Asymptotically minimax estimation of a function with jumps

C.G.M. Oudshoorn *

January 9, 1997

Abstract

Asymptotically minimax nonparametric estimation of a regression function observed in white Gaussian noise over a bounded interval is considered, with respect to a $\mathbf{L}_2$-loss function. The unknown function f is assumed to be m times differentiable except for an unknown, though finite, number of jumps, with piecewise mth derivative bounded in $\mathbf{L}_2$-norm. An estimator is constructed, attaining the same optimal risk bound, known as Pinsker's constant, as in the case of smooth functions (without jumps).

Key words: Jump-point estimation; Nonparametric regression; Optimal constant; Tapered orthogonal series estimator

1 Introduction

In the eighties optimal rates of convergence in nonparametric regression estimation problems have been thoroughly examined, following the book of Ibragimov and Hasminskii (1981), the ground-breaking papers due to Stone (1982) and Birgé (1983) and others. Later the interest has shifted to finding not only the optimal rates, but also the asymptotic optimal constants, determining the risk of optimal estimators. This interest was greatly initiated by the pioneering paper of Pinsker (1980). Such results have been obtained for different observation schemes, involving smooth functions confined to balls in a Sobolev space, with $\mathbf{L}_2$-losses (Nussbaum (1985), Golubev and Nussbaum (1990), Speckman (1985), Efroimovich (1994), etc.); for functions restricted to Hölder balls in case of

*Institute for Business and Industrial Statistics, University of Amsterdam, the Netherlands karino@fwi.uva.nl

$\mathbf{L}_\infty$-losses (Korostelev (1994), Donoho (1994)); for analytic functions with different types of loss functions (Golubev, Levit and Tsybakov (1995)).

Recently there is a growing interest in estimating functions with isolated singularities, stimulated by a variety of applications, like change-point problems, spatially inhomogeneous data, image fragments restoration as well by their mathematical meaningfulness (see Müller (1992) and further references therein). At first glance, the known results on the optimal rates of convergence may suggest that only slow rates of convergence can be achieved for such functions with discontinuities. This however turns out not to be the case. For functions which are smooth except for a few points the discontinuities do not affect the convergence rate of the optimal estimators (Hall and Patil (1995)).

In this paper we show that, in the Sobolev-type setting with $\mathbf{L}_2$-losses, regardless of the presence of an unknown, although finite, number of jumps in the unknown regression function even the same asymptotic optimal constant can be attained as in the case of smooth functions without jumps. This optimal constant is known as Pinsker's constant.

In order to obtain this optimal result we need accurate estimators of the jump-points. To detect the number and location of the jumps we construct an estimator depending on a sample version of the jump sizes $f(x+) - f(x-)$. For a similar kind of estimator we refer to (among others) Müller (1992), where boundary kernels are used for estimating the location of a jump and its size. In Wang (1995) jumps are detected using wavelets.

We will work with the Gaussian white noise model, as will be defined below in (2.1). Signal recovery in Gaussian white noise with variance tending to zero has served for already some time as a representative model for non-parametric curve estimation, having all the essential traits in a pure form. In contrast in particular with the nonparametric regression model, with observations on a discrete grid, it entails minimal technical nuisance. This is reflected by the fact that, roughly speaking for corresponding derivations one makes in these models, in the discrete model one has to deal with summations and in the Gaussian model with integrals, hence the latter gives more elegant and transparent arguments. We conjecture that an approximation in the sense of Le Cam's deficiency distance should make it precise. The models are then asymptotically equivalent for all purposes of statistical decision with bounded loss. A first result of this kind has recently been established by Brown and Low (1990). They proved that the nonparametric regression model with observations on a discrete grid is asymptotically equivalent with the Gaussian white noise model for Hölder classes with smoothness $\beta > 1/2$. Although our results concern regression for Sobolev classes with discontinuities we might expect from their work that similar results should hold also in our context. Hence the results obtained in this paper should hold also in the discrete type model, but that requires even more technical proofs. Moreover it gives the idea what the corresponding asymptotically minimax estimator is in that model.

2 The model and main result

Suppose we observe a random process $X_\varepsilon(t)$ satisfying the stochastic differential equation

$$dX_\varepsilon(t) = f(t)\,dt + \varepsilon\,d\mathcal{W}(t) \qquad\qquad t \in [0,1], \tag{2.1}$$

with some prescribed initial value $X(0)$, which is either constant or a random variable independent of $\mathcal{W}$, $\mathcal{W}(t)$ is a standard Wiener process and ε is a known parameter, assumed to be small. The space of square-integrable functions (or signals) on $[0,1]$ is denoted by $\mathbf{L}_2 = \mathbf{L}_2[0,1]$ and $\|\cdot\|$ is the usual $\mathbf{L}_2$-norm.

Let m be a positive integer and $Q > 0$ a constant, both given. Assume that the unknown signal f belongs to a class of functions $\Phi_{B,\kappa,L}(m,Q)$, for which there exist (not necessarily known) positive constants B, κ, L such that for all $f \in \Phi_{B,\kappa,L}(m,Q)$:

A1 $\sup_{t\in[0,1]} |f(t)| < B$

A2 there exist jump-points b_j, $j = 0,\ldots,S+1$ (unknown) , $0 = b_0 < \cdots < b_{S+1} = 1$, such that

A2.1 the points are at least distance κ apart;

A2.2 the jumps have at least size L, that is,

$$L \leq |\lim_{t\downarrow b_j} f(t) - \lim_{t\uparrow b_j} f(t)| =: |f(b_j+) - f(b_j-)| \qquad\qquad j = 1,\ldots,S;$$

A2.3 f is m-times differentiable on $]\,b_j, b_{j+1}[$, $j = 0,\ldots,S$;

A2.4 f belongs to a piecewise Sobolev 'ellipsoid', that is,

$$\sum_{j=0}^{S} \int_{b_j}^{b_{j+1}} (f^{(m)}(t))^2 \, dt \leq Q.$$

Note that existence of the limits in A2.2 follows from A2.4. Moreover, conditions A1 and A2.4 imply that the derivatives $f^{(i)}(t)$, $i = 1, \ldots, m-1$ are bounded everywhere except at the jump-points b_j. Obviously the functions in $\Phi_{B,\kappa,L}(m, Q)$ are square-integrable.

Define the quadratic risk for an estimator $\widehat{f}$ of f as follows

$$R\,(\widehat{f}, f) = \mathbf{E}_f \|\widehat{f} - f\|^2. \tag{2.2}$$

Furthermore denote the minimax quadratic risk in estimating f with respect to the class of functions Φ by

$$r_\varepsilon(\Phi) = \inf_{\widehat{f}} \sup_{f \in \Phi} R\,(\widehat{f}, f), \tag{2.3}$$

where the infimum is taken over all estimators $\widehat{f}$.

We derive the exact asymptotic behaviour of this minimax quadratic risk for the class $\Phi_{B,\kappa,L}(m, Q)$, described by the conditions A1 - A2.4, formulated in the following theorem

Theorem 2.1 *Let $\gamma(m, Q) = (Q(2m+1))^{1/(2m+1)} \left(\frac{m}{\pi(m+1)}\right)^{2m/(2m+1)}$. The minimax quadratic risk of the above defined model satisfies*

$$\lim_{\varepsilon \downarrow 0} \varepsilon^{\frac{-4m}{2m+1}} r_\varepsilon(\Phi_{B,\kappa,L}(m, Q)) = \gamma(m, Q), \tag{2.4}$$

for arbitrary though fixed B, κ and L.

Notice that the right hand-side of (2.4) is independent of B, κ and L. The proof of this theorem is outlined in Section 2.4. There also an projection-type estimator $\widetilde{f}$ is given which attains this optimal constant $\gamma(m, Q)$, Pinsker's constant.

For the lower bound on the minimax risk in the setting (2.1) we refer to Pinsker (1980), and to a more recent paper by Belitser and Levit (1996) for the corresponding discrete setting. Note that our model allows functions f without jumps, while the additional restriction A -1 does not affect the technique used in these papers for obtaining the lower bounds in estimating such functions.

