing ϕ_g.

(b) Theorems 1 and 2 demonstrate that the proposed estimator $\hat{f}_{h_*, N_*}(x_0)$ is rate optimal in both *light tail* and *heavy tail regimes*. We note that on the boundary between two regimes, $q = 2m - 1$, there is a logarithmic gap between the upper and lower bounds of Theorems 1 and 2.

Thus far, the risk evaluations are under the functional class $\mathscr{W}_{\alpha,q}(A, B)$ defined in Assumption (F). Although these conditions are pretty reasonable in the context of the density deconvolution, they involve an extra assumption on the tail behavior of f, and it is natural to ask what happens when the tail condition does not hold. The next result provides an answer to this question.

Corollary 1. *Let* $\phi_g(z) = [\sinh(\theta z)/(\theta z)]^m$, $m \in \mathbb{N}$; *then*

$$\liminf_{n \to \infty} \left\{ \psi_n^{-1} \mathcal{R}_n^* [\mathscr{H}_\alpha(A)] \right\} \geq C_3; \tag{3.6}$$

$$\limsup_{n \to \infty} \left\{ \psi_n^{-1} \mathcal{R}_n^* [\mathscr{H}_\alpha(A) \cap \mathscr{N}_1(B)] \right\} \leq C_4, \tag{3.7}$$

where $\psi_n := (A^{(2m+1)/\alpha}/n)^{\frac{\alpha}{2m\alpha + 2m + 1}}$, *and* C_3 *and* C_4 *do not depend on* A.

Remark. In view of (3.6), the rate of convergence ψ_n on the functional class $\mathscr{H}_\alpha(A)$ is significantly slower than the one achieved on $\mathscr{H}_\alpha(A)$ in the setting with non–vanishing characteristic function ϕ_g. Note that the upper bound in (3.7) is achieved on a slightly smaller functional class. The assumption $f \in \mathscr{N}_1(B)$ is very mild and is fulfilled for virtually any probability density. However it does not hold uniformly for all densities. We were not able to derive the upper bound (3.7) without this additional condition.

4. Adaptive procedure

The minimax results in the previous section can only be achieved when the information on the functional class is known to us in advance. This is evident by observing that the optimal choice of tuning parameters h_* and N_* requires knowledge of the functional class. However, in most of applications, it is extremely rare to have the advance information about the functional class where the target function f resides in. Therefore, it is natural to ask whether one can construct an estimator that guarantees an equivalent or comparable accuracy without knowing the parameters of functional class.

In this section we develop an adaptive estimator of $f(x_0)$ whose construction is based on the idea of data–driven selection from a family of estimators $\{\hat{f}_{h,N}(x_0) : (h, N) \in \mathcal{H} \times \mathcal{N}\}$, where $\hat{f}_{h,N}(x_0)$ is defined in the previous section, and $\mathcal{H}$ and $\mathcal{N}$ are some fixed sets of bandwidths and cut–off parameters. Since the estimator $\hat{f}_{h,N}(x_0)$ depends on two tuning parameters, we adopt the general method of adaptive estimation proposed in [14].

4.1. Selection rule

Let $\mathcal{H}$ and $\mathcal{N}$ be discrete sets defined as follows: for real numbers $0 < h_{\min} < h_{\max} = \theta$ and integer number $N_{\max}$ to be specified later

$$\mathcal{H} := \big\{ h \in [h_{\min}, h_{\max}] : h = 2^{-j} h_{\max}, \ j = 0, \ldots, M_h \big\};$$
$$\mathcal{N} := \big\{ j : j = 1, \ldots, N_{\max} =: M_N \big\},$$

where $M_h := \lfloor \log_2(h_{\max}/h_{\min}) \rfloor$ and $M_N := N_{\max}$ denote the cardinality of $\mathcal{H}$ and $\mathcal{N}$ respectively. Let $\mathcal{T} := \mathcal{H} \times \mathcal{N}$ with a corresponding element $\tau := (h, N)$, and consider the family of estimators $\mathcal{F}(\mathcal{T}) = \{ \hat{f}_\tau^\pm(x_0), \ \tau \in \mathcal{T} \}$, where $\hat{f}_\tau^\pm(x_0) = \hat{f}_{h,N}^\pm(x_0)$ is defined in (2.11) and (3.3). The adaptive estimator is based on a data–driven selection from the family $\mathcal{F}(\mathcal{T})$. For the sake of definiteness in the sequel we assume that $x_0 \geq 0$ and consider estimators $\hat{f}_\tau^+(x_0)$ only; the case $x_0 < 0$ and $\hat{f}_\tau^-(x_0)$ can be handled in exactly the same way.

The selection rule is based on auxiliary estimators that are constructed as follows. We first define an operation of coordinate–wise maximum and minimum:

$$\tau \Cap \tau' := (h \vee h', N \wedge N'), \quad \forall \tau, \tau' \in \mathcal{T}.$$

Then, we associate the estimator [cf. (2.11)] with any pair $\tau, \tau' \in \mathcal{T}$

$$\hat{f}_{\tau \Cap \tau'}^+(x_0) := \frac{1}{n} \sum_{i=1}^n \frac{(2\theta)^m}{(h \vee h')^{m+1}} \sum_{j=0}^{N \wedge N'} C_{j,m} K^{(m)} \left(\frac{Y_i - x_0 - \theta(2j + m)}{h \vee h'} \right).$$

Observe that $\hat{f}_{\tau \Cap \tau'}^+(x_0) = \hat{f}_{\tau' \Cap \tau}^+(x_0)$ for all $\tau, \tau' \in \mathcal{T}$.

Selection rules based on convolution–type auxiliary kernel estimators are developed in [14, 15], while Lepski [22] uses auxiliary estimators that are based on

the operation of point–wise maximum of multi–bandwidths. Our construction is close in spirit to the latter one; it is dictated by the structure of estimators $\hat{f}_{h,N}^{\pm}(x_0)$ in the deconvolution problem.

An important ingredient in the construction of the proposed selection rule is a uniform upper bound on the stochastic error of estimator $\hat{f}_{\tau}^{+}(x_0)$, $\tau \in \mathcal{T}$. For $\tau \in \mathcal{T}$ the stochastic error of $\hat{f}_{\tau}^{+}(x_0)$ is

$$\xi_\tau(x_0) := \frac{1}{n} \sum_{i=1}^{n} L_\tau^+(Y_i - x_0) - \mathrm{E}_f\left[L_\tau^+(Y_1 - x_0)\right], \tag{4.1}$$

where

$$L_\tau^+(y) := \frac{(2\theta)^m}{h^{m+1}} \sum_{j=0}^{N} C_{j,m} K^{(m)}\left(\frac{y - \theta(2j + m)}{h}\right);$$

Cf., (2.9). Define

$$\sigma_\tau^2 := \frac{(2\theta)^{2m}}{h^{2m+2}} \sum_{j=0}^{N} C_{j,m}^2 \int_{-\infty}^{\infty} \left| K^{(m)}\left(\frac{y - x_0 - \theta(2j + m)}{h}\right) \right|^2 f_Y(y)\mathrm{d}y. \tag{4.2}$$

The proof of Theorem 1 shows that $\mathrm{var}_f[\xi_\tau(x_0)] \leq \sigma_\tau^2/n$. Let

$$u_\tau := 2^{m+1} \theta^m C_{N,m} \|K^{(m)}\|_\infty h^{-m-1}, \tag{4.3}$$

and for real number $\varkappa > 0$ that will be specified later we put

$$\Lambda_\tau(\varkappa) := \sigma_\tau \sqrt{\frac{2\varkappa}{n}} + \frac{2u_\tau \varkappa}{3n}. \tag{4.4}$$

In Lemma 1 in Appendix we demonstrate that $\Lambda_\tau(\varkappa)$ is a uniform upper bound on $|\xi_\tau(x_0)|$ in the sense that all moments of the random variable $\sup_{\tau \in \mathcal{T}}[|\xi_\tau(x_0)| - \Lambda_\tau(\varkappa)]_+$ are suitably small as $\varkappa$ increases. Note however that $\Lambda_\tau(\varkappa)$ cannot be used in the selection rule because it depends on the unknown density. In order to overcome this problem we consider a data–driven uniform upper bound on $\xi_\tau(x_0)$ that is constructed as follows.

For $\tau \in \mathcal{T}$ let

$$\hat{\sigma}_\tau^2 := \frac{1}{n} \sum_{i=1}^{n} \frac{(2\theta)^{2m}}{h^{2m+2}} \sum_{j=0}^{N} C_{j,m}^2 \left| K^{(m)}\left(\frac{Y_i - x_0 - \theta(2j + m)}{h}\right) \right|^2.$$

Note that $\hat{\sigma}_\tau^2$ is the empirical estimator of σ_τ^2. Let

$$\hat{\Lambda}_\tau(\varkappa) := 7\left(\hat{\sigma}_\tau \sqrt{\frac{2\varkappa}{n}} + \frac{2u_\tau \varkappa}{3n}\right). \tag{4.5}$$

With the introduced notations the selection rule is defined for any $\tau \in \mathcal{T}$

$$\hat{R}_\tau(x_0) := \sup_{\tau' \in \mathcal{T}} \left[\left| \hat{f}_{\tau \bowtie \tau'}^+(x_0) - \hat{f}_{\tau'}^+(x_0) \right| - \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa) - \hat{\Lambda}_{\tau'}(\varkappa) \right]_+$$

$$+ \hat{\Lambda}_\tau(\varkappa) + \sup_{\tau' \in \mathcal{T}} \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa). \qquad (4.6)$$

Then, the adaptive estimator $\hat{f}_*(x_0)$ can be chosen to be

$$\hat{f}_*(x_0) := \hat{f}_{\hat{\tau}}^+(x_0), \quad \hat{\tau} = (\hat{h}, \hat{N}) := \arg\min_{\tau \in \mathcal{T}} \hat{R}_\tau(x_0). \qquad (4.7)$$

Remark. The defined selection rule is fully data–driven; it only requires specification of parameter $\varkappa$ in (4.5). This parameter provides a uniform control of the stochastic errors for the family of estimators $\mathcal{F}(\mathcal{T})$, and has no relation to the properties of the density to be estimated. In addition, the parameters $h_{\min}$ and $N_{\max}$ should be chosen; they determine the sets of admissible bandwidths $\mathcal{H}$ and cut–off parameters $\mathcal{N}$.

