Bernoulli **8**(4), 2002, 475–489

Application of structural risk minimization to multivariate smoothing spline regression estimates

MICHAEL KOHLER[1*], ADAM KRZYŻAK[2] and DOMINIK SCHÄFER[1**]

[1]*Mathematisches Institut A, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany.*
E-mail: *kohler@mathematik.uni-stuttgart.de; **schaefdk@mathematik.uni-stuttgart.de*
[2]*Department of Computer Science, Concordia University, 1455 De Maisonneuve West, Montreal, Canada H3G 1M8. E-mail: krzyzak@cs.concordia.ca*

Estimation of regression functions from bounded, independent and identically distributed data is considered. Motivated by Vapnik's principle of structural risk minimization, a data-dependent choice of the smoothing parameter of multivariate smoothing spline estimates is proposed. The corresponding smoothing spline estimates automatically adapt to the unknown smoothness of the regression function and their L_2 errors achieve the optimal rate of convergence up to a logarithmic factor. The result is valid without any regularity conditions on the distribution of the design.

Keywords: empirical process theory; rate of convergence; regression estimate; smoothing splines; structural risk minimization

1. Introduction

Let $(X, Y), (X_1, Y_1), (X_2, Y_2), \ldots$ be independent, identically distributed $(\mathbb{R}^d \times \mathbb{R})$-valued random vectors with $\mathbf{E}Y^2 < \infty$. No assumptions are made about the distribution of X, which can be discrete or continuous or a mixture of the two. In regression estimation the distribution of (X, Y) is unknown. Given a sequence $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of independent observations of (X, Y), our goal is to construct an estimate $m_n(x) = m_n(x, \mathcal{D}_n)$ of the regression function $m(x) := \mathbf{E}\{Y|X = x\}$ such that the L_2 error

$$\int |m_n(x) - m(x)|^2 P_X(\mathrm{d}x)$$

is small.

It was demonstrated by Stone (1977) that there exist estimates m_n with the property that $\int |m_n(x) - m(x)|^2 P_X(\mathrm{d}x) \to 0$ in probability for all distributions of (X, Y) with $\mathbf{E}Y^2 < \infty$. Unfortunately, distribution-free rates of convergence for the L_2 error do not exist. In order to obtain non-trivial rates one has to impose restrictions on the class of distributions considered; see Devroye *et al.* (1996, Chapter 7) and Devroye and Wagner (1980).

For p-smooth regression functions, by the results of Stone (1982), there is no estimate

1350–7265 © 2002 ISI/BS

which converges in the minimax sense in L_2 faster than $n^{-2p/(2p+d)}$. He demonstrated that the L_2 error of a local polynomial kernel estimate converges to zero in probability with the rate $n^{-2p/(2p+d)}$ for all distributions of (X, Y) such that the regression function is p-smooth, X has a density bounded away from zero and infinity and $\mathrm{E}Y^2$ is finite.

Kohler (2000) showed that the L_2 error of suitably defined least-squares spline estimates achieves the rate $n^{-2p/(2p+d)}$ for all distributions of (X, Y) with p-smooth regression function (p bounded above by some fixed upper bound on the smoothness) and (X, Y) bounded with probability one. The definition of his estimates does not depend on p, hence the estimates automatically adapt to the smoothness of the regression function. This adaptation is achieved by an application of Vapnik's principle of structural risk minimization. The basic idea is to define a family of least-squares estimates, to derive upper bounds on the L_2 error of these estimates depending on the empirical L_2 risk $(1/n)\sum_{i=1}^{n}|f(X_i) - Y_i|^2$ together with a measure of the complexity of the underlying function space, and finally to choose the estimate whose upper bound on the L_2 error for a given set of data is minimal.

In this paper we use a similar idea to define adaptive smoothing spline estimates. We show that the estimates automatically adapt to the unknown smoothness of the regression function and that their L_2 errors achieve the optimal rate of convergence up to a logarithmic factor. This result is valid without any regularity assumption on the distribution of the design.

1.1. Discussion of related results

Smoothing spline estimates have been studied by many authors; see, for example, the monographs by Eubank (1988) and Wahba (1990) and the literature cited therein. Most of the results in the literature are derived for fixed design regression (where the X_i are non-random) and cover the case $d = 1$ only.

In the context of random design regression, consistency and rate of convergence of univariate smoothing spline estimates have been studied by means of empirical process theory by van de Geer (1987; 1988; 1990). Further results can be found in the monograph of van de Geer (2000), which is also an excellent source for the techniques of empirical process theory used in this paper. The results of van de Geer cited above are derived for estimators which use parameters dependent on the smoothness of the regression function. For fixed design regression, estimates similar to the one in this paper are studied by van de Geer (2001).

The principle of structural risk minimization which is behind the definition of the adaptive estimates in this paper was introduced by Vapnik and Chervonenkis (1974) in the context of pattern recognition; see also the recent monograph by Vapnik (1998). In Krzyżak and Linder (1998) and Kohler (1998; 2000) it was applied to various least-squares estimates. These estimates are similar to adaptive least-squares estimates investigated in Barron *et al.* (1999) and Baraud (1997), although the estimates in these papers have different motivations. They obtain the optimal rate of convergence but under some regularity condition on the distribution of the design. (One should note that Baraud (1997) and in part also Barron *et al.* (1999) use the L_2 error with integration with respect to the Lebesgue–Borel measure, whose analysis requires regularity conditions on the design.) Krzyżak and Linder (1998) and Kohler (1998) impose no conditions besides boundedness on the distribution of the design, but, as in the present paper, they obtain rates of convergence that are optimal up to a logarithmic factor. In Kohler (2000) the optimal rate of convergence is shown for adaptive least-squares spline estimates under no assumptions on the distribution of the design. There, in contrast to the results in the present paper, it is assumed that the derivatives of order $p-1$ satisfy a global Lipschitz condition (rather than that the derviatives of order p are square-integrable) and that p is bounded from above by some fixed upper bound on the smoothness.