First in Section 2.1 we give the motivation of the chosen basis. Second in Section 2.2 we generalise the ideas of Pinsker (1980), such that an asymptotically minimax estimator $\widehat{f}(t) = \widehat{f}(t; b_1, \ldots, b_S)$, in the case the number and locations of the jump-points are known, is obtained. Of course this estimator depends on the (unknown) jump-points b_j. Then we estimate these jump-points (see Section 2.3). Finally in Section 2.4 we argue that replacement, in the estimator $\widehat{f}$ for the case of known jump-points, of the unknown jump-points b_j by their estimators, results in an asymptotically minimax estimator $\widetilde{f}$ in the case when the jump-points are unknown (cf. (2.20)). We stress again that we do not have to know the number of jump-points, but due to the choice of κ this number is assumed to be finite.

2.1 Choice of the basis

Let us consider the basis arising from the following boundary value problem on the interval $[a, b]$, with Neumann conditions on the boundary

$$\begin{cases} (-\mathrm{d}^2/\mathrm{d}t^2)^m u = \lambda u & t \in [a, b] \\ u^{(s)}(a) = u^{(s)}(b) = 0 & s = m, \ldots, 2m-1. \end{cases} \tag{2.5}$$

The corresponding differential operator $L = (-\mathrm{d}^2/\mathrm{d}t^2)^m$ is self-adjoint and semi positive-definite on the space $D = \{u \in C^{2m}[a, b] : u^{(s)}(a) = u^{(s)}(b) = 0 \text{ for } s = m, \ldots, 2m-1\}$. Furthermore the null-space D_0 of L is spanned by the polynomials of degree at most $m-1$. On $D \backslash D_0$, L is compactly invertible and positive definite. Hence due to the spectral theorem the normed eigenfunctions $\varphi_m, \varphi_{m+1}, \ldots$ of (2.5) corresponding to positive eigenvalues $\lambda_m, \lambda_{m+1}, \ldots$ and supplemented with an orthonormal set of polynomials $\varphi_0, \ldots, \varphi_{m-1}$ on $[a, b]$ with degree at most $m-1$, with corresponding eigenvalues $\lambda_0 = 0, \ldots, \lambda_{m-1} = 0$, provide a basis for D. Moreover it is an orthonormal basis for the whole of $\mathbf{L}_2[a, b]$, as D is dense in $\mathbf{L}_2[a, b]$ (cf. e.g. Coddington and Levinson (1955), Ch. 7).

The statistical importance of the eigenfunctions of the boundary value problem (2.5), in the Sobolev-type setting, was apparently first recognized by Golubev and Nussbaum (1990). Its discrete counterpart, the so called Demmler-Reinsch basis, appeared in the nonparametric methods based on splines even earlier (see e.g. Speckman (1985)).

In the theory of differential equations extensive study of the asymptotic behaviour of the eigenvalues and eigenfunctions of the boundary value problem (2.5), usually in the broader setting of general linear differential operators of order m, was initiated by G.D.Birkhoff in 1908. A detailed description of these properties is presented in the following monographs: Neumark (1967), Sect. II.4; Dunford and Schwartz (1971), Sect. XIX.

The particular form of (2.5) allows a more straightforward and detailed account of these properties. Prof. J.J. Duistermaat from the University of Utrecht kindly agreed to review these properties at our request. In his recent report (1995) sharper asymptotics for the eigenvalues and eigenfunctions of the boundary value problem (2.5), as compared to those presented in the references above, are obtained and a further study of the asymptotic properties of the corresponding eigenfunctions is made. Especially the sharper asymptotics of the eigenfunctions is necessary for proving our result.

Here we summarize the results we use further on:

Theorem 2.2 *(cf. Duistermaat (1995)) The non-zero eigenvalues λ_k of the boundary value problem (2.5) satisfy the relation*

$$\lambda_k = \rho_k^{2m} = (k\,\pi(b-a)^{-1})^{2m}(1+o(1)) \qquad\qquad k \to \infty \tag{2.6}$$

and the corresponding eigenfunctions φ_k equal (for coefficients A_k, B_k such that they are orthonormalised)

$$A_k(\cos(\rho_k t) + r_k(t)),\ \textit{for } k \textit{ odd, and } B_k(\sin(\rho_k t) + r_k(t)) \textit{ for } k \textit{ even,}$$

where the functions $r_k(t)$ satisfy

$$|\, r_k^{(i)}(t)| \le C_{i,m}\rho_k^i \left(e^{-\rho_k\alpha_m(t-a)} + e^{-\rho_k\alpha_m(b-t)}\right),$$

for any i and some constants $C_{i,m}$ depending on i, m and $(b-a)$ and $\alpha_m > 0$ depending only on m.

Note that for $m = 1$ the eigenvalues equal $(k\pi/(b-a))^2$ and the eigenfunctions are exactly cosines.

Observe that the functions $\{(\lambda_k)^{-1/2}\varphi_k^{(m)}(\cdot)\}_{k\geq m}$ are also orthonormal. Therefore for a function $f \in C^m[a,b]$ satisfying

$$\int_a^b \left(f^{(m)}(t)\right)^2 dt \leq Q,$$

we can derive by partial integration and Bessel's inequality that the Fourier coefficients $\theta_k = \int_a^b f(t)\varphi_k(t)\,dt$ belong to the ellipsoid

$$\sum_{k=m}^{\infty} \lambda_k \theta_k^2 = \sum_{k=0}^{\infty} \lambda_k \theta_k^2 \leq Q. \tag{2.7}$$

We note that in Pinsker (1980), among other things, an optimal estimator is constructed for signals with a restriction on Fourier coefficients of precisely this type. Our goal here is to develop a similar but broader technique, incorporating the piece-wise smooth functions.

2.2 Optimal estimation with known jump-points

Given relation (2.7) we follow and generalise the ideas of Pinsker in deriving an optimal estimator of f (cf. also Belitser and Levit (1996)). Denote the corresponding orthonormal eigenvalues and eigenfunctions of the equation (2.5) on the interval $[b_j, b_{j+1}]$ by $\lambda_{k,j} = \lambda_k(b_{j+1}-b_j)^{-2m}$ and $\varphi_{k,j}(t) = (b_{j+1}-b_j)^{-1/2}\varphi_k((b_{j+1}-b_j)^{-1}(t-b_j))$ for $j = 0,\ldots,S$, where φ_k and λ_k denote here the corresponding quantities related to the standard interval $[0,1]$.

Let us emphasize the fact that the jump-points b_1 up to b_S are unknown. However pretend for a moment that they are known. Rewrite the observation process (2.1) into the following equivalent sequence of observations, $(j = 0,\ldots,S)$,

$$Y_{k,j} = \theta_{k,j} + \varepsilon\,\xi_{k,j}, \qquad\qquad k = 0,1,\ldots, \tag{2.8}$$

with

$$Y_{k,j} = \int_{b_j}^{b_{j+1}} \varphi_{k,j}(t)\,dX_\varepsilon(t), \qquad\qquad \theta_{k,j} = \int_{b_j}^{b_{j+1}} f(t)\varphi_{k,j}(t)\,dt$$

and

$$\xi_{k,j} = \int_{b_j}^{b_{j+1}} \varphi_{k,j}(t)\,d\mathcal{W}(t).$$

Here $\xi_{1,0},\ldots,\xi_{1,S},\xi_{2,0},\ldots,\xi_{2,S},\ldots$ are independent and standard Gaussian random variables.

