ELSEVIER

Journal of Statistical Planning and Inference 108 (2002) 219–253

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Free knot splines in concave extended linear modeling

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

Abstract

Many problems of practical interest can be formulated as the estimation of a certain function such as a regression function, logistic or other generalized regression function, density function, conditional density function, hazard function, or conditional hazard function. Extended linear modeling provides a convenient framework for using polynomial splines and their tensor products in such function estimation problems. Huang (Statist. Sinica 11 (2001) 173) has given a general treatment of the rates of convergence of maximum likelihood estimation in the context of concave extended linear modeling. Here these results are generalized to let the approximation space used in the fitting procedure depend on a vector of parameters. More detailed treatments are given for density estimation and generalized regression (including ordinary regression) on the one hand and for approximation spaces whose components are suitably regular free knot splines and their tensor products on the other hand.

MSC: primary 62G07; secondary 62J12

Keywords: Density estimation; Generalized regression; Free knot splines; Tensor products; Maximum likelihood; Method of sieves; Rates of convergence

1. Introduction

Extended linear modeling has been introduced in Hansen (1994), Stone et al. (1997), and Huang (2001) to synthesize the theory and methodology that uses polynomial splines and their tensor products to model functions of interest. One prominent feature of such an approach of functional modeling is the ease of incorporating functional

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

PII: S0378-3758(02)00280-X

analysis of variance (ANOVA) decompositions. Typically, only selected low-order terms in the ANOVA decomposition of the target function are considered and the selected terms are unspecified except for the requirement of smoothness. Thus the desire for flexibility in nonparametric modeling is balanced with the desire for interpretability and for taming the "curse of dimensionality". This approach provides a useful tool for a broad range of statistical problems; see Stone et al. (1997) and the references therein.

In extended linear modeling, the function of interest, such as a regression function, logistic regression function, density function, or conditional hazard function, is modeled as a member of a finite- or infinite-dimensional linear model space, and maximum "likelihood" estimation over a finite-dimensional linear approximation subspace is used to fit the data. Here, the likelihood could be a true likelihood, pseudo-likelihood, conditional likelihood, or partial likelihood, depending on the problem under consideration, and letting the unknown function have specified terms in its ANOVA decomposition corresponds to choosing an appropriate model space.

Theoretical properties of maximum likelihood estimation in extended linear modeling have been obtained in a number of statistical contexts, including regression in Stone (1985, 1994) and Huang (1998a); generalized regression in Stone (1986, 1994) and Huang (1998b), density estimation in Stone (1990, 1994); conditional density estimation in Stone (1991, 1994) and Hansen (1994); hazard regression in Kooperberg et al. (1995a); spectral density estimation in Kooperberg et al. (1995b); event history analysis in Huang and Stone (1998); and proportional hazards regression in Huang et al. (2000). Hansen (1994) synthesized the theory as it existed at the time.

Recently, Huang (2001) developed a unified theory to deal simultaneously with the various statistical contexts. In this theory, the overall error of estimation is decomposed into two parts—a stochastic part (estimation error) and a systematic part (approximation error). Let N_n denote the dimension of the linear approximation space $\mathbb{G}$, which is finite and positive and may tend to infinity with the sample size n, and let ρ_n denote the L_∞ approximation rate corresponding to $\mathbb{G}$ (that is, the minimum L_∞ norm of the error when the target function is approximated by a function in $\mathbb{G}$). Then the L_2 norms of the estimation and approximation errors are bounded in probability by multiples of $\sqrt{N_n/n}$ and ρ_n, respectively. These results provide considerable insight: the error bound of the stochastic part can be obtained by a heuristic variance calculation, while that of the systematic part is reduced to a problem in approximation theory. Huang (2001) also considered estimation of the components in ANOVA decompositions of unknown functions and the issue of model misspecification. Moreover, he worked out the additional details for applying the general theory in the specific contexts of counting process regression and conditional density estimation.

The main purpose of the present paper is to generalize the theory developed in Huang (2001) by letting the approximation space depend on a vector of nonlinear parameters. Specifically, we consider a collection $\mathbb{G}_\gamma$, $\gamma \in \Gamma$, of linear estimation spaces having a common dimension that may vary with the sample size. For each fixed γ, the maximum likelihood estimate is obtained. We let the data pick which estimation space $\mathbb{G}_\gamma$ to use, again using the maximum likelihood. As an important application, γ can be thought as the knot positions when the estimation space consists of spline functions, and our interest lies in choosing the knot positions using the data.

Let N_n denote the common dimension of $\mathbb{G}_\gamma$, $\gamma \in \Gamma$, which may increase with the sample size n. Let $\rho_{n\gamma}$ denote the L_∞ approximation rate corresponding to $\mathbb{G}_\gamma$. We show in this paper that, under regularity conditions that are satisfied by suitably constructed polynomial spline spaces, the L_2 norms of the estimation and approximation errors corresponding to $\mathbb{G}_\gamma$ are bounded in probability, uniformly over $\gamma \in \Gamma$, by multiples of $\sqrt{N_n/n}$ and $\rho_{n\gamma}$ respectively. As an important consequence, the empirical selection of γ does not influence the magnitude of the estimation error. It is also shown that, when γ is appropriately selected by a data-driven method, the corresponding approximation rate is close to the best possible approximation rate $\inf_{\gamma\in\Gamma} \rho_{n\gamma}$. See Proposition 2.1 for precise statement of these results. Our results give additional insight into the various knot placement methodologies (free knot splines) that have been discussed in the literature; see Stone et al. (1997) and the references cited therein, Zhou and Shen (2001), and Lindstrom (1999).

There is considerable recent interest in developing a general rate of convergence theory of nonparametric estimation with the method of sieves using empirical process theory; see, for example, Barron et al. (1999) and the references therein. Since polynomial splines and their tensor products can be viewed as a sieve of particular type, the aims of this paper overlap with that literature, although none of the papers in that literature studied free knot splines explicitly. We should note that the formulation and treatment are different in the two approaches. This paper follows the line of development in Stone (1994) and Huang (2001). In our approach the variance-bias trade-off and the effect of parameter selection are seen more explicitly. One difference in the formulation of this paper and that in Barron et al. (1999) is that, for each γ, we consider maximum likelihood estimation over the entire linear space $\mathbb{G}_\gamma$, while they consider maximum likelihood over a subset of functions in $\mathbb{G}_\gamma$ having a common bound.

Section 2 of this paper contains the basic setup and the main results on rates of convergence. In Section 3, we discuss the various properties of spaces of free knot splines and tensor products of such spaces that are needed to verify the conditions in our general results on rates of convergence. In Section 4 we verify the conditions in the main results in Section 2 in the contexts of density estimation and generalized regression, including ordinary regression as a special case. There, for simplicity, we avoid the explicit consideration of ANOVA models and restrict attention to spaces $\mathbb{G}_\gamma$, $\gamma \in \Gamma$, that are tensor products of polynomial spline spaces. The Appendix contains the proofs of results in Section 3.

We conclude this section by introducing some notation. For a function f on $\mathscr{U}$, set $\|f\|_\infty = \sup_{\boldsymbol{u}\in\mathscr{U}} |f(\boldsymbol{u})|$. 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 W_n for $n \geqslant 1$, let $W_n = \mathrm{O_P}(b_n)$ mean that $\lim_{c\to\infty} \limsup_n P(|W_n| \geqslant cb_n) = 0$, and let $W_n = \mathrm{o_P}(b_n)$ mean that $\limsup_n P(|W_n| \geqslant cb_n) = 0$ for all $c > 0$. These notions can be extended to hold uniformly over $\gamma \in \Gamma$. In particular, we let $W_{n\gamma} = \mathrm{O_P}(b_{n\gamma})$ uniformly over $\gamma \in \Gamma$ mean that $\lim_{c\to\infty} \limsup_n P(|W_{n\gamma}| \geqslant cb_{n\gamma}$ for some $\gamma \in \Gamma)$=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 $M_1, M_2, \ldots$ to denote positive numbers that do not depend on n, $\gamma \in \Gamma$, or $g \in \mathbb{G}_\gamma$.

2. Main results

2.1. Basic setup and statement of main results

Consider a $\mathscr{W}$-valued random variable $\boldsymbol{W}$, where $\mathscr{W}$ is an arbitrary set. Let $\mathscr{U}$ be a compact subset of $\mathbb{R}^d$ for some positive integer d, where $\mathscr{U}$ may or may not coincide with $\mathscr{W}$. For a (real-valued) function h on $\mathscr{U}$, let $l(h, \boldsymbol{W})$ be a log-likelihood and let $\Lambda(h) = E[l(h, \boldsymbol{W})]$ be the corresponding expected log-likelihood. There may be some mild restrictions on h for the log-likelihood to be defined. We assume that, subject to such restrictions, there is an essentially unique function η on $\mathscr{U}$ that maximizes the expected log-likelihood.

Let $\mathbb{H}$ be a finite- or infinite-dimensional linear space of functions on $\mathscr{U}$. We say that the space $\mathbb{H}$ and the log-likelihood function $l(h, \boldsymbol{W})$, $h \in \mathbb{H}$, together define an extended linear model. Suppose the set of functions in $\mathbb{H}$ whose log-likelihood and expected log-likelihood are well-defined is convex. The extended linear model is said to be concave if $l(h, \boldsymbol{w})$ is a concave function of h for each $\boldsymbol{w} \in \mathscr{W}$ and $\Lambda(h)$ is a strictly concave function of h when restricted to those functions $h \in \mathbb{H}$ such that $\Lambda(h) > -\infty$. Typically, when the model is concave, there is an essentially unique function η^* that maximizes the expected log-likelihood over $\mathbb{H}$, which we refer to as the best approximation in $\mathbb{H}$ to η; moreover, if the function η is in $\mathbb{H}$, then $\eta^* = \eta$ almost everywhere with respect to an appropriate measure on $\mathscr{U}$.

The class of concave extended linear models is extremely rich, containing many estimation problems as special cases, including ordinary and generalized regression, density and conditional hazard estimation, hazard and conditional density estimation, polychotomous regression, marked counting process regression, and proportional hazards regression. Many structural models can be dealt with in this framework. By choosing $\mathbb{H}$ appropriately, we can get additive models, partly linear models, varying coefficient models, and functional ANOVA models. See Stone et al. (1997) and Huang (2001) for more discussion. In this paper we restrict our attention to concave extended linear models.

Let $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ be a random sample of size n from the distribution of $\boldsymbol{W}$. When it is well defined, the (normalized) log-likelihood corresponding to this random sample is given by $\ell(h) = n^{-1} \sum_i l(h, \boldsymbol{W}_i)$. Let $\mathbb{G}_\gamma$, $\gamma \in \Gamma$, be a collection of finite-dimensional linear subspaces of $\mathbb{H}$. We assume that each function in every such space $\mathbb{G}_\gamma$ is bounded and that if it equals zero almost everywhere on $\mathscr{U}$, then it equals zero everywhere on $\mathscr{U}$. We call $\mathbb{G}_\gamma$ an estimation space. For each fixed $\gamma \in \Gamma$, the maximum likelihood estimate is given by $\hat{\eta}_\gamma = \operatorname{argmax}_{g \in \mathbb{G}_\gamma} \ell(g)$. We let the data pick which estimation space to use. To be specific, we choose $\hat{\gamma} \in \Gamma$ such that $\ell(\hat{\eta}_{\hat{\gamma}}) = \max_{\gamma \in \Gamma} \ell(\hat{\eta}_\gamma)$. (Such a $\hat{\gamma}$ exists under mild conditions; see Lemma 2.1 below.) We will study the benefit of allowing the flexibility to pick estimation spaces among a big collection. Specifically we will study the rate of convergence of $\hat{\eta}_{\hat{\gamma}} - \eta^*$, where η^* is the best approximation in $\mathbb{H}$ to the function η of interest.

In the above setup, we assume that $\mathbb{G}_\gamma$, $\gamma \in \Gamma$, have the same dimension and that the index set Γ is a compact subset of $\mathbb{R}^J$ for some positive integer J. The dimension of $\mathbb{G}_\gamma$, Γ and J are allowed to vary with the sample size n. Let $\|\cdot\|$ be a norm on $\mathbb{H}$ such that $\|h\| < \infty$ and $\|h\| \leqslant C_0\|h\|_\infty$ for $h \in \mathbb{H}$ and a positive constant C_0. This norm is used to measure the distance between two functions in $\mathbb{H}$. Typically, it is chosen to be an L_2-norm on $\mathscr{U}$ relative to an appropriate measure that depends on the estimation problem. In the regression context, for example, a natural choice is given by $\|h\|^2 = E[h^2(\boldsymbol{U})]$ where $\boldsymbol{U}$ is the random vector of covariates. In the following, we assume without loss of generality that $C_0 = 1$ since, otherwise, we can apply the same arguments to the norm $\|\cdot\|/C_0$. For $\gamma \in \Gamma$, set

$$N_n = \dim(\mathbb{G}_\gamma),$$

$$A_{n\gamma} = \sup_{g\in\mathbb{G}_\gamma} \frac{\|g\|_\infty}{\|g\|} := \sup_{\substack{g\in\mathbb{G}_\gamma \\ \|g\|\neq 0}} \frac{\|g\|_\infty}{\|g\|},$$

and

$$\rho_{n\gamma} = \inf_{g\in\mathbb{G}_\gamma} \|g - \eta^*\|_\infty.$$

Fix $n \geqslant 1$ and suppose that $A_n = \sup_{\gamma\in\Gamma} A_{n\gamma} < \infty$. Then the norms $\|\cdot\|$ and $\|\cdot\|_\infty$ are uniformly equivalent on $\mathbb{G}_\gamma$, $\gamma \in \Gamma$, in the sense that $\|g\| \leqslant \|g\|_\infty \leqslant A_n\|g\|$ for $\gamma \in \Gamma$ and $g \in \mathbb{G}_\gamma$.

It follows from Theorem 2.1 of Huang (2001) that, under regularity conditions, $\|\hat\eta_\gamma - \eta^*\|^2 = \mathrm{O}_\mathrm{P}(\rho_{n\gamma}^2 + N_n/n)$ 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 under mild conditions; see Lemma 2.1 below.) Then $\|\hat\eta_{\gamma^*} - \eta^*\|^2 = \mathrm{O}_\mathrm{P}(\rho_{n\gamma^*}^2 + N_n/n) = \mathrm{O}_\mathrm{P}(\inf_{\gamma\in\Gamma} \rho_{n\gamma}^2 + N_n/n)$. Thus $\inf_{\gamma\in\Gamma} \|\hat\eta_\gamma - \eta^*\|^2 \leqslant \|\hat\eta_{\gamma^*} - \eta^*\|^2 = \mathrm{O}_\mathrm{P}(\inf_{\gamma\in\Gamma} \rho_{n\gamma}^2 + N_n/n)$. It is natural to expect that, with γ estimated by $\hat\gamma$, the squared L_2 norm of the difference between the estimator and the target, i.e., $\|\hat\eta_{\hat\gamma} - \eta^*\|^2$ will be not much larger than the ideal quantity $\inf_{\gamma\in\Gamma} \|\hat\eta_\gamma - \eta^*\|^2$. Hence we hope that $\|\hat\eta_{\hat\gamma} - \eta^*\|^2$ will be not much larger than $\inf_{\gamma\in\Gamma} \rho_{n\gamma}^2 + N_n/n$ in probability. The main results stated in the following proposition are concerned with justifying this heuristic under suitable conditions.

Let $V_n = \bar{\mathrm{O}}_\mathrm{P}(b_n)$ mean that $\lim_n P(|V_n| \geqslant cb_n) = 0$ for some $c > 0$, where $b_n > 0$ for $n \geqslant 1$. Let $V_{n\gamma} = \mathrm{O}_\mathrm{P}(b_{n\gamma})$ uniformly over $\gamma \in \Gamma$ mean that $\lim_{c\to\infty} \limsup_n P(|V_{n\gamma}| \geqslant cb_{n\gamma}$ for some $\gamma \in \Gamma) = 0$, where $b_{n\gamma} > 0$ for $n \geqslant 1$ and $\gamma \in \Gamma$.

Proposition 2.1. *Suppose Conditions* 2.1–2.2 *and* 2.4–2.6 *hold and that* $\lim_n \sup_{\gamma\in\Gamma} A_{n\gamma}\rho_{n\gamma} = 0$ *and* $\lim_n \sup_{\gamma\in\Gamma} A_{n\gamma}^2 N_n/n = 0$. *Then, for n sufficiently large,* $\bar\eta_\gamma = \mathrm{argmax}_{g\in\mathbb{G}_\gamma} \Lambda(g)$ *exists uniquely for* $\gamma \in \Gamma$ *and* $\|\bar\eta_\gamma - \eta^*\|^2 = \mathrm{O}(\rho_{n\gamma}^2)$ *uniformly over* $\gamma \in \Gamma$. *Moreover, except on an event whose probability tends to zero as* $n \to \infty$, $\hat\eta_\gamma$ *exists uniquely for* $\gamma \in \Gamma$ *and* $\sup_{\gamma\in\Gamma} \|\hat\eta_\gamma - \bar\eta_\gamma\|^2 = \mathrm{O}_\mathrm{P}(N_n/n)$. *Consequently,* $\|\hat\eta_\gamma - \eta^*\|^2 = \mathrm{O}_\mathrm{P}(\rho_{n\gamma}^2 + N_n/n)$ *uniformly over* $\gamma \in \Gamma$. *In addition,*

$$\|\hat\eta_{\hat\gamma} - \eta^*\|^2 = \mathrm{O}_\mathrm{P}\left(\inf_{\gamma\in\Gamma} \rho_{n\gamma}^2\right) + \bar{\mathrm{O}}_\mathrm{P}\left((\log n)\frac{N_n}{n}\right).$$

The proof of this result is broken up into three theorems (Theorems 2.1–2.3) that will be given in the following subsections where technical conditions are stated explicitly. The technical conditions will be verified in the contexts of density estimation and generalized regression in Section 4 when $\mathbb{G}_\gamma$, $\gamma \in \Gamma$, are spaces of tensor product splines. The $\log n$ term in the final result of Proposition 2.1 plays an essential role in the proof of that result, but we do not know whether it is essential to the result itself.

2.2. Uniformity in rates of convergence

If γ is predetermined (independent of data) but $N_n = \dim(\mathbb{G}_\gamma)$ is allowed to increase with the sample size, then the rate of convergence of $\hat{\eta}_\gamma$ in the context of concave extended linear models is thoroughly treated in Huang (2001). In this section, we show that the rates of convergence results in Huang (2001) hold uniformly in $\gamma \in \Gamma$ if the sufficient conditions in those results hold in a uniform sense. Theorems 2.1 and 2.2 below are in parallel to Theorems A.1 and A.2 of the cited paper and can be proven by similar arguments (details of proof are omitted to save space).