1.2. Notation

Throughout the paper we will use the following notation: $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$ denote the set of natural numbers, real numbers and non-negative real numbers, respectively. For $x \in \mathbb{R}$, we denote the greatest integer less than or equal to x by $\lfloor x \rfloor$. For $L > 0$ and $z \in \mathbb{R}$, set $T_L z = \operatorname{sgn}(z)\min\{L, |z|\}$. For a function $f: \mathbb{R}^d \to \mathbb{R}$, define $T_L f: \mathbb{R}^d \to \mathbb{R}$ by $(T_L f)(x) = T_L(f(x))$, $x \in \mathbb{R}^d$. The partial derivative of order $(\alpha_1, \ldots, \alpha_d)$ of a function $f: \mathbb{R}^d \to \mathbb{R}$ is denoted by $\partial^{\alpha_1+\ldots+\alpha_d} f/\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}$. $L_2([0, 1]^d)$ is the set of all square-integrable functions $f: [0, 1]^d \to \mathbb{R}$, and for $k \in \mathbb{N}$, $W^k([0, 1]^d)$ is the Sobolev space containing all functions $f: [0, 1]^d \to \mathbb{R}$ whose derivatives of total order k are in $L_2([0, 1]^d)$. $\log(x)$ denotes the natural logarithm of $x > 0$. $|A|$ is the cardinality of the set A and I_A the corresponding indicator function.

1.3. Outline of the paper

In Section 2 we give the exact definition of the estimate. The main result is stated in Section 3 and proven in Section 4.

2. Definition of the estimate

We will assume that (X, Y) takes with probability one only values in some bounded subset of $\mathbb{R}^d \times \mathbb{R}$. Without loss of generality this bounded subset is $[0, 1]^d \times [-L, L]$, that is, almost surely, $(X, Y) \in [0, 1]^d \times [-L, L]$ for some $L \in \mathbb{R}_+$.

Let $k \in \mathbb{N}$ with $2k > d$. The condition $2k > d$ implies that the functions in $W^k([0, 1]^d)$ are continuous and hence the value of a function at a point is well defined. Set

$$J_k^2(f) = \sum_{\substack{\alpha_1,\ldots,\alpha_d \in \mathbb{N},\\ \alpha_1+\ldots+\alpha_d=k}} \frac{k!}{\alpha_1! \cdot \ldots \cdot \alpha_d!} \int_{\mathbb{R}^d} \left| \frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(x) \right|^2 \mathrm{d}x.$$

Let $\lambda \in \mathbb{R}_+$. First define the smoothing spline estimate $\tilde{m}_{n,(k,\lambda)}$ by

$$\tilde{m}_{n,(k,\lambda)}(\cdot) = \arg \min_{f \in W^k([0,1[^d)} \left(\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 + \lambda J_k^2(f) \right). \tag{1}$$

By the results in Duchon (1976, Section V) the minimum exists; for our purposes, however, we do not require it to be unique. Also observe that $\tilde{m}_{n,(k,\lambda)}$ depends on the data $\mathcal{D}_n$ and that we have suppressed this in our notation.

The estimate $\tilde{m}_{n,(k,\lambda)}$ is parametrized by $k \in \mathbb{N}$ and $\lambda \in \mathbb{R}_+$. We next describe how one can use the data $\mathcal{D}_n$ to choose these parameters. The basic idea is related to Vapnik's structural risk minimization principle: in Lemma 1 below we give an upper bound on L_2 error of (a truncated version of) the estimate $\tilde{m}_{n,(k,\lambda)}$. We then choose (k^*, λ^*) by minimizing this upper bound.

Lemma 1. *Let $1 \leqslant L < \infty$, $\lambda \in \mathbb{R}_+$ and $\eta \in (0, 1]$. Then for*

$$pen_n(k, \lambda) = \frac{L^5(\log(n))^2}{n \cdot \lambda^{d/2k}}$$

and n sufficiently large, one has, with probability greater than or equal to $1 - \eta$,

$$\int_{\mathbb{R}^d} |T_L \tilde{m}_{n,(k,\lambda)}(x) - m(x)|^2 P_X(\mathrm{d}x) \leqslant L^4 \frac{\log(n)}{n} + 2\, pen_n(k, \lambda) + 2 \left\{ \frac{1}{n} \sum_{i=1}^{n} |T_L \tilde{m}_{n,(k,\lambda)}(X_i) - Y_i|^2 \right.$$
$$\left. + \lambda J_k^2(\tilde{m}_{n,(k,\lambda)}) - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right\}$$

for every distribution of (X, Y) with $(X, Y) \in [0, 1]^d \times [-L, L]$ almost surely.

The proof of Lemma 2 is similar to the proof of Theorem 1 given below and is therefore omitted.