Consider the following tapered orthogonal series estimator of f:

$$\widehat{f}(t)=\sum_{k=0}^{\infty}\overline{\theta}_{k,j}\varphi_{k,j}(t), \qquad t\in[b_j,b_{j+1}[, \qquad j=0,\ldots,S \tag{2.9}$$

(the last subinterval is $[b_S,1]$ instead of $[b_S,1[$), where

$$\overline{\theta}_{k,j}=h_{k,j}Y_{k,j}, \qquad 0\le h_{k,j}\le 1,$$

is an estimate of $\theta_{k,j}$ and the tapering coefficients $h_{k,j}$ are chosen such that

$$h_{k,j}=\begin{cases}1 & k=0,\ldots,m-1\\ (1-c_j\lambda_{k,j}^{1/2})_+ & k\ge m,\end{cases} \tag{2.10}$$

with c_j the solutions of the equations

$$c_j(b_{j+1}-b_j)\,Q=\varepsilon^2\sum_{k=0}^{\infty}\lambda_{k,j}^{1/2}(1-c_j\lambda_{k,j}^{1/2})_+. \tag{2.11}$$

These coefficients are shown below to be optimal in the sense that the maximal quadratic risk of the pseudo-estimator $\widehat{f}$ over $\Phi_{B,\kappa,L}(m,Q)$ attains asymptotically the optimal bound (2.4). In order to explain this observe first that the risk $R(\widehat{f},f)$, defined in (2.2), of an estimator $\widehat{f}$ of the form (2.9) with arbitrary tapering coefficients $h_{k,j}$ is equal to

$$\sum_{j=0}^{S}\sum_{k=0}^{\infty}\mathbf{E}_f(\overline{\theta}_{k,j}-\theta_{k,j})^2,$$

or written out

$$\sum_{j=0}^{S}\sum_{k=0}^{\infty}(\varepsilon^2h_{k,j}^2+\theta_{k,j}^2(h_{k,j}-1)^2).$$

Therefore the maximal quadratic risk of $\widehat{f}$ over $\Phi_{B,\kappa,L}(m,Q)$ for arbitrary tapering coefficients $h_{k,j}$ can be bounded as follows (cf. (2.7))

$$\begin{aligned}\sup_{f\in\Phi_{B,\kappa,L}(m,Q)}R(\widehat{f},f)&=\sup_{f\in\Phi_{B,\kappa,L}(m,Q)}\sum_{j=0}^{S}\sum_{k=0}^{\infty}(\varepsilon^2h_{k,j}^2+\theta_{k,j}^2(h_{k,j}-1)^2)\\ &\le\sup_{\{\theta_{k,j}:\sum_j\sum_k\lambda_{k,j}\theta_{k,j}^2\le Q\}}\sum_{j=0}^{S}\sum_{k=0}^{\infty}(\varepsilon^2h_{k,j}^2+\theta_{k,j}^2(h_{k,j}-1)^2)\\ &\le\sup_{\{Q_j:\sum_{j=0}^{S}Q_j=Q\}}\sup_{\{\theta_{k,j}:\sum_k\lambda_{k,j}\theta_{k,j}^2\le Q_j\}}\sum_{j=0}^{S}\sum_{k=0}^{\infty}(\varepsilon^2h_{k,j}^2+\theta_{k,j}^2(h_{k,j}-1)^2)\\ &\le\sup_{\{Q_j:\sum_{j=0}^{S}Q_j=Q\}}\sum_{j=0}^{S}(\varepsilon^2\sum_{k=0}^{\infty}h_{k,j}^2+Q_j\sup_{k\ge m}\lambda_{k,j}^{-1}(h_{k,j}-1)^2).\end{aligned}$$

Taking the infimum over all tapering coefficients $0 \le h_{k,j} \le 1$ for $k \ge m$ in the last expression we see that the minimax risk in estimating f, as defined in (2.3), is bounded by

$$r_\varepsilon(\Phi_{B,\kappa,L}(m,Q)) \le \inf_{c_j} \inf_{\{h_{k,j}:(h_{k,j}-1)^2 \le c_j^2 \lambda_{k,j}\}} \sup_{Q_j} \sum_{j=0}^{S} (\varepsilon^2 \sum_{k=m}^{\infty} h_{k,j}^2 + c_j^2 Q_j) + (S+1)\, m\varepsilon^2. \tag{2.12}$$

If the class $\Phi_{B,\kappa,L}(m,Q)$ comprises only those functions f with given values of $Q_0,\ldots,Q_S$ (i.e. the distribution of the power of the signal for the intervals $[b_j, b_{j+1}[$ is known) then it is not difficult to see (cf. Belitser and Levit (1996)) that the optimal tapering coefficients $h_{k,j}$ in (2.12) are given by (2.10) and

$$c_j Q_j = \varepsilon^2 \sum_{k=m}^{\infty} \lambda_{k,j}^{1/2} (1 - c_j \lambda_{k,j}^{1/2})_+ \tag{2.13}$$

(cf. (2.11)). However since the Q_j are not known we replace them by

$$\widetilde{Q_j} = (b_{j+1} - b_j)\, Q \tag{2.14}$$

in relation (2.13), i.e., we presume for the moment that the 'worst' possible distribution of the signals power Q_j over the intervals $[b_j, b_{j+1}[$ occurs when the Q_j are proportional to their lengths (constant power per unit of time). This hypothesis can be easily backed up by an elementary calculation. In Section 3 we show that, with c_j and $h_{k,j}$ defined above, asymptotically the c_j do not depend on j (cf. (3.13)) and

$$\sup_{\{Q_j:\sum_{j=0}^{S} Q_j = Q\}} \sum_{j=0}^{S} (\varepsilon^2 \sum_{k=m}^{\infty} h_{k,j}^2 + c_j^2 Q_j) \;=\; \gamma(m,Q)\, \varepsilon^{\frac{4m}{2m+1}} (1 + o(1)), \qquad \varepsilon \to 0. \tag{2.15}$$

Denote therefore from now on c_j by c.

Thus the estimator $\widehat{f}$ is an asymptotically minimax estimator, in the case when the jump-points are known, that is,

$$\sup_{f \in \Phi_{B,\kappa,L}(m,Q)} R(\widehat{f}, f) \le \gamma(m,Q)\, \varepsilon^{\frac{4m}{2m+1}} (1 + o(1)) \qquad \varepsilon \to 0. \tag{2.16}$$

To summarize: using the method of Pinsker on each of the subintervals $[b_j, b_{j+1}[\subset [0,1]$ we obtain, in the case of *known* jump-points, the same rate and even the same constant as in the case *without* jump-points.

However the jump-points on which the construction of the estimator $\widehat{f}$ above heavily depends (note that all the quantities involved, namely $h_{k,j}$, $\varphi_{k,j}$, and $\lambda_{k,j}$ in the above relations (2.5), (2.8), (2.10), (2.13) and (2.14) depend on them) are actually not known.

Our next goal is to show that combining the method described above with the use of sufficiently accurate estimators $\hat{b}_j$ of b_j will enable us to obtain asymptotically the same quality of estimation as in (2.16). Therefore we now explain how we estimate the (number of) jump-points.

2.3 Estimation of the jump-points

There exist estimators of the jump-points that converge with rate ε^2 to the true jump-points (cf. Korostelev (1987)). However we define here jump-point estimators which are easy to calculate and are nearly optimal, but converge fast enough to the true jump-points for our purposes.

In the sequel the bandwidth parameter h equals $\varepsilon^2(\ln \varepsilon^{-2})^{1+\delta}$ for an arbitrary but fixed $\delta > 0$. Furthermore suppose ε is small enough such that $2h < \kappa$. Define a set of grid-points, separated at distance ε^2 by

$$A = \{a_i = i\varepsilon^2, i = \lceil h\varepsilon^{-2} \rceil, \ldots, \lfloor (1-h)\varepsilon^{-2} \rfloor\}$$

on the interval $[h, 1-h]$ and let N be the number of elements of A. For every grid-point $a_i \in A$ we calculate the quantity

$$T(a_i) = h^{-1} \int\limits_{a_i - h}^{a_i} dX_\varepsilon(t) - h^{-1} \int\limits_{a_i}^{a_i+h} dX_\varepsilon(t).$$

We expect that $|T(a_i)|$ is 'large' only if there is a jump-point b_j in the interval $[a_i - h, a_i + h]$. More precisely the following result will be proved in Section 3)

Lemma 2.3 *For every $\gamma > 0$ and $0 < \alpha < \delta/2$, uniformly in $f \in \Phi_{B,\kappa,L}(m, Q)$*

(i)

$$\mathbf{P}_f\{\,|\,T(a_i)\,| > (\ln \varepsilon^{-2})^{-\alpha}\} = o(\varepsilon^\gamma) \qquad \varepsilon \to 0,$$

for every grid-point $a_i \in A$ such that there are no jump-points b_j in the interval $[a_i - h, a_i + h]$;

(ii)

$$\mathbf{P}_f\{\,|T(a_i)\,| < (\ln\varepsilon^{-2})^{-\alpha}\} = o(\varepsilon^{\gamma}) \qquad\qquad \varepsilon\to 0$$

for every grid-point $a_i \in A$ such that for some $j = 1,\dots S$, $|a_i - b_j| < h/2$.

However it is possible that there are more than one grid-points a_i 'near' a jump-point b_j that have large value $|T(a_i)|$. Thus we have to classify those a_i detecting the same b_j.