For each fixed $\gamma \in \Gamma$, decompose the error into a stochastic part and a systematic part:

$$\hat{\eta}_\gamma - \eta^* = (\hat{\eta}_\gamma - \bar{\eta}_\gamma) + (\bar{\eta}_\gamma - \eta^*),$$

where $\hat{\eta}_\gamma - \bar{\eta}_\gamma$ is referred to as the *estimation error* and $\bar{\eta}_\gamma - \eta^*$ as the *approximation error*.

Condition 2.1. The best approximation η^* in $\mathbb{H}$ to η exists and there is a positive constant K_0 such that $\|\eta^*\|_\infty \leqslant K_0$.

Condition 2.2. For each pair h_1, h_2 of bounded functions in $\mathbb{H}$, $\Lambda(h_1 + \alpha(h_2 - h_1))$ is twice continuously differentiable with respect to α. (i) For any positive constant K, there is a fixed positive number M such that if $h_1, h_2 \in \mathbb{H}$, $\|h_1\|_\infty \leqslant K$, and h_2 is bounded, then

$$\left| \frac{\mathrm{d}}{\mathrm{d}\alpha} \Lambda(h_1 + \alpha h_2)|_{\alpha=0} \right| \leqslant M\|h_2\|.$$

(ii) For any positive constant K, there are fixed positive numbers M_1 and $M_2 \leqslant M_1$ such that

$$-M_1\|h_2 - h_1\|^2 \leqslant \frac{\mathrm{d}^2}{\mathrm{d}\alpha^2} \Lambda(h_1 + \alpha(h_2 - h_1)) \leqslant -M_2\|h_2 - h_1\|^2$$

for $h_1, h_2 \in \mathbb{H}$ with $\|h_1\|_\infty \leqslant K$ and $\|h_2\|_\infty \leqslant K$ and $0 \leqslant \alpha \leqslant 1$.

Condition 2.1 is the same as Condition A.1 of Huang (2001). Condition 2.2 strengthens Condition A.2 of Huang (2001) by putting an additional requirement on the first derivative of $\Lambda(\cdot)$. Condition 2.2(ii) implies that the restriction of $\Lambda(\cdot)$ to the bounded functions in $\mathbb{H}$ is strictly concave. The following result extends Theorem A.1 of Huang (2001).

Theorem 2.1 (Approximation error). *Suppose Conditions* 2.1 *and* 2.2 *hold and that* $\lim_n \sup_{\gamma\in\Gamma} A_{n\gamma}\rho_{n\gamma}=0$. *Let* K_1 *be a positive constant such that* $K_1 > K_0$ *with* K_0 *as in Condition* 2.1. *Then, for n sufficiently large,* $\bar\eta_\gamma$ *exists uniquely and* $\|\bar\eta_\gamma\|_\infty \leqslant K_1$ *for* $\gamma\in\Gamma$. *Moreover,* $\|\bar\eta_\gamma - \eta^*\|^2 = \mathrm{O}(\rho_{n\gamma}^2)$ *uniformly over* $\gamma\in\Gamma$.

Condition 2.3. There is a positive constant K_0 such that, for n sufficiently large, $\bar\eta_\gamma$ exists uniquely and $\|\bar\eta_\gamma\|_\infty \leqslant K_0$ for $\gamma\in\Gamma$.

Condition 2.4. For $\gamma\in\Gamma$ and $g_1, g_2\in\mathbb{G}_\gamma, \ell(g_1+\alpha(g_2-g_1))$ is twice continuously differentiable with respect to $\alpha\in[0,1]$. (i) The following holds:

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

(ii) For any positive constant K, there is a fixed positive number M such that

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

for $\gamma\in\Gamma$ and $g_1, g_2\in\mathbb{G}_\gamma$ with $\|g_1\|_\infty\leqslant K$ and $\|g_2\|_\infty \leqslant K$, except on an event whose probability tends to zero as $n\to\infty$; moreover,

$$\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\ell(g_1+\alpha(g_2-g_1)) \leqslant 0, \quad -\infty<\alpha<\infty,$$

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

The above two conditions are strengthened versions of A.3 and A.4 of Huang (2001). Condition 2.3 is in fact a consequence of Theorem 2.1. It is convenient to state it as a separate condition in order to avoid having to specify conditions on the expected log-likelihood when we study the estimation error in Theorem 2.2 below. Condition 2.4(ii) implies that $\ell(\cdot)$ is concave and largely strictly concave on each $\mathbb{G}_\gamma$. The following result extends Theorem A.2 of Huang (2001).

Theorem 2.2 (Estimation error). *Suppose Conditions* 2.3 *and* 2.4 *hold and that* $\lim_n \sup_{\gamma\in\Gamma} A_{n\gamma}^2 N_n/n=0$. *Let* K_1 *be a positive constant such that* $K_1 > K_0$ *with* K_0 *as in Condition* 2.3. *Then* $\hat\eta_\gamma$ *exists uniquely and* $\|\hat\eta_\gamma\|_\infty \leqslant K_1$ *for* $\gamma\in\Gamma$, *except on an event whose probability tends to zero as* $n\to\infty$. *Moreover*, $\sup_{\gamma\in\Gamma}\|\hat\eta_\gamma-\bar\eta_\gamma\|^2 = \mathrm{O_P}(N_n/n)$.

We have the decomposition

$$\hat\eta_{\hat\gamma} - \eta^* = (\hat\eta_{\hat\gamma} - \bar\eta_{\hat\gamma}) + (\bar\eta_{\hat\gamma} - \eta^*).$$

Note that Theorem 2.2 implies that $\|\hat\eta_{\hat\gamma} - \bar\eta_{\hat\gamma}\|^2 = \mathrm{O_P}(N_n/n)$. It remains to study the rate of convergence of $\bar\eta_{\hat\gamma} - \eta^*$, which is given in the next subsection.

2.3. Adaptive parameter selection

Condition 2.5. For $K<\infty$, the set $\{(\gamma, g)\colon \gamma\in\Gamma,\ g\in\mathbb{G}_\gamma, \text{ and } \|g\|_\infty\leqslant K\}$ is compact and $\ell(\cdot)$ is continuous on this set.

When $\mathbb{G}_\gamma$, $\gamma \in \Gamma$, are spaces of tensor product splines as in Section 3, the first part of Condition 2.5 follows from Lemmas 2.1 and 4.1 of Chapter 5 of DeVore and Lorentz (1993). Under the further restriction to density estimation and generalized regression in Section 4, the second part of Condition 2.5 follows from the corresponding explicit forms of the log-likelihood function.

Lemma 2.1. *Suppose Condition* 2.5 *holds. Then there is a* $\gamma^* \in \Gamma$ *such that* $\rho_{n\gamma^*} = \inf_{\gamma\in\Gamma} \rho_{n\gamma}$. *Moreover, on the event that* $\hat{\eta}_\gamma$ *exists uniquely and* $\|\hat{\eta}_\gamma\|_\infty \leqslant K_1$ *for* $\gamma \in \Gamma$, *where* K_1 *is a positive constant, there is a* $\hat{\gamma} \in \Gamma$ *such that* $\ell(\hat{\eta}_{\hat{\gamma}}) = \sup_{\gamma\in\Gamma} \ell(\hat{\eta}_\gamma)$.

Proof. Given $\gamma \in \Gamma$, choose $g_\gamma \in \mathbb{G}_\gamma$ such that $\|g_\gamma - \eta^*\|_\infty = \rho_\gamma$. By Condition 2.5, we can choose $\gamma_v \in \Gamma$ such that $\gamma_v \to \gamma^* \in \Gamma$, $\rho_{\gamma_v} \to \inf_{\gamma\in\Gamma} \rho_\gamma$, and $\|g_{\gamma_v} - g^*\|_\infty \to 0$ as $v \to \infty$, where $g^* \in \mathbb{G}_{\gamma^*}$. Then $\|g^* - \eta^*\|_\infty = \inf_{\gamma\in\Gamma} \rho_\gamma$, so γ^* has its desired property.

It follows from Condition 2.5 that, on the indicated event, we can choose $\gamma_v \in \Gamma$ such that $\gamma_v \to \hat{\gamma} \in \Gamma$, $\ell(\hat{\eta}_{\gamma_v}) \to \sup_{\gamma\in\Gamma} \ell(\hat{\eta}_\gamma)$, and $\hat{\eta}_{\gamma_v} \to g$ as $v \to \infty$, where $g \in \mathbb{G}_{\hat{\gamma}}$. Since $\ell(\cdot)$ is continuous, $\ell(g) = \sup_{\gamma\in\Gamma} \ell(\hat{\eta}_\gamma)$, so $g = \hat{\eta}_{\hat{\gamma}}$ and hence $\hat{\gamma}$ has its desired property. □

Let $V_{n\gamma} = \bar{\mathrm{O}}_\mathrm{P}(b_{n\gamma})$ uniformly over $\gamma \in \Gamma$ mean that, for some $c \in (0, \infty)$, $\lim_n P(|V_{n\gamma}| \geqslant cb_{n\gamma}$ for some $\gamma \in \Gamma) = 0$, where $b_{n\gamma} > 0$ for $n \geqslant 1$ and $\gamma \in \Gamma$.

Condition 2.6. (i) $|\ell(\bar{\eta}_{\gamma^*}) - \ell(\eta^*) - [\Lambda(\bar{\eta}_{\gamma^*}) - \Lambda(\eta^*)]| = \mathrm{O}_\mathrm{P}(\inf_{\gamma\in\Gamma}\rho^2_{n\gamma} + \frac{N_n}{n})$ and

(ii) $|\ell(\bar{\eta}_\gamma) - \ell(\eta^*) - [\Lambda(\bar{\eta}_\gamma) - \Lambda(\eta^*)]| = \bar{\mathrm{O}}_\mathrm{P}((\log^{1/2} n)\|\bar{\eta}_\gamma - \eta^*\|(\frac{N_n}{n})^{1/2} + (\log n)\frac{N_n}{n})$ uniformly over $\gamma \in \Gamma$.

In Section 4, we will verify that Condition 2.6 holds under reasonable conditions in the contexts of density estimation and generalized regression. There, we will actually verify a slight strengthening of the second property of Condition 2.6:

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

uniformly over $\gamma \in \Gamma$.

Theorem 2.3 (Parameter selection). *Suppose Conditions* 2.1–2.6 *hold and that* $\lim_n \sup_{\gamma\in\Gamma} A_{n\gamma}\rho_{n\gamma} = 0$ *and* $\lim_n \sup_{\gamma\in\Gamma} A^2_{n\gamma} N_n/n = 0$. *Then* $\|\bar{\eta}_{\hat{\gamma}} - \eta^*\|^2 = \mathrm{O}_\mathrm{P}(\inf_{\gamma\in\Gamma}\rho^2_{n\gamma}) + \bar{\mathrm{O}}_\mathrm{P}((\log n)N_n/n)$.

Proof. We first show that

$$\ell(\hat{\eta}_\gamma) - \ell(\bar{\eta}_\gamma) = \mathrm{O}_\mathrm{P}\left(\frac{N_n}{n}\right) \quad \text{uniformly in } \gamma \in \Gamma. \tag{2.1}$$

Write

$$f(\alpha) = \ell(\bar{\eta}_\gamma + \alpha(\hat{\eta}_\gamma - \bar{\eta}_\gamma)), \quad \gamma \in \Gamma.$$

By Condition 2.4, $f''(\alpha) \leqslant 0$ (except on an event whose probability tends to zero as $n \to \infty$). Thus,

$$0 \leqslant \ell(\hat{\eta}_\gamma) - \ell(\bar{\eta}_\gamma) = f(1) - f(0) = f'(0) + \int_0^1 (1-\alpha) f''(\alpha)\,\mathrm{d}\alpha \leqslant f'(0).$$

On the other hand, by Condition 2.4(i) and Theorem 2.2,

$$f'(0) = \left.\frac{\mathrm{d}}{\mathrm{d}\alpha}\ell(\bar{\eta}_\gamma + \alpha(\hat{\eta}_\gamma - \bar{\eta}_\gamma))\right|_{\alpha=0} = \mathrm{O_P}\left(\left(\frac{N_n}{n}\right)^{1/2}\right)\|\hat{\eta}_\gamma - \bar{\eta}_\gamma\| = \mathrm{O_P}\left(\frac{N_n}{n}\right)$$

uniformly in $\gamma \in \Gamma$. The desired result follows.

By Theorem 2.1, $\bar{\eta}_{\hat{\gamma}}$ is bounded. Thus it follows from Lemma A.1 of Huang (2001) that, for some positive constant M,

$$M\|\bar{\eta}_{\hat{\gamma}} - \eta^*\|^2 \leqslant \Lambda(\eta^*) - \Lambda(\bar{\eta}_{\hat{\gamma}}).$$

Since $\gamma^* \in \Gamma$ satisfies $\rho_{n\gamma^*} = \inf_{\gamma\in\Gamma}\rho_{n\gamma}$, $\|\bar{\eta}_{\gamma^*} - \eta^*\|^2 = \mathrm{O}(\inf_{\gamma\in\Gamma}\rho^2_{n\gamma})$ by Theorem 2.1. We have the decomposition

$$\begin{aligned}\Lambda(\eta^*) - \Lambda(\bar{\eta}_{\hat{\gamma}}) &= \Lambda(\eta^*) - \Lambda(\bar{\eta}_{\gamma^*}) + \Lambda(\bar{\eta}_{\gamma^*}) - \Lambda(\bar{\eta}_{\hat{\gamma}})\\ &= I_1 + I_2 - I_3 + I_4,\end{aligned}$$

where

$$I_1 = \Lambda(\eta^*) - \Lambda(\bar{\eta}_{\gamma^*}),$$

$$I_2 = \Lambda(\bar{\eta}_{\gamma^*}) - \Lambda(\eta^*) - [\ell(\bar{\eta}_{\gamma^*}) - \ell(\eta^*)],$$

$$I_3 = \Lambda(\bar{\eta}_{\hat{\gamma}}) - \Lambda(\eta^*) - [\ell(\bar{\eta}_{\hat{\gamma}}) - \ell(\eta^*)],$$

$$I_4 = \ell(\bar{\eta}_{\gamma^*}) - \ell(\bar{\eta}_{\hat{\gamma}}).$$

Note that $I_1 = \mathrm{O}(\inf_{\gamma\in\Gamma}\rho^2_{n\gamma})$ by Theorem 2.1 and Lemma A.1 of Huang (2001). The terms I_2 and I_3 can be bounded using Condition 2.6. Moreover, by using (2.1) and $\ell(\hat{\eta}_{\gamma^*}) \leqslant \ell(\hat{\eta}_{\hat{\gamma}})$ [which follows from the definition of $\hat{\gamma}$], we get that

$$I_4 = \ell(\hat{\eta}_{\gamma^*}) - \ell(\hat{\eta}_{\hat{\gamma}}) + \mathrm{O_P}\left(\frac{N_n}{n}\right) \leqslant \mathrm{O_P}\left(\frac{N_n}{n}\right).$$

Hence

$$\begin{aligned}\|\bar{\eta}_{\hat{\gamma}} - \eta^*\|^2 \leqslant{}& \bar{\mathrm{O}}_\mathrm{P}\left((\log^{1/2} n)\|\bar{\eta}_{\hat{\gamma}} - \eta^*\|\left(\frac{N_n}{n}\right)^{1/2} + (\log n)\frac{N_n}{n}\right)\\ &+ \mathrm{O_P}\left(\inf_{\gamma\in\Gamma}\rho^2_{n\gamma} + \frac{N_n}{n}\right).\end{aligned} \tag{2.2}$$

Observe that, for positive numbers B and C, $z^2 \leqslant Bz + C$ implies that $2z^2 \leqslant (B^2 + z^2) + 2C$ and hence that $z^2 \leqslant B^2 + 2C$. Therefore (2.2) yields the desired result. □

3. Free knot splines and their tensor products

In this section we will develop some properties of spaces of free knot splines and tensor products of such spaces, which will be used in Section 4 to verify Conditions 2.4 and 2.6.

For $1 \leqslant l \leqslant L$, let $\mathscr{U}_l = [a_l, b_l]$ be a compact subinterval of $\mathbb{R}$ having positive length $b_l - a_l$ and let $\mathscr{U}$ denote the Cartesian product of $\mathscr{U}_1, \ldots, \mathscr{U}_L$. For each l, let m_l be an integer with $m_l \geqslant 2$, 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 $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L)$, let $\mathbb{G}_\gamma$ be the tensor product of $\mathbb{G}_{l\gamma_l}$, $1 \leqslant l \leqslant L$ (that is, the linear space spanned by $g_1(u_1) \cdots g_L(u_L)$ as g_l runs over $G_{l\gamma_l}$), which has dimension $N_n = \prod_l N_{nl}$.

For $1 \leqslant l \leqslant L$, 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

$$\frac{\gamma_{l,j_2-1} - \gamma_{l,j_2-m_l}}{\gamma_{l,j_1-1} - \gamma_{l,j_1-m_l}} \leqslant \bar{M}_l, \quad 2 \leqslant j_1, j_2 \leqslant J_l + m_l, \tag{3.1}$$

where $\gamma_{l,1-m_l} = \cdots = \gamma_{l0} = a$ and $\gamma_{l,J_l+1} = \cdots = \gamma_{l,J_l+m_l} = b$. Let Γ denote the Cartesian product of Γ_l, $1 \leqslant l \leqslant L$, which can be viewed as a subset of $\mathbb{R}^J$ with $J = \sum_l J_l$. We consider the use of the collection $\mathbb{G}_\gamma$, $\boldsymbol{\gamma} \in \Gamma$, in fitting an extended linear model. Such a collection of free knot splines has some properties that we will list below. (The proofs will be given in Appendix A.) In the technical arguments, we need to approximate Γ by a finite subset of a larger set $\tilde{\Gamma}$, which is defined in the same way as Γ, but with $\bar{M}_l$ in (3.1) replaced by the larger constant $3\bar{M}_l$.