Set

$$\mathcal{K}_n := \left\{ \left\lfloor \frac{d}{2} \right\rfloor, \ldots, \left\lfloor \frac{d}{2} \right\rfloor + \lfloor (\log(n))^{1/2d} \rfloor \right\}$$

and

$$\Lambda_n := \left\{ \frac{\log(n)}{2^n}, \frac{\log(n)}{2^{n-1}}, \ldots, \frac{\log(n)}{1} \right\}.$$

For $(k, \lambda) \in \mathcal{K}_n \times \Lambda_n$, define $m_{n,(k,\lambda)}$ by

$$m_{n,(k,\lambda)}(x) = T_L \tilde{m}_{n,(k,\lambda)}(x), \qquad x \in \mathbb{R}^d. \tag{2}$$

Depending on the data $\mathcal{D}_n$, we choose the member of the family $\{m_{n,(k,\lambda)}: (k, \lambda) \in \mathcal{K}_n \times \Lambda_n\}$ that minimizes the upper bound in Lemma 1. More precisely, we choose

$$(k^*, \lambda^*) = (k^*(\mathcal{D}_n), \lambda^*(\mathcal{D}_n)) \in \mathcal{K}_n \times \Lambda_n$$

such that

$$\frac{1}{n}\sum_{i=1}^{n}|m_{n,(k^*,\lambda^*)}(X_i)-Y_i|^2+\lambda^* J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)})+pen_n(k^*,\lambda^*)$$

$$=\min_{(k,\lambda)\in\mathcal{K}_n\times\Lambda_n}\left\{\frac{1}{n}\sum_{i=1}^{n}|m_{n,(k,\lambda)}(X_i)-Y_i|^2+\lambda J_k^2(\tilde{m}_{n,(k,\lambda)})+pen_n(k,\lambda)\right\}$$

and define our adaptive smoothing spline estimate by

$$m_n(x)=m_n(x,\mathcal{D}_n)=m_{n,(k^*(\mathcal{D}_n),\lambda^*(\mathcal{D}_n))}(x,\mathcal{D}_n).$$

An upper bound on the L_2 error of the estimate is given in the next section.

3. Main result

In Section 4 we prove the oracle-style inequality

$$\mathbf{E}\int|m_n(x)-m(x)|^2P_X(\mathrm{d}x)\leqslant 2\left(\inf_{\lambda\in\Lambda_n}(\lambda J_p^2(m)+pen_n(m,\lambda))\right)+\frac{\text{const.}}{n} \tag{3}$$

$(2p>d)$ for the above estimate m_n from which the following theorem can be derived:

Theorem 1. *Let* $p\in\mathbb{N}$ *with* $2p>d$ *be arbitrary. Let the estimate* m_n *be defined as in Section 2.*

(*a*)

$$\mathbf{E}\int|m_n(x)-m(x)|^2P_X(\mathrm{d}x)=O\left(\frac{(\log(n))^2}{n}\right)$$

for any $p>d/2$ *and any distribution of* (X,Y) *with* $(X,Y)\in[0,1]^d\times[-L,L]$ *almost surely,* $m\in W^p([0,1]^d)$ *and* $J_p^2(m)=0$.

(*b*)

$$\mathbf{E}\int|m_n(x)-m(x)|^2P_X(\mathrm{d}x)=O\left((J_p^2(m))^{d/(2p+d)}(\log n)^2n^{-2p/(2p+d)}\right)$$

for any $p>d/2$ *and any distribution of* (X,Y) *with* $(X,Y)\in[0,1]^d\times[-L,L]$ *almost surely,* $m\in W^p([0,1]^d)$ *and* $0<J_p^2(m)<\infty$.

Remark 1. It follows from the proof given in Section 4 that in inequality (3) the factor 2 can be replaced by $1+\eta$, where $\eta>0$ is arbitrarily small.

Remark 2. The condition $J_p^2(m)=0$ in Theorem 1(a) implies that m is a multivariate polynomial of degree $p-1$ (or less, in each coordinate). In this case the estimate is within a logarithmic factor of the parametric rate $O(1/n)$. Furthermore, it follows from Stone (1982) that the rate of convergence in Theorem 1(b) is optimal up to the logarithmic factor $(\log(n))^2$.

Remark 3. The definition of the estimate does not depend on p or $J_p^2(m)$, hence it automatically adapts to the unknown smoothness of the regression function measured by p and $J_p^2(m)$.

Remark 4. We wish to emphasize that in Theorem 1 there is no assumption on the underlying distribution of X besides $X \in [0, 1]^d$ with probability one. In particular, X is not required to have a density with respect to the Lebesgue–Borel measure.

Remark 5. We use truncation of the estimate in order to ensure that the estimate is bounded in supremum norm on $[0, 1]^d$, a property which is required in Lemma 3 below. It should be noted that in the general setting considered here truncation is indeed necessary. For example, consider random variables X with $P\{X = 0\} = \frac{1}{2}$ and $P\{X \leqslant x\} = (1 + x)/2$ for $x \in [0, 1]$ and Y independent of X with $P\{Y = -1\} = P\{Y = +1\} = \frac{1}{2}$. Hence $m(x) = 0$ for all x. Now draw an independent, identically distributed sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of (X, Y). If the event $A := \{X_1 = \ldots = X_{n-1} = 0;\ Y_1, \ldots, Y_{n-1} = -1;\ X_n \neq 0;\ Y_n = 1\}$ occurs, then the smoothing spline m_n, obtained with penalty J_k^2 for $k \geqslant 2$, is the straight line through $(0, -1)$ and $(X_n, 1)$, $m_n(x) = -1 + 2x/X_n$. Using this, the L_2 error satisfies

$$\begin{aligned}
\mathrm{E}\int |m_n(x) - m(x)|^2 P_X(\mathrm{d}x) &\geqslant \mathrm{E}\left(I_A \cdot \frac{1}{2}\int_0^1 \left| -1 + \frac{2x}{X_n}\right|^2 \mathrm{d}x \right) \\
&= \frac{p}{2} \cdot \mathrm{E}\left(I_{\{X_n \neq 0\}} \cdot \frac{X_n}{6}\left(\left(-1 + \frac{2}{X_n}\right)^3 + 1\right)\right) \\
&= \frac{p}{4}\int_0^1 \frac{u}{6}\left(\left(-1 + \frac{2}{u}\right)^3 + 1\right) \mathrm{d}u = \infty,
\end{aligned}$$

where $p = P\{X_1 = \ldots = X_{n-1} = 0;\ Y_1 = \ldots = Y_{n-1} = -1;\ Y_n = 1\} > 0$. In other words, if no restrictions are imposed on the distribution of the regression design, the untruncated smoothing spline estimate with penalty of order $k \geqslant 2$ is not even weakly universally consistent. In view of this example, some sort of truncation cannot be avoided.

