



Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: [www.elsevier.com/locate/jspi](http://www.elsevier.com/locate/jspi)

## Convergence rates for smoothing spline estimators in varying coefficient models

P.P.B. Eggermont<sup>a,\*</sup>, R.L. Eubank<sup>b,1</sup>, V.N. LaRiccia<sup>a</sup>

<sup>a</sup>Food and Resource Economics, University of Delaware, Newark, DE 19717-1303, USA

<sup>b</sup>Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804, USA

### ARTICLE INFO

#### Article history:

Received 15 November 2006

Received in revised form

18 December 2007

Accepted 17 June 2009

#### MSC:

62G08

62G20

#### Keywords:

Time-varying coefficient model

spline smoothing

Convergence rates

Reproducing kernel Hilbert space

### ABSTRACT

We consider the estimation of a multiple regression model in which the coefficients change slowly in “time”, with “time” being an additional covariate. Under reasonable smoothness conditions, we prove the usual expected mean square error bounds for the smoothing spline estimators of the coefficient functions.

© 2009 Published by Elsevier B.V.

## 1. Introduction

Since their introduction by [Hastie and Tibshirani \(1993\)](#), varying coefficient models have become an increasingly popular option for dimension reduction in nonparametric regression with multiple predictors. An important special case of the general varying coefficient formulation is the time-varying coefficient model which utilizes only one effect-modifying covariate (“time”). These later models have applications in various contexts such as functional regression analysis, see, e.g., [Hoover et al. \(1998\)](#) or [Eubank et al. \(2004\)](#), and the analysis of longitudinal data, e.g., [Wu and Chiang \(2000\)](#). In this paper, we study the efficacy of smoothing spline estimation in time-varying coefficient models. To be precise, we consider the data

$$(Y_{in}, X_{in}, t_{in}), \quad i = 1, 2, \dots, n, \quad (1.1)$$

where the  $Y_{in}$  are the responses, the  $X_{in} \in \mathbb{R}^{1 \times p}$  are the predictors (here,  $p$  is some fixed positive integer), and the time points  $t_{in}$  are the additional covariates. It is assumed that locally (for nearby  $t_{in}$ ), the usual linear model

$$Y_{in} = X_{in}\beta + \varepsilon_{in}, \quad i = 1, 2, \dots, n, \quad (1.2)$$

\* Corresponding author.

E-mail address: [eggermon@udel.edu](mailto:eggermon@udel.edu) (P.P.B. Eggermont).

<sup>1</sup> Research supported by NSF.

provides a good fit for some fixed  $\beta \in \mathbb{R}^p$ , but not globally in that  $\beta$  varies with time. Then, it is sensible to let the predictors change with time as well, so that the new and improved model is

$$Y(t_{in}) = X(t_{in})\beta(t_{in}) + \varepsilon(t_{in}), \quad i = 1, 2, \dots, n, \tag{1.3}$$

where  $\beta(t)$  is a smooth vector-valued function of time,  $X(t)$  is a suitable stochastic process modeling a random design, and  $\varepsilon(t)$  is white noise independent of the  $X(t)$ -process with (unknown) variance  $\sigma^2$ .

Thus, we have a family of linear models (1.3), with one observation per model. If the model is changing smoothly, then we can pretend that observations for nearby models are observations for the “current” model and we can perform the usual multiple regression estimation. Note that the preceding implies that the  $X(t_{in})$  should vary *anything but* smoothly with time. As a matter of fact, we want the nearby  $X(t_{in})$  to be as far apart as possible. The technical assumption is that  $X(t_{1,n}), X(t_{2,n}), \dots, X(t_{n,n})$  are independent. (If there are replicate time points, there is a slight notational hitch we shall ignore.)

In this paper, we prove convergence rates for the smoothing spline estimator of  $\beta(t)$  in the model (1.3). This estimator was proposed (but not further studied) by Hoover et al. (1998), and is defined as the solution to

$$\text{minimize } \frac{1}{n} \sum_{i=1}^n |X(t_{in})b(t_{in}) - Y(t_{in})|^2 + \sum_{j=1}^p h_j^{2m} \|b_j^{(m)}\|^2, \tag{1.4}$$

where  $b(t) = (b_1(t), b_2(t), \dots, b_p(t))^T$  and the minimization is over all smooth functions  $b_j(t), j = 1, 2, \dots, p$ . Here,  $\|b_j^{(m)}\|$  denotes the  $L^2$ -norm of the  $m$ -th derivative of  $b_j$ . The existence and uniqueness of the estimator follows, e.g., from Eubank et al. (2004), provided that the matrix  $H \in \mathbb{R}^{n \times mp}$  with  $i$ -th row defined as the Kronecker product

$$X(t_{in}) \otimes (1, t_{in}, t_{in}^2, \dots, t_{in}^{m-1}), \tag{1.5}$$

has full column rank. As an aside, the same authors show that the resulting estimator can be computed efficiently using the Kalman filter (with associated Bayesian confidence intervals) in  $\mathcal{O}(n)$  operations. Also, note that the usual spline smoothing problem is subsumed in (1.4) by taking  $p = 1$  and  $X(t) = 1$  for all  $t$ . Taking some of the  $b_j(t)$  constant over time would cover partially linear models, see, e.g., Green et al. (1985). Finally, note that in (1.4), each coefficient of  $b(t)$  has its own smoothing parameter. It is a question of practical importance whether these  $h_j$  can be chosen (near) optimally by data-driven methods, but we shall not address this issue here.

A number of authors have studied the large sample properties of kernel estimators for  $\beta(t)$  in varying coefficient models. Convergence rates for Nadaraya–Watson kernel estimators in longitudinal versions of (1.2) with time varying covariates have been derived in Wu et al. (1998, 2000), Hoover et al. (1998), and Wu and Chiang (2000). Similar results for local polynomial based estimators are provided in Fan and Zhang (1999, 2000), and Cai et al. (2000).

Smoothing spline estimators are easily adapted to the analysis of longitudinal data with correlated errors. Here, one typically has observations like (1.1)–(1.3) on many individuals. For the scalar case, see, e.g., Wang (1998). Thus, the model (1.3) holds conditionally on  $\beta(t)$ , a realization of a stochastic process with mean denoted by  $\beta^*(t)$  and covariance (matrix) operator

$$V(t, \tau) = \mathbb{E}[(\beta(t) - \beta^*(t))(\beta(\tau) - \beta^*(\tau))^T]. \tag{1.6}$$