Let ψ denote the uniform distribution on $\mathscr{U}$ and let vol$(\mathscr{U})$ denote the volume of $\mathscr{U}$. Let $\mathbb{H}$ denote the space of (real-valued) functions on $\mathscr{U}$ that are square-integrable with respect to ψ, and let $\langle \cdot, \cdot \rangle_\psi$ and $\| \cdot \|_\psi$ denote the inner product and norm on $\mathbb{H}$ given by

$$\langle h_1, h_2 \rangle_\psi = \int_{\mathscr{U}} h_1(\boldsymbol{u}) h_2(\boldsymbol{u})\, \psi(\mathrm{d}\boldsymbol{u}) = \frac{1}{\mathrm{vol}(\mathscr{U})} \int_{\mathscr{U}} h_1(\boldsymbol{u}) h_2(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}$$

and $\|h\|_\psi^2 = \langle h, h \rangle_\psi$.

In the statement of the main results, we were rather vague about the form of the norm $\| \cdot \|$ used. To verify the technical conditions we need to be more specific. Let $\boldsymbol{U}$ denote a $\mathscr{U}$-valued random variable that is a transform (function) of $\boldsymbol{W}$ (for example, $\boldsymbol{W} = (\boldsymbol{X}, Y)$ and $\boldsymbol{U} = \boldsymbol{X}$). Partly for simplicity, we consider the theoretical inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$ on $\mathbb{H}$ given by $\langle h_1, h_2 \rangle = E[h_1(\boldsymbol{U}) h_2(\boldsymbol{U})]$ and $\|h\|^2 = \langle h, h \rangle = E[h^2(\boldsymbol{U})]$. Define the empirical inner product and empirical norm by $\langle h_1, h_2 \rangle_n = E_n(h_1 h_2) = n^{-1} \sum_i h_1(\boldsymbol{U}_i) h_2(\boldsymbol{U}_i)$ and $\|h\|_n^2 = \langle h, h \rangle_n = n^{-1} \sum_i h^2(\boldsymbol{U}_i)$.

Condition 3.1. The random variable $\boldsymbol{U}$ has a density function $f_{\boldsymbol{U}}$ such that $M_1/\mathrm{vol}(\mathscr{U}) \leqslant f_{\boldsymbol{U}} \leqslant M_2/\mathrm{vol}(\mathscr{U})$ on $\mathscr{U}$, where M_1 and M_2 are fixed positive numbers.

It follows from Condition 3.1 that $M_1 \leqslant 1 \leqslant M_2$ and

$$M_1\|h\|_\psi^2 \leqslant \|h\|^2 \leqslant M_2\|h\|_\psi^2, \quad h\in\mathbb{H}. \tag{3.2}$$

Let $|\cdot|_\infty$ denote the l_∞ norm on any Euclidean space. Let ζ denote the metric on $\mathbb{R}^J$ given by $\zeta(\gamma,\tilde\gamma) = \max_l 9\bar{M}_l N_{nl}|\gamma_l - \tilde\gamma_l|_\infty/(b_l - a_l)$. The following lemmas will be proved in the Appendix A.

Lemma 3.1. *Let $0<\varepsilon\leqslant 1/2$ and let K be a positive integer. There is a positive constant M and there are subsets Ξ_k, $0\leqslant k\leqslant K$, of $\tilde\Gamma$ such that*

$$\#(\Xi_k) \leqslant (M\varepsilon^{-k})^{N_n}, \quad 1\leqslant k\leqslant K;$$

every point in Γ is within ε^K of some point in Ξ_K (in ζ distance); and, for $1\leqslant k\leqslant K$, every point in Ξ_k is within ε^{k-1} of some point in Ξ_{k-1}.

Let $0<\varepsilon\leqslant 1/2$ and let Ξ_k, $0\leqslant k\leqslant K$ be as in Lemma 3.1. Given $\gamma\in\tilde\Gamma$, set $\mathbb{B}_\gamma = \{g\in\mathbb{G}_\gamma\colon \|g\|\leqslant 1\}$. Let k be a nonnegative integer. If $k=0$, set $\mathbb{B}_{\gamma k}=\{0\}$; otherwise, let $\mathbb{B}_{\gamma k}$ be a maximal subset of $\mathbb{B}_\gamma$ such that any two functions in $\mathbb{B}_{\gamma k}$ are at least ε^k apart in the norm $\|\ \|$. Then $\min_{\tilde g\in\mathbb{B}_{\gamma k}}\|g-\tilde g\|\leqslant\varepsilon^k$ for $g\in\mathbb{B}_\gamma$. Moreover,

$$\#(\mathbb{B}_{\gamma k}) \leqslant \left(\frac{1+\varepsilon^k/2}{\varepsilon^k/2}\right)^{N_n} \leqslant (3\varepsilon^{-k})^{N_n}.$$

Set $\mathbb{B}_k = \bigcup_{\gamma\in\Xi_k}\mathbb{B}_{\gamma k}$. Then, by Lemma 3.1,

$$\#(\mathbb{B}_k) \leqslant (M'\varepsilon^{-2k})^{N_n}, \quad 1\leqslant k\leqslant K, \tag{3.3}$$

for some constant $M'\geqslant 1$. Also, set $\mathbb{B} = \{g\in\bigcup_{\gamma\in\Gamma}\mathbb{G}_\gamma\colon \|g\|\leqslant 1\} = \bigcup_{\gamma\in\Gamma}\mathbb{B}_\gamma$ and $\tilde{\mathbb{B}} = \{g\in\bigcup_{\gamma\in\tilde\Gamma}\mathbb{G}_\gamma\colon \|g\|\leqslant 1\} = \bigcup_{\gamma\in\tilde\Gamma}\mathbb{B}_\gamma$.

Lemma 3.2. *Suppose, for a given positive integer n, that $\bar\eta_\gamma$ exists uniquely and is bounded for $\gamma\in\tilde\Gamma$ and that $\|\bar\eta_\gamma-\eta^*\|$ is a continuous function of $\gamma\in\tilde\Gamma$. There is a positive constant M such that, for $0<\varepsilon\leqslant 1$, there is a subset $\tilde\Gamma'$ of $\tilde\Gamma$ such that*

$$\#(\tilde\Gamma') \leqslant \exp(M[\log(2/\varepsilon)]N_n)$$

and every point γ in Γ is within ε (in ζ distance) of some point $\tilde\gamma$ in $\tilde\Gamma'$ such that $\|\bar\eta_{\tilde\gamma}-\eta^*\|\leqslant\|\bar\eta_\gamma-\eta^*\|$.

The condition that $\|\bar\eta_\gamma-\eta^*\|$ is a continuous function of $\gamma\in\tilde\Gamma$, which is used in the above lemma, follows from the first conclusion of Lemma 3.5.

Lemma 3.3. *Suppose Condition 3.1 holds. There is a positive constant M such that*

$$\|g\|_\infty \leqslant MN_n^{1/2}\|g\|, \quad \gamma\in\tilde\Gamma \text{ and } g\in\mathbb{G}_\gamma. \tag{3.4}$$

Lemma 3.4. *There are positive numbers M_1 and M_2 such that, for $\gamma,\tilde\gamma\in\tilde\Gamma$ and $g\in\mathbb{G}_\gamma$, there is a function $\tilde g\in\mathbb{G}_{\tilde\gamma}$ such that $\|\tilde g\|\leqslant\|g\|$, $\|\tilde g-g\|\leqslant M_1\zeta(\gamma,\tilde\gamma)\|g\|$, and $\|\tilde g-g\|_\infty\leqslant M_2\zeta(\gamma,\tilde\gamma)\|g\|_\infty$. Suppose Condition 3.1 holds and that $\lim_n N_n^2/n=0$.*

Then there is a positive number M_3 *and an event* Ω_n *such that* $\lim_n P(\Omega_n) = 1$ *and the functions* $\tilde{g}$ *above can be chosen to satisfy the additional property that* $\|g - \tilde{g}\|_n \leqslant M_3\zeta(\gamma, \tilde{\gamma})\|g\|$ *on* Ω_n *for* $\gamma, \tilde{\gamma} \in \tilde{\Gamma}$ *and* $g \in \mathbb{G}_\gamma$.

Lemma 3.5. *Suppose Condition* 2.2 *holds. Let K be a positive number. There are positive numbers* M_1 *and* M_2 *such that if* $\gamma, \tilde{\gamma} \in \tilde{\Gamma}$, $\zeta(\gamma, \tilde{\gamma}) \leqslant 1$, $\|\bar{\eta}_\gamma\|_\infty \leqslant K$, *and* $\|\bar{\eta}_{\tilde{\gamma}}\| \leqslant K$, *then* $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\| \leqslant M_1[\zeta(\gamma, \tilde{\gamma})]^{1/2}$, $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\|_\infty \leqslant M_2 N_n^{1/2}[\zeta(\gamma, \tilde{\gamma})]^{1/2}$. *Suppose, in addition Condition* 3.1 *holds and that* $\lim_n N_n^2/n = 0$. *Then there is an event* Ω_n *such that* $\lim_n P(\Omega_n) = 1$ *and* $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\|_n \leqslant M_1[\zeta(\gamma, \tilde{\gamma})]^{1/2}$ *for* $\gamma, \tilde{\gamma} \in \tilde{\Gamma}$ *on* Ω_n.

4. Verification of technical conditions

In this section we verify Conditions 2.2, 2.4 and 2.6 using primitive assumptions in some specific statistical contexts. As a consequence, the conclusions of Proposition 2.1 hold under these primitive assumptions. For simplicity, we focus on two contexts: density estimation in Section 4.2 and generalized regression, which includes ordinary regression as a special case, in Section 4.3. Again for simplicity, we also restrict attention to the saturated model (that is, there is no structural assumption and $\mathbb{H}$ is the collection of all square integrable functions on $\mathscr{U}$), so that $\eta^* = \eta$. Thus Condition 2.1 amounts to the assumption that η is bounded. The case of unsaturated models can be treated similarly at the expense of more complicated notation.

Throughout this section, we take $\mathbb{G}_\gamma$, $\gamma \in \tilde{\Gamma}$, to be tensor product free-knot spline spaces as defined in Section 3. It follows from (3.4) that $A_{n\gamma} \leqslant MN_n^{1/2}$ for some constant M, where N_n is the common dimension of $\mathbb{G}_\gamma$. Thus the requirements $\lim_n \sup_{\gamma \in \Gamma} A_{n\gamma}\rho_{n\gamma} = 0$ and $\lim_n \sup_{\gamma \in \Gamma} A_{n\gamma}^2 N_n/n = 0$, which are used in Proposition 2.1, reduce to $\lim_n \sup_{\gamma \in \Gamma} \rho_{n\gamma} N_n^{1/2} = 0$ and $\lim_n N_n^2/n = 0$, respectively.

Condition 4.1. $N_n^{-(c-1/2)} \lesssim \log^{-1/2} n$ and $N_n^c \sup_{\gamma \in \tilde{\Gamma}} \rho_{n\gamma} \lesssim 1$ for some $c > 1/2$.

4.1. Preliminary lemmas

Let $\xi_1, \ldots, \xi_n$ be independent random variables, and set $\bar{\xi} = (\xi_1 + \cdots + \xi_n)/n$. Suppose that, for $1 \leqslant i \leqslant n$, $E\xi_i = 0$ and

$$|E\xi_i^m| \leqslant \frac{m!}{2} b_i^2 H^{m-2}, \quad m \geqslant 2, \tag{4.1}$$

where $H > 0$. Set $B_n^2 = (b_1^2 + \cdots + b_n^2)/n$. Then, by Bernstein's inequality (see Yurinskiĭ, 1976),

$$P(|\bar{\xi}| \geqslant t) \leqslant 2\exp\left(-\frac{nt^2}{2(B_n^2 + tH)}\right) \tag{4.2}$$

for $t > 0$. Suppose, in particular, that $E\xi_i = 0$, $\operatorname{var}(\xi_i) \leqslant \sigma^2$, and $P(|\xi_i| \leqslant b) = 1$ for $1 \leqslant i \leqslant n$, where $b > 0$. Then (4.1) and hence (4.2) hold with $b_i = \sigma$ for $1 \leqslant i \leqslant n$, $B_n = \sigma$, and H replaced by b. In this case, however, (4.2) also holds with H replaced by $b/3$ (see (2.13) in Hoeffding, 1963). If we drop the assumption that $E\xi_i = 0$, we need to multiply b by 2. (Note that $|E\xi_i| \leqslant b$ and hence $|\xi_i - E\xi_i| \leqslant 2b$.) It follows easily from (4.2) that

$$P(|\bar{\xi}| \geqslant tH^{-1}[B_n(N_n/n)^{1/2} + N_n/n]) \leqslant 2\exp\left(-\frac{tH^{-2}N_n}{2}\right) \tag{4.3}$$

for $t \geqslant 1$. (Note that $B_n^2 + t[B_n(N_n/n)^{1/2} + N_n/n] \leqslant t[B_n + (N_n/n)^{1/2}]$ for $t \geqslant 1$.)

In the proofs of Lemmas 4.2 and 4.6, we will use a "chaining argument" that is well known in the empirical process theory literature; see Pollard (1984). For convenience, we summarize a portion of this argument in the form of the following result.

Lemma 4.1 (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 K be a positive integer; let $\mathbb{S}_k$ be a finite, nonempty subset of $\tilde{\mathbb{S}}$ for $0 \leqslant k \leqslant K$ such that $V_s = 0$ for $s \in \mathbb{S}_0$; let $C_1, \ldots, C_6$ be positive numbers; let $0 < \delta \leqslant 1/4$; and let Ω be an event. Suppose that*

$$P\left(\sup_{s\in\mathbb{S}} \min_{\tilde{s}\in\mathbb{S}_K} |V_s - V_{\tilde{s}}| > C_1; \Omega\right) \leqslant C_2; \tag{4.4}$$

$$\#(\mathbb{S}_k) \leqslant C_3 \exp(C_4 k), \quad 1 \leqslant k \leqslant K; \tag{4.5}$$

and

$$\max_{s\in\mathbb{S}_k} \min_{\tilde{s}\in\mathbb{S}_{k-1}} P(|V_s - V_{\tilde{s}}| > 2^{-(k-1)}C_5; \Omega)$$
$$\leqslant C_6 \exp(-2C_4(2\delta)^{-(k-1)}), \quad 1 \leqslant k \leqslant K. \tag{4.6}$$

Then

$$P\left(\sup_{s\in\mathbb{S}} |V_s| > C_1 + 2C_5\right) \leqslant C_2 + \frac{C_3C_6}{\exp(C_4) - 1} + P(\Omega^c)$$
$$\leqslant C_2 + \frac{C_3C_6}{C_4} + P(\Omega^c).$$

Proof. Observe that

$$\sup_{s\in\mathbb{S}} |V_s| \leqslant \sup_{s\in\mathbb{S}} \min_{\tilde{s}\in\mathbb{S}_K} |V_s - V_{\tilde{s}}| + \sup_{s\in\mathbb{S}_K} |V_s|.$$

So, in light of (4.4), it suffices to verify that

$$P\left(\max_{s\in\mathbb{S}_K} |V_s| \geqslant 2C_5; \Omega\right) \leqslant C_3C_6 \frac{\exp(-C_4)}{1 - \exp(-C_4)}. \tag{4.7}$$

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

$$P(|V_s - V_{\sigma_{k-1}(s)}| \geqslant 2^{-(k-1)}C_5; \Omega)$$
$$\leqslant C_6 \exp(-2C_4(2\delta)^{-(k-1)}), \quad 1 \leqslant k \leqslant K \text{ and } s \in \mathbb{S}_k;$$

the existence of σ_{k-1} follows from (4.6). Then, by (4.5),

$$P(|V_s - V_{\sigma_{k-1}(s)}| > 2^{-(k-1)}C_5 \text{ for some } k \in \{1,\ldots,K\} \text{ and } s \in \mathbb{S}_k; \Omega)$$
$$\leqslant \sum_{k=1}^{K} C_3 \exp(C_4 k) C_6 \exp(-2C_4(2\delta)^{-(k-1)}).$$

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

$$C_3C_6\sum_{k=1}^{K}\exp(-C_4(2\delta)^{-(k-1)}) \leqslant C_3C_6\sum_{k=1}^{K}\exp(-C_4k)$$
$$\leqslant C_3C_6\,\frac{\exp(-C_4)}{1-\exp(-C_4)}.$$

Suppose that $|V_s - V_{\sigma_{k-1}(s)}| \leqslant 2^{-(k-1)}C_5$ for $1 \leqslant k \leqslant K$ and $s \in \mathbb{S}_k$. Choose $s \in \mathbb{S}_K$ and set $s_K = s$, $s_{K-1} = \sigma_{K-1}(s_K), \ldots, s_0 = \sigma_0(s_1)$. (We refer to $s_K, \ldots, s_0$ 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 $|V_{s_k} - V_{s_{k-1}}| \leqslant 2^{-(k-1)}C_5$ for $1 \leqslant k \leqslant K$, so

$$|V_s| = \left|\sum_{k=1}^{K}(V_{s_k} - V_{s_{k-1}})\right| \leqslant 2C_5.$$

Consequently

$$P\left(\max_{s\in\mathbb{S}_K}|V_s| > 2C_5; \Omega\right)$$
$$\leqslant P(|V_s - V_{\sigma_{k-1}(s)}| > 2^{-(k-1)}C_5 \text{ for } k \in \{1,\ldots,K\} \text{ and } s \in \mathbb{S}_k; \Omega).$$

Thus (4.7) holds as desired. □

Lemma 4.2. *Suppose Condition* 3.1 *holds and that* $\lim_n N_n^2/n = 0$. *Then*

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

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

$$\tfrac{1}{2}\|g\|^2 \leqslant \|g\|_n^2 \leqslant 2\|g\|^2, \quad \gamma \in \tilde{\Gamma} \text{ and } g \in \mathbb{G}_\gamma.$$

This lemma extends Lemma 10 of Huang (1998a,b), which applies to fixed knot splines and other such linear approximation spaces, except that Condition 3.1 is not required there.

Proof of Lemma 4.2. It suffices to verify the lemma with $\tilde{\Gamma}$ replaced by Γ. Let $0 < \delta \leqslant 1/4$, let $0 < t < \infty$, let $K = K_n$ be a positive integer to be specified later, and let $\mathbb{B}$ and let Ξ_k and $\mathbb{B}_k$, $0 \leqslant k \leqslant K$, be as in Lemma 3.1 and the following paragraph with $\varepsilon = \delta$. We will apply Lemma 4.1 with $s = (f,g)$, $V_s = \langle f,g\rangle_n - \langle f,g\rangle = (E_n - E)(fg)$,

$\mathbb{S}=\{(f,g)\colon f,g\in\mathbb{B}\}$, $\mathbb{S}_k=\{(f,g)\colon f,g\in\mathbb{B}_k\}$ for $0\leqslant k\leqslant K$, and $\Omega^c=\emptyset$. It follows from (3.3) that

$$\#(\mathbb{S}_k)\leqslant(M'\delta^{-2k})^{2N_n},\quad 1\leqslant k\leqslant K,$$

and hence that (4.5) holds with $C_3=1$ and any $C_4\geqslant 4\log(M'\delta^{-1})N_n$.