4. Proofs

In the proof of Theorem 1 we will need two auxiliary results. To formulate these results we need the concept of covering numbers.

Definition 1. *Let $q \geqslant 1$, $l \in \mathbb{N}$ and let $\mathcal{F}$ be a class of functions $f\colon \mathbb{R}^l \to \mathbb{R}$. The covering number $\mathcal{N}_q(\epsilon, \mathcal{F}, x_1^n)$ is defined for any $\epsilon > 0$ and $x_1^n = (x_1, \ldots, x_n) \in (\mathbb{R}^l)^n$ as the smallest integer k such that there exist functions $g_1, \ldots, g_k\colon \mathbb{R}^l \to \mathbb{R}$ with*

$$\min_{1 \leqslant i \leqslant k} \left(\frac{1}{n} \sum_{j=1}^{n} |f(x_j) - g_i(x_j)|^q \right)^{1/q} \leqslant \epsilon$$

for each $f \in \mathcal{F}$.

Lemma 2 (Kohler 2000, Theorem 2). *Let* $Z_1^n = (Z_1, \ldots, Z_n)$ *be independent, identically distributed random variables with values in some set* $\mathcal{X}$. *Let* $K_1, K_2 \geqslant 1$ *and let* $\mathcal{F}$ *be a class of functions* $f: \mathcal{X} \to [-K_1, K_1]$ *such that*

$$\mathrm{E}\{f(Z_1)^2\} \leqslant K_2 \mathrm{E} f(Z_1). \tag{4}$$

For $0 < \epsilon < 1$ *and* $\alpha > 0$, *let*

$$\sqrt{n}\sqrt{1-\epsilon}\sqrt{\alpha} \geqslant 288 \max\{2K_1, \sqrt{2K_2}\} \tag{5}$$

and, for all $z_1, \ldots, z_n \in \mathcal{X}$ *and all* $\delta \geqslant \alpha/4$, *let*

$$\sqrt{n}(1-\epsilon)\delta \geqslant 288 \max\{2K_1, K_2\} \int_{\alpha/8K_2}^{\sqrt{\delta}} \left(\log \mathcal{N}_2 \left(u, \left\{ f \in \mathcal{F}: \frac{1}{n} \sum_{i=1}^{n} f^2(z_i) \leqslant 4\delta \right\}, z_1^n \right) \right)^{1/2} \mathrm{d}u. \tag{6}$$

Then

$$P \left\{ \sup_{f \in \mathcal{F}} \frac{\left| (1/n) \sum_{i=1}^{n} f(Z_i) - \mathrm{E} f(Z_1) \right|}{\alpha + \mathrm{E} f(Z_1)} > \epsilon \right\} \leqslant 50 \exp\left(\frac{n \alpha \epsilon^2 (1-\epsilon)}{128 \cdot 2304 \max\{K_1^2, K_2\}} \right). \tag{7}$$

If one replaces $\alpha/8K_2$ by zero (6) then Lemma 2 follows from Theorem 2 in Kohler (2000). The version given here can be proven analogously by an application of Lemma 3.2 in the general version of van de Geer (2000), instead of the more specific form given in Lemma 3 in Kohler (2000).

Lemma 3. *Let* $L, c > 0$ *and set*

$$\mathcal{F} = \{T_L f: f \in W^k([0, 1]^d) \text{ and } J_k^2(f) \leqslant c\}.$$

Then there exists a constant $c_d \in \mathbb{R}_+$ *depending only on d, such that, for any* $\epsilon > 0$ *and all* $x_1, \ldots, x_n \in [0, 1]^d$,

$$\log \mathcal{N}_2(\epsilon, \mathcal{F}, x_1^n) \leqslant c_d(k^d + 1)\left(\left(\frac{\sqrt{c}}{\epsilon} \right)^{d/k} + 1 \right) \cdot \log\left(\frac{64 \mathrm{e} L^2 n}{\epsilon^2} \right) \cdot I_{\{\epsilon \leqslant L\}}. \tag{8}$$

Proof. The proof is a straightforward modification of the proof of Lemma 3 in Kohler and Krzyżak (2001), therefore we give only an outline.

For $\epsilon > L$ the bound (8) is trivially satisfied for the cover $\{0\}$, so assume $0 < \epsilon \leqslant L$. Let

$\mathcal{G}$ be the set of all piecewise polynomials of degree $k-1$ (or less, in each coordinate) with respect to a rectangular partition of $[0, 1]^d$ consisting of at most

$$K \leqslant c_1(2^d+1)\left(\sqrt{\frac{c}{\epsilon}}\right)^{d/k} + 2^d$$

rectangles, and set $T_L\mathcal{G} = \{T_L g\colon g \in \mathcal{G}\}$. It follows from the proof of Lemma 3 in Kohler and Krzyżak (2001) that, for $c_1 = c_1(d) \in \mathbb{R}_+$ sufficiently large, for any $f \in \mathcal{F}$, there exists $p_f \in T_L\mathcal{G}$ such that

$$\sup_{x\in[0,1]^d} |f(x) - p_f(x)| < \frac{\epsilon}{2}.$$

From this one easily concludes

$$\mathcal{N}_2(\epsilon, \mathcal{F}, x_1^n) \leqslant \mathcal{N}_2\left(\frac{\epsilon}{2}, T_L\mathcal{G}, x_1^n\right). \tag{9}$$