Thus, if we have a random sample of  $L$  individuals, one observes

$$(Y_\ell(t_{\ell i}), X_\ell(t_{\ell i}), t_{\ell i}), \quad i = 1, 2, \dots, n_\ell, \tag{1.7}$$

for  $\ell = 1, 2, \dots, L$ , according to the model(s)

$$Y_\ell(t_{\ell i}) = X_\ell(t_{\ell i})\beta_{[\ell]}(t_{\ell i}) + \varepsilon_\ell(t_{\ell i}), \quad i = 1, 2, \dots, n_\ell, \tag{1.8}$$

where  $X_\ell(t_{\ell i}) \in \mathbb{R}^{1 \times p}$ ,  $\beta_{[\ell]} \in \mathbb{R}^p$ , and conditional on the  $\beta_{[\ell]}$ , one has that  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_L$  are independent realizations of a white noise process with  $\mathbb{E}[|\varepsilon_\ell(t)|^2] = \sigma^2$  say. Note that the number of observations and the observation times need not be the same for each individual. Also note that in the present paper, we are concerned with estimating the individual  $\beta_{[\ell]}$ . If one wishes to estimate the population mean function  $\beta^*$ , the model (1.8) must be replaced by

$$Y_\ell(t_{\ell i}) = X_\ell(t_{\ell i})\beta^*(t_{\ell i}) + \delta_\ell(t_{\ell i}), \tag{1.9}$$

for  $i = 1, 2, \dots, n_\ell$  and  $\ell = 1, 2, \dots, L$ , where  $\delta_1, \delta_2, \dots, \delta_L$  are independent, identically distributed mean 0 processes with covariance operator conditional on the design,

$$\begin{aligned} W(t, \tau) &= \mathbb{E}[\delta(t)\delta(\tau)|X(t), X(\tau)] \\ &= X(t)V(t, \tau)X(\tau)^T + \sigma^2 I. \end{aligned} \tag{1.10}$$

The smoothing spline approach is then readily adapted to estimating  $\beta^*$ , viz., as the minimizer  $b = \hat{\beta}^*$  of

$$\frac{1}{L} \sum_{\ell=1}^L (Y_\ell - \mathbb{X}_\ell S_\ell b)^T W_\ell^{-1} (Y_\ell - \mathbb{X}_\ell S_\ell b) + \sum_{j=1}^p h_j^{2m} \|b_j^{(m)}\|^2, \tag{1.11}$$

where, for  $\ell = 1, 2, \dots, L$ , the vector  $Y_\ell \in \mathbb{R}^{n_\ell}$  is defined as

$$Y_\ell = (Y_\ell(t_{\ell,1}), Y_\ell(t_{\ell,2}), \dots, Y_\ell(t_{\ell,n_\ell}))^T, \tag{1.12}$$

the matrix  $\mathbb{X}_\ell \in \mathbb{R}^{n_\ell \times pn_\ell}$  is block-diagonal with diagonal blocks  $X_\ell(t_{\ell,i}) \in \mathbb{R}^{1 \times p}$

$$\mathbb{X}_\ell = \text{block-diag}(X_\ell(t_{\ell,1}), X_\ell(t_{\ell,2}), \dots, X_\ell(t_{\ell,n_\ell})), \tag{1.13}$$

and the operator  $S_\ell : W^{m,2}(0, 1) \rightarrow \mathbb{R}^{pn_\ell}$  samples the vector-valued function  $b(t)$  at the appropriate points,

$$S_\ell b = (b(t_{\ell,1})^T, b(t_{\ell,2})^T, \dots, b(t_{\ell,n_\ell})^T)^T. \tag{1.14}$$

Finally,  $W_\ell = \mathbb{X}_\ell V_\ell \mathbb{X}_\ell^T + \sigma^2 I$  with  $V_\ell$  a block matrix, with the  $(i, p)$ -th block given by

$$[V_\ell]_{ip} = V(t_{\ell,i}, t_{\ell,p}), \quad i, p = 1, 2, \dots, n_\ell. \tag{1.15}$$

In general,  $V$  must be estimated first, e.g., by

$$\widehat{V}(t, \tau) = \frac{1}{L} \sum_{\ell=1}^L (\widehat{\beta}_{[\ell]}(t) - \bar{\beta}(t))(\widehat{\beta}_{[\ell]}(\tau) - \bar{\beta}(\tau))^T, \tag{1.16}$$

where  $\widehat{\beta}_{[\ell]}$  is the smoothing spline estimator of  $\beta_{[\ell]}$ , and

$$\bar{\beta}(t) = \frac{1}{L} \sum_{\ell=1}^L \widehat{\beta}_{[\ell]}(t).$$

The methods and results of this paper are readily extended to provide convergence rates for  $\widehat{\beta}^*(t)$ , although the effect of estimating the covariance structure will cause some difficulties.

Other estimators that have been used for the analysis of longitudinal studies are kernel and local polynomial estimators, but they may not be as convenient. Since these estimators are local, this precludes them from taking the covariance structure into account. Thus, [Fan and Zhang \(1999, 2000\)](#) and [Hoover et al. \(1998\)](#) “collect” the data for each observation time  $t_i$  into a multiple linear regression model (with  $t_i = t_{i_\ell}$  for appropriate indices  $i_\ell$ )

$$Y(t_i) = \mathbb{X}(t_i)\beta^*(t_i) + \delta(t_i), \quad i = 1, 2, \dots, n, \tag{1.17}$$

where

$$\begin{aligned} \mathbb{X}(t_i) &= (X_1(t_i)^T, X_2(t_i)^T, \dots, X_L(t_i)^T)^T, \\ Y(t_i) &= (Y_1(t_i), Y_2(t_i), \dots, Y_L(t_i))^T, \end{aligned} \tag{1.18}$$

and similarly for  $\delta(t_i)$ . Then,  $\delta(t_i) \sim \text{Normal}(0, \sigma_i^2 I_{L \times L})$  for a suitable  $\sigma_i^2$ . One may then compute the (least squares) estimator of  $\beta^*$  in the usual way, and compute the covariance  $\text{Cov}[\widehat{\beta}^*(t_i), \widehat{\beta}^*(t_p)]$  in terms of the covariance of  $\delta$  (which may be estimated). Since in (1.17), consecutive observations times are decoupled, it is then natural to smooth the  $\widehat{\beta}^*(t_i)$ . However, in these local methods there is no choice but to ignore the covariance structure of the  $\widehat{\beta}^*(t_i)$ ,  $i = 1, 2, \dots, n$ , even if one performs smoothing for each coefficient  $\beta_j$  separately. See also [Chiang et al. \(2001\)](#).

The efficacy of smoothing spline estimators in varying coefficient problems has been demonstrated by [Eubank et al. \(2004\)](#). The purpose of this paper is to provide a new approach to determining rates of convergence for smoothing spline estimators for nonparametric regression in general, and varying coefficient problems in particular.

The paper is laid out as follows. In the next section, we formulate the assumptions and state the theorem on the convergence rates. The remaining sections develop the proof of this theorem. Most notably, in Section 3, we introduce the Sobolev spaces  $W^{m,2}(0, 1)$  with suitable inner products depending on the (scalar) smoothing parameter  $\lambda$  and the associated reproducing kernels. These reproducing kernels play a crucial role in the analysis. In Section 4, we prove the main theorem on convergence rates, formulating various lemmas which are then proved in later sections.

## 2. The problem, assumptions and main result

In this section, we give a precise description of the estimation problem and state the main result on the convergence rates of the estimator.

