CONSISTENCY OF RIDGE FUNCTION FIELDS FOR VARYING NONPARAMETRIC REGRESSION

Robert FROUIN [a] and Bruno PELLETIER [b,*]

[a] Scripps Institution of Oceanography
University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0224, USA
rfrouin@ucsd.edu

[a] Institut de Mathématiques et de Modélisation de Montpellier
UMR CNRS 5149, Equipe de Probabilités et Statistique
Université Montpellier II, CC 051
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
pelletier@math.univ-montp2.fr

Abstract

A nonparametric regression model proposed in [Pelletier and Frouin, *Applied Optics*, 2006] as a solution to the geophysical problem of ocean color remote sensing is studied. The model, called *ridge function field*, combines a regression estimate in the form of a superposition of ridge functions, or equivalently a neural network, with the idea pertaining to varying-coefficients models, where the parameters of a parametric family are allowed to vary with other variables. Under mild assumptions on the underlying distribution of the data, the strong universal consistency of the least-squares ridge function fields estimate is established.

Index Terms — Varying coefficients model, ridge function approximation, nonparametric regression, universal consistency, least squares.

AMS 2000 Classification: 62G08, 62G05.

[*] Corresponding author.

1 Introduction

The motivation for the present work comes from the geophysical problem of ocean color remote sensing from space, which we shall describe first. In this problem, the aim is to predict the value of an oceanic parameter Y (e.g., the surface phytoplankton pigment concentration) from a vector X of radiometric measurements at several wavelengths acquired by a sensor onboard a satellite platform. This is in fact an inverse problem where the measurements X depend on Y and other parameters, such as the aerosol optical properties, through a given forward operator (governing the radiative transfer in the ocean-atmosphere system) contaminated by a random measurement noise. In particular, X depends on a vector T of angular variables describing the relative positions of the Sun and the sensor with respect to the target; as such, T is observed simultaneously with X. The difficulty, though, is that the operator to be inverted, resulting from modeling of scattering and absorption processes at the micro-physical scale, is rather complex. To circumvent this issue, the approach taken up in Pelletier and Frouin (2004, 2006) and Frouin and Pelletier (2007) consists in i) sampling the forward operator according to some prior distribution, ii) selecting a reasonable noise model, and iii) estimating the regression function m of Y on X and T from the data $(X_1, T_1, Y_1), \dots, (X_n, T_n, Y_n)$, where m is defined by

$$m(x,t) = \mathbb{E}_{X=x,T=t}\big[Y|X,T\big].$$

In this display, the variable T is independent from the response Y but not from the predictor X. As such, T acts as an *effect modifier* in the sense of Hastie and Tibshirani (1993) in the context of varying coefficient models, i.e., the shape of the relation explaining Y from X is affected by the values taken by T. Let us recall that a varying-coefficient model is a linear model the parameters of which are allowed to vary with other variables. Since the work of Hastie and Tibshirani (1993), varying-coefficient models have been studied by several authors (see e.g., Fan and Zhang, 1999, 2000; Fan, Yao, and Cai, 2003, Wong, Ip, and Zhang, 2008, and the references therein). By relaxing to some extent the parametric constraint, these models have proved useful in various application contexts, and in particular in a high-dimensional setting.

As pointed out by Hastie and Tibshirani (1993) as well as Fan and Zhang (1999), the idea of allowing the parameters of a linear model to vary with other variables is not new, and is widely used in the literature. More generally, one may start with an arbitrary parametric family in place of a linear model, and allow its parameters to vary with other variables. This idea is developed in Pelletier and Frouin (2004, 2006) and Frouin and Pelletier (2007) where we considered models of the form

$$\zeta_n^* : (x,t) \mapsto \zeta_n^*(x,t) := f\big(x;\theta_n(t)\big). \tag{1.1}$$

In model (1.1), the function $f(.;\theta_n)$ belongs to a set $\mathcal{R}_n$ of functions, about which more will be said later, parameterized by a vector θ_n, which in (1.1) is allowed to vary with t. Additionally, to obtain an implementable model, each coordinate map of the application $t \mapsto \theta_n(t)$ in (1.1) is taken to lie in a parametric set $\mathcal{T}_n$ of functions of t. Now suppose that X and T are taking values in some subsets $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{S} \subseteq \mathbb{R}^p$, respectively. Another way of formalizing this construction is to view the model (1.1) as a function of x being "attached" to each point t in $\mathcal{S}$, i.e., in mathematical terms, as a map

$$\zeta_n : \mathcal{S} \to \mathcal{R}_n, \tag{1.2}$$

such that the map ζ_n^* in (1.1) is the representation of ζ_n over the product space $\mathcal{X} \times \mathcal{S}$, i.e., we let $\zeta_n^*(x,t) = \zeta_n(t)(x)$. Thus the representation (1.2) highlights the fact that the variable T acts as an effect modifier.