4.2. Oracle inequality and rates of convergence

For $h, h' \in \mathcal{H}$ and $N, N' \in \mathcal{N}$ define

$$\bar{B}_h(f) \quad := \quad \sup_{h' \leq h} \sup_{x \in \mathbb{R}} \left| \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{t - x}{h}\right)[f(t) - f(x)]\mathrm{d}t \right|; \qquad (4.8)$$

$$\bar{B}_N(x_0; f) \quad := \quad \max_{1 \leq j \leq m} \sup_{|t| \leq \theta} \sup_{N' \geq N} \left[f\big(t + x_0 + 2\theta(N' + 1)j\big) \right], \qquad (4.9)$$

and let

$$\bar{B}_\tau(x_0; f) := 2^{m+1} \Big[\bar{B}_h(f) + (1 + \|K\|_1)\bar{B}_N(x_0; f) \Big]. \qquad (4.10)$$

Theorem 3. *Let $\hat{f}_*(x_0)$ be the estimator defined in (4.6)–(4.7) and associated with parameter $\varkappa > 0$; then*

$$|\hat{f}_*(x_0) - f(x_0)| \leq C_1 \inf_{\tau \in \mathcal{T}} \left\{ \bar{B}_\tau(x_0; f) + \Lambda_\tau(\varkappa) \right\} + C_2\Big(\delta(x_0) + \frac{\varkappa}{n}\Big), \qquad (4.11)$$

where C_1 is an absolute constant, C_2 depends only on m and θ, and $\delta(x_0)$ is a non–negative random variable whose moments admit the following bound: for any $p \geq 1$

$$\mathrm{E}_f\big[\delta(x_0)\big]^p \leq C_3 M_h M_N[\bar{\Lambda}(\varkappa)]^p \varkappa^{-p} e^{-\varkappa}, \qquad (4.12)$$

where $\bar{\Lambda}(\varkappa) := \sup_{\tau \in \mathcal{T}}\{(1 + u_\tau)\Lambda_\tau(\varkappa)\}$, and constant C_3 depends on p only.

Remark. Explicit expressions for constants C_1, C_2 and C_3 appear in the proofs of Theorem 3 and Lemma 2. Note that the oracle inequality holds for any probability density f, without any functional class assumptions.

The oracle inequality (4.11) allows us to derive the following result on the accuracy of the adaptive estimator $\hat{f}_*(x_0)$ on the class $\mathscr{W}_{\alpha,q}(A, B)$.

Corollary 2. *Let $\mathcal{F}(\mathcal{T})$ be the family of estimators $\big\{\hat{f}_{h,N}^+(x_0), (h, N) \in \mathcal{H} \times \mathcal{N}\big\}$ with*

$$h_{\min} := \Big(\frac{\log n}{n}\Big)^{1/(2m+1)}, \quad h_{\max} := \theta, \quad N_{\max} := \Big(\frac{n}{\log n}\Big)^{1/(2m)}. \qquad (4.13)$$

Let $\hat{f}_*(x_0)$ be the estimator defined by selection rule (4.6)–(4.7) and associated with parameter $\varkappa = \varkappa_* := 5 \log n$; then for any $\alpha > 0$, $q \geq 1$, $A > 0$ and $B > 0$ one has

$$\limsup_{n \to \infty} \left\{ \left[\varphi \left(\frac{n}{\log n} \right) \right]^{-1} \mathcal{R}_n[\hat{f}_{h_*, N_*}; \mathscr{W}_{\alpha, q}(A, B)] \right\} \leq C,$$

where $\varphi(\cdot)$ is defined in (3.5), and C does not depend on A and B.

Remark. Note that the resulting rate is the same as the rate of convergence in Theorem 1 except for the extra $\log n$ factor. It is a well-known fact by Lepski [21] that this factor cannot be avoided in the adaptive nonparametric estimation of a function at a single point.

5. Concluding remarks

We close this paper with a few concluding remarks.

In this paper we concentrated on the setting when the error distribution is the m–fold convolution of the uniform distribution on $[-\theta, \theta]$. Here the error characteristic function has infinite number of isolated zeros on the imaginary axis, each of them has the same multiplicity m. Note that the results of Theorems 1, 2, and Corollary 2 also hold for the binomial error distribution $\mathrm{Bin}(m, 1/2)$ with the following minor changes in notation: in (3.4) parameter ν should be redefined as $\nu = 1/(2\alpha + 1 + r)$, and in (3.5) and in the statement of Theorem 2 expression $A^{(2m+1)/\alpha}$ should be replaced by $A^{1/\alpha}$. The specific form of the error characteristic functions used in this paper facilitates derivation of lower bounds on the minimax risk. However, in general, the proposed technique is applicable to other error distributions whose characteristic function has zeros on the imaginary axis.

We developed rate optimal estimators with respect to the point–wise risk. It is worth noting that there is a significant difference between settings with the point–wise and $\mathbb{L}_2$–risks when the error characteristic function has zeros on the imaginary axis. In particular, the minimax rates of convergence and the *heavy tail regimes* in these settings are determined by completely different formulas. This fact has been already noticed in [2]. Some results for density deconvolution with $\mathbb{L}_2$–risk for non–standard error distributions appeared in [24] and [18]. However, results in these papers are not directly comparable to ours. In general, deconvolution problems under global losses with non–standard error distributions deserve a thorough study.

Appendix A: Proofs

A.1. Proof of Theorem 1

Proof. In the subsequent proof $c_1, c_2, \ldots$ stand for positive constants independent of A and B. Without loss of generality we assume that $x_0 \geq 0$; the proof for the case $x_0 < 0$ is identical in every detail. We follow the ideas of the proof of Theorem 2 in [2].

(a). We begin with bounding the variance of $\hat{f}^+_{h,N}(x_0)$. It is shown in [2] that the variance of $\hat{f}^+_{h,N}(x_0)$ is bounded from above as follows

$$\mathrm{var}_f\big[\hat{f}^+_{h,N}(x_0)\big] \leq \frac{(2\theta)^{2m}}{nh^{2m+2}} \sum_{j=0}^{N} C^2_{j,m} \int_{-\infty}^{\infty} \Big|K^{(m)}\Big(\frac{y - x_0 - \theta(2j+m)}{h}\Big)\Big|^2 f_Y(y)\mathrm{d}y$$

$$\leq \frac{c_1\theta^{2m}}{nh^{2m+1}} \sum_{j=0}^{N} \frac{C^2_{j,m}}{h} \int_{I_j(x_0)} f_Y(t)\mathrm{d}t, \tag{A.1}$$

where $I_j(x_0) := [x_0 + \theta(2j+m) - h, x_0 + \theta(2j+m) + h]$. Moreover, by [2, (A.16)],

$$\frac{1}{h} \int_{I_j(x_0)} f_Y(y)\mathrm{d}y \leq \frac{c_2}{\theta} \int_{-h}^{h} f(t + x_0 + 2(j+m)\theta)\mathrm{d}t + \frac{c_3}{\theta} \int_{-h}^{h} f(t + x_0 + 2j\theta)\mathrm{d}t$$

$$+ \frac{c_4}{\theta} \int_{-m\theta}^{m\theta} f(t + x_0 + (2j+m)\theta)\mathrm{d}t =: S_{1,j} + S_{2,j} + S_{3,j}.$$

We have

$$\sum_{j=0}^{N} C^2_{j,m} S_{1,j} = \frac{c_2}{\theta} \sum_{j=0}^{N} C^2_{j,m} \int_{-h}^{h} f(t + x_0 + 2(j+m)\theta)\mathrm{d}t$$

$$\leq c_5 \sum_{j=0}^{N} \frac{j^{2m-2}}{\theta} \int_{x_0+2(j+m)\theta-h}^{x_0+2(j+m)\theta+h} \frac{t^q f(t)}{(x_0 + 2\theta j)^q}\mathrm{d}t \leq \frac{c_6 B h}{\theta^{q+1}} \sum_{j=0}^{N} j^{2m-q-2}, \quad \text{(A.2)}$$

where we have used that $C_{j,m} = \binom{j+m-1}{m-1} \leq c_0 j^{m-1}$, $f \in \mathcal{N}_q(B)$ and $\theta > h$ for large n. The term $\sum_{j=0}^{N} C^2_{j,m} S_{2,j}$ is also bounded from above by the same expression as on the right hand side of (A.2). Furthermore,

$$\sum_{j=0}^{N} C^2_{j,m} S_{3,j} = c_4 \sum_{j=0}^{N} \frac{C^2_{j,m}}{\theta} \int_{-m\theta}^{m\theta} f(t + x_0 + (2j+m)\theta)\mathrm{d}t$$

$$\leq \frac{c_8}{\theta} \sum_{j=0}^{N} j^{2m-2} \int_{x_0+2j\theta}^{x_0+2(j+m)\theta} \frac{t^q f(t)}{(x_0 + 2\theta j)^q}\mathrm{d}t \leq \frac{c_9 B}{\theta^q} \sum_{j=0}^{N} j^{2m-q-2}. \tag{A.3}$$

Combining (A.3), (A.2) and (A.1) we conclude that

$$\mathrm{var}_f\big[\hat{f}^+_{h,N}(x_0)\big] \leq \frac{c_{10}\theta^{2m-q}B\psi_N}{nh^{2m+1}}, \quad \psi_N := \begin{cases} 1, & q > 2m - 1; \\ \log N, & q = 2m - 1; \\ N^{2m-q-1}, & q < 2m - 1. \end{cases}$$

$$\tag{A.4}$$

(b). Now we bound the bias of $\hat{f}^+_{h,N}(x_0)$. It is shown in [2] that

$$\mathrm{E}_f\big[\hat{f}^+_{h,N}(x_0)\big] = \frac{1}{h} \int_{-\infty}^{\infty} K\Big(\frac{t - x_0}{h}\Big) f(t)\mathrm{d}t + T_N(f; x_0),$$

where

$$T_N(f; x_0) = \sum_{j=1}^{m} \binom{m}{j} \int_{-1}^{1} K(y) f(yh + x_0 + 2\theta(N+1)j) \mathrm{d}y.$$