We consider the data

$$(Y(t_{in}), X(t_{in}), t_{in}), \quad i = 1, 2, \dots, n, \tag{2.1}$$

following the model:

$$Y(t) = X(t)\beta(t) + \varepsilon(t), \quad 0 \leq t \leq 1. \tag{2.2}$$

Here,  $\beta(t)$  is a deterministic vector-valued function with values in  $\mathbb{R}^p$  for some fixed integer  $p$ , and  $X(t)$  is a random function with values in  $\mathbb{R}^{1 \times p}$ . The noise process  $\varepsilon(t)$  is independent of  $X(t)$  and satisfies

$$\mathbb{E}[\varepsilon(t)] = 0, \quad \mathbb{E}[|\varepsilon(t)|^2] = \sigma^2, \quad \mathbb{E}[\varepsilon(t)\varepsilon(s)] = 0 \text{ for } t \neq s. \tag{2.3}$$

The “design” process  $X(t)$  is assumed to be independent and bounded, i.e.,

$$X(t_{1,n}), X(t_{2,n}), \dots, X(t_{n,n}) \text{ are mutually independent,} \tag{2.4}$$

and there exists a constant  $C$  such that for all  $t$ ,

$$\|X(t)\|_{\mathbb{R}^p} \leq C \text{ almost surely.} \tag{2.5}$$

Moreover, it is required to be “full” in that  $\mathbb{E}[X(t)^T X(t)]$  should be positive definite, uniformly in  $t$ , i.e., for some positive constant  $\rho$ ,

$$\mathbb{E}[X(t)^T X(t)] - \rho I \text{ is semi-positive-definite for all } t. \tag{2.6}$$

Finally, the time points  $t_{1,n}, t_{2,n}, \dots, t_{n,n}$  are assumed to be deterministic and super-quasi-uniformly distributed on  $[0, 1]$ . (More or less equally spaced time points will do. For the precise statement, see Definition 1 in Section 4.) The smoothness requirement on  $\beta$  is interpreted as

$$\|\beta_j\|^2 + \|\beta_j^{(m)}\|^2 \leq C_1, \quad j = 1, 2, \dots, p, \tag{2.7}$$

for a (known) integer  $m \geq 1$  and an (unknown) constant  $C_1$ . Here,  $\|\cdot\|$  and  $\langle \cdot, \cdot \rangle$  denote the  $L^2$  norm and inner product,

$$\|f\|^2 = \langle f, f \rangle \text{ where } \langle f, g \rangle = \int_0^1 f(t)g(t) dt. \tag{2.8}$$

Thus, the function space of interest is, for  $m \geq 1$ ,

$$W^{m,2}(0, 1) = \left\{ f \in C[0, 1] : \begin{array}{l} f^{(m-1)} \text{ absolutely continuous} \\ f^{(m)} \in L^2(0, 1) \end{array} \right\}. \tag{2.9}$$

We may then concisely state the smoothness assumption (2.7) on  $\beta$  as

$$\beta \in (W^{m,2}(0, 1))^p. \tag{2.10}$$

The estimator of  $\beta$  under consideration here is the solution  $b^{nh}$  of the spline smoothing problem

$$\text{minimize } S_n(b) + J_h(b) \text{ subject to } b \in (W^{m,2}(0, 1))^p, \tag{2.11}$$

where

$$S_n(b) = \frac{1}{n} \sum_{i=1}^n |X(t_{in})b(t_{in}) - Y(t_{in})|^2, \tag{2.12}$$

$$J_h(b) = \sum_{j=1}^p h_j^{2m} \|b_j^{(m)}\|^2. \tag{2.13}$$

Here, the  $h_j$  are the smoothing parameters, and as is typical, have to be chosen appropriately. The notation  $b^{nh}$  is succinct, but hides many details, such as the fact that  $b^{nh} \in (W^{m,2}(0, 1))^p$ , and  $h = (h_1, h_2, \dots, h_p)$ .

The problem (2.11) always has solutions, while the condition (1.5) guarantees uniqueness. Also, the usual considerations show that the coefficients of the solution(s) are natural splines of degree  $2m - 1$ . For more on spline smoothing, see, e.g., Wahba (1990), Eubank (1999), or Eggermont and LaRiccia (2009).

We have the following result on the mean squared error of  $b^{nh}$ .

**Theorem 1** (Convergence rates). (a) Under the model (2.1)–(2.7) and super-quasi-uniformly distributed, deterministic time points, the solution  $b^{nh}$  of (2.11) satisfies

$$\mathbb{E} \left[ \sum_{j=1}^p \|b_j^{nh} - \beta_j\|^2 \right] = \mathcal{O} \left( \sum_{j=1}^m h_j^{2m} + (nh_j)^{-1} \right),$$

for deterministic  $h_j$  with  $h_j \rightarrow 0$  and  $nh_j^2/\log n \rightarrow \infty$ .

(b) If in addition,  $\beta \in (W^{2m,2}(0, 1))^p$  and the coefficients of  $\beta$  satisfy the natural boundary conditions

$$\beta_j^{(k)}(0) = \beta_j^{(k)}(1) = 0, \quad k = m, m + 1, \dots, 2m - 1,$$

for all  $j$ , then the bias terms  $h_j^{2m}$  in the solution  $b^{nh}$  of (2.11) improve to  $h_j^{4m}$ ,

$$\mathbb{E} \left[ \sum_{j=1}^p \|b_j^{nh} - \beta_j\|^2 \right] = \mathcal{O} \left( \sum_{j=1}^m h_j^{4m} + (nh_j)^{-1} \right).$$

It follows that if  $h_j \asymp n^{-1/(2m+1)}$  for all  $j$ , then we get the usual rate  $n^{-2m/(2m+1)}$  for the mean integrated squared error, and the rate  $n^{-4m/(4m+1)}$  for  $h_j \asymp n^{-1/(4m+1)}$  when the natural boundary conditions apply.

### 3. The setting

In this section, we lay the groundwork for the proof of the main theorem. The starting point is the observation that the coefficient functions  $\beta_j(t)$  live in a reproducing kernel Hilbert space (not a surprise in the spline world) with inner products depending on a smoothing parameter (a new twist). The associated reproducing kernels turn out to be the right tool for studying the random sums that pop-up in various places. This is the approach for ordinary spline smoothing taken in Eggermont and LaRiccia (2006, 2009), that carries over nicely to the present problem.

To motivate what follows, it is instructive to consider the plain nonparametric regression problem, i.e., the problem (2.2) with  $p = 1$  and  $X(t) = 1$  for all  $t$ . Then the smoothing spline problem (2.11) takes the form, with  $b_1$  replaced by  $f$  and the smoothing parameter  $h_1$  by  $\lambda$ ,

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n} \sum_{i=1}^n |f(t_{in}) - Y(t_{in})|^2 + \lambda^{2m} \|f^{(m)}\|^2 \\ &\text{such that} \quad f \in W^{m,2}(0, 1). \end{aligned} \tag{3.1}$$

The solution  $f^{n\lambda}$  is an estimator of  $f_0(t) = \beta_1(t)$ , the true mean function. In Section 2, we already made the case for  $W^{m,2}(0, 1)$  being the appropriate space of functions. The smoothness conditions on  $f$  may then be expressed as

$$\|f\|_{W^{m,2}(0,1)}^2 < \infty,$$

where

$$\|f\|_{W^{m,2}(0,1)} = \{\|f\|^2 + \|f^{(m)}\|^2\}^{1/2}. \tag{3.2}$$

Inspection of the problem (3.1) suggests that perhaps, the term  $\|f^{(m)}\|^2$  should be weighted by  $\lambda^{2m}$ . This becomes even more clear when proving convergence rates, as we now outline. The following approach to penalized least-squares problems goes back at least as far as Ribière (1967). See also van de Geer (2000).

