Sankhyā: The Indian Journal of Statistics
Special issue on Wavelets
2001, Volume 63, Series B, Pt. 2, pp. 149–180

NUMERICAL METHODS FOR ASYMPTOTICALLY MINIMAX NON-PARAMETRIC FUNCTION ESTIMATION WITH POSITIVITY CONSTRAINTS I

By LUBOMIR DECHEVSKY
Université de Montréal, Canada
BRENDA MACGIBBON
Université du Québec à Montréal, Canada
and
SPIRIDON PENEV
University of New South Wales, Australia

SUMMARY. One important challenge in nonparametric density and regression-function estimation is spatially inhomogeneous smoothness. This is often modelled by Besov-type smoothness constraints. With this type of constraint, Donoho and Johnstone (1992), Delyon and Juditsky (1993) studied asymptotic-minimax optimal wavelet estimators with thresholding, while Lepski, Mammen and Spokoiny (1995) proposed a variable-bandwidth selection for kernel estimators that also achieved the asymptotic-minimax rates. However, a second challenge in many applications of nonparametric curve estimation is that the function must be nonnegative or order-constrained. Dechevsky and MacGibbon (1999) constructed wavelet- and kernel-based estimators under positivity constraints that satisfied these constraints and also achieved asymptotic-minimax rates over the appropriate smoothness classes. Here we show how to replace the integral in their definition by a quadrature formula in order to numerically construct the estimators, so that the new "quadrature" estimators enjoy the positivity- and smoothness-preserving properties of the ones in Dechevsky and MacGibbon (1999), and are also asymptotic-minimax optimal.

Paper received July 2000.
AMS (2000) *subject classification.* Primary 62G07, 62G08, 65C60; secondary 06A05, 06A06, 41A29, 46B42, 65D30.
Key words and phrases. Density estimation, nonparametric regression, wavelet estimator, kernel estimator, order constraint, asymptotic minimax rate, numerical integration.

1. Introduction

One important challenge in nonparametric density and regression-function estimation is spatially inhomogeneous smoothness. This is often modelled by Besov-type smoothness constraints. For this problem Donoho and Johnstone Donoho and Johnstone (1998), Delyon and Juditsky Delyon and Juditsky (1996) studied wavelet estimators with thresholding, while Lepski, Mammen and Spokoiny Lepski, Mammen and Spokoiny (1977) proposed a variable-bandwidth selection for kernel estimators. All these wavelet and kernel estimators achieve asymptotic-minimax optimal rates over Besov classes. However, a second challenge in many applications of nonparametric curve estimation is that the function must also be nonnegative, or, more generally, it must be bounded from below and/or above by known functions v and/or w. Dechevsky and MacGibbon (1999) proposed a general method which can be extended to achieve the following: for an unknown density or regression function f whose graph is *a priori* known to be bounded by those of v and/or w, known functions with certain regularity, given an estimator $\hat{f}$ whose graph need not necessarily obey these bound(s), construct a *sufficiently smooth* estimator $\hat{f}^+$ whose graph obeys the bound(s) and achieves *the same* asymptotic rates as $\hat{f}$, with respect to *the same* metric. The definition of the new estimator involves a convolution with a smooth nonnegative symmetric kernel with bandwidth ε, chosen in an optimal way to preserve the rates and so that the multiplicative constant for these rates increases only by a constant factor, greater than, but close to, 1. In some cases, it is possible to compute the convolution integral exactly, in closed form, but, in general, numerical integration by a quadrature rule must be done. (This essentially replaces the convolution by a kernel-type estimator.)

The aim of the present paper is to study the properties of the resulting new "quadrature" kernel estimator of f and to compare it to the properties of the underlying $\hat{f}$ and $\hat{f}^+$. In particular: (1) the quadrature formula must be selected so that the new estimator is *as smooth as* $\hat{f}^+$; (2) the quadrature formula depends on *additional small parameters* which determine the precision with which the integral is approximated; these additional parameters, together with ε, must be chosen in an optimal way so as to preserve *the same* asymptotic rates; (3) the choice of these parameters can be made optimal only with respect to certain values of the metric indices p, q and the smoothness index s of the Besov space B^s_{pq} to which f is assumed to belong. In practice, only a "confidence" range of approximate values p, q, s

is known; therefore, optimal (smaller) values of the quadrature parameters and of ε have to be determined in such a way that the "quadrature" estimator attains *the same* rates, with a prescribed constant factor for these rates, as $\hat{f}$ and $\hat{f}^+$ *simultaneously for the entire* "confidence" range of p, q and s.

Here we study only the case when v and w are constants. The general case of spatially variable lower and upper bounds $v = v(x)$, $w = w(x)$, will be considered in future. The methods can also be used to develop an analogous approach to *local pointwise* order-constrained statistical estimation; that is to develop estimators which are not only (locally and globally) asymptotically rate-optimal, but also locally highly spatially adaptive for small and moderate samples. (More details about this will also be given in future papers.)

Although in Dechevsky and MacGibbon (1999) the multivariate case $f : \mathbb{R}^d \to \mathbb{R}$ was considered, here, for technical simplicity, we only address the univariate case $d = 1$. For the multivariate case we refer to Dechevsky, MacGibbon and Penev (2000, Sections 1, 3, 8 and Appendix).

The organization of this article is as follows. All preliminary notation and definitions not given in the main text can be found in the appendix. In section 2 we consider a general class of variable-kernel estimators which includes both asymptotically minimax-optimal wavelet and kernel estimators as well as their respective $\hat{f}^+$ from Dechevsky and MacGibbon (1999) and their new quadrature versions. Section 3 contains the main results. In Section 4 some applications are discussed in the context of Delyon and Juditsky (1996) and Lepski, Mammen and Spokoiny (1977). Section 5 addresses the important practical question of selecting an appropriate concrete quadrature rule. The graphical results for some simulated examples are displayed in Section 6. The proofs are given in Section 7.

2. Generalized Kernel Estimators

Our aim in this section is to propose a unified systematic approach to kernel and wavelet methods for density and regression-function estimation in which new positivity-preserving methods can also be included in a natural way. We shall show that the attempts to unite the notions of a kernel and a wavelet estimator, and in particular, their asymptotic-minimax versions, lead to a natural generalization which we call a "generalized kernel estimator" (GKE) – see Definition 1 below. It will then be shown that the convolution-based constrained estimator $\widehat{f}_{+,\varepsilon}$ proposed in Dechevsky and MacGibbon (1999) (its definition is given here in Section 3), as well as the version of $\widehat{f}_{+,\varepsilon}$

based on quadrature approximation of the convolution integral, are both GKE.

Henceforth, we assume that $\widehat{h} = \widehat{h}(x), \quad x \in \mathbb{R}^d$, is a measurable function of the random vector $(X_1, \ldots, X_N)$ of N (not necessarily uncorrelated) random variables with cumulative distribution function (c.d.f.) $F^N : \mathbb{R}^{Nd} \to [0,1]$. In applications $(X_1, \ldots, X_N)$ are most often independent and identically distributed (i.i.d.) (see, e.g., Delyon and Juditsky, 1996 and Lepski, Mammen and Spokoiny, 1977). We will be interested in the two cases $\widehat{h} = \widehat{f}$ or $h = \widehat{g}$, where $\widehat{f}$ and $\widehat{g}$ are estimators of a density f and a regression function g, respectively.

DEFINITION **1.** *(a) Density estimation. Assume that the c.d.f.* $F : \mathbb{R}^d \to [0,1]$ *is such that* F *is absolutely continuous relative to the Lebesgue measure on* $\mathbb{R}^d$*, with Radon-Nikodym derivative* f*.*

A generalized kernel estimator (GKE) of a density f *is:*

$$\widehat{f}(x) := \frac{1}{N} \sum_{\nu=1}^{N} K(x, X_\nu; X_1, \ldots, X_N), \qquad x \in \mathbb{R}^{\mathrm{d}}, \tag{2.1}$$

which is an empirical version of

$$f_1(x) = \int\limits_{\mathbb{R}^{\mathrm{d}}} K(x, t; X_1, \ldots, X_N)\ f(t)\ dt; \tag{2.2}$$

(b) Regression-function estimation. In this case the c.d.f. F *need not have a density. The unknown regression function* $g : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ *is assumed to be estimated on the basis of* N *(not necessarily uncorrelated) noisy observations*

$$Y_i = g(X_i) + \Delta_i, \qquad i = 1, \ldots, N,$$

where Δ_i *are the error terms. A generalized kernel estimator (GKE) of the regression function* g *is*

$$\widehat{g}(x) := \frac{1}{N} \sum_{\nu=1}^{N} Y_\nu . K(x, X_\nu; X_1, \ldots, X_N), \tag{2.3}$$

which is an empirical version of

$$g_1(x) = \int\limits_{\mathbb{R}^{\mathrm{d}}} K(x, t; X_1, \ldots, X_N)\ g(t)\ dF(t). \tag{2.4}$$

In (2.2,4), for any given sample $(X_1, \ldots, X_N)$

$$K(x, t; X_1, \ldots, X_N) \in L_{1,loc}(\mathbb{R}^{2\mathrm{d}}) \text{ as a function of } (\mathrm{x}, \mathrm{t}). \tag{2.5}$$

In other words, (2.1,3) are kernel-type estimators whose kernel $K(x,t)$ is variable in t and sample-dependent.

EXAMPLE **1.** *(Asymptotically minimax estimators)*

(A) Variable-bandwidth kernel estimator with kernel Ψ

(a) density: $\widehat{f}$ *is defined by (2.1);*

(b) regression (see Lepski, Mammen and Spokoiny, 1977): $\widehat{g}$ *is defined by (2.3);*

$$\begin{aligned} K(x,t;X_1,\dots,X_N) &= \frac{1}{\widehat{\chi}(X_1,\dots,X_N)}.\Psi\left[\frac{x-t}{\widehat{\chi}(X_1,\dots,X_N)}\right] \\ &= \Psi_1(x-t;X_1,\dots,X_N), \end{aligned} \tag{2.6}$$

where $\widehat{\chi}$ *is an optimally selected bandwidth (in Lepski, Mammen and Spokoiny, 1977, it is denoted by* $\widehat{h}$ *– not to be confused with* $\widehat{h}$ *in our notation!).*

(B) Thresholded wavelet estimator (see Delyon and Juditsky, 1996)

(a) density: $\widehat{f}$ *is defined via (2.1),*

$$\begin{aligned} K(x,t;X_1,\dots,X_N) &= \sum_{k\in\mathbb{Z}^d} \varphi_{j_0k}(x)\ \varphi_{jk}(t) \\ &\quad + \sum_{j=j_0}^{j_1}\sum_{k\in\mathbb{Z}^d}\sum_{i=1}^{2^d-1} \mu_{jk}^{(i)}(X_1,\dots,X_N)\ \psi_{jk}^{(i)}(x)\ \psi_{jk}^{(i)}(t), \end{aligned} \tag{2.7}$$

$$\mu_{jk}^{(i)}(X_1,\dots,X_N) = \begin{cases} \dfrac{\delta\left[\frac{1}{N}\sum\limits_{\nu=1}^{N}\psi_{jk}^{(i)}(X_\nu);\lambda_j\right]}{\frac{1}{N}\sum\limits_{\nu=1}^{N}\psi_{jk}^{(i)}(X_\nu)}, & \text{if } \frac{1}{N}\sum\limits_{\nu=1}^{N}\psi_{jk}^{(i)}(X_\nu)\neq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{2.8}$$

where $\delta(\alpha;\lambda_j)$ *is the threshold rule with threshold* λ_j*, as considered in Delyon and Juditsky (1996);*

(b) regression: $\widehat{g}$ *is defined by (2.3);*

$K(x,t;X_1,\dots,X_N)$ *is defined by (2.7), but with modified values of* $\mu_{jk}^{(i)}$*:*

$$\begin{aligned} &\mu_{jk}^{(i)}(X_1,\dots,X_N) \\ &= \begin{cases} \dfrac{\delta\left[\frac{1}{N}\sum\limits_{\nu=1}^{N} Y_\nu.\gamma_{jk,\nu}^{(i)}\psi_{jk}^{(i)}(X_\nu);\lambda_j\right]}{\frac{1}{N}\sum\limits_{\nu=1}^{N} Y_\nu.\gamma_{jk,\nu}^{(i)}\psi_{jk}^{(i)}(X_\nu)}, & \text{if } \dfrac{1}{N}\sum\limits_{\nu=1}^{N} Y_\nu\ .\ \gamma_{jk,\nu}^{(i)}\ .\ \psi_{jk}^{(i)}(X_\nu)\neq 0, \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{2.9}$$

$$\gamma_{jk,\nu}^{(i)} = \begin{cases} 1, & \text{if } |Y_\nu . \psi_{jk}^{(i)}(X_\nu)| \leqq M \\ 0, & \text{otherwise,} \end{cases} \tag{2.10}$$

$M = \sqrt{\dfrac{C_{tr}.N}{\ln N}}$ *(with C_{tr} = constant) is the truncation value, as considered in Delyon and Juditsky (1996).*

REMARK **1.** From the point of view of Definition 1, the only essential difference between homologous kernel and wavelet estimators is that kernel estimators $K(x,t)$ depend on $x-t$ only, which implies that kernel estimators commute with translations over the whole of $\mathbb{R}^d$, whereas the wavelet estimators only commute with shifts over the discrete group $2^{-j_1}\mathbb{Z}^d$.