Fix an α between 0 and $\delta/2$. Neighbouring grid-points a_i and a_{i+1} are called connected if both $|T(a_i)|$ and $|T(a_{i+1})|$ exceed $(\ln\varepsilon^{-2})^{-\alpha}$. Consider all subsets of A, consisting of more than $\lceil(\log\varepsilon^{-2})^{1+\delta}\rceil$ connected grid-points and denote these subsets (or so-called cluster sets) by T_j, $j = 1,\dots,\widetilde{S}$. We use here the convention that if such subsets do not exist we put $\widetilde{S} = 0$. Loosely speaking these sets contain the candidates for jump-points. Finally we define estimator(s) $\hat{b}_j$ of possible jump-points as follows

$$\hat{b}_j = \underset{a_i\in T_j}{\arg\max}\{\,|T(a_i)|\,\} \qquad\qquad j = 1,\dots,\widetilde{S}, \tag{2.17}$$

and $\hat{b}_0 = 0$, $\hat{b}_{\widetilde{S}+1} = 1$. Notice that if $\widetilde{S}$ is bigger than S then at least for one a_i there is no jump b_j in the interval $[a_i - h, a_i + h]$ and nevertheless $|T(a_i)| > (\ln\varepsilon^{-2})^{-\alpha}$. Fortunately due to Lemma 2.3(i), with $\gamma = 2 + 4m/(2m+1)$, this happens only with small probability, namely

$$\mathbf{P}_f\{\,S < \widetilde{S}\,\} \le \sum_{a_i:\text{ no jump in }[a_i-h,a_i+h]} \mathbf{P}_f\{\,|T(a_i)\,| > (\ln\varepsilon^{-2})^{-\alpha}\} = o((\varepsilon^2)^{\frac{2m}{2m+1}}). \tag{2.18}$$

In the case $\widetilde{S} < S$ there is at least one b_j such that there exists an a_i within distance $h/2$ of b_j having $|\,T(a_i)\,|$-value smaller than $(\ln\varepsilon^{-2})^{-\alpha}$. Applying Lemma 2.3(ii) with $\gamma = 4m/(2m+1)$ yields

$$\mathbf{P}_f\{\,S > \widetilde{S}\,\} \le \sum_{b_j} \mathbf{P}_f\{\,|\,T(a_i)\,|\, < (\ln\varepsilon^{-2})^{-\alpha}\} = o((\varepsilon^2)^{\frac{2m}{2m+1}}). \tag{2.19}$$

Indeed with high probability the right number of jump-points is estimated. Notice that since $2h < \kappa$, the minimal distance of the jump-points, it can not happen that two real jump-points are 'seen' as one. The following bound concerning the accuracy of estimating b_j is frequently used in the sequel

Corollary 2.4 *Uniformly in $f \in \Phi_{B,\kappa,L}(m,Q)$ we have*

$$\mathbf{E}_f|\hat{b}_j - b_j|^2 \le O\left((\varepsilon^2(\ln\varepsilon^{-2})^{1+\delta})^2\right) \qquad \varepsilon \to 0.$$

Observe that the above expectation is obviously bounded by $(2h)^2 + \mathbf{P}_f\{|\hat{b}_j - b_j| > 2h\}$. Use Lemma 2.3 for finishing the proof of this corollary.

2.4 The proposed estimator and the framework of the proof

Substitute $\hat{b}_j$ and $\hat{b}_{j+1}$ in relations (2.5) and (2.8). Denote the eigenfunctions $\varphi_k(\cdot, \hat{b}_j, \hat{b}_{j+1})$ by $\widehat{\varphi}_{k,j}$ and $\lambda_{k,j}(\hat{b}_j, \hat{b}_{j+1})$ by $\widehat{\lambda}_{k,j}$. The observation process (2.1) can be rewritten into the sequence of observations

$$\widehat{Y}_{k,j} = \widehat{\theta}_{k,j} + \varepsilon\,\widehat{\xi}_{k,j}, \qquad k = 0, 1, \dots,$$

where

$$\widehat{Y}_{k,j} = \int_{\hat{b}_j}^{\hat{b}_{j+1}} \widehat{\varphi}_{k,j}(t)\, dX_\varepsilon(t), \qquad \widehat{\theta}_{k,j} = \int_{\hat{b}_j}^{\hat{b}_{j+1}} f(t)\widehat{\varphi}_{k,j}(t)\, dt$$

and

$$\widehat{\xi}_{k,j} = \int_{\hat{b}_j}^{\hat{b}_{j+1}} \widehat{\varphi}_{k,j}(t)\, d\mathcal{W}(t).$$

Furthermore substitute $\hat{b}_j$ and $\hat{b}_{j+1}$ also in relations (2.10), (2.11)) and (2.14). As suggested in the discussion of Sect. 2.2 we propose the following estimator $\widetilde{f}$ for the unknown regression function f

$$\widetilde{f}(t) = \sum_{k=0}^{\infty} \widetilde{\theta}_{k,j}\widehat{\varphi}_{k,j}(t), \qquad t \in [\,\hat{b}_j, \hat{b}_{j+1}[, \qquad j = 0, \dots, \widetilde{S}, \tag{2.20}$$

where the estimator $\widetilde{\theta}_{k,j}$ is equal to $\widehat{h}_{k,j}\widehat{Y}_{k,j}$ with the (estimated) tapering coefficients $\widehat{h}_{k,j}$ defined by the relations

$$\widehat{h}_{k,j} = (1 - (c\widehat{\lambda}_{k,j})^{1/2})_+, \qquad k = 0, 1, \dots, \qquad j = 0, \dots, \widetilde{S},$$

and c is given by (3.13) below (note that again the first m tapering coefficients equal 1).

As the regression function is supposed to be bounded by some (unknown) constant B we continue the proof with the following truncated version of $\widetilde{f}$:

$$\widetilde{f}_{tr}(t) = \begin{cases} -B(\varepsilon) & \widetilde{f} < -B(\varepsilon) \\ \widetilde{f}(t) & |\,\widetilde{f}\,| \le B(\varepsilon) \\ B(\varepsilon) & \widetilde{f} > B(\varepsilon). \end{cases}$$

Here $B(\varepsilon)$ is a sequence tending 'slowly' to infinity, e.g. $B(\varepsilon) = (\ln \varepsilon^{-2})$, thus for ε small enough it exceeds B. The risk of this estimator is bounded by $(2B(\varepsilon))^2$. Due to inequalities (2.18) and (2.19) we know that the probability that S is not equal to $\widetilde{S}$ is sufficiently small. Namely,

$$\mathbf{P}_f\,\{S \ne \widetilde{S}\,\} \;=\; o((\varepsilon^2)^{\frac{2m}{2m+1}}),.$$

as ε tends to zero. Moreover we have according to Corollary 2.4

$$\mathbf{E}_f|\hat{b}_j - b_j|^2 \le O\left((\varepsilon^2(\ln\varepsilon^{-2})^{1+\delta})^2\right) \;=\; o((\varepsilon^2)^{\frac{2m}{2m+1}}), \qquad \varepsilon \to 0. \tag{2.21}$$

Combining these two facts we see that the complement of the event $F = \{\,S = \widetilde{S}\} \cap \{\,|\hat{b}_j - b_j| < 2h, j = 1, \ldots, S\}$ has small probability. Thus we derive for the risk of $\widetilde{f}_{tr}$, provided ε is small enough,

$$\begin{aligned} R\,(\widetilde{f}_{tr}, f) &= \mathbf{E}_f\|\widetilde{f}_{tr} - f\,\|^2\mathbf{I}_F + \mathbf{E}_f\|\widetilde{f}_{tr} - f\,\|^2\mathbf{I}_{F^c} \\ &\le \mathbf{E}_f\|\widetilde{f} - f\,\|^2\mathbf{I}_F + o((\varepsilon^2)^{\frac{2m}{2m+1}}) \qquad\qquad \varepsilon \to 0. \end{aligned}$$