Taking into account that $f \in \mathcal{N}_q(B)$ we obtain for any $j = 1, \ldots, m$

$$\int_{-1}^{1} |K(y)| f(yh + x_0 + 2\theta(N+1)j) dy \leq \frac{c_{11}}{h} \int_{x_0+2\theta(N+1)j-h}^{x_0+2\theta(N+1)j+h} f(y) dy$$

$$\leq \frac{c_{12} B h}{h(x_0 + 2\theta N)^q} \leq \frac{c_{13} B}{(\theta N)^q}.$$

This leads to the following upper bound on the bias of $\hat{f}_{h,N}(x_0)$:

$$\left| \mathrm{E}_f \left[\hat{f}_{h,N}^+(x_0) \right] - f(x_0) \right| \leq c_{14} \left(A h^\alpha + \frac{B}{\theta^q N^q} \right). \tag{A.5}$$

(c). We complete the proof by combining the bounds in (A.4) and (A.5) for the cases $q > 2m-1$, $q = 2m-1$ and $q < 2m-1$. Straightforward algebra shows that the following choice of $h = h_*$ and $N = N_*$ yields the theorem result:

(i) if $q > 2m - 1$ then we set

$$h_* = c_1 \left(\frac{B}{A^2 n} \right)^{\frac{1}{2\alpha+2m+1}}, \quad N_* \geq c_2 \left(\frac{B^{\alpha+2m+1} n^\alpha}{A^{2m+1}} \right)^{\frac{1}{q(2\alpha+2m+1)}}; \tag{A.6}$$

(ii) if $q = 2m - 1$ then

$$h_* = c_3 \left(\frac{B \log n}{A^2 n} \right)^{\frac{1}{2\alpha+2m+1}}, \quad N_* = c_4 \left\{ \frac{B^{\alpha+2m+1}}{A^{2m+1}} \left(\frac{n}{\log n} \right)^\alpha \right\}^{\frac{1}{q(2\alpha+2m+1)}}; \tag{A.7}$$

(iii) if $q < 2m - 1$ then

$$h_* = c_5 \left(\frac{B^{(2m-1)/q}}{A^{(2m+q-1)/q}} \frac{1}{n} \right)^{\frac{1}{2\alpha+2m+1+r}}, \quad N_* = c_6 (B/A)^{1/q} h_*^{-\alpha/q}, \tag{A.8}$$

where constants $c_1, \ldots c_6$ do not depend on A and B. $\qquad \square$

A.2. Proof of Theorem 2

Proof. Without loss of generality we fix x_0 to be 0. The proof is split into a few steps: (i) defines two functions in $\mathscr{W}_{\alpha,q}(A, B)$ and provides their point–wise distance; (ii) bounds the χ^2–divergence between densities of the observations; (iii) specifies the proper tuning parameters and provides the rate for the lower bound, and (iv) deals with derivation of the lower bound for the light tail regime.

(i). For $s > 1/2$ define

$$f_0(x) := \frac{C(s)}{(1 + x^2)^s}, \quad x \in \mathbb{R}, \tag{A.9}$$

where $C(s)$ is a normalizing constant depending on s. Then, $f_0 \in \mathcal{N}_q(B)$ for $1 < q \le 2s$ since $f_0(x) \le C(s)/x^{2s} \le B/x^q$ for $x > 1$ with properly chosen $B > 0$. In addition, since f_0 is infinitely differentiable, $f_0 \in \mathscr{H}_\alpha(A)$ for any α with properly chosen A.

Define a function η_0 on $\mathbb{R}$ via its Fourier transform $\phi_{\eta_0}(\omega) = \int_{-\infty}^{\infty} \eta_0(x) e^{-i\omega x} dx$ as follows. Let ϕ_{η_0} be an infinitely differentiable function on $\mathbb{R}$ with the following properties:

(a) ϕ_{η_0} is supported on $[-1, 1]$;
(b) ϕ_{η_0} is symmetric, $\phi_{\eta_0}(\omega) = \phi_{\eta_0}(-\omega)$, $\forall \omega \in \mathbb{R}$;
(c) given some fixed $\delta \in (0, 1/8)$, $\phi_{\eta_0}(\omega) = 1$ for $\omega \in [0, 1 - \delta)$, $\phi_{\eta_0}(\omega) = 0$ for $\omega \ge 1$, and ϕ_{η_0} is monotone decreasing on $[1 - \delta, 1)$.

Given $h \in (0, \pi/\theta)$ and $N \in \mathbb{N}$, define

$$\phi_\eta(\omega) := \sum_{k=N+1}^{2N} \left\{ \phi_{\eta_0}\left(\frac{\omega - \pi k/\theta}{h}\right) + \phi_{\eta_0}\left(\frac{\omega + \pi k/\theta}{h}\right) \right\}. \tag{A.10}$$

Note that ϕ_η is supported on:

$$\bigcup_{k=N+1}^{2N} A_k(h), \quad A_k(h) := \left[\frac{-\pi k}{\theta} - h, \frac{-\pi k}{\theta} + h\right] \cup \left[\frac{\pi k}{\theta} - h, \frac{\pi k}{\theta} + h\right]. \tag{A.11}$$

Then, define a function η through the inverse Fourier transform as follows:

$$\eta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_\eta(\omega) e^{i\omega x} d\omega = 2h\eta_0(hx) \sum_{k=N+1}^{2N} \cos\left(\frac{\pi k x}{\theta}\right) \text{ for } x \in \mathbb{R}. \tag{A.12}$$

In the subsequent proof the parameters h and N are specified so that $h \to 0$ and $N \to \infty$ as $n \to \infty$; thus, we tacitly assume that N is large and h is small for large enough sample size n.

Given real numbers $M > 0$ and $c_0 > 0$, define

$$f_1(x) := f_0(x) + c_0 M \eta(x). \tag{A.13}$$

We demonstrate that under appropriate choice of c_0 and M, f_1 is a probability density from $\mathscr{W}_{\alpha,q}(A, B)$ for any h and N. Observe that $\phi_\eta(0) = 0$ implies $\int_{-\infty}^{\infty} \eta(x) dx = 0$ so that f_1 integrates to one. Moreover, since ϕ_{η_0} is infinitely differentiable and compactly supported, η_0 is a rapidly decreasing function, i.e., $|\eta_0^{(j)}(x) x^\ell| \le c_{j,l}$ for any $j, \ell = 0, 1, 2, \ldots$. In particular, for some constant $c_1(s)$ depending on s only one has $|\eta_0(x)| \le c_1(s)|x|^{-2s}$ for all $x \in \mathbb{R}$. It follows from (A.12) that $|\eta(x)| \le c_2 h^{-2s+1}|x|^{-2s} N$ for $x \in \mathbb{R}$. Therefore choosing

$$M = h^{2s-1} N^{-1}$$

we obtain $c_0 M |\eta(x)| \le f_0(x)$ for c_0 small enough. Therefore f_0 is non–negative, and it is a probability density. Moreover, $f_1 \in \mathcal{N}_q(B)$ for $q \le 2s$. If α is a

positive integer then it follows from (A.12) that

$$\left|\eta^{(\alpha)}(x)\right| = \left|2h\sum_{i=0}^{\alpha}\binom{\alpha}{i}h^i\eta_0^{(i)}(xh)\sum_{k=N+1}^{2N}\cos^{(\alpha-i)}(\pi kx/\theta)\right|$$

$$\leq c_2 h\sum_{i=0}^{\alpha}h^i N^{\alpha-i+1} \leq c_3 h N^{\alpha+1}.$$

Therefore, we can ensure $f_1 \in \mathscr{H}_\alpha(A)$ by selecting h and N so that

$$MhN^{\alpha+1} = h^{2s}N^\alpha \leq A. \tag{A.14}$$

Thus, under (A.14) we have $f_0, f_1 \in \mathscr{W}_{\alpha,q}(A,B)$. In addition,

$$|f_1(0) - f_0(0)| = c_0 M\eta(0) = c_0 Mh\eta_0(0)N = c_4 h^{2s}. \tag{A.15}$$

(ii). Now we derive an upper bound on the χ^2–divergence between the densities of observations $f_{Y,0} = g \star f_0$ and $f_{Y,1} = g \star f_1$ that correspond to f_0 and f_1. Observe the following expression:

$$\chi^2(f_{Y,1}, f_{Y,0}) := \int_{-\infty}^{\infty}\frac{(f_{Y,1}(x) - f_{Y,0}(x))^2}{f_{Y_0}(x)}dx \overset{(A.13)}{=} c_0^2 M^2\int_{-\infty}^{\infty}\frac{|(g\star\eta)(x)|^2}{(g\star f_0)(x)}dx.$$

Consider the denominator, $g \star f_0$, of the integrand. We have

$$(g\star f_0)(x) = C(s)\int_{-\infty}^{\infty}\frac{g(y)}{[1+(x-y)^2]^s}dy$$

$$\geq C(s)\int_{-\infty}^{\infty}\frac{g(y)}{2^s(1+y^2)^s(1+x^2)^s}dy \geq \frac{c_5}{(1+x^2)^s},$$

where we have used the elementary inequality $1 + |x-y|^2 \leq 2(1+|x|^2)(1+|y|^2)$, $\forall x, y$. Then the χ^2–divergence can be bounded:

$$\chi^2(f_{Y,1}; f_{Y,0}) \leq c_6 M^2\int_{-\infty}^{\infty}|(g\star\eta)(x)|^2 dx + c_7 M^2\int_{-\infty}^{\infty}x^{2s}|(g\star\eta)(x)|^2 dx. \tag{A.16}$$

Let us handle the second integral on the right-hand side. For any positive integer number s we have

$$\int_{-\infty}^{\infty}x^{2s}|(g\star\eta)(x)|^2 dx = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left|\frac{d^s}{d\omega^s}\phi_g(\omega)\phi_\eta(\omega)\right|^2 d\omega. \tag{A.17}$$