The following is an important property of the convolution-based constrained estimator $\widehat{h}_{+,\varepsilon}$, defined by (3.1) in the next section, in both the case of density ($h = f$) and the case of regression ($h = g$) estimation.

PROPOSITION **1.** *If $\widehat{f}$ (or $\widehat{g}$) is a GKE of the density f (resp. the regression function g), with kernel $K(x,t;X_1,\ldots,X_N)$, then $\widehat{f}_{+,\varepsilon}$ (resp. $\widehat{g}_{+,\varepsilon}$) is also a GKE, with kernel*

$$K_\varepsilon(x,t;X_1,\ldots,X_N) = \left[I_{\{\xi\in\mathbb{R}^d \,:\, h(\xi)>0\}}(\cdot)\, K(.,t;X_1,\ldots,X_N) \right]_\varepsilon (x), \tag{2.11}$$

where $h := \widehat{f}$ (resp. $h := \widehat{g}$), and $K_\varepsilon(.,t) = [K(.,t)]_\varepsilon$ is given by the LHS of (3.16) (see also Appendix).

3. Main Results

In this section, we first recall the main idea on which the constrained estimator proposed in Dechevsky and MacGibbon (1999) is based. We are given an "unconstrained" estimator $\widehat{h}$ attaining a certain rate of estimation of the function f and satisfying certain smoothness constraints, but not necessarily a certain order constraint to which h is subject (say, $h \geq v$, where v is a constant). The aim is to construct a new "constrained" estimator $\widehat{h}^+$ based on $\widehat{h}$ which attains essentially the same rate of estimation for h, satisfies the same smoothness constraints and, additionally, obeys the order constraint: $\widehat{h}^+ \geq v$. In Dechevsky and MacGibbon (1999) it is shown that

there is a natural general way to define an estimator $\widehat{h}^+$ with the required properties, namely,

$$\widehat{h}^+ = \widehat{h}_{+,\varepsilon} := \Phi_\varepsilon * \widehat{h}_+ = \frac{1}{\varepsilon^d} \int_{\mathbb{R}^d} \Phi\left(\frac{\cdot - t}{\varepsilon}\right) \widehat{h}_+(t)dt \tag{3.1}$$

where $\widehat{h}_+ = \max(\widehat{h}, v)$, Φ is a "suitably smooth" symmetric non-negative Lebesgue-measurable kernel with integral 1 on $\mathbb{R}^d$).

Compliance of $\widehat{h}_{+,\varepsilon}$ with the smoothness and the order constraints follows in a natural way from the properties of Φ and the definition of $\widehat{h}_+$. The main result of Dechevsky and MacGibbon (1999) (see Theorem 1, given below) is to determine an upper bound ε_N so that, $\widehat{h}_{+,\varepsilon}$ preserves the same rate of estimation for h as $\widehat{h}$ whenever $\varepsilon \in (0, \varepsilon_N]$ holds.

The present paper considers numerical methods for computation of $\widehat{h}_{+,\varepsilon}$ by approximating the integral in the convolution in (3.1) with a quadrature formula. In this way new quadrature-based estimators $\widehat{h}_{+,\varepsilon,quadr}$ are obtained; in Proposition 2 and Theorem 2 we show that they satisfy the smoothness and the order constraints. Besides ε, these new estimators depend on (at least one) additional parameter related to the discretization step of the quadrature formula. In Theorem 3, which is the main technical result in this paper, we determine upper bound(s) on the additional parameter(s), so that the resulting quadrature-based estimators achieve essentially the same rates as $\widehat{h}$ and $\widehat{h}_{+,\varepsilon}$.

We pay particular attention to two limiting cases of Theorem 3. They correspond to the simplest (but most computationally expensive), and to the most complex (and computationally least expensive) algorithm in the whole family of quadrature methods considered in Theorem 3. Finally, in this section, we briefly discuss the possibility of computing exactly, in closed form, the estimator $\widehat{h}_{+,\varepsilon}$ for some special choices of the kernel Φ.

The main result of Dechevsky and MacGibbon (1999) is

THEOREM **1.** *(Theorem 2.1 in Dechevsky and MacGibbon, 1999) Assume that*

$$d \in \mathbb{N}; \tag{3.2}$$

$h : \mathbb{R}^d \to \mathbb{R}$ *is such that there exists a constant* $v \in \mathbb{R}$, *and*

$$h(x) \geqq v \text{ holds for Lebesgue-almost all } x \in \mathbb{R}^d. \tag{3.3}$$

$N \in \mathbb{N}$; X_j, $(j = 1, \ldots, N)$ *are identically distributed with c.d.f.* F *(not necessarily uncorrelated);* (3.4)

$$\widehat{h} = \widehat{h}_N : \mathbb{R}^{\mathrm{d}} \to \mathbb{R} \tag{3.5}$$

is an estimator of h *based on the* X_j, $j = 1, \ldots, N$;

$$1 \leqq p_1 \leqq \infty; \quad 0 < \rho < \infty; \tag{3.6}$$

$$h \in L_{p_1}(\mathbb{R}^{\mathrm{d}}), \ \widehat{\mathrm{h}} \in \mathcal{L}_\rho(\mathbb{R}^{\mathrm{d}}, \mathrm{L}_{\mathrm{p}_1}(\mathbb{R}^{\mathrm{d}}))^\rho \tag{3.7}$$

(see Appendix for the definitions); W *is a (quasi-)Banach space such that*

$$W \hookrightarrow L_{p_1}(\mathbb{R}^{\mathrm{d}}) \ + \ \dot{W}^2_{p_1}(\mathbb{R}^{\mathrm{d}}) \ , \tag{3.8}$$

there exist $\delta_0 > 0$, $\beta \in \Omega(\delta_0)$, *such that*

$$\omega_2(h_1, \delta)_{L_{p_1}} \leqq \beta(\delta) \ \|h_1\|_W, \ \forall h_1 \in W \qquad \forall \delta \in (0, \delta_0], \tag{3.9}$$

where the integral modulus of smoothness (on $\mathbb{R}^d$*), for* $\delta > 0$ *and* $k \in \mathbb{N}$, *is defined by*

$$\omega_k(f; \delta)_{L_p(\mathbb{R}^d)} = \sup_{|h| \leq \delta} \|\Delta^k_h f\|_{L_p(\mathbb{R}^d)},$$

where $h \in \mathbb{R}^d$, $\Delta^k_h f(x)$ *is the* k*-th finite difference with step* h *at* x :

$$\Delta^k_h f(x) = \sum_{\nu=0}^{k} \binom{\nu}{k} (-1)^{k+\nu} f(x + \nu h), \quad \Delta^k_h f(x) = \Delta^1_h \Delta^{k-1}_h f(x));$$

$$h \in W \text{ and } \|h\|_W \leqq L \text{ for some } L \in (0, \infty). \tag{3.10}$$

Then there exist

$$\alpha \in \Omega(\delta_0) = \{\omega : [0, \delta_0] \to [0, \infty) \, ; \ \omega(t_1) > \omega(t_2), \ t_1 > t_2; \\ \exists \lim_{t \to 0+} \omega(t) = \omega(0+) = 0; \ \omega(0) = 0\} \, ,$$

$$K = K(p_1, \rho, W, L, \alpha) \in (0, \infty)$$

$$\text{and } N_0 = N_0(p_1, \rho, W, L, \alpha) \in (0, \infty), \tag{3.11}$$

$$\text{such that } E_{F^N} \|h - \widehat{h}_N\|^\rho_{L_{p_1}} \leqq K \ . \ \alpha\left(\frac{1}{N}\right), \qquad \forall N > N_0. \tag{3.12}$$

It is also true that for $\widehat{h}_{+,\varepsilon}$, *as defined in* (3.1), *then,*

$$E_{F^N} \|h - \widehat{h}_{+,\varepsilon}\|^\rho_{L_{p_1}} \leqq 2^{\rho_1} \ (3 + 2^{\rho/\rho_1})^{\rho_1} \ . \ K \ . \ \alpha\left(\frac{1}{N}\right), \tag{3.13}$$

$\forall N > N_0, \ \forall \varepsilon \in (0, \varepsilon_N]$, *where*

$$\rho_1 = \max\{1, \rho\}, \tag{3.14}$$

$$\varepsilon_N = \beta^{-1}\left(\frac{2}{L}. \ (3 + 2^{\rho/\rho_1})^{\rho_1/\rho}.K^{1/\rho}.\alpha\left(\frac{1}{N}\right)^{1/\rho}\right). \tag{3.15}$$

In the context of density estimator, $h = f$ and $v \equiv 0$, for regression estimation, $h = g$.

REMARK **2.** It is essential that the quasi-norm in W be homogeneous, therefore, without loss of generality in (3.10) we may assume $\|h\|_W = L$. (See also Remark 3.1 in Dechevsky, MacGibbon and Penev, 2000.)

In most cases it is only possible to compute $\widehat{h}_{+,\varepsilon}$ numerically. For conciseness of presentation the case $d = 1$ and $v \equiv 0$ is assumed throughout. The numerical approximation to $\widehat{h}_{+,\varepsilon}$ is achieved by replacing the convolution integral in the definition of $\widehat{h}_{+,\varepsilon}$ by a composite quadrature formula in the following way:

$$\widehat{h}_{+,\varepsilon}(x) = (\widehat{h}W_+)_\varepsilon = \frac{1}{\varepsilon} \int\limits_{-\infty}^{\infty} \Phi\left(\frac{x-t}{\varepsilon}\right) \widehat{h}_+(t)\, dt$$

$$= \frac{1}{\varepsilon} \sum_{\mu=-\infty}^{\infty} \int\limits_{\Delta_\mu} \Phi\left(\frac{x-t}{\varepsilon}\right) \widehat{h}_+(t)\, dt \;\mapsto\; \widehat{h}_{+,\varepsilon,\mathrm{quadr}}(x) := (\widehat{h}_+)_{\varepsilon,\mathrm{quadr}}(x)$$

$$= \frac{1}{\varepsilon} \sum_{\mu=-\infty}^{\infty} \sum_{k=1}^{n_\mu} A_{\mu k} \,.\, \Phi\left(\frac{x-t_{\mu k}}{\varepsilon}\right) .\, \widehat{h}_+(t_{\mu k}), \tag{3.16}$$

where:

$$\{\Delta_\mu\}_{\mu\in\mathbb{Z}} \text{ is a decomposition of } \mathbb{R} \text{ into non-intersecting intervals;} \tag{3.17}$$

$$0 < \inf_\mu |\Delta_\mu| \;\text{ and }\; \sup_\mu |\Delta_\mu| < \infty; \tag{3.18}$$

$$t_{\mu k} \in \Delta_\mu,\; A_{\mu k} \geqq 0; \tag{3.19}$$

$$\inf_{\mu,\mu_1,k,k_1} |t_{\mu_1 k_1} - t_{\mu k}| = C_0 > 0. \tag{3.20}$$

PROPOSITION **2.** *If $\widehat{f}$ (or $\widehat{g}$) is a GKE of the density f (resp. the regression function g), with kernel $K(x,t;X_1,\ldots,X_N)$, then $\widehat{f}_{+,\varepsilon,quadr}$ (resp. $\widehat{g}_{+,\varepsilon,quadr}$) is also a GKE, with kernel*

$$K_{\varepsilon,quadr}(x,t;X_1,\ldots,X_N) = \Big[I_{\{\xi\in\mathbb{R}\,:\,\mathrm{h}(\xi)>0\}}(\cdot)\; K(.,t;X_1,\ldots,X_N)\Big]_{\varepsilon,quadr}(x), \tag{3.21}$$

where $h := \widehat{f}$ (resp. $h := \widehat{g}$) and $(.)_{\varepsilon,quadr}$ is given in the RHS of (3.16); moreover,

$$\widehat{f}_{+,\varepsilon,quadr} \text{ (resp. } \widehat{g}_{+,\varepsilon,quadr}) \text{ is non-negative on } \mathbb{R}. \tag{3.22}$$