Suppose Condition 3.1 holds, let $0<\varepsilon=\delta\leqslant 1/4$, let k be a positive integer, let $\gamma,\tilde{\gamma}\in\tilde{\Gamma}$ with $\zeta(\gamma,\tilde{\gamma})\leqslant\delta^{k-1}$, and let $g\in\mathbb{B}_\gamma$. Then, by (3.4) and Lemma 3.4, there is a function $g'\in\mathbb{B}_{\tilde{\gamma}}$ such that $\|g-g'\|\leqslant M_1\delta^{k-1}$ and $\|g-g'\|_\infty\leqslant MM_2N_n^{1/2}\delta^{k-1}$. Also, there is a function $\tilde{g}\in\mathbb{B}_{\tilde{\gamma},k-1}$ such that $\|g'-\tilde{g}\|\leqslant\delta^{k-1}$ and hence $\|g'-\tilde{g}\|_\infty\leqslant MN_n^{1/2}\delta^{k-1}$. Observe that $\|g\|_\infty\leqslant c_1N_n^{1/2}$, $\|\tilde{g}\|_\infty\leqslant c_1N_n^{1/2}$, $\|g-\tilde{g}\|\leqslant c_2\delta^{k-1}$, and $\|g-\tilde{g}\|_\infty\leqslant c_3N_n^{1/2}\delta^{k-1}$, where $c_1=M$, $c_2=M_1+1$, and $c_3=M(M_2+1)$.

Let k be a positive integer, and let $f,\tilde{f},g,\tilde{g}$ be functions on $\mathscr{U}$ such that $\|\tilde{f}\|_\infty\leqslant c_1N_n^{1/2}$, $\|f-\tilde{f}\|\leqslant c_2\delta^{k-1}$, $\|f-\tilde{f}\|_\infty\leqslant c_3N_n^{1/2}\delta^{k-1}$, $\|g\|_\infty\leqslant c_1N_n^{1/2}$, $\|g-\tilde{g}\|\leqslant c_2\delta^{k-1}$, and $\|g-\tilde{g}\|_\infty\leqslant c_3N_n^{1/2}\delta^{k-1}$. Then

$$\|fg-\tilde{f}\tilde{g}\|_\infty\leqslant\|f-\tilde{f}\|_\infty\|g\|_\infty+\|\tilde{f}\|_\infty\|g-\tilde{g}\|_\infty\leqslant 2c_1c_3N_n\delta^{k-1},$$

so $|(E_n-E)(fg-\tilde{f}\tilde{g})|\leqslant 4c_1c_3N_n\delta^{k-1}$. Moreover,

$$\begin{aligned}\operatorname{var}(fg-\tilde{f}\tilde{g})&\leqslant 2\operatorname{var}((f-\tilde{f})g)+2\operatorname{var}(\tilde{f}(g-\tilde{g}))\\&\leqslant 2\|g\|_\infty^2\|f-\tilde{f}\|^2+2\|\tilde{f}\|_\infty^2\|g-\tilde{g}\|^2\\&\leqslant 4c_1^2c_2^2N_n\delta^{2(k-1)}.\end{aligned}$$

Since $0<2\delta\leqslant 1$, it now follows from Bernstein's inequality (4.2) that, for $t>0$,

$$P(|(E_n-E)(fg-\tilde{f}\tilde{g})|\geqslant t2^{-(k-1)})\leqslant 2\exp\left(-\frac{nt^2(2\delta)^{-(k-1)}}{8c_1[c_1c_2^2+tc_3]N_n}\right).\tag{4.8}$$

Let K be such that $4c_1c_3N_n\delta^K\leqslant t$. Given $f,g\in\mathbb{B}$, let $\tilde{f},\tilde{g}\in\mathbb{B}_K$ be such that $\|f-\tilde{f}\|_\infty\leqslant c_3N_n^{1/2}\delta^K$ and $\|g-\tilde{g}\|_\infty\leqslant c_3N_n^{1/2}\delta^K$. Then $\|fg-\tilde{f}\tilde{g}\|_\infty\leqslant 2c_1c_3N_n\delta^K$, so $|(E_n-E)(fg-\tilde{f}\tilde{g})|\leqslant 4c_1c_3N_n\delta^K\leqslant t$. Consequently, (4.4) holds with $C_1=t$ and $C_2=0$.

Let $1\leqslant k\leqslant K$. For $f,g\in\mathbb{B}_k$, let $\tilde{f},\tilde{g}\in\mathbb{B}_{k-1}$ be such that $\|f-\tilde{f}\|\leqslant c_2\delta^{k-1}$, $\|f-\tilde{f}\|_\infty\leqslant c_3N_n^{1/2}\delta^{k-1}$, $\|g-\tilde{g}\|\leqslant c_2\delta^{k-1}$, and $\|g-\tilde{g}\|_\infty\leqslant c_3N_n^{1/2}\delta^{k-1}$. Since $N_n=o(n^{1/2})$, we now conclude from (4.8) that (4.6) holds with $C_5=t$, $C_6=2$, $\Omega^c=\emptyset$, and

$$C_4=\frac{nt^2}{16c_1[c_1c_2^2+tc_3]N_n}\geqslant 4\log(M'\delta^{-1})N_n$$

for n sufficiently large. It now follows from Lemma 4.1 that, for n sufficiently large,

$$P\left(\sup_{\gamma,\tilde{\gamma}\in\Gamma}\sup_{f\in\mathbb{B}_\gamma}\sup_{g\in\mathbb{B}_{\tilde{\gamma}}}|\langle f,g\rangle_n-\langle f,g\rangle|\geqslant 3t\right)\leqslant\frac{32c_1[c_1c_2^2+tc_3]N_n}{nt^2},$$

which tends to zero as $n\to\infty$. Since t can be made arbitrarily small, the first conclusion of the lemma is valid, from which the second conclusion follows easily. □

Lemma 4.3. *Suppose Condition* 3.1 *holds and that* $\lim_n N_n^2/n = 0$, *and let* h_n *be uniformly bounded functions on* $\mathscr{U}$. *Then*

$$\sup_{\gamma\in\Gamma}\sup_{g\in\mathbb{G}_\gamma}\frac{|\langle h_n,g\rangle_n-\langle h_n,g\rangle|}{\|g\|}=\mathrm{O_P}\left(\left(\frac{N_n}{n}\right)^{1/2}\right).$$

Proof. The proof of this result is a slight simplification of the Proof of Lemma 4.2. □

The next, obviously valid, lemma is useful in verifying the second property of Condition 2.6 in a variety of contexts.

Lemma 4.4. *Let* $C_1,\dots,C_4$ *be fixed positive numbers with* $C_3>1$. *Let* A_γ, $\gamma\in\tilde{\Gamma}$, *be positive numbers that depend on n, and let* V_γ, $\gamma\in\tilde{\Gamma}$, *be random variables that depend on n. Suppose that, for n sufficiently large,* $P(|V_\gamma|\geqslant C_1A_\gamma)\leqslant C_2\exp(-2C_3N_n\log n)$ *for* $\gamma\in\tilde{\Gamma}$. *Let* $\tilde{\Gamma}''$ *be a subset of* $\tilde{\Gamma}$ *such that*

$$\#(\tilde{\Gamma}'')\leqslant\exp(C_3N_n\log n)\quad \textit{for n sufficiently large.}\tag{4.9}$$

Suppose that, except on an event whose probability tends to zero as $n\to\infty$, *for every point* $\gamma\in\Gamma$, *there is a point* $\tilde{\gamma}\in\tilde{\Gamma}''$ *such that* $A_{\tilde{\gamma}}\leqslant A_\gamma$ *and* $|V_\gamma-V_{\tilde{\gamma}}|\leqslant C_4A_\gamma$. *Then* $|V_\gamma|=\bar{\mathrm{O}}_\mathrm{P}(A_\gamma)$ *uniformly over* $\gamma\in\Gamma$.

4.2. Density estimation

Let $\boldsymbol{Y}=\boldsymbol{W}$ have an unknown density function $f_{\boldsymbol{Y}}$ on $\mathscr{Y}=\mathscr{U}$, and let $\phi=\log f_{\boldsymbol{Y}}$ denote the corresponding log-density function. Let $\mathbb{H}_1$ be a linear space of functions on $\mathscr{Y}$ that contains all constant functions. We model the log-density function ϕ as a member of $\mathbb{H}_1$. Note that ϕ satisfies the nonlinear constraint $c(\phi)=\log\int_{\mathscr{Y}}\exp\phi(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}=0$. It is convenient to write $\phi=\eta-c(\eta)$ such that η satisfies a linear constraint. To this end, set $\mathbb{H}=\{h\in\mathbb{H}_1\colon \int_{\mathscr{Y}}h(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}=0\}$. If $\phi\in\mathbb{H}_1$, then there is a unique function $\eta\in\mathbb{H}$ such that $\phi=\eta-c(\eta)$. Thus the original problem is transformed to the estimation of $\eta\in\mathbb{H}$. The log-likelihood is given by $l(h;\boldsymbol{Y})=h(\boldsymbol{Y})-c(h)$, and the expected log-likelihood is given by $\Lambda(h)=E[l(h;\boldsymbol{Y})]=E[h(\boldsymbol{Y})]-c(h)$.

Assumption 4.1. The density $f_{\boldsymbol{Y}}$ is bounded away from zero and infinity on $\mathscr{Y}$.

In this section, we assume that $\eta^*=\eta$ and that Assumption 4.1 holds or, equivalently, that η is bounded. Thus Condition 2.1 holds. We also take $\boldsymbol{U}=\boldsymbol{W}=\boldsymbol{Y}$, so that Condition 3.1 holds. In addition, we assume that Condition 4.1 holds. We will verify Conditions 2.2, 2.4 and 2.6

Define the empirical inner product as $\langle h_1,h_2\rangle_n=E_n[h_1(\boldsymbol{Y})h_2(\boldsymbol{Y})]$ with corresponding norm $\|h\|_n^2=\langle h,h\rangle_n$. The theoretical inner product and norms are defined as $\langle h_1,h_2\rangle=E[h_1(\boldsymbol{Y})h_2(\boldsymbol{Y})]$ and $\|h\|^2=\langle h,h\rangle$. Let $h_1,h_2\in\mathbb{H}$ be a pair of bounded functions on $\mathscr{Y}$.

Set $h_\alpha = h_1 + \alpha(h_2 - h_1)$ for $0 \leqslant \alpha \leqslant 1$. Then

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\, l(h_\alpha; \boldsymbol{y}) = h_2(\boldsymbol{y}) - h_1(\boldsymbol{y}) - E[h_2(\boldsymbol{Y}_\alpha) - h_1(\boldsymbol{Y}_\alpha)]$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\, l(h_\alpha; \boldsymbol{y}) = -\mathrm{var}[h_2(\boldsymbol{Y}_\alpha) - h_1(\boldsymbol{Y}_\alpha)],$$

where $\boldsymbol{Y}_\alpha$ has the density $f_{\boldsymbol{Y}_\alpha}(\boldsymbol{y}) = \exp(h_\alpha(\boldsymbol{y}) - c(h_\alpha))$.

Verification of Condition 2.2. Part (i) of this condition follows from the Cauchy–Schwarz inequality. To verify part (ii), note that

$$\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2} \Lambda(h_1 + \alpha(h_2 - h_1)) = -\mathrm{var}[h_2(\boldsymbol{Y}_\alpha) - h_1(\boldsymbol{Y}_\alpha)].$$

Since $f_{\boldsymbol{Y}_\alpha}(\boldsymbol{y})$ is bounded away from zero and infinity,

$$\begin{aligned}\mathrm{var}[h_2(\boldsymbol{Y}_\alpha) - h_1(\boldsymbol{Y}_\alpha)] &= \inf_c \int_{\mathscr{Y}} [h_2(\boldsymbol{y}) - h_1(\boldsymbol{y}) - c]^2 f_{\boldsymbol{Y}_\alpha}(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \\ &\asymp \inf_c \frac{1}{|\mathscr{Y}|} \int_{\mathscr{Y}} [h_2(\boldsymbol{y}) - h_1(\boldsymbol{y}) - c]^2\, \mathrm{d}\boldsymbol{y} \\ &= \frac{1}{|\mathscr{Y}|} \int_{\mathscr{Y}} [h_2(\boldsymbol{y}) - h_1(\boldsymbol{y})]^2\, \mathrm{d}\boldsymbol{y};\end{aligned}$$

here, we use the fact that $\int_{\mathscr{Y}} h_2(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \int_{\mathscr{Y}} h_2(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = 0$. Now the density of $\boldsymbol{Y}$ is bounded away from zero and infinity, so the above right side is bounded above and below by multiples of

$$E[(h_2(\boldsymbol{Y}) - h_1(\boldsymbol{Y}))^2] = \|h_2 - h_1\|^2.$$

Verification of Condition 2.4. Note that, for $g \in \mathbb{G}_\gamma$,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \ell(\bar{\eta}_\gamma + \alpha g)\bigg|_{\alpha=0} = E_n \left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\, l(\bar{\eta}_\gamma + \alpha g)\bigg|_{\alpha=0} \right) = E_n[g(\boldsymbol{Y})] - E[g(\bar{\boldsymbol{Y}})],$$

where $\bar{\boldsymbol{Y}}$ has the density $\exp(\bar{\eta}_\gamma(\boldsymbol{y}) - c(\bar{\eta}_\gamma))$. Since $\bar{\eta}_\gamma \in \mathbb{G}_\gamma$ maximizes $\Lambda(g)$ over $g \in \mathbb{G}_\gamma$, we have that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \Lambda(\bar{\eta}_\gamma + \alpha g)\bigg|_{\alpha=0} = 0, \quad g \in \mathbb{G}_\gamma,$$

which implies that $E[g(\boldsymbol{Y})] - E[g(\bar{\boldsymbol{Y}})] = 0$ for $g \in \mathbb{G}_\gamma$. Consequently,

$$\frac{(\mathrm{d}/\mathrm{d}\alpha)\ell(\bar{\eta}_\gamma + \alpha g)|_{\alpha=0}}{\|g\|} = \frac{(E_n - E)[g(\boldsymbol{Y})]}{\|g\|}.$$

Condition 2.4(i) now follows from Lemma 4.3.

Observe that, for $g_1, g_2 \in \mathbb{G}_\gamma$,

$$\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2} \ell(g_1 + \alpha(g_2 - g_1)) = \frac{\mathrm{d}^2}{\mathrm{d}\alpha^2} \Lambda(g_1 + \alpha(g_2 - g_1)).$$

Condition 2.4(ii) now follows from Condition 2.2.

Verification of Condition 2.6. Observe that

$$\ell(\bar{\eta}_\gamma) - \ell(\eta) - [\Lambda(\bar{\eta}_\gamma) - \Lambda(\eta)] = (E_n - E)(\bar{\eta}_\gamma - \eta). \tag{4.10}$$

The first property of Condition 2.6 follows from (4.10) with $\gamma = \gamma^*$, Theorem 2.1, and the consequence of Chebyshev's inequality that

$$(E_n - E)(\bar{\eta}_{\gamma^*} - \eta) = \mathrm{O_P}\left(\frac{\|\bar{\eta}_{\gamma^*} - \eta\|}{\sqrt{n}}\right) = \mathrm{O_P}\left(\frac{\inf_\gamma \rho_{n\gamma}}{\sqrt{n}}\right) = \mathrm{O_P}\left(\inf_\gamma \rho_{n\gamma}^2 + \frac{1}{n}\right).$$

We claim that

$$|(E_n - E)(\bar{\eta}_\gamma - \eta)| = \bar{\mathrm{O}}_{\mathrm{P}}\left((\log^{1/2} n)\left[\|\bar{\eta}_\gamma - \eta\|\left(\frac{N_n}{n}\right)^{1/2} + \frac{N_n}{n}\right]\right) \tag{4.11}$$

uniformly over $\gamma \in \Gamma$. The second property of Condition 2.6 follows from (4.10) and (4.11).

Let us now verify (4.11). Condition 4.1 implies that $N_n^{1/2} \sup_\gamma \rho_{n\gamma} \lesssim \log^{-1/2} n$. Now $\|\bar{\eta}_\gamma - \eta\| \lesssim \sup_\gamma \rho_{n\gamma}$ (uniformly over $\gamma \in \tilde{\Gamma}$) by Theorem 2.1 and $\|\bar{\eta}_\gamma - \eta\|_\infty \lesssim N_n^{1/2} \sup_\gamma \rho_{n\gamma} \lesssim \log^{-1/2} n$ by (3.4). [Choose $g_\gamma^* \in \mathbb{G}_\gamma$ such that $\|g_\gamma^* - \eta\|_\infty = \rho_{n\gamma}$]. Let c be a fixed positive number. It follows from Bernstein's inequality (4.3) that, for c' a sufficiently large positive number,

$$\begin{aligned} &P(|(E_n - E)(\bar{\eta}_\gamma - \eta)| \geqslant c'(\log^{1/2} n)\{\|\bar{\eta}_\gamma - \eta\|(N_n/n)^{1/2} + N_n/n\}) \\ &\quad \leqslant 2\exp(-2cN_n \log n) \end{aligned} \tag{4.12}$$

for $\gamma \in \tilde{\Gamma}$.

Let c be sufficiently large. Then, according to Lemma 3.2, there is a subset $\tilde{\Gamma}_n''$ of $\tilde{\Gamma}$ such that (4.9) holds with $C_3 = c$ and every point $\gamma \in \Gamma$ is within n^{-2} of some point $\tilde{\gamma} \in \tilde{\Gamma}_n''$ such that $\|\bar{\eta}_{\tilde{\gamma}} - \eta\| \leqslant \|\bar{\eta}_\gamma - \eta\|$. Let γ and $\tilde{\gamma}$ be as just described. Then, by Theorem 2.1 and Lemma 3.5,

$$|(E_n - E)(\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}})| \leqslant 2\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\|_\infty \lesssim \frac{N_n^{1/2}}{n} \leqslant \frac{N_n}{n}. \tag{4.13}$$

The desired result (4.11) follows from (4.12), (4.13) and Lemma 4.4. This completes the verification of Condition 2.6.