Hence it suffices to restrict ourselves to the case S equals $\widetilde{S}$ and we can assume that $\hat{b}_j$ estimates b_j within distance $2h$. From now on we take expectations conditioned on the event F without to mention. Applying Parseval's equality we derive for the risk of $\widetilde{f}$

$$\begin{aligned} R\,(\widetilde{f}, f) &= \sum_{j=0}^{S}\sum_{k=0}^{\infty} \mathbf{E}_f(\widetilde{\theta}_{k,j} - \widehat{\theta}_{k,j})^2 \\ &= \sum_{j=0}^{S}\sum_{k=0}^{\infty} \mathbf{E}_f(\varepsilon\widehat{h}_{k,j}\widehat{\xi}_{k,j} - (1 - \widehat{h}_{k,j})\,\widehat{\theta}_{k,j})^2 \\ &= \sum_{j=0}^{S} \mathbf{E}_f\,\widehat{L}_{\varepsilon,j}(f) + R_{1,j} + R_{2,j} - R_{3,j}, \end{aligned}$$

where

- $\widehat{L}_{\varepsilon,j}(f) = \sum_{k\geq m} (1-\widehat{h}_{k,j})^2 \widehat{\theta}_{k,j}^2 + \varepsilon^2 \widehat{h}_{k,j}^2$
- $R_{1,j} = R_{1,j}(f, \hat{b}_j, \hat{b}_{j+1}) = \varepsilon^2 \sum_{k=0}^{m-1} \mathbf{E}_f \widehat{\xi}_{k,j}^2$
- $R_{2,j} = R_{2,j}(\hat{b}_j, \hat{b}_{j+1}) = \varepsilon^2 \sum_{k\geq m} \mathbf{E}_f \widehat{h}_{k,j}^2 (\widehat{\xi}_{k,j}^2 - 1)$
- $R_{3,j} = R_{3,j}(f, \hat{b}_j, \hat{b}_{j+1}) = 2\varepsilon \sum_{k\geq m} \mathbf{E}_f (1-\widehat{h}_{k,j}) \widehat{h}_{k,j} \widehat{\theta}_{k,j} \widehat{\xi}_{k,j}$.

For the remainder terms $R_{l,j}$, $(l = 1,2,3)$, we deduce in Section 3 the following bounds for ε tending to 0, all uniformly for $f \in \Phi_{B,\kappa,L}(m,Q)$,

$$R_{1,j}(f, \hat{b}_j, \hat{b}_{j+1}) = o(\varepsilon^2 \ln^3(\varepsilon^{-2})), \tag{2.22}$$

$$R_{2,j}(f, \hat{b}_j, \hat{b}_{j+1}) = o(1)\, \mathbf{E}_f \widehat{L}_{\varepsilon,j}(f) + O\left(\varepsilon^2 \ln^3(\varepsilon^{-2})\right) \tag{2.23}$$

and

$$|\, R_{3,j}(f, \hat{b}_j, \hat{b}_{j+1}) \,| = o(1)\, \mathbf{E}_f \widehat{L}_{\varepsilon,j}(f) + O\left(\varepsilon^2 \ln^3(\varepsilon^{-2})\right). \tag{2.24}$$

This implies uniformly in $f \in \Phi_{B,\kappa,L}(m,Q)$

$$R(\widetilde{f}, f) = (1+o(1)) \sum_{j=0}^{S} \mathbf{E}_f \widehat{L}_{\varepsilon,j}(f) + O\left(\varepsilon^2 \ln^3(\varepsilon^{-2})\right), \qquad \varepsilon \to 0.$$

Denote by $L_{\varepsilon,j}(f)$ the quantity

$$L_{\varepsilon,j}(f) = \sum_{k=m}^{\infty} (1-h_{k,j})^2 \theta_{k,j}^2 + \varepsilon^2 h_{k,j}^2,$$

with $h_{k,j}$ defined in (2.10) and (2.11). The next step in the proof is to show that, uniformly in $f \in \Phi_{B,\kappa,L}(m,Q)$, we have for $j = 0,\ldots,S$

$$\mathbf{E}_f \widehat{L}_{\varepsilon,j}(f) \leq L_{\varepsilon,j}(f)(1+o(1)) + o(\varepsilon^{\frac{4m}{2m+1}}), \qquad \varepsilon \to 0, \tag{2.25}$$

(cf. Section 3). Thus for the maximal quadratic risk of $\widetilde{f}$ we have

$$\sup_{f\in\Phi_{B,\kappa,L}(m,Q)} R(\widetilde{f}, f) \leq \sup_{\Phi_{B,\kappa,L}(m,Q)} \sum_{j=0}^{S} L_{\varepsilon,j}(f)(1+o(1)) + o(\varepsilon^{\frac{4m}{2m+1}}), \qquad \varepsilon \to 0. \tag{2.26}$$

Finally note that in Section 2.2 we have explained that

$$\sup_{\Phi_{B,\kappa,L}(m,Q)} \sum_{j=0}^{S} L_{\varepsilon,j}(f) \leq \gamma(m,Q)\, \varepsilon^{\frac{4m}{2m+1}} (1+o(1)), \qquad \varepsilon \to 0. \tag{2.27}$$

Combination of (2.26) and (2.27) finishes the proof of Theorem 2.1.

3 Proofs

Proof of Lemma 2.3

Suppose $a_i \in A$, the set of grid-points, is such that there are no jump-points b_j in the interval $[a_i - h, a_i + h]$. For any such grid-point we bound the bias of $T(a_i)$, uniformly in $f \in \Phi_{B,\kappa,L}(m, Q)$, as follows

$$|\,\mathbf{E}_f T(a_i)\,| = h^{-1}\Big|\int\limits_{a_i-h}^{a_i} f(t)\,dt - \int\limits_{a_i}^{a_i+h} f(t)\,dt\Big| = \text{const}\cdot h \;\le 1/2(\ln\varepsilon^{-2})^{-\alpha}, \tag{3.1}$$

provided ε is sufficiently small. $T(a_i)$ is normally distributed with expectation $\mathbf{E}_f T(a_i)$ and variance $2\varepsilon^2 h^{-1}$. Therefore, for any $\gamma > 0$, using the 'tail' approximation of a normal distribution and (3.1) we derive Lemma 2.3(i) with the following steps

$$\begin{aligned}
\mathbf{P}_f\Big\{\,|\,T(a_i)\,| > (\ln\varepsilon^{-2})^{-\alpha}\Big\} &\le \mathbf{P}_f\Big\{\;|\,T(a_i) - \mathbf{E}_f T(a_i)\,| \;> (\ln\varepsilon^{-2})^{-\alpha} - |\mathbf{E}_f T(a_i)|\,\Big\}\\
&\le \mathbf{P}_f\Big\{\,|\,T(a_i) - \mathbf{E}_f T(a_i)\,| \;> 1/2(\ln\varepsilon^{-2})^{-\alpha}\Big\}\\
&\le 4\Big((\pi h)^{-1}\varepsilon^2\Big)^{1/2}(\ln\varepsilon^{-2})^{\alpha}\exp\Big\{-(4\varepsilon)^{-2}(\ln\varepsilon^{-2})^{-2\alpha}h\Big\}\\
&= 4\pi^{-1/2}(\ln\varepsilon^{-2})^{\alpha-\delta/2-1/2}\exp\Big\{-4^{-2}(\ln\varepsilon^{-2})^{1+\delta-2\alpha}\Big\}\\
&= o(\,\varepsilon^{\gamma}),
\end{aligned}$$

if α is chosen smaller than $\delta/2$.