For arbitrary functions f, g: $\mathbb{R}^d \to [-L, L]$,

$$\left(\frac{1}{n}\sum_{i=1}^{n}|f(x_i) - g(x_i)|^2\right)^{1/2} \leqslant \sqrt{2L}\left(\frac{1}{n}\sum_{i=1}^{n}|f(x_i) - g(x_i)|\right)^{1/2},$$

which implies

$$\mathcal{N}_2\left(\frac{\epsilon}{2}, T_L\mathcal{G}, x_1^n\right) \leqslant \mathcal{N}_1\left(\frac{\epsilon^2}{8L}, T_L\mathcal{G}, x_1^n\right). \tag{10}$$

Finally, it is shown in the proof of Lemma 3 in Kohler and Krzyżak (2001) that

$$\log\mathcal{N}_1\left(\frac{\epsilon^2}{8L}, T_L\mathcal{G}, x_1^n\right) \leqslant c_2\left(\left(\frac{\sqrt{c}}{\epsilon}\right)^{d/k} + 1\right)(k^d+1)\log\left(\frac{64\mathrm{e}L^2 n}{\epsilon^2}\right). \tag{11}$$

Then (9), (10) and (11) imply the assertion. □

Proof of Theorem 1. Let $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be the collection of observed data. We divide the proof into five steps.

Step 1. We start with the error decomposition

$$\int_{\mathbb{R}^d}|m_n(x) - m(x)|^2 P_X(\mathrm{d}x) = T_{1,n} + T_{2,n},$$

where

$$T_{1,n} = \mathrm{E}[|m_n(X) - Y|^2|\mathcal{D}_n] - \mathrm{E}(|m(X) - Y|^2) - 2\Bigg\{\frac{1}{n}\sum_{i=1}^{n}|m_n(X_i) - Y_i|^2$$
$$+ \lambda^* J^2_{k^*}(\tilde{m}_{n,(k^*,\lambda^*)}) - \frac{1}{n}\sum_{i=1}^{n}|m(X_i) - Y_i|^2 + pen_n(k^*, \lambda^*)\Bigg\}$$

and

$$T_{2,n} = 2\Bigg\{\frac{1}{n}\sum_{i=1}^{n}|m_n(X_i) - Y_i|^2 + \lambda^* J^2_{k^*}(\tilde{m}_{n,(k^*,\lambda^*)}) - \frac{1}{n}\sum_{i=1}^{n}|m(X_i) - Y_i|^2 + pen_n(k^*, \lambda^*)\Bigg\}.$$

We show that

$$T_{2,n} \leqslant 2 \inf_{\lambda\in\Lambda_n}\Big\{\lambda J^2_p(m) + pen_n(p, \lambda)\Big\}. \tag{12}$$

By the definition of m_n, the Lipschitz property of T_L and $|Y_i| \leqslant L$ almost surely, which implies $Y_i = T_L Y_i$ $(i = 1, \ldots, n)$ almost surely and $J^2_p(m) < \infty$,

$$\begin{aligned}
T_{2,n} &\leqslant 2 \inf_{\lambda\in\Lambda_n}\Bigg\{\frac{1}{n}\sum_{i=1}^{n}|T_L\tilde{m}_{n,(p,\lambda)}(X_i) - Y_i|^2 + \lambda J^2_p(\tilde{m}_{n,(p,\lambda)}) \\
&\qquad - \frac{1}{n}\sum_{i=1}^{n}|m(X_i) - Y_i|^2 + pen_n(p, \lambda)\Bigg\} \\
&\leqslant 2 \inf_{\lambda\in\Lambda_n}\Bigg\{\frac{1}{n}\sum_{i=1}^{n}|\tilde{m}_{n,(p,\lambda)}(X_i) - Y_i|^2 + \lambda J^2_p(\tilde{m}_{n,(p,\lambda)}) \\
&\qquad - \frac{1}{n}\sum_{i=1}^{n}|m(X_i) - Y_i|^2 + pen_n(p, \lambda)\Bigg\} \\
&\leqslant 2 \inf_{\lambda\in\Lambda_n}\Bigg\{\frac{1}{n}\sum_{i=1}^{n}|m(X_i) - Y_i|^2 + \lambda J^2_p(m) - \frac{1}{n}\sum_{i=1}^{n}|m(X_i) - Y_i|^2 + pen_n(p, \lambda)\Bigg\} \\
&= 2 \inf_{\lambda\in\Lambda_n}\Big\{\lambda J^2_p(m) + pen_n(p, \lambda)\Big\}.
\end{aligned}$$

Step 2. Let $t > 0$ be arbitrary. We show that

$$\begin{aligned} P\{T_{1,n} > t\} \leqslant \sum_{(k,\lambda)\in\mathcal{K}_n\times\Lambda_n} \sum_{l=1}^{\infty} P\Bigg\{ &\exists f = T_L g,\ g \in W^k([0,1]^d),\ J_k^2(g) \leqslant \frac{2^l\, pen_n(k,\lambda)}{\lambda}: \\ &\frac{\mathrm{E}|f(X)-Y|^2 - \mathrm{E}|m(X)-Y|^2 - (1/n)\sum_{i=1}^{n}\{|f(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\}}{t + 2^l\, pen_n(k,\lambda) + \mathrm{E}|f(X)-Y|^2 - \mathrm{E}|m(X)-Y|^2} > \frac{1}{2}\Bigg\}. \end{aligned}$$

