Available at WWW.MATHEMATICSWEB.ORG
POWERED BY SCIENCE DIRECT

Journal of Statistical Planning and Inference 116 (2003) 451 – 474

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Statistical modeling of diffusion processes with free knot splines

Charles J. Stone[a,*], Jianhua Z. Huang[b]

[a] *Department of Statistics, University of California, Berkeley, CA 94720-3860, USA*
[b] *Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104-6302, USA*

Received 18 July 2001; accepted 7 May 2002

Abstract

We consider the nonparametric estimation of the drift coefficient in a diffusion type process in which the diffusion coefficient is known and the drift coefficient depends in an unknown manner on a vector of time-dependent covariates. Based on many continuous realizations of the process, the estimator is constructed using the method of maximum likelihood, where the maximization is taken over a finite dimensional estimation space whose dimension grows with the sample size n. We focus on estimation spaces of polynomial splines. We obtain rates of convergence of the spline estimates when the knot positions are prespecified but the number of knots increases with the sample size. We also give the rates of convergence for free knot spline estimates, in which the knot positions of splines are treated as free parameters that are determined by data.

MSC: primary 62G07; secondary 62J12

Keywords: Drift coefficient; Maximum partial likelihood; Method of Sieves; Nonparametric estimation; Rates of convergence; Stochastic integrals; Time-dependent covariates

1. Introduction

Diffusion type processes form a large class of continuous time processes that are widely used for stochastic modeling with application to physical, biological, medical, economic, and social sciences. Statistical estimation and inference for diffusion type processes are of considerable importance and have been extensively studied in the past 20 years; see Prakasa Rao (1999a,b) and the many references cited therein, especially in the latter book.

* Corresponding author.
E-mail address: stone@stat.berkeley.edu (C.J. Stone).

PII: S0378-3758(02)00349-X

In this paper we consider the nonparametric estimation of the drift coefficient in a diffusion type process with time-dependent covariates. As is usual in this context (see Prakasa Rao, 1999b), we assume that the diffusion coefficient is a known, possibly random, function of time. We are interested in estimating the drift coefficient as an unknown function of time and a vector of possibly internal time-dependent covariates. The estimator is constructed based on many continuous realizations of the process. To study the asymptotic properties of estimators, we let the number n of independent realizations of the process observed over a fixed time period $[0,T]$ tend to infinity. The same type of asymptotics for estimation in diffusion type processes has been considered by Nguyen and Pham (1982), McKeague (1986), Beder (1987), and Leskow and Rozanski (1989), among others.

To estimate the drift coefficient, we employ the method of maximum likelihood, where the maximization is taken over a finite dimensional estimation space $\mathbb{G}$ whose dimension grows with the sample size n. The estimation space can be chosen as a space of polynomials, trigonometric polynomials, or polynomial splines. Our approach is a special case of the method of sieves (Grenander, 1981). We focus on polynomial splines in this paper. First, we give rates of convergence for spline estimates when the knot positions are prespecified. To allow better approximation for large samples, the number of knots will increase with the sample size. Secondly, we will consider free knot splines, that is, when the knot positions are treated as free parameters that are determined by data. Our theoretical study of free knot splines is motivated by various methodological considerations in the literature involving free knot splines; see, for example, Stone et al. (1997) and the references therein, Lindstrom (1999), Hansen and Kooperberg (2002), Zhou and Shen (2001), and Kooperberg and Stone (2002a,b). From the theoretical point view, the benefit of using free knot splines versus fixed knot splines lies in the potential for better rates of convergence.

It turns out that the estimation problem we consider can be cast in the framework of extended linear models; see Stone et al. (1997). The general theory of extended linear modeling has developed gradually over the last couple of decades: see Stone (1985, 1986, 1990, 1991, 1994); Hansen (1994); Kooperberg et al. (1995a,b); Huang (1998a,b); Huang and Stone (1998); Huang et al. (2000). A general treatment is given in Huang (2001). Recently, Stone and Huang (2002), referred to in the sequel as SH (2002), have adapted the theory to handle free knot splines.

An important advantage of the extended linear modeling framework is that it allows for a common approach to a wide variety of seemingly separate contexts: ordinary regression; generalized regression, including logistic and Poisson regression; multiple logistic regression and multiple classification; density estimation; conditional density estimation; spectral density estimation; hazard regression; and counting process regression, with or without marks. It is also very convenient to discuss dimensionality reduction through structural models such as additive models in extended linear models. However in the paper, for brevity, we will restrict attention to saturated models; that is, we will not consider additive models or ANOVA models involving selected low-order interaction terms (see the Remark at the end of Section 2 for discussion of results along this line).

Huang (2001) and SH (2002) laid down the theoretical framework for extended linear modeling and provided general results under very broad conditions, the verification of which in each specific context (ordinary regression, density estimation, and so forth) using primitive conditions in that context is necessary and could be a challenging task. The main contributions of this paper lie in realizing that estimating the dependence of the drift coefficient of a diffusion type process on a vector of time-dependent covariates is a mathematically tractable special case of extended linear modeling and in establishing results under primitive conditions on the process. This paper can be thought of as a novel application of the extended linear model framework to a new context.

There are several extensions of the results in this paper that would be of practical interest. One would be to construct an estimator based on one long continuous realization of the process; that is, the process would be observed over the time interval $[0, T]$, where T tends to infinity. Another interesting extension would be to processes that are observed at discrete time instants. We anticipate that substantial further technical advances would be necessary to treat these extensions, but the techniques developed in this paper should be helpful.

A more straightforward extension would be to handle diffusion processes with jumps (Prakasa Rao, 1999a) and time-dependent covariates. The special case of pure jump processes (that is, without a continuous component) was treated theoretically by Li (2000), where an application to continuous-time currency exchange rate data was also given. Li's dissertation and the possible extension that it suggested provided the initial motivation for the present paper.

In Section 2 we describe the basic setup and state the main results of estimation of the drift coefficient. Sections 2.2 and 2.3 give rates of convergence for fixed knot spline estimates and free knot spline estimates, respectively. Section 3 provides some preliminary lemmas. The proofs of the main results are given in Section 4.

We conclude this section by introducing some notation. Given positive numbers a_n and b_n for $n \geqslant 1$, let $a_n \lesssim b_n$ mean that a_n/b_n is bounded and let $a_n \asymp b_n$ mean that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Given random variables V_n for $n \geqslant 1$, let $V_n = \mathrm{O}_P(b_n)$ mean that $\lim_{c\to\infty} \limsup_n P(|V_n| \geqslant cb_n) = 0$, and let $V_n = \mathrm{o}_P(b_n)$ mean that $\limsup_n P(|V_n| \geqslant cb_n) = 0$ for all $c > 0$. For a random variable V, let E_n denote expectation relative to its empirical distribution; that is, $E_n(V) = n^{-1} \sum_i V_i$, where V_i, $1 \leqslant i \leqslant n$, is a random sample from the distribution of V. We use $c_1, c_2, \ldots$ to denote positive constants that may change from context to context.

2. Statement of results

2.1. Setup

Consider a one-dimensional diffusion type process $Y(t)$ that satisfies

$$\mathrm{d}Y(t) = \eta(t, \boldsymbol{X}(t)) + \sigma(t)\,\mathrm{d}\boldsymbol{W}(t), \qquad 0 \leqslant t \leqslant \tau,$$

where $0 < \tau < \infty$ and $\boldsymbol{W}(t)$ is a Wiener process. It is assumed that the diffusion coefficient $\sigma^2(t)$ at time t is a known, predictable, random function of time. It is also assumed that the value at time t of the drift coefficient is an unknown function $\eta(t, \boldsymbol{X}(t))$ of t and the value at time t of a predictable covariate process $\boldsymbol{X}(t) = (X_1(t), \ldots, X_L(t))$, $0 \leqslant t \leqslant \tau$. We refer to η as the *regression function*. Let $Z(t)$, $0 \leqslant t \leqslant \tau$, be a predictable $\{0, 1\}$-valued process. The process $Z(t)$ can be thought of as a censoring indicator: the processes $\boldsymbol{X}(t)$ and $Y(t)$ are only observed when $Z(t) = 1$. We will consider inference for the regression function based on a random sample of n realizations of $\{(\boldsymbol{X}(t), Y(t))\colon\ 0 \leqslant t \leqslant \tau \text{ and } Z(t) = 1\}$. Nguyen and Pham (1982) and McKeague (1986) have considered the special case of this setup when $\eta(t, \boldsymbol{X}(t)) = \boldsymbol{X}(t)^T \alpha(t)$ with $\alpha(t) = (\alpha_1(t), \ldots, \alpha_L(t))$ and $Z(t) \equiv 1$.

The (partial) log-likelihood corresponding to a candidate h for η based on a single observation is given by

$$l(h) = \int Z(t)\, \frac{h(t, \boldsymbol{X}(t))}{\sigma^2(t)}\, \mathrm{d}Y(t) - \frac{1}{2} \int Z(t)\, \frac{h^2(t, \boldsymbol{X}(t))}{\sigma^2(t)}\, \mathrm{d}t.$$

This can be seen either by passing to the limit from a discrete-time approximation or by modeling $(\boldsymbol{X}(t), Y(t))$, $0 \leqslant t \leqslant \tau$, as a multidimensional diffusion process and determining the appropriate partial log-likelihood. The expected log-likelihood is given by

$$\begin{aligned} \Lambda(h) &= E\left(\int Z(t)\, \frac{h(t, \boldsymbol{X}(t))}{\sigma^2(t)}\, \mathrm{d}Y(t) - \frac{1}{2} \int Z(t) \frac{h^2(t, \boldsymbol{X}(t))}{\sigma^2(t)}\, \mathrm{d}t\right) \\ &= E\left(\int Z(t)\, \frac{h(t, \boldsymbol{X}(t))\eta(t, \boldsymbol{X}(t))}{\sigma^2(t)}\, \mathrm{d}t - \frac{1}{2} \int Z(t)\, \frac{h^2(t, \boldsymbol{X}(t))}{\sigma^2(t)}\, \mathrm{d}t\right). \end{aligned}$$

For a random sample of size n from the process $(\boldsymbol{X}(t), Y(t), Z(t))$, $0 \leqslant t \leqslant \tau$, the corresponding (normalized) log-likelihood is given by

$$\ell(h) = \frac{1}{n} \sum_i \int Z_i(t)\, \frac{h(t, \boldsymbol{X}_i(t))}{\sigma_i^2(t)}\, \mathrm{d}Y_i(t) - \frac{1}{2n} \sum_i \int Z_i(t)\, \frac{h^2(t, \boldsymbol{X}_i(t))}{\sigma_i^2(t)}\, \mathrm{d}t.$$

As in the framework of extended linear models (Huang, 2001), we will estimate the unknown function η by maximizing this log-likelihood over a finite dimensional estimation space (or approximation space).

Let $[a_l, b_l]$ be a compact interval for $1 \leqslant l \leqslant L$, let $\mathscr{X}$ denote the Cartesian product of $[a_l, b_l]$, $1 \leqslant l \leqslant L$, and set $\mathscr{U} = [0, \tau] \times \mathscr{X}$. Suppose that, for $0 \leqslant t \leqslant \tau$, $Z(t) = 0$ unless $\boldsymbol{X}(t) \in \mathscr{X}$. Then we can think of the regression function η and the candidate functions h as being functions on $\mathscr{U}$. Given such a function h, set $\|h\|_\infty = \sup_{\boldsymbol{u} \in \mathscr{U}} |h(\boldsymbol{u})|$ and let $\|h\|_{L_2}$ denote the standard L_2 norm relative to Lebesgue measure on $\mathscr{U}$. We restrict attention to bounded candidates h for η.

The following technical assumptions will be used in this paper.

Assumption 2.1. $\|\eta\|_\infty \leqslant M_0$ for some constant $M_0 > 0$.

Assumption 2.2. There are constants $M_1 > 0$ and $M_2 \geqslant M_1$ such that $M_1 \leqslant \sigma^{-2}(t) \leqslant M_2$ for $t \in [0, \tau]$ such that $Z(t) = 1$.