THEOREM **2.** *Let $r \in \mathbb{N}$ and $\Phi \in C_0^r(\mathbb{R})$, $\operatorname{supp} \Phi = [-1, 1]$, be the kernel in the definition of approximate identity. Then,*

$$\widehat{f}_{+,\varepsilon,quadr} \in C^r(\mathbb{R}), \ \textit{resp.} \ \widehat{\mathrm{g}}_{+,\varepsilon,quadr} \in \mathrm{C}^{\mathrm{r}}(\mathbb{R}). \tag{3.23}$$

If, moreover, $\widehat{f} \in C_0^r(\mathbb{R})$, resp. $\widehat{g} \in C_0^r(\mathbb{R})$, then also

$$\widehat{f}_{+,\varepsilon,quadr} \in C_0^r(\mathbb{R}), \ \textit{resp.} \ \widehat{\mathrm{g}}_{+,\varepsilon,quadr} \in \mathrm{C}_0^{\mathrm{r}}(\mathbb{R}). \tag{3.24}$$

In other words, quadrature approximations of $\widehat{f}_{+,\varepsilon}$ and $\widehat{g}_{+,\varepsilon}$ generated via the rule given by (3.16) have the same regularity as the smoothing kernel Φ, and have compact support if $\widehat{f}$ and $\widehat{g}$ are compactly supported.

Under the assumptions of Theorem 2, $\widehat{h}_+$ is C^r-regular just as is $\widehat{h}$ itself, everywhere on $\mathbb{R}$ except for those x where $\widehat{h}$ changes its sign. Now, $\widehat{h}_+$ is continuous but not necessarily C^1-smooth at these points which will be called "critical points of $\widehat{h}$ (of the first kind)". These critical points can be encountered in the wavelet estimator considered in Delyon and Juditsky (1996). In the case of the kernel estimator in Lepski, Mammen and Spokoiny, 1977, however, the set of critical points must also include "critical points of the second kind": those x where a jump in the bandwidth has occurred (see the definition in Lepski, Mammen and Spokoiny, 1977). At these special points the estimator $\widehat{h}$ of Lepski, Mammen and Spokoiny (1977) can be non-smooth, even discontinuous. Thus, no matter how large the regularity index r of $\widehat{h}$ is, $\widehat{h}_+$ is at most absolutely continuous on $\mathbb{R}$ in the case of Delyon and Juditsky (1996) and piecewise absolutely continuous with bounded variation on $\mathbb{R}$ in the case of Lepski, Mammen and Spokoiny, 1977.

Consider the uniform cover with disjoint interiors

$$\mathbb{R} = \ \cup_{k=-\infty}^{\infty} \ [k\varepsilon, (k+1)\varepsilon], \tag{3.25}$$

and assume that it is known for which $k \in \mathbb{Z}$, $[k\varepsilon, (k+1)\varepsilon]$ contains at least one critical point of $\widehat{h}$.

Let us define $\{\Delta_\mu\}$. Consider δ, δ_1: $\delta_1 \geqq \delta > 0$, $\frac{1}{\delta} \in \mathbb{N}$, $\frac{1}{\delta_1} \in \mathbb{N}$, whose values will be specified later. For any $k \in \mathbb{Z}$ such that $[k\varepsilon, (k+1)\varepsilon]$ contains a critical point of $\widehat{h}$, consider the refinement

$$[k\varepsilon, (k+1)\varepsilon] \ = \ \cup_{\nu=1}^{1/\delta} \ [(k + (\nu-1)\delta)\varepsilon, \ (k + \nu\delta)\varepsilon]. \tag{3.26}$$

For any $k_1 \in \mathbb{Z}$ such that $[k_1\varepsilon, (k_1+1)\varepsilon]$ does not contain a critical point of $\widehat{h}$, consider

$$[k_1\varepsilon, (k_1+1)\varepsilon] \ = \ \cup_{\nu_1=1}^{1/\delta_1} \ [(k_1 + (\nu_1-1)\delta_1)\varepsilon, \ (k_1 + \nu_1\delta_1)\varepsilon]. \tag{3.27}$$

Denote $\Lambda = \{k \in \mathbb{Z}:\ (3.26) \text{ holds}\}$, $\Lambda_1 = \{k_1 \in \mathbb{Z}:\ (3.27) \text{ holds}\}$. The intervals $\{\Delta_\mu\}$, $\mu \in \mathbb{Z}$, are defined to be the intervals $[(k+(\nu-1)\delta)\varepsilon,\ (k+\nu\delta)\varepsilon]$, $k \in \Lambda$, and $[(k_1+(\nu_1-1)\delta)\varepsilon,\ (k_1+\nu_1\delta)\varepsilon]$, $k_1 \in \Lambda_1$, reordered together, by increasing abscissae.

Define

$$I := \{\mu \in \mathbb{Z}:\ \Delta_\mu \text{ is of type } [(k+(\nu-1)\delta)\varepsilon,\ (k+\nu\delta)\varepsilon],\ k \in \Lambda\ \}, \tag{3.28}$$

$$I_1 := \{\mu \in \mathbb{Z}:\ \Delta_\mu \text{ is of type } [(k_1+(\nu-1)\delta_1)\varepsilon,\ (k_1+\nu\delta_1)\varepsilon],\ k_1 \in \Lambda_1\ \}. \tag{3.29}$$

In each Δ_μ, $\mu \in I$, consider a quadrature formula which is accurate of order 1; if $\mu \in I_1$, then consider a quadrature formula accurate of order r_1. (A quadrature formula on Δ is called accurate of order $r_1 \in \mathbb{N}$, if the error of replacing the integral over Δ by the quadrature formula is zero for any integrand which is an algebraic polynomial of degree not exceeding $r_1 - 1$ on Δ, but there exists a polynomial integrand of degree r_1 on Δ for which this error is not zero.) Here r_1 is any integer satisfying $1 \leqq r_1 \leqq r$, and is the same for all $\mu \in I_1$. Given $\widehat{h}$, denote

$$\widehat{h}_{+,\varepsilon,\delta,\delta_1} \ := \ \widehat{h}_{+,\varepsilon,\text{quadr}}, \tag{3.30}$$

where $\widehat{h}_{+,\varepsilon,\text{quadr}}$ corresponds to the above-defined mesh and composite quadrature formula.

THEOREM **3.** *Under the conditions of Theorem 1, where $\widehat{h}$ denotes the wavelet density or regression-function estimator from Delyon and Juditsky (1996), assume: $d = 1$, $v \equiv 0$ on $\mathbb{R}$;*

$$W = B^s_{pq} \tag{3.31}$$

where B^s_{pq} is the Besov space with quasi-norm

$$\|f\|_{B^s_{pq}(\mathbb{R})} \sim \|f\|_{L_p(\mathbb{R})} + \left[\int_0^\infty \left(\xi^{-s}\omega_r(f;\xi)_{L_p(\mathbb{R})}\right)^q \frac{d\xi}{\xi}\right]^{\frac{1}{q}},$$

$1 \leqq p \leqq \infty$, $1 \leqq q \leqq \infty$, $0 < s_0 \leqq s < r$, *where $q = 1$ if $s = s_0$, and*

$$r \in \mathbb{N},\ r \geqq 2,\ 1 \geqq s_0 \geqq \max\{1/p,\ r_0\},\quad 0 < r_0 \leqq 1; \tag{3.32}$$

$$\Phi \in C^r_0(\mathbb{R}),\quad \text{supp } \Phi = [-1,1], \tag{3.33}$$

as in the definition of approximate identity;

$$\text{supp } h \ \cup\ \text{supp } \widehat{h} \ \subset [-R/2, R/2],\ \ 0 < R < \infty, \tag{3.34}$$

where R does not depend on the sample size N;

$$0 < \varepsilon \leqq 1,\ 0 < \delta \leqq 1. \tag{3.35}$$

Assume also that (3.25-30) hold ($\widehat{h}$ being the wavelet estimator in Delyon and Juditsky, 1996). Suppose further that

$$r_1 \in \mathbb{N},\ 1 \leqq r_1 \leqq r; \tag{3.36}$$

the quadrature formula on Δ_μ is accurate of order 1 $\forall \mu \in I$; (3.37)

the quadrature formula on Δ_{μ_1} is accurate of order r_1 $\forall \mu_1 \in I_1$. (3.38)

Also assume that $s:\ \max\{s_0,\ \sigma\} < s < r$, $\sigma:\ \max\{1/p,\ r_0\} \leqq \sigma < r$.

Then, $\exists N_2 \geqq N_1$ such that the following is true :

$$E_{F^N} \|h - \widehat{h}_{+,\varepsilon,\delta,\delta_1}\|^\rho_{L_{p_1}} \ \leqq\ 2^{2\rho_1}\ (3 + 2^{\rho/\rho_1})^{\rho_1}\ .\ K\ .\ \alpha\left(\frac{1}{N}\right), \tag{3.39}$$

$\forall N > N_2,\ \forall \varepsilon \in (0, \varepsilon_N],\ \forall \delta \in (0, \delta_N],\ \forall \delta_1 \in (0, \delta_{1,N}]$, *where*

$$\varepsilon_N = \min\{1,\ \varepsilon'_N\}, \tag{3.40}$$

ε'_N being the value of ε_N given in (3.15) in Theorem 1;

$$\delta_N = \min\Bigg\{1,\ C^{1+\left(1/s_0 - 1/s_0^*\right)_+}\ .$$
$$\left[\alpha(1/N)^{1/\rho}\beta^{-1}\ \left[C_1\ \alpha(1/N)^{1/\rho}\right]^{1-1/p_1}\right]^{1+\left(1/s_0-1/s_0^*\right)_+}\Bigg\}; \tag{3.41}$$

$$\delta_{1,N} = \min\Bigg\{1,\ C_2^{1/r_1+(1/\sigma^*-1/\sigma^*)_+}\ .$$
$$\left[\alpha(1/N)^{1/\rho}\beta^{-1}\ \left[C_1\ \alpha(1/N)^{1/\rho}\right]^{1-1/p_1}\right]^{1/r_1+(1/\sigma-1/\sigma^*)_+}\Bigg\}; \tag{3.42}$$

$$C_1 = \frac{2}{L}\ .\ (3 + 2^{\rho/\rho_1})^{\rho_1/\rho}\ .\ K^{1/\rho} \tag{3.43}$$

(see (4) in Dechevsky and MacGibbon, 1999); L is as in Theorem 1;

$$C = C'_2(p, p_1, \rho, r_0, s, L)\ .\ \frac{K^{1/\rho}}{R^{1-1/p}\ \|\Phi\|_{W^1_{p_1}}}; \tag{3.44}$$

$$C_2 = C_2(r, r_1, p, q, \rho, R, \Phi) \tag{3.45}$$

can be evaluated from the proof, similarly to C;

$$\sigma^* = \left[1 - \frac{r_1\ \log_N \varepsilon_N}{\log_N C_2 + \frac{1}{\rho}\log_N \alpha\left(\frac{1}{N}\right) + \left(r_1 + 1 - \frac{1}{p_1}\right)\log_N \varepsilon_N}\right] \cdot r_1 \in (0, r_1); \tag{3.46}$$

$$s_0^* = 1 - \frac{\log_N \left[\beta^{-1}\left(C_1\alpha\left(\frac{1}{N}\right)^{1/\rho}\right)\right]}{\log_N C + \frac{1}{\rho}\log_N \alpha\left(\frac{1}{N}\right) + \left(2 - \frac{1}{p_1}\right)\ \log_N \left[\beta^{-1}\left(C_1\alpha\left(\frac{1}{N}\right)^{1/\rho}\right)\right]} \in (0,1). \tag{3.47}$$