4.3. Generalized regression

Consider an exponential family of distributions on $\mathbb{R}$ of the form $P(Y \in \mathrm{d}y) = \exp[B(\eta)y - C(\eta)]\Psi(\mathrm{d}y)$, where $B(\cdot)$ is a known, twice continuously differentiable function on $\mathbb{R}$ whose first derivative is strictly positive on $\mathbb{R}$, Ψ is a nonzero measure on $\mathbb{R}$ that is not concentrated at a single point, and $C(\eta) = \log \int_{\mathbb{R}} \exp[(B(\eta)y]\Psi(\mathrm{d}y) < \infty$ for $\eta \in \mathbb{R}$. Observe that $B(\cdot)$ is strictly increasing and $C(\cdot)$ is twice continuously differentiable on $\mathbb{R}$. The mean of the distribution is given by $\mu = A(\eta) = C'(\eta)/B'(\eta)$ for $\eta \in \mathbb{R}$. It follows from the information inequality that $E[B(h)Y - C(h)] = B(h)\mu - C(h)$ is uniquely maximized at $h = \eta$. If $B(\eta) = \eta$ for $\eta \in \mathbb{R}$, then η is referred to as the *canonical parameter* of the exponential family; here $\mu = A(\eta) = C'(\eta)$.

Consider also a random pair $\boldsymbol{W}=(\boldsymbol{X},Y)$, where the random vector $\boldsymbol{X}$ of covariates is $\mathscr{X}$-valued with $\mathscr{X}=\mathscr{U}$ and Y is real-valued. Suppose the conditional distribution of Y given that $\boldsymbol{X}=\boldsymbol{x}\in\mathscr{X}$ has the form

$$P(Y\in \mathrm{d}y|\boldsymbol{X}=\boldsymbol{x})=\exp[B(\eta(\boldsymbol{x}))y-C(\eta(\boldsymbol{x}))]\Psi(\mathrm{d}y). \tag{4.14}$$

Here the function of interest is the response function $\eta(\cdot)$, which specifies the dependence on $\boldsymbol{x}$ of the conditional distribution of the response Y given that the value of the vector $\boldsymbol{X}$ of covariates equals $\boldsymbol{x}$. The mean of this conditional distribution is given by

$$\mu(\boldsymbol{x})=E(Y|\boldsymbol{X}=\boldsymbol{x})=A(\eta(\boldsymbol{x})),\quad \boldsymbol{x}\in\mathscr{X}. \tag{4.15}$$

The (conditional) log-likelihood is given by

$$l(h,\boldsymbol{X},Y)=B(h(\boldsymbol{X}))Y-C(h(\boldsymbol{X})),$$

and its expected value is given by

$$\Lambda(h)=E[B(h(\boldsymbol{X}))\mu(\boldsymbol{X})-C(h(\boldsymbol{X}))],$$

which is essentially uniquely maximized at $h=\eta$. This property of the response function depends only on (4.15), not on the stronger assumption (4.14). In the application of the theory developed in this paper to generalized regression, we require (4.15), but not (4.14).

When the underlying exponential family is the Bernoulli distribution with parameter π and canonical parameter $\eta=\mathrm{logit}(\pi)$, we get logistic regression. Here $\mu(\boldsymbol{x})=\pi(\boldsymbol{x})=P(Y=1|\boldsymbol{X}=\boldsymbol{x})$ and $\eta(\boldsymbol{x})=\mathrm{logit}(\pi(\boldsymbol{x}))=\mathrm{logit}(\mu(\boldsymbol{x}))$. When the underlying exponential family is the Poisson distribution with parameter λ and canonical parameter $\eta=\log\lambda$, we get Poisson regression. Here $\mu(x)=\lambda(x)$ and $\eta(\boldsymbol{x})=\log\lambda(x)$. When the underlying exponential family is the normal distribution with canonical parameter $\eta=\mu$ and known variance, we get ordinary regression as discussed above.

In this subsection we verify the technical conditions required in Proposition 2.1 under five auxiliary assumptions.

Assumption 4.2. $B(\cdot)$ is twice continuously differentiable and its first derivative $B'(\cdot)$ is strictly positive on $\mathbb{R}$. There is a subinterval S of R such that Ψ is concentrated on S and

$$B''(\xi)y-C''(\xi)<0,\quad -\infty<\xi<\infty, \tag{4.16}$$

for all $y\in\overset{\circ}{S}$, where $\overset{\circ}{S}$ denotes the interior of S. If S is bounded, (4.16) holds for at least one of its endpoints.

Note that $A(\eta)\in\overset{\circ}{S}$ for $-\infty<\eta<\infty$. Thus by Assumption 4.2,

$$B''(\xi)A(\eta)-C''(\xi)<0,\quad -\infty<\xi,\eta<\infty. \tag{4.17}$$

If η is the canonical parameter of the exponential family, then $B(\eta)=\eta$ and hence $B''(\xi)=0$ and $C''(\xi)>0$ for $-\infty<\xi<\infty$, so Assumption 4.2 automatically holds with $S=\mathbb{R}$. This assumption is satisfied by many familiar exponential families, including normal, binomial-probit, binomial-logit, Poisson, gamma, geometric and negative binomial distributions; see Stone (1986).

Assumption 4.3. $P(Y\in S)=1$ and $E(Y|\boldsymbol{X}=\boldsymbol{x})=A(\eta(\boldsymbol{x}))$ for $\boldsymbol{x}\in\mathscr{X}$.

Observe that Assumption 4.3 is implied by the stronger assumption that the conditional distribution of Y given that $\boldsymbol{X}=\boldsymbol{x}$ has the exponential family form given by (4.14).

Assumption 4.4. The response function $\eta(\cdot)$ is bounded.

Assumption 4.5. There are positive constants M_1 and M_2 such that $E[\mathrm{e}^{|Y-\mu(\boldsymbol{X})|/M_1}|\boldsymbol{X}=\boldsymbol{x}]\leqslant M_2$ for $\boldsymbol{x}\in\mathscr{X}$.

It follows from Assumption 4.5 that there is a positive constant D such that $\operatorname{var}(Y|\boldsymbol{X}=\boldsymbol{x})\leqslant D$ for $\boldsymbol{x}\in\mathscr{X}$.

Assumption 4.6. The distribution of $\boldsymbol{X}$ is absolutely continuous and its density function $f_{\boldsymbol{X}}$ is bounded away from zero and infinity on $\mathscr{X}$.

Throughout this section we assume that $\eta^*=\eta$ and that Assumptions 4.2–4.6 hold. Now η is bounded by Assumption 4.4, so Condition 2.1 holds. We take $\boldsymbol{W}=(\boldsymbol{X},Y)$ and $\boldsymbol{U}=\boldsymbol{X}$, so Condition 3.1 follows from Assumption 4.6. We also assume that Condition 4.1 holds. We will verify Conditions 2.2, 2.4 and 2.6.

Define the empirical inner product as $\langle h_1,h_2\rangle_n=E_n[h_1(\boldsymbol{X})h_2(\boldsymbol{X})]$ with corresponding norm $\|h\|_n^2=\langle h,h\rangle_n$. The theoretical inner product and norms are defined as $\langle h_1,h_2\rangle=E[h_1(\boldsymbol{X})h_2(\boldsymbol{X})]$ and $\|h\|^2=\langle h,h\rangle$. Recall that the log-likelihood based on the random sample and its expected value are given by $\ell(h)=E_n[B(h)Y-C(h)]$ and $\Lambda(h)=E[B(h)Y-C(h)]$.

Let $(\boldsymbol{X}_1,Y_1),\ldots,(\boldsymbol{X}_n,Y_n)$ be a random sample of size n from the joint distribution of $\boldsymbol{X}$ and Y. Choose $M_1'\in(M_1,\infty)$. It follows from Assumption 4.5 that $P(|Y-\mu(\boldsymbol{X})|\geqslant M_1'\log n)\leqslant M_2 n^{-M_1'/M_1}$ and hence that

$$\lim_n P\left(\max_{1\leqslant i\leqslant n}|Y_i-\mu(\boldsymbol{X}_i)|\geqslant M_1'\log n\right)=0. \tag{4.18}$$

Moreover, by the power series expansion of the exponential function, for $m\geqslant 2$ and $1\leqslant i\leqslant n$,

$$E[|Y_i-\mu(\boldsymbol{X}_i)|^m|\boldsymbol{X}_i]\leqslant\frac{m!}{2}(2M_1^2M_2)M_1^{m-2}. \tag{4.19}$$

Thus, by Bernstein's inequality (4.2), if h is a bounded function on $\mathscr{X}$, then

$$P(|\langle h,Y-\mu\rangle_n|\geqslant t|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)\leqslant 2\exp\left(-\frac{nt^2}{2M_1(2M_1M_2\|h\|_n^2+t\|h\|_\infty)}\right) \tag{4.20}$$

for $t>0$.

Verification of Condition 2.2. Observe that

$$\left|\frac{\mathrm{d}}{\mathrm{d}\alpha}\Lambda(h_1+\alpha h_2)\bigg|_{\alpha=0}\right| = E(h_2(\boldsymbol{X})\{B'(h_1(\boldsymbol{X}))\mu(\boldsymbol{X}) - C'(h_1(\boldsymbol{X}))\},$$

where $\mu(\boldsymbol{x}) = E(Y|\boldsymbol{X}=\boldsymbol{x})$. By Assumptions 4.2–4.4, $\mu(\cdot)$ is bounded. Since $B'(\cdot)$ and $C'(\cdot)$ are continuous, they are bounded on finite intervals. Condition 2.2(i) then follows from the Cauchy–Schwarz inequality. Let $h_1, h_2 \in \mathbb{H}$ be a pair of bounded functions on $\mathscr{Y}$. Set $h_\alpha = h_1 + \alpha(h_2 - h_1)$ for $0 \leqslant \alpha \leqslant 1$. Then

$$\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\Lambda(h_\alpha) = E\{(h_2(\boldsymbol{X}) - h_1(\boldsymbol{X}))^2[B''(h_\alpha(\boldsymbol{X}))A(\eta(\boldsymbol{X})) - C''(h_\alpha(\boldsymbol{X}))]\}.$$

Condition 2.2(ii) now follows from (4.17), the boundedness of $\eta(\cdot)$, and the continuity of $A(\cdot)$, $B''(\cdot)$, and $C''(\cdot)$.

Verification of Condition 2.4.

Lemma 4.5. *Suppose* $\lim_n N_n^2/n = 0$. *Then Condition* 2.4(ii) *holds.*

Proof. The desired result follows from Lemma 4.2 and the argument used to prove Lemma 4.3 of Huang (1998b). The requirement that $\lim_n N_n^2/n = 0$ ensures the applicability of Lemma 4.2. □

Lemma 4.6. *Suppose* $\lim_n N_n^2/n = 0$ and $\sup_{\gamma\in\tilde{\Gamma}} \rho_{n\gamma} = \mathrm{O}(N_n^{-c})$ *for some* $c > 1/2$. *Then Condition* 2.4(i) *holds.*

Proof. In this proof, set $\bar{\rho}_n = \sup_{\gamma\in\tilde{\Gamma}} \rho_{n\gamma}$. By Theorem 2.1 applied to $\tilde{\Gamma}$ there is a positive constant K_1 such that, for n sufficiently large, $\bar{\eta}_\gamma$ exists uniquely and $\|\bar{\eta}_\gamma\|_\infty \leqslant K_1$ for $\gamma \in \tilde{\Gamma}$. Let $\gamma \in \tilde{\Gamma}$ and $g \in \mathbb{G}_\gamma$. Then

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\ell(\bar{\eta}_\gamma + \alpha g)\bigg|_{\alpha=0} = E_n[gD(\bar{\eta}_\gamma)] + E_n[gB'(\bar{\eta}_\gamma)(Y-\mu)],$$

where $D(\bar{\eta}_\gamma) = B'(\bar{\eta}_\gamma)\mu - C'(\bar{\eta}_\gamma)$ and $E[gD(\bar{\eta}_\gamma)] = 0$.

Let $0 < \delta \leqslant 1/4$. Since $N_n^{1/2}\bar{\rho}_n \lesssim N_n^{-(c-1/2)}$ for some $c > 1/2$ by Condition 4.1, there is an $\varepsilon \in (0, \delta^2)$ and there is a fixed positive number c_1 such that, for n sufficiently large,

$$N_n^{1/2}\bar{\rho}_n \leqslant c_1(N_n^{1/2})^{-(\log 1/\delta)/(\log \delta/\varepsilon^{1/2})}$$

and hence

$$\min(c_1^{-1}N_n^{1/2}\bar{\rho}_n, N_n^{1/2}\varepsilon^{(k-1)/2}) \leqslant \delta^{k-1} \tag{4.21}$$

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

Let Ω_n, $\lim_n P(\Omega_n) = 1$, be an event that depends only on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and is such that the statements in Lemma 3.4 and Lemma 3.5 hold.

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 Lemma 3.5, $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\| \leqslant c_2\varepsilon^{(k-1)/2}$, $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\|_n \leqslant c_2\varepsilon^{(k-1)/2}$ on Ω_n, and $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\|_\infty$

$\leqslant c_3 N_n^{1/2} \varepsilon^{(k-1)/2}$ (for some fixed positive constants c_2, c_3). Let $\mathbb{B}_{\tilde{\gamma},k}$ be as in Section 3 and let $g \in \mathbb{B}_\gamma$. Then, by Lemmas 3.3 and 3.4 (see the proof of Lemma 4.2), there is a $\tilde{g} \in \mathbb{B}_{\tilde{\gamma},k-1}$ such that $\|g - \tilde{g}\| \leqslant c_4 \varepsilon^{k-1}$, $\|g - \tilde{g}\|_n \leqslant c_5 \varepsilon^{k-1}$ on Ω_n, and $\|g - \tilde{g}\|_\infty \leqslant c_6 N_n^{1/2} \varepsilon^{k-1}$. Now

$$gB'(\bar{\eta}_\gamma) - \tilde{g}B'(\bar{\eta}_{\tilde{\gamma}}) = (g - \tilde{g})B'(\bar{\eta}_\gamma) + \tilde{g}[B'(\bar{\eta}_\gamma) - B'(\bar{\eta}_{\tilde{\gamma}})]. \tag{4.22}$$

Observe that $\|(g-\tilde{g})B'(\bar{\eta}_\gamma)\|_n \leqslant c_7 \varepsilon^{k-1}$ on Ω_n and $\|(g-\tilde{g})B'(\bar{\eta}_\gamma)\|_\infty \leqslant c_7 N_n^{1/2} \varepsilon^{k-1}$. Observe also that, $\|\tilde{g}[B'(\bar{\eta}_\gamma) - B'(\bar{\eta}_{\tilde{\gamma}})]\|_n \leqslant c_8 N_n^{1/2} \varepsilon^{(k-1)/2}$ on Ω_n and $\|\tilde{g}[B'(\bar{\eta}_\gamma) - B'(\bar{\eta}_{\tilde{\gamma}})]\|_\infty \leqslant c_8 N_n \varepsilon^{(k-1)/2}$. Consequently, $\|gB'(\bar{\eta}_\gamma) - \tilde{g}B'(\bar{\eta}_{\tilde{\gamma}})\|_n \leqslant c_9 N_n^{1/2} \varepsilon^{(k-1)/2}$ on Ω_n and $\|gB'(\bar{\eta}_\gamma) - \tilde{g}B'(\bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant c_9 N_n \varepsilon^{(k-1)/2}$. By the same argument, c_9 can be chosen so that, in addition, $\|gD(\bar{\eta}_\gamma) - \tilde{g}D(\bar{\eta}_{\tilde{\gamma}})\| \leqslant c_9 N_n^{1/2} \varepsilon^{(k-1)/2}$ and $\|gD(\bar{\eta}_\gamma) - \tilde{g}D(\bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant c_9 N_n \varepsilon^{(k-1)/2}$.

Alternatively, by Theorem 2.1 and Lemma 4.2,

$$\sup_{\gamma \in \tilde{\Gamma}} \frac{\|\bar{\eta}_\gamma - \eta\|}{\rho_{n\gamma}} = \mathrm{O}(1) \quad \text{and} \quad \sup_{\gamma \in \tilde{\Gamma}} \frac{\|\bar{\eta}_\gamma - \eta\|_n}{\rho_{n\gamma}} = \mathrm{O}(1)[1 + \mathrm{o}_P(1)].$$

(Choose $g^* \in \mathbb{G}_\gamma$ such that $\|g^* - \eta\|_\infty = \rho_{n\gamma}$.) Consequently, for n sufficiently large, $\|\bar{\eta}_\gamma - \eta\| \leqslant c_{10} \bar{\rho}_n$ and $\|\bar{\eta}_\gamma - \eta\|_n \leqslant c_{10} \bar{\rho}_n$ on Ω_n for $\gamma \in \tilde{\Gamma}$ (provided that Ω_n is suitably chosen).

Given $\gamma, \tilde{\gamma} \in \tilde{\Gamma}$, we have that $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\| \leqslant 2c_{10} \bar{\rho}_n$ and $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\|_n \leqslant 2c_{10} \bar{\rho}_n$ on Ω_n. Choose $\eta'_\gamma \in \mathbb{G}_\gamma$ and $\eta'_{\tilde{\gamma}} \in \mathbb{G}_{\tilde{\gamma}}$ such that $\|\eta'_\gamma - \eta\|_\infty \leqslant \bar{\rho}_n$ and $\|\eta'_{\tilde{\gamma}} - \eta\|_\infty \leqslant \bar{\rho}_n$. It follows from the triangle inequality and (3.4) that $\|\eta'_\gamma - \bar{\eta}_\gamma\|_\infty \leqslant M(c_{10}+1) N_n^{1/2} \bar{\rho}_n$. Thus $\|\bar{\eta}_\gamma - \eta\|_\infty \leqslant [M(c_{10}+1) N_n^{1/2} + 1]\bar{\rho}_n$. Similarly, $\|\bar{\eta}_{\tilde{\gamma}} - \eta\|_\infty \leqslant [M(c_{10}+1) N_n^{1/2} + 1]\bar{\rho}_n$. Hence $\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\|_\infty \leqslant 2[M(c_{10}+1) N_n^{1/2} + 1]\bar{\rho}_n$.