Note that

$$\frac{d^s}{d\omega^s}\phi_g(\omega)\phi_\eta(\omega) = \sum_{j=0}^{s}\binom{s}{j}\phi_g^{(j)}(\omega)\phi_\eta^{(s-j)}(\omega)$$

$$= \sum_{j=0}^{s} \binom{s}{j} \frac{\phi_g^{(j)}(\omega)}{h^{s-j}} \sum_{k=N+1}^{2N} \left\{ \phi_{\eta_0}^{(s-j)}\left(\frac{\omega - \pi k/\theta}{h}\right) + \phi_{\eta_0}^{(s-j)}\left(\frac{\omega + \pi k/\theta}{h}\right) \right\}.$$

Furthermore, $\phi_g^{(j)}$ can be expanded by Faá di Bruno formula for $j \in \mathbb{N}$: if $\phi_{g_0}(\omega) := \sin(\theta\omega)/(\theta\omega)$ then $\phi_g(\omega) = [\phi_{g_0}(\omega)]^m$ and

$$\phi_g^{(j)}(\omega) = \frac{d^j}{d\omega^j} \left(\frac{\sin\theta\omega}{\theta\omega}\right)^m$$
$$= \sum_{l=1}^{j} j\cdots(j-l+1)\left(\frac{\sin\theta\omega}{\theta\omega}\right)^{m-l} B_{j,l}\left(\phi_{g_0}'(\omega), \ldots, \phi_{g_0}^{(j-l+1)}(\omega)\right),$$

where $B_{j,l}$ denotes the Bell polynomials. Recall that $B_{j,l}$ is a homogeneous polynomial in j variables of degree l, and note that $|\phi_{g_0}^{(j)}(\omega)| \le c_8(|\omega|^{-1} \wedge 1)$, $\forall j$. Then,

$$\left|\phi_g^{(j)}(\omega)\right| \le c_9 \sum_{l=1}^{j} \left|\frac{\sin\theta\omega}{\theta\omega}\right|^{m-l} |\theta\omega|^{-l} = \frac{c_9}{|\theta\omega|^m} \sum_{l=1}^{j} |\sin\theta\omega|^{m-l}. \qquad (A.18)$$

Combining the above results and the fact that sets $A_k(h)$ in (A.10) are disjoint for $k = N+1, \ldots, 2N$, we bound the integral in (A.17) as follows:

$$\int_{-\infty}^{\infty} \left| \sum_{j=0}^{s} \binom{s}{j} \frac{\phi_g^{(j)}(\omega)}{h^{s-j}} \sum_{k=N+1}^{2N} \left\{ \phi_{\eta_0}^{(s-j)}\left(\frac{\omega - \pi k/\theta}{h}\right) + \phi_{\eta_0}^{(s-j)}\left(\frac{\omega + \pi k/\theta}{h}\right) \right\} \right|^2 d\omega$$

$$\le c_{10} h^{-2s} \sum_{k=N+1}^{2N} \int_{A_k(h)} \left| \sum_{j=0}^{s} h^j \phi_g^{(j)}(\omega) \right|^2 d\omega$$

$$\le c_{11} h^{-2s} \sum_{k=N+1}^{2N} \int_{A_k(h)} \left(\left|\frac{\sin\theta\omega}{\theta\omega}\right|^{2m} + \frac{1}{|\theta\omega|^{2m}} \sum_{j=1}^{s} h^{2j} \sum_{l=1}^{j} |\sin\theta\omega|^{2m-2l} \right) d\omega$$

$$\le c_{12} h^{2m+1-2s} \sum_{k=N+1}^{2N} \frac{1}{k^{2m}} = c_{13} h^{2m-2s+1} N^{-2m+1}.$$

In addition, the first integral on the left-hand side in (A.16) can be bounded with $s = 0$, so that

$$\int_{-\infty}^{\infty} |(g \star \eta)(x)|^2 dx \le c_{14} h^{2m+1} N^{-2m+1}.$$

Therefore, for positive integer s,

$$\chi^2(f_{Y,1}; f_{Y,0}) \le c_{14} M^2 h^{2m+1} N^{-2m+1} + c_{13} M^2 h^{2m-2s+1} N^{-2m+1}$$
$$\le c_{15} h^{2m+2s-1} N^{-2m-1}. \qquad (A.19)$$

The same upper bound holds for any non-integer $s \geq 0$; this fact is due to the interpolation inequality for the Sobolev spaces, see, e.g., Aubin [1] for the details.

(iii). Now, based on (A.14) and (A.19), we specify parameters $h = h_*$ and $N = N_*$ as follows:

$$N_* := \left(\frac{A}{h_*^{2s}} \right)^{1/\alpha}, \qquad h_* := \left(\frac{A^{\frac{2m+1}{\alpha}}}{n} \right)^{\frac{\alpha}{(2m+2s-1)\alpha+2s(2m+1)}}.$$

Under this choice (A.14) holds, and $\chi^2(f_{Y,1}, f_{Y,0}) \leq c_{15}/n$. Then the lower bound on the minimax risk is obtained by plugging these expressions in (A.15) and letting $2s = q > 1$:

$$\mathcal{R}_n^*[\mathscr{W}_{\alpha,q}(A,B)] \geq c_4 \left(\frac{A^{\frac{2m+1}{\alpha}}}{n} \right)^{\frac{\alpha}{2\alpha+2m+1+(\alpha/q)(2m-1-q)}}. \tag{A.20}$$

(iv). To complete the proof of the theorem it remains to observe that in the considered problem the following standard lower bound on the minimax risk can be also established:

$$\mathcal{R}_n^*[\mathscr{W}_{\alpha,q}(A,B)] \geq c_4 \left(\frac{A^{\frac{2m+1}{\alpha}}}{n} \right)^{\frac{\alpha}{2\alpha+2m+1}}. \tag{A.21}$$

For completeness, we provide the proof sketch. Let f_0 be given by (A.9), and let η be the function defined via its Fourier transform ϕ_η as follows

$$\phi_\eta(\omega) = \phi_{\eta_0}(2\omega h - 3) + \phi_{\eta_0}(2\omega h + 3),$$

where ϕ_{η_0} is a function with properties (a)–(c). Obviously, ϕ_η is symmetric, supported on $[-2/h, -1/h] \cup [1/h, 2/h]$, and

$$\eta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\phi_{\eta_0}(2\omega h - 3) + \phi_{\eta_0}(2\omega h + 3) \right] e^{i\omega x} \mathrm{d}\omega = \frac{2}{h} \eta_0 \left(\frac{x}{2h} \right) \cos \left(\frac{3x}{2} \right).$$

The function f_1 is defined by (A.13), and the choice $M = Ah^{\alpha+1}$ and properties of function η_0 guarantee that f_1 is a density from the class $\mathscr{W}_{\alpha,q}(A,B)$ with $q \leq 2s$. With this construction $|f_0(0) - f_1(0)| = c_0 M \eta(0) = c_{16} A h^\alpha$. The upper bound on the χ^2–divergence between $f_{Y,0}$ and $f_{Y,1}$ is computed along the same lines as above with the following modifications. Now we apply (A.18) to get

$$\left| \frac{d^s}{d\omega^s} \phi_g(\omega) \phi_\eta(\omega) \right| \leq \sum_{j=0}^{s} \binom{s}{j} \left| \phi_g^{(j)}(\omega) \phi_\eta^{(s-j)}(\omega) \right| \leq c_{17} |\theta\omega|^{-m} \sum_{j=0}^{s} |\phi_\eta^{(s-j)}(\omega)|,$$

and, by properties of function ϕ_η,

$$\int_{-\infty}^{\infty} x^{2s} |(g \star \eta)(x)|^2 \mathrm{d}x \leq c_{18} \int_{1/h}^{2/h} |\omega|^{-2m} \mathrm{d}\omega = c_{19} h^{2m-1}.$$

The same upper bound holds for the integral $\int_{-\infty}^{\infty} |(g \star \eta)(x)|^2 \mathrm{d}x$ which leads to

$$\chi^2(f_{Y,1}; f_{Y,0}) \leq c_{20} M^2 h^{2m-1} = c_{20} A^2 h^{2\alpha+2m+1}.$$

Then (A.21) follows from the choice $h_* = (A^2 n)^{-1/(2\alpha+2m+1)}$.

Combining (A.20) and (A.21) and noting that the following relation holds for $1 < q < 2m - 1$

$$\frac{\alpha}{2\alpha + 2m + 1 + (\alpha/q)(2m - 1 - q)} \leq \frac{\alpha}{2\alpha + 2m + 1},$$

we complete the proof.   □

A.3. Proof of Corollary 1

Proof. The upper bound (3.7) is obtained directly from Theorem 1 applied with $q = 1$. We need to establish (3.6) only. The proof goes along the lines of the proof of Theorem 2 with minor modifications that are indicated below.