In the experiments reported in Pelletier and Frouin (2004, 2006) and Frouin and Pelletier (2007), the sequence $\mathcal{R}_n$ of functions in (1.2) is taken as the set spanned by functions of the ridge form or, equivalently, neural networks, and the resulting models (1.1)-(1.2) is called a *ridge function field* over $\mathcal{S}$. The sets $\mathcal{R}_n$ form a nested sequence of the form $\mathcal{R}_1 \subseteq \mathcal{R}_2 \subseteq \ldots$. There exists a vast literature on regression estimation with these ridge functions and neural networks. In particular, it is known that the union of sets spanned by ridge functions is dense in $L_2(\mu)$, for any probability measure μ (see, Cybenko, 1989; Hornik, Stinchcombe, and White, 1989; Barron, 1993; Lin and Pinkus, 1993; Burger and Neubaeur, 2001; Maiorov, 1999), and that they yield consistent nonparametric regression procedures (see e.g. Gyorfi, Kohler, Krzyzak, and Walk, 2002, Chapter 16). Similarly, the sets $\mathcal{T}_n$ have been chosen to form an increasing sequence $\mathcal{T}_1 \subseteq \mathcal{T}_2 \subseteq \ldots$ of sets of real-valued functions on $\mathcal{S}$.

Density results for ridge function fields have been obtained in Pelletier (2004) in the compact-open topology. In particular, under mild assumptions on the sets $(\mathcal{R}_n)_n$ and $(\mathcal{T}_n)_n$, the union of ridge function fields is dense. Thus under these conditions, ridge function fields are suitable candidates for nonparametric regression, since the approximation error of the estimate will decrease to 0 as n increases. The purpose of the present work is to establish the strong universal consistency of the least-squares regression estimate in the form of (1.1)-(1.2).

The paper is organized as follows. Section 2 introduces ridge function fields and presents our main result (Theorem 2.1), which states that ridge function fields adjusted by least squares are *strongly universally consistent*, i.e., that for any distribution of (X,T,Y) satisfying mild assumptions, the L_2 error of the regression estimate converges to 0 with probability 1 as the sample size tends to infinity. Section 3 is devoted to the proof of Theorem 2.1.

2 Ridge function fields

Let (X,T,Y) be a random object, where X, T, and Y take values in $\mathbb{R}^d$, $\mathbb{R}^p$, and $\mathbb{R}$, respectively. We shall assume that X, T, and Y are bounded, i.e., that there exists positive constants C_X, C_T and C_Y such that

$$|X| \le C_X, \quad |T| \le C_T, \quad |Y| \le C_Y, \quad \text{with probability } 1. \tag{2.1}$$

Given n independent copies $(X_1,T_1,Y_1),\dots,(X_n,T_n,Y_n)$ of (X,T,Y), our aim is to construct an estimate of the regression function m of Y on the pair (X,T), i.e.,

$$m(x,t) = \mathbb{E}\big[Y|(X,T)=(x,t)\big].$$

Let us start with the definition of a ridge function. A *ridge function* on $\mathbb{R}^d$ is a function of the form $\sigma\big(\langle a,x\rangle\big)$ where $\sigma : \mathbb{R} \to \mathbb{R}$ is a given function, where $a \in \mathbb{R}^d$, and where $\langle .,.\rangle$ denotes the usual scalar product on $\mathbb{R}^d$. Approximation by ridge functions refers to approximation by linear combinations of ridge functions. There exists several variants of approximation by ridge functions, according to whether the function σ is fixed or not. In the following, we shall consider the family $\mathcal{R}_n$ of functions on $\mathbb{R}^d$ defined by

$$\begin{aligned}\mathcal{R}_n = \Big\{ x \mapsto \sum_{j=1}^{K_n} c_j \sigma\big(\langle a_j,x\rangle + b_j\big) + c_0 \,:\, & a_j \in \mathbb{R}^d\,,\, b_j \in \mathbb{R}\,,\, c_0, c_j \in \mathbb{R}\,,\\ & j=1,\dots,K_n \Big\},\end{aligned}$$

where K_n is an integer, and where $\sigma : \mathbb{R} \to \mathbb{R}$ is a fixed *squashing function*, i.e., σ is nondecreasing, and satisfies the following two conditions:

$$\lim_{u \to -\infty} \sigma(u) = 0 \quad \text{and} \quad \lim_{u \to \infty} \sigma(u) = 1.$$

Next, let $\Psi_1, \Psi_2, \ldots : \mathbb{R}^p \to \mathbb{R}$ be bounded basis functions, the linear span of which is dense in $\mathcal{C}\big(\mathbb{R}^p\big)$ in the topology of uniform convergence on compact sets, i.e., the set

$$\bigcup_{k=1}^{\infty} \Big\{ t \mapsto \sum_{j=1}^{k} a_j \Psi_j(t) \ : \ a_1, \ldots, a_k \in \mathbb{R} \Big\}$$

is dense in $\mathcal{C}\big(\mathbb{R}^p\big)$. Without loss of generality, we assume that $|\Psi_i| \leq 1$ for all $i \geq 1$. Next, given an integer $L \in \mathbb{N}$ and a real number $\rho > 0$, we shall consider the following families of functions on $\mathbb{R}^p$:

$$\begin{aligned} \mathcal{T}(L) &= \Big\{ t \mapsto \sum_{j=1}^{L} a_j \Psi_j(t) \ : \ a_1, \ldots, a_k \in \mathbb{R} \Big\}, \\ \mathcal{T}_\rho(L) &= \Big\{ t \mapsto \sum_{j=1}^{L} a_j \Psi_j(t) \ : \ a_1, \ldots, a_k \in \mathbb{R}\,, \sum_{j=1}^{L} |a_j| \leq \rho \Big\}. \end{aligned}$$