Denote the objective function in (3.1) by  $L_n(f)$ . Now, expand  $L_n(f)$  around its minimizer  $f^{n\lambda}$ , and take  $f = f_0$ . Then, with  $v = f_0 - f^{n\lambda}$ ,

$$L_n(f_0) = L_n(f^{n\lambda}) + \partial L_n(f^{n\lambda}, v) + \frac{1}{n} \sum_{i=1}^n |v(t_{in})|^2 + \lambda^{2m} \|v^{(m)}\|^2, \tag{3.3}$$

where  $\partial L_n(f^{n\lambda}, v)$  is the directional derivative of  $L_n$  at  $f^{n\lambda}$  in the direction  $v$ ,

$$\partial L_n(f^{n\lambda}, v) = \frac{2}{n} \sum_{i=1}^n (f^{n\lambda}(t_{in}) - Y(t_{in}))v(t_{in}) + 2\lambda^{2m} \langle f_0^{(m)}, v^{(m)} \rangle. \tag{3.4}$$

Since  $f^{n\lambda}$  is the minimizer of  $L_n$ , the directional derivative vanishes in every direction. In particular,  $\partial L_n(f^{n\lambda}, v) = 0$  and (3.3) takes the form

$$\frac{1}{n} \sum_{i=1}^n |v(t_{in})|^2 + \lambda^{2m} \|v^{(m)}\|^2 = L_n(f_0) - L_n(f^{n\lambda}). \tag{3.5}$$

Now, in the right hand side, expand  $L_n(f^{n\lambda})$  around  $f_0$ . This gives

$$L_n(f^{n\lambda}) = L_n(f_0) + \frac{1}{n} \sum_{i=1}^n |v(t_{in})|^2 + \lambda^{2m} \|v^{(m)}\|^2 - \frac{2}{n} \sum_{i=1}^n (f_0(t_{in}) - Y(t_{in}))v(t_{in}) - 2\lambda^{2m} \langle f_0^{(m)}, v^{(m)} \rangle. \tag{3.6}$$

Substituting this into the right hand side of (3.5) and moving terms around gives the fundamental equality for the error  $v=f_0-f^{n\lambda}$ ,

$$\frac{1}{n} \sum_{i=1}^n |v(t_{in})|^2 + \lambda^{2m} \|v^{(m)}\|^2 = \frac{1}{n} \sum_{i=1}^n \varepsilon(t_{in})v(t_{in}) + \lambda^{2m} \langle f_0^{(m)}, v^{(m)} \rangle, \tag{3.7}$$

where  $\varepsilon(t_{in})=f_0(t_{in})-Y(t_{in})$  is just the negative of the noise in the observations.

Formula (3.7) has important consequences. With the Cauchy–Schwarz inequality,

$$\lambda^{2m} \langle f_0^{(m)}, v^{(m)} \rangle \leq \lambda^{2m} \|f_0^{(m)}\| \|v^{(m)}\|, \tag{3.8}$$

which gives rise to the bias in the estimator.

The more interesting part is the sum on the right hand side of the inequality (3.7). Note that this sum is/should be small for two reasons. First, the errors  $v(t_{in})=f_0(t_{in})-f^{n\lambda}(t_{in})$  should be small (although this is begging the question). Second, it is a (weighted) sum of the random noise  $\varepsilon(t_{in})$ , although the randomness of the weights  $v(t_{in})$  is cause of concern. Anyway, the objective is to bound this sum in terms of (the square root) of the expression on the left of (3.7). This would be a convenient norm on  $v \in W^{m,2}(0,1)$ , except that the sums are unwieldy. However, viewing the sum as a Riemann sum, and viewing this as an approximation of an integral, all of this suggest the norms  $\|\cdot\|_{m,\lambda}$  and associated inner products  $\langle \cdot, \cdot \rangle_{m,\lambda}$ ,

$$\begin{aligned} \|f\|_{m,\lambda} &= \{\|f\|^2 + \lambda^{2m} \|f^{(m)}\|^2\}^{1/2}, \\ \langle f, g \rangle_{m,\lambda} &= \langle f, g \rangle + \lambda^{2m} \langle f^{(m)}, g^{(m)} \rangle. \end{aligned} \tag{3.9}$$

The goal is then to obtain an inequality of the form

$$\frac{1}{n} \sum_{i=1}^n \varepsilon(t_{in})v(t_{in}) \leq \eta \|v\|_{m,\lambda}, \tag{3.10}$$

with  $\eta \rightarrow 0$  in an appropriate sense. Then, ignoring the difference between the sum and the integral for now, substituting the bounds (3.8) and (3.10) into the fundamental equality (3.7) results in the bound

$$\|v\|_{m,\lambda} \leq \eta + \lambda^m \|f_0\|. \tag{3.11}$$

Thus, we need to get a handle on the inequality (3.10). Here is where reproducing kernel Hilbert spaces and their reproducing kernels enter the scene.

It is well-known that  $W^{m,2}(0,1)$  is a reproducing kernel Hilbert space but that the reproducing kernel depends on the choice of the inner product. For the inner product (3.9), we denote the reproducing kernel as  $\mathfrak{R}_{m\lambda}(\cdot, \cdot)$ . Thus,  $\mathfrak{R}_{m\lambda}(t, \cdot) \in W^{m,2}(0,1)$  for all  $t$  and satisfies

$$f(t) = \langle \mathfrak{R}_{m\lambda}(t, \cdot), f \rangle_{m,\lambda} \quad \text{for all } t \in [0,1], f \in W^{m,2}(0,1). \tag{3.12}$$

It is not too hard to show that uniformly in  $t$ ,

$$\|\mathfrak{R}_{m\lambda}(t, \cdot)\|_{m,\lambda} = \mathcal{O}(\lambda^{-1/2}), \quad \lambda \rightarrow 0, \tag{3.13}$$

e.g., by proving that  $W^{m,2}(0,1)$  is a reproducing kernel Hilbert space in the form of the inequalities

$$|f(t)| \leq c\lambda^{-1/2} \|f\|_{1,\lambda} \leq c_m \lambda^{-1/2} \|f\|_{m,\lambda}, \tag{3.14}$$

for suitable constants  $c$  and  $c_m$ , independent of  $t$ . See, e.g., Adams and Fournier (2003) or Eggermont and LaRiccia (2009). The reproducing kernels  $\mathfrak{R}_{m\lambda}$  are precisely what is needed for the inequality (3.10).