Let $\psi(A)$ denote the Lebesgue measure of a Borel subset A of $\mathscr{U}$.

Assumption 2.3. There are constants $M_3 > 0$ and $M_4 \geqslant M_3$ such that

$$M_3\psi(A) \leqslant E\left(\int Z(t)\mathrm{ind}((t, \boldsymbol{X}(t)) \in A)\,\mathrm{d}t\right) \leqslant M_4\psi(A)$$

for every Borel subset A of $\mathscr{U}$.

Suppose that, for $0 \leqslant t \leqslant \tau$, the conditional distribution of $\boldsymbol{X}(t)$ given that $Z(t) = 1$ has a density $f_{\boldsymbol{X}(t)|Z(t)=1}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathscr{X}$. Given a Borel subset A of $\mathscr{U}$, set $A_t = \{\boldsymbol{x} \in \mathscr{X}\colon (\boldsymbol{x}, t) \in A\}$ for $0 \leqslant t \leqslant \tau$. Then

$$E\left(\int Z(t)\mathrm{ind}((t, \boldsymbol{X}(t)) \in A)\,\mathrm{d}t\right) = \int P(Z(t) = 1)\left(\int_{A_t} f_{\boldsymbol{X}(t)|Z(t)=1}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right)\mathrm{d}t.$$

Thus a sufficient condition for Assumption 2.3 is that (i) $P(Z(t) = 1)$ is bounded away from zero uniformly over $t \in [0, \tau]$ and (ii) $f_{\boldsymbol{X}(t)|Z(t)=1}(\boldsymbol{x})$ is bounded away from zero and infinity uniformly over $(t, \boldsymbol{x}) \in \mathscr{U}$.

2.2. *Spline estimation: fixed knots*

We consider in this section the maximum likelihood estimation over an appropriate finite dimensional space of spline functions. Here, the number of knots is allowed to increase deterministically with the sample size while, for each sample size, the knot positions are fixed. We first provide a result for general estimation space and then specify to spline estimation.

Consider an estimation space $\mathbb{G} = \mathbb{G}_n$ whose dimension N_n is finite and may tend to infinity with the sample size n. Let $\hat{\eta}$ denote the maximum likelihood estimate in $\mathbb{G}$ of η; that is, $\hat{\eta} = \mathrm{argmax}_{g \in \mathbb{G}}\, \ell(g)$, where $\ell(g)$ is the normalized log-likelihood defined above. We now give conditions for the consistency of $\hat{\eta}$ and derive the rate of convergence.

Since $\hat{\eta}$ maximizes the scaled log-likelihood $\ell(g)$, which should be close to the expected log-likelihood $\Lambda(g)$ for $g \in \mathbb{G}$ when sample size is large, it is natural to think that $\hat{\eta}$ is directly estimating the best approximation $\bar{\eta} = \mathrm{argmax}_{g \in \mathbb{G}} \Lambda(g)$ in $\mathbb{G}$ to η. If $\mathbb{G}$ is chosen such that $\bar{\eta}$ is close to η then $\hat{\eta}$ should provide a reasonable estimate of η. This motivates the decomposition

$$\hat{\eta} - \eta = (\bar{\eta} - \eta) + (\hat{\eta} - \bar{\eta}),$$

where $\bar{\eta} - \eta$ and $\hat{\eta} - \bar{\eta}$ are referred to, respectively, as the *approximation error* and *estimation error*.

Set $A_n = \sup_{g \in \mathbb{G}, \|g\|_{L_2} \neq 0} \{\|g\|_\infty / \|g\|_{L_2}\} \geqslant 1$, and $\rho_n = \inf_{g \in \mathbb{G}} \|g - \eta\|_\infty$.

Theorem 2.1. *Suppose Assumptions* 2.1–2.3 *hold and that* $\lim_n A_n \rho_n = 0$ *and* $\lim_n A_n^2 N_n/n = 0$. *Then* $\bar{\eta}$ *exists uniquely for* n *sufficiently large and* $\|\bar{\eta} - \eta\|_{L_2}^2 = O(\rho_n^2)$. *Moreover*, $\hat{\eta}$ *exists uniquely except on an event whose probability tends to zero as* $n \to \infty$ *and* $\|\hat{\eta} - \bar{\eta}\|_{L_2}^2 = O_P(N_n/n)$. *Consequently*, $\|\hat{\eta} - \eta\|_{L_2}^2 = O_P(N_n/n + \rho_n^2)$. *In particular*, $\|\hat{\eta} - \eta\|_{L_2} = o_P(1)$; *that is*, $\hat{\eta}$ *is consistent.*

The bounds for the magnitudes of the estimation and approximation errors can be interpreted intuitively as follows: N_n/n is just the inverse of the number of observations per parameter, and ρ_n is the best possible approximation rate in the estimation space to the target function η.

Let us now specialize to spline estimation. To this end, write $\mathscr{U} = \mathscr{U}_0 \times \cdots \times \mathscr{U}_L$, where $\mathscr{U}_0 = [0, \tau]$ and $\mathscr{U}_l = [a_l, b_l]$ for $1 \leqslant l \leqslant L$. For each l, let m_l be an integer with $m_l \geqslant 2$ (not depending on n), let J_l be a positive integer, and let γ_{lj}, $1 \leqslant j \leqslant J_l$, be such that $a < \gamma_{l1} \leqslant \cdots \leqslant \gamma_{lJ_l} < b$ and $\gamma_{l,j-1} > \gamma_{l,j-m}$ for $2 \leqslant j \leqslant J_l + m_l$, where $\gamma_{lj} = a$ for $1 - m_l \leqslant j \leqslant 0$ and $\gamma_{lj} = b$ for $J_l + 1 \leqslant j \leqslant J_l + m_l$. Let $\mathbb{G}_{l\gamma_l}$ be the space of polynomial splines of order m_l (degree $m_l - 1$) on $\mathscr{U}_l$ with the interior knot sequence $\boldsymbol{\gamma}_l = (\gamma_{l1}, \ldots, \gamma_{lJ_l})$, whose dimension $J_l + m_l$ is denoted by N_{nl} to indicate its possible dependence on the sample size n. For $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0, \ldots, \boldsymbol{\gamma}_L)$, let $\mathbb{G}_{\boldsymbol{\gamma}}$ be the tensor product of $\mathbb{G}_{l\gamma_l}$, $0 \leqslant l \leqslant L$, (that is, the linear space spanned by $g_0(u_0) \cdots g_L(u_l)$ as g_l runs over $\mathbb{G}_{l\gamma_l}$) which has dimension $N_n = \prod_l N_{nl}$.

Let the knot specification $\boldsymbol{\gamma}$ possibly depend on the sample size but otherwise be fixed. For $0 \leqslant l \leqslant L$, let $\bar{M}_l \geqslant 1$ be a fixed positive number and suppose the knot sequence $\boldsymbol{\gamma}_l = (\gamma_{l1}, \ldots, \gamma_{lJ_l})$ on $\mathscr{U}_l$ satisfies

$$\frac{\gamma_{l,j_2-1} - \gamma_{l,j_2-m_l}}{\gamma_{l,j_1-1} - \gamma_{l,j_1-m_l}} \leqslant \bar{M}_l, \quad 2 \leqslant j_1, j_2 \leqslant J_l + m_l, \tag{2.1}$$

where $\gamma_{l,1-m_l} = \cdots = \gamma_{l0} = a$ and $\gamma_{l,J_l+1} = \cdots = \gamma_{l,J_l+m_l} = b$. It follows from (2.1) and (3.1) that $A_n \lesssim N_n^{1/2}$; see SH (2002, proof of Lemma 3.3). Thus the condition $\lim_n A_n^2 N_n/n = 0$ reduces to $\lim_n N_n^2/n = 0$.

To get a precise result, let us introduce a commonly used smoothness condition. Let $0 < \beta \leqslant 1$, let k be a nonnegative integer, and set $p = k + \beta$. A function on $\mathscr{U}$ is said to be *p-smooth* if it is k times continuously differentiable on $\mathscr{U}$ and D^i satisfies a Hölder condition with exponent β for all i with $[i] = k$. Suppose η is p-smooth and that $N_{nl} \asymp N_{nl'}$ for $0 \leqslant l, l' \leqslant L$. Then $\rho_n \asymp N_{nl}^{-p} \asymp N_n^{-p/(L+1)}$ (see Schumaker, 1981, Equation 13.69 and Theorem 12.8). It follows from Theorem 2.1 that, if $p > (L+1)/2$, $\lim_n N_n = 0$ and $\lim_n N_n^2/n = 0$, then $\|\hat{\eta} - \eta\|_{L_2}^2 = O_P(N_n/n + N_n^{-2p/(L+1)})$. In particular, for $N_n \asymp n^{(L+1)/(2p+L+1)}$, we have that $\|\hat{\eta} - \eta\|_{L_2}^2 = O_P(n^{-2p/(2p+L+1)})$.

2.3. Spline estimation: free knots

We now apply the maximum likelihood method to free knot splines to estimate the regression function.

For $0 \leqslant l \leqslant L$, let $\bar{M}_l \geqslant 1$ be a fixed positive number and let Γ_l denote the collection of free knot sequences $\boldsymbol{\gamma}_l = (\gamma_{l1}, \ldots, \gamma_{lJ_l})$ on $\mathscr{U}_l$ such that (2.1) holds. Let Γ denote the Cartesian product of Γ_l, $0 \leqslant l \leqslant L$, which can be viewed as a subset of $\mathbb{R}^J$ with $J = \sum_l J_l$. We consider the use of the collection $\mathbb{G}_\gamma, \gamma \in \Gamma$, which can be viewed as a (nonlinear) space of free knot splines, in estimating the regression function.

For each fixed $\gamma \in \Gamma$, the maximum likelihood estimate is given by $\hat{\eta}_\gamma = \max_{g \in \mathbb{G}_\gamma} \ell(g)$. To let the data pick the knot positions, we choose $\hat{\gamma} \in \Gamma$ such that $\ell(\hat{\eta}_{\hat{\gamma}}) = \max_{\gamma \in \Gamma} \ell(\hat{\eta}_\gamma)$. We refer to $\hat{\eta}_{\hat{\gamma}}$ as the free knot spline estimator. We will study the benefit of using free knot splines through the rate of convergence of $\hat{\eta}_{\hat{\gamma}} - \eta$. The following theorem is about the existence of $\hat{\gamma}$.

Theorem 2.2. *Suppose Assumptions* 2.1–2.3 *hold. There is a* $\gamma^* \in \Gamma$ *such that* $\rho_{n\gamma^*} = \inf_{\gamma \in \Gamma} \rho_{n\gamma}$. *Moreover, almost surely there is a* $\hat{\gamma} \in \Gamma$ *such that* $\ell(\hat{\eta}_{\hat{\gamma}}) = \max_{\gamma \in \Gamma} \ell(\hat{\eta}_\gamma)$.

It follows from Theorem 2.1 that, $\|\hat{\eta}_\gamma - \eta\|_{L_2}^2 = O_P(\rho_{n\gamma}^2 + N_n/n)$ for each fixed $\gamma \in \Gamma$. Let γ^* be such that $\rho_{n\gamma^*} = \inf_{\gamma \in \Gamma} \rho_{n\gamma}$. (Such a γ^* exists; see Theorem 2.2.) Then

$$\inf_{\gamma \in \Gamma} \|\hat{\eta}_\gamma - \eta\|_{L_2}^2 \leqslant \|\hat{\eta}_{\gamma^*} - \eta\|_{L_2}^2 = O_P(\rho_{n\gamma^*}^2 + N_n/n) = O_P\left(\inf_{\gamma \in \Gamma} \rho_{n\gamma}^2 + N_n/n\right).$$

It is natural to expect that, with high probability, $\|\hat{\eta}_{\hat{\gamma}} - \eta\|_{L_2}^2$ will be not much larger than $\inf_{\gamma \in \Gamma} \|\hat{\eta}_\gamma - \eta\|_{L_2}^2$ or $\inf_{\gamma \in \Gamma} \rho_{n\gamma}^2 + N_n/n$. The next theorem is concerned with justifying this heuristic under suitable conditions. Ideally, we would like to show that $\|\hat{\eta}_{\hat{\gamma}} - \eta\|_{L_2}^2 = O_P(\inf_{\gamma \in \Gamma} \rho_{n\gamma}^2 + N_n/n)$, but as in SH (2002) we are able only to prove a somewhat weaker result.