Define

$$f_0(x) := \frac{h}{\pi(1 + h^2 x^2)}, \quad x \in \mathbb{R},$$

where $h > 0$ is a parameter to be specified. Obviously, $f_0 \in \mathscr{H}_\alpha(A)$ for small enough h. Using the function η defined in (A.10), (A.11), and (A.12), let

$$f_1(x) := f_0(x) + c_0 M \eta(x) \text{ for } x \in \mathbb{R}.$$

Similarly to the proof of Theorem 2, $|\eta(x)| \leq c_1 h^{-1} N |x|^{-2}$. Set $M := N^{-1}$, so that $c_0 M |\eta(x)| = c_0 c_1 / (h|x|^2) \leq f_0(x)$ holds for sufficiently small c_0. Since we use the same function η in Theorem 2, we can ensure $f_1 \in \mathscr{H}_\alpha(A)$ by setting

$$Mh N^{\alpha+1} = h N^\alpha \leq A. \tag{A.22}$$

Therefore, for $x_0 = 0$, we have the following point-wise distance

$$|f_1(0) - f_0(0)| = c_1 M \eta(0) = c_1 M h \eta_0(0) N = c_2 h.$$

The bound on the χ^2–divergence takes the following form

$$\begin{aligned}
\chi^2(f_{Y,1}; f_{Y,0}) \leq & c_3 \frac{M^2}{h} \int_{-\infty}^{\infty} |(g \star \eta)(x)|^2 dx + c_4 M^2 h \int_{-\infty}^{\infty} x^2 |(g \star \eta)(x)|^2 dx \\
\leq & c_5 (M^2/h) h^{2m+1} N^{-2m+1} + c_6 (M^2 h) h^{2m-1} N^{-2m+1} \\
\leq & c_7 h^{2m} N^{-2m-1}.
\end{aligned} \tag{A.23}$$

Based on (A.22) and (A.23), we choose $h = h_*$ and $N = N_*$ as follows:

$$N_* := (A/h_*)^{1/\alpha}, \quad h_* := (A^{2m+1}/n^\alpha)^{\frac{1}{2m\alpha+2m+1}}$$

which leads to the announced result.   □

A.4. Proof of Theorem 3

Proof. (I). The error of estimator $\hat{f}_\tau^+(x_0)$ is

$$|\hat{f}_\tau^+(x_0) - f(x_0)| \leq |B_\tau(x_0; f)| + |\xi_\tau(x_0)|,$$

where $B_\tau(x_0; f)$ is the bias term, and $\xi_\tau(x_0)$ is the stochastic error given by (4.1). The bias term is expressed as follows (see the proof of Theorem 1):

$$
\begin{aligned}
B_\tau(x_0; f) :=& \mathrm{E}_f\big[\hat{f}_{h,N}^+(x_0)\big] - f(x_0) \\
=& \frac{1}{h} \int_{-\infty}^{\infty} K\Big(\frac{t - x_0}{h}\Big)[f(t) - f(x_0)]\mathrm{d}t \\
& + \sum_{j=1}^{m} \binom{m}{j}(-1)^j \int_{-1}^{1} K(y)f(yh + x_0 + 2\theta(N+1)j)\mathrm{d}y \\
=& \sum_{j=0}^{m} \binom{m}{j}(-1)^j \int_{-1}^{1} K(y)\big[f(yh + x_0 + 2\theta(N+1)j) \\
& - f(x_0 + 2\theta(N+1)j)\big]\mathrm{d}y + \sum_{j=1}^{m} \binom{m}{j}(-1)^j f(x_0 + 2\theta(N+1)j).
\end{aligned}
$$

Therefore by definitions of $\bar{B}_h(f)$ and $\bar{B}_N(x_0; f)$ [see (4.8), (4.9)] we have

$$|B_\tau(x_0; f)| \leq 2^m \bar{B}_h(f) + 2^m \bar{B}_N(x_0; f) \leq \bar{B}_\tau(x_0; f),$$

where $\bar{B}_\tau(x_0; f)$ is defined in (4.10).

(II). Now we demonstrate that

$$|B_{\tau \boxtimes \tau'}(x_0; f) - B_{\tau'}(x_0; f)| \leq \bar{B}_\tau(x_0; f), \quad \forall \tau, \tau' \in \mathcal{T}.$$

For this purpose denote

$$
\begin{aligned}
S_h(x) &:= \frac{1}{h} \int_{-\infty}^{\infty} K\Big(\frac{t - x}{h}\Big)[f(t) - f(x)]\mathrm{d}t \\
T_N(x) &:= \sum_{j=1}^{m} \binom{m}{j}(-1)^j f(x + 2\theta(N+1)j)
\end{aligned}
$$

and write

$$B_\tau(x_0; f) = S_h(x_0) + T_N(x_0) + \sum_{j=1}^{m} \binom{m}{j}(-1)^j S_h(x_0 + 2\theta(N+1)j). \quad \text{(A.24)}$$

In view of (A.24) for any pair $\tau = (h, N)$, $\tau' = (h', N')$ we have

$$B_{\tau \boxtimes \tau'}(x_0; f) - B_{\tau'}(x_0; f) = \big[S_{h \vee h'}(x_0) - S_{h'}(x_0)\big] + \big[T_{N \wedge N'}(x_0) - T_{N'}(x_0)\big]$$

$$+ \sum_{j=1}^{m} \binom{m}{j}(-1)^j \Big[S_{h \vee h'}(x_0 + 2\theta(N \wedge N' + 1)j) - S_{h'}(x_0 + 2\theta(N' + 1)j) \Big].$$
$$(A.25)$$

We consider the three terms on the right hand side of (A.25):

$$\sup_{h' \in \mathcal{H}} \big| S_{h \vee h'}(x_0) - S_{h'}(x_0) \big| = \sup_{h' \leq h} \big| S_{h \vee h'}(x_0) - S_{h'}(x_0) \big|$$
$$\leq \big| S_h(x_0) \big| + \sup_{h' \leq h} \big| S_{h'}(x_0) \big| \leq 2 \sup_{h' \leq h} \big| S_{h'}(x_0) \big|, \qquad (A.26)$$

and similarly

$$\sup_{N' \in \mathcal{N}} \big| T_{N \wedge N'}(x_0) - T_{N'}(x_0) \big| \leq 2 \sup_{N' \geq N} \sum_{j=1}^{m} \binom{m}{j} f(x_0 + 2\theta(N' + 1)j). \quad (A.27)$$

Furthermore

$$\sup_{h', N'} \big| S_{h \vee h'}(x_0 + 2\theta(N \wedge N' + 1)j) - S_{h'}(x_0 + 2\theta(N' + 1)j) \big|$$
$$\leq \sup_{h', N'} \big| S_{h \vee h'}(x_0 + 2\theta(N \wedge N' + 1)j) - S_{h'}(x_0 + 2\theta(N' \wedge N + 1)j) \big|$$
$$\quad + \sup_{h', N'} \big| S_{h'}(x_0 + 2\theta(N \wedge N' + 1)j) - S_{h'}(x_0 + 2\theta(N' + 1)j) \big|$$
$$\leq 2 \sup_{h' \leq h} \big\| S_{h'} \big\|_{\infty} + 2 \sup_{h' \in \mathcal{H}} \sup_{N' \geq N} \big| S_{h'}(x_0 + 2\theta(N' + 1)j) \big|$$
$$\leq 2 \sup_{h' \leq h} \big\| S_{h'} \big\|_{\infty} + 2 \sup_{h \in \mathcal{H}} \sup_{N' \geq N} \bigg| \int_{-1}^{1} K(y) f(yh + x_0 + 2\theta(N' + 1)j) \mathrm{d}y \bigg|$$
$$\quad + 2 \sup_{N' \geq N} f(x_0 + 2\theta(N' + 1)j)$$
$$\leq 2 \sup_{h' \leq h} \big\| S_{h'} \big\|_{\infty} + 2(1 + \|K\|_1) \sup_{|t| \leq \theta} \sup_{N' \geq N} f(t + x_0 + 2\theta(N' + 1)j). \quad (A.28)$$

Combining (A.26)–(A.28) with (A.25) we obtain

$$\sup_{\tau' \in \mathcal{T}} \big| B_{\tau \bowtie \tau'}(x_0; f) - B_{\tau'}(x_0; f) \big| \leq 2^{m+1} \sup_{h' \leq h} \big\| S_{h'} \big\|_{\infty}$$
$$+ 2^{m+1}(1 + \|K\|_1) \max_{1 \leq j \leq m} \sup_{|t| \leq \theta} \sup_{N' \geq N} f(t + x_0 + 2\theta(N' + 1)j)$$
$$= 2^{m+1} \bar{B}_h(f) + 2^{m+1}(1 + \|K\|_1) \bar{B}_N(x_0; f) \leq \bar{B}_\tau(x_0; f), \qquad (A.29)$$

where $\bar{B}_h(f)$, $\bar{B}_N(x_0; f)$ and $\bar{B}_\tau(x_0; f)$ are defined in (4.8), (4.9), and (4.10) respectively.

(III). Let $\hat{\tau} = (\hat{h}, \hat{N})$ be the parameter selected by the rule (4.6)–(4.7). For any $\tau \in \mathcal{T}$ we have the following triangle inequality

$$|\hat{f}_{\hat{\tau}}^+(x_0) - f(x_0)| \leq |\hat{f}_{\hat{\tau}}^+(x_0) - \hat{f}_{\hat{\tau} \bowtie \tau}^+(x_0)| + |\hat{f}_{\tau \bowtie \hat{\tau}}^+(x_0) - \hat{f}_{\tau}^+(x_0)| + |\hat{f}_{\tau}^+(x_0) - f(x_0)|.$$
$$(A.30)$$

Now we bound the terms on the right hand side separately.

We begin with the following simple observation: it follows from (4.6) that

$$
\hat{R}_\tau(x_0) - \hat{\Lambda}_\tau(\varkappa) - \sup_{\tau' \in \mathcal{T}} \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa)
$$
$$
= \sup_{\tau' \in \mathcal{T}} \left[\left| \hat{f}^+_{\tau \bowtie \tau'}(x_0) - \hat{f}^+_{\tau'}(x_0) \right| - \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa) - \hat{\Lambda}_{\tau'}(\varkappa) \right]_+
$$
$$
\leq \sup_{\tau' \in \mathcal{T}} \left| B_{\tau \bowtie \tau'}(x_0; f) - B_{\tau'}(x_0; f) \right|
$$
$$
+ \sup_{\tau' \in \mathcal{T}} \left[\left| \xi_{\tau \bowtie \tau'}(x_0) - \xi_{\tau'}(x_0) \right| - \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa) - \hat{\Lambda}_{\tau'}(\varkappa) \right]_+ .
$$

Hence by (A.29)

$$
\hat{R}_\tau(x_0) \leq \bar{B}_\tau(x_0; f) + 2\hat{\zeta}(x_0) + \hat{\Lambda}_\tau(\varkappa) + \sup_{\tau' \in \mathcal{T}} \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa), \tag{A.31}
$$

where

$$
\hat{\zeta}(x_0) := \sup_{\tau \in \mathcal{T}} \left[|\xi_\tau(x_0)| - \hat{\Lambda}_\tau(\varkappa) \right]_+ .
$$