REMARK **3.** The above theorem is also valid if $\widehat{h}$ is the kernel estimator from Lepski, Mammen and Spokoiny (1977), with the following modifications: (3.39), (3.41) and (3.42) are respectively replaced by

$$E_{F^N}\|h - \widehat{h}_{+,\varepsilon,\delta,\delta_1}\|_{L_{p_1}}^{\rho} \ = \ O\left(\alpha\left(\frac{1}{N}\right)\right);\ N \to \infty; \tag{3.48}$$

$$\delta_N = O\left\{\left[\alpha(1/N)^{1/\rho}\beta^{-1}\ \left(C_1\ \alpha(1/N)^{1/\rho}\right)^{1-1/p_1}\right]^{1+\left(1/\sigma_0 - 1/s_0^*\right)_+}\right\},\quad \delta_N \leqq 1; \tag{3.49}$$

$$\delta_{1,N} = O\left\{\left[\alpha(1/N)^{1/\rho}\beta^{-1}\ \left(C_1\ \alpha(1/N)^{1/\rho}\right)^{1-1/p_1}\right]^{1/r_1+(1/\sigma-1/\sigma^*)_+}\right\},\quad \delta_{1,N} \leqq 1. \tag{3.50}$$

The constants in the O-estimates in (3.48-50) depend on the expectation of the Wiener-Young variation $\bigvee_{p_{-\infty}}^{+\infty}\widehat{h}$ (see Dechevsky and Penev, 1997, 1998 for its relevant properties). The proof of Theorem 3 allows us to obtain explicit bounds for these constants involving a bound for the expectation of $\bigvee_{p_{-\infty}}^{+\infty}\widehat{h}$; however, this derivation is more technical than in the case of wavelets, because this estimator may have points of jumps which also contribute to the value of $\bigvee_{p_{-\infty}}^{+\infty}\widehat{h}$ and requires a modified approach (see Remark 7.1).

REMARK **4.** The assumptions $\varepsilon_N \leqq 1$, $\delta_N \leqq 1$, $\delta_{1,N} \leqq 1$ are made only for the sake of simplicity of presentation. If more general bounds are considered, these will also appear in the expressions for the constants involved in the theorem.

The two-parameter family of quadrature discretizations of $\mathbb{R}$ considered in Theorem 3 has two extreme limiting cases : $I = \mathbb{Z}, I_1 = \emptyset$ and $I = \emptyset, I_1 = \mathbb{Z}$. They correspond to one-parameter families of discretization which are of special importance. Let us denote the resulting versions of $\widehat{h}_{+,\varepsilon,\delta,\delta_1}$ for them by $\widehat{h}_{+,\varepsilon,\delta}$ and $\widehat{h}_{+,\varepsilon,\delta_1}$, respectively. The former, $\widehat{h}_{+,\varepsilon,\delta}$, is the simplest algorithmically, because no information whatsoever is required for the critical points of $\widehat{h}$; however, the price to pay is that it is also the most computationally expensive. For its computational complexity we have

$$\text{comput. compl.}(\widehat{h}_{+,\varepsilon,\delta}) \ \sim \ N^{\frac{2r}{2r+1}} \times \text{comput. compl.}(\widehat{h}), \tag{3.51}$$

In contrast to this, $\widehat{h}_{+,\varepsilon,\delta_1}$ is the computationally least expensive in the family with

$$\text{comput. compl.}(\widehat{h}_{+,\varepsilon,\delta_1}) \ \sim \ N^{\frac{2}{2r+1}} \times \text{comput. compl.}(\widehat{h}). \tag{3.52}$$

In this case, the price to pay is that its algorithm is the most complex, and requires most information about $\widehat{h}$. Namely, it requires not only a rough localization of the critical points of $(\widehat{h})$ within the intervals $\Delta_\mu,\ \mu \in I$ (which suffices for $\widehat{h}_{+,\varepsilon,\delta,\delta_1}$ with $I \neq \emptyset$), but also their computation with sufficiently high precision, because every such critical point must be situated on the boundary of an interval from the partition of $\mathbb{R}$ corresponding to $\widehat{h}_{+,\varepsilon,\delta_1}$. This requires also a more complex partition algorithm, as follows.

Each interval between two neighbouring critical points $x_1,\ x_2 :\ x_1 < x_2$ is being subdivided into ν intervals of equal length, where

$$\nu \ = \ \left[\left[\frac{x_2 - x_1}{\delta_1}\right]\right] \ + \ 1, \tag{3.53}$$

where $[[\tau]]$ is the Gaussian bracket (i.e., the largest integer less than, or equal to, the real argument τ). The intervals obtained by this procedure form a cover of $[x_1, x_l]$, (where x_1 denotes the smallest and x_l, the largest critical point of $\widehat{h}$). This cover is completed to form a cover of $\mathbb{R}$ by uniform δ_1-partition of $(-\infty, x_1]$ and $[x_l, \infty)$. Assuming that $l = l_N$ remains bounded as $N \to \infty$ (which should be the case, unless the graph of h has fractal properties); the cover obtained is a δ_1-quasi-uniform partition of $\mathbb{R}$ into $\{\Delta_\mu\}$: there exist $0 < K_1(l_0) \leqq K_2(l_0) < \infty$, independent of N, such that

$$K_1(l_0)\ \delta_1 \ \leq \ |\Delta_\mu| \ \leq \ K_2(l_0)\ \delta_1, \quad \forall \mu \in \mathbb{Z}, \tag{3.54}$$

where l_0 is the upper bound for l_N. The Δ_μ all enjoy the property of containing no critical points of $\widehat{h}$ in their interiors. Therefore a quadrature formula accurate of high order can be applied for any Δ_μ.

Finally in this section, we note that if the estimator $\widehat{h}$ is a (polynomial or exponential) spline-kernel or spline-wavelet estimator (see Chui, 1992 and Daubechies, 1992), and if Φ in the definition of $\widehat{h}_{+,\varepsilon}$ is chosen to be a (polynomial or exponential) spline-kernel, then $\widehat{h}_{+,\varepsilon}$ can be computed exactly and explicitly in a closed form. More details about the numerical implementation of this estimator, which ensures that its computational complexity is $O_N(1) \times \text{comp.compl.}(\widehat{h})$, can be found at the end of Section 3 in Dechevsky, MacGibbon and Penev (2000).

4. **Applications**

For a given estimator the results in the previous section can be applied to design optimal strategies of selecting ε, δ and δ_1, so that the quadrature approximation of $\hat{h}_{+,\varepsilon}$ has the same asymptotic rates as $\hat{h}_{+,\varepsilon}$ itself, for a broad class of error measures and function spaces simultaneously. Examples of such results can be obtained, e.g., by combining Theorem 3 with Corollaries 3.1.1-8 in Dechevsky and MacGibbon (1999) for wavelet estimators, and with Corollaries 3.2.1,2 in the same paper for kernel estimators. In Corollary 4.1 in Dechevsky, MacGibbon and Penev (2000) we have considered some of these examples under the assumption that the parameters p, q, s in (3.31) in Theorem 3 are known, together with the metric index p_1. Other problems can be solved analogously.

In this section, $\widehat{h}^{(w)}$ is the wavelet estimator from Theorem 1 in Delyon and Juditsky (1996) and $\widehat{h}^{(k)}$ is the variable-bandwidth kernel estimator from Lepski, Mammen and Spokoiny (1977); $\widehat{h}_{+,\varepsilon,\delta,\delta_1} = \widehat{h}^{(w)}_{+,\varepsilon,\delta,\delta_1}$ or $\widehat{h}^{(k)}_{+,\varepsilon,\delta,\delta_1}$ is the respective quadrature estimator from Theorem 3.

The choice of ε_N, δ, δ_1 in Corollary 4.1 in Dechevsky, MacGibbon and Penev (2000) depends on the explicit knowledge of p_1, p, q and s. Of these, only p_1 is *a priori* known, in general. The following corollary suggests (smaller) values for $\varepsilon_N = \varepsilon_N(p_1)$, $\delta_N = \delta_N(p_1)$ and $\delta_{1,N} = \delta_{1,N}(p_1)$, so that the conclusions of Theorem 3 and of Corollary 4.1 in Dechevsky, MacGibbon and Penev (2000) hold simultaneously for the whole range of admissible p, q and s.

COROLLARY **1.** *Assume that*

$$p_1 : 1 \leq p_1 \leq \infty; \quad 0 < r_0 \leq 1; \quad r \in \mathbb{N}, \quad r \geq 2; \quad r_1 \in \mathbb{N}, \tag{4.1}$$

$$1 \leq r_1 \leq r; \quad s_0 : r_0 \leq s_0 \leq 1; \quad \sigma : r_0 \leq \sigma < r;$$

$$\varepsilon_N = O\left[N^{-\frac{r}{r_0(2r+1)}}\right]; \tag{4.2}$$

$$\delta_N = O\left[N^{-\frac{r}{r_0(2r+1)}\cdot(r_0+1-\frac{1}{p_1})\cdot\max\left\{\frac{1}{s_0}-\frac{1}{o(1)+r_0+1-\frac{1}{p_1}},1\right\}}\right]; \tag{4.3}$$

$$\delta_{1,N} = O\left[N^{-\frac{r}{r_0(2r+1)}\cdot(r_0+1-\frac{1}{p_1})\cdot\max\left\{\frac{1}{\sigma}-\frac{1}{o(1)+r_0+1-\frac{1}{p_1}},\frac{1}{r_1}\right\}}\right]. \tag{4.4}$$

Then,
(*A*) $\hat{h}^{(k)}_{+,\varepsilon,\delta,\delta_1}$ *is asymptotically-minimax optimal within the "asymptopia" paradigm (see Donoho, Johnstone, Kerkyacharian and Picard, 1995) in the risk norm* $\mathcal{L}_{p_1}(L_{p_1})$, $1 \le p_1 < \infty$, *uniformly in:*

$$p : 1 \le p \le \infty, \quad \max\left\{\frac{1}{p}, r_0 + \left(\frac{1}{p} - \frac{1}{p_1}\right)_+\right\} \le \min\{s_0, \sigma\}; \tag{4.5}$$

$$q : 1 \le q \le \infty, \quad \max\{s_0, \sigma\} < s < r; \tag{4.6}$$

(*B*) $\hat{h}^{(w)}_{+,\varepsilon,\delta,\delta_1}$ *is also such in* $\mathcal{L}_2(L_{p_1})$, *uniformly in* q, s *as in (4.6) and*

$$p : 1 \le p \le p_1, \quad \max\left\{\frac{1}{p}, r_0 + \frac{1}{p} - \frac{1}{p_1}\right\} \le \min\{s_0, \sigma\}. \tag{4.7}$$

The following corollary gives (smaller) values of δ_N and $\delta_{1,N}$, so that asymptotically-minimax optimality is preserved uniformly in p_1, also.

COROLLARY **2.** *Under the assumptions of (4.1) for* r_0, r, s_0 *and* σ, *suppose that* ε_N *is given by (4.2) and*

$$\delta_N = O\left[N^{-\frac{r}{r_0(2r+1)}\cdot(r_0+1)\cdot\max\left\{\frac{1}{s_0} - \frac{1}{o(1)+r_0+1}, 1\right\}}\right]; \tag{4.8}$$

$$\delta_{1,N} = O\left[N^{-\frac{r}{r_0(2r+1)}\cdot(r_0+1)\cdot\max\left\{\frac{1}{\sigma} - \frac{1}{o(1)+r_0+1}, \frac{1}{r_1}\right\}}\right]. \tag{4.9}$$

Then, the conclusions (A,B) of Corollary 1 hold, uniformly in $p_1 : 1 \le p_1 < \infty$ *for* $\hat{h}^{(k)}_{+,\varepsilon,\delta,\delta_1}$ *and* $1 \le p_1 \le \infty$ *for* $\hat{h}^{(w)}_{+,\varepsilon,\delta,\delta_1}$ *respectively.*

From (4.3,4) it is seen that the "critical" choices of s_0 and σ, when the rate "changes orders", are:

$$s_0 = 1 + \frac{1}{o(1) + r_0 + 1 - \frac{1}{p_1}}, \quad \sigma = r_1 + \frac{1}{o(1) + r_0 + 1 - \frac{1}{p_1}}, \tag{4.10}$$

which are exactly the respective values of s_0^* and σ^*, computed at $s = r$ and $p_1 = \infty$.