For $\gamma \in \Gamma$, set $N_n = \dim(\mathbb{G}_\gamma)$, $\rho_{n\gamma} = \inf_{g \in \mathbb{G}_\gamma} \|g - \eta\|_\infty$, and $\rho_n = \sup_{\gamma \in \Gamma} \rho_{n\gamma}$. Also, let $\tilde{\Gamma}$ be defined in the same way as Γ, but with $\bar{M}_l$ in (2.2) replaced by $3\bar{M}_l$. Set $\tilde{\rho}_n = \sup_{\gamma \in \tilde{\Gamma}} \rho_{n\gamma}$.

Condition 2.1. $N_n = o(n^{1/2})$, $N_n^{-(c-1/2)} \lesssim \log^{-1/2} n$, and $\tilde{\rho}_n = O(N_n^{-c})$ for some $c > \frac{1}{2}$.

If η satisfies a Hölder condition with exponent $c > \frac{1}{2}$ (that is, if there is a positive number γ such that $|h(\boldsymbol{u}_2) - h(\boldsymbol{u}_1)| \leqslant \gamma |\boldsymbol{u}_2 - \boldsymbol{u}_1|^c$ for $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathscr{U}$), then $\tilde{\rho}_n = O(N_n^{-c})$ (see Schumaker, 1981, Eq. (13.69) and Theorem 12.8). In this case, a sufficient condition for Condition 2.1 is that $N_n \asymp n^\delta$ for some δ with $0 < \delta < \frac{1}{2}$.

Let $V_{n\gamma} = O_P(b_{n\gamma})$ uniformly over $\gamma \in \Gamma$ mean that $\lim_{c \to \infty} \limsup_n P(|V_{n\gamma}| \geqslant c b_{n\gamma}$ for some $\gamma \in \Gamma) = 0$. Let $V_n = \bar{O}_P(b_n)$ mean that $\lim_n P(|V_n| \geqslant c b_n) = 0$ for some $c > 0$, where $b_n > 0$ for $n \geqslant 1$. Note that $V_n = \bar{O}_P(b_n)$ is a slightly stronger statement than $V_n = O_P(b_n)$. Let M_0' be any constant with $M_0' > M_0$ for M_0 in Assumption 2.1.

Theorem 2.3. *Suppose Assumptions* 2.1–2.3 *and Condition* 2.1 *hold.*

(i) *For* $\gamma \in \tilde{\Gamma}$ *there is a unique function* $\bar{\eta}_\gamma \in \mathbb{G}_\gamma$ *that maximizes the expected log-likelihood over* $\mathbb{G}_\gamma$. *Moreover,* $\sup_{\gamma \in \Gamma} \|\bar{\eta}_\gamma\|_\infty \leqslant M_0'$ *for n sufficiently large and* $\|\bar{\eta}_\gamma - \eta\|_{L_2} = O(\rho_{n\gamma})$ *uniformly over* $\gamma \in \Gamma$.

(ii) $\hat{\eta}_\gamma$ *exists uniquely for* $\gamma \in \Gamma$ *and* $\sup_{\gamma \in \Gamma} \|\hat{\eta}_\gamma\|_\infty \leqslant M_0'$ *on an event whose probability tends to one as* $n \to \infty$. *Moreover,* $\sup_{\gamma \in \Gamma} \|\hat{\eta}_\gamma - \bar{\eta}_\gamma\|_{L_2}^2 = O_P(N_n/n)$.

(iii) $\|\bar{\eta}_{\hat{\gamma}} - \eta\|_{L_2}^2 = O_P(\inf_{\gamma \in \Gamma} \rho_{n\gamma}^2) + \bar{O}_P((\log n) N_n/n)$.

Note that it follows from (ii) and (iii) of Theorem 2.3 that $\|\hat{\eta}_{\hat{\gamma}}-\eta\|_{L_2}^2=\mathrm{O}_P(\inf_{\gamma\in\Gamma}\rho_{n\gamma}^2)+\bar{\mathrm{O}}_P((\log n)N_n/n)$.

It is interesting to compare the rate of convergence here for free knot splines with that obtained in Section 2.2 for fixed knots splines. The use of free knot splines provides smaller approximation error: free knot splines can achieve the smallest possible approximation error ($\inf_{\gamma\in\Gamma}\rho_{n\gamma}$) over a class of knot configurations. The cost is a small inflation of the variance, as there is an extra log$n$ term in the variance bound for free knot splines.

To illustrate of the improved rate of convergence of free knot spline estimates over fixed knot spline estimates, consider, for simplicity, estimating the regression function $\eta(x)$ in the regression model $Y=\eta(X)+\varepsilon$, where X has a uniform distribution on $[-1,1]$ and ε has a normal distribution with mean 0 and variance σ^2. Suppose $\eta(x)=x^2+|x|$. Clearly, the first derivative of η at 0 does not exist, which implies a slow fixed knot spline approximation rate if there is no knot very close to 0. Specifically, let the estimation space $\mathbb{G}$ be the space of linear splines on $[-1,1]$ with $2J_n$ equally spaced knots located at $\pm(2k-1)/(2J_n-1)$, $k=1,\dots,J_n$, and let $\hat{\eta}$ be the least squares estimate on $\mathbb{G}$. It can be shown (see Appendix A) that, for some positive constant c and large n, $P(\|\hat{\eta}-\eta\|^2>cJ_n/n+cJ_n^{-3})>\frac{1}{3}$ and thus for all choices of J_n, $P(\|\hat{\eta}-\eta\|^2>cn^{-3/4})>\frac{1}{3}$. On the other hand, let $\mathbb{G}_\gamma$ be the space of linear splines on $[-1,1]$ with $2J_n-1$ knots located at γ and $\pm(2k-1)/(2J_n-1)$, $k=2,\dots,J_n$, where $-1/(2J_n-1)\leqslant\gamma\leqslant 1/(2J_n-1)$. Here, we simply replace two fixed knots $\pm 1/(2J_n-1)$ in the previous setup by one free knot at γ. Let $\hat{\eta}_{\hat{\gamma}}$ be the free knot spline least squares estimate. It is easily seen that $\inf_\gamma \rho_{n\gamma}=\mathrm{O}(J_n^{-2})$. Using an analogue of Proposition 2.3 for the regression context (SH, 2002), we obtain that $\|\hat{\eta}_{\hat{\gamma}}-\eta\|^2=\mathrm{O}_P(J_n^{-4}+J_n\log n/n)$. Hence, for $J_n\asymp(n/\log n)^{1/5}$, the convergence rate of the free knot spline estimate satisfies $\|\hat{\eta}_{\hat{\gamma}}-\eta\|^2=\mathrm{O}_P((n^{-1}\log n)^{4/5})$, which is faster than that of the fixed knot spline estimate.

Remark. At the expense of much additional notation, Theorems 2.1 and 2.3 could be extended to additive and more general unsaturated ANOVA models. In the statement of such results, it would be necessary to replace η by its best approximation having the specified form. For estimation based on fixed knot splines of the components of the corresponding ANOVA decomposition, the results in Corollary 2.2 of Huang (2001) hold in the current context. Moreover, these results can be extended to hold uniformly over a class of knot configurations. Here the uniformity is in the same sense as in Theorem 2.3(i) and (ii). However, it is not obvious how to obtain a satisfactory analog of Theorem 2.3(iii) that would apply to the ANOVA components of a free knot spline estimate.

3. Preliminary lemmas

3.1. Theoretical and empirical inner products and norms

As in the general theory of extended linear models (Huang, 2001), we introduce some inner products and norms that are convenient for our technical arguments. Consider first the theoretical inner product and norm given by

$$\langle h_1, h_2\rangle = E\left(\int Z(t)\,\frac{h_1(t,\boldsymbol{X}(t))h_2(t,\boldsymbol{X}(t))}{\sigma^2(t)}\,\mathrm{d}t\right)$$

and $\|h\|^2 = \langle h, h\rangle$. Under Assumptions 2.2 and 2.3,

$$M_5\|h\|_{L_2}^2 \leqslant \|h\|^2 \leqslant M_6\|h\|_{L_2}^2 \tag{3.1}$$

for some constants $M_5 > 0$ and $M_6 \geqslant M_5$. Hence in the statements of Theorems 2.1 and 2.3, the norm $\|\cdot\|_{L_2}$ can be replaced by $\|\cdot\|$.

The theoretical inner product and norm are connected to the expected log-likelihood and its derivatives. In fact,

$$\Lambda(h) = \langle h, \eta\rangle - \frac{\langle h, h\rangle}{2} = \frac{\|\eta\|^2 - \|h-\eta\|^2}{2},$$

which is essentially uniquely maximized at $h = \eta$; that is, h maximizes the expected log-likelihood if and only if $h = \eta$ almost everywhere. Moreover,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\,\Lambda(h_1 + \alpha h_2)\bigg|_{\alpha=0} = \langle h_2, \eta - h_1\rangle \tag{3.2}$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\,\Lambda(h_1 + \alpha(h_2 - h_1)) = -\|h_2 - h_1\|^2. \tag{3.3}$$

We next introduce a data version of the theoretical inner product and norm. The empirical inner product and norm are defined by

$$\begin{aligned}\langle h_1, h_2\rangle_n &= E_n\left(\int Z(t)\,\frac{h_1(t,\boldsymbol{X}(t))h_2(t,\boldsymbol{X}(t))}{\sigma^2(t)}\,\mathrm{d}t\right)\\ &= \frac{1}{n}\sum_i \int Z_i(t)\,\frac{h_1(t,\boldsymbol{X}_i(t))h_2(t,\boldsymbol{X}_i(t))}{\sigma_i^2(t)}\,\mathrm{d}t\end{aligned}$$

and $\|h\|_n^2 = \langle h, h\rangle_n$. The next result, which is SH (2002, Lemma 4.2) with slightly different definitions of the theoretical and empirical inner products and norms, says that the empirical inner product and its theoretical counterpart are close uniformly in the estimation spaces.

Lemma 3.1. *Suppose Assumptions* 2.2 *and* 2.3 *hold and that* $N_n = \mathrm{o}(n^{1/2})$. *Then*

$$\sup_{\gamma,\tilde{\gamma}\in\Gamma}\ \sup_{f\in\mathbb{G}_\gamma}\ \sup_{g\in\mathbb{G}_{\tilde{\gamma}}} \frac{|\langle f, g\rangle_n - \langle f, g\rangle|}{\|f\|\|g\|} = \mathrm{o}_P(1).$$

Consequently, except on an event whose probability tends to zero as $n \to \infty$,

$$\frac{\|g\|^2}{2} \leqslant \|g\|_n^2 \leqslant 2\|g\|^2, \quad \gamma\in\Gamma \text{ and } g\in\mathbb{G}_\gamma.$$

The empirical norm is related to second derivative of the log-likelihood; that is,

$$\frac{d^2}{d\alpha^2}\ell(g_1+\alpha(g_2-g_1)) = -\frac{1}{n}\sum_i \int Z_i(t)\,\frac{[g_2(t,\boldsymbol{X}_i(t))-g_1(t,\boldsymbol{X}_i(t))]^2}{\sigma_i^2(t)}\,dt$$

$$= -\|g_2-g_1\|_n^2.$$

Suppose that Assumptions 2.2 and 2.3 hold and that $\lim_n N_n^2/n=0$. Then, by Lemma 3.1, except on an event whose probability tends to zero as $n\to\infty$,

$$\frac{d^2}{d\alpha^2}\ell(g_1+\alpha(g_2-g_1)) \leqslant -\frac{\|g_2-g_1\|^2}{2}, \quad 0\leqslant\alpha\leqslant 1, \tag{3.4}$$

for $\gamma\in\Gamma$ and $g_1,g_2\in\mathbb{G}_\gamma$.