Then, given integers L_n^a, L_n^b, and L_n^c, and a positive number ρ_n, we define the family $\mathcal{F}_n$ of ridge function fields by

$$\begin{aligned} \mathcal{F}_n = \Big\{ (x,t) \mapsto \sum_{j=1}^{K_n} c_j(t) \sigma\big(\langle a_j(t), x\rangle + b_j(t)\big) + c_0(t) \ : \ a_j \in \big(\mathcal{T}(L_n^a)\big)^d, \\ b_j \in \mathcal{T}(L_n^b)\,,\, c_0, c_j \in \mathcal{T}_{\rho_n}(L_n^c)\,,\, j = 1, \ldots, K_n \Big\}. \end{aligned}$$

We shall also consider the subset $\tilde{\mathcal{F}}_n$ of $\mathcal{F}_n$ where only the coefficients c_j are allowed to vary with the variable t, i.e., we define

$$\begin{aligned} \tilde{\mathcal{F}}_n = \Big\{ (x,t) \mapsto \sum_{j=1}^{K_n} c_j(t) \sigma\big(\langle a_j, x\rangle + b_j\big) + c_0(t) \ : \ a_j, b_j \in \mathbb{R}, c_0, c_j \in \mathcal{T}_{\rho_n}(L_n^c)\,, \\ j = 1, \ldots, K_n \Big\}. \end{aligned}$$

Both sets $\mathcal{F}_n$ and $\tilde{\mathcal{F}}_n$ are dense in the topology of uniform convergence on compact sets (Pelletier, 2004). Consequently, $\mathcal{F}_n$ and $\tilde{\mathcal{F}}_n$ are also dense in L_2 for any distribution with bounded support and, as exposed in the Introduction, the approximation error of the estimates will decrease to 0 as $n \to \infty$.

Let us first define the maps $m_n^\dagger$ and $\tilde{m}_n^\dagger$ as any minimizer of the empirical L_2 risk over $\mathcal{F}_n$ and $\tilde{\mathcal{F}}_n$, respectively, i.e., m_n and $\tilde{m}_n$ are such that

$$\frac{1}{n}\sum_{i=1}^{n}\Big(m_n^\dagger(X_i,T_i)-Y_i\Big)^2 = \inf_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}\Big(f(X_i,T_i)-Y_i\Big)^2, \tag{2.2}$$

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\tilde{m}_n^\dagger(X_i,T_i)-Y_i\Big)^2 = \inf_{\tilde{f}\in\tilde{\mathcal{F}}_n}\frac{1}{n}\sum_{i=1}^{n}\Big(\tilde{f}(X_i,T_i)-Y_i\Big)^2. \tag{2.3}$$

Then we define the estimates m_n and $\tilde{m}_n$ as truncated versions of $m_n^\dagger$ and $\tilde{m}_n^\dagger$ respectively, i.e.,

$$m_n(x,t) = T_{\beta_n}m_n^\dagger(x,t), \tag{2.4}$$

$$\tilde{m}_n(x,t) = T_{\beta_n}\tilde{m}_n^\dagger(x,t), \tag{2.5}$$

where $(\beta_n)_n$ is a sequence of positive numbers such that $\beta_n \to \infty$, and where T_{β_n} is the truncation operator defined by $T_{\beta_n}u = \min\{|u|,\beta_n\}\,\mathrm{sign}(u)$.

We are now in a position to state our main result.

Theorem 2.1 *Let m_n and $\tilde{m}_n$ be the truncated least-squares estimates defined by* (2.2)*,* (2.3)*,* (2.4)*, and* (2.5)*.*

(i) If

$$K_n \to \infty,\quad L_n^a \to \infty,\quad L_n^b \to \infty,\quad L_n^c \to \infty,\quad \beta_n \to \infty,\quad \rho_n \to \infty,$$

as $n \to \infty$ in such a way that

$$\frac{K_n\beta_n^4(L_n^a+L_n^b+L_n^c)\log(\beta_n\rho_nK_n)}{n} \to 0 \quad \textit{and} \quad \frac{\beta_n^4}{n^{1-\delta}} \to 0,$$

for some $\delta > 0$ as $n \to \infty$, then

$$\lim_{n\to\infty}\int\big(m_n(x,t)-m(x,t)\big)^2\mu(dx,dt) = 0 \quad \textit{with probability 1},$$

for every distribution of (X,T,Y) satisfying (2.1).

(ii) If

$$K_n \to \infty, \quad L_n^c \to \infty, \quad \beta_n \to \infty, \quad \textit{and} \quad \rho_n \to \infty,$$

as $n \to \infty$ in such a way that

$$\frac{K_n \beta_n^4 L_n^c \log(\beta_n \rho_n K_n)}{n} \to 0 \quad \textit{and} \quad \frac{\beta_n^4}{n^{1-\delta}} \to 0,$$

for some $\delta > 0$ as $n \to \infty$, then

$$\lim_{n\to\infty} \int \big(\tilde{m}_n(x,t) - m(x,t)\big)^2 \mu(dx, dt) = 0 \quad \textit{with probability 1},$$

for every distribution of (X, T, Y) satisfying (2.1).