Therefore for any $\tau, \tau' \in \mathcal{T}$

$$
\left| \hat{f}^+_{\tau \bowtie \tau'}(x_0) - \hat{f}^+_{\tau'}(x_0) \right| \leq \left| B_{\tau \bowtie \tau'}(x_0; f) - B_{\tau'}(x_0; f) \right| + \left| \xi_{\tau \bowtie \tau'}(x_0) - \xi_{\tau'}(x_0) \right|
$$
$$
\leq \bar{B}_\tau(x_0; f) + 2\hat{\zeta}(x_0) + \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa) + \hat{\Lambda}_{\tau'}(\varkappa) \leq \bar{B}_\tau(x_0; f) + 2\hat{\zeta}(x_0) + \hat{R}_{\tau'}(x_0),
$$

where the last inequality follows from the definition of $\hat{R}_\tau(x_0)$. This inequality together with (A.31) imply the following bound on the first term on the right hand side of (A.30):

$$
|\hat{f}^+_{\hat{\tau} \bowtie \tau}(x_0) - \hat{f}^+_{\hat{\tau}}(x_0)| \leq \bar{B}_\tau(x_0; f) + 2\hat{\zeta}(x_0) + \hat{R}_{\hat{\tau}}(x_0)
$$
$$
\leq \bar{B}_\tau(x_0; f) + 2\hat{\zeta}(x_0) + \hat{R}_\tau(x_0) \leq 2\bar{B}_\tau(x_0; f) + 4\hat{\zeta}(x_0) + \hat{\Lambda}_\tau(\varkappa) + \sup_{\tau'} \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa), \tag{A.32}
$$

where in the penultimate inequality we have used that $\hat{R}_{\hat{\tau}}(x_0) \leq \hat{R}_\tau(x_0)$ for any $\tau \in \mathcal{T}$.

We proceed with bounding the second term on the right hand side of (A.30): by definition of $\hat{R}_{\hat{\tau}}(x_0)$ we have

$$
|\hat{f}^+_{\tau \bowtie \hat{\tau}}(x_0) - \hat{f}^+_\tau(x_0)| \pm [\hat{\Lambda}_{\tau \bowtie \hat{\tau}}(\varkappa) + \hat{\Lambda}_\tau(\varkappa)]
$$
$$
\leq \hat{R}_{\hat{\tau}}(x_0) + \sup_{\tau' \in \mathcal{T}} \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa) + \hat{\Lambda}_\tau(\varkappa) \leq \hat{R}_\tau(x_0) + \sup_{\tau' \in \mathcal{T}} \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa) + \hat{\Lambda}_\tau(\varkappa)
$$
$$
\leq \bar{B}_\tau(x_0; f) + 2\hat{\zeta}(x_0) + 2 \sup_{\tau' \in \mathcal{T}} \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa) + 2\hat{\Lambda}_\tau(\varkappa). \tag{A.33}
$$

Finally

$$
|\hat{f}^+_\tau(x_0) - f(x_0)| \leq |B_\tau(x_0; f)| + |\xi_\tau(x_0)| \leq \bar{B}_\tau(x_0; f) + \Lambda_\tau(\varkappa) + \zeta(x_0), \tag{A.34}
$$

where we recall that

$$\zeta(x_0) := \sup_{\tau \in \mathcal{T}} \left[|\xi_\tau(x_0)| - \Lambda_\tau(\varkappa) \right]_+.$$

Combining (A.32), (A.33), (A.34) and (A.30) we obtain

$$\left| \hat{f}_{\hat{\tau}}^+(x_0) - f(x_0) \right| \le \inf_{\tau \in \mathcal{T}} \left\{ 4\bar{B}_\tau(x_0; f) + 3\hat{\Lambda}_\tau(\varkappa) + 3 \sup_{\tau' \in \mathcal{T}} \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa) + \Lambda_\tau(\varkappa) \right\}$$
$$+ 6\hat{\zeta}(x_0) + \zeta(x_0).$$

(IV). We complete the proof using Lemmas 2 and 1 in Appendix. Observing that $\hat{\Lambda}_\tau(\varkappa) = 7\tilde{\Lambda}_\tau(\varkappa)$ and applying the first inequality in (A.41) we have

$$\hat{\zeta}(x_0) \le \zeta(x_0) + \sup_{\tau \in \mathcal{T}} \left[\Lambda_\tau(\varkappa) - 7\tilde{\Lambda}_\tau(\varkappa) \right]_+ \le \zeta(x_0) + 2c\eta(x_0),$$

where $c = 2^{-m-2}\theta \|K^{(m)}\|_\infty^{-1}$ [cf. Lemma 2]. Then using the second inequality in (A.41) in order to bound $\hat{\Lambda}_\tau(\varkappa)$ and $\sup_{\tau' \in \mathcal{T}} \hat{\Lambda}_{\tau \bowtie \tau'}(\varkappa)$ in terms of $\Lambda_\tau(\varkappa)$ we obtain

$$\left| \hat{f}_{\hat{\tau}}^+(x_0) - f(x_0) \right| \le \inf_{\tau \in \mathcal{T}} \left\{ 4\bar{B}_\tau(x_0; f) + 127\Lambda_\tau(\varkappa) + 126 \sup_{\tau' \in \mathcal{T}} \Lambda_{\tau \bowtie \tau'}(\varkappa) \right\}$$
$$+ 7\zeta(x_0) + (42 + 12c)\eta(x_0) + \frac{42\varkappa}{n}.$$

By definition of the opeartion $\bowtie$ and by definition of σ_τ^2 and u_τ [see (4.2) and (4.3)] we have that $\sigma_{\tau \bowtie \tau'}^2 \le \sigma_\tau^2$ and $u_{\tau \bowtie \tau'} \le u_\tau$ for any $\tau, \tau' \in \mathcal{T}$; therefore $\sup_{\tau' \in \mathcal{T}} \Lambda_{\tau \bowtie \tau'}(\varkappa) \le \Lambda_\tau(\varkappa)$ for all $\tau \in \mathcal{T}$. We complete the proof by setting $\delta(x_0) = \zeta(x_0) + \eta(x_0)$ and using Lemma 1. $\qquad\square$

A.5. Proof of Corollary 2

Proof. Below $c_1, c_2, \ldots$ stand for positive constants independent of n, A and B. The proof goes along the following lines. We select values of h and N from $\mathcal{H} \times \mathcal{N}$ and apply the oracle inequality of Theorem 3.

The proof of Theorem 1 shows that if $f \in \mathcal{W}_{\alpha,q}(A, B)$ then

$$\bar{B}_h(f) \le c_1 A h^\alpha, \quad \bar{B}_N(x_0; f) \le c_2 B \theta^{-q} N^{-q}.$$

Furthermore, by (A.4)

$$\sigma_\tau^2 \le \frac{c_3 \theta^{2m-q} B \psi_N}{h^{2m+1}}, \quad \psi_N := \begin{cases} 1, & q > 2m - 1; \\ \log N, & q = 2m - 1; \\ N^{2m-q-1}, & q < 2m - 1. \end{cases}$$

In addition, with $\varkappa_* = \varkappa_0 \log n$ we have

$$\Lambda_\tau(\varkappa_*) \le c_4 \left(\frac{B^{1/2} \psi_N^{1/2}}{h^{m+1/2}} \sqrt{\frac{\varkappa_0 \log n}{n}} + \frac{N^{m-1}}{h^{m+1}} \frac{\varkappa_0 \log n}{n} \right). \tag{A.35}$$

First we note that for all $h_{\min} \le h \le h_{\max}$ and $N \le N_{\max}$ and all sufficiently large n

$$\Lambda_\tau(\varkappa_*) \le c_5 \frac{B^{1/2}\psi_N^{1/2}}{h^{m+1/2}} \sqrt{\frac{\varkappa_0 \log n}{n}}.$$

Indeed, this inequality follows from (A.35) because by the choice of $h_{\min}$ and $N_{\max}$ for large n one has

$$h_{\min}\Big(\frac{n}{\log n}\Big) = \Big(\frac{n}{\log n}\Big)^{2m/(2m+1)} \ge N_{\max}^{2m-2} = \Big(\frac{n}{\log n}\Big)^{(2m-2)/(2m)}.$$

Thus, using (4.11) we have

$$|\hat{f}_*(x_0) - f(x_0)| \le c_6 \inf_{(h,N)\in\mathcal{H}\times\mathcal{N}} \left\{ Ah^\alpha + \frac{B}{\theta^q N^q} + \frac{B^{1/2}\psi_N^{1/2}}{h^{m+1/2}} \sqrt{\frac{\log n}{n}} \right\}$$
$$+ c_7\left(\delta(x_0) + \frac{\varkappa_0 \log n}{n} \right).$$

Now we set h_* and N_* to be defined by formulas (A.8), (A.7) and (A.6) with n replaced by $n/\log n$. Note that these values of h and N balance the bias and stochastic error bounds on the right hand side of the previous display formula [for details see the proof of Theorem 1]. We need to verify that h_* and N_* satisfy $h_* \ge h_{\min}$ and $N_* \le N_{\max}$ for large n. The first inequality is evident because $1/(2\alpha + 2m + 1) \ge 1/(2m + 1)$ for all $\alpha > 0$. To check the inequality $N_* \le N_{\max}$ we note that $N_* = O\big((n/\log n)^{\frac{\alpha}{q(2m+2\alpha+1+r)}}\big)$ in the case $1 \le q < 2m - 1$ and

$$\frac{\alpha}{q(2m + 2\alpha + 1 + r)} = \frac{\alpha}{\alpha(2m - 1 + q) + q(2m + 1)} \le \frac{1}{2m}$$

for all $\alpha > 0$. If $q > 2m - 1$ then $N_* = O\big((n/\log n)^{\frac{\alpha}{q(2m+2\alpha+1)}}\big)$, and

$$\frac{\alpha}{q(2m + 2\alpha + 1)} \le \frac{\alpha}{(2m - 1)(2m + 2\alpha + 1)} \le \frac{1}{4m - 2}, \quad \forall \alpha > 0.$$

Thus, we always have $N_* \le N_{\max}$ for large n. The inequalities $h_* \ge h_{\min}$ and $N_* \le N_{\max}$ imply that sets $\mathcal{H}$ and $\mathcal{N}$ contain elements that bound h_* and N_* from below and from above within constant factors. This yields

$$|\hat{f}_*(x_0) - f(x_0)| \le c_8\varphi(n/\log n) + c_7\left(\delta(x_0) + \frac{\varkappa_0 \log n}{n} \right),$$

where function $\varphi(\cdot)$ is defined in (3.5).