3.2. Chaining argument

The following result and its proof are modifications of SH (2002, statement and proof of Lemma 4.1). This chaining argument, which is useful for proving the main results of this paper, is well known in the empirical process literature; see Pollard (1984). However, we develop our own version of the chaining argument since we cannot find a version in the literature that is suitable to our purpose.

Lemma 3.2 (Chaining argument). *Let $\mathbb{S}$ be a nonempty subset of $\tilde{\mathbb{S}}$; let V_s, $s\in\tilde{\mathbb{S}}$, be random variables; let $\mathbb{S}_k$ be a finite, nonempty subset of $\tilde{\mathbb{S}}$ for $k\geqslant 0$ such that $V_s=0$ for $s\in\mathbb{S}_0$; let $C_1,\dots,C_4$ be positive numbers; and let Ω be an event. Suppose that*

$$P\left(\lim_{k\to\infty}\min_{\tilde{s}\in\mathbb{S}_k}|V_s-V_{\tilde{s}}|=0 \ \textit{for}\ s\in\mathbb{S}\right)=1, \tag{3.5}$$

$$\#(\mathbb{S}_k)\leqslant C_1\exp(C_2k), \quad k\geqslant 1, \tag{3.6}$$

and

$$\max_{s\in\mathbb{S}_k}\min_{\tilde{s}\in\mathbb{S}_{k-1}} P(|V_s-V_{\tilde{s}}|>2^{-(k-1)}C_3;\Omega)\leqslant C_4\exp(-2C_22^{k-1}), \quad k\geqslant 1. \tag{3.7}$$

Then

$$P\left(\sup_{s\in\mathbb{S}}|V_s|>2C_3\right)\leqslant\frac{C_1C_4}{C_2}+P(\Omega^c).$$

Proof. Let Ω_0 be the event that $\lim_{k\to\infty}\min_{\tilde{s}\in\mathbb{S}_k}|V_s-V_{\tilde{s}}|=0$ for $s\in\mathbb{S}$. Then $P(\Omega_0)=1$. Let $0<\varepsilon<\infty$. Given $\omega\in\Omega_0$, choose $s_0=s_0(\omega)\in\mathbb{S}$ such that $|V_{s_0}|>\sup_{s\in\mathbb{S}}|V_s|-\varepsilon$. Then

$$\liminf_{k\to\infty}\max_{\tilde{s}\in\mathbb{S}_k}|V_{\tilde{s}}|\geqslant|V_{s_0}|>\sup_{s\in\mathbb{S}}|V_s|-\varepsilon.$$

Since ε can be made arbitrarily small, we conclude that

$$\sup_{s\in\mathbb{S}}|V_s|\leqslant\liminf_{k\to\infty}\max_{\tilde{s}\in\mathbb{S}_k}|V_{\tilde{s}}|$$

on Ω_0. Hence

$$\liminf_{k\to\infty} \operatorname{ind}\left(\max_{\tilde{s}\in\mathbb{S}_k} |V_{\tilde{s}}| > 2C_3\right) \geqslant \operatorname{ind}\left(\sup_{s\in\mathbb{S}} |V_s| > 2C_3\right)$$

on Ω_0. Consequently, by Fatou's lemma,

$$P\left(\sup_{s\in\mathbb{S}} |V_s| > 2C_3; \Omega\right) \leqslant \liminf_{k\to\infty} P\left(\max_{\tilde{s}\in\mathbb{S}_k} |V_{\tilde{s}}| > 2C_3; \Omega\right).$$

Therefore it suffices to verify that, for $K \geqslant 1$,

$$P\left(\max_{s\in\mathbb{S}_K} |V_s| > 2C_3; \Omega\right) \leqslant \frac{C_1 C_4}{C_2}. \tag{3.8}$$

To this end, for $1 \leqslant k \leqslant K$, let σ_{k-1} be a map from $\mathbb{S}_k$ to $\mathbb{S}_{k-1}$ such that

$$P(|V_s - V_{\sigma_{k-1}(s)}| > 2^{-(k-1)}C_3; \Omega) \leqslant C_4 \exp(-2C_2 2^{k-1}), \quad 1 \leqslant k \leqslant K \text{ and } s \in \mathbb{S}_k;$$

the existence of σ_{k-1} follows from (3.7). Then, by (3.6),

$$P(|V_s - V_{\sigma_{k-1}(s)}| > 2^{-(k-1)}C_3 \text{ for some } k \in \{1,\dots,K\} \text{ and } s \in \mathbb{S}_k; \Omega)$$

$$\leqslant \sum_{k=1}^{K} C_1 \exp(C_2 k) C_4 \exp(-2C_2 2^{k-1}).$$

Since $k \leqslant 2^{k-1}$ for $k \geqslant 1$, the right side of the above inequality is bounded above by

$$C_1 C_4 \sum_{k=1}^{K} \exp(-C_2 k) \leqslant C_1 C_4 \frac{\exp(-C_2)}{1-\exp(-C_2)} \leqslant \frac{C_1 C_4}{C_2}.$$

Suppose that $|V_s - V_{\sigma_{k-1}(s)}| \leqslant 2^{-(k-1)}C_3$ for $1 \leqslant k \leqslant K$ and $s \in \mathbb{S}_k$. Choose $s \in \mathbb{S}_K$ and set $s_K = s$, $s_{K-1} = \sigma_{K-1}(s_K), \dots, s_0 = \sigma_0(s_1)$. (We refer to $s_K, \dots, s_0$ as forming a "chain" from the point $s \in \mathbb{S}_K$ to a point $s_0 \in \mathbb{S}_0$.) Then $V_{s_0} = 0$ and $|V_{s_k} - V_{s_{k-1}}| \leqslant 2^{-(k-1)}C_3$ for $1 \leqslant k \leqslant K$, so

$$|V_s| = \left|\sum_{k=1}^{K} (V_{s_k} - V_{s_{k-1}})\right| \leqslant 2C_3.$$

Consequently,

$$P\left(\max_{s\in\mathbb{S}_K} |V_s| > 2C_3; \Omega\right) \leqslant P(|V_s - V_{\sigma_{k-1}(s)}| > 2^{-(k-1)}C_3$$
$$\text{for some } k \in \{1,\dots,K\} \text{ and } s \in \mathbb{S}_k; \Omega).$$

Thus (3.8) holds as desired. □

We next introduce a useful device for checking condition (3.5) when applying the chaining argument (Lemma 3.2).

By assumption, the predictable variation of the martingale

$$Y(t) - \int_0^t \eta(s, \boldsymbol{X}(s))\,\mathrm{d}s, \quad 0 \leqslant t \leqslant \tau,$$

is given by $\int_0^t \sigma^2(s)\,\mathrm{d}s$, $0 \leqslant t \leqslant \tau$. Let h be a bounded function on $\mathscr{U}$. Then the predictable variation of the martingale

$$\int_0^t Z(s)\,\frac{h(s, \boldsymbol{X}(s))}{\sigma^2(s)}[\mathrm{d}Y(s) - \eta(s, \boldsymbol{X}(s))\,\mathrm{d}s], \quad 0 \leqslant t \leqslant \tau,$$

is given by

$$\int_0^t Z(s)\,\frac{h^2(s, \boldsymbol{X}(s))}{\sigma^2(s)}\,\mathrm{d}s, \quad 0 \leqslant t \leqslant \tau.$$

Given a positive integer n and a bounded function h on $\mathscr{U}$, set

$$\begin{aligned} W_n(h) &= E_n\left(\int Z(t)\,\frac{h(t, \boldsymbol{X}(t))}{\sigma^2(t)}[\mathrm{d}Y(t) - \eta(t, \boldsymbol{X}(t))\,\mathrm{d}t]\right) \\ &= \frac{1}{n}\sum_{k=1}^{n}\int Z_i(t)\,\frac{h(t, \boldsymbol{X}_i(t))}{\sigma^2(t)}[\mathrm{d}Y_i(t) - \eta(t, \boldsymbol{X}_i(t))\,\mathrm{d}t], \end{aligned}$$

which is uniquely defined up to events of probability zero. Observe that $W_n(h_2) - W_n(h_1) = W_n(h_2 - h_1)$. It follows from van de Geer (1995, Lemma 2.1), which is taken from Shorack and Wellner (1986, p. 899), that, for $a, b > 0$,

$$P(|W_n(h)| \geqslant a \text{ and } \|h\|_n \leqslant b) \leqslant 2\exp\left(-\frac{na^2}{2b^2}\right). \tag{3.9}$$

As in SH (2002), let $|\cdot|_\infty$ denote the l_∞ norm on any Euclidean space and let ζ denote the metric on $\mathbb{R}^J$ given by $\zeta(\gamma, \tilde{\gamma}) = \max_l 9\bar{M}_l N_n |\gamma_l - \tilde{\gamma}_l|_\infty/(b_l - a_l)$.

Lemma 3.3. *Suppose Assumptions* 2.1–2.3 *hold, let n be a positive integer, and let* $0 < \beta < 1$. *Then, for each j, there are versions of the random variables $W_n(B_{\gamma j})$, $\gamma \in \Gamma$, such that $W_n(B_{\gamma j})$ is Hölder continuous in γ with index β.*

Proof. Write $W_n(B_{\gamma j})$ as $W(\gamma)$. Let $\gamma, \gamma' \in \tilde{\Gamma}$. According to SH (2002, Lemma A.1), $\|B_\gamma - B_{\gamma'}\|_\infty \leqslant L\zeta(\gamma, \gamma')$ and hence $\|B_\gamma - B_{\gamma'}\|_n^2 \leqslant c_1\zeta^2(\gamma, \gamma')$, where $c_1 = L^2 M_2 \tau$. Thus, by (3.9), for $a > 0$,

$$P(|W(\gamma) - W(\gamma')| \geqslant a) \leqslant 2\exp\left(-\frac{na^2}{2c_1\zeta^2(\gamma, \gamma')}\right).$$

Let k be a positive integer. Suppose that $\zeta(\gamma, \gamma') \leqslant 2^{-(k-3)}$. Then

$$P(|W(\gamma) - W(\gamma')| > a2^{-\beta(k-3)}) \leqslant 2\exp\left(-\frac{na^2 2^{2(1-\beta)(k-3)}}{2c_1}\right).$$

For $k \geqslant 1$, let Ξ_k be a subset of $\tilde{\Gamma}$ such that $\#(\Xi_k) \leqslant \exp(c_2 k N_n)$ and every point in Γ is within $2^{-(k-1)}$ of some point in Ξ_k (in the metric ζ); here c_2 is a sufficiently large positive constant. Such subsets exist according to SH (2002, Lemma 3.1). Let $A_k^{(1)}$ denote the event that $|W(\gamma_k) - W(\tilde{\gamma}_k)| > a2^{-\beta(k-3)}$ for some $\gamma_k, \tilde{\gamma}_k \in \Xi_k$ such that $\zeta(\gamma_k, \tilde{\gamma}_k) \leqslant 2^{-(k-3)}$, and let $A_k^{(2)}$ denote the event that $|W(\gamma_{k+1}) - W(\gamma_k)| > a2^{-\beta(k-3)}$ for some $\gamma_k \in \Xi_k$ and $\gamma_{k+1} \in \Xi_{k+1}$ such that $\zeta(\gamma_k, \gamma_{k+1}) \leqslant 2^{-(k-2)}$. Then

$$P(A_k^{(1)}) \leqslant 2\exp(2c_2 k N_n)\exp\left(-\frac{na^2 2^{2(1-\beta)(k-3)}}{2c_1}\right)$$

and

$$P(A_k^{(2)}) \leqslant 2\exp(2c_2 (k+1) N_n)\exp\left(-\frac{na^2 2^{2(1-\beta)(k-3)}}{2c_1}\right).$$

Consequently (for fixed n),

$$\sum_k P(A_k^{(1)} \cup A_k^{(2)}) < \infty.$$

Set $\Omega_K = (A_K^{(1)} \cup A_K^{(2)} \cup A_{K+1}^{(1)} \cup A_{K+1}^{(2)} \cup \cdots)^c$. Then $\Omega_1 \subset \Omega_2 \subset \cdots$ and $\lim_K P(\Omega_K) = 1$. Consequently, $P(\Omega) = 1$, where $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots$.