**Theorem 2** (Bounds on random sums). Let  $\theta_n = (\theta_{1,n}, \theta_{2,n}, \dots, \theta_{n,n})^T$  be a vector of random variables satisfying

$$\mathbb{E}[\theta_n] = 0, \quad \mathbb{E}[\theta_n \theta_n^T] = \sigma_\theta^2 I_{n \times n},$$

with  $\sigma_\theta < \infty$ . Then, there exists a constant  $c$  such that for all (random)  $f \in W^{m,2}(0,1)$ , and all deterministic  $\lambda, 0 < \lambda < 1$ ,

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \theta_{in} f(t_{in}) \right] \leq c(n\lambda)^{-1/2} (\mathbb{E}[\|f\|_{m,\lambda}^2])^{1/2}.$$

**Proof.** With (3.12), we have for all  $f \in W^{m,2}(0,1)$ , random or not,

$$\frac{1}{n} \sum_{i=1}^n \theta_{in} f(t_{in}) = \left\langle \frac{1}{n} \sum_{i=1}^n \theta_{in} \mathfrak{R}_{m\lambda}(t_{in}, \cdot), f \right\rangle_{m,\lambda} \leq \left\| \frac{1}{n} \sum_{i=1}^n \theta_{in} \mathfrak{R}_{m\lambda}(t_{in}, \cdot) \right\|_{m,\lambda} \|f\|_{m,\lambda}. \tag{3.15}$$

Now, employing (3.13), one shows that

$$\mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^n \theta_{in} \mathfrak{R}_{m\lambda}(t_{in}, \cdot) \right\|_{m,\lambda}^2 \right] = \mathcal{O}((n\lambda)^{-1}), \tag{3.16}$$

and then an application of Cauchy–Schwarz to (3.15) gives the required result.  $\square$

Applying this to the inequality (3.10) gives that

$$\eta = \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon(t_{in}) \mathfrak{R}_{m\lambda}(t_{in}, \cdot) \right\|_{m,\lambda}, \tag{3.17}$$

and so  $\mathbb{E}[\eta^2] = \mathcal{O}(n\lambda)^{-1}$ . From (3.11), one then gets the bound

$$\mathbb{E}[\|v\|_{m,\lambda}^2] = \mathcal{O}((n\lambda)^{-1} + \lambda^{2m}). \tag{3.18}$$

Minimizing the asymptotic bound over  $\lambda$  gives

$$\mathbb{E}[\|v\|_{m,\lambda}^2] = \mathcal{O}(n^{2m/(2m+1)}) \quad \text{for } \lambda \asymp n^{-1/(2m+1)}. \tag{3.19}$$

The above approach to random sums should be contrasted with the metric entropy approach. See, e.g., van de Geer (2000) and references therein.

#### 4. The proof of Theorem 1

The starting point in the proof is the following inequality. It is useful to introduce the error

$$\delta(t) = b^{nh}(t) - \beta(t), \quad 0 \leq t \leq 1. \tag{4.1}$$

Recall that  $\delta$  is vector valued :  $\delta(t) = (\delta_1(t), \delta_2(t), \dots, \delta_p(t))^T$ .

**Lemma 1.** *With  $\varepsilon(t)$  as in (2.2), the error  $\delta(t)$  satisfies*

$$\frac{2}{n} \sum_{i=1}^n |X(t_{in})\delta(t_{in})|^2 + J_h(\delta) \leq \frac{2}{n} \sum_{i=1}^n \varepsilon(t_{in})X(t_{in})\delta(t_{in}) + J_h(\beta) - J_h(b^{nh}). \tag{4.2}$$

The proof of this lemma goes along the lines of that of the equalities (3.5)–(3.7), and is omitted. The hard work is to show that the inequality implies that

$$\Delta^2 \leq c(n\mathcal{H})^{-1/2} \Delta + c\mathcal{H}^{(2m)} \quad \text{where } \Delta^2 = \mathbb{E} \left[ \sum_{j=1}^p \|\delta_j\|_{m,h_j}^2 \right]$$

and  $\mathcal{H}$  is the maximum of the  $p$  smoothing parameters  $h_1, h_2, \dots, h_p$ . This would imply part (a) of the theorem.

We first state bounds on the various parts appearing in the inequality of Lemma 1, and then prove the theorem. The bounds themselves are proven in the following sections.

**Lemma 2.** *Let  $m \geq 1$ . For the model (2.1)–(2.7), there exists a constant  $c$  such that for all random  $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_p)^T \in (W^{m,2}(0, 1))^p$ ,*

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \varepsilon(t_{in})X(t_{in})\varphi(t_{in}) \right] \leq c \sum_{j=1}^p (nh_j)^{-1/2} (\mathbb{E}[\|\varphi_j\|_{m,h_j}^2])^{1/2},$$

provided  $nh_j \rightarrow \infty, h_j \rightarrow 0$  for all  $j$ .

Scanning our progress so far, we now have the  $L^2$  integral of  $\delta$  on the right of the inequality (4.2), but a sum on the left. We need a quadrature result for “regular” distributions of the time points  $t_{in}$ .

**Definition 1.** The family of time points  $\{t_{in} : i = 1, 2, \dots, n\}$  is quasi-uniform if there exists a constant  $c$  such that for all  $f$  with integrable derivative,

$$\left| \frac{1}{n} \sum_{i=1}^n f(t_{in}) - \int_0^1 f(t) dt \right| \leq cn^{-1} \|f'\|_{L^1(0,1)}.$$

The family of time points  $\{t_{in} : i = 1, 2, \dots, n\}$  is super-quasi-uniform if there exists positive constants  $c_1$  and  $c_2$  such that for all  $f$  with integrable derivative,

$$\frac{1}{n} \sum_{i=1}^n f(t_{in}) \geq c_1 \int_0^1 f(t) dt - c_2 n^{-1} \|f'\|_{L^1(0,1)}.$$

In Section 7 we show that both of the uniform time point designs

$$t_{in} = \frac{i-1}{n-1} \quad \text{and} \quad t_{in} = \frac{i-1/2}{n}, \quad i = 1, 2, \dots, n \tag{4.3}$$

are quasi-uniform in the sense of Definition 1. A family of time points is super-quasi-uniform if it “contains” a quasi-uniform family consisting asymptotically of at least  $rn$  points for some fixed  $r > 0$ .

The above two lemmas get combined into the following.

**Lemma 3.** *Suppose the time points  $\{t_{in} : i = 1, 2, \dots, n\}$  are super-quasi-uniform. Then, there exists positive constants  $c$  and  $c_1$  such that for all random  $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_p)^T \in (W^{m,2}(0, 1))^p$  and all  $h_j$  satisfying  $nh_j^2/\log n \rightarrow \infty, h_j \rightarrow 0$ ,*

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n |X(t_{in})\varphi(t_{in})|^2 \right] \geq \sum_{j=1}^p \{c\mathbb{E}[\|\varphi_j\|^2] - c_1(nH^2/\log n)^{-1/2} \mathbb{E}[\|\varphi_j\|_{m,h_j}^2]\},$$

with  $H^{-1} = \sum_{j=1}^p h_j^{-1}$ .

**Proof of Theorem 1.** Consider the inequality (4.2). Applying Lemma 3 with  $\varphi = \delta = (\delta_1, \delta_2, \dots, \delta_p)^T$ , where  $\delta_j = b_j^{nh} - \beta_j$ , to the left hand side gives the lower bound, for a positive constant  $c$ ,

$$(c - c_1(nH^2/\log n)^{-1/2}) \sum_{j=1}^n \eta_j^2, \tag{4.4}$$

where

$$\eta_j^2 = \mathbb{E}[\|\delta_j\|_{m,h_j}^2]. \tag{4.5}$$

Of course, under the conditions of Theorem 1 on the  $h_j$ , we may ignore the  $(nH^2/\log n)^{-1/2}$  term in the above lower bound.

Then, applying Lemma 2 (with  $\nu = \delta$ ) to the right hand side of (4.2) gives

$$c \sum_{j=1}^p \eta_j^2 \leq c_1 \sum_{j=1}^p (nh_j)^{-1/2} \eta_j + J_h(\beta) - \mathbb{E}[J_h(b^{nh})]. \tag{4.6}$$

Since we may obviously drop the term  $-\mathbb{E}[J_h(b^{nh})]$  in (4.6), and since the assumption  $\beta \in (W^{m,2}(0, 1))^p$  implies that

$$J_h(\beta) \leq c \sum_{j=1}^p h_j^{2m},$$

then part (a) of Theorem 1 follows.