To see this, write

$$R(f) = |f(X)-Y|^2 - |m(X)-Y|^2, \qquad R_n(f) = \frac{1}{n}\sum_{i=1}^{n}\{|f(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\}$$

and observe that

$$\begin{aligned} &P\{T_{1,n} > t\} \\ &\leqslant P\Bigg\{\mathrm{E}[R(m_{n,(k^*,\lambda^*)})|\mathcal{D}_n] - R_n(m_{n,(k^*,\lambda^*)}) \\ &\quad > \frac{1}{2}\Big(t + 2\lambda^* J_{k^*}^2(\tilde{m}_{n,(k^*,\lambda^*)}) + 2\,pen_n(k^*,\lambda^*) + \mathrm{E}[R(m_{n,(k^*,\lambda^*)})|\mathcal{D}_n]\Big)\Bigg\} \\ &\leqslant \sum_{(k,\lambda)\in\mathcal{K}_n\times\Lambda_n} P\Bigg\{\exists f = T_L g,\ g\in W^k([0,1]^d): \frac{\mathrm{E}R(f) - R_n(f)}{t + 2\lambda J_k^2(g) + 2\,pen_n(k,\lambda) + \mathrm{E}R(f)} > \frac{1}{2}\Bigg\} \\ &\leqslant \sum_{(k,\lambda)\in\mathcal{K}_n\times\Lambda_n} \sum_{l=1}^{\infty} P\Bigg\{\exists f = T_L g,\ g\in W^k([0,1]^d), \\ &\qquad 2^l\, pen_n(k,\lambda) \leqslant 2\lambda J_k^2(g) + 2\,pen_n(k,\lambda) < 2^{l+1}\, pen_n(k,\lambda): \\ &\quad \frac{\mathrm{E}R(f) - R_n(f)}{t + 2\lambda J_k^2(g) + 2\,pen_n(k,\lambda) + \mathrm{E}R(f)} > \frac{1}{2}\Bigg\} \\ &\leqslant \sum_{(k,\lambda)\in\mathcal{K}_n\times\Lambda_n} \sum_{l=1}^{\infty} P\Bigg\{\exists f = T_L g,\ g\in W^k([0,1]^d),\ J_k^2(g) \leqslant 2^l\, \frac{pen_n(k,\lambda)}{\lambda}: \\ &\qquad \frac{\mathrm{E}R(f) - R_n(f)}{t + 2^l\, pen_n(k,\lambda) + \mathrm{E}R(f)} > \frac{1}{2}\Bigg\}. \end{aligned}$$

Step 3. Now fix $(k,\lambda)\in\mathcal{K}_n\times\Lambda_n$ and $l\in\mathbb{N}$. We show, for n sufficiently large, that

$$P\left\{\exists f = T_L g,\ g \in W^k([0, 1]^d),\ J_k^2(g) \leqslant \frac{2^l pen_n(k, \lambda)}{\lambda}: \frac{\mathrm{E}R(f) - R_n(f)}{t + 2^l pen_n(k, \lambda) + \mathrm{E}R(f)} > \frac{1}{2}\right\}$$

$$\leqslant c_3 \exp\left(-c_4 \frac{n \cdot (t + 2^l pen_n(k, \lambda))}{L^4}\right).$$

This inequality follows directly from Lemma 2 provided we can show that the assumptions of Lemma 2 are satisfied. Set

$$\mathcal{F} = \Big\{f: \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}: f(x, y) = |T_L g(x) - T_L y|^2 - |m(x) - T_L y|^2 \qquad ((x, y) \in \mathbb{R}^d \times \mathbb{R})$$

$$\text{for some } g \in W^k([0, 1]^d),\ J_k^2(g) \leqslant 2^l \frac{pen_n(k, \lambda)}{\lambda}\Big\}$$

and $Z_i = (X_i, Y_i)$ $i = 1, \ldots, n$. Then the above probability can be rewritten as

$$P\left\{\sup_{f \in \mathcal{F}} \frac{\mathrm{E}f(Z_1) - (1/n)\sum_{i=1}^n f(Z_i)}{t + 2^l pen_n(k, \lambda) + \mathrm{E}f(Z_1)} > \frac{1}{2}\right\}.$$

Hence it suffices to show that, for the set $\mathcal{F}_n$ of functions, $\alpha = t + 2^l pen_n(k, \lambda)$, $\epsilon = \frac{1}{2}$ and suitable values of K_1 and K_2, the assumptions of Lemma 2 are satisfied. Recall that the penalty is defined by

$$pen_n(k, \lambda) = \frac{L^5(\log(n))^2}{n \cdot \lambda^{d/2k}} \qquad (k, \lambda) \in \mathcal{K}_n \times \Lambda_n.$$

We first determine K_1 and K_2. For $f \in \mathcal{F}$, we have $|f(z)| \leqslant 4L^2$, $z \in \mathbb{R}^d \times \mathbb{R}$, and

$$\begin{aligned}
\mathrm{E}|f(Z)|^2 &= \mathrm{E}\{|\,|(T_L g)(X) - Y|^2 - |m(X) - Y|^2|^2\} \\
&= \mathrm{E}\{|((T_L g)(X) - Y) - (m(X) - Y)|^2 \cdot |((T_L g)(X) - Y) + (m(X) - Y)|^2\} \\
&\leqslant 16L^2 \mathrm{E}|(T_L g)(X) - m(X)|^2 = 16L^2 \mathrm{E}f(Z).
\end{aligned}$$

So we can choose $K_1 = 4L^2$ and $K_2 = 16L^2$.

Next we show that (5) holds for n sufficiently large. Using $\lambda \leqslant \log(n)$ for $\lambda \in \Lambda_n$, this follows from

$$\left(\sqrt{n}\sqrt{1-\epsilon}\sqrt{\alpha}\right)^2 = \frac{n}{2} \cdot \left(t + 2^l \frac{L^5(\log(n))^2}{n \cdot \lambda^{d/2k}}\right) \geqslant \frac{L^5(\log(n))^2}{(\log(n))^{d/2k}} \to \infty \qquad n \to \infty,$$

because of $2k > d$ for $k \in \mathcal{K}_n$.