Given $\gamma, \tilde{\gamma} \in \Gamma$ and $k \geqslant 1$, let $\gamma_k, \tilde{\gamma}_k \in \Xi_k$ be such that $\zeta(\gamma, \gamma_k) \leqslant 2^{-(k-1)}$ and $\zeta(\tilde{\gamma}, \tilde{\gamma}_k) \leqslant 2^{-(k-1)}$. Then $\zeta(\gamma_k, \tilde{\gamma}_k) \leqslant \zeta(\gamma, \tilde{\gamma}) + 2^{-(k-2)}$, $\zeta(\gamma_k, \gamma_{k+1}) \leqslant 2^{-(k-2)}$, and $\zeta(\tilde{\gamma}_k, \tilde{\gamma}_{k+1}) \leqslant 2^{-(k-2)}$. Observe that $\lim_k \gamma_k = \gamma$ and hence that $\lim_k \|B_{\gamma j} - B_{\gamma_k j}\|_\infty = 0$. Consequently, $W(\gamma_k)$ converges to $W(\gamma)$ in probability as $k \to \infty$. Since $\lim_k W(\gamma_k)$ exists on Ω, we conclude that $\lim_k W(\gamma_k) = W(\gamma)$ with probability one. Similarly, $\lim_k W(\tilde{\gamma}_k) = W(\tilde{\gamma})$ with probability one. Thus, for $K \geqslant 1$,

$$W(\gamma) - W(\tilde{\gamma}) = W(\gamma_K) - W(\tilde{\gamma}_K) + \sum_{k=K}^{\infty} \{W(\gamma_{k+1}) - W(\gamma_k) - [W(\tilde{\gamma}_{k+1}) - W(\tilde{\gamma}_k)]\} \tag{3.10}$$

with probability one. By deleting an event of probability zero and redefining each $W(\gamma)$ on an event of probability zero, we can assume that (3.10) holds on Ω for $\gamma, \tilde{\gamma} \in \Gamma$ and $K \geqslant 1$.

Suppose the event Ω occurs. Choose $\gamma, \tilde{\gamma} \in \Gamma$ such that $\zeta(\gamma, \tilde{\gamma}) \leqslant 2$. Then there exists $k \geqslant 1$ such that $2^{-(K-1)} < \zeta(\gamma, \tilde{\gamma}) \leqslant 2^{-(K-2)}$. Thus $\zeta(\gamma_K, \tilde{\gamma}_K) \leqslant 2^{-(K-3)}$. Consequently, $|W(\gamma_K) - W(\tilde{\gamma}_K)| \leqslant a2^{-\beta(K-3)}$; moreover, $|W(\gamma_{k+1}) - W(\gamma_k)| \leqslant a2^{-\beta(K-3)}$ and $|W(\tilde{\gamma}_{k+1}) - W(\tilde{\gamma}_k)| \leqslant a2^{-\beta(K-3)}$ for $k \geqslant K$. Thus, by (3.10),

$$|W(\gamma) - W(\tilde{\gamma})| \leqslant a2^{2\beta}\left(1 + \frac{2}{1-2^{-\beta}}\right)2^{-\beta(K-1)}$$

$$\leqslant a2^{2\beta}\left(1 + \frac{2}{1-2^{-\beta}}\right)[\zeta(\gamma, \tilde{\gamma})]^{\beta}.$$

We conclude that $W(\gamma)$, $\gamma \in \Gamma$, is Hölder continuous with index β on Ω_K for $K \geqslant 1$ and hence on all of Ω. □

Given $\gamma \in \tilde{\Gamma}$, set $\mathbb{B}_\gamma = \{g \in \mathbb{G}_\gamma \colon \|g\| \leqslant 1\}$. The following lemma will be used in the proof of Lemma 4.2 to check (3.5) when applying the chaining argument.

Lemma 3.4. *Suppose Assumptions* 2.1–2.3 *hold. Then, except on an event having probability zero, the following is true: Let* $\gamma \in \Gamma$, $\gamma_k \in \tilde{\Gamma}$ *for* $k \geqslant 1$, $\lim_{k\to\infty} \gamma_k = \gamma$, $g \in \mathbb{B}_\gamma$, $g_k \in \mathbb{B}_{\gamma_k}$ *for* $k \geqslant 1$, *and* $\lim_{k\to\infty} \|g_k - g\|_\infty = 0$. *Then*

$$\lim_{k\to\infty} \frac{\mathrm{d}}{\mathrm{d}\alpha}\, l(\bar{\eta}_{\gamma_k} + \alpha g_k)\bigg|_{\alpha=0} = \frac{\mathrm{d}}{\mathrm{d}\alpha}\, l(\bar{\eta}_\gamma + \alpha g)\bigg|_{\alpha=0}.$$

Proof. Now

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\, l(\bar{\eta}_\gamma + \alpha g)\bigg|_{\alpha=0} = W_n(g) + \int Z(t)\,\frac{g(t,\boldsymbol{X}(t))[\eta(t,\boldsymbol{X}(t)) - \bar{\eta}_\gamma(t,\boldsymbol{X}(t))]}{\sigma^2(t)}\,\mathrm{d}t.$$

Let γ, γ_k, g, and g_k be as in the statement of the lemma. According to SH (2002, Lemma 3.5), $\lim_{k\to\infty} \|\bar{\eta}_{\gamma_k} - \bar{\eta}_\gamma\|_\infty = 0$. Consequently,

$$\lim_{k\to\infty} \int Z(t)\,\frac{g_k(t,\boldsymbol{X}(t))[\eta(t,\boldsymbol{X}(t)) - \bar{\eta}_{\gamma_k}(t,\boldsymbol{X}(t))]}{\sigma^2(t)}\,\mathrm{d}t$$
$$= \int Z(t)\,\frac{g(t,\boldsymbol{X}(t))[\eta(t,\boldsymbol{X}(t)) - \bar{\eta}_\gamma(t,\boldsymbol{X}(t))]}{\sigma^2(t)}\,\mathrm{d}t.$$

Next, we apply Lemma 3.3 with Γ replaced by $\tilde{\Gamma}$. Let Ω be an event such that $P(\Omega)=1$ and, on Ω, $W_n(B_{\gamma j})$ is continuous in γ over $\tilde{\Gamma}$ for each j. Write $g = \sum_j b_{\gamma j} B_{\gamma j}$ and $g_k = \sum_j b_{\gamma_k j} B_{\gamma_k j}$. Then $W_n(g) = \sum_j b_{\gamma j} W_n(B_{\gamma j})$ and $W_n(g_k) = \sum_j b_{\gamma_k j} W_n(B_{\gamma_k j})$. Thus, to prove the desired result, it suffices to show that $\lim_{k\to\infty} b_{\gamma_k j} = b_{\gamma j}$ for each j.

Consider the inner product given by $\langle h_1, h_2\rangle_\psi = \int_{\mathscr{U}} h_1 h_2\,\mathrm{d}\psi$. Let $\boldsymbol{M}_\gamma$ denote the Gram matrix of the basis functions $B_{\gamma j}$ relative to this inner product, the entry in row j and column k of which equals $\langle B_{\gamma j}, B_{\gamma k}\rangle_\psi$; let $\boldsymbol{A}_\gamma(g)$ denote the column vector the jth entry of which equals $\int B_{\gamma j} g\,\mathrm{d}\psi$; and let $\boldsymbol{b}_\gamma$ denote the column vector the jth entry of which equals $b_{\gamma j}$. Then $\boldsymbol{b}_\gamma = \boldsymbol{M}_\gamma^{-1}\boldsymbol{A}_\gamma(g)$ and $\boldsymbol{b}_{\gamma_k} = \boldsymbol{M}_{\gamma_k}^{-1}\boldsymbol{A}_{\gamma_k}(g_k)$. Now $\lim_{k\to\infty} \boldsymbol{A}_{\gamma_k}(g_k) = \boldsymbol{A}_\gamma(g)$ and $\lim_{k\to\infty} \boldsymbol{M}_{\gamma_k} = \boldsymbol{M}_\gamma$, so $\lim_{k\to\infty} \boldsymbol{b}_{\gamma_k} = \boldsymbol{b}_\gamma$ as desired. □

4. Proofs of the main results

4.1. Proof of Theorem 2.1

The theorem follows by applying Huang (2001, Theorems A.1 and A.2). Note that Conditions A.2 and A.4(ii) in that paper are implied, respectively, by (3.3) and (3.4) in the present paper. Condition A.4(i) in that paper is a simplified version of Lemma 4.2 in the present paper. Its proof is simpler than that of Lemma 4.2 (no need for a chaining argument) and is thus omitted.

4.2. Proof of Theorem 2.2

The theorem follows from SH (2002, Lemma 2.1) and the following lemma.

Lemma 4.1. *Suppose Assumptions* 2.1–2.3 *hold. Then, for any constant* $0 < c < \infty$, *the set* $\{(\gamma, g) : \gamma \in \Gamma, g \in \mathbb{G}_\gamma,$ *and* $\|g\|_\infty \leqslant c\}$ *is compact and* $\ell(\cdot)$ *is almost surely continuous on this set.*

Proof. The first conclusion of the lemma follows from DeVore and Lorentz (1993, Chapter 5, Lemmas 2.1 and 4.1). It follows from Assumption 2.2 that

$$\int Z(t)\,\frac{g^2(t, \boldsymbol{X}(t))}{\sigma^2(t)}\,\mathrm{d}t$$

is continuous in g. The second conclusion of the lemma now follows from Lemma 3.3 by arguing as in the proof of Lemma 3.4. □

4.3. Proof of Theorem 4.3(i)

Clearly, $\bar{\eta}_\gamma$ is the orthogonal projection of η onto $\mathbb{G}_\gamma$ corresponding to the theoretical inner product. The result follows by arguing as in Huang (2001, proof of Theorem A.1) or SH (2002, proof of Theorem 2.1). Note that SH (2002, Condition 2.2) is implied by (3.2) and (3.3).

4.4. Proof of Theorem 2.3(ii)

The result follows from SH (2002, Theorem 2.2) and the following lemma. Note that SH (2002, Condition 2.4(ii)) is implied by (3.4).

Lemma 4.2. *Suppose Assumptions* 2.1–2.3 *and Condition* 2.1 *hold. Then*

$$\sup_{\gamma\in\Gamma}\sup_{g\in\mathbb{G}_\gamma}\frac{|\mathrm{d}/\mathrm{d}\alpha\,\ell(\bar{\eta}_\gamma + \alpha g)|_{\alpha=0}|}{\|g\|} = \mathrm{O}_P\left(\left(\frac{N_n}{n}\right)^{1/2}\right).$$

Observe next that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}\ell(\bar{\eta}_\gamma + \alpha g)\Big|_{\alpha=0} &= \frac{1}{n}\sum_i \int Z_i(t)\,\frac{g(t, \boldsymbol{X}_i(t))}{\sigma_i^2(t)}[\mathrm{d}Y_i(t) - \bar{\eta}_\gamma(t, \boldsymbol{X}_i(t))\,\mathrm{d}t]\\ &= E_n\left(\int Z(t)\,\frac{g(t, \boldsymbol{X}(t))}{\sigma^2(t)}\,[\mathrm{d}Y(t) - \eta(t, \boldsymbol{X}(t))\,\mathrm{d}t]\right)\\ &\quad + E_n\left(\int Z(t)\,\frac{g(t, \boldsymbol{X}(t))[\eta(t, \boldsymbol{X}(t)) - \bar{\eta}_\gamma(t, \boldsymbol{X}(t))]}{\sigma^2(t)}\,\mathrm{d}t\right)\end{aligned}$$

for $\gamma \in \Gamma$ and $g_1, g_2 \in \mathbb{G}_\gamma$. We will use the chaining argument (Lemma 3.2) to show that the two terms on the above right side are of order $(N_n/n)^{1/2}$ uniformly over g and γ.

In the proof of Lemma 4.2 below, we will use repeatedly the result that, for some positive constant M_7,

$$\|g\|_\infty \leqslant M_7 N_n^{1/2}\|g\|, \quad g \in \tilde{\Gamma} \text{ and } g \in \mathbb{G}_\gamma; \tag{4.1}$$

see SH (2002, proof of Lemma 3.3).