To prove part (b), we take a closer look at the term  $J_h(\beta) - \mathbb{E}[J_h(b^{nh})]$  on the right hand side of the inequality (4.6). Consider the identity, valid for all  $f, g \in W^{m,2}(0, 1)$ ,

$$\|f^{(m)}\|^2 - \|g^{(m)}\|^2 = 2 \langle f^{(m)}, f^{(m)} - g^{(m)} \rangle - \|f^{(m)} - g^{(m)}\|^2.$$

Obviously, we need not worry about the term  $-\|f^{(m)} - g^{(m)}\|^2$ . Now, suppose that  $f \in W^{2m,2}(0, 1)$  and that  $f$  satisfies the natural boundary conditions

$$f^{(k)}(0) = f^{(k)}(1) = 0, \quad k = m, m+1, \dots, 2m-1.$$

Then, integration by parts  $m$  times yields

$$\langle f^{(m)}, f^{(m)} - g^{(m)} \rangle = (-1)^m \langle f^{(2m)}, f - g \rangle,$$

which may be bounded by  $\|f^{(2m)}\| \|f - g\|$  (using Cauchy-Schwarz).

Applying this to each coefficient of  $J_h(\beta) - J_h(b^{nh})$  gives

$$J_h(\beta) - J_h(b^{nh}) \leq \sum_{j=1}^p 2h_j^{(2m)} \|\beta_j^{(2m)}\| \|\delta_j\|,$$

so that after taking expectations

$$J_h(\beta) - \mathbb{E}[J_h(b^{nh})] \leq \sum_{j=1}^p 2h_j^{(2m)} \|\beta_j^{(2m)}\| \eta_j,$$

with  $\eta_j$  as in (4.5). Substituting the above bound into (4.6) then gives

$$c \sum_{j=1}^p \eta_j^2 \leq c_2 \sum_{j=1}^p \{(nh_j)^{-1/2} + h_j^{2m}\} \eta_j,$$

and part (b) follows.  $\square$

This completes the proof of the theorem. In the remaining sections, we prove the lemmas.

### 5. Quadrature

In this section, we first prove that the uniform time points (4.3) are indeed quasi-uniform in the sense of Definition 1. We also prove Lemma 3.

**Lemma 4.** For the uniform time points  $t_{in} = (i - 1)/(n - 1)$ ,  $i = 1, 2, \dots, n$ , and for every  $f$  with integrable derivative,

$$\left| \frac{1}{n} \sum_{i=1}^n f(t_{in}) - \int_0^1 f(t) dt \right| \leq \frac{1}{n-1} \int_0^1 |f'(t)| dt.$$

For the time points  $t_{in} = (i - 1/2)/n$ ,  $i = 1, 2, \dots, n$ , the same inequality holds with the factor  $1/(n - 1)$  replaced by  $1/n$ .

**Proof.** We only consider the first part. The first step is the following amusing identity:

$$\frac{1}{n} \sum_{i=1}^n c_{in} = \frac{1}{n-1} \sum_{i=1}^{n-1} \{a_{in} c_{in} + b_{in} c_{i+1,n}\}, \tag{5.1}$$

for all  $c_{in}$ ,  $i = 1, 2, \dots, n$ , where  $a_{in} = (n - i)/n$ ,  $b_{in} = i/n$ . Then, with the intervals  $\omega_{in} = (t_{in}, t_{i+1,n})$ ,

$$\frac{1}{n-1} \{a_{in} f(t_{in}) + b_{in} f(t_{i+1,n})\} - \int_{\omega_{in}} f(t) dt = a_{in} \int_{\omega_{in}} \{f(t_{in}) - f(t)\} dt + b_{in} \int_{\omega_{in}} \{f(t) - f(t_{i+1,n})\} dt.$$

Now, for  $t \in \omega_{in}$ ,

$$|f(t) - f(t_{in})| = \left| \int_{t_{in}}^t f'(s) ds \right| \leq \int_{\omega_{in}} |f'(s)| ds,$$

so

$$\int_{\omega_{in}} |f(t) - f(t_{in})| dt \leq \frac{1}{n-1} \int_{\omega_{in}} |f'(t)| dt.$$

The same bound applies to  $\int_{\omega_{in}} |f(t) - f(t_{i+1,n})| dt$ . Then, adding these bounds gives

$$\left| \frac{1}{n-1} \{a_{in} f(t_{in}) + b_{in} f(t_{i+1,n})\} - \int_{\omega_{in}} f(t) dt \right| \leq \frac{1}{n-1} \int_{\omega_{in}} |f'(t)| dt,$$

so that adding them over  $i = 1, 2, \dots, n - 1$ , gives the required result.  $\square$

**Lemma 5.** Let the time points  $t_{1,n}, t_{2,n}, \dots, t_{n,n}$  be super-quasi-uniform and let  $m \geq 1$ . Then, there exists positive constants  $c$  and  $c_1$  such that for all functions  $f \in W^{m,2}(0, 1)$ , and  $n\lambda \rightarrow \infty$ ,  $\lambda \rightarrow 0$ ,

$$\frac{1}{n} \sum_{i=1}^n |f(t_{in})|^2 \geq c \int_0^1 |f(t)|^2 dt - c_1 (n\lambda)^{-1} \|f\|_{m,\lambda}^2.$$

**Proof.** By the super-quasi-uniformity of the time points, we obtain

$$\frac{1}{n} \sum_{i=1}^n |f(t_{in})|^2 \geq c \int_0^1 |f(t)|^2 dt - cn^{-1} \|f^2\|_{L^1(0,1)}.$$

Since

$$\begin{aligned} \|f^2\|_{L^1(0,1)} &= 2\|ff'\|_{L^1(0,1)} \leq 2\lambda^{-1} \|f\| \{\lambda \|f'\|\} \\ &\leq \lambda^{-1} \|f\|_{1,\lambda}^2 \leq c\lambda^{-1} \|f\|_{m,\lambda}^2, \end{aligned}$$

the last inequality by (3.14), the lemma follows.  $\square$

We are now ready for Lemma 3.

**Proof of Lemma 3.** First, write

$$|X(t_{in})\varphi(t_{in})|^2 = \varphi(t_{in})^T X(t_{in})^T X(t_{in}) \varphi(t_{in}).$$

Then,

$$\frac{1}{n} \sum_{i=1}^n |X(t_{in})\varphi(t_{in})|^2 = \frac{1}{n} \sum_{i=1}^n \varphi(t_{in})^T \mathbb{E}[X(t_{in})^T X(t_{in})] \varphi(t_{in}) + \text{rem},$$

where the remainder “rem” is given by

$$\text{rem} = \frac{1}{n} \sum_{i=1}^n \varphi(t_{in})^T \Theta(t_{in}) \varphi(t_{in}),$$

with  $\Theta(t) = X(t)^T X(t) - \mathbb{E}[X(t)^T X(t)]$ .