Theorem 2.1 states that nonparametric regression estimation with ridge function fields is a consistent procedure as long as the underlying distribution satisfies (2.1). Note first that the extension to a multivariate response Y is straightforward, and second that the family $\tilde{\mathcal{F}}_n$ is a subset of the family $\mathcal{F}_n$. Consequently, $\mathcal{F}_n$ offers more flexibility than $\tilde{\mathcal{F}}_n$ but, as expected, at the expense of an increased complexity of the fitting algorithm. More generally, one may consider replacing the family $\mathcal{R}_n$ in the definition of a set of function fields by any other nested sequence of models used in nonparametric regression. That said, regarding the ocean color remote sensing problem, the choice of using ridge functions has been dictated not only by their approximation properties, but more importantly by the speed of execution of these models, which should be high for processing large data sets, as provided by satellite imagery. Models belonging to $\mathcal{F}_n$ have been implemented and evaluated in Pelletier and Frouin (2006) for the retrieval of the chlorophyll-a concentration, i.e., a univariate response, while the family $\tilde{\mathcal{F}}_n$ has been used in Frouin and Pelletier (2007) for the retrieval of the spectral marine reflectance, a multivariate output. In practice, the minimization of the empirical L_2 risk in (2.2) and (2.3) may be solved using a stochastic gradient descent algorithm (see Pelletier and Frouin, 2006 for details).

Let us also mention that assumption (2.1), which is reasonable from a physical perspective, may be weakened at the price of a little extra work. As a matter of fact, one may assume that the response Y is unbounded and use a truncation argument, as in, e.g., Kohler and Krzyzak (2005), and Gyorfi, Kohler, Krzyzak, and Walk (2002, Theorem 10.2). Next, if X and T are unbounded, it is immediate to show that the union of the sets $\mathcal{F}_n$ is dense in $L_2(\mu)$ for any distribution μ

provided that the sets $\mathcal{T}(L)$ contain the constant and affine functions of t. On the other hand, a little extra work would be needed to derive the density in L_2 of the union of the $\tilde{\mathcal{F}}_n$'s. The study of the convergence rates is left for future research. In this perspective, let us emphasize that the respective influences of the variables X and T on the response Y are well separated in a ridge function field, by construction, which may prove useful for adaptation over anisotropic classes of regression functions (see e.g. Hofmann and Lepski, 2002).

3 Proofs

3.1 Technical Lemmas

We shall need the following technical Lemmas to derive an upper-bound on the covering number of $\mathcal{F}_n$. Let us start by introducing the notations and definitions used hereafter.

First of all, let $\mathcal{G}$ be a set of real valued functions on $\mathbb{R}^d$, and let $x_1, \ldots, x_n$ be n fixed points in $\mathbb{R}^d$. For any functions $g_1, g_2 : \mathbb{R}^d \to \mathbb{R}$, set

$$\operatorname{dist}_{1,n}(g_1, g_2) = \mathbb{P}_n\big(|g_1 - g_2|\big),$$

where $\mathbb{P}_n$ is the discrete uniform measure on $\{x_1, \ldots, x_n\}$. The ε-covering number of $\mathcal{G}$ with respect to $\operatorname{dist}_{1,n}$ is called the *L_1 ε-covering number of $\mathcal{G}$ on* $\{x_1, \ldots, x_n\}$ and will be denoted by $\mathcal{N}_1\big(\varepsilon, \mathcal{G}, x_1^n\big)$, i.e., $\mathcal{N}_1\big(\varepsilon, \mathcal{G}, x_1^n\big)$ is the smallest integer N such that there exists functions $g_1, \ldots, g_n$ such that

$$\min_{1 \leq j \leq N} \frac{1}{n} \sum_{i=1}^{n} |g(x_i) - g_j(x_i)| \leq \varepsilon,$$

for all $g \in \mathcal{G}$.

Similarly to the above, the *L_1 ε-packing number of $\mathcal{G}$ on* $\{x_1, \ldots, x_n\}$, further denoted by $\mathcal{M}_1(\varepsilon, \mathcal{G}, x_1^n)$, is the maximal integer N such that there exists functions $g_1, \ldots, g_N$ in $\mathcal{G}$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \big|g_j(x_i) - g_k(x_i)\big| \geq \varepsilon,$$

for all $1 \le j < k \le N$.

Finally, $\mathcal{G}^+$ will denote the set of all subgraphs of functions in $\mathcal{G}$, i.e.,

$$\mathcal{G}^+ = \big\{(x,u) \in \mathbb{R}^d \times \mathbb{R} \quad : \quad u \le g(x)\,,\, g \in \mathcal{G}\big\},$$

and the Vapnik-Chervonenkis dimension of $\mathcal{G}^+$ will be denoted by $V_{\mathcal{G}^+}$ (Vapnik and Chervonenkis, 1971).