The most important choice of r_1 is $r_1 = r$.

5. Numerical Methods

In this section we propose a class of quadrature formulae which are most appropriate to apply in our context. In Dechevsky, MacGibbon and Penev (2000), Section 5, there is a detailed consideration of this topic, where Hermit-Birkhoff, Padé, Newton-Cotes and Gaussian quadratures have been discussed and compared to quadrature processes based on the Runge principle (Richardson extrapolation) (see Sendov and Popov, 1988, Berezin and Zhidkov, 1965, Ralston and Rabinowitz, 1978, Davis and Rabinowitz, 1975 and Sendov and Popov, 1976 for references on quadratures). This comparison has shown that in our context it is most advantageous to use quadrature processes (i.e., iterative methods for generating the knots and coefficients of the quadrature formula). To ensure that the quadrature process converges to the exact value of the integral with sufficiently high order of accuracy r_1 (presumably, $r_1 = r$), our general recommendation is to repeatedly apply Aitken-Steffensen's method of accelerating the convergence of the iterative process. This approach results in a procedure which can be found under different names in various references, e.g., as Runge principle or Richardson extrapolation. Various types of Richardson extrapolation techniques can be used in our context but, to be concrete, we suggest Romberg integration with the "trapezoid rule" as an initial formula for the iterative process. Thanks to exploiting the structure of the remainder in the Euler-McLaurin summation formula, the recursive formulae for this particular type of Richardson extrapolation are very simple. Briefly, given the order of accuracy r_1, and assuming that f has integrable derivative of even order not less than r_1, $\int_a^b f(x)\,dx$ can be written as

$$\int_a^b f(x)\,dx \;=\; \frac{b-a}{\mu}\left(\frac{1}{2}f_0 \;+\; \sum_{i=1}^{\mu-1} f_i \;+\; \frac{1}{2}f_\mu\right) + \sum_{i=1}^{\nu} a_i \mu^{-2i} \;+\; R_\mu \mu^{-2(\nu+1)}, \tag{5.1}$$

where

$$\nu = \left[\left[\frac{r_1+1}{2}\right]\right] - 1,$$

$$f_i = f\left(a + i\frac{b-a}{\mu}\right), i = 0, 1, \ldots, \mu, \tag{5.2}$$

$$a_i = (b-a)^{2i} \,.\, B_{2i} \,.\, [f^{(2i-1)}(b) - f^{(2i-1)}(a)]/(2i)! \tag{5.3}$$

$$R_\nu = \frac{(b-a)^{2\nu+2}\mu^{-2\nu-2}}{(2\nu+2)!} \int_0^{b-a} B_{2\nu+2}^{\mathrm{per}} \left(\frac{\mu x}{b-a}\right) f^{(2\nu+2)}(a+x)\, dx, \qquad (5.4)$$

B_l is the l-th Bernoulli number, $B_l^{\mathrm{per}}(x)$ is the 1-periodic version of the Bernoulli polynomials:

$$B_l^{\mathrm{per}}(x) = \begin{cases} B_l(x), & 0 \leqq x \leqq 1 \\ B_l^{\mathrm{per}}(x-1), & x > 1 \\ B_l^{\mathrm{per}}(x+1), & x < 0 \end{cases} \qquad (5.5)$$

Note that the expression containing $f_0,\ f_1, \ldots, f_\mu$ in (5.1) is exactly the composite quadrature formula obtained by applying the "trapezoid rule" to every interval of the form $\left[a + (i-1)\dfrac{b-a}{\mu},\ a + i\dfrac{b-a}{\mu}\right]$, $i = 1, \ldots, \mu$.
For $\mu = 2^n$, $n = 0, 1, 2, \ldots$, consider

$$I_{0,n} = (b-a)\ 2^{-n} \left(\frac{1}{2}f_0 + \sum_{i=1}^{2^n-1} f_i + \frac{1}{2}f_{2^n}\right), \qquad (5.6)$$

$$I_{\nu,n} = \frac{1}{2^{2\nu}-1}\,(2^{2\nu} I_{\nu-1,\ n+1} - I_{\nu-1,\ n}). \qquad (5.7)$$

It can be shown that the iterative computation of $I_{\nu,\ n}$ is equivalent to iterative computation of the coefficients and knots of a quadrature formula, starting with (5.6). Computing $I_{\nu,\ n}$ is equivalent to eliminating the coefficients a_i in (5.1) and the resulting quadrature formula for $I_{\nu,\ n}$ is accurate of order $2\nu + 2$: $r_1 \leqq 2\nu + 2 \leqq r_1 + 1$. The organization of the computation is usually the following: first I_{00} is computed, then $I_{0,1}$, then I_{10} via (5.7), etc. On the ν-th step, first $I_{0,\ \nu}$ is computed via (5.6), and then $I_{k,\ \nu-k}$, $k = 1, 2, \ldots, \nu$ are being calculated consecutively via (5.7). The procedure stops at $I_{n,0}$, $n = \left[\left[\frac{r_1+1}{2}\right]\right] - 1$, which is accepted as the value of the quadrature formula for $\int_a^b f(x)\, dx$. An additional advantage is that the values of f used in the computation of $I_{0,\nu}$ can be used in the computation of $I_{0,\nu+1}$, too.

We recommend this type of Romberg integration as the algorithm for numerical integration in all cases discussed in this paper. Here are some of our arguments in its favour:

- it can be shown that the coefficients in the quadrature formula corresponding to $I_{\nu,\ n}$ are non-negative for any ν, n.
- this algorithm works well for both small and large values of r_1. This allows us to always choose the most favourable value of r_1 : $r_1 = r$ in the model of

Theorem 3.

- it works equally well with Daubechies wavelets and spline-wavelets and, in fact, with any kind of dyadic wavelets, since it uses only values of f involved in iterative refinement via the dyadic functional equation for which the wavelet f is fundamental interpolant.
- essentially, this is a *dyadic pyramidal algorithm* which can in a certain sense be considered as an extension of Mallat's pyramidal algorithms "on a subcoefficient level".
- if the quadrature formula for numerical integration on intervals in I in Theorem 3 (see (3.28)) is chosen to be the "trapezoid rule", then all numerical integration can be done via a unified algorithm. Indeed, in the model considered in Theorem 3, if $\Delta_\mu \in I$, then 0 iterations of Romberg integration are being carried out (i.e., the "trapezoid rule" is applied); if $\Delta_\mu \in I_1$ (see (3.29)), then (for $r_1 = r$) $[[(r+1)/2]] - 1$ iterations of Romberg integration are performed.

6. Some Numerical Examples

The purpose of the numerical examples in this section is not to demonstrate the superiority of the fits of a fully automated locally or globally adaptive estimator, but rather *to show typical cases in which the constrained estimators are expected to perform visibly better than the unconstrained versions from which they have been derived.*

In Figures 1,2, the dotted line depicts the standard (hard or soft) thresholded wavelet density estimator, the dashed line corresponds to the new estimator satisfying the constraints and the unbroken line refers to the true curve.

We tested the same numerical data with three types of unconstrained estimators : spline, kernel and wavelet estimators. Although there were certain differences between their performances, that of the new constrained estimators was visibly better than that of the original unconstrained versions, most of the time. We used the simplest quadrature-based estimator $\widehat{h}_{+,\varepsilon,\delta}$, with the *"trapezoid" rule*. Here we give the results for the wavelet estimators using Daubechies' extremal phase wavelets with compact support (Daubechies, 1992, p. 195). (For a given integer value M, the length of the support is $2M-1$ and the smoothness increases with M increasing.) The selected values of ε and δ obey the respective bounds in Theorem 3.

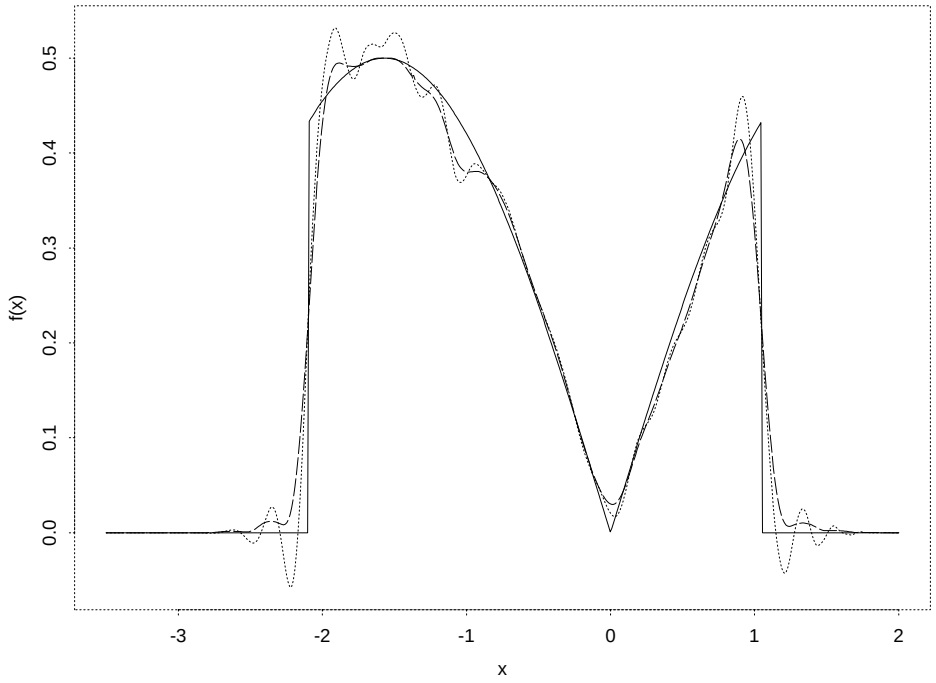

Figure 1. Estimating a sinusoidal density f with a natural lower constraint 0 and an additional upper constraint 0.5: sample size $N = 8000$; wavelet estimator; initial resolution level $j_0 = 2$, highest resolution level $j_1 = 7$; regularity parameter of the Daubechies' wavelet $M = 9$; soft thresholding; estimated ISE's: 6.6×10^{-3} for the standard unconstrained estimator $\widehat{f}$, 7.2×10^{-3} for the new constrained estimator $\widehat{f}_{+,\varepsilon}$; smoothing parameter $\epsilon = 0.15$; quadrature rule – trapezoid.

In Dechevsky, MacGibbon and Penev (2000), we have estimated a mixture of two normal densities given by the formula $f(x) = 0.78\mathcal{N}(0.1, 1) + 0.22\mathcal{N}(1.6, 0.16)$. We found out that even with sample sizes of $N = 1024$, the unconstrained estimator may have negative values in the tails and the application of our method may bring significant benefits, including reduction of the ISE. Here we consider two much more geometrically challenging examples.

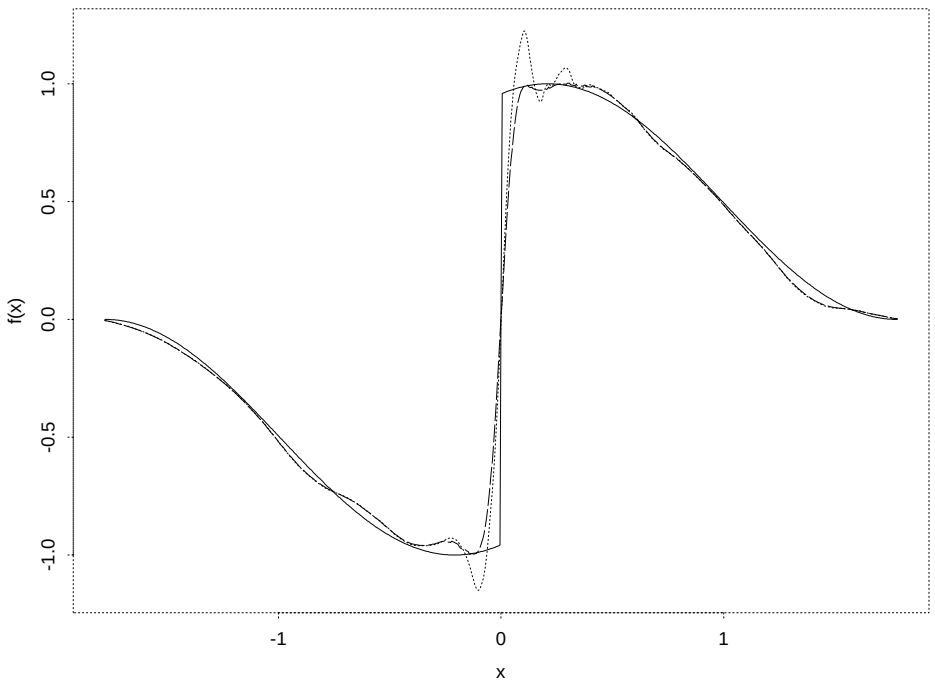

Figure 2. Estimating a regression curve g with a lower and an upper constraint -1 and +1, resp.; sample size $N = 1024$; wavelet estimator; initial resolution level $j_0 = 0$, highest resolution level $j_1 = 2$; regularity parameter of the Daubechies' wavelet $M = 5$; hard thresholding; estimated ISE's: 3.1×10^{-2} for the standard unconstrained estimator $\widehat{g}$, 3.7×10^{-2} for the new constrained estimator $\widehat{g}_{+,\varepsilon}$; smoothing parameter $\epsilon = 0.08$; quadrature rule – trapezoid.