So it remains to show that (6) holds for n sufficiently large. In order to bound the covering number, we observe that

$$\frac{1}{n}\sum_{i=1}^{n}\left|(|T_L g_1(x_i) - T_L y_i|^2 - |m(x_i) - T_L y_i|^2) - (|T_L g_2(x_i) - T_L y_i|^2 - |m(x_i) - T_L y_i|^2)\right|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}|T_L g_1(x_i) - T_L g_2(x_i)|^2 \cdot |T_L g_1(x_i) + T_L g_2(x_i) - 2T_L y_i|^2$$

$$\leqslant 16L^2\frac{1}{n}\sum_{i=1}^{n}|T_L g_1(x_i) - T_L g_2(x_i)|^2$$

and obtain

$$\mathcal{N}_2(u, \mathcal{F}, z_1^n) \leqslant \mathcal{N}_2\left(\frac{u}{4L}, \left\{T_L g\colon g \in W^k([0, 1]^d),\ J_k^2(g) \leqslant 2^l\frac{pen_n(k, \lambda)}{\lambda}\right\}, x_1^n\right).$$

This, together with Lemma 3, implies, for any $u \geqslant 1/n$:

$$\log\mathcal{N}_2\left(u, \left\{f \in \mathcal{F}: \frac{1}{n}\sum_{i=1}^{n}f(z_i)^2 \leqslant 4\delta\right\}, z_1^n\right) \leqslant \log\mathcal{N}_2(u, \mathcal{F}, z_1^n)$$

$$\leqslant \log\mathcal{N}_2\left(\frac{u}{4L}, \left\{T_L g\colon g \in W^k([0, 1]^d),\ J_k^2(g) \leqslant 2^l\frac{pen_n(k, \lambda)}{\lambda}\right\}, x_1^n\right)$$

$$\leqslant c_d(k^d + 1)\left(\left(\frac{\sqrt{2^l pen_n(k, \lambda)/\lambda}}{u/4L}\right)^{d/k} + 1\right)\log\left(\frac{64\mathrm{e}L^2 n}{u^2/16L^2}\right)$$

$$\leqslant c_d(k^d + 1)\left(\left(4L\frac{\sqrt{2^l pen_n(k, \lambda)/\lambda}}{u}\right)^{d/k} + 1\right)\log(1024\mathrm{e}L^4 n^3).$$

Using $k \leqslant d/2 + (\log(n))^{1/2d}$ for $k \in \mathcal{K}_n$, we obtain, for n sufficiently large,

$$\left(\log\mathcal{N}_2\left(u, \left\{f \in \mathcal{F}: \frac{1}{n}\sum_{i=1}^{n}f(z_i)^2 \leqslant 4\delta\right\}, z_1^n\right)\right)^{1/2}$$

$$\leqslant c_5(\log(n))^{3/4}\left(\left(2^l pen_n(k, \lambda)/\lambda\right)^{d/4k} \cdot u,{}^{-d/2k} + 1\right)$$

and

$$\int_{\alpha/8K_2}^{\sqrt{\delta}}\left(\log\mathcal{N}_2\left(u, \left\{f \in \mathcal{F}: \frac{1}{n}\sum_{i=1}^{n}f(z_i)^2 \leqslant 4\delta\right\}, z_1^n\right)\right)^{1/2}\mathrm{d}u$$

$$\leqslant c_5(\log(n))^{3/4}\left(\left(2^l pen_n(k, \lambda)/\lambda\right)^{d/4k}\int_0^{\sqrt{\delta}}u^{-d/2k}\mathrm{d}u + \sqrt{\delta}\right)$$

$$= c_6(\log(n))^{3/4}((2^l pen_n(k, \lambda)/\lambda)^{d/4k}\delta^{1/2-d/(4k)} + \sqrt{\delta}).$$

Hence (6) is implied by

$$\sqrt{n}\delta \geqslant c_7(\log(n))^{3/4}\Big(\big(2^l pen_n(k,\lambda)/\lambda\big)^{d/4k}\delta^{1/2-d/4k}+\sqrt{\delta}\Big) \tag{13}$$

for all $\delta \geqslant \alpha/4$. Since

$$\begin{aligned}\frac{\sqrt{n}\alpha}{(\log(n))^{3/4}\big(2^l pen_n(k,\lambda)/\lambda\big)^{d/4k}\alpha^{1/2-d/(4k)}} &= \frac{\sqrt{n}\alpha^{1/2+d/4k}\lambda^{d/4k}}{(\log(n))^{3/4}(2^l pen_n(k,\lambda))^{d/4k}}\\ &\geqslant \frac{\sqrt{n}(pen_n(k,\lambda))^{1/2}\lambda^{d/(4k)}}{(\log(n))^{3/4}} = L^{5/2}(\log(n))^{1/4}\to\infty, \qquad n\to\infty,\end{aligned}$$

and

$$\begin{aligned}\frac{\sqrt{n}\alpha}{\alpha^{1/2}} = \sqrt{n}\alpha^{1/2} \geqslant \sqrt{n}(2^l pen_n(k,\lambda))^{1/2} &= \sqrt{n}\left(2^l\frac{L^5(\log(n))^2}{n\cdot\lambda^{d/2k}}\right)^{1/2}\\ &\geqslant 2^{l/2}L^{5/2}(\log(n))^{1-d/4k}\to\infty, \qquad n\to\infty,\end{aligned}$$

(13) and hence also (6) hold for n sufficiently large.