Let now a_i be a grid-point such that $|\,a_i - b_j| < h/2$. For any such a_i we bound the bias of $T(a_i)$ from below by $L/3$

$$|\,\mathbf{E}_f T(a_i)\,| = h^{-1}\Big|\int\limits_{a_i-h}^{a_i} f(t)\,dt - \int\limits_{a_i}^{a_i+h} f(t)\,dt\,\Big| \;=\; |f(b_j+) - f(b_j-)|\,(1 + o(1)) \;>\; L/3. \tag{3.2}$$

Using (3.2) and taking arbitrary $\gamma > 0$ we finally prove Lemma 2.3(ii) as follows

$$\begin{aligned}
\mathbf{P}_f\Big\{|\,T(a_i)| < (\ln\varepsilon^{-2})^{-\alpha}\Big\} &\le \mathbf{P}_f\left\{\,|\,T(a_i) - \mathbf{E}_f T(a_i)\,| \;> L/4\right\}\\
&\le 8\Big((\pi h)^{-1}\varepsilon^2\Big)^{1/2}L^{-1}\exp\Big\{-(L/\,8\varepsilon)^2 h\Big\}\\
&= 8\Big(\pi^{1/2}L(\ln\varepsilon^{-2})^{(1+\delta)}\Big)^{-1}\exp\Big\{-(L/8)^2(\ln\varepsilon^{-2})^{1+\delta}\Big\}\\
&= o(\,\varepsilon^{\gamma}).
\end{aligned}$$

□

Proof of (2.22)

Define for i, j, each varying from $\lceil h\varepsilon^{-2} \rceil$ to $\lfloor (1-h)\varepsilon^{-2} \rfloor$, the (not necessarily independent), but standard Gaussian distributed random variables

$$\xi_{k,i,l} = \int_{a_i}^{a_l} \varphi_k(t; a_i, a_l)\, d\mathcal{W}(t) \qquad\qquad k = 0, \ldots, m-1, m, \ldots$$

where a_k, a_l are grid-points taken from the set A defined in Section 2.3. Observe that for every $j = 0, \ldots, S$ we have

$$R_{1,j}(f, \hat{b}_j, \hat{b}_{j+1}) = 4\varepsilon^2 \sum_{k=0}^{m-1} \mathbf{E}_f \ln e^{\widehat{\xi}^2_{k,j}/4}.$$

Remark that for every j there are indices i and l such that $\widehat{\xi}_{k,j}$ is equal to $\xi_{k,i,l}$ as the estimators $\hat{b}_j$ and $\hat{b}_{j+1}$ belong to the set of grid-points A. Therefore we have

$$R_{1,j}(f, \hat{b}_j, \hat{b}_{j+1}) \leq 4\varepsilon^2 \sum_{k=0}^{m-1} \mathbf{E}_f \ln \sum_{i,l} e^{\xi^2_{k,i,l}/4}.$$

By Jensen's inequality we finally obtain that the remainder term R_1 is of order $o(\varepsilon^2 \ln^3(\varepsilon^{-2}))$ (for ε tending to 0 and because $N \leq \varepsilon^{-2}$)

$$\begin{aligned} R_{1,j}(f, \hat{b}_j, \hat{b}_{j+1}) &\leq 4\varepsilon^2 \sum_{k=0}^{m-1} \ln \sum_{i,l} \mathbf{E}_f e^{\xi^2_{k,i,l}/4} \\ &= 4\varepsilon^2 m \ln(\sqrt{2}N^2) = o(\varepsilon^2 \ln^3(\varepsilon^{-2})). \end{aligned}$$

Proof of (2.23)

Using relations (2.10) and (2.11) we can associate tapering coefficients $h_{k,i,l}$ with any pair a_i, a_l in the same way it has been done for the pairs b_j, b_{j+1}. In particular, $\widehat{h}_{k,j} = h_{k,i,l}$ if $\hat{b}_j = a_i$ and $\hat{b}_{j+1} = a_l$. Denote by $\|x\|$ the l_2-norm of a sequence $(x_k)_{k \geq m}$. Recall the random variables $\xi_{k,i,l}$ for $k = m, \ldots,$ defined above.

Applying first Cauchy-Schwarz inequality and then Jensen's inequality to the concave function $\ln^2(x+e)$ (for x positive) we bound $R_{2,j}$ as follows

$$
\begin{aligned}
R_{2,j}^2 &\le (2\varepsilon)^4 \mathbf{E}_f \|\widehat{h}^2_{\cdot,j}\|^2 \mathbf{E}_f \Big((4\|\widehat{h}^2_{\cdot,j}\|)^{-1} \sum_{k=m}^{\infty} \widehat{h}^2_{k,j} (\widehat{\xi}^2_{k,j} - 1) \Big)^2 \\
&\le (2\varepsilon)^4\, \mathbf{E}_f \|\widehat{h}_{\cdot,j}\|^2\, \mathbf{E}_f \ln^2 \sum_{i,l} \exp\Big\{ (4\,\|h^2_{\cdot,i,l}\|)^{-1} \sum_{k=m}^{\infty} h^2_{k,i,l}(\xi^2_{k,i,l} - 1) \Big\} \\
&\le (2\varepsilon)^4\, \mathbf{E}_f \|\widehat{h}_{\cdot,j}\|^2 \ln^2 \Big(\sum_{i,l} \mathbf{E}_f \exp\Big\{ (4\,\|h^2_{\cdot,i,l}\|)^{-1} \sum_{k=m}^{\infty} h^2_{k,i,l}(\xi^2_{k,i,l} - 1) \Big\} + e \Big) \\
&\le (2\varepsilon)^4\, \mathbf{E}_f \|\widehat{h}_{\cdot,j}\|^2 \ln^2 \Big(N^2 \max_{i,l} \mathbf{E}_f \exp\Big\{ (4\,\|h^2_{\cdot,i,l}\|)^{-1} \sum_{k=m}^{\infty} h^2_{k,i,l}(\xi^2_{k,i,l} - 1) \Big\} + e \Big).
\end{aligned}
\tag{3.3}
$$

According to the distribution of the random variables $\xi_{k,i,l}$ we see that the second expectation $\mathbf{E}_f \exp\Big\{ (4\,\|h^2_{\cdot,i,l}\|)^{-1} \sum_{k \ge m} h^2_{k,i,l}(\xi^2_{k,i,l} - 1) \Big\}$ equals

$$
\prod_{k=m}^{\infty} \exp\Big\{ - h^2_{k,i,l}(4\,\|h^2_{\cdot,i,l}\|)^{-1} - 1/2 \ln(1 - h^2_{k,i,l}(2\,\|h^2_{\cdot,i,l}\|)^{-1}) \Big\}
$$

and using the elementary inequality $\ln(1-x) \ge -x - x^2$ for $|x| \le 1/2$ this is bounded by $\prod_{k \ge m} \exp\Big\{ h^4_{k,i,l}(\,8\,\|h^2_{\cdot,i,l}\|^2)^{-1} \Big\}$. Substituting the last expression into (3.3) we obtain

$$
\begin{aligned}
R^2_{2,j} &\le (2\varepsilon)^4\, \mathbf{E}_f\, \|\widehat{h}_{\cdot,j}\|^2 \ln^2 \Big(N^2 \max_{i,l} \exp\Big\{ (8\|h^2_{\cdot,i,l}\|^2)^{-1} \sum_{k=m}^{\infty} h^4_{k,i,l} \Big\} + e \Big) \\
&\le (2\varepsilon)^4\, \mathbf{E}_f\, \|\widehat{h}_{\cdot,j}\|^2\, \ln^2 \Big(e\,(N^2+1) \Big).
\end{aligned}
$$

Using Cauchy's inequality,

$$
2ab \le \gamma^{-1} a^2 + \gamma\, b^2, \qquad \gamma > 0, \qquad a, b \in \mathbf{R} \tag{3.4}
$$

and choosing $\gamma = \ln(e\,(N^2+1))$ we finish the proof of (2.23) as follows

$$
\begin{aligned}
R_{2,j} &\le 4\varepsilon^2\, \ln(e\,(N^2+1)) (\mathbf{E}_f \|\widehat{h}_{\cdot,j}\|^2)^{1/2} \\
&\le 2\varepsilon^2 \Big(\ln(e\,(N^2+1)) \Big)^{-1} \mathbf{E}_f \|\widehat{h}_{\cdot,j}\|^2 + 2\varepsilon^2 \ln^3(e\,(N^2+1)) \\
&= o(1)\, \mathbf{E}_f \widehat{L}_{\varepsilon,j}(f) + O\left(\varepsilon^2 \ln^3 \varepsilon^{-2} \right).
\end{aligned}
$$

□

proof of (2.24)

Denote the sequence $((1 - \widehat{h}_{k,j})\, \widehat{h}_{k,j}\, \widehat{\theta}_{k,j})_{\,k \ge m}$ by $(\widehat{\mu}_{k,j})_{\,k \ge m}$. Furthermore define, for $k = m, \ldots,$

$$
\mu_{k,i,l} = (1 - h_{k,i,l})\, h_{k,i,l} \int_{a_i}^{a_l} f(t)\, \varphi_k(t, a_i, a_l)\, dt.
$$