The first example (Figure 1) here is the estimation of the following compactly supported sinusoidal density: $f(x) = 0.5|\sin(x)|, x \in (-2\pi/3, \pi/3)$, with the natural lower bound $v = 0$ and an *a priori* known upper bound $w = 0.5$, i.e., $0 \le f(x) \le 0.5$ for every real x. At the boundary of the support, the density is discontinuous and the standard estimator exhibits strong Gibbs' effect even for very large sample sizes. This is expressed in strong oscillations around the points $-2\pi/3$ and $\pi/3$ and negative values of the estimated density around these points. Applying the smoothed constrained estimator makes the estimated curve much more acceptable.

Accomodating the additional information about the upper bound on f

into our smoothed constrained estimator brings additional visual benefits (see Figure 1). (The non-negativity of the new estimator makes it possible also to enhance its definition in such a way that the resulting rate-optimal estimator not only obeys the lower and upper constraints, but also has integral 1 on $\mathbb{R}$. Details about this additional enhancement in the case of density estimation will be given in a future paper.)

In the second example (Figure 2) in this section, we estimate the regression curve

$$g(x) = \begin{cases} \cos^2(x - \frac{\pi}{15}), \ x \in (0, \frac{17\pi}{30}] \\ 0, \ \text{if } x = 0 \\ -\cos^2(-x - \frac{\pi}{15}), \ x \in (-\frac{17\pi}{30}, 0) \end{cases}$$

Additive Gaussian noise with $\mathcal{N}(0, 0.04)$ has been added to $N = 1024$ observations from the above curve at equidistant points in the interval $(-\frac{17\pi}{30}, \frac{17\pi}{30}]$. The regression curve is known to be constrained from above *and* from below by 1 and -1, respectively. The standard wavelet estimation method does not take these constraints into account. Because of this, and due to the Gibbs' phenomenon, the estimated curve takes values outside the interval $[-1, 1]$ in a neighbourhood of the critical point $x = 0$. These effects are corrected by the smoothed version of our constrained estimator. It can be seen that it satisfies the constraints and at the same time, improves the visual performance of the estimated curve around the critical point. In terms of ISE, the quality of the fit is only slightly worse.

7. Proofs

Since the proofs of the main results of this paper (especially Theorem 3) are very technical and, necessarily, rather long, we can only give short outlines here. For the missing details we refer to Dechevsky, MacGibbon and Penev (2000, Section 7).

PROOF OF PROPOSITION 1 *(outline).* Follows from the definition of $\widehat{h}_{+,\varepsilon}$, $\widehat{h}_+$ and $\widehat{h}$, a change of the order of summation and integration, and then using the definition of GKE (see Dechevsky, MacGibbon and Penev, 2000, 7.1). □

PROOF OF PROPOSITION 2 *(outline).* The proof of (3.22) is analogous to the proof of (2.11) in Proposition 1. The proof of (3.23) follows from the definition of $\widehat{f}_{+,\varepsilon,\text{quadr}}$ ($\widehat{g}_{+,\varepsilon,\text{quadr}}$) and $A_{\mu k} \geqq 0$ in (3.20). □

PROOF OF THEOREM 2 *(outline).* We give the proof for $\widehat{g}$; the one for $\widehat{f}$ is analogous.

$$\begin{aligned}\widehat{g}_{+,\varepsilon,\text{quadr}}(x) &= (\widehat{g}_+)_{\varepsilon,\text{quadr}} \\ &= \frac{1}{N}\sum_{\nu=1}^{N} Y_\nu \cdot \left[\frac{1}{\varepsilon}\sum_{\mu=-\infty}^{\infty}\sum_{k=1}^{n_\mu} A_{\mu k}\,\Phi\left(\frac{x-t_{\mu k}}{\varepsilon}\right)\right. \\ &\qquad \left.\cdot I_{\{\xi\in\mathbb{R}\,:\,\widehat{g}(\xi)>0\}}(t_{\mu k}) \cdot K(t_{\mu k}, X_\nu; X_1, \ldots, X_N)\right] \quad (7.1)\end{aligned}$$

where only a finite number of summands in $\sum_{\mu}\sum_{k}$ is non-zero. Namely, $\Phi\left(\frac{x-t_{\mu k}}{\varepsilon}\right)$ is non-zero for only those $t_{\mu k}$ for which $|x - t_{\mu k}| < \varepsilon$.

For every $x_0 \in \mathbb{R}$ consider $I_{x_0,\,\varepsilon} \subset \mathbb{Z}\times\mathbb{N}$, defined by

$$I_{x_0,\,\varepsilon} = \{(\mu,k) :\ |x_0 - t_{\mu k}| \leqq \varepsilon\}. \quad (7.2)$$

Observe that, for C_0 defined in (3.20),

$$\forall x :\ |x - x_0| \leqq C_0/2 \implies I_{x,\,\varepsilon} = I_{x_0,\,\varepsilon}\ . \quad (7.3)$$

Therefore, in a $C_0\ \varepsilon/2$-neighbourhood of any $x_0 \in \mathbb{R}$, $(\widehat{g}_+)_{\varepsilon,\text{quadr}}$ is the same finite linear combination of translates $\Phi\left(\frac{x-t_{\mu k}}{\varepsilon}\right)$, hence

$$\widehat{g}_{+,\varepsilon,\text{quadr}}\Big|_{[x_0-\varepsilon C_0/2,\ x_0+\varepsilon C_0/2]} \in\ C^r([x_0 - \varepsilon C_0/2,\ x_0 + \varepsilon C_0/2]), \quad (7.4)$$

$\forall x_0 \in \mathbb{R}$,

which easily implies

$$\widehat{g}_{+,\varepsilon,\text{quadr}} \in C^r(\mathbb{R}) \implies (3.23). \quad (7.5)$$

If $\operatorname{supp}\widehat{g} \subset [-R, R]$ for some $R :\ 0 < R < \infty$, then $\operatorname{supp}\Phi = [-1, 1]$ and (7.2) imply

$$\operatorname{supp}\widehat{g}_{+,\varepsilon,\text{quadr}} \subset [R - \varepsilon, R + \varepsilon] \quad \implies (3.24). \quad (7.6)$$

□

PROOF OF THEOREM 3 *(outline).*

$$\left(E_{F^N}\|h-\widehat{h}_{+,\varepsilon,\delta,\delta_1}\|^{\rho}_{L_{p_1}}\right)^{1/\rho_1} \leqq 2(3+2^{\rho/\rho_1})K^{1/\rho_1}\,\alpha\left(\frac{1}{N}\right)^{1/\rho_1} + \left(E_{F^N}\|\widehat{h}_{+,\varepsilon}-\widehat{h}_{+,\varepsilon,\delta,\delta_1}\|^{\rho}_{L_{p_1}}\right)^{1/\rho_1}, \quad (7.7)$$

$$\|\widehat{h}_{+,\varepsilon}-\widehat{h}_{+,\varepsilon,\delta,\delta_1}\|_{L_{p_1}} \leqq \widehat{\mathcal{J}}+\widehat{\mathcal{J}}_1, \quad (7.8)$$

$$\text{where } \widehat{\mathcal{J}} = \left(\int_{-\infty}^{\infty}\left|\frac{1}{\varepsilon}\sum_{\mu\in I}\left[\int_{\Delta_\mu}\Phi\left(\frac{x-t}{\varepsilon}\right)\,\widehat{h}_+(t)\,dt - \sum_{\nu=1}^{n_\mu}A_{\mu\nu}\,\Phi\left(\frac{x-t_{\mu\nu}}{\varepsilon}\right)\widehat{h}_+(t_{\mu\nu})\right]\right|^{p_1}dx\right)^{1/p_1} \quad (7,9)$$

$$\widehat{\mathcal{J}}_1 = \left(\int_{-\infty}^{\infty}\left|\frac{1}{\varepsilon}\sum_{\mu_1\in I_1}\left[\int_{\Delta_{\mu_1}}\Phi\left(\frac{x-t}{\varepsilon}\right)\,\widehat{h}_+(t)\,dt - \sum_{\nu=1}^{n_{\mu_1}}A_{\mu_1\nu}\,\Phi\left(\frac{x-t_{\mu_1\nu}}{\varepsilon}\right)\widehat{h}_+(t_{\mu_1\nu})\right]\right|^{p_1}dx\right)^{1/p_1} \quad (7.10)$$

Here I and I_1 are defined in (3.28, 29), respectively.

To bound $\widehat{\mathcal{J}}$, we use the fact that the quadrature formula is accurate of order at least 1, and then apply consecutively the generalized and the "standard" Minkowski's inequality, a change of variable in Δ_μ, the translation-invariance of $L_{p_1}(\mathbb{R})$, the definition of local, integral and average moduli of smoothness, the positivity of $A_{\mu k}$, and the inequality

$$\sum_{\mu\in I}\int_{\Delta_\mu}\omega_1(\widehat{h}_+,\theta,\delta\varepsilon)\,d\theta \leqq \sum_{\mu=-\infty}^{\infty}\int_{\Delta_\mu}\omega_1(\widehat{h}_+,\theta,\delta\varepsilon)\,d\theta = \tau_1(\widehat{h}_+,\delta\varepsilon)_{L_1}. \quad (7.11)$$

to obtain $\widehat{\mathcal{J}} \leq \frac{1}{\varepsilon}\left[\left\|\Phi\left(\frac{\cdot}{\varepsilon}\right)\right\|_{L_{p_1}}.\,\tau_1(\widehat{h}_+;2\delta\varepsilon)_{L_1} + T_1\,.\,\omega_1\left(\Phi\left(\frac{\cdot}{\varepsilon}\right);\delta\varepsilon\right)_{L_{p_1}}\right]$, where

$$T_1 = \sum_{\mu=-\infty}^{\infty}\sum_{\nu=1}^{n_\mu}|A_{\mu\nu}|\,.\,|\widehat{h}_+(t_{\mu\nu})|\,. \quad (7.12)$$

(For details, see Dechevsky, MacGibbon and Penev, 2000, 7.9 – 13, $9'_1$,38.)

The bound for T_1 is relatively simple to derive, in view of $|\Delta_\mu| = \delta\varepsilon$.

$$T_1 \leq \tau_1(\widehat{h}, \delta\varepsilon)_{L_1} + \|\widehat{h}\|_{L_1} \tag{7.13}$$

(see Dechevsky, MacGibbon and Penev, 2000, 7.14-16).