The following results, extracted from Lemmas 16.3, 16.4, and 16.5 in Gyorfi, Kohler, Krzyzak, and Walk (2002), will be needed in the following, and are given to make the paper self-contained. Let $\mathcal{F}$ be a set of functions $\mathbb{R}^d \to \mathbb{R}$. Given a squashing function σ, define $\mathcal{G} = \{\sigma \circ f \,:\, f \in \mathcal{F}\}$. Then

$$V_{\mathcal{G}^+} \le V_{\mathcal{F}^+}. \tag{3.1}$$

Given two sets $\mathcal{F}$ and $\mathcal{G}$ of functions $\mathbb{R}^d \to \mathbb{R}$, set $\mathcal{F} \oplus \mathcal{G} = \{f + g \,:\, f \in \mathcal{F},\ g \in \mathcal{G}\}$, and let $\{x_1, \dots, x_n\}$ be n points in $\mathbb{R}^d$. Then

$$\mathcal{N}_1\big(\varepsilon + \delta, \mathcal{F} \oplus \mathcal{G}, x_1^n\big) \le \mathcal{N}_1\big(\varepsilon, \mathcal{F}, x_1^n\big)\mathcal{N}_1\big(\delta, \mathcal{G}, x_1^n\big). \tag{3.2}$$

Given two sets $\mathcal{F}$ and $\mathcal{G}$ of functions $\mathbb{R}^d \to \mathbb{R}$, uniformly bounded over $\mathbb{R}^d$ by some constants $C_{\mathcal{F}}$ and $C_{\mathcal{G}}$, respectively, set $\mathcal{F} \odot \mathcal{G} = \{fg \,:\, f \in \mathcal{F},\ g \in \mathcal{G}\}$, and let $\{x_1, \dots, x_n\}$ be n points in $\mathbb{R}^d$. Then

$$\mathcal{N}_1\big(\varepsilon + \delta, \mathcal{F} \odot \mathcal{G}, x_1^n\big) \le \mathcal{N}_1\big(\varepsilon / C_{\mathcal{G}}, \mathcal{F}, x_1^n\big)\mathcal{N}_1\big(\delta / C_{\mathcal{F}}, \mathcal{G}, x_1^n\big). \tag{3.3}$$

At last, given a set $\mathcal{G}$ of functions $\mathbb{R}^d \to [0; B]$ with $V_{\mathcal{G}^+} \ge 2$, n points $\{x_1, \dots, x_n\}$ in $\mathbb{R}^d$, and $0 < \varepsilon < B/4$, we have (Gyorfi, Kohler, Krzyzak, and Walk (2002), Theorem 9.4)

$$\mathcal{M}_1\big(\varepsilon, \mathcal{G}, x_1^n\big) \le 3\left(\frac{2eB}{\varepsilon} \log \frac{3eB}{\varepsilon}\right)^{V_{\mathcal{G}^+}}. \tag{3.4}$$

We may now state an upper-bound on the covering number of $\mathcal{F}_n$.

Lemma 3.1 *Let $\mathcal{N}_1\big(\varepsilon, \mathcal{F}_n, (X,T)_1^n\big)$ be the L_1 ε-covering number of $\mathcal{F}_n$ on the sample $(X_1, T_1), \dots, (X_n, T_n)$. Then*

$$\mathcal{N}_1\big(\varepsilon, \mathcal{F}_n, (X,T)_1^n\big) \le 3\left(\frac{108e\rho_n(K_n + 1)}{\varepsilon}\right)^{2(L_n^c+1)+2K_n(L_n^a d+L_n^b+L_n^c+2)}. \tag{3.5}$$

Proof. Consider the following sets of functions:

$$\begin{aligned}
\mathcal{G}_1 &= \big\{(x,t)\mapsto \langle a(t),x\rangle + b(t) \,:\, a_1,\ldots,a_d\in\mathcal{T}(L_n^a)\,,\, b\in\mathcal{T}(L_n^b)\big\},\\
\mathcal{G}_2 &= \big\{(x,t)\mapsto \sigma\big(\langle a(t),x\rangle + b(t)\big) \,:\, a_1,\ldots,a_d\in\mathcal{T}(L_n^a)\,,\, b\in\mathcal{T}(L_n^b)\big\},\\
\mathcal{G}_3 &= \big\{(x,t)\mapsto c(t)\sigma\big(\langle a(t),x\rangle + b(t)\big) \,:\, a_1,\ldots,a_d\in\mathcal{T}(L_n^a)\,,\, b\in\mathcal{T}(L_n^b)\,,\\
&\qquad c\in\mathcal{T}_{\rho_n}(L_n^c)\big\}.
\end{aligned}$$

Since $\mathcal{G}_1$ is a vector space of dimension $L_n^a d + L_n^b$, we have

$$V_{\mathcal{G}_1^+}\le L_n^a d + L_n^b + 1.$$

Next, (3.1) yields

$$V_{\mathcal{G}_2^+}\le V_{\mathcal{G}_1^+}.$$

Consequently, since $|\sigma(u)|\le 1$ for all $u\in\mathbb{R}$, and using (3.4), we obtain

$$\begin{aligned}
\mathcal{N}_1\big(\varepsilon,\mathcal{G}_2,(X,T)_1^n\big) &\le \mathcal{M}_1\big(\varepsilon,\mathcal{G}_2,(X,T)_1^n\big)\\
&\le 3\left[\frac{2e}{\varepsilon}\log\left(\frac{3e}{\varepsilon}\right)\right]^{V_{\mathcal{G}_2^+}}\\
&\le 3\left(\frac{3e}{\varepsilon}\right)^{2(L_n^a d + L_n^b+1)}. \qquad (3.6)
\end{aligned}$$