Proof of Lemma 4.2. By Theorem 2.3(i) applied to $\tilde{\Gamma}$, $\sup_{\gamma\in\tilde{\Gamma}} \|\bar{\eta}_\gamma\|_\infty \leqslant M_0'$ for n sufficiently large.

Since $N_n^{1/2}\tilde{\rho}_n \lesssim N_n^{-(c-1/2)}$ for some $c > \frac{1}{2}$ by assumption, there is an $\varepsilon \in (0, \frac{1}{16})$ and there is a fixed positive number c_1 such that, for n sufficiently large,

$$N_n^{1/2}\tilde{\rho}_n \leqslant c_1 (N_n^{1/2})^{-(\log 4)/\log[(4\varepsilon^{1/2})^{-1}]}$$

and hence

$$\min(c_1^{-1} N_n^{1/2}\tilde{\rho}_n, N_n^{1/2}\varepsilon^{(k-1)/2}) \leqslant 4^{-(k-1)} \tag{4.2}$$

for $k \geqslant 1$. (If $N_n^{1/2}\varepsilon^{(k-1)/2} \geqslant 4^{-(k-1)}$, then $N_n^{1/2}\tilde{\rho}_n \leqslant c_1 4^{-(k-1)}$.)

Let Ω_n, $n \geqslant 1$, be events such that $\lim_n P(\Omega_n) = 1$ and SH (2002, statement of Lemma 3.4) holds.

Let k be a positive integer, and let $\gamma, \tilde{\gamma} \in \tilde{\Gamma}$ be such that $\zeta(\gamma, \tilde{\gamma}) \leqslant \varepsilon^{k-1}$. Then, by SH (2002, Lemma 3.5), $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\| \leqslant c_2\varepsilon^{(k-1)/2}$ and $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\|_\infty \leqslant c_2 N_n^{1/2}\varepsilon^{(k-1)/2}$ (for some fixed positive constants c_2). Let $\mathbb{B}_{\tilde{\gamma},k}$ be as in SH (2002, Section 3) and let $g \in \mathbb{B}_\gamma$. Then, by Lemmas 3.3 and 3.4, as in the proof of SH (2002, Lemma 4.2), there is a $\tilde{g} \in \mathbb{B}_{\tilde{\gamma},k-1}$ such that $\|g - \tilde{g}\| \leqslant c_3\varepsilon^{k-1}$, $\|g - \tilde{g}\|_n \leqslant c_3\varepsilon^{k-1}$ on Ω_n, and $\|g - \tilde{g}\|_\infty \leqslant c_3 N_n^{1/2}\varepsilon^{k-1}$. Now

$$g(\eta - \bar{\eta}_\gamma) - \tilde{g}(\eta - \bar{\eta}_{\tilde{\gamma}}) = (g - \tilde{g})(\eta - \bar{\eta}_\gamma) + \tilde{g}(\bar{\eta}_{\tilde{\gamma}} - \bar{\eta}_\gamma).$$

Observe that $\|(g-\tilde{g})(\eta-\bar{\eta}_\gamma)\| \leqslant c_4\varepsilon^{k-1}$ and $\|(g-\tilde{g})(\eta-\bar{\eta}_\gamma)\|_\infty \leqslant c_4 N_n^{1/2}\varepsilon^{k-1}$. It follows from (4.1) that $\|\tilde{g}(\bar{\eta}_{\tilde{\gamma}} - \bar{\eta}_\gamma)\| \leqslant c_5 N_n^{1/2}\varepsilon^{(k-1)/2}$ and $\|\tilde{g}(\bar{\eta}_{\tilde{\gamma}} - \bar{\eta}_\gamma)\|_\infty \leqslant c_5 N_n\varepsilon^{(k-1)/2}$. Consequently, $\|g(\eta-\bar{\eta}_\gamma) - \tilde{g}(\eta-\bar{\eta}_{\tilde{\gamma}})\| \leqslant c_6 N_n^{1/2}\varepsilon^{(k-1)/2}$ and $\|g(\eta-\bar{\eta}_\gamma) - \tilde{g}(\eta-\bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant c_6 N_n\varepsilon^{(k-1)/2}$. Moreover, c_6 can be chosen so that, in addition, $\|g-\tilde{g}\| \leqslant c_6 N_n^{1/2}\varepsilon^{(k-1)/2}$ and $\|g-\tilde{g}\|_\infty \leqslant c_6 N_n\varepsilon^{(k-1)/2}$.

Alternatively, by Theorem 2.3(i), $\|\bar{\eta}_\gamma - \eta\| \leqslant c_7\tilde{\rho}_n$ for $\gamma \in \tilde{\Gamma}$. Given $\gamma, \tilde{\gamma} \in \tilde{\Gamma}$, we have that $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\| \leqslant 2c_7\tilde{\rho}_n$. Choose $\eta'_\gamma \in \mathbb{G}_\gamma$ and $\eta'_{\tilde{\gamma}} \in \mathbb{G}_{\tilde{\gamma}}$ such that $\|\eta'_\gamma - \eta\|_\infty \leqslant \tilde{\rho}_n$ and $\|\eta'_{\tilde{\gamma}} - \eta\|_\infty \leqslant \tilde{\rho}_n$. It follows from the triangle inequality and (4.1) that $\|\eta'_\gamma - \bar{\eta}_\gamma\|_\infty \leqslant c_8 N_n^{1/2}\tilde{\rho}_n$. Thus $\|\bar{\eta}_\gamma - \eta\|_\infty \leqslant (c_8 N_n^{1/2} + 1)\tilde{\rho}_n$. Similarly, $\|\bar{\eta}_{\tilde{\gamma}} - \eta\|_\infty \leqslant (c_8 N_n^{1/2} + 1)\tilde{\rho}_n$. Hence $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\|_\infty (c_8 N_n^{1/2} + 1)\tilde{\rho}_n$. Consequently, $\|g(\eta - \bar{\eta}_\gamma) - \tilde{g}(\eta - \bar{\eta}_{\tilde{\gamma}})\| \leqslant c_9 N_n^{1/2}\tilde{\rho}_n$ and $\|g(\eta - \bar{\eta}_\gamma) - \tilde{g}(\eta - \bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant c_9 N_n\tilde{\rho}_n$.

Let $\zeta(\gamma, \tilde{\gamma}) \leqslant \varepsilon^{k-1}$ and let $g \in \mathbb{B}_\gamma$ and $\tilde{g} \in \mathbb{B}_{\tilde{\gamma},k-1}$ be as above. It follows from (4.2) that $\|g(\eta - \bar{\eta}_\gamma) - \tilde{g}(\eta - \bar{\eta}_{\tilde{\gamma}})\| \leqslant c_{10}4^{-(k-1)}$, $\|g(\eta - \bar{\eta}_\gamma) - \tilde{g}(\eta - \bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant c_{10}N_n^{1/2}4^{-(k-1)}$, $\|g - \tilde{g}\| \leqslant c_{10}4^{-(k-1)}$, $\|g - \tilde{g}\|_n \leqslant c_{10}4^{-(k-1)}$ on Ω_n, and $\|g - \tilde{g}\|_\infty \leqslant c_{10}N_n^{1/2}4^{-(k-1)}$.

Let Ξ_k, $\mathbb{B}_{\gamma k}$ for $\gamma \in \tilde{\Gamma}$, and $\mathbb{B}_k$, $k \geqslant 0$, be as in SH (2002, Lemma 3.1 and the following paragraph) with the current value of ε. We will apply Lemma 3.4 of the present paper with $s = (\gamma, g)$,

$$V_s = \frac{\mathrm{d}}{\mathrm{d}\alpha}\,\ell(\bar{\eta}_\gamma + \alpha g)\bigg|_{\alpha=0},$$

$\mathbb{S} = \{(\gamma, g) : \gamma \in \Gamma \text{ and } g \in \mathbb{B}_\gamma\}$, and $\mathbb{S}_k = \{(\gamma, g) \colon \gamma \in \Xi_k \text{ and } g \in \mathbb{B}_{\gamma k}\}$. Now (3.5) follows from Lemma 3.4. Moreover, $\#(\mathbb{S}_k) \leqslant (M'\varepsilon^{-2k})^{N_n}$ for $k \geqslant 1$ by SH (2002, Equation (3.3)), so (3.6) holds with $C_1 = 1$ and any $C_2 \geqslant 2\log(M'\varepsilon^{-1})N_n$.

Given $\gamma \in \Xi_k$ and $g \in \mathbb{B}_{\gamma k}$, choose $\tilde{\gamma} \in \Xi_{k-1}$ and $\tilde{g} \in \mathbb{B}_{\tilde{\gamma}, k-1}$ such that $\zeta(\gamma, \tilde{\gamma}) \leqslant \varepsilon^{k-1}$, $\|g(\eta - \bar{\eta}_\gamma) - \tilde{g}(\eta - \bar{\eta}_{\tilde{\gamma}})\| \leqslant c_{10}4^{-(k-1)}$, $\|g(\eta - \bar{\eta}_\gamma) - \tilde{g}(\eta - \bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant c_{10}N_n^{1/2}4^{-(k-1)}$, $\|g - \tilde{g}\| \leqslant c_{10}4^{-(k-1)}$, $\|g - \tilde{g}\|_n \leqslant c_{10}4^{-(k-1)}$ on Ω_n, and $\|g - \tilde{g}\|_\infty \leqslant c_{10}N_n^{1/2}4^{-(k-1)}$. Write $s = (\gamma, g)$ and $V_s = V_{1s} + V_{2s}$, where

$$V_{1s} = E_n\left(\int Z(t)\,\frac{g(t, \boldsymbol{X}(t))[\eta(t, \boldsymbol{X}(t)) - \bar{\eta}_\gamma(t, \boldsymbol{X}(t))]}{\sigma^2(t)}\,\mathrm{d}t\right)$$

and

$$V_{2s} = E_n\left(\int Z(t)\,\frac{g(t, \boldsymbol{X}(t))[dY(t) - \eta(t, \boldsymbol{X}(t))\,\mathrm{d}t]}{\sigma^2(t)}\right).$$

Similarly, write $\tilde{s} = (\tilde{\gamma}, \tilde{g})$ and $V_{\tilde{s}} = V_{1\tilde{s}} + V_{2\tilde{s}}$. Observe that

$$\begin{aligned} V_{1s} - V_{1\tilde{s}} &= E_n\left(\int Z(t)\,\frac{h(t, \boldsymbol{X}(t))}{\sigma^2(t)}\,\mathrm{d}t\right) \\ &= (E_n - E)\left(\int Z(t)\,\frac{h(t, \boldsymbol{X}(t))}{\sigma^2(t)}\,\mathrm{d}t\right), \end{aligned}$$

where $h = g(\eta - \bar{\eta}_\gamma) - \tilde{g}(\eta - \bar{\eta}_{\tilde{\gamma}})$. It follows from Bernstein's inequality (see Hoeffding, 1963, Equation 2.13) that, for $C > 0$,

$$P(|V_{1s} - V_{1\tilde{s}}| \geqslant C2^{-(k-1)}(N_n/n)^{1/2}) \leqslant 2\exp\left(-\frac{C^2 2^{k-1} N_n}{2M_2\tau c_{10}(c_{10} + Cn^{-1/2}N_n)}\right).$$

Similarly,

$$V_{2s} - V_{2\tilde{s}} = E_n\left(\int Z(t)\,\frac{[g(t, \boldsymbol{X}(t)) - \tilde{g}(t, \boldsymbol{X}(t))][\mathrm{d}Y(t) - \eta(t, \boldsymbol{X}(t))\,\mathrm{d}t]}{\sigma^2(t)}\right),$$

so it follows from (3.9) that

$$P(|V_{2s} - V_{2\tilde{s}}| \geqslant C2^{-(k-1)}(N_n/n)^{1/2}; \Omega_n) \leqslant 2\exp\left(-\frac{C^2 2^{2(k-1)} N_n}{2c_{10}^2}\right).$$

Hence

$$P(|V_s - V_{\tilde{s}}| \geqslant 2C2^{-(k-1)}(N_n/n)^{1/2}; \Omega_n) \leqslant 4\exp\left(-\frac{2C^2 2^{k-1} N_n}{c_{11}(c_{10} + Cn^{-1/2}N_n)}\right),$$

so (3.7) holds with

$$C_2 = \frac{C^2 N_n}{c_{11}(c_{10} + Cn^{-1/2}N_n)} \geqslant 2\log(M'\varepsilon^{-1})N_n$$

for C sufficiently large, $C_3 = 2C(N_n/n)^{1/2}$, and $C_4 = 4$. Consequently, by Lemma 3.2,

$$P\left(\sup_{\gamma\in\Gamma}\sup_{g\in\mathbb{B}_\gamma}\left|\frac{\mathrm{d}}{\mathrm{d}\alpha}\ell(\bar\eta_\gamma + \alpha g)\Big|_{\alpha=0}\right| \geqslant 4C(N_n/n)^{1/2}\right)$$

$$\leqslant \frac{4c_{11}(c_{10} + Cn^{-1/2}N_n)}{C^2 N_n} + P(\Omega_n^c),$$

which can be made arbitrarily close to zero by making n and C sufficiently large. □

4.5. *Proof of Theorem 2.3(iii)*

The result follows from SH (2002, Theorem 2.3) and the following lemma.