Using Cauchy-Schwarz inequality and Jensen's inequality we obtain as above

$$\begin{aligned}
R_{3,j}^2 &\le (2\varepsilon)^2\,\mathbf{E}_f\|\widehat{\mu}_{\cdot,j}\|^2\mathbf{E}_f\Big(\|\widehat{\mu}_{\cdot,j}\|^{-1}\sum_{k=m}^{\infty}\widehat{\mu}_{k,j}\widehat{\xi}_{k,j}\Big)^2\\
&\le (2\varepsilon)^2\,\mathbf{E}_f\|\widehat{\mu}_{\cdot,j}\|^2\ln^2\Big(N^2\max_{i,l}\mathbf{E}_f\exp\Big\{\|\mu_{\cdot,i,l}\|^{-1}\sum_{k=m}^{\infty}\mu_{k,i,l}\,\xi_{k,i,l}\Big\}+e\Big)\\
&\le (2\varepsilon)^2\,\ln^2(e\,(N^2+1))\,\mathbf{E}_f\|\widehat{\mu}_{\cdot,j}\|^2.
\end{aligned}$$

Hence, according to (3.4) with $\gamma=\ln(e\,(N^2+1))$ we have

$$\begin{aligned}
|R_{3,j}| &\le 2\varepsilon\,\ln(e\,(N^2+1))\,\Big(\mathbf{E}_f\|\widehat{\mu}_{\cdot,j}\|^2\Big)^{1/2}\\
&\le \ln^{-1}(e\,(N^2+1))\mathbf{E}_f\|\,\widehat{\mu}_{\cdot,j}\|^2+\varepsilon^2\,\ln^3\,(e\,(N^2+1))\\
&= o(1)\,\mathbf{E}_f\sum_{k=m}^{\infty}(1-\widehat{h}_{k,j})^2\,\widehat{\theta}_{k,j}^2+O\left(\varepsilon^2\ln^3\varepsilon^{-2}\right)\\
&= o(1)\,\mathbf{E}_f\widehat{L}_{\varepsilon,j}(f)+O\left(\varepsilon^2\ln^3\left(\varepsilon^{-2}\right)\right).
\end{aligned}$$

$\square$

Proof of (2.25)

In this section $j=0,\dots,S$ is arbitrary but fixed. Obviously the expectation of $\widehat{L}_{\varepsilon,j}(f)$ is equal to

$$\mathbf{E}_f\sum_{k=m}^{\infty}(1-h_{k,j}+h_{k,j}-\widehat{h}_{k,j})^2\,\widehat{\theta}_{k,j}^2+\varepsilon^2(h_{k,j}-h_{k,j}+\widehat{h}_{k,j})^2.$$

For arbitrary $0<\gamma<1$ the simple inequality $(a+b)^2\le(1-\gamma)^{-1}a^2+\gamma^{-1}b^2$, $a,b\in\mathbf{R}$, holds. Combining this inequality and the fact that the squared Fourier-coefficient $\widehat{\theta}_{k,j}^2$ is bounded by $\|f\|^2$ we bound $\mathbf{E}_f\widehat{L}_{\varepsilon,j}(f)$ by

$$\sum_{k=m}^{\infty}\Big((1-\gamma)^{-1}\mathbf{E}_f\Big((1-h_{k,j})^2\,\widehat{\theta}_{k,j}^2+\varepsilon^2h_{k,j}^2\Big)+\gamma^{-1}(\,\|f\|^2+\varepsilon^2)\,\mathbf{E}_f(\widehat{h}_{k,j}-h_{k,j})^2\Big).$$

Rewrite this last expression into

$$(1-\gamma)^{-1}\Big(L_{\varepsilon,j}(f)+R_1\Big)+\gamma^{-1}R_2, \tag{3.5}$$

with

$$R_1=\sum_{k=m}^{\infty}(1-h_{k,j})^2(\mathbf{E}_f\widehat{\theta}_{k,j}^2-\theta_{k,j}^2)\ \text{and}\ R_2=(\,\|f\|^2+\varepsilon^2)\sum_{k=m}^{\infty}\mathbf{E}_f(\widehat{h}_{k,j}-h_{k,j})^2.$$

Observe that the desired term $L_{\varepsilon,j}(f)$ turns up in (3.5), besides two (remainder) terms R_1 and R_2 we have to bound.

Because $\mathbf{E}_f\widehat{\theta}^2_{k,j} - \theta^2_{k,j} = \mathbf{E}_f(\widehat{\theta}_{k,j} - \theta_{k,j})^2 + 2\theta_{k,j}(\mathbf{E}_f\widehat{\theta}_{k,j} - \theta_{k,j})$ and $h^2_{k,j} - 2h_{k,j} \leq 0$ for all k the first remaining term R_1 obviously does not exceed

$$\sum_{k=m}^{\infty} (\mathbf{E}_f\widehat{\theta}^2_{k,j} - \theta^2_{k,j}) + 2\sum_{k=m}^{\infty} (h^2_{k,j} - 2h_{k,j})\ \theta_{k,j}\,(\mathbf{E}_f\widehat{\theta}_{k,j} - \theta_{k,j}),$$

which we rewrite into

$$\sum_{k=0}^{m-1} (\theta^2_{k,j} - \mathbf{E}_f\widehat{\theta}^2_{k,j}) + 2\sum_{k=m}^{\infty} (h^2_{k,j} - 2h_{k,j})\ \theta_{k,j}\,(\mathbf{E}_f\widehat{\theta}_{k,j} - \theta_{k,j}), \tag{3.6}$$

as both $\sum_{k\geq 0}\widehat{\theta}^2_{k,j}$ and $\sum_{k\geq 0}\theta^2_{k,j}$ equal $\|f\|^2$. In order to estimate further we establish some approximations of the differences $|\widehat{\theta}_{k,j} - \theta_{k,j}|$.

Remind that the jump-points b_j are at least at distance κ from each other. Moreover $\hat{b}_j$ differs from b_j more than $2h$ only with small probability (cf. (2.21) and (2.17)). Therefore we can assume that the estimated jump-points $\hat{b}_j$ are also separated from each other at least by $\kappa/2$.

In view of the definitions of the functions $\varphi_{k,j}$ and $\widehat{\varphi}_{k,j}$ on the subintervals $[\,b_j, b_{j+1}]$ and $[\,\hat{b}_j, \hat{b}_{j+1}]$ respectively and Theorem 2.2 we deduce that there exists a constant C_1 such that for $j = 0, \ldots, S$ and $t \in [\,\max(b_j, \hat{b}_j), \min(b_{j+1}, \hat{b}_{j+1})\,]$ the following approximations hold

$$|\,\widehat{\varphi}_{k,j}(t) - \varphi_{k,j}(t)\,| \leq C_1\left(|\,\hat{b}_{j+1} - b_{j+1}\,| + |\,\hat{b}_j - b_j\,|\right), \qquad k = 0, \ldots, m-1$$

and

$$|\,\widehat{\varphi}_{k,j}(t) - \varphi_{k,j}(t)\,| \leq C_1\,k\left(|\,\hat{b}_{j+1} - b_{j+1}\,| + |\,\hat{b}_j - b_j\,|\right), \qquad k = m, \ldots.$$

On the intervals $[\min(\hat{b}_j, b_j), \max(\hat{b}_j, b_j)]$ and $[\min(\hat{b}_{j+1}, b_{j+1}), \max(\hat{b}_{j+1}, b_{j+1})]$ either $\widehat{\varphi}_{k,j}$ or $\varphi_{k,j}$ is zero and the other is bounded. This implies that there exists a constant C_2 such that

$$|\,\widehat{\theta}_{k,j} - \theta_{k,j}\,| \leq C_2 \sup_{j=0,\ldots,S} |\hat{b}_j - b_j|, \qquad k = 0, \ldots, m-1 \tag{3.7}$$

and

$$|\,\widehat{\theta}_{k,j} - \theta_{k,j}\,| \leq C_2\,k \sup_{j=0,\ldots,S} |\hat{b}_j - b_j|, \qquad k = m, \ldots. \tag{3.8}$$