To complete the proof we note that $M_h = O(\log_2 n)$, $M_N = O(n^{1/(2m)})$, and

$$\bar{\Lambda}(\varkappa_*) \le c_9 \frac{N_{\max}^{m-1}}{h_{\min}^{m+1/2}} \sqrt{\frac{\log n}{n}} \left(1 + \frac{N_{\max}^{m-1}}{h_{\min}^{m+1}}\right) \le c_{10}\Big(\frac{n}{\log n}\Big)^{3/2},$$

so that if $\varkappa_0 \ge 5$ then in view of (4.12) for large n

$$\mathrm{E}_f[\delta(x_0)]^2 \le c_{11}(\log_2 n) n^{1/2m}\Big(\frac{n}{\log n}\Big)^3 e^{-\varkappa_0 \log n} \le c_{12} n^{-1}.$$

This completes the proof. $\qquad\square$

A.6. Auxiliary results

Denote

$$L_\tau^+(y) := \frac{(2\theta)^m}{h^{m+1}} \sum_{j=0}^N C_{j,m} K^{(m)} \left(\frac{y - x_0 - \theta(2j+m)}{h} \right).$$

Then

$$\mathrm{var}_f[\hat{f}_\tau^+(x_0)] = \mathrm{E}_f[\xi_\tau(x_0)]^2, \quad \xi_\tau(x_0) := \frac{1}{n} \sum_{i=1}^n \left[L_\tau^+(Y_i) - \mathrm{E}_f L_\tau^+(Y_i) \right].$$

Let

$$\zeta(x_0) := \sup_{\tau \in \mathcal{T}} \left[|\xi_\tau(x_0)| - \Lambda_\tau(\varkappa) \right]_+; \tag{A.36}$$

$$\eta(x_0) := \sup_{\tau \in \mathcal{T}} \left[|\hat{\sigma}_\tau^2 - \sigma_\tau^2| - u_\tau \Lambda_\tau(\varkappa) \right]_+. \tag{A.37}$$

Lemma 1. *For any $p \geq 1$ and $\varkappa > 0$ one has*

$$\mathrm{E}_f[\zeta(x_0)]^p \leq 2\Gamma(p+1) M_h M_N \left[\Lambda_\tau(\varkappa) \right]^p \varkappa^{-p} e^{-\varkappa};$$
$$\mathrm{E}_f[\eta(x_0)]^p \leq 2\Gamma(p+1) M_h M_N \left[u_\tau \Lambda_\tau(\varkappa) \right]^p \varkappa^{-p} e^{-\varkappa}.$$

Proof. (i). Observe that $|L_\tau^+(Y_j)| \leq u_\tau/2$, where u_τ is defined in (4.3); hence $|\xi_\tau| \leq u_\tau$. In addition, it follows from (A.1) that $\mathrm{var}_f \left[L_\tau^+(Y_1) \right]$ is bounded above by:

$$\sigma_\tau^2 := \frac{(2\theta)^{2m}}{h^{2m+2}} \sum_{j=0}^N C_{j,m}^2 \int_{-\infty}^\infty \left| K^{(m)} \left(\frac{y - x_0 - \theta(2j+m)}{h} \right) \right|^2 f_Y(y) \mathrm{d}y.$$

By Bernstein's inequality for any $z > 0$

$$\mathrm{P}_f \left[|\xi_\tau(x_0)| \geq z \right] \leq 2 \exp \left\{ - \frac{nz^2}{2\sigma_\tau^2 + \frac{2}{3} u_\tau z} \right\}.$$

Therefore for $\Lambda_\tau(\varkappa)$ defined in (4.4) we obtain

$$\mathrm{P}_f \left[|\xi_\tau(x_0)| \geq \Lambda_\tau(\varkappa) \right] \leq 2 \exp \left\{ - \frac{\left(\sigma_\tau \sqrt{\frac{2\varkappa}{n}} + \frac{2}{3} u_\tau \varkappa n^{-1} \right)^2}{2\sigma_\tau^2/n + \frac{2u_\tau}{3n} \left(\sigma_\tau \sqrt{\frac{2\varkappa}{n}} + \frac{2\varkappa u_\tau}{3n} \right)} \right\} \leq 2e^{-\varkappa}, \tag{A.38}$$

where we have used the following elementary inequality: for any $a > 0, b > 0$ and $\varkappa > 0$

$$\frac{(\sqrt{\varkappa}a + \varkappa b)^2}{a^2 + b(\sqrt{\varkappa}a + \varkappa b)} \geq \varkappa. \tag{A.39}$$

Therefore, for any $p \geq 1$

$$
\begin{aligned}
\mathrm{E}_f\big[|\xi_\tau(x_0)| - \Lambda_\tau(\varkappa)\big]_+^p &= p \int_0^\infty t^{p-1} \mathrm{P}_f\big[|\xi_\tau(x_0)| \geq \Lambda_\tau(\varkappa) + t\big] \mathrm{d}t \\
&\leq p[\Lambda_\tau(\varkappa)]^p \int_0^\infty y^{p-1} \mathrm{P}_f\big[|\xi_\tau(x_0)| \geq \Lambda_\tau(\varkappa(1+y))\big] \mathrm{d}y \\
&\leq 2p[\Lambda_\tau(\varkappa)]^p \int_0^\infty y^{p-1} e^{-\varkappa(1+y)} \mathrm{d}y = 2\Gamma(p+1)\big[\Lambda_\tau(\varkappa)\big]^p \varkappa^{-p} e^{-\varkappa},
\end{aligned}
$$

$$(A.40)$$

where the second line follows from the change of variables and the fact that $\Lambda_\tau(a\varkappa) \leq a\Lambda_\tau(\varkappa)$ for $a \geq 1$; and the third line is a consequence of (A.38).

(ii). Let $\hat{\sigma}_\tau^2$ be the empirical estimator for σ_τ^2 based on the sample $Y_1, \ldots, Y_n$:

$$
\hat{\sigma}_\tau^2 := \frac{(2\theta)^{2m}}{nh^{2m+2}} \sum_{i=1}^n \sum_{j=0}^N C_{j,m}^2 \left| K^{(m)}\left(\frac{Y_i - x_0 - \theta(2j+m)}{h}\right) \right|^2.
$$

Then

$$
\hat{\sigma}_\tau^2 - \sigma_\tau^2 = \frac{1}{n} \sum_{i=1}^n \Big(\psi_\tau(Y_i) - \mathrm{E}_f[\psi_\tau(Y_i)]\Big),
$$

where we put

$$
\psi_\tau(y) := \frac{(2\theta)^{2m}}{h^{2m+2}} \sum_{j=0}^N C_{j,m}^2 \left| K^{(m)}\left(\frac{y - x_0 - \theta(2j+m)}{h}\right) \right|^2.
$$

It is evident that

$$
|\psi_\tau(y)| \leq \frac{(2\theta)^{2m}}{h^{2m+2}} C_{N,m}^2 \|K^{(m)}\|_\infty^2 = \tfrac{1}{4} u_\tau^2, \quad \forall y;
$$

hence $\big|\psi_\tau(Y_i) - \mathrm{E}_f[\psi_\tau(Y_i)]\big| \leq u_\tau^2/4$, and

$$
\mathrm{var}_f[\psi_\tau(Y_i)] \leq \mathrm{E}_f\big[\psi_\tau^2(Y_i)\big] \leq \tfrac{1}{4}\sigma_\tau^2 u_\tau^2.
$$

Therefore by the Bernstein inequality for any $z \geq 0$

$$
\mathrm{P}_f\Big[|\hat{\sigma}_\tau^2 - \sigma_\tau^2| \geq z\Big] \leq 2\exp\bigg\{-\frac{nz^2}{\tfrac{1}{2}\sigma_\tau^2 u_\tau^2 + \tfrac{1}{6}u_\tau^2 z}\bigg\}.
$$

This inequality together with (A.39) implies that

$$
\mathrm{P}_f\Big[|\hat{\sigma}_\tau^2 - \sigma_\tau^2| \geq u_\tau \Lambda_\tau(\varkappa)\Big] \leq \mathrm{P}_f\bigg[|\hat{\sigma}_\tau^2 - \sigma_\tau^2| \geq u_\tau\Big(\sigma_\tau\sqrt{\frac{\varkappa}{2n}} + \frac{u_\tau\varkappa}{6n}\Big)\bigg] \leq 2e^{-\varkappa}.
$$

Similarly to the derivation in (A.40) we have for any $p \geq 1$

$$
\mathrm{E}_f\big[|\hat{\sigma}_\tau^2 - \sigma^2| - u_\tau\Lambda_\tau(\varkappa)\big]_+^p = p \int_0^\infty t^{p-1} \mathrm{P}_f\big[|\hat{\sigma}_\tau^2 - \sigma_\tau^2| \geq u_\tau\Lambda_\tau(\varkappa) + t\big] \mathrm{d}t
$$

$$\leq p\big[u_\tau\Lambda_\tau(\varkappa)\big]^p \int_0^\infty y^{p-1} \mathrm{P}_f\big[|\hat\sigma_\tau^2 - \sigma_\tau^2| \geq u_\tau\Lambda_\tau(\varkappa(1+y))\big]\mathrm{d}y$$

$$\leq 2p[u_\tau\Lambda_\tau(\varkappa)]^p \int_0^\infty y^{p-1}e^{-\varkappa(1+y)}\mathrm{d}y = 2\Gamma(p+1)\big[u_\tau\Lambda_\tau(\varkappa)\big]^p\varkappa^{-p}e^{-\varkappa}.$$

This completes the proof. $\square$

Denote

$$\tilde\Lambda_\tau(\varkappa) := \hat\sigma_\tau\sqrt{\frac{2\varkappa}{n}} + \frac{2u_\tau\varkappa}{3n}$$

and observe that $\tilde\Lambda_\tau(\varkappa) = \frac{1}{7}\hat\Lambda_\tau(\varkappa)$, where $\hat\Lambda_\tau(\varkappa)$ is defined in (4.5).