Now, using the inequality (3.3) leads to

$$\mathcal{N}_1\big(\varepsilon,\mathcal{G}_3,(X,T)_1^n\big)\le \mathcal{N}_1\Big(\frac{\varepsilon}{2},\mathcal{T}_{\rho_n}(L_n^c),(X,T)_1^n\Big)\mathcal{N}_1\Big(\frac{\varepsilon}{2\rho_n},\mathcal{G}_2,(X,T)_1^n\Big).$$

But, using (3.4), we have

$$\begin{aligned}
\mathcal{N}_1\big(\varepsilon,\mathcal{T}_{\rho_n}(L_n^c),(X,T)_1^n\big) &\le \mathcal{M}_1\big(\varepsilon,\mathcal{T}_{\rho_n}(L_n^c),(X,T)_1^n\big)\\
&\le 3\left[\frac{2e(2\rho_n)}{\varepsilon}\log\left(\frac{3e(2\rho_n)}{\varepsilon}\right)\right]^{V_{\mathcal{T}_{\rho_n}(L_n^c)^+}}\\
&\le 3\left(\frac{6e\rho_n}{\varepsilon}\right)^{2(L_n^c+1)}, \qquad (3.7)
\end{aligned}$$

since

$$V_{\mathcal{T}_{\rho_n}(L_n^c)^+}\le V_{\mathcal{T}(L_n^c)^+}\le L_n^c+1,$$

and since $\mathcal{T}(L_n^c)$ is a vector space of dimension L_n^c. Then from (3.6) and (3.7), it follows that

$$\begin{aligned}\mathcal{N}_1\big(\varepsilon, \mathcal{G}_3, (X,T)_1^n\big) &\leq 3\left(\frac{6e\rho_n}{\varepsilon/2}\right)^{2(L_n^c+1)} 3\left(\frac{3e}{\varepsilon/(2\rho_n)}\right)^{2(L_n^a d+L_n^b+1)} \\ &= 9\left(\frac{12e\rho_n}{\varepsilon}\right)^{2(L_n^c+1)} \left(\frac{6e\rho_n}{\varepsilon}\right)^{2(L_n^a d+L_n^b+1)} \\ &\leq 9\left(\frac{12e\rho_n}{\varepsilon}\right)^{2(L_n^a d+L_n^b+L_n^c+2)}. \end{aligned} \tag{3.8}$$

Applying (3.2) yields

$$\mathcal{N}_1\big(\varepsilon, \mathcal{F}_n, (X,T)_1^n\big) \leq \mathcal{N}_1\left(\frac{\varepsilon}{K_n+1}, \mathcal{T}_{\rho_n}(L_n^c), (X,T)_1^n\right)\left[\mathcal{N}_1\left(\frac{\varepsilon}{K_n+1}, \mathcal{G}_3, (X,T)_1^n\right)\right]^{K_n}.$$

Finally, by reporting (3.7) and (3.8) in the equation above, we obtain the upper bound:

$$\begin{aligned}\mathcal{N}_1\big(\varepsilon, \mathcal{F}_n, (X,T)_1^n\big) &\leq 3\left(\frac{6e\rho_n}{\varepsilon/(K_n+1)}\right)^{2(L_n^c+1)}\left[9\left(\frac{12e\rho_n}{\varepsilon/(K_n+1)}\right)^{2(L_n^a d+L_n^b+L_n^c+2)}\right]^{K_n} \\ &\leq 3\left(\frac{108e\rho_n(K_n+1)}{\varepsilon}\right)^{2(L_n^c+1)+2K_n(L_n^a d+L_n^b+L_n^c+2)}.\end{aligned}$$

□

3.2 Proof of Theorem 2.1

By Lemma 10.2 in Gyorfi, Kohler, Krzyzak, and Walk (2002), we have the inequality

$$\begin{aligned}&\int \big|m_n(x,t) - m(x,t)\big|^2 \mu(dx,dt) \\ &\quad\leq \inf_{\{f\in\mathcal{F}_n\,:\,\|f\|_\infty \leq \beta_n\}} \int \big|f(x,t)-m(x,t)\big|\mu(dx,dt) \\ &\qquad + 2\sup_{f\in T_{\beta_n}\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n \big(f(X_i,T_i)-Y_i\big)^2 - \mathbb{E}\Big\{\big(f(X_i,T_i)-Y_i\big)^2\Big\}\right|, \end{aligned} \tag{3.9}$$

where the first term is the approximation error, and where the second term is a uniform upper-bound over $T_{\beta_n}\mathcal{F}_n$ of the estimation error. Since the union of the $\mathcal{F}_n$'s is dense in $\mathcal{C}(\mathbb{R}^d \times \mathbb{R}^p)$ in the topology of uniform convergence on compact sets, and since X and T are bounded by assumption, the union of the $\mathcal{F}_n$'s is dense in $L_2(\mu)$ for every μ with bounded support. Then it follows that the approximation error of m by an element of $\{f \in \mathcal{F}_n : \|f\|_\infty \leq \beta_n\}$ converges to zero as $K_n, L_n^a, L_n^b, L_n^c, \beta_n, \rho_n \to \infty$, and part (i) of the theorem will be proved if we show that

$$\sup_{f \in T_{\beta_n}\mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^{n} \big(f(X_i, T_i) - Y_i\big)^2 - \mathbb{E}\Big\{\big(f(X_i, T_i) - Y_i\big)^2\Big\} \right| \to 0, \qquad (3.10)$$

a.s. as $K_n, L_n^a, L_n^b, L_n^c, \beta_n, \rho_n \to \infty$. The same arguments hold for the family $\tilde{\mathcal{F}}_n$, so that part (ii) of the theorem will be proved if we show that (3.10) is satisfied with $T_{\beta_n}\mathcal{F}_n$ replaced by $T_{\beta_n}\tilde{\mathcal{F}}_n$.