Denote $\sup_{j=0,\dots,S} |\hat{b}_j - b_j|$ further on by D. Given relations (3.7) and (3.8) we return to the estimation of R_1. The first term of (3.6) equals

$$\sum_{k=0}^{m-1} \mathbf{E}_f\,(\theta_{k,j} - \widehat{\theta}_{k,j})^2 + 2\theta_{k,j}(\theta_{k,j} - \mathbf{E}_f\widehat{\theta}_{k,j})$$

and is therefore bounded by $C_2\, m\,(\,C_2\,\mathbf{E}_f D^2 + 2\,\|f\|\,\mathbf{E}_f D\,)$. As $|h(h-2)| \le 2h$ the second term of (3.6) is bounded by $4C_2\,\mathbf{E}_f D \sum_{k\ge m} h_{k,j} k\,|\,\theta_{k,j}\,|$. According to condition A2.3 and (2.6) the sum $\sum_{k\ge m} k^2\theta_{k,j}^2$ is bounded. Furthermore Corollary 2.4 gives a bound for $\mathbf{E}_f D^2$. Hence for some constant C_3 we have, by Cauchy-Schwarz and (3.4) with $\gamma = \varepsilon^2(\ln\varepsilon^{-2})^{-1}$ that

$$\begin{aligned}
R_1 &\le C_2\;\mathbf{E}_f D\,(\,4\sum_{k=m}^{\infty} h_{k,j}\,k|\theta_{k,j}| + 2m\,\|f\|\,) + m\,C_2^2\,\mathbf{E}_f D^2\,) \\
&\le C_3\,\|\,h_j\,\|\,\Big(\,\sum_{k=m}^{\infty} k^2\,\theta_{k,j}^2\Big)^{1/2}\Big(\mathbf{E}_f D^2\Big)^{1/2} + O\,(\,\mathbf{E}_f D\,) \\
&\le C_3\Big((\ln\varepsilon^{-2})^{-1}\varepsilon^2\sum_{k=m}^{\infty} h_{k,j}^2 + \varepsilon^{-2}\ln\varepsilon^{-2}\,\mathbf{E}_f\,D^2\Big) + O\,(\,\mathbf{E}_f D\,) \\
&= o(1)L_{\varepsilon,j}(f) + O\left(\varepsilon^2\ln^{3+2\delta}(\varepsilon^{-2})\right) \qquad\qquad \varepsilon\to 0.
\end{aligned} \tag{3.9}$$

It remains to estimate R_2. Below we will see that there exists a constant C_4 such that the coefficients $h_{k,j}$ vanish for $k > C_4\varepsilon^{-2/(2m+1)}$ (cf. (3.11) and (3.13)). Furthermore from the formulas (2.6) and (2.10) it is clear that the tapering coefficients $h_{k,j}$ behave well in the sense that we can assume that $h_{k,j}$ is Lipschitz with respect to $b_{j+1} - b_j$. Hence for ε tending to zero we have

$$\begin{aligned}
R_2 &= (\,\|f\| + \varepsilon^2\,)\sum_{k=m}^{C_4\varepsilon^{-2/(2m+1)}} \mathbf{E}_f\,(\widehat{h}_{k,j} - h_{k,j})^2 \\
&= O\left(\varepsilon^{\frac{-2}{2m+1}}\,\mathbf{E}_f\,D^2\right) = O\left(\varepsilon^{2+\frac{4m}{2m+1}}\ln^{2+2\delta}(\varepsilon^{-2})\right).
\end{aligned} \tag{3.10}$$

Substituting (3.9), (3.10) and $\gamma = \varepsilon$ in (3.5) we obtain (2.25).

□

Proof of (2.15)

According to (2.6), the number of non-zero summations in (2.11) are finite. Denote these numbers by $N_j = N_j(c_j)$. Note that the solutions $c_j = c_j(\varepsilon)$ of the equations (2.11) tend to zero, as ε does. Indeed if $c_j(\varepsilon)$ stayed away from zero, the same would happen to the left-hand sides of (2.11) while their respective right-hand sides would tend to zero. Therefore, again according to (2.6),

$$N_j = \frac{b_{j+1} - b_j}{\pi} c_j^{-1/m}(1 + o(1)), \qquad \varepsilon \to 0 \tag{3.11}$$

and using this, together with the asymptotic relations, (for $j = 0, \ldots, S$ and $\alpha > 0$),

$$\sum_{k=m}^{N_j} \lambda_{k,j}^{\alpha} = \left(\frac{\pi}{b_{j+1} - b_j}\right)^{2m\alpha} \frac{N_j^{2m\alpha+1}}{2m\alpha + 1}(1 + o(1)), \qquad N_j \to \infty, \tag{3.12}$$

the equations (2.11) become

$$c_j^{\frac{2m+1}{m}} Q = \frac{m\varepsilon^2}{\pi(m+1)(2m+1)}(1 + o(1)), \qquad \varepsilon \to 0.$$

Note that the c_j asymptotically do not depend on j, i.e., $c_j = c\,(1 + o(1))$ where

$$c = c(\,\varepsilon) = \left(\frac{m\varepsilon^2}{\pi Q(m+1)(2m+1)}\right)^{\frac{m}{2m+1}} \tag{3.13}$$

and as the calculations below show, we can just substitute this value of c into (2.15). Indeed with such a choice of c, we have, according to (3.11) – (3.13), for ε tending to zero,

$$\begin{aligned}
\sum_{j=0}^{S}(\varepsilon^2 \sum_{k=m}^{\infty} h_{k,j}^2 + c_j^2 Q_j) &= \sum_{j=0}^{S}(\varepsilon^2 \sum_{k=m}^{N_j} (1 - c\lambda_{k,j}^{1/2})^2 + c^2 Q_j) \\
&= \varepsilon^2 c^{-1/m} \frac{2m^2}{\pi(m+1)(2m+1)} \sum_{j=0}^{S}(b_{j+1} - b_j) + c^2 Q \\
&= \gamma(m, Q)\, \varepsilon^{\frac{4m}{2m+1}}(1 + o(1)).
\end{aligned}$$

□

Acknowledgements

I thank B. Levit for proposing me this problem and for his supervising during this work, Y. Golubev for the stimulating discussions and his indispensable help with the proofs. Moreover I want to thank H. Duistermaat and R. Gill for their helpful comments and useful discussions.

References

Belitser, E. and B. Levit (1996). Asymptotically minimax nonparametric regression in L_2. *Statistics 28*, 105–122.

Birgé, L. et al. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete 65*, 181–237.

Brown, L. and M. Low (1990). Asymptotic equivalence of nonparametric regression and white noise. Technical report, University of Pennsylvania. to appear in Ann. Statist.

Coddington, E. and N. Levinson (1955). *Theory of ordinary differential equations.* McGraw-Hill Book Company, Inc. New York.

Donoho, D. (1994). Asymptotic Risk for Sup-Norm Loss: Solution via Optimal Recovery. *Probab. Theory Related Fields 99*, 145–170.

Duistermaat, J. (1995, February). The Sturm-Liouville problem for the Operator $(-\mathrm{d}^2/\mathrm{d}x^2)^m$, with Neumann or Dirichlet Boundary conditions. Technical Report 899, Department of Mathematics, University Utrecht.

Dunford, N. and J. Schwartz (1971). *Linear operators, part III: Spectral Operators.* Wiley Interscience, New York, etc.

Efroimovich, S. Y. (1994). On nonparametric regression for i.i.d. observations in general setting. Technical report, University of New Mexico.

Golubev, Y. K., B. Levit, and A. Tsybakov (1995). Asymptotically efficient estimation of analytic functions in Gaussian noise. Technical Report 894, Dept. Mathematics, University of Utrecht.

Golubev, Y. K. and M. Nussbaum (1990). A risk bound in Solobev class regression. *Ann. Statist. 18*(2), 758–778.

Hall, P. and P. Patil (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist. 23*(3), 905–928.

Ibragimov, I. and R. Hasminskii (1981). *Statistical Estimation: Asymptotic Theory.* Springer Verlag, New York–Heidelberg–Berlin.

Korostelev, A. (1987). On minimax estimation of a discontinuous signal. *Theory Probab. Appl. 32*, 727–730.

Korostelev, A. (1994). Exact asymptotic minimax estimate of a nonparametric regression in the uniform norm. *Theory Probab. Appl. 38*, 737–743.

Müller, H. (1992). Change-points in nonparametric regression analysis. *Ann. Statist. 20*, 737–761.

Neumark, M. (1967). *Lineare differential operatoren.* Akademie Verlag Berlin.

Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist. 13*(3), 984–997.

Pinsker, M. (1980). Optimal Filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission 16*(2), 120–133.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist. 13*, 970–983.

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist. 10*(4), 1040–1053.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika 82*(2), 385–397.