Step 4. We show, for n sufficiently large, that $\mathrm{E}T_{1,n} \leqslant c_9/n$. Using the results of steps 2 and 3, we obtain, for n sufficiently large, that

$$\begin{aligned}\mathrm{E}T_{1,n} &\leqslant \int_0^\infty P\{T_{1,n}>t\}\mathrm{d}t\\ &\leqslant \sum_{(k,\lambda)\in\mathcal{K}_n\times\Lambda_n}\sum_{l=1}^{\infty}\int_0^\infty c_3\exp\left(-\frac{c_4 n\cdot(t+2^l pen_n(k,\lambda))}{L^4}\right)\mathrm{d}t\\ &= \sum_{(k,\lambda)\in\mathcal{K}_n\times\Lambda_n}\sum_{l=1}^{\infty}\exp\left(-\frac{c_4 n 2^l pen_n(k,\lambda)}{L^4}\right)\cdot\frac{c_3\cdot L^4}{c_4 n}\\ &\leqslant \sum_{(k,\lambda)\in\mathcal{K}_n\times\Lambda_n}\sum_{l=1}^{\infty}\exp\big(-c_4 2^l L^5\cdot(\log(n))^{2-d/2k}\big)\cdot\frac{c_3 L^4}{c_4 n}\\ &\leqslant \sum_{(k,\lambda)\in\mathcal{K}_n\times\Lambda_n}\exp\big(-c_4 L^5\cdot(\log(n))^{2-d/2k}\big)\cdot\frac{c_3 L^4}{c_4 n}\\ &\leqslant c_8 n\cdot(\log(n))^{1/2d}\exp(-2\log(n))\cdot\frac{c_3 L^4}{c_4 n}\\ &\leqslant \frac{c_9}{n}.\end{aligned}$$

Step 5. By the results of steps 1 and 4, we obtain, for n sufficiently large,

$$\mathrm{E}\int |m_n(x) - m(x)|^2 P_X(\mathrm{d}x) \leqslant 2 \inf_{\lambda \in \Lambda_n} \left\{ \lambda \cdot J_p^2(m) + \frac{L^5(\log(n))^2}{n \cdot \lambda^{d/2p}} \right\} + \frac{c_9}{n}.$$

Clearly, this implies the assertion of part (a). Concerning (b), assume $0 < J_p^2(m) < \infty$ and set

$$\lambda^* = \left(\frac{L^5(\log(n))^2}{n \cdot J_p^2(m)} \right)^{2p/(2p+d)}.$$

Then, for n sufficiently large, there exists $\bar{\lambda} \in \Lambda_n$ such that $\lambda^* \leqslant \bar{\lambda} \leqslant 2\lambda^*$. It follows that

$$\begin{aligned}
\mathrm{E}\int |m_n(x) - m(x)|^2 P_X(\mathrm{d}x) &\leqslant 2\left\{ \bar{\lambda} \cdot J_p^2(m) + \frac{L^5(\log(n))^2}{n \cdot \bar{\lambda}^{d/2p}} \right\} + \frac{c_9}{n} \\
&\leqslant 2\left\{ 2\lambda^* \cdot J_p^2(m) + \frac{L^5(\log(n))^2}{n \cdot (\lambda^*)^{d/2p}} \right\} + \frac{c_9}{n} \\
&\leqslant 6 \cdot (J_p^2(m))^{d/(2p+d)} \left(\frac{L^5(\log(n))^2}{n} \right)^{2p/(2p+d)} + \frac{c_9}{n} \\
&= O\Big((J_p^2(m))^{d/(2p+d)} (\log(n))^2 n^{-2p/(2p+d)} \Big).
\end{aligned}$$

□

Acknowledgements

The first author's research was supported by the Deutsche Forschungsgemeinschaft. The second author was supported by the National Sciences and Engineering Research Council of Canada and the Alexander von Humboldt Foundation, the third author by the German National Scholarship Foundation. The authors wish to thank two anonymous referees for helpful comments.

References

Baraud, Y. (1997) Model selection for regression on random design. Submitted.

Barron, A.R., Birgé, L. and Massart, P. (1999) Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, **113**, 301–413.

Devroye, L., Györfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag.

Devroye, L.P. and Wagner, T.J. (1980) Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.*, **8**, 231–239.

Duchon, J. (1976) Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *RAIRO Analyse Numérique*, **10**, 5–12.

Eubank, R.L. (1988) *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker.

Kohler, M. (1998) Nonparametric regression function estimation using interaction least squares splines and complexity regularization. *Metrika*, **47**, 147–163.

Kohler, M. (2000) Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *J. Statist. Plann. Inference*, **89**, 1–23.

Kohler, M. and Krzyżak, A. (2001) Nonparametric regression estimation using penalized least-squares. *IEEE Trans. Inform. Theory*, **47**, 3054–3058.

Krzyżak, A. and Linder, T. (1998) Radial basis function networks and complexity regularization in function learning. *IEEE Trans. Neural Networks*, **9**, 247–256.

Stone, C.J. (1977) Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–645.

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

van de Geer, S. (1987) A new approach to least-squares estimation, with applications. *Ann. Statist.*, **15**, 587–602.

van de Geer, S. (1988) *Regression Analysis and Empirical Processes,* CWI Tract 45. Amsterdam: Centrum voor Wiskunde en Informatica.

van de Geer, S. (1990) Estimating a regression function. *Ann. Statist.*, **18**, 907–924.

van de Geer, S. (2000) *Applications of Empirical Process Theory.* Cambridge: Cambridge University Press.

van de Geer, S. (2001) Least squares estimation with complexity penalties. *Math. Meth. Statist.*, **10**, 355–374.

Vapnik, V.N. (1998) *Statistical Learning Theory.* New York: Wiley.

Vapnik, V.N. and Chervonenkis, A.Y. (1974) *The Theory of Pattern Recognition*. Moscow: Nauka.

Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society of Industrial and Applied Mathematics.

Received March 2000 and revised February 2001