Statistical Modeling with Spline Functions
Methodology and Theory

Mark H. Hansen
University of California at Los Angeles

Jianhua Z. Huang
University of Pennsylvania

Charles Kooperberg
Fred Hutchinson Cancer Research Center

Charles J. Stone
University of California at Berkeley

Young K. Truong
University of North Carolina at Chapel Hill

January 5, 2006

11

Rates of Convergence in Extended Linear Modeling

In previous chapters, we have separately developed spline-based methodology in the contexts of regression, generalized regression, polychotomous regression, density estimation, spectral density estimation, and hazard regression. Throughout the remainder of the book, spline based procedures will be studied in a general mathematical framework and their fundamental properties will be investigated. Specifically, we consider unknown functions and maximum likelihood estimates of such functions in finite-dimensional estimation spaces that mainly are built up from polynomial splines and their selected tensor products. We will develop a large sample asymptotic theory on consistency and rates of convergence for the resulting estimates. It turns out that the asymptotic results can be obtained simultaneously for various statistical contexts by introducing the framework of "extended linear modeling" as a broad synthesis.

In the present chapter, the dimension of the estimation spaces is allowed to depend on the sample size, but not on the sample data itself. This corresponds to fixed knot spline estimation in which the knot positions are prespecified. However, in our asymptotic theory, the number of knots is allowed to increase with the sample size, reflecting improved flexibility for increasing sample size. In the next chapter we will extend the results in this chapter to handle free knot splines, that is, the knot positions are treated as "free" parameters to be determined by the data.

In Section 11.1.1 we describe the theoretical framework of extended linear modeling, which involves a model space $\mathbb{H}$, a log-likelihood function on $\mathbb{H}$, and maximum likelihood estimation corresponding to a finite-dimensional subspace $\mathbb{G}$ of $\mathbb{H}$. In Section 11.1.2 we give an informal discussion of the corresponding asymptotic theory on consistency and rates of convergence. In Section 11.1.3 we specialize these results by considering spaces $\mathbb{H}$ and $\mathbb{G}$ constructed using functional analysis of variance (ANOVA) decompositions. In Sections 11.2 and 11.3 we give formal presentations of the results in Sections 11.1.2 and 11.1.3, respectively. Section 11.4 contains verifications of some of the technical conditions for the main asymptotic results either in general or separately for generalized regression, density estimation, and hazard regression.

11.1 Theoretical Framework and Basic Results

In this section we give a detailed description of the theoretical framework of extended linear models and present some general asymptotic results applicable to a variety of contexts.

11.1.1 Extended Linear Models

Consider a $\mathcal{W}$-valued random variable $\boldsymbol{W}$, where $\mathcal{W}$ is an arbitrary set. The probability density $p(\eta, \boldsymbol{w})$ of $\boldsymbol{W}$ depends on an unknown function η. The function η is defined on a domain $\mathcal{U}$, which may or may not be the same as $\mathcal{W}$. We assume that $\mathcal{U}$ is a compact subset of some Euclidean space and that it has positive volume $\operatorname{vol}(\mathcal{U})$. The goal is to estimate η based on a random sample from the distribution of $\boldsymbol{W}$.

Corresponding to a candidate function h for η, the log-likelihood is given by $l(h, \boldsymbol{w}) = \log p(h, \boldsymbol{w})$. The expected log-likelihood is defined by $\Lambda(h) = E[l(h, \boldsymbol{W})]$, where the expectation is taken with respect to the true function η. There may be some mild restrictions on h for $l(h, \boldsymbol{w})$, $\boldsymbol{w} \in \mathcal{W}$, and $\Lambda(h)$ to be well-defined. It follows from the information inequality (Rao, 1973) that η is the essentially unique function on $\mathcal{U}$ that maximizes the expected log-likelihood. (Here two functions on $\mathcal{U}$ are regarded as essentially equal if their difference equals zero except on a subset of $\mathcal{U}$ having Lebesgue measure zero.)