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. Then [recall (3.4), (4.21), and (4.22)], $\|gB'(\bar{\eta}_\gamma) - \tilde{g}B'(\bar{\eta}_{\tilde{\gamma}})\|_n \leqslant c_{11} \delta^{k-1}$ on Ω_n, $\|gB'(\bar{\eta}_\gamma) - \tilde{g}B'(\bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant c_{11} N_n^{1/2} \delta^{k-1}$, $\|gD(\bar{\eta}_\gamma) - \tilde{g}D(\bar{\eta}_{\tilde{\gamma}})\| \leqslant c_{11} \delta^{k-1}$, and $\|gD(\bar{\eta}_\gamma) - \tilde{g}D(\bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant c_{11} N_n^{1/2} \delta^{k-1}$.

Let $K = K_n$ be a positive integer satisfying the two inequalities specified in the next paragraph, and let Ξ_k, $\mathbb{B}_{\gamma k}$ for $\gamma \in \tilde{\Gamma}$, and $\mathbb{B}_k$, $0 \leqslant k \leqslant K$, be as in Lemma 3.1 and the following paragraph with the current value of ε. We will apply Lemma 4.1 with $s = (\gamma, g)$, $V_s = E_n\{g[D(\bar{\eta}_\gamma) + B'(\bar{\eta}_\gamma)(Y - \mu)]\}$, $\mathbb{S} = \{(\gamma, g)\colon \gamma \in \Gamma \text{ and } g \in \mathbb{B}_\gamma\}$, and $\mathbb{S}_k = \{(\gamma, g)\colon \gamma \in \Xi_k \text{ and } g \in \mathbb{B}_{\gamma k}\}$. Now $\#(\mathbb{S}_k) \leqslant (M' \varepsilon^{-2k})^{N_n}$ for $1 \leqslant k \leqslant K$ by (3.3), so (4.5) holds with $C_3 = 1$ and any $C_4 \geqslant 2 \log(M' \varepsilon^{-1}) N_n$.

Let Ω_{n0} denote the event that $\max_{1 \leqslant i \leqslant n} |Y_i - \mu(\boldsymbol{X}_i)| \leqslant M'_1 \log n$ with M'_1 as in (4.18). Then $\lim_n P(\Omega_{n0}) = 1$. Choose $\gamma \in \Gamma$ and $g \in \mathbb{B}_\gamma$. Let $\tilde{\gamma} \in \Xi_K$ be such that $\zeta(\gamma, \tilde{\gamma}) \leqslant \varepsilon^K$. Then there is a $\tilde{g} \in \mathbb{B}_{\tilde{\gamma} K}$ such that $\|gD(\bar{\eta}_\gamma) - \tilde{g}D(\bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant c_{11} N_n^{1/2} \delta^K$ and $\|gB'(\bar{\eta}_\gamma) - \tilde{g}B'(\bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant c_{11} N_n^{1/2} \delta^K$. Thus $\|gD(\bar{\eta}_\gamma) - \tilde{g}D(\bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant (N_n/n)^{1/2}$ provided that K satisfies the inequality $c_{11} \delta^K \leqslant n^{-1/2}$ and $|E_n\{[gB'(\bar{\eta}_\gamma) - \tilde{g}B'(\bar{\eta}_{\tilde{\gamma}})](Y - \mu)\}| \leqslant (N_n/n)^{1/2}$ on Ω_{n0} provided that K satisfies the inequality $M'_1 c_{11} \delta^K \leqslant 1/(n^{1/2} \log n)$. Let K satisfy both inequalities. Then (4.4) holds with $C_1 = 2(N_n/n)^{1/2}$, $C_2 = 0$, and $\Omega = \Omega_{n0}$.

Let $1 \leqslant k \leqslant K$. 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 $\zeta(\gamma, \tilde{\gamma}) \leqslant \varepsilon^{k-1}$, $\|gB'(\bar{\eta}_\gamma) - \tilde{g}B'(\bar{\eta}_{\tilde{\gamma}})\|_n \leqslant c_{11} \delta^{k-1}$ on Ω_n, $\|gB'(\bar{\eta}_\gamma) - \tilde{g}B'(\bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant c_{11} N_n^{1/2} \delta^{k-1}$, $\|gD(\bar{\eta}_\gamma) - \tilde{g}D(\bar{\eta}_{\tilde{\gamma}})\| \leqslant c_{11} \delta^{k-1}$, and $\|gD(\bar{\eta}_\gamma) - \tilde{g}D(\bar{\eta}_{\tilde{\gamma}})\|_\infty \leqslant c_{11} N_n^{1/2} \delta^{k-1}$.

Write $s=(\gamma,g)$ and $V_s=V_{1s}+V_{2s}$, where $V_{1s}=E_n[gD(\bar\eta_\gamma)]$ and $V_{2s}=E_n[gB'(\bar\eta_\gamma)(Y-\mu)]$. Similarly, write $\tilde s=(\tilde\gamma,\tilde g)$ and $V_{\tilde s}=V_{1\tilde s}+V_{2\tilde s}$, where $V_{1\tilde s}=E_n[\tilde gD(\bar\eta_{\tilde\gamma})]$ and $V_{2\tilde s}=E_n[\tilde gB'(\bar\eta_{\tilde\gamma})(Y-\mu)]$. Observe that $V_{1s}-V_{1\tilde s}=(E_n-E)[gD(\bar\eta_\gamma)-\tilde gD(\bar\eta_{\tilde\gamma})]$. Since $0<2\delta\leqslant 1$, it follows from Bernstein's inequality (4.2) that, for $C>0$,

$$P(|V_{1s}-V_{1\tilde s}|\geqslant C2^{-(k-1)}(N_n/n)^{1/2})\leqslant 2\exp\left(-\frac{C^2(2\delta)^{-(k-1)}N_n}{2c_{11}(c_{11}+Cn^{-1/2}N_n)}\right).$$

Similarly, $V_{2s}-V_{2\tilde s}=E_n\{[gB'(\bar\eta_\gamma)-\tilde gB'(\bar\eta_{\tilde\gamma})](Y-\mu)\}$, so it follows from (4.20) that

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

provided that c_{11} is sufficiently large. Hence

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

so (4.6) holds with

$$C_4=\frac{C^2N_n}{4c_{11}(c_{11}+Cn^{-1/2}N_n)}\geqslant 2\log(M'\varepsilon^{-1})N_n$$

for C sufficiently large, $C_5=2C(N_n/n)^{1/2}$, $C_6=4$, and $\Omega=\Omega_n$. Consequently, by Lemma 4.1,

$$P\left(\sup_{\gamma\in\Gamma}\sup_{g\in\mathbb{B}_\gamma}\left|\frac{\mathrm{d}}{\mathrm{d}\alpha}\ell(\bar\eta_\gamma+\alpha g)\bigg|_{\alpha=0}\right|\geqslant 2(1+2C)(N_n/n)^{1/2}\right)$$
$$\leqslant\frac{16c_{11}(c_{11}+Cn^{-1/2}N_n)}{C^2N_n}+P((\Omega_n\cap\Omega_{n0})^c),$$

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

Verification of Condition 2.6. It follows from (4.19) and Bernstein's inequality (4.3) (with $H=M_1A$) that if h is a bounded function on $\mathscr{X}$ and $A\geqslant\|h\|_\infty$, then

$$P(|E_n\{h(Y-\mu)\}|\geqslant tM_1^{-1}A^{-1}[(2M_1^2M_2)^{1/2}\|h\|_n(N_n/n)^{1/2}+N_n/n]|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$$
$$\leqslant 2\exp\left(-\frac{tM_1^{-2}A^{-2}N_n}{2}\right)\tag{4.23}$$

for $t\geqslant 1$.

Observe that

$$\ell(\bar\eta_\gamma)-\ell(\eta)-[\Lambda(\bar\eta_\gamma)-\Lambda(\eta)]$$
$$=(E_n-E)\{[B(\bar\eta_\gamma)-B(\eta)]\mu-[C(\bar\eta_\gamma)-C(\eta)]\}$$
$$+E_n\{[B(\bar\eta_\gamma)-B(\eta)](Y-\mu)\}.\tag{4.24}$$

Lemma 4.7. *Suppose Condition* 4.1 *holds. Then*

$$(E_n - E)\{[B(\bar{\eta}_\gamma) - B(\eta)]\mu - [C(\bar{\eta}_\gamma) - C(\eta)]\}$$
$$= \bar{\mathrm{O}}_{\mathrm{P}}\left((\log^{1/2} n)\left[\|\bar{\eta}_\gamma - \eta\|\left(\frac{N_n}{n}\right)^{1/2} + \frac{N_n}{n}\right]\right)$$

uniformly over $\gamma \in \Gamma$.

Proof. The proof of this result is similar to that of Condition 2.6(ii) in the density estimation context. □

Lemma 4.8. *Suppose Condition* 4.1 *holds. Then*

$$|E_n\{[B(\bar{\eta}_\gamma) - B(\eta)](Y - \mu)\}| = \bar{\mathrm{O}}_{\mathrm{P}}\left((\log^{1/2} n)\left[\|\bar{\eta}_\gamma - \eta\|\left(\frac{N_n}{n}\right)^{1/2} + \frac{N_n}{n}\right]\right)$$

uniformly over $\gamma \in \Gamma$.

Proof. Note that $\|\bar{\eta}_\gamma - \eta\| \lesssim \sup_{\gamma\in\tilde{\Gamma}} \rho_{n\gamma}$ and $\|\bar{\eta}_\gamma - \eta\|_\infty \lesssim \log^{-1/2} n$ uniformly over $\gamma \in \tilde{\Gamma}$ (see the arguments in Section 4.2). Set $h_\gamma = B(\bar{\eta}_\gamma) - B(\eta)$ for $\gamma \in \tilde{\Gamma}$. Then $\|h_\gamma\| \lesssim \|\bar{\eta}_\gamma - \eta\|$, $\|h_\gamma^2\| \lesssim (\log^{-1/2} n)\|\bar{\eta}_\gamma - \eta\|$ and $\|h_\gamma^2\|_\infty \lesssim \log^{-1} n$ uniformly over $\gamma \in \tilde{\Gamma}$. Let c_1 be a fixed positive number. It now follows from Bernstein's inequality (4.2) (note that $\|h_\gamma\|_n^2 = E_n(h_\gamma^2)$) that, for c_2 a sufficiently large positive number,

$$P\left(\|h_\gamma\|_n^2 - \|h_\gamma\|^2 \geqslant c_2^2\left[\|\bar{\eta}_\gamma - \eta\|\left(\frac{N_n}{n}\right)^{1/2} + \frac{N_n}{n}\right]\right) \leqslant 2\exp(-2c_1 N_n \log n)$$

for $\gamma \in \tilde{\Gamma}$ and hence that, for c_2 a sufficiently large positive number,

$$P(\Omega_{n\gamma}^c) \leqslant 2\exp(-2c_1 N_n \log n), \quad \gamma \in \tilde{\Gamma},$$

where $\Omega_{n\gamma}$ denotes the event that $\|h_\gamma\|_n \leqslant c_2[\|\bar{\eta}_\gamma - \eta\| + (N_n/n)^{1/2}]$. It follows from (4.23) that, for a sufficiently large positive number c_3,

$$P(|E_n\{h_\gamma(Y - \mu)\}| \geqslant c_3(\log^{1/2} n)[\|\bar{\eta}_\gamma - \eta\|(N_n/n)^{1/2} + N_n/n]|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$$
$$\leqslant 2\exp(-2c_1 N_n \log n)$$

on $\Omega_{n\gamma}$ for $\gamma \in \tilde{\Gamma}$ and hence that

$$P(|E_n\{h_\gamma(Y - \mu)\}| \geqslant c_3(\log^{1/2} n)[\|\bar{\eta}_\gamma - \eta\|(N_n/n)^{1/2} + N_n/n])$$
$$\leqslant 4\exp(-2c_1 N_n \log n) \tag{4.25}$$

for $\gamma \in \tilde{\Gamma}$.

Let c_1 be sufficiently large. Then, according to Lemma 3.2, there is a subset $\tilde{\Gamma}_n''$ of $\tilde{\Gamma}$ such that (4.9) holds with $C_3 = c_1$ and every point $\gamma \in \Gamma$ is within n^{-3} of some point $\tilde{\gamma} \in \tilde{\Gamma}_n''$ such that $\|\bar{\eta}_{\tilde{\gamma}} - \eta\| \leqslant \|\bar{\eta}_\gamma - \eta\|$. Let γ and $\tilde{\gamma}$ be as just described.

By Lemma 3.5,

$$|E_n\{[B(\bar\eta_\gamma) - B(\bar\eta_{\tilde\gamma})](Y-\mu)\}| \lesssim \|\bar\eta_\gamma - \bar\eta_{\tilde\gamma}\|_\infty \max_{1\leqslant i\leqslant n} |Y_i - \mu(\boldsymbol{X}_i)|$$

$$\lesssim N_n^{1/2} n^{-3/2}\log n \lesssim N_n/n \tag{4.26}$$

provided that $|Y_i - \mu(\boldsymbol{X}_i)| \leqslant M_1' \log n$ for $1 \leqslant i \leqslant n$.

The desired result follows from (4.9), (4.18), (4.25), (4.26), and Lemma 4.4.

Lemma 4.9. *Suppose Condition* 4.1 *holds. Then Condition* 2.6 *holds.*

Proof. Now $E(E_n\{[B(\bar\eta_{\gamma^*}) - B(\eta)](Y-\mu)\}|\boldsymbol{X}_1,\dots,\boldsymbol{X}_n) = 0$ and

$$\operatorname{var}(E_n\{[B(\bar\eta_{\gamma^*}) - B(\eta)](Y-\mu)\}|\boldsymbol{X}_1,\dots,\boldsymbol{X}_n) = \mathrm{O}\left(\frac{\|\bar\eta_{\gamma^*}-\eta\|_n^2}{n}\right),$$

so

$$E[(E_n\{[B(\bar\eta_{\gamma^*}) - B(\eta)](Y-\mu)\})^2] = \mathrm{O_P}\left(\frac{\|\bar\eta_{\gamma^*}-\eta\|^2}{n}\right).$$

Since $\|\bar\eta_{\gamma^*}-\eta\| = \inf_{\gamma\in\tilde\Gamma}\rho_{n\gamma}$, it follows from Chebyshev's inequality that

$$E_n\{[B(\bar\eta_{\gamma^*}) - B(\eta)](Y-\mu)\} = \mathrm{O_P}\left(\frac{\inf_{\gamma\in\tilde\Gamma}\rho_{n\gamma}}{\sqrt{n}}\right).$$

Similarly,

$$(E_n - E)\{[B(\bar\eta_{\gamma^*}) - B(\eta)]\mu - [C(\bar\eta_{\gamma^*}) - C(\eta)]\} = \mathrm{O_P}\left(\frac{\inf_{\gamma\in\tilde\Gamma}\rho_{n\gamma}}{\sqrt{n}}\right).$$

The first property of Condition 2.6 now follows from (4.24) with $\gamma=\gamma^*$. The second property follows from (4.24) and Lemmas 4.7 and 4.8. □

Ordinary regression: The framework of generalized regression, as considered above, includes ordinary regression as a special case. Specifically, let $B(\eta)=2\eta$ for $\eta\in\mathbb{R}$ and $\Psi(\mathrm{d}y)=\pi^{-1/2}\mathrm{e}^{-y^2}\mathrm{d}y$ for $y\in\mathbb{R}$. Then $S=\mathbb{R}$. Also, $C(\eta)=\eta^2$ and $A(\eta)=\eta$ for $\eta\in\mathbb{R}$, so the regression function μ equals the response function η. Suppose that Y has finite second moment. The pseudo-log-likelihood and its expectation are given, respectively, by $l(h;\boldsymbol{X},Y)=2h(\boldsymbol{X})Y-h^2(\boldsymbol{X})=-[Y-h(\boldsymbol{X})]^2+Y^2$ and $\Lambda(h)=-E\{[Y-h(\boldsymbol{X})]^2\}+E(Y^2)$. Assumption 4.4 is that the regression function is bounded. Let h_1 and h_2 be bounded functions on $\mathscr{X}$. Then

$$\left.\frac{\mathrm{d}}{\mathrm{d}\alpha}\Lambda(h_1+\alpha h_2)\right|_{\alpha=0} = 2E\{h_2(\boldsymbol{X})[\mu(\boldsymbol{X})-h_1(\boldsymbol{X})]\}$$

and

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

so Condition 2.2 follows from the boundedness of the regression function and of the density function of $\boldsymbol{X}$. Also,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\ell(\bar{\mu}_\gamma + \alpha g)\bigg|_{\alpha=0} = 2E_n\{g[Y - \bar{\mu}_\gamma(\boldsymbol{X})]\}$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\ell(g_1 + \alpha(g_2 - g_1)) = -2\|g_2 - g_1\|_n^2.$$

Thus Condition 2.4(ii) follows from Lemma 4.2, while Condition 2.4(i) requires Lemma 4.6 for its verification.

Acknowledgements

The first author's research was supported in part by NSF grants DMS-9504463 and DMS-9802071. The authors wish to thank Larry Schumaker for pointing out the relevance of Theorem 4.27 of his book to the verification of (A.20) in Appendix A.

Appendix A. Proofs of lemmas in Section 3

In this appendix we prove Lemmas 3.1–3.5.

Consider a free knot sequence $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)$ such that $a < \gamma_1 \leqslant \cdots \leqslant \gamma_J < b$ and

$$\frac{\gamma_{j_2-1} - \gamma_{j_2-m}}{\gamma_{j_1-1} - \gamma_{j_1-m}} \leqslant \bar{M}, \quad 2 \leqslant j_1, j_2 \leqslant J + m, \tag{A.1}$$

where $\gamma_{1-m} = \cdots = \gamma_0 = a$ and $\gamma_{J+1} = \cdots = \gamma_{J+m} = b$.