Lemma 4.3. *Suppose Assumptions* 2.1–2.3 *and Condition* 2.1 *hold. Then*

$$(i)\ |\ell(\bar\eta_{\gamma^*}) - \ell(\eta) - [\Lambda(\bar\eta_{\gamma^*}) - \Lambda(\eta)]| = \mathrm{O}_P\left(\inf_{\gamma\in\Gamma}\rho_{n\gamma}^2 + \frac{N_n}{n}\right)$$

and

$$(ii)\ |\ell(\bar\eta_\gamma) - \ell(\eta) - [\Lambda(\bar\eta_\gamma) - \Lambda(\eta)]| = \bar{\mathrm{O}}_P\left((\log^{1/2} n)\left[\|\bar\eta_\gamma - \eta\|\left(\frac{N_n}{n}\right)^{1/2} + \frac{N_n}{n}\right]\right)$$

uniformly over $\gamma\in\Gamma$.

Proof. The first conclusion of the lemma follows by the argument used in SH (2002) to verify the first part of Condition 2.6 of that paper. To verify the second conclusion we observe that

$$\ell(\bar\eta_\gamma) - \ell(\eta) - [\Lambda(\bar\eta_\gamma) - \Lambda(\eta)]$$

$$= E_n\left(\int Z(t)\frac{[\bar\eta_\gamma(t,\boldsymbol{X}(t)) - \eta(t,\boldsymbol{X}(t))][\mathrm{d}Y(t) - \eta(t,\boldsymbol{X}(t))\,\mathrm{d}t]}{\sigma^2(t)}\right)$$

$$-\frac{1}{2}(E_n - E)\left(\int Z(t)\frac{[\bar\eta_\gamma(t,\boldsymbol{X}(t)) - \eta(t,\boldsymbol{X}(t))]^2}{\sigma^2(t)}\,\mathrm{d}t\right). \tag{4.3}$$

We claim that, uniformly over $\gamma\in\Gamma$,

$$(E_n - E)\left(\int Z(t)\frac{[\bar\eta_\gamma(t,\boldsymbol{X}(t)) - \eta(t,\boldsymbol{X}(t))]^2}{\sigma^2(t)}\,\mathrm{d}t\right)$$

$$= \bar{\mathrm{O}}_P((\log^{1/2} n)[\|\bar\eta_\gamma - \eta\|(N_n/n)^{1/2} + N_n/n]) \tag{4.4}$$

and

$$E_n\left(\int Z(t)\,\frac{[\bar\eta_\gamma(t,\boldsymbol{X}(t))-\eta(t,\boldsymbol{X}(t))][\mathrm{d}Y(t)-\eta(t,\boldsymbol{X}(t))\,\mathrm{d}t]}{\sigma^2(t)}\right)$$
$$=\bar{\mathrm{O}}_P((\log^{1/2} n)[\|\bar\eta_\gamma-\eta\|(N_n/n)^{1/2}+N_n/n]). \tag{4.5}$$

The desired result follows from (4.3)–(4.5). The proof of (4.4) is similar to the verification in SH (2002) of Condition 2.6(ii) in the density estimation context.

To verify (4.5) we observe first of all that, by Bernstein's inequality, for $a>0$,

$$P(\|h\|_n^2-\|h\|^2\geqslant a)\leqslant \exp\left(-\frac{na^2}{2M_2\tau\|h\|_\infty^2(\|h\|^2+a)}\right). \tag{4.6}$$

We will apply (3.9) and (4.6) first to $h=\bar\eta_\gamma-\eta$. According to arguments in SH (2002, Section 4.2), $\|\bar\eta_\gamma-\eta\|\lesssim\tilde\rho_n$ and $\|\bar\eta_\gamma-\eta\|_\infty\lesssim\log^{-1/2}n$ uniformly over $\gamma\in\tilde\Gamma$. Let c_1 be a fixed positive number. Then, by (4.6), for c_2 a sufficiently large positive number,

$$P(\|\bar\eta_\gamma-\eta\|_n^2-\|\bar\eta_\gamma-\eta\|^2\geqslant c_2^2[\|\bar\eta_\gamma-\eta\|(N_n/n)^{1/2}+N_n/n])$$
$$\leqslant \exp(-2c_1N_n\log n),\qquad \gamma\in\tilde\Gamma.$$

Thus, for c_2 a sufficiently large positive number,

$$P(\|\bar\eta_\gamma-\eta\|_n\geqslant c_2[\|\bar\eta_\gamma-\eta\|+(N_n/n)^{1/2}])\leqslant\exp(-2c_1N_n\log n),\qquad\gamma\in\tilde\Gamma.$$

According to (3.9), for c_3 a sufficiently large positive number,

$$P\left(\left|E_n\left(\int Z(t)\,\frac{[\bar\eta_\gamma(t,\boldsymbol{X}(t))-\eta(t,\boldsymbol{X}(t))][\mathrm{d}Y(t)-\eta(t,\boldsymbol{X}(t))\,\mathrm{d}t]}{\sigma^2(t)}\right)\right|\right.$$
$$\geqslant c_3(\log^{1/2}n)[\|\bar\eta_\gamma-\eta\|(N_n/n)^{1/2}+N_n/n]\quad\text{and}$$
$$\left.\|\bar\eta_\gamma-\eta\|_n\leqslant c_2[\|\bar\eta_\gamma-\eta\|+(N_n/n)^{1/2}]\right)$$
$$\leqslant 2\exp(-2c_1N_n\log n)$$

for $\gamma\in\tilde\Gamma$. Consequently,

$$P\left(\left|E_n\left(\int Z(t)\,\frac{[\bar\eta_\gamma(t,\boldsymbol{X}(t))-\eta(t,\boldsymbol{X}(t))][\mathrm{d}Y(t)-\eta(t,\boldsymbol{X}(t))\,\mathrm{d}t]}{\sigma^2(t)}\right)\right|\right.$$
$$\left.\geqslant c_3(\log^{1/2}n)[\|\bar\eta_\gamma-\eta\|(N_n/n)^{1/2}+N_n/n]\right)$$
$$\leqslant 3\exp(-2c_1N_n\log n),\qquad\gamma\in\tilde\Gamma. \tag{4.7}$$

Let $\varepsilon>0$, let k be a positive integer, and let $\gamma,\tilde\gamma\in\tilde\Gamma$ be such that $\zeta(\gamma,\tilde\gamma)\leqslant c_4\varepsilon^k$. Then, as in the proof of Lemma 4.2, $\|\bar\eta_\gamma-\bar\eta_{\tilde\gamma}\|\leqslant c_5\varepsilon^{k/2}$, $\|\bar\eta_\gamma-\bar\eta_{\tilde\gamma}\|_n\leqslant c_5\varepsilon^{k/2}$ on Ω_n, and $\|\bar\eta_\gamma - \bar\eta_{\tilde\gamma}\|_\infty \leqslant c_5 N_n^{1/2}\varepsilon^{k/2}$. It follows from (3.9) that

$$P\left(\left|E_n\left(\int Z(t)\,\frac{[\bar\eta_\gamma(t,\boldsymbol{X}(t)) - \bar\eta_{\tilde\gamma}(t,\boldsymbol{X}(t))][\mathrm{d}Y(t) - \eta(t,\boldsymbol{X}(t))\,\mathrm{d}t]}{\sigma^2(t)}\right)\right| \geqslant c_6\varepsilon^{k/4};\Omega_n\right)$$
$$\leqslant 2\exp\left(-\frac{nc_6^2\varepsilon^{-k/2}}{2c_5^2}\right). \tag{4.8}$$

By Theorem 2.3(i) of this paper and SH (2002, Lemma 3.5), $\|\bar\eta_\gamma - \bar\eta_{\tilde\gamma}\| \leqslant c_7[\zeta(\gamma,\tilde\gamma)]^{1/2}$ and $\|\bar\eta_\gamma - \bar\eta_{\tilde\gamma}\|_\infty \leqslant c_7 N_n^{1/2}[\zeta(\gamma,\tilde\gamma)]^{1/2}$ for $\gamma,\tilde\gamma \in \tilde\Gamma$. Let $0 < \varepsilon \leqslant \frac{1}{2}$. Then, as in SH (2002), there exist subsets Ξ_k of $\tilde\Gamma$ for $k \geqslant 1$ such that

$$\#(\Xi_k) \leqslant \exp(c_8 k N_n), \quad k \geqslant 1, \tag{4.9}$$

and, for $k \geqslant 1$, every point in Γ is within ε^k (in ζ distance) of some point in Ξ_k. Let σ_k denote a function from Γ to Ξ_k such that $\sigma_k(\gamma)$ is within ε^k of γ for $\gamma \in \Gamma$. Then $\sigma_k(\gamma)$ is within $\varepsilon^k + \varepsilon^{k+1}$ of $\sigma_{k+1}(\gamma)$.

Set

$$W_n(\gamma) = E_n\left(\int Z(t)\,\frac{[\bar\eta_\gamma(t,\boldsymbol{X}(t)) - \eta(t,\boldsymbol{X}(t))][\mathrm{d}Y(t) - \eta(t,\boldsymbol{X}(t))\,\mathrm{d}t]}{\sigma^2(t)}\right) \quad \gamma \in \tilde\Gamma.$$

It follows from Lemma 3.3 in this paper together with SH (2002, Lemma 3.5), by arguing as in the proof of Lemma 3.4 of this paper, that

$$P\left(\lim_{K\to\infty} W_n(\sigma_K(\gamma)) = W_n(\gamma) \text{ for } \gamma \in \Gamma\right) = 1. \tag{4.10}$$

We want to verify that, for some $b > 0$,

$$\lim_{n\to\infty} P(|W_n(\gamma)| > b(\log^{1/2} n)[\|\bar\eta_\gamma - \eta\|(N_n/n)^{1/2} + N_n/n] \text{ for some } \gamma \in \Gamma) = 0.$$

To this end, by (4.10) and Fatou's lemma, it suffices to show that, for some $b > 0$,

$$\lim_{n\to\infty} \limsup_{K\to\infty} P(|W_n(\sigma_K(\gamma))| > b(\log^{1/2} n)[\|\bar\eta_\gamma - \eta\|(N_n/n)^{1/2} + N_n/n]$$
$$\text{for some } \gamma \in \Gamma; \Omega_n) = 0. \tag{4.11}$$

Let $1 \leqslant K_0 < K$. Then

$$|W_n(\sigma_K(\gamma))| \leqslant |W_n(\sigma_{K_0}(\gamma))| + \sum_{k=K_0}^{K-1} |W_n(\sigma_{k+1}(\gamma)) - W_n(\sigma_k(\gamma))|. \tag{4.12}$$