In many applications, we are interested in a function η that is related to but need not totally specify the probability distribution of $\boldsymbol{W}$. In such applications, we can modify the above setup by taking $l(h, \boldsymbol{w})$ to be the logarithm of a conditional likelihood, a pseudo-likelihood, or a partial likelihood, depending on the problem under consideration.

Consider, for example, the estimation of a regression function $\eta(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$. In terms of the above notation, $\boldsymbol{W}$ consists of a pair of random variables $\boldsymbol{X}$ and Y, and $\mathcal{U}$ is the range of $\boldsymbol{X}$. We can take $l(h, \boldsymbol{w})$ to be the negative of the residual sum of squares; that is, $l(h, \boldsymbol{W}) = -[Y - h(\boldsymbol{X})]^2$ with $\boldsymbol{W} = (\boldsymbol{X}, Y)$. If the conditional distribution of Y given $\boldsymbol{X}$ is assumed to be normal with constant variance, then l is (up to additive and multiplicative constants) the conditional log-likelihood. Even if this conditional distribution is not assumed to be normal, we can still think of l as the logarithm of a pseudo-likelihood. In either case, the true regression function η maximizes $\Lambda(h) = E[l(h, \boldsymbol{W})] = -E[\eta(\boldsymbol{X}) - h(\boldsymbol{X})]^2$.

From now on, we will adopt this broad view of $l(h, \boldsymbol{w})$. For simplicity, we will still call $l(h, \boldsymbol{w})$ the log-likelihood and $\Lambda(h)$ the expected log-likelihood. To relate the function of interest to the log-likelihood, we assume that, subject to mild conditions on $l(h, \boldsymbol{w})$, the function η is the essentially unique function that maximizes the expected log-likelihood.

We now give three examples of this setup:

Generalized Regression. Consider an exponential family of distributions on $\mathbb{R}$ of the form $P(Y \in dy) = \exp[B(\eta)y - C(\eta)]\Psi(dy)$, 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 $C(\eta) = \log \int_{\mathbb{R}} \exp[(B(\eta)y]\Psi(dy) < \infty$ 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 $\mu = A(\eta) = C'(\eta)$.

The Bernoulli distribution with parameter $\pi \in (0, 1)$ forms an exponential family with canonical parameter $\eta = \operatorname{logit}(\pi) = \log \pi/(1 - \pi)$. Here $P(Y = 1) = \pi$, $P(Y = 0) = 1 - \pi$, Ψ is concentrated on $\{0, 1\}$ with $\Psi(\{0\}) = \Psi(\{1\}) = 1$, $B(\eta) = \eta$, $C(\eta) = \log\big(1 + \exp(\eta)\big)$, and $\mu = \pi = (\exp \eta)/(1 + \exp \eta)$.

The Poisson distribution with parameter $\lambda \in (0, \infty)$ forms an exponential family with canonical parameter $\eta = \log \lambda$. Here $P(Y = y) = \lambda^y \exp(-\lambda)/y!$ for $y \in \mathcal{Y} = \{0, 1, 2, \ldots\}$, Ψ is concentrated on $\mathcal{Y}$ with $\Psi(\{y\}) = 1/y!$ for $y \in \mathcal{Y}$, $B(\eta) = \eta$, $C(\eta) = \exp \eta$, and $\mu = \lambda = C'(\eta) = \exp \eta$.

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

$$P(Y \in dy | \boldsymbol{X} = \boldsymbol{x}) = \exp[B(\eta(\boldsymbol{x}))y - C(\eta(\boldsymbol{x}))]\Psi(dy). \qquad (11.1.1)$$

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

$$\mu(\boldsymbol{x}) = E(Y | \boldsymbol{X} = \boldsymbol{x}) = A(\eta(\boldsymbol{x})), \qquad \boldsymbol{x} \in \mathcal{X}. \qquad (11.1.2)$$

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 (11.1.2), not on the stronger assumption (11.1.1). In the application of the theory developed in this chapter and the next one to generalized regression, we require (11.1.2), but not (11.1.1).

When the underlying exponential family is the Bernoulli distribution with parameter π and canonical parameter $\eta = \operatorname{logit}(\pi)$, we get logistic regression. Here $\mu(\boldsymbol{x}) = \pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ and $\eta(\boldsymbol{x}) = \operatorname{logit}(\pi(\boldsymbol{x})) = \operatorname{logit}(\mu(\boldsymbol{x}))$.