Observe that

$$\sum_{j=1}^{J+m}(\gamma_{j-1} - \gamma_{j-m}) = (m-1)(b-a).$$

Thus it follows from (A.1) that

$$\gamma_{j-1} - \gamma_{j-m} \geqslant \frac{(m-1)(b-a)}{\bar{M}(J+m)}, \quad 2 \leqslant j \leqslant J + m. \tag{A.2}$$

The requirement (A.1) is stronger than the bound on the global mesh ratio of $\boldsymbol{\gamma}$ that was considered by de Boor (1976). To see this, let $\boldsymbol{\gamma} \in \Gamma$ and note that $\gamma_1 - \gamma_{1-m} = \gamma_1 - \gamma_{2-m}$, $\gamma_{J+m} - \gamma_J = \gamma_{J+m-1} - \gamma_J$, and

$$\frac{\gamma_{j_2} - \gamma_{j_2-m}}{\gamma_{j_1} - \gamma_{j_1-m}} \leqslant \frac{\gamma_{j_2-1} - \gamma_{j_2-m} + \gamma_{j_2} - \gamma_{j_2+1-m}}{(\gamma_{j_1-1} - \gamma_{j_1-m})/2 + (\gamma_{j_1} - \gamma_{j_1+1-m})/2}$$

for $1 \leqslant j_1, j_2 \leqslant J + m$ (the numerator is increased and the denominator is decreased), so it follows from (A.1) that

$$\frac{\gamma_{j_2} - \gamma_{j_2-m}}{\gamma_{j_1} - \gamma_{j_1-m}} \leqslant 2\bar{M}, \quad 1 \leqslant j_1, j_2 \leqslant J + m. \tag{A.3}$$

Observe that $\sum_{j=1}^{J+m}(\gamma_j - \gamma_{j-m}) = m(b-a)$. Thus it follows from (A.3) that

$$\gamma_j - \gamma_{j-m} \geqslant \frac{m(b-a)}{2\bar{M}(J+m)}, \quad 1 \leqslant j \leqslant J+m, \tag{A.4}$$

and

$$\gamma_j - \gamma_{j-m} \leqslant \frac{2\bar{M}m(b-a)}{J+m}, \quad 1 \leqslant j \leqslant J+m, \tag{A.5}$$

Proof of Lemma 3.1. We first verify this result when $L=1$, $J=J_1 \geqslant 1$, $\gamma_j = \gamma_{1j}$, $\boldsymbol{\gamma} = \boldsymbol{\gamma}_1$, $\mathscr{U} = \mathscr{U}_1 = [a,b] = [a_1,b_1]$, $m = m_1 \geqslant 2$, and $N_n = N_{n1} = J+m$. Here the metric ζ is given by $\zeta(\boldsymbol{\gamma},\tilde{\boldsymbol{\gamma}}) = 9\bar{M}N_n|\boldsymbol{\gamma}-\tilde{\boldsymbol{\gamma}}|_\infty/(b-a)$. Let $0 \leqslant \varepsilon_1 \leqslant 2$, let $\boldsymbol{\gamma} \in \Gamma$, and let $\tilde{\boldsymbol{\gamma}}$ be a free knot sequence such that $\zeta(\boldsymbol{\gamma},\tilde{\boldsymbol{\gamma}}) \leqslant \varepsilon_1$ and hence

$$2|\boldsymbol{\gamma}-\tilde{\boldsymbol{\gamma}}|_\infty \leqslant \frac{\varepsilon_1(b-a)}{4\bar{M}N_n}.$$

Thus, by (A.2), $\tilde{\boldsymbol{\gamma}}$ satisfies (A.1) with $\bar{M}$ replaced by

$$\bar{M}\,\frac{m-1+\varepsilon_1/4}{m-1-\varepsilon_1/4} \leqslant 3\bar{M},$$

so $\tilde{\boldsymbol{\gamma}} \in \tilde{\Gamma}$. Let $\tilde{\Gamma}_{\varepsilon_1}$ denote the collection of all such free knot sequences $\tilde{\boldsymbol{\gamma}}$ as $\boldsymbol{\gamma}$ ranges over Γ. Then $\tilde{\Gamma}_{\varepsilon_1} \subset \tilde{\Gamma}$ and $\tilde{\Gamma}_0 = \Gamma$.

Given a positive integer Λ, let $\phi(u;\Lambda)$ denote the function on $[a,b]$ defined by

$$\phi(u;\Lambda) = a + \frac{b-a}{\Lambda}\left[\Lambda\frac{u-a}{b-a} + \frac{1}{2}\right], \quad a \leqslant u \leqslant b,$$

where $[\,\cdot\,]$ denotes the greatest integer function. Observe that $\phi(u;\Lambda)$ is nondecreasing in u, $\phi(a;\Lambda) = a$, $\phi(b;\Lambda) = b$, $\phi(u;\Lambda) \in \{a + i(b-a)/\Lambda\colon\, i = 0,\ldots,\Lambda\}$, and

$$u - \frac{b-a}{2\Lambda} < \phi(u;\Lambda) \leqslant u + \frac{b-a}{2\Lambda}, \quad a \leqslant u \leqslant b.$$

Given the free knot sequence $\boldsymbol{\gamma}$, consider the transformed sequence $\phi(\boldsymbol{\gamma};\Lambda) = (\phi(\gamma_j;\Lambda))$. Let $0 < \varepsilon \leqslant 1$. Observe that

$$|\phi(\boldsymbol{\gamma};\Lambda) - \boldsymbol{\gamma}|_\infty \leqslant \frac{b-a}{2\Lambda}$$

and hence that if

$$\Lambda \geqslant 4\varepsilon^{-1}\bar{M}N_n, \tag{A.6}$$

then $\zeta(\boldsymbol{\gamma},\phi(\boldsymbol{\gamma},\Lambda)) \leqslant \varepsilon$. Let Λ be the smallest integer satisfying (A.6). Then $\Lambda - 1 \leqslant 4\varepsilon^{-1}\bar{M}N_n$. [Observe also that if (A.6) holds, $u_1, u_2 \in [a,b]$, and

$$u_2 - u_1 \geqslant \frac{(b-a)\varepsilon}{4\bar{M}N_n},$$

then

$$\frac{\Lambda(u_2-u_1)}{b-a} \geqslant 1$$

and hence $\phi(u_2;\Lambda) > \phi(u_1;\Lambda)$.]

Suppose that (A.6) holds and let $0 \leqslant \varepsilon_0 \leqslant 1$. Set $\tilde{\Gamma}'_{\varepsilon_0,\varepsilon} = \{\phi(\gamma; \Lambda)\colon \gamma \in \tilde{\Gamma}_{\varepsilon_0}\} \subset \tilde{\Gamma}_{\varepsilon_0+\varepsilon}$. Then every point in $\tilde{\Gamma}_{\varepsilon_0}$ is within ε of some point in $\tilde{\Gamma}'_{\varepsilon_0,\varepsilon}$. Observe that

$$\#(\tilde{\Gamma}'_{\varepsilon_0,\varepsilon}) \leqslant \binom{(m-1)(\Lambda-1)}{J}.$$

(Note that the multiplicity of each free knot is at most $m-1$.)

Let I be an integer with $I \geqslant J$. Then

$$1 = \sum_{y=0}^{I} \binom{I}{y} \left(\frac{J}{I}\right)^{y} \left(1-\frac{J}{I}\right)^{I-y} \geqslant \binom{I}{J} \left(\frac{J}{I}\right)^{J} \left(1-\frac{J}{I}\right)^{I-J},$$

so

$$\binom{I}{J} \leqslant \left(\frac{I}{J}\right)^{J} \left(1-\frac{J}{I}\right)^{J-I} = \left(\frac{I}{J}\right)^{J} \left(\left(1-\frac{J}{I}\right)^{-(\frac{I}{J}-1)}\right)^{J} \leqslant \left(\frac{I}{J}\right)^{J} \mathrm{e}^{J}.$$

(Observe that $(\mathrm{d}/\mathrm{d}x)[x+(1-x)\log(1-x)] > 0$ for $0 < x < 1$, so $x+(1-x)\log(1-x) > 0$ for $0 < x < 1$ and hence $(1-x)^{-(1/x-1)} < e$ for $0 < x < 1$.) Consequently,

$$\#(\tilde{\Gamma}'_{\varepsilon_0,\varepsilon}) \leqslant \left[4e\varepsilon^{-1}\bar{M}(m-1)\left(1+\frac{m}{J}\right)\right]^{J} \leqslant (4e\varepsilon^{-1}m^2\bar{M})^{N_n}.$$

Consider now the general case $L \geqslant 1$. Here $\zeta(\gamma, \tilde{\gamma}) = \max_l \zeta_l(\gamma_l, \tilde{\gamma}_l)$ and

$$\Lambda_l \geqslant 4\varepsilon^{-1}\bar{M}_l N_{nl}, \quad 1 \leqslant l \leqslant L. \tag{A.7}$$

Let $\tilde{\Gamma}$ be the Cartesian product of $\tilde{\Gamma}_l$, $1 \leqslant l \leqslant L$, and let $\tilde{\Gamma}_{\varepsilon_1}$ denote the Cartesian product of $\tilde{\Gamma}_{l\varepsilon_1}$, $1 \leqslant l \leqslant L$. Then $\tilde{\Gamma}_{\varepsilon_1} \subset \Gamma$ and $\tilde{\Gamma}_0 = \Gamma$. Let $\tilde{\Gamma}'_{\varepsilon_0,\varepsilon} \subset \tilde{\Gamma}_{\varepsilon_0+\varepsilon}$ denote the Cartesian product of $\tilde{\Gamma}'_{l\varepsilon_0,\varepsilon}$, $1 \leqslant l \leqslant L$. Then every point in $\tilde{\Gamma}_{\varepsilon_0}$ is within ε of some point in $\tilde{\Gamma}'_{\varepsilon_0,\varepsilon}$. Now $N_n = \prod_l N_{nl} \geqslant \sum_l N_{nl}$, so

$$\#(\tilde{\Gamma}_{\varepsilon_0,\varepsilon}) \leqslant \left(4e\varepsilon^{-1}\max_l \bar{M}_l m_l^2\right)^{N_n}.$$

Let $0 < \varepsilon \leqslant 1/2$, let K be a positive integer, and set $\Xi_K = \tilde{\Gamma}_{0,\varepsilon^K} \subset \tilde{\Gamma}$ and $\Xi_k = \tilde{\Gamma}_{\varepsilon^K+\cdots+\varepsilon^{k+1},\varepsilon^k} \subset \tilde{\Gamma}$ for $0 \leqslant k \leqslant K-1$. Then

$$\#(\Xi_k) \leqslant \left(4e\varepsilon^{-k}\max_l \bar{M}_l m_l^2\right)^{N_n}, \quad 1 \leqslant k \leqslant K.$$

Moreover, every point in $\Gamma = \tilde{\Gamma}_0$ is within ε^K of some point in $\tilde{\Gamma}_{0,\varepsilon^K} = \Xi_K$; and, for $1 \leqslant k \leqslant K$, every point in $\Xi_k \subset \tilde{\Gamma}_{\varepsilon^K+\cdots+\varepsilon^k}$ is within ε^{k-1} of some point in $\tilde{\Gamma}_{\varepsilon^K+\cdots+\varepsilon^k,\varepsilon^{k-1}} = \Xi_{k-1}$. □

Proof of Lemma 3.2. For each point $\gamma' \in \tilde{\Gamma}'_{0,\varepsilon/2}$ (which is defined as in the proof of Lemma 3.1), there is a point $\tilde{\gamma}$ in the compact set $\{\gamma \in \tilde{\Gamma}\colon \zeta(\gamma', \gamma) \leqslant \varepsilon/2\}$ that minimizes the function $\|\bar{\eta}_\gamma - \eta^*\|$ over this set. Let $\tilde{\Gamma}_{0,\varepsilon}$ denote the collection of all such points $\tilde{\gamma}$. Then $\tilde{\Gamma}_{0,\varepsilon} \subseteq \tilde{\Gamma}_\varepsilon$ and $\#(\tilde{\Gamma}_{0,\varepsilon}) \leqslant \#(\tilde{\Gamma}'_{0,\varepsilon/2}) \leqslant (8e\varepsilon^{-1}\max_l \bar{M}_l m_l^2)^{N_n}$. Given $\gamma \in \Gamma$, choose

$\gamma' \in \tilde{\Gamma}'_{0,\varepsilon/2}$ such that $\zeta(\gamma,\gamma') \leqslant \varepsilon/2$ and let $\tilde{\gamma} \in \tilde{\Gamma}_{0,\varepsilon}$ be as defined above. Then $\zeta(\gamma,\tilde{\gamma}) \leqslant \varepsilon$ and $\|\bar{\eta}_{\tilde{\gamma}} - \eta^*\| \leqslant \|\bar{\eta}_\gamma - \eta^*\|$. □

Suppose that $L = 1$. Let $B_{\gamma j}$ be the normalized B-spline corresponding to the knot sequence $\gamma_{j-m}, \ldots, \gamma_j$. According to Theorem 4.2 of DeVore and Lorentz (1993, Chapter 5), there is a positive constant $D_m \leqslant 1$ such that

$$\frac{D_m^2}{m(b-a)} \sum_j b_j^2(\gamma_j - \gamma_{j-m}) \leqslant \left\| \sum_j b_j B_{\gamma j} \right\|_\psi^2 \leqslant \frac{1}{m(b-a)} \sum_j b_j^2(\gamma_j - \gamma_{j-m}) \tag{A.8}$$

and

$$D_m \max_j |b_j| \leqslant \left\| \sum_j b_j B_{\gamma j} \right\|_\infty \leqslant \max_j |b_j|. \tag{A.9}$$

It follows from (A.4), (A.5) and (A.8) that

$$\frac{D_m^2}{2\bar{M}(J+m)} \sum_j b_j^2 \leqslant \left\| \sum_j b_j B_{\gamma j} \right\|_\psi^2 \leqslant \frac{2\bar{M}}{J+m} \sum_j b_j^2, \quad \gamma \in \Gamma. \tag{A.10}$$

For general L, set $m = \prod_l m_l$, $D = \prod_l D_{m_l}$, and $\bar{M} = \prod_l \bar{M}_l$, and note that $N_n = \prod_l (J_l + m_l)$. Also, let $\mathscr{J}$ denote the Cartesian product of the sets $\{1, \ldots, J_{l+m_l}\}$, $1 \leqslant l \leqslant L$ and, for $j = (j_1, \ldots, j_L) \in \mathscr{J}$, consider the tensor product B-spline $B_{\gamma j}(\boldsymbol{u}) = B_{\gamma_1 j_1}(u_1) \cdots B_{\gamma_L j_L}(u_L)$. The *support* $\operatorname{supp}(h)$ of a function h on a set $\mathscr{U}$ is defined by $\operatorname{supp}(h) = \{\boldsymbol{u} \in \mathscr{U} : h(\boldsymbol{u}) \neq 0\}$.

Lemma A.1. *Let $\gamma, \tilde{\gamma} \in \tilde{\Gamma}$ and $j \in \mathscr{J}$. Then*

$$\frac{D^2}{6^L \bar{M} N_n} \sum_j b_j^2 \leqslant \left\| \sum_j b_j B_{\gamma j} \right\|_\psi^2 \leqslant \frac{6^L \bar{M}}{N_n} \sum_j b_j^2; \tag{A.11}$$

$$D \max_j |b_j| \leqslant \left\| \sum_j b_j B_{\gamma j} \right\|_\infty \leqslant \max_j |b_j|; \tag{A.12}$$

$$\psi(\operatorname{supp}(B_{\gamma j})) \leqslant \frac{6^L \bar{M} m}{N_n}; \tag{A.13}$$

$$\#\{j \in \mathscr{J} : B_{\gamma j}(\boldsymbol{u}) \neq 0\} \leqslant m \quad \textit{for } \boldsymbol{u} \in \mathscr{U}; \tag{A.14}$$

$$\#\{k \in \mathscr{J} : B_{\gamma j} B_{\tilde{\gamma} k} \textit{ is not identically zero on } \mathscr{U}\} \leqslant 38^L \bar{M}^2 m; \tag{A.15}$$

$$\|B_{\gamma j} - B_{\tilde{\gamma} j}\|_\infty \leqslant L\zeta(\gamma, \tilde{\gamma}); \tag{A.16}$$

$$\|B_{\gamma j} - B_{\tilde{\gamma} j}\|_\psi^2 \leqslant \frac{L^2 6^L 2\bar{M} m}{N_n}\,\zeta^2(\gamma,\tilde{\gamma}); \tag{A.17}$$

$$\left\|\sum_j b_j B_{\gamma j} - \sum_j b_j B_{\tilde{\gamma} j}\right\|_\psi^2 \leqslant \frac{8L^2 6^{2L} 38^L \bar{M}^4 m^2}{D^2}\,\zeta^2(\gamma,\tilde{\gamma})\left\|\sum_j b_j B_{\gamma j}\right\|_\psi^2; \tag{A.18}$$

and

$$\left\|\sum_j b_j B_{\gamma j} - \sum_j b_j B_{\tilde{\gamma} j}\right\|_\infty \leqslant \frac{2mL}{D}\,\zeta(\gamma,\tilde{\gamma})\left\|\sum_j b_j B_{\gamma j}\right\|_\infty. \tag{A.19}$$

Proof. Eq. (A.11) follows from (A.8), with $\bar{M}$ replaced by $3\bar{M}$, and induction; (A.12) follows from (A.9) and induction; since $\psi(\mathrm{supp}(B_{\gamma j})) = \prod_l [(\gamma_{l,j} - \gamma_{l,j-m_l})/(b_l - a_l)]$, (A.13) follows from (A.5) with $\bar{M}_l$ replaced by $3\bar{M}_l$.

To verify (A.14), let $u_l \in \mathscr{U}_l$ and suppose first that u_l is not a knot. Then $\gamma_{l,j_0} < u_l < \gamma_{l,j_0+1}$ for some j_0. If $B_{\gamma_l j}(u_l) > 0$, then $\gamma_{l,j-m_l} < u_l < \gamma_{l,j}$ and hence $j_0+1 \leqslant j \leqslant j_0 + m_l$. Suppose, instead, that $u = \gamma_{l,j_0}$. If $B_{\gamma_l j}(u_l) > 0$, then $\gamma_{l,j-m_l} < \gamma_{l,j_0} < \gamma_{l,j}$, so $j_0 + 1 \leqslant j \leqslant j_0 + m_l - 1$. In either case,

$$\#\{j \in \mathscr{J}\colon B_{\gamma j} \neq 0\} = \prod_l \#\{j \in \mathscr{J}_l\colon B_{\gamma_l j}(u_l) \neq 0\} \leqslant \prod_l m_l = m.$$

To verify (A.15), given $j \in \mathscr{J}_l$, let k_1 (k_2) be the smallest (largest) value of k in $\mathscr{J}_l$ such that $B_{\gamma_l j} B_{\tilde{\gamma}_l k}$ is not identically zero. Then $\tilde{\gamma}_{l,k_1} > \gamma_{l,j-m_l}$ and $\tilde{\gamma}_{l,k_2-m_l} < \gamma_{l,j}$. It follows from (A.5) (with $\bar{M}_l$ replaced by $3\bar{M}_l$) that

$$\gamma_{l,k_2-m_l} < \gamma_{l,j} \leqslant \gamma_{l,j-m_l} + \frac{6\bar{M}_l m_l (b-a)}{J_l + m_l}.$$

Let I be the smallest integer such that $I \geqslant 6^2 \bar{M}_l^2$. It follows from (A.4) that

$$\tilde{\gamma}_{l,k_1+Im_l} \geqslant \tilde{\gamma}_{l,k_1} + \frac{Im_l(b-a)}{2\bar{M}_l(J_l+m_l)} > \gamma_{l,j-m_l} + \frac{Im_l(b-a)}{6\bar{M}_l(J_l+m_l)} \geqslant \gamma_{l,k_2-m_l}$$

and hence that $k_2 < k_1 + (I+1)m_l$. Consequently,

$$\begin{aligned}&\#\{k \in \mathscr{J}_l\colon B_{\gamma_l j} B_{\tilde{\gamma}_l k} \text{ is not identically zero on } \mathscr{U}_l\}\\ &\quad \leqslant (I+1)m_l \leqslant (6^2\bar{M}_l^2 + 2)m_l \leqslant 38\bar{M}_l^2 m_l,\end{aligned}$$

which yields the desired result.