To bound the second term in the right hand side of (3.9), let us first define the set $\mathcal{H}_n$ of functions by

$$\mathcal{H}_n = \big\{h : \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R} \ni (x, t, y) \mapsto |f(x,t) - y|^2 \mathbf{1}_{[-C_Y;C_Y]}(y) \ : \ f \in T_{\beta_n}\mathcal{F}_n\big\},$$

where $\mathbf{1}_{[-C_Y;C_Y]}(.)$ is the characteristic function of the interval $[-C_Y; C_Y]$. Next, we shall set $Z_i = (X_i, T_i, Y_i)$, for $i = 1, \ldots, n$.

For every f in $T_{\beta_n}\mathcal{F}_n$, we have $|f(x,t)| \leq \beta_n$. Hence for n large enough such that $C_Y \leq \beta_n$,

$$|f(x,t) - y|^2 \mathbf{1}_{[-C_Y;C_Y]}(y) \leq 4\beta_n^2,$$

so that the range of each function h in $\mathcal{H}_n$ is included in the interval $\big[0; 4\beta_n^2\big]$. Then from the proof of Theorem 24 in Pollard (1984, p. 25), or Theorem 9.1 in Gyorfi, Kohler, Krzyzak, and Walk (2002, p. 136), we obtain for any $\varepsilon > 0$ the following inequality:

$$\begin{aligned} \mathbb{P}&\left\{ \sup_{f \in T_{\beta_n}\mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^{n} |f(X_i, T_i) - Y_i|^2 - \mathbb{E}\big\{|f(X, T) - Y|^2\big\} \right| > \varepsilon \right\} \\ &= \mathbb{P}\left\{ \sup_{h \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^{n} h(Z_i) - \mathbb{E}\big\{h(Z)\big\} \right| > \varepsilon \right\} \\ &\leq 8\mathbb{E}\mathcal{N}_1\Big(\frac{\varepsilon}{8}, \mathcal{H}_n, Z_1^n\Big) \exp\left(- \frac{n\varepsilon^2}{128\big(4\beta_n^2\big)^2} \right). \end{aligned} \qquad (3.11)$$

Now we proceed to bound the covering number in (3.11). Given h_1 and h_2 in $\mathcal{H}_n$, corresponding respectively to f_1 and f_2 in $T_{\beta_n}\mathcal{F}_n$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}|h_1(Z_i)-h_2(Z_i)| &= \frac{1}{n}\sum_{i=1}^{n}\Big||f_1(X_i,T_i)-Y_i|^2-|f_2(X_i,T_i)-Y_i|^2\Big| \quad \text{a.s.}\\ &\le 4\beta_n\frac{1}{n}\sum_{i=1}^{n}|f_1(X_i,T_i)-f_2(X_i,T_i)|, \qquad (3.12)\end{aligned}$$

since $|f(x,t)|\le\beta_n$ for every f in $T_{\beta_n}\mathcal{F}_n$, and since $|Y|\le C_Y\le\beta_n$ a.s. for n large enough. Therefore, if $f_1,\dots,f_N$ is a L_1 ε-cover of $T_{\beta_n}\mathcal{F}_n$ on $(X_1,T_1),\dots,(X_n,T_n)$ with

$$N=\mathcal{N}_1\big(\varepsilon,T_{\beta_n}\mathcal{F}_n,(X,T)_1^n\big),$$

there exists a L_1 2ε-cover $f_1',\dots,f_N'$ of $T_{\beta_n}\mathcal{F}_n$ on $(X,T)_1^n$ with $f_j'\in T_{\beta_n}\mathcal{F}_n$, for all $j=1,\dots,n$. Then using the inequality (3.12) we conclude that

$$\mathcal{N}_1\Big(\frac{\varepsilon}{8},\mathcal{H}_n,Z_1^n\Big)\le\mathcal{N}_1\left(\frac{\varepsilon}{64\beta_n},T_{\beta_n}\mathcal{F}_n,(X,T)_1^n\right). \qquad (3.13)$$

Using Lemma 3.1 together with the fact that $\mathcal{N}_1\big(\varepsilon,T_{\beta_n}\mathcal{F}_n,(X,T)_1^n\big)\le\mathcal{N}_1\big(\varepsilon,\mathcal{F}_n,(X,T)_1^n\big)$, it follows that

$$\begin{aligned}&\mathbb{P}\Bigg\{\sup_{f\in T_{\beta_n}\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}|f(X_i,T_i)-Y_i|^2-\mathbb{E}\{|f(X,T)-Y|^2\}\right|>\varepsilon\Bigg\}\\ &\le 24\left(\frac{6912e\beta_n\rho_n(K_n+1)}{\varepsilon}\right)^{2(L_n^c+1)+2K_n(L_n^a d+L_n^b+L_n^c+2)}\\ &\quad\times\exp\left(-\frac{n\varepsilon^2}{2048\beta_n^4}\right)\\ &=24\exp\Bigg\{-n^{\delta}\frac{n^{1-\delta}}{\beta_n^4}\Bigg[\frac{\varepsilon^2}{2048}-\frac{2\beta_n^4\big[L_n^c+1+K_n(L_n^a d+L_n^b+L_n^c+2)\big]}{n}\\ &\qquad\times\log\left(\frac{6912\beta_n\rho_n(K_n+1)}{\varepsilon}\right)\Bigg]\Bigg\}, \qquad (3.14)\end{aligned}$$

for some $\delta>0$.