Next we show that

$$\tau_1(\widehat{h}_+; 2\delta\varepsilon)_{L_1} \ \leqq \ 2\tau_1(\widehat{h}; \delta\varepsilon)_{L_1} \tag{7.14}$$

These results imply, together with Hölder's inequality and (3.34),

$$\begin{aligned} \widehat{\mathcal{J}} \ \leqq \ & \frac{R^{1-1/p}}{\varepsilon} \left[2 \left\| \ \Phi\left(\frac{\cdot}{\varepsilon}\right) \right\|_{L_{p_1}} \ \tau_1(\widehat{h}; \delta\varepsilon)_{L_p} \right. \\ & \left. + \ \left(\|\widehat{h}\|_{L_p} \ + \ \tau_1(\widehat{h}; \delta\varepsilon)_{L_p} \right) \ \omega_1 \left(\ \Phi\left(\frac{\cdot}{\varepsilon}\right); \delta\varepsilon \right)_{L_{p_1}} \right] . \end{aligned} \tag{7.15}$$

(See Dechevsky, MacGibbon and Penev, 2000, 7.17-21). Obtaining an analogous bound for $\widehat{\mathcal{J}}_1$ is much more technical. First of all, observe that

$$\widehat{h}_+\Big|_{\Delta_{\mu_1}} \ \equiv \ \widehat{h}\Big|_{\Delta_{\mu_1}}, \qquad \forall \mu_1 \in I_1. \tag{7.16}$$

Therefore, $\widehat{\mathcal{J}}_1$ can be rewritten as

$$\begin{aligned} \widehat{\mathcal{J}}_1 \ = \ & \left(\int_{-\infty}^{\infty} \left| \frac{1}{\varepsilon} \sum_{\mu_1 \in I_1} \left[\int_{\Delta_{\mu_1}} \Phi\left(\frac{x-t}{\varepsilon}\right) \ \widehat{h}(t) \ dt - \right. \right. \right. \\ & \left. \left. \left. \sum_{\nu=1}^{n_{\mu_1}(r_1)} A_{\mu_1\nu} \ \Phi\left(\frac{x - t_{\mu_1\nu}}{\varepsilon}\right) \ \widehat{h}(t_{\mu_1\nu}) \right] \right|^{p_1} dx \right)^{1/p_1} . \end{aligned} \tag{7.17}$$

We use the Steklov mean $(\widehat{h})_{r_1,\, \delta_1\varepsilon/r_1}$ of $\widehat{h}$ and expand the function

$$G(x,t) \ := \ \Phi\left(\frac{x-t}{\varepsilon}\right) \ (\widehat{h})_{r_1,\, \delta_1\varepsilon/r_1}(t) \tag{7.18}$$

as a function of t in a local Taylor expansion with integral remainder involving $\dfrac{\partial^{r_1}}{\partial t^{r_1}} G(x,t)$, then use the definitions of Φ and the Steklov mean, invoke the accuracy of the quadrature formula of order r_1, the Leibniz rule and the properties of the Steklov mean to obtain an intermediate result

(see Dechevsky, MacGibbon and Penev, 2000, 7.11', 40-43, 11", 44-46, 11"', 11^{iv}), after which the derivations are similar to those of (7.13), and the final result is (see Dechevsky, MacGibbon and Penev, 2000, 7.12')

$$\begin{aligned}\widehat{\mathcal{J}}_1 \;\leqq\; & 2\varepsilon^{-(1-1/p_1)}\, R^{1-1/p_1} \Bigg[C_1'(r_1)\, \|\Phi\|_{L_{p_1}} \tau_{r_1}(\widehat{h};\, \delta_1\varepsilon)_{L_p} \\ & + \frac{2^{-r_1}\,.\,(r_1!)}{(2r_1-1)!}\, \delta_1^{r_1} \sum_{j=0}^{r_1} \binom{r_1}{j} \delta_1^{-j} \\ & .\, C'(r_1,j)\,.\, C_1'(j)\,.\, \left\| \Phi^{(r_1-j)} \right\|_{L_{p_1}} .\tau_j(\widehat{h};\, \delta_1\varepsilon)_{L_p} \Bigg] \end{aligned} \tag{7.19}$$

Now we obtain the final bounds for $\left(E_{F^N}\widehat{\mathcal{J}}^{\rho}\right)^{\frac{1}{\rho_1}}$ and $\left(E_{F^N}\widehat{\mathcal{J}}_1^{\rho}\right)^{\frac{1}{\rho_1}}$. For the former, we use (7.15) and apply some of the properties of the integral and average moduli of smoothness, then invoke some embedding theorems about Besov spaces and apply Theorem 1 in Delyon and Juditsky (1996), to arrive, after computation, at

$$\begin{aligned}\left(E_{F^N}\widehat{\mathcal{J}}^{\rho}\right)^{\frac{1}{\rho_1}} \;\leq\; & \Big[C_3(p,p_1,\rho,r_0,s,L)\,.\, R^{1-1/p}\,. \\ & \|\Phi\|_{W^1_{p_1}}\,.\, L\,.\, \varepsilon^{-(1-1/p_1)}\,.\; \delta^{s_0}\,.\max\{\varepsilon^{s_0},\; \delta^{1-s_0}\}\,\Big]^{\frac{\rho}{\rho_1}} \end{aligned} \tag{7.20}$$

$\forall N > N_1$, (for the details, see Dechevsky, MacGibbon and Penev, 2000, 7.21-28, 3.31). For the bound of $\left(E_{F^N}\widehat{\mathcal{J}}_1^{\rho}\right)^{\frac{1}{\rho_1}}$ we use additionally Marchaud-type inequalities of the first kind for the average moduli of smoothness (see Johnen and Scherer, 1977, Sendov and Popov, 1988 and Dechevsky, 1988) and, after computation, obtain

$$\begin{aligned}\left(E_{F^N}\widehat{\mathcal{J}}_1^{\rho}\right)^{\frac{1}{\rho_1}} \;\leq\; & C_4(p,q,\sigma,s,r,\rho,L)\,\Big(R^{1-1/p}\,\|\Phi\|_{W^{r_1}_{p_1}}\Big)^{\rho/\rho_1} \\ & \cdot\Big[\varepsilon^{-(1-1/p_1)}\,\delta_1^{\min\{\sigma,\; r_1\}}\,\max\Big\{\varepsilon^{\min\{\sigma,\; r_1\}},\; \delta_1^{(r_1-\sigma)_+}\Big\}\Big]^{\rho/\rho_1}, \end{aligned} \tag{7.21}$$

$\forall N > N_1$. (For the details of this derivation, see Dechevsky, MacGibbon and Penev, 2000, 7.47, 47', 24', 48-53, 21', 28'.)

Now s_0^*, δ_N and $\sigma^*, \delta_{1,N}$ are obtained from the requirement that the contribution of each of the two quantities in the RHS of (7.20) and (7.21) does not exceed that of the first term in the RHS of (7.7), which leads to the

requirement that

$$\begin{array}{rcl} \varepsilon_N^{s_0^*} & = & \delta_N^{1-s_0^*}, \\ \varepsilon_N^{-(1-1/p_1)} \ . \ \delta_N & = & C_5^{\rho_1/\rho} \ . \ \alpha(1/N)^{1/\rho} \end{array} \tag{7.22}$$

hold simultaneously, and

$$\begin{array}{rcl} \varepsilon_N^{\min\{\sigma,\ r_1\}} & = & \delta_{1,N}^{(r_1-\sigma)_+} \\ \varepsilon_N^{-(1-1/p_1)} \ \delta_{1,\ N}^{r_1} & = & C_6^{\rho_1/\rho} \ . \ \alpha(1/N)^{1/\rho} \end{array} \tag{7.23}$$

hold simultaneously, too. Here

$$\begin{array}{rcl} C_5 & = & C_5(p, p_1, \rho, r_0, s, L, K, R, \phi) \\ & = & \dfrac{2(3 + 2^{\rho/\rho_1}) \ . \ K^{1/\rho_1}}{[C_3(p, p_1, \rho, r_0, s, L) \ . \ R^{1-1/p} \ . \ \|\Phi\|_{W_{p_1}^1} \ . \ L]^{\rho/\rho_1}} \end{array} \tag{7.24}$$

and

$$C_6 \quad = \quad \frac{2(3 + 2^{\rho/\rho_1}) \ . \ K^{1/\rho_1}}{\left[C_7(p, q, \sigma, s, \rho, L) \ \ R^{1-1/p} \ . \ \|\Phi\|_{W_{p_1}^{r_1}} \ . \ L\right]^{\rho/\rho_1}}. \tag{7.25}$$

After computation, using also that $\varepsilon_N \to 0$ as $N \to \infty$, we obtain (3.41-47) which completes the proof of the theorem. (For the details of these derivations, see Dechevsky, MacGibbon and Penev, 2000, 7.29-37 and 7.30', 31', 31", 32'.) □

REMARK **5.** For the variable-kernel regression-function estimators from Lepski, Mammen and Spokoiny (1977) asymptotic minimax rates are available only when the loss function in the risk is the L_{p_1}-norm. It cannot be replaced by any $B_{p_1q_1}^{s_1}$-norm for any $s_1 : \ 0 < s_1 < s$, because the estimators from Lepski, Mammen and Spokoiny (1977) may be non-smooth, even discontinuous at a finite number of points (if h has compact support) (or locally finite, if supp h is not compact). Because of this, the bounds about the constant factors for the rates (see Dechevsky, MacGibbon and Penev, 2000, 7.25–27) are unavailable for this estimator. However, if $s_0 = 1/p$ (cf. (3.31)), then $\tau_1(\widehat{h}; \delta\varepsilon)_{L_p}^{\rho/\rho_1}$ can be evaluated in the following way (compare with Dechevsky, MacGibbon and Penev, 2000, 7.24–26):

$$\tau_1(\widehat{h}; \delta\varepsilon)_{L_p} \ \leqq \ C_p \ . \ (\delta\varepsilon)^{1/p} \ . \ \left[\bigvee_{p\ -\infty}^{+\infty} \widehat{h}\right]^{1/p} \ = \ C_p \ . \ (\delta\varepsilon)^{s_0} \ . \ O_N(1) \tag{7.26}$$

In other words, although the estimator may be discontinuous, it always has bounded Jordan variation, hence p-variation, also. Note that in this case $\rho = p_1$.

PROOF OF COROLLARY 1. The highest rate for $\alpha(1/N)^{1/\rho}$ is obtained for $s = r$:

$$\alpha(1/N)^{1/\rho} = O(N^{-\frac{r}{2r+1}}); \tag{7.27}$$

the lowest possible rate for $\beta(\delta)$ is obtained for $s = r_0$:

$$\beta(\delta) = O(\delta^{r_0}). \tag{7.28}$$

(For the details of these computations, see Dechevsky, MacGibbon and Penev, 2000, the proof of Corollary 1).

Now s_0^* and σ^*, as computed in (3.46,47), depend on s, and increase together with s. Therefore, for fixed s_0 and σ,

$$1 + \left(\frac{1}{s_0} - \frac{1}{s_0^*}\right)_+ \quad , \quad \frac{1}{r_1} + \left(\frac{1}{\sigma} - \frac{1}{\sigma^*}\right)_+ \tag{7.29}$$

also increase together with s, so they are largest for $s = r$.

On the other hand, denoting $Q = Q(\beta) : \ \beta(\delta) = \delta^Q$, it can be shown (see Dechevsky, MacGibbon and Penev, 2000, 7.53-57) that $(1 - 1/p_1)/Q$ decreases with the increase of s, and is bounded from above by its value at $s = r_0$, uniformly in all admissible choices of s_0 and σ. Combining (7.27-29) yields $(4.2 - 4)$, and (A, B) are fulfilled. □

PROOF OF COROLLARY 2. Follows from Corollary 1 by noticing that the quantities in (7.29), as well as $(1 - 1/p_1)/Q$, all attain their maximal values at $p_1 = \infty$. □

Acknowledgment. This research has been supported by the Natural Sciences and Engineering Research Council of Canada and by the Australian Research Council. We wish to thank GERAD, where parts of this paper were written in a very hospitable atmosphere while the third author was visiting the first two in Montreal. We also thank the referees whose remarks helped us to make this exposition more concise.

Appendix: Preliminaries

For the notions of quasi-norm, semi-norm, quasi-(semi-)normed abelian group, and the power A^ρ of the space A, we refer to Dechevsky, MacGibbon and Penev (2000, Appendix 0), and to Bergh and Löfström (1976) and Dechevsky and Penev (1997).

A is a quasi-Banach space, if A is a complete quasi-normed abelian group, a linear space, and the quasi-norm is homogeneous: $\|\alpha a\|_A = |\alpha|\|a\|_A$.

$A \hookrightarrow B$ ($B \hookleftarrow A$): continuous embedding of A in B (for quasi-normed abelian groups) - see Bergh and Löfström (1976).

$A \rightleftharpoons B : A \hookrightarrow B$ and $B \hookrightarrow A$. In this case $\|\cdot\|_A$ and $\|\cdot\|_B$ are equivalent - denoted by $\|a\|_A \sim \|a\|_B$.