To verify (A.16), we first observe that, as a consequence of Definitions 4.12 and 4.19 and Theorems 2.51, 2.55, and 4.27 of Schumaker (1981), the partial derivative of $B_{\gamma_l j}$ with respect to the knot $\gamma_{l,k}$ for $j - m_l \leqslant k \leqslant j$ is bounded in absolute value by

$$\max\left\{\frac{1}{\gamma_{l,j-1} - \gamma_{l,j-m_l}}, \frac{1}{\gamma_{l,j} - \gamma_{l,j+1-m_l}}\right\}.$$

Thus, by (A.2),

$$\|B_{\gamma_l j} - B_{\tilde{\gamma}_l j}\|_\infty \leqslant \frac{3\bar{M}_l(m_l+1)N_l}{(m_l-1)(b_l-a_l)}|\gamma_l - \tilde{\gamma}_l|_\infty$$
$$\leqslant \frac{m_l+1}{3(m_l-1)}\zeta_l(\gamma,\tilde{\gamma}) \leqslant \zeta_l(\gamma,\tilde{\gamma}). \tag{A.20}$$

The desired result now follows from the observation that normalized B-splines lie between 0 and 1.

Eq. (A.17) follows from (A.13) and (A.16).

Set

$$A_{\gamma\tilde{\gamma}j} = \{k \in \mathscr{J}: \langle B_{\gamma j} - B_{\tilde{\gamma}j}, B_{\gamma k} - B_{\tilde{\gamma}k}\rangle_\psi \neq 0\}, \quad \gamma, \tilde{\gamma} \in \tilde{\Gamma} \text{ and } j \in \mathscr{J}.$$

Then $\#(A_{\gamma\tilde{\gamma}j}) \leqslant 38^L 4\bar{M}^2 m$ by (A.15). Consequently, by (A.11) and(A.17),

$$\begin{aligned}
&\left\|\sum_j b_j B_{\gamma j} - \sum_j b_j B_{\tilde{\gamma}j}\right\|_\psi^2 \\
&= \sum_j \sum_{k \in A_{\gamma\tilde{\gamma}j}} b_j b_k \langle B_{\gamma j} - B_{\tilde{\gamma}j}, B_{\gamma k} - B_{\tilde{\gamma}k}\rangle_\psi \\
&\leqslant \sum_j \sum_{k \in A_{\gamma\tilde{\gamma}j}} \left(\frac{b_j^2 + b_k^2}{2}\right)\left(\frac{\|B_{\gamma j} - B_{\tilde{\gamma}j}\|_\psi^2 + \|B_{\gamma k} - B_{\tilde{\gamma}k}\|_\psi^2}{2}\right) \\
&\leqslant \frac{8L^2 6^L 38^L \bar{M}^3 m^2}{N_n}\zeta^2(\gamma,\tilde{\gamma})\sum_j b_j^2 \\
&\leqslant \frac{8L^2 6^{2L} 38^L \bar{M}^4 m^2}{D^2}\zeta^2(\gamma,\tilde{\gamma})\left\|\sum_j b_j B_{\gamma j}\right\|_\psi^2,
\end{aligned}$$

so (A.18) holds.

It follows from (A.14) that, for $\gamma, \tilde{\gamma} \in \Gamma$ and $\boldsymbol{u} \in \mathscr{U}$, there are at most $2m$ values of $j \in \mathscr{J}$ such that $B_{\gamma j}(\boldsymbol{u}) - B_{\tilde{\gamma}j}(\boldsymbol{u}) \neq 0$. Thus, by (A.12) and (A.16),

$$\left\|\sum_j b_j B_{\gamma j} - \sum_j b_j B_{\tilde{\gamma}j}\right\|_\infty \leqslant \frac{2mL}{D}\zeta(\gamma,\tilde{\gamma})\left\|\sum_j b_j B_{\gamma j}\right\|_\infty,$$

so (A.19) holds. □

Proof of Lemma 3.3. It follows from (A.11) and (A.12) that

$$\left\|\sum_j b_j B_{\gamma j}\right\|_\infty^2 \leqslant \max_j b_j^2 \leqslant \frac{6^L \bar{M} N_n}{D^2}\left\|\sum_j b_j B_{\gamma j}\right\|_\psi^2.$$

The desired result now follows from (3.2). □

Recall that $\boldsymbol{U}$ is defined as a transform of $\boldsymbol{W}$. Let $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n$ be the corresponding transforms of $\boldsymbol{W}_1,\ldots,\boldsymbol{W}_n$, respectively. Recall the definition of empirical inner product and empirical norm in Section 3. Observe that $E_n(h)=\langle 1,h\rangle_n$.

Lemma A.2. *Suppose Condition* 3.1 *holds and that* $N_n=o(n^{1-\varepsilon})$ *for some* $\varepsilon>0$. *Then there is a constant* M *and there is an event* Ω_n *such that* $\lim_n P(\Omega_n)=1$ *and*

$$\left\|\sum_j \beta_j B_{\gamma j}-\sum_j \beta_j B_{\tilde\gamma j}\right\|_n^2 \leqslant M\zeta^2(\gamma,\tilde\gamma)\left\|\sum_j \beta_j B_{\gamma j}\right\|^2 \quad \text{on } \Omega_n$$

for $\gamma,\tilde\gamma\in\tilde\Gamma$ *and* $\beta_j\in\mathbb{R}$ *for* $j\in\mathscr{J}$.

Proof. It follows from (A.16) that

$$\begin{aligned}\|B_{\gamma j}-B_{\tilde\gamma j}\|_n^2 &\leqslant \|B_{\gamma j}-B_{\tilde\gamma j}\|_\infty^2\frac{1}{n}\#(\{i\colon \boldsymbol{U}_i\in\operatorname{supp}(B_{\gamma j})\cup\operatorname{supp}(B_{\tilde\gamma j})\})\\ &\leqslant L^2\zeta^2(\gamma,\tilde\gamma)\frac{1}{n}\#(\{i\colon \boldsymbol{U}_i\in\operatorname{supp}(B_{\gamma j})\cup\operatorname{supp}(B_{\tilde\gamma j})\})\end{aligned}$$

for $\gamma,\tilde\gamma\in\tilde\Gamma$ and $j\in\mathscr{J}$. It follows from Condition 3.1, (A.13), and the assumption on N_n by a straightforward application of Bernstein's inequality (4.2) [or by Theorem 12.2 of Breiman et al. (1984)] that

$$\sup_{\gamma,\tilde\gamma\in\tilde\Gamma}\max_{j\in\mathscr{J}}\frac{1}{n}\#(\{i\colon \boldsymbol{U}_i\in\operatorname{supp}(B_{\gamma j})\cup\operatorname{supp}(B_{\tilde\gamma j})\})\leqslant\frac{6^L 2\bar{M}M_2 m}{N_n}[1+o_P(1)].$$

Let Ω_n denote the event that

$$\sup_{\gamma,\tilde\gamma\in\tilde\Gamma}\max_{j\in\mathscr{J}}\frac{1}{n}\#(\{i\colon \boldsymbol{U}_i\in\operatorname{supp}(B_{\gamma j})\cup\operatorname{supp}(B_{\tilde\gamma j})\})\leqslant\frac{6^L 4\bar{M}M_2 m}{N_n}.$$

Then $\lim_n P(\Omega_n)=1$ and

$$\|B_{\gamma j}-B_{\tilde\gamma j}\|_n^2\leqslant\frac{4L^2 6^L\bar{M}M_2 m}{N_n}\zeta^2(\gamma,\tilde\gamma)\quad\text{on }\Omega_n \tag{A.21}$$

for $\gamma,\tilde\gamma\in\tilde\Gamma$ and $j\in\mathscr{J}$.

Set $A_{\gamma\tilde\gamma jn}=\{k\in\mathscr{J}\colon\langle B_{\gamma j}-B_{\tilde\gamma j},B_{\gamma k}-B_{\tilde\gamma k}\rangle_n\neq 0\}$ for $\gamma,\tilde\gamma\in\tilde\Gamma$ and $j\in\mathscr{J}$. Then $\#(A_{\gamma\tilde\gamma jn})\leqslant 38^L 4\bar{M}^2 m$ by (A.15). Consequently, by (A.11), (A.21), and Condition 3.1,

$$\begin{aligned}&\left\|\sum_j \beta_j B_{\gamma j}-\sum_j \beta_j B_{\tilde\gamma j}\right\|_n^2\\ &\quad=\sum_k\sum_{k\in A_{\gamma\tilde\gamma jn}}\beta_j\beta_k\langle B_{\gamma j}-B_{\tilde\gamma j},B_{\gamma k}-B_{\tilde\gamma k}\rangle_n\\ &\quad\leqslant\sum_j\sum_{k\in A_{\gamma\tilde\gamma jn}}\left(\frac{\beta_j^2+\beta_k^2}{2}\right)\left(\frac{\|B_{\gamma j}-B_{\tilde\gamma j}\|_n^2+\|B_{\gamma k}-B_{\tilde\gamma k}\|_n^2}{2}\right)\end{aligned}$$

$$\leqslant \frac{16L^2 6^L 38^L \bar{M}^3 M_2 m^2}{N_n} \zeta^2(\gamma,\tilde{\gamma}) \sum_j \beta_j^2$$

$$\leqslant \frac{16L^2 6^{2L} 38^L \bar{M}^4 M_2 m^2}{D^2 M_1} \zeta^2(\gamma,\tilde{\gamma}) \left\| \sum_j \beta_j B_{\gamma j} \right\|^2$$

on Ω_n for $\gamma, \tilde{\gamma} \in \tilde{\Gamma}$ and $\beta_j \in \mathbb{R}$ for $j \in \mathscr{J}$, as desired. □

Proof of Lemma 3.4. Set $\varepsilon = \zeta(\gamma,\tilde{\gamma})$. Write $g = \sum_j \beta_j B_{\gamma j}$ and set $g' = \sum_j \beta_j B_{\tilde{\gamma} j}$. It follows from (3.2) and (A.18) that $\|g - g'\| \leqslant c_1 \varepsilon \|g\|$ for some constant c_1, and it follows from (A.19) that $\|g - g'\|_\infty \leqslant c_2 \varepsilon \|g\|_\infty$ for some constant c_2.

If $\|g'\| \leqslant \|g\|$, then $\tilde{g} = g'$ has the properties specified in the first result of the lemma. Suppose, instead, that $\|g'\| > \|g\|$ and set $\lambda = \|g\|/\|g'\|$. Then $(1 + c_1\varepsilon)^{-1} \leqslant \lambda < 1$, $\|\lambda g'\| = \|g\|$, and $\|g - \lambda g'\| \leqslant \|g - g'\| \leqslant c_1 \varepsilon \|g\|$. (Note that $\langle g, g' \rangle \leqslant \|g\| \|g'\|$ by the Cauchy–Schwarz inequality.) Moreover,

$$\begin{aligned} \|g' - \lambda g'\|_\infty &= (\|g'\| - \|g\|) \frac{\|g'\|_\infty}{\|g'\|} \\ &\leqslant \frac{\|g' - g\|(\|g\|_\infty + \|g' - g\|_\infty)}{\|g\|} \\ &\leqslant c_1 \varepsilon (1 + c_2) \|g\|_\infty, \end{aligned}$$

so $\|g - \lambda g'\|_\infty \leqslant (c_1 + c_2 + c_1 c_2)\varepsilon \|g\|_\infty$ and hence $\tilde{g} = \lambda g'$ has the properties specified in the first result.

Let Ω_{n1} be the event Ω_n in Lemma A.2, let Ω_{n2} be the event that $\|g\|_n^2 \leqslant 2\|g\|^2$ for $\gamma \in \tilde{\Gamma}$, and set $\Omega_n = \Omega_{n1} \cup \Omega_{n2}$. It follows from Lemmas A.2 and 4.2 that $\lim_n P(\Omega_n) = 1$. Let ε, g', and λ be as in the proof of the first result of the lemma. Then for some constant c_3, $\|g - g'\|_n \leqslant c_3 \varepsilon \|g\|$ on Ω_n. If $\|g'\| \leqslant \|g\|$, then $\tilde{g} = g'$ satisfies the desired additional property. Otherwise,

$$\begin{aligned} \|g - \lambda g'\|_n &\leqslant \|g - g'\|_n + (1 - \lambda)\|g'\|_n \\ &\leqslant c_3 \varepsilon \|g\| + 2\left(\frac{1}{\lambda} - 1\right)\|g\| \\ &\leqslant (c_3 + 2c_1)\varepsilon \|g\| \end{aligned}$$

on Ω_n, so $\tilde{g} = \lambda g'$ satisfies the desired additional property. □

Proof of Lemma 3.5. Let $K_1 > K$. Choose $\gamma, \tilde{\gamma} \in \tilde{\Gamma}$ such that $\|\bar{\eta}_\gamma\|_\infty \leqslant K$ and $\|\bar{\eta}_{\tilde{\gamma}}\|_\infty \leqslant K$, and set $\varepsilon = \zeta(\gamma,\tilde{\gamma})$. By Lemma 3.4, there is a fixed positive number c_1 (not depending on $\gamma, \tilde{\gamma}$) and there are functions $\eta'_\gamma \in \mathbb{G}_\gamma$ and $\eta'_{\tilde{\gamma}} \in \mathbb{G}_{\tilde{\gamma}}$ such that $\|\eta'_\gamma - \bar{\eta}_{\tilde{\gamma}}\|_\infty \leqslant c_1\varepsilon$ and $\|\eta'_{\tilde{\gamma}} - \bar{\eta}_\gamma\|_\infty \leqslant c_1\varepsilon$. Without loss of generality, we can assume that $\varepsilon \leqslant 1$ and that ε is sufficiently small that $\|\eta'_\gamma\|_\infty \leqslant K_1$ and $\|\eta'_{\tilde{\gamma}}\|_\infty \leqslant K_1$. Then, by Condition 2.2(ii), there is a fixed positive number c_2 such that $\Lambda(\bar{\eta}_\gamma) - \Lambda(\eta'_{\tilde{\gamma}}) \leqslant c_2\varepsilon$ and $\Lambda(\bar{\eta}_{\tilde{\gamma}}) - \Lambda(\eta'_\gamma) \leqslant c_2\varepsilon$. Since $\Lambda(\eta'_{\tilde{\gamma}}) \leqslant \Lambda(\bar{\eta}_{\tilde{\gamma}})$, we conclude that $\Lambda(\bar{\eta}_\gamma) - \Lambda(\eta'_\gamma) \leqslant 2c_2\varepsilon$. On the other hand, by Condition 2.2(ii), $\Lambda(\bar{\eta}_\gamma) - \Lambda(\eta'_\gamma) \geqslant c_3\|\bar{\eta}_\gamma - \eta'_\gamma\|^2$ for some constant c_3, so $\|\bar{\eta}_\gamma - \eta'_\gamma\| \leqslant (2c_2c_3^{-1}\varepsilon)^{1/2}$ and hence

$$\begin{aligned}\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\| &\leqslant \|\bar{\eta}_\gamma - \eta'_\gamma\| + \|\eta'_\gamma - \bar{\eta}_{\tilde{\gamma}}\| \\ &\leqslant (2c_2c_3^{-1}\varepsilon)^{1/2} + c_1\varepsilon \\ &\leqslant [(2c_2c_3^{-1})^{1/2} + c_1]\varepsilon^{1/2}.\end{aligned}$$

Moreover, by (3.2), (A.11), and (A.12), $\|\bar{\eta}_\gamma - \eta'_\gamma\|_\infty \leqslant c_4 N_n^{1/2}\|\bar{\eta}_\gamma - \eta'_\gamma\|$. So,

$$\begin{aligned}\|\bar{\eta}_\gamma - \bar{\eta}_{\tilde{\gamma}}\|_\infty &\leqslant \|\bar{\eta}_\gamma - \eta'_\gamma\|_\infty + \|\eta'_\gamma - \bar{\eta}_{\tilde{\gamma}}\|_\infty \\ &\leqslant \left(\frac{2c_2c_4^2}{c_3}\right)^{1/2} N_n^{1/2}\varepsilon^{1/2} + c_1\varepsilon.\end{aligned}$$

By Lemma 4.2, there is an event Ω_n such that $\lim_n P(\Omega_n)=1$ and $\|g\|_n \leqslant 2\|g\|$ on Ω_n for $\gamma \in \tilde{\Gamma}$ and $g \in \mathbb{G}_\gamma$. Thus, by the first paragraph of this proof, $\|\bar{\eta}_\gamma - \eta'_\gamma\|_n \leqslant 2\|\bar{\eta}_\gamma - \eta'_\gamma\| \leqslant 2(2c_2c_3^{-1}\varepsilon)^{1/2}$ and hence $\|\bar{\eta}_{\tilde{\gamma}} - \bar{\eta}_\gamma\|_n \leqslant [2(2c_2c_3^{-1})^{1/2} + c_1)]\varepsilon^{1/2}$ on Ω_n for $\gamma, \tilde{\gamma}$ as in the first paragraph. □

References

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

de Boor, C., 1976. A bound on the L_∞-norm of the L_2-approximation by splines in terms of a global mesh ratio. Math. Comp. 30, 765–771.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

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

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

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, 961–999.

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.

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

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

Schumaker, L.L., 1981. Spline Functions: Basic Theory. 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, 18–184.
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.
Yurinskiǐ, V.V., 1976. Exponential inequalities for sums of random vectors. Multivariate Anal. 6, 473–499.
Zhou, S.G., Shen, X.T., 2001. Spatially adaptive regression splines and accurate knot selection schemes. J. Amer. Statist. Asssoc. 96, 247–259.