Now, by assumption (2.6) and Lemma 5,

$$\frac{1}{n} \sum_{i=1}^n \varphi(t_{in})^T \mathbb{E}[X(t_{in})^T X(t_{in})] \varphi(t_{in}) \geq \rho \frac{1}{n} \sum_{i=1}^n \varphi(t_{in})^T \varphi(t_{in}) \geq \sum_{j=1}^p \{c\|\varphi_j\|^2 - c_1(nh_j)^{-1} \|\varphi_j\|_{m,h_j}^2\}.$$

So, this term is better than advertised.

For the remainder “rem”, we write  $\text{rem} = \sum_{j,\ell=1}^p S_{j\ell}$ , with

$$S_{j\ell} = \frac{1}{n} \sum_{i=1}^n \varphi_j(t_{in}) \varphi_\ell(t_{in}) [\Theta(t_{in})]_{j,\ell}.$$

Note that by (2.4) and (2.5), the random variables  $[\Theta(t_{in})]_{j,\ell}$ ,  $i = 1, 2, \dots, n$ , are independent and bounded (uniformly in the time points). Now, by our reproducing kernel Hilbert spaces trick, as in Section 3, for all  $\lambda$  (actually, for  $\lambda = (h_j h_\ell)^{1/2}$ ),

$$|S_{j\ell}| \leq \left\| \frac{1}{n} \sum_{i=1}^n [\Theta(t_{in})]_{j\ell} \mathfrak{R}_{1,\lambda}(t_{in}, \cdot) \right\|_{1,\lambda} \|\varphi_j \varphi_\ell\|_{1,\lambda}.$$

The McDiarmid–Devroye exponential inequality, see the proof of Lemma 7 below, yields the almost sure bound

$$\left\| \frac{1}{n} \sum_{i=1}^n [\Theta(t_{in})]_{j\ell} \mathfrak{R}_{1,\lambda}(t_{in}, \cdot) \right\|_{1,\lambda} \stackrel{\text{a.s.}}{\leq} \mathcal{O}((n\lambda/\log n)^{-1/2}). \tag{5.2}$$

Next, below in Lemma 6, we show for  $\lambda = (h_j h_\ell)^{1/2}$  that

$$\lambda^{-1/2} \|\varphi_j \varphi_\ell\|_{1,\lambda} \leq c(h_j^{-1} + h_\ell^{-1})(\|\varphi_j\|_{m,h_j}^2 + \|\varphi_\ell\|_{m,h_\ell}^2).$$

This gives a bound on  $\mathbb{E}[|S_{j\ell}|]$ , and hence

$$\mathbb{E}[\text{rem}] \leq c(n/\log n)^{-1/2} H^{-1} \sum_{j=1}^p \|\varphi_j\|_{m,h_j}^2,$$

with  $H^{-1} = h_1^{-1} + h_2^{-1} + \dots + h_p^{-1}$ . The lemma follows.  $\square$

**Lemma 6.** *There exists a constant  $c_m$  such that for all  $f, g \in W^{1,2}(0, 1)$  and all positive  $\lambda, \kappa, \nu$ , with  $\nu = (\lambda\kappa)^{1/2}$ ,*

$$\nu^{-1/2} \|fg\|_{1,\nu} \leq c(\lambda^{-1} + \kappa^{-1})(\|f\|_{1,\lambda}^2 + \|g\|_{1,\kappa}^2).$$

**Proof.** First, we have the inequality

$$\nu^{-1/2} \|fg\|_{1,\nu} \leq (\lambda\kappa)^{-1/4} \|fg\| + (\lambda\kappa)^{1/4} \|(fg)'\|.$$

The first term on the right is easy: by the Arithmetic–Geometric Mean inequality, for all  $r > 0$ , we have  $2fg \leq rf^2 + r^{-1}g^2$  (pointwise), so that for  $r = (\kappa/\lambda)^{1/2}$ ,

$$2(\lambda\kappa)^{-1/4} \|fg\| \leq \lambda^{-1/2} \|f\|^2 + \kappa^{-1/2} \|g\|^2. \tag{5.3}$$

For the remaining term, obviously

$$\|(fg)'\| \leq \|fg'\| + \|f'g\|.$$

Now,

$$\|fg'\| \leq \|f\|_{L^\infty(0,1)} \|g'\| \leq c\lambda^{-1/2} \|f\|_{1,\lambda} \|g'\|,$$

the last inequality by (3.14). Using the Arithmetic–Geometric Mean inequality once (no, twice) more, then gives

$$\begin{aligned} \nu^{1/2} \|fg'\| &\leq c\lambda^{-1/4} \kappa^{1/4} \|f\|_{1,\lambda} \|g'\| \\ &\leq c\lambda^{-1/2} \kappa^{-1/2} \|f\|_{1,\lambda}^2 + c\kappa \|g\|^2 \\ &\leq c(\lambda^{-1} + \kappa^{-1}) \|f\|_{1,\lambda}^2 + c\kappa^{-1} \|g\|_{1,\kappa}^2. \end{aligned}$$

Adding the corresponding inequality for  $\nu^{1/2} \|f'g\|$  proves the lemma.  $\square$

The example  $\lambda = \kappa = \nu$  and  $f(x) = g(x) = \max(1 - x/\lambda, 0)$  on the interval  $(0, 1)$  shows that Lemma 6 cannot easily be improved. We leave the details to the interested reader.

## 6. Random sums

**Proof of Lemma 2.** First we write

$$\frac{1}{n} \sum_{i=1}^n \varepsilon(t_{in}) X(t_{in}) \varphi(t_{in}) = \sum_{j=1}^p \left\{ \frac{1}{n} \sum_{i=1}^n \varepsilon(t_{in}) X_j(t_{in}) \varphi_j(t_{in}) \right\}.$$

Then, as in the *proof* of Theorem 2, we write, with  $\lambda = h_j$ ,

$$\frac{1}{n} \sum_{i=1}^n \varepsilon(t_{in}) X_j(t_{in}) \varphi_j(t_{in}) = \left\langle \frac{1}{n} \sum_{i=1}^n \varepsilon(t_{in}) X_j(t_{in}) \mathfrak{R}_{m,\lambda}(t_{in}, \cdot), \varphi_j \right\rangle_{m,\lambda} \leq S \|\varphi_j\|_{m,\lambda}, \tag{6.1}$$

where

$$S = \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon(t_{in}) X_j(t_{in}) \mathfrak{R}_{m,\lambda}(t_{in}, \cdot) \right\|_{m,h}.$$

Now,  $\mathbb{E}[S^2]$  is equal to

$$n^{-2} \sum_{i,k=1}^n \mathbb{E}[\varepsilon(t_{in}) \varepsilon(t_{kn}) X_j(t_{in}) X_j(t_{kn})] \langle \mathfrak{R}_{m,\lambda}(t_{in}, \cdot), \mathfrak{R}_{m,\lambda}(t_{kn}, \cdot) \rangle_{m,\lambda}.$$

Further, from (2.3),

$$\mathbb{E}[\varepsilon(t_{in}) \varepsilon(t_{kn}) X_j(t_{in}) X_j(t_{kn})] = 0 \quad \text{for } i \neq k,$$

and with (2.5),

$$\mathbb{E}[|\varepsilon(t_{in})|^2 |X_j(t_{in})|^2] \leq c \mathbb{E}[|\varepsilon(t_{in})|^2] \leq c\sigma^2.$$

Finally, with (3.13), we have

$$(\mathfrak{R}_{m\lambda}(t_{in}, \cdot), \mathfrak{R}_{m\lambda}(t_{in}, \cdot))_{m,\lambda} = \|\mathfrak{R}_{m\lambda}(t_{in}, \cdot)\|_{m,\lambda}^2 \asymp (n\lambda)^{-1},$$

so that  $\mathbb{E}[S^2] \leq c(nh)^{-1}$ , and the lemma follows with Cauchy–Schwarz.  $\square$

The left hand side of the inequality (4.2) also contains a random sum. This was handled in (5.2) by referring to the following lemma.