By (4.7)

$$P(|W_n(\sigma_{K_0}(\gamma))| \geqslant c_3(\log^{1/2} n)[\|\bar\eta_{\sigma_{K_0}(\gamma)} - \eta\|(N_n/n)^{1/2} + N_n/n])$$
$$\leqslant 3\exp(-2c_1 N_n \log n)$$

for $\gamma \in \Gamma$. Suppose that $c_9 \log n \leqslant K_0 \leqslant c_{10} \log n$, where $c_8 c_{10} \leqslant c_1$. Then, by (4.9),

$$\lim_{n\to\infty} P(|W_n(\sigma_{K_0}(\gamma))| \geqslant c_3(\log^{1/2} n)[\|\bar{\eta}_{\sigma_{K_0}(\gamma)} - \eta\|(N_n/n)^{1/2} + N_n/n] \text{ for some } \gamma \in \Gamma) = 0.$$

Moreover, by SH (2002, Lemma 3.5),

$$\begin{aligned}\|\bar{\eta}_{\sigma_{K_0}(\gamma)} - \eta\| &\leqslant \|\bar{\eta}_\gamma - \eta\| + \|\bar{\eta}_\gamma - \bar{\eta}_{\sigma_{K_0}(\gamma)}\| \\ &\leqslant \|\bar{\eta}_\gamma - \eta\| + c_7 \varepsilon^{K_0/2} \\ &\leqslant \|\bar{\eta}_\gamma - \eta\| + c_7 n^{-c_9(\log \varepsilon^{-1})/2} \\ &\leqslant \|\bar{\eta}_\gamma - \eta\| + (N_n/n)^{1/2}\end{aligned}$$

provided that c_9 is sufficiently large. Consequently, by choosing c_3 larger if necessary, we conclude that

$$\lim_{n\to\infty} P(|W_n(\sigma_{K_0}(\gamma))| \geqslant c_3(\log^{1/2} n)[\|\bar{\eta}_\gamma - \eta\|(N_n/n)^{1/2} + N_n/n] \text{ for some } \gamma \in \Gamma) = 0. \quad (4.13)$$

It follows from (4.8) that

$$P(|W_n(\sigma_{k+1}(\gamma)) - W_n(\sigma_k(\gamma))| \geqslant c_6 \varepsilon^{k/4}; \Omega_n) \leqslant 2\exp(-nc_{11}\varepsilon^{-k/2})$$

for $K_0 \leqslant k \leqslant K-1$ and $\gamma \in \Gamma$. Consequently, by (4.9),

$$\begin{aligned}&P(|W_n(\sigma_{k+1}(\gamma)) - W_n(\sigma_k(\gamma))| \geqslant c_6 \varepsilon^{k/4} \text{ for some } \gamma \in \Gamma; \Omega_n) \\ &\quad \leqslant 2\exp(2c_8(k+1)N_n)\exp(-nc_{11}\varepsilon^{-k/2})\end{aligned}$$

for $K_0 \leqslant k \leqslant K-1$. Observe that

$$\sum_{k=K_0}^{K-1} c_6 \varepsilon^{k/4} \leqslant \frac{c_6 n^{-c_9(\log 1/\varepsilon)/4}}{1-\varepsilon^{1/4}} \leqslant \frac{1}{n}$$

for n sufficiently large provided that c_9 and ε are chosen appropriately. Observe also that

$$\exp(2c_8(k+1)N_n) \leqslant \exp(nc_{11}\varepsilon^{-k/2}/2), \quad k \geqslant c_9 \log n,$$

provided that n is sufficiently large, in which case

$$\begin{aligned}\sum_{k=K_0}^{K-1} 2\exp(2c_8(k+1)N_n)\exp(-nc_{11}\varepsilon^{-k/2}) &\leqslant 2\sum_{k=K_0}^{\infty} \exp(-nc_{11}\varepsilon^{-k/2}/2) \\ &\leqslant \frac{4}{nc_{11}e}\sum_{k=K_0}^{\infty} \varepsilon^{k/2} \\ &\leqslant \frac{4n^{-c_9\log(1/\varepsilon)/2}}{nc_{11}e(1-\varepsilon^{1/2})},\end{aligned}$$

which tends to 0 as $n \to \infty$. (Note that $e^{-x} \leqslant 1/(xe)$ for $x > 0$.) Therefore,

$$\lim_{n\to\infty} \limsup_{K\to\infty} P\left(\sum_{k=K_0}^{K-1} |W_n(\sigma_{k+1}(\gamma)) - W_n(\sigma_k(\gamma))| \geqslant \frac{1}{n} \text{ for some } \gamma \in \Gamma \right) = 0. \tag{4.14}$$

Equation (4.11) follows from (4.12)–(4.14). □

Acknowledgements

The first author's research was supported in part by NSF grant DMS-9802071.

Appendix A. Improved convergence rate of free knot spline estimate over fixed-knot spline estimate

In this Appendix, we give the necessary technical details for the example given at the end of Section 2.3. Suppose (X_i, Y_i), $i = 1, \ldots, n$, is an iid sample of (X, Y) which follows the regression model $Y = \eta(X) + \varepsilon$. Define the empirical and theoretical inner products as $\langle f_1, f_2 \rangle_n = (1/n) \sum_{i=1}^n f_1(X_i) f_2(X_i)$ and $\langle f_1, f_2 \rangle = E[f_1(X) f_2(X)]$. The induced norms are denoted as $\|\cdot\|_n$ and $\|\cdot\|$, respectively. Let $\tilde{\eta}$ and $\bar{\eta}$ denote the orthogonal projection of η on the estimation space $\mathbb{G}$ relative to the empirical and theoretical inner product respectively. Since $\tilde{\eta} - \eta$ is orthogonal to $\mathbb{G}$ relative to the empirical inner product,

$$\|\hat{\eta} - \eta\|_n^2 = \|\hat{\eta} - \tilde{\eta}\|_n^2 + \|\tilde{\eta} - \eta\|_n^2. \tag{A.1}$$

Set $b = 1/(2J_n - 1)$. We have that

$$\|\tilde{\eta} - \eta\|^2 \geqslant \|\bar{\eta} - \eta\|^2 \geqslant \frac{1}{2} \min_{a_0, a_1} \int_{-b}^{b} \{(a_0 + a_1 x) - |x|\}^2 \, dx = \frac{b^3}{24}.$$

Using the Bernstein inequality and noting that $\eta(x) = x^2 + |x|$, we can show that

$$\sup_{g \in \mathbb{G}} \left| \frac{\|g - \eta\|_n^2}{\|g - \eta\|^2} - 1 \right| = o_P(1). \tag{A.2}$$

(See Lemma 4 of Huang 1998a.) Hence, for some positive constant c,

$$\|\tilde{\eta} - \eta\|_n \geqslant c J_n^{-3}, \tag{A.3}$$

with probability tending to 1. On the other hand, $n\sigma^{-2} \|\hat{\eta} - \tilde{\eta}\|_n^2$ has a χ^2-distribution with J_n degrees of freedom. This implies that

$$\lim_n P(n \|\hat{\eta} - \tilde{\eta}\|_n^2 \geqslant \sigma^2 J_n) = \frac{1}{2}. \tag{A.4}$$

Combining (A.1), (A.3) and (A.4), we obtain that

$$\lim_n P\left(\|\hat{\eta}-\eta\|_n^2 \geqslant \sigma^2 \frac{J_n}{n} + cJ_n^{-3}\right) = \frac{1}{2}.$$

Hence, by (A.2),

$$\lim_n P\left(\|\hat{\eta}-\eta\|^2 \geqslant \sigma^2 \frac{J_n}{n} + cJ_n^{-3}\right) = \frac{1}{2}.$$

References

Beder, J.H., 1987. A sieve estimator for the mean of gaussian process. Ann. Statist. 15, 59–78.

DeVore, R.A., Lorentz, G.G., 1993. Constructive Approximation. Springer, Berlin.

Grenander, U., 1981. Abstract Inference. Wiley, New York.

Hansen, M., 1994. Extended linear models, multivariate splines, and ANOVA. Ph.D. Dissertation, Department of Statistics, University of California at Berkeley.

Hansen, M.H., Kooperberg, C., 2002. Spline adaption in extended linear models. With discussion. Statist. Sci. 17, 2–51.

Hoeffding, W., 1963. Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. 58, 13–30.

Huang, J.Z., 1998a. Projection estimation in multiple regression with application to functional ANOVA models. Ann. Statist. 26, 242–272.

Huang, J.Z., 1998b. Functional ANOVA models for generalized regression. J. Multivariate Anal. 67, 49–71.

Huang, J.Z., 2001. Concave extended linear modeling: a theoretical synthesis. Statist. Sinica 11, 173–197.

Huang, J.Z., Stone, C.J., 1998. The L_2 rate of convergence for event history regression with time-dependent covariates. Scand. J. Statist. 25, 603–620.

Huang, J.Z., Kooperberg, C., Stone, C.J., Truong, Y.K., 2000. Functional ANOVA modeling for proportional hazards regression. Ann. Statist. 28, 960–999.

Kooperberg, C., Stone, C.J., 2002a. Confidence intervals for logspline density estimation. In: Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, B. (Eds.), Nonlinear Estimation and Classification. Springer, New York, pp. 285–296.

Kooperberg, C., Stone, C.J., 2002b. Comparison of parametric, bootstrap, and Bayesian approaches to obtaining confidence intervals for logspline density estimation. Manuscript.

Kooperberg, C., Stone, C.J., Truong, Y.K., 1995a. The L_2 rate of convergence for hazard regression. Scand. J. Statist. 22, 143–157.

Kooperberg, C., Stone, C.J., Truong, Y.K., 1995b. Rate of convergence for logspline spectral density estimation. J. Time Ser. Anal. 16, 389–401.

Leskow, J., Rozanski, R., 1989. Maximum likelihood estimation of the drift function for a diffusion process. Statist. Decisions 7, 243–262.

Li, W., 2000. Modeling marked point processes with an application to currency exchange rates. Ph.D. Dissertation, Dept. Statist., Univ. California, Berkeley.

Lindstrom, M.J., 1999. Penalized estimation of free-knot splines. J. Comput. Graph. Statist. 8, 333–352.

McKeague, I.W., 1986. Estimation for a semimartingale regression model using the method of sieves. Ann. Statist. 14, 579–589.

Nguyen, H.T., Pham, D.T., 1982. Identification of nonstationary diffusion model by the method of sieves. SIAM J. Control Optimiz. 20, 603–611.

Pollard, D., 1984. Convergence of Stochastic Processes. Springer, New York.

Prakasa Rao, B.L.S., 1999a. Semimartingales and their Statistical Inference. Chapman & Hall/CRC, Boca Raton, FL.

Prakasa Rao, B.L.S., 1999b. Statistical Inference for Diffusion Type Processes. Arnold, London.

Schumaker, L.L., 1981. Spline Functions: Basic Theory. Wiley, New York.

Shorack, G.R., Wellner, J.A., 1986. Empirical Processes with Applications to Statistics. Wiley, New York.

Stone, C.J., 1985. Additive regression and other nonparametric models. Ann. Statist. 13, 689–705.
Stone, C.J., 1986. The dimensionality reduction principle for generalized additive models. Ann. Statist. 14, 590–606.
Stone, C.J., 1990. Large-sample inference for log-spline models. Ann. Statist. 18, 717–741.
Stone, C.J., 1991. Asymptotics for doubly flexible logspline response models. Ann. Statist. 19, 1832–1854.
Stone, C.J., 1994. The use of polynomial splines and their tensor products in multivariate function estimation. With discussion. Ann. Statist. 22, 118–184.
Stone, C.J., Huang, J.Z., 2002. Free knot splines in concave extended linear modeling. J. Statist. Plann. Inference 108, 219–253.
Stone, C.J., Hansen, M., Kooperberg, C., Truong, Y.K., 1997. Polynomial splines and their tensor products in extended linear modeling. with discussion. Ann. Statist. 25, 1371–1470.
van de Geer, S., 1995. Exponential inequalities for martingales with application to maximum likelihood estimation for counting processes. Ann. Statist. 23, 1779–1801.
Zhou, S., Shen, X., 2001. Spatially adaptive regression splines and accurate knot selection schemes. J. Amer. Statist. Assoc. 96, 247–259.