Consequently, if

$$K_n \to \infty, \quad L_n^a \to \infty, \quad L_n^b \to \infty, \quad L_n^c \to \infty, \quad \beta_n \to \infty, \quad \rho_n \to \infty,$$

as $n \to \infty$ in such a way that

$$\frac{K_n \beta_n^4 (L_n^a + L_n^b + L_n^c) \log(\beta_n \rho_n K_n)}{n} \to 0 \quad \text{and} \quad \frac{\beta_n^4}{n^{1-\delta}} \to 0,$$

for some $\delta > 0$ as $n \to \infty$, we deduce from (3.14) that

$$\sum_{n \geq 1} \mathbb{P}\left\{ \sup_{f \in T_{\beta_n} \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^{n} |f(X_i, T_i) - Y_i|^2 - \mathbb{E}\{|f(X,T) - Y|^2\} \right| > \varepsilon \right\} < \infty. \tag{3.15}$$

Part (i) of Theorem 2.1 then follows from the Borel-Cantelli Lemma.

To prove part (ii), it suffices to set $L_n^a = 1$ and $L_n^b = 1$. Then one obtains (3.15) from (3.14) if

$$K_n \to \infty, \quad L_n^c \to \infty, \quad \beta_n \to \infty, \quad \text{and} \quad \rho_n \to \infty,$$

as $n \to \infty$ in such a way that

$$\frac{K_n \beta_n^4 L_n^c \log(\beta_n \rho_n K_n)}{n} \to 0 \quad \text{and} \quad \frac{\beta_n^4}{n^{1-\delta}} \to 0,$$

for some $\delta > 0$ as $n \to \infty$, which together with the Borel-Cantelli Lemma concludes the proof of the second part of Theorem 2.1. □

Acknowledgment

This work was supported by the National Aeronautics and Space Administration under Grant No. NNX08AF65A. The authors are indebted to an anonymous referee for a careful reading of the paper and insightful comments.

References

[1] Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, **Vol. 39**, pp. 930-945.

[2] Burger, M. and Neubauer, A. (2001). Error bounds for approximation with neural networks. *Journal of Approximation Theory*, **Vol. 112**, pp. 235-250.

[3] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function *Mathematics of Control, Signals, and Systems*, **Vol. 2**, pp. 303-314.

[4] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficients models. *The Annals of Statistics*, **Vol. 27**, pp. 1491-1518.

[5] Fan, J. and Zhang, J.-T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Series B*, **Vol. 62**, pp.303-322.

[6] Fan, J., Yao, Q., and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society, Series B*, **Vol. 65**, pp 57-80.

[7] Frouin, R. and Pelletier, B. (2007). Fields of nonlinear regression models for atmospheric correction of satellite ocean-color imagery. *Remote Sensing of Environment*, **Vol. 111**, pp. 450-465.

[8] Gyorfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New-York.

[9] Hastie, T. and Tibshirani, R. (1993). Varying coefficient models, *Journal of the Royal Statistical Society, Series B*, **Vol. 55**, pp. 757-796.

[10] Hoffmann, M. and Lepski, O. (2002). Random rates in anisotropic regression. *The Annals of Statistics*, **Vol. 30**, pp. 325-358.

[11] Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **Vol. 2**, pp. 359-366.

[12] Kohler, M. and Krzyzak, A. (2005). Adaptive regression estimation with multilayer feedforward neural networks. *Nonparametric Statistics*, **Vol. 17**, pp. 891-913.

[13] Lin, V.Y. and Pinkus, A. (1993). Fundamentality of ridge functions. *Journal of Approximation Theory*, **Vol. 75**, pp. 295-311.

[14] Maiorov, V. (1999). On best approximation by ridge functions. *Journal of Approximation Theory*, **Vol. 99**, pp. 68-94.

[15] Pelletier, B. (2004). Approximation by ridge function fields over compact sets. *Journal of Approximation theory*, **Vol. 129**, pp. 230-239.

[16] Pelletier, B. and Frouin, R. (2006). Remote sensing of phytoplankton chlorophyll-a concentration by ridge function fields. *Applied Optics*, **Vol. 45**, pp784-798.

[17] Pelletier, B. and Frouin, R. (2004). Fields of nonlinear regression models for inversion of satellite data. *Geophysical Research Letters*, **Vol. 31**, L16304, doi 10.1029/2004GL019840.

[18] Pollard, D. (1984). *Convergence of Stochastic Processes*, Springer-Verlag, New-york.

[19] Vapnik, V.N. and Chervonenkis, A.Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, **Vol. 16**, pp. 264-280.

[20] Wong, H., Ip, W., and Zhang, R. (2008). Varying-coefficient single-index model. *Computational Statistics & Data Analysis*, **Vol. 52**, pp. 1458-1476.