**Lemma 7.** Let  $\theta_{1,n}, \theta_{2,n}, \dots, \theta_{n,n}$  be independent random variables with mean zero and

$$|\theta_{in}| \leq \gamma, \quad i = 1, 2, \dots, n.$$

Then, for all  $\lambda$  with  $0 < \lambda < 1$ ,

$$\left\| \frac{1}{n} \sum_{i=1}^n \theta_{in} \mathfrak{R}_{m\lambda}(t_{in}, \cdot) \right\|_{m,\lambda} = \mathcal{O}((n\lambda/\log n)^{-1/2}) \quad \text{almost surely.}$$

**Proof.** Let  $\theta_n = (\theta_{1,n}, \theta_{2,n}, \dots, \theta_{n,n})^T$  and define

$$\psi(\theta_{1,n}, \theta_{2,n}, \dots, \theta_{n,n}) \equiv \psi(\theta_n) = \left\| \frac{1}{n} \sum_{i=1}^n \theta_{in} \mathfrak{R}_{m\lambda}(t_{in}, \cdot) \right\|_{m,\lambda}. \tag{6.2}$$

Then, for all  $|\zeta_i| \leq \gamma, i = 2, \dots, n$ , the maximal variation of  $\psi(\theta_n)$  over its first argument,

$$s_{1,n} = \sup_{|a| \leq \gamma, |b| \leq \gamma} |\psi(a, \zeta_2, \dots, \zeta_n) - \psi(b, \zeta_2, \dots, \zeta_n)|,$$

satisfies

$$\begin{aligned} s_{1,n} &\leq \sup_{|a| \leq \gamma, |b| \leq \gamma} \frac{1}{n} \|(a - b)\mathfrak{R}_{m\lambda}(t_{in}, \cdot)\|_{m,\lambda} \\ &\leq 2\gamma n^{-1} \|\mathfrak{R}_{m\lambda}(t_{in}, \cdot)\|_{m,\lambda} \leq cn^{-1} \lambda^{-1/2}, \end{aligned}$$

for the appropriate constant  $c$ , not depending on  $n$ . The same bound with the same  $c$  applies to  $s_{in}$ , the total variation of  $\psi$  over the  $i$ -th argument, for all  $i$ .

The McDiarmid–Devroye exponential inequality, see, e.g., Devroye (1991) or Devroye et al. (1996), now implies that

$$\mathbb{P}[|\psi(\theta_n) - \mathbb{E}[\psi(\theta_n)]| \geq u] \leq 2 \exp(-\frac{1}{2}u^2/s_n^2), \tag{6.3}$$

where

$$s_n^2 = \sum_{i=1}^n s_{in}^2 \leq c(n\lambda)^{-1}.$$

Now, take

$$u = u_n = 2s_n(\log n)^{1/2},$$

to conclude that

$$\mathbb{P}[|\psi(\theta_n) - \mathbb{E}[\psi(\theta_n)]| \geq u_n] \leq 2n^{-2}, \tag{6.4}$$

so that with Borel–Cantelli,

$$|\psi(\theta_n) - \mathbb{E}[\psi(\theta_n)]| = \mathcal{O}(u_n) = \mathcal{O}((n\lambda/\log n)^{-1/2}) \quad \text{almost surely.}$$

Since  $\mathbb{E}[\psi(\theta_n)] \leq \{\mathbb{E}[|\psi(\theta_n)|^2]\}^{1/2} \leq c(n\lambda)^{-1/2}$ , this proves the lemma.  $\square$

## References

- Adams, R.A., Fournier, J.J.F., 2003. Sobolev Spaces. second ed. Academic Press, Amsterdam.
- Cai, Z., Fan, J., Li, R., 2000. Efficient estimation and inference for varying coefficient models. *J. Amer. Statist. Assoc.* 95, 888–902.
- Chiang, C., Rice, J., Wu, C., 2001. Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Amer. Statist. Assoc.* 96, 605–619.
- Devroye, L., 1991. Exponential inequalities in nonparametric estimation. In: Roussas, G. (Ed.), *Nonparametric Functional Estimation and Related Topics*. Kluwer Academic Publishers, Dordrecht, pp. 31–44.
- Devroye, L., Györfi, L., Lugosi, G., 1996. *A Probabilistic Theory of Pattern Recognition*. Springer, New York.
- Eggermont, P.P.B., LaRiccia, V.N., 2006. Uniform error bounds for smoothing splines. In: Koltchinskii, V.I., Li, W., Zinn, J. (Eds.), *IMS Lecture Notes—Monograph Series*, vol. 51, pp. 220–237.
- Eggermont, P.P.B., LaRiccia, V.N., 2009. *Maximum Penalized Likelihood Estimation. Volume II: Regression*. Springer, New York.
- Eubank, R.L., 1999. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.
- Eubank, R.L., Huang, C., Muñoz Maldonado, Y., Wang, N., Wang, S., Buchanan, R.J., 2004. Smoothing spline estimation in varying-coefficient models. *J. Roy. Statist. Soc. B* 66, 653–667.
- Fan, J., Zhang, W., 1999. Statistical estimation in varying coefficient models. *Ann. Statist.* 27, 1491–1518.
- Fan, J., Zhang, J., 2000. Two-step estimation of functional linear models with applications to longitudinal data. *J. Roy. Statist. Soc. B* 62, 303–322.
- Green, P.J., Jennison, C., Seheult, A., 1985. Analysis of field experiments by least squares smoothing. *J. Roy. Statist. Soc. B* 47, 299–315.
- Hastie, T., Tibshirani, R., 1993. Varying coefficient models. *J. Roy. Statist. Soc. B* 55, 757–796.
- Hoover, D., Rice, J., Wu, C., Yang, L., 1998. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* 85, 809–822.
- Ribièrè, G., 1967. Régularisation d'opérateurs. *Rev. Informat. Recherche Opérationnelle* 1, 57–79.
- van de Geer, S.A., 2000. *Applications of Empirical Process Theory*. Cambridge University Press, Cambridge.
- Wahba, G., 1990. *Spline Models for Observational Data*. SIAM, Philadelphia.
- Wang, Y., 1998. Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.* 93, 341–348.
- Wu, C., Chiang, C., 2000. Kernel smoothing of varying coefficient models with longitudinal dependent variables. *Statist. Sinica* 10, 433–456.
- Wu, C., Chiang, C., Hoover, D., 1998. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Amer. Statist. Assoc.* 93, 1388–1402.
- Wu, C., Yu, K., Chiang, C., 2000. A two-step smoothing method for varying-coefficient models with repeated measurements. *J. Inst. Statist. Math.* 52, 519–543.