Lemma 2. *For any $\tau \in \mathcal{T}$ one has*

$$\big[\Lambda_\tau(\varkappa) - 7\tilde\Lambda_\tau(\varkappa)\big]_+ \leq 2c\eta(x_0), \quad \big[\tilde\Lambda_\tau(\varkappa) - 6\Lambda_\tau(\varkappa)\big]_+ \leq \eta(x_0) + \frac{\varkappa}{n}, \quad \text{(A.41)}$$

where $\eta(x_0)$ is defined in (A.37) and $c := 2^{-m-2}\theta\|K^{(m)}\|_\infty^{-1}$.

Proof. We have $\tilde\Lambda_\tau(\varkappa) - \Lambda_\tau(\varkappa) = (\hat\sigma_\tau - \sigma_\tau)\sqrt{2\varkappa/n}$. Define

$$\mathcal{T}_1 := \Big\{\tau \in \mathcal{T} : \sigma_\tau\sqrt{\frac{2\varkappa}{n}} \geq \frac{4u_\tau\varkappa}{n}\Big\}.$$

If $\tau \in \mathcal{T}_1$ then $\sigma_\tau \geq 2\sqrt{2}u_\tau(\varkappa/n)^{1/2}$ and

$$|\hat\sigma_\tau - \sigma_\tau| = \frac{|\hat\sigma_\tau^2 - \sigma_\tau^2|}{\hat\sigma_\tau + \sigma_\tau} \leq \frac{1}{\sigma_\tau}|\hat\sigma_\tau^2 - \sigma_\tau^2| \leq \frac{1}{2u_\tau}\sqrt{\frac{n}{2\varkappa}}\big[\eta(x_0) + u_\tau\Lambda_\tau(\varkappa)\big];$$

hence for any $\tau \in \mathcal{T}_1$

$$|\tilde\Lambda_\tau(\varkappa) - \Lambda_\tau(\varkappa)| \leq \tfrac{1}{2}\Lambda_\tau(\varkappa) + \frac{\eta(x_0)}{2u_\tau} \leq \tfrac{1}{2}\Lambda_\tau(\varkappa) + c\eta(x_0) \qquad \text{(A.42)}$$

where we have used that $u_\tau \geq 2^{m+1}\theta^{-1}\|K^{(m)}\|_\infty$ for all $\tau \in \mathcal{T}$, and denoted for brevity $c := 2^{-m-2}\theta\|K^{(m)}\|_\infty^{-1}$. Thus (A.42) implies that

$$\big[\tilde\Lambda_\tau(\varkappa) - \tfrac{3}{2}\Lambda_\tau(\varkappa)\big]_+ \leq c\eta(x_0) \quad \text{and} \quad \big[\Lambda_\tau(\varkappa) - 2\tilde\Lambda_\tau(\varkappa)\big]_+ \leq 2c\eta(x_0), \quad \forall\tau \in \mathcal{T}_1.$$
$$\text{(A.43)}$$

Now assume that $\tau \in \mathcal{T}_2 := \mathcal{T} \setminus \mathcal{T}_1$; for such τ, $\Lambda_\tau(\varkappa) \leq \frac{14}{3}u_\tau\varkappa/n$. Note also that by definition $\tilde\Lambda_\tau(\varkappa) \geq \frac{2}{3}u_\tau\varkappa/n$; therefore

$$[\Lambda_\tau(\varkappa) - 7\tilde\Lambda_\tau(\varkappa)]_+ = 0, \quad \forall\tau \in \mathcal{T}_2. \qquad \text{(A.44)}$$

Furthermore, we bound $|\hat\sigma_\tau - \sigma_\tau|$ as follows:

$$|\hat\sigma_\tau - \sigma_\tau| \leq |\hat\sigma_\tau^2 - \sigma_\tau^2|^{1/2} \leq \sqrt{\eta(x_0)} + \sqrt{u_\tau\Lambda_\tau(\varkappa)} \leq \sqrt{\eta(x_0)} + \sqrt{5}u_\tau\sqrt{\frac{\varkappa}{n}}.$$

Therefore for any $\tau \in \mathcal{T}_2$

$$\left|\tilde{\Lambda}_\tau(\varkappa) - \Lambda_\tau(\varkappa)\right| \leq \sqrt{\frac{2\varkappa}{n}\eta(x_0)} + \sqrt{10}\,\frac{u_\tau \varkappa}{n} \leq \frac{\varkappa}{n} + \eta(x_0) + 5\Lambda_\tau(\varkappa),$$

where the last bound follows from the elementary inequality $\sqrt{2ab} \leq \sqrt{a^2 + b^2} \leq a + b$ for $a, b \geq 0$. This implies that

$$\left[\tilde{\Lambda}_\tau(\varkappa) - 6\Lambda_\tau(\varkappa)\right]_+ \leq \frac{\varkappa}{n} + \eta(x_0), \quad \forall \tau \in \mathcal{T}_2. \tag{A.45}$$

Combining (A.43), (A.44) and (A.45) we complete the proof. $\qquad\square$

Acknowledgments

We thank two anonymous reviewers and the Associate Editor for useful remarks.

References

[1] AUBIN, J.-P. (2000). *Applied functional analysis*, 2 ed. John Wiley & Sons. MR0549483
[2] BELOMESTNY, D. and GOLDENSHLUGER, A. (2021). Density deconvolution under general assumptions on the distribution of measurement errors. *The Annals of Statistics* **49** 615–649. MR4255101
[3] BERTERO, M. and BOCCACCI, P. (1998). *Introduction to inverse problems in imaging*. Institute of Physics Publishing, Bristol. MR1640759
[4] BUTUCEA, C. and TSYBAKOV, A. B. (2007a). Sharp optimality in density deconvolution with dominating bias. I. *Teoriya Veroyatnosteĭ i ee Primeneniya* **52** 111–128. MR2354572
[5] BUTUCEA, C. and TSYBAKOV, A. B. (2007b). Sharp optimality in density deconvolution with dominating bias. II. *Teoriya Veroyatnosteĭ i ee Primeneniya* **52** 336–349. MR2742504
[6] CARROLL, R. J. and HALL, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association* **83** 1184–1186. MR0997599
[7] CAVALIER, L. (2011). Inverse problems in statistics. In *Inverse problems and high-dimensional estimation. Lect. Notes Stat. Proc.* **203** 3–96. Springer, Heidelberg. MR2868199
[8] DELAIGLE, A. and MEISTER, A. (2011). Nonparametric function estimation under Fourier-oscillating noise. *Statistica Sinica* **21** 1065–1092. MR2827515
[9] DEVROYE, L. (1989). Consistent deconvolution in density estimation. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* **17** 235–239. MR1033106
[10] FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics* **19** 1257–1272. MR1126324

[11] FAN, J. and TRUONG, Y. K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21** 1900–1925. MR1245773

[12] FEUERVERGER, A., KIM, P. T. and SUN, J. (2008). On optimal uniform deconvolution. *Journal of Statistical Theory and Practice* **2** 433–451. MR2528791

[13] GOLBERG, M. A. (1979). A method of adjoints for solving some ill-posed equations of the first kind. *Applied Mathematics and Computation* **5** 123–129. MR0522852

[14] GOLDENSHLUGER, A. and LEPSKI, O. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics* **39** 1608–1632. MR2850214

[15] GOLDENSHLUGER, A. and LEPSKI, O. (2014). On adaptive minimax density estimation on $\mathbb{R}^d$. *Probability Theory and Related Fields* **159** 479–543. MR3230001

[16] GROENEBOOM, P. and JONGBLOED, G. (2003). Density estimation in the uniform deconvolution model. *Statistica Neerlandica* **57** 136–157. MR2035863

[17] HALL, P. and KOCH, I. (1990). On continuous image models and image analysis in the presence of correlated noise. *Adv. in Appl. Probab.* **22** 332–349. MR1053234

[18] HALL, P. and MEISTER, A. (2007). A ridge-parameter approach to deconvolution. *The Annals of Statistics* **35** 1535–1558. MR2351096

[19] JOHNSTONE, I. M. and RAIMONDO, M. (2004). Periodic boxcar deconvolution and Diophantine approximation. *The Annals of Statistics* **32** 1781–1804. MR2102493

[20] JOHNSTONE, I. M., KERKYACHARIAN, G., PICARD, D. and RAIMONDO, M. (2004). Wavelet deconvolution in a periodic setting. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 547–573. MR2088290

[21] LEPSKI, O. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications* **35** 454–466. MR1091202

[22] LEPSKI, O. (2015). Adaptive estimation over anisotropic functional classes via oracle approach. *The Annals of Statistics* **43** 1178–1242. MR3346701

[23] LOUNICI, K. and NICKL, R. (2011). Global uniform risk bounds for wavelet deconvolution estimators. *The Annals of Statistics* **39** 201–231. MR2797844

[24] MEISTER, A. (2007). Deconvolution from Fourier-oscillating error densities under decay and smoothness restrictions. *Inverse Problems* **24** 015003. MR2384762

[25] MEISTER, A. (2009). *Deconvolution problems in nonparametric statistics. Lecture Notes in Statistics* **193**. Springer-Verlag, Berlin. MR2768576

[26] STANLEY, R. P. (2011). *Enumerative combinatorics*, second ed. *Cambridge studies in advanced mathematics* **1**. Cambridge University Press. MR2868112

[27] STEFANSKI, L. A. and CARROLL, R. J. (1990). Deconvolving kernel density estimators. *Statistics* **21** 169–184. MR1054861

[28] WIDDER, D. V. (1946). *The Laplace transform.* Princeton University Press. MR0005923

[29] ZHANG, C.-H. (1990). Fourier methods for estimating mixing densities and distributions. *The Annals of Statistics* **18** 806–831. MR1056338