The error measure in which the risk will be measured is the same as in Dechevsky and MacGibbon (1999), Delyon and Juditsky (1996) and Lepski, Mammen and Spokoiny (1977), that is, the quasi-norm of the quasi-normed abelian group $\mathcal{L}_\rho(L_p)^\rho = \mathcal{L}_\rho\left(\mathbb{R}^d, L_p(\mathbb{R}^d)\right)^\rho$, where $0 < p \leq \infty$, $0 < \rho < \infty$, defined in Dechevsky and MacGibbon (1999), Appendix 0. For the sake of completeness we recall the definition of this quasi-norm:

$$\|\widehat{h}\|_{\mathcal{L}_\rho(L_p)^\rho} = E_{F^N}\left(\|\widehat{h}\|_{L_p}^\rho\right),$$

where $\widehat{h} = \widehat{h}(x)$, $x \in \mathbb{R}^d$, is defined in Section 2.

In our consideration of wavelet estimators based on orthonormal wavelets on $\mathbb{R}^d$ (see 2.11, 13, 15) we use exactly the same notation as in Delyon and Juditsky (1996). (In particular, φ and ψ are the respective scaling function and "mother" wavelet; $\varphi, \psi \in C_0^r(\mathbb{R}^d)$, where $C_0^r(\mathbb{R}^d)$ consists, as usual, of all $f \in C^r(\mathbb{R}^d)$ (r-times continuously differentiable) with compact support. Here $r \in \mathbb{N}$ is the upper bound for the smoothness index s of the Besov spaces under consideration (as in Delyon and Juditsky, 1996.)

The kernels $\Phi(x)$ and $\Phi_\varepsilon(x)$ (sometimes called approximate identity), $x \in \mathbb{R}^d$, $\varepsilon > 0$, are defined in Dechevsky and MacGibbon (1999), Section 1 and Appendix 0. Note that $\Phi \in C_0^r(\mathbb{R}^d)$.

As usual, $L_{1,loc}(\mathbb{R}^d) = \{f : \mathbb{R}^d \to \mathbb{R},\ f|_\Omega \in L_1(\Omega)$ for any compact $\Omega \subset \mathbb{R}^d\}$. For $f \in L_{1,loc}(\mathbb{R}^d)$, the convolution $f_\varepsilon(x) = \Phi_\varepsilon * f(x)$, $0 < \varepsilon < \infty$, is in $C^r(\mathbb{R}^d)$ (and in $L_p(\mathbb{R}^d)$ if $f \in L_p(\mathbb{R}^d)$). f_ε is called the Sobolev $\varepsilon-$mean of f (see Nikol'skiĭ, 1975 and Dechevsky and MacGibbon, 1999).

For $f \in L_{1,loc}(\mathbb{R})$, $0 < \varepsilon < \infty$,

$$f_{r,\varepsilon}(x) = (-\varepsilon)^{-r} \underbrace{\int_0^{\varepsilon} \cdots \int_0^{\varepsilon}}_{r} \left[\sum_{\nu=0}^{r-1} (-1)^{\nu+r} f\left(x + \frac{r-\nu}{r} \sum_{\mu=1}^{r} \theta_\mu \right) \right] d\theta_1 \cdots d\theta_r,$$

is the Steklov (r, ε)-mean of f (see Sendov and Popov, 1988, Petrushev and Popov, 1987 and Dechevsky and Penev, 1997; for the multivariate case $d > 1$, see Johnen and Scherer, 1977).

For the properties of the *integral modulus* of smoothness of $f \in L_{1,loc}(\mathbb{R})$, of order r, in $L_p(\mathbb{R}^d)$, $1 \leq p \leq \infty$, with step $\varepsilon > 0$: $\omega_r(f;\varepsilon)_p = \omega_r(f;\varepsilon)_{L_p(\mathbb{R}^d)}$, and its properties, we refer to Nikol'skiĭ (1975), Sendov and Popov (1988), Petrushev and Popov (1987), Dechevsky and Penev (1997), Dechevsky and Penev (1998), Dechevsky and MacGibbon (1999), Johnen and Scherer (1977), Bergh and Löfström (1976) and Triebel (1983).

For $\varepsilon > 0$, $0 < p \leq \infty$, (see Dechevsky, 1988a,b) consider $A_{p,\varepsilon}(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R},\ \|f\|_{A_{p,\varepsilon}(\mathbb{R})} < \infty\}$, where $\|f\|_{A_{p,\varepsilon}(\mathbb{R})} = \|S(\varepsilon, |f|;\cdot)\|_{L_p(\mathbb{R})}$, $S(\varepsilon, f; x) = \sup\{f(y) : y \in [x-\varepsilon, x+\varepsilon]\}$, often called *an upper Baire's function*. The *local modulus of smoothness* is a generalisation of $S(\frac{\varepsilon}{2}, |f|; x)$ (which corresponds to $r = 0$): with $\omega_r(f, x; \varepsilon) = \sup\{|\Delta_\theta^r f(y)| : y, y + r\theta \in [x - \frac{r\varepsilon}{2}, x + \frac{r\varepsilon}{2}]\}$, (see Sendov and Popov, 1988, Dechevsky, 1988 and Ptrushev and Popov, 1987).

The *averaged modulus of smoothness* (or τ-*modulus*) of $f : \mathbb{R} \to \mathbb{R}$, of order r, in $L_p(\mathbb{R})$, $0 < p \leq \infty$, with step $\varepsilon > 0$, is defined by $\tau_r(f;\varepsilon)_p = \tau_r(f;\varepsilon) = \|\omega(f,\cdot;\varepsilon)\|_{L_p(\mathbb{R})}$ (see Sendov and Popov, 1988, Dechevsky, 1988 and Ptrushev and Popov, 1987).

For the definition of the *Wiener-Young variation* $\bigvee_{p-\infty}^{+\infty}$ of f, $1 \leq p \leq \infty$, and for its relevant properties, we refer to Dechevsky and Penev (1997), Dechevsky and Penev (1998), Sendov and Popov (1988), Petrushev and Popov (1987), as well as Dechevsky and MacGibbon (1999), Dechevsky (1988a). In this paper we make implicit use of some of the properties of $A_{p,\varepsilon}$ and the three types of moduli of smoothness; all these properties, including the relevant Marchaud-type inequalities and bounds by the Wiener - Young variation, can be found in the sources listed above.

More relevant details about the definition of some equivalent quasi-norms in the Besov spaces $B^s_{pq}(\mathbb{R})$ defined in Section 3 can be found in Dechevsky, MacGibbon and Penev (2000, Appendix 0). For the relevant properties of Besov spaces we refer to Bergh and Löfström (1976), Triebel (1983) and Triebel (1992). The so-called *A-spaces* A^s_{pq} have been defined by V. A. Popov analogously to the above definition of B^s_{pq}, with the integral modulus being

replaced by the respective averaged one; In Dechevsky (1988b), Section 2, if has been proved that $A^s_{pq}(\mathbb{R}) \rightleftharpoons B^s_{pq}(\mathbb{R})$, $0 < p \leq \infty$, $0 < q \leq \infty$, $s > \frac{1}{p}$. (It can also be proved that this isomorphism continues to hold true for $s = \frac{1}{p}$, $q = \min\{1, p\}$.) This is very essential for our proofs. (For the multivariate case, see the relevant comment in Dechevsky, MacGibbon and Penev (2000, Appendix 0).)

In our proofs we also exploit the close relationship between the Steklov (r, ε)-mean and the integral, local and averaged moduli of smoothness. For a discussion of this relationship, we refer to Dechevsky, MacGibbon and Penev (2000, Appendix A), and for the relevant details – to Sendov and Popov (1988), Dechevsky (1988a,b), Bergh and Löfström (1976), Dechevsky and Penev (1997), Triebel (1983) and Petrushev and Popov (1987).

References

Berezin, I.S. and Zhidkov, N.P. (1965). *Computing Methods.* Pergamon Press, Oxford.

Bergh, J. and Löfström, S. (1976). *Interpolation Spaces: An Introduction.* Grundlehren der Mathematischen Wissenschaften, **223**, Springer-Verlag, Berlin.

Chui, C.K.. (1992). On Cardinal spline wavelets. In *Wavelets and their Applications*, M.B. Ruskai *et al.*, eds., Jones and Barlett, 419–438.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, Philadelphia.

Davis, P.J. and Rabinowitz, P. (1975). *Methods of Numerical Integration.* Academic Press, New York.

Dechevsky, L.T. (1988a). *Ph.D. Dissertation,* Sofia University, 1988.

– – – – (1988b). τ-Moduli and interpolation. In *Function Spaces and Applications*, M. Cwikel, J. Peetre, Y. Sagher and H. Wallin, eds., Lecture Notes in Mathematics, **1302**, Springer-Verlag, Berlin, 177-190.

Dechevsky, L.T. and MacGibbon, B. (1999). Asymptotically minimax non-parametric function estimation with positivity constraints I. *Les Cahiers du GERAD* (ISSN0711-2440) G–99–24.

L. T. Dechevsky, B. MacGibbon, S.I. Penev (2000). Numerical methods for asymptotically minimax non-parametric function estimation with positivity constraints I. *Les Cahiers du GERAD* (ISSN0711-2440) G-2000-33, Montréal, Canada.

L.T. Dechevsky, S.I. Penev (1997). On shape-preserving probabilistic wavelet approximators, *Stochastic Analysis and Application*, **15**, 187–215.

– – – – (1998). On shape-preserving wavelet estimators of cumulative distribution functions and densities, *Stochastic Analysis and Application*, **16**, 428–469.

Delyon, B. and Juditsky, A. (1996). On minimax wavelet estimators, *Applied and Computational Harmonic Analysis*, **3**, 215–228.

Donoho, D.L., Johnstone, I., Kerkyacharian, G. and Picard, D. (1995). Wavelet Shrinkage : asymptopia? (with discussion), *Journal of the Royal Statistical Society, Series B*, **57**, 301–369.

Donoho, D.L., Johnstone, I.M. (1998). Minimax estimation via wavelet shrinkage, *Annals of Statistics*, **26**, 879–921.

JOHNEN, H. and SCHERER, K. (1977). On the equivalence of the K-functional and moduli of continuity and some applications. In *Constructive Theory of Functions of Several Variables*, W. Schempp and K. Zeller, eds., Lecture Notes in Mathematics, **571**, Springer, 119–140.

LEPSKI, O.V., MAMMEN, E. and SPOKOINY, V.G. (1977). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimators with variable bandwidth selectors, *Annals of Statistics*, **25**, 929–947.

NIKOL'SKIĬ, S.M. (1975). *Approximation of Functions of Several Variables and Imbedding Theorems.* Grundlehren der Mathematischen Wissenschaften, **205**, Springer-Verlag, Berlin.

PETRUSHEV, P.P. and POPOV, V.A. (1987). *Rational Approximation of Real Functions.* Cambridge University Press, Cambridge.

RALSTON, A. and RABINOWITZ, P. (1978). *A First Course in Numerical Analysis.* McGraw-Hill, New York.

SENDOV, BL. and POPOV, V.A. (1988). *The Averaged Moduli of Smoothness: Applications in Numerical Methods and Approximation.* Wiley, New York.

– – – – (1976). *Numerical Methods, vol. 1.* Nauka i Izkustvo. (In Bulgarian).

TRIEBEL, H. (1983). *Theory of Function Spaces.* Monographs in Mathematics, **78**, Birkhäuser Verlag, Basel.

– – – – (1992). *Theory of Function Spaces II.* Monographs in Mathematics, **84**, Birkhäuser Verlag, Basel.

LUBOMIR DECHEVSKY
DÉPARTEMENT DE MATHÉMATIQUES
ET DE STATISTIQUE
UNIVERSITÉ DE MONTRÉAL
C. P. 6128, SUCCURSALE A
MONTRÉAL, QUÉBEC, CANADA H3C 3J7
Email: dechevsk@dms.umontreal.ca

BRENDA MACGIBBON
DÉPARTEMENT DE MATHÉMATIQUES
UNIVERSITÉ DU QUÉBEC À MONTRÉAL
C. P. 8888, SUCCURSALE CENTRE-VILLE
MONTRÉAL, QUÉBEC, CANADA H3C 3P8
Email: brenda@math.uqam..ca

SPIRIDON PENEV
DEPARTMENT OF STATISTICS
SCHOOL OF MATHEMATICS
THE UNIVERSITY OF NEW SOUTH WALES
SYDNEY NSW 2052 AUSTRALIA
Email: spiro@maths.unsw.edu.au