When the underlying exponential family is the Poisson distribution with parameter λ and canonical parameter $\eta = \log\lambda$, we get Poisson regression. Here $\mu(\boldsymbol{x}) = \lambda(\boldsymbol{x})$ and $\eta(\boldsymbol{x}) = \log\lambda(\boldsymbol{x})$.

When the underlying exponential family is the normal distribution with canonical parameter $\eta = \mu$ and known variance, we get ordinary regression as discussed above.

Density Estimation. Let $\boldsymbol{Y} = \boldsymbol{W}$ have an unknown density function $f_{\boldsymbol{Y}}$ on $\mathcal{Y} = \mathcal{U}$, and let $\phi = \log f_{\boldsymbol{Y}}$ denote the corresponding log-density function. Since ϕ is controlled by the intrinsic nonlinear constraint $\int_{\mathcal{Y}} \exp\phi(\boldsymbol{y})\,d\boldsymbol{y} = 1$, it is convenient to write $\phi = \eta - c(\eta)$; here $c(\eta) = \log\int_{\mathcal{Y}} \exp\eta(\boldsymbol{y})\,d\boldsymbol{y}$. By imposing a linear constraint such as $\int_{\mathcal{Y}} \eta(\boldsymbol{y})\,d\boldsymbol{y}$, we can make the map $\sigma : \eta \mapsto \phi$ one-to-one. The problem of estimating ϕ is then transformed to that of estimating η. The log-likelihood corresponding to a candidate h for η is given by $l(h, \boldsymbol{Y}) = h(\boldsymbol{Y}) - c(h)$.

Hazard Regression. Consider a positive survival time T, a positive censoring time C, the observed time $\min(T, C)$, and an $\mathcal{X}$-valued random vector $\boldsymbol{X}$ of covariates. Let $\delta = \operatorname{ind}(T \le C)$ be the indicator random variable that equals one or zero according as $T \le C$ (T is uncensored) or $T > C$ (T is censored), and set $Y = \min(T, C)$ and $\boldsymbol{W} = (\boldsymbol{X}, Y, \delta)$. Suppose T and C are conditionally independent given $\boldsymbol{X}$. Suppose also that $P(C \le \tau) = 1$ for a known positive constant τ. Let

$$\eta(\boldsymbol{x}, t) = \log\frac{f(t|\boldsymbol{x})}{1 - F(t|\boldsymbol{x})}$$

denote the logarithm of the conditional hazard function, where $f(t|\boldsymbol{x})$ and $F(t|\boldsymbol{x})$ are the conditional density function and conditional distribution function, respectively, of T given that $\boldsymbol{X} = \boldsymbol{x}$. Then

$$1 - F(t|\boldsymbol{x}) = \exp\Bigl(-\int_0^t \exp\eta(\boldsymbol{x}, u)\,du\Bigr)$$

and hence

$$f(t|\boldsymbol{x}) = \exp\Bigl(\eta(\boldsymbol{x}, t) - \int_0^t \exp\eta(\boldsymbol{x}, u)\,du\Bigr).$$

The log-likelihood for a candidate h for η is given by

$$l(h, \boldsymbol{W}) = \delta h(\boldsymbol{X}, Y) - \int_0^Y \exp h(\boldsymbol{X}, t)\, dt.$$

Here, $\mathcal{U} = \mathcal{X} \times [0, \tau]$.

Conditional density estimation (Huang 2001), counting process regression (Huang 2001), marked point process regression (Li 2001), proportional hazards regression (Huang, Kooperberg, Stone and Truong 2000), robust regression (Stone 2001), and spectral density estimation (Kooperberg, Stone and Truong 1995d) can be treated in the present framework. Polychotomous regression (Hansen 1994) and event history analysis (Huang and Stone 1998) can also be treated in this framework provided that we consider vector-valued instead of real-valued functions on $\mathcal{U}$.

Let $\mathbb{H}$ be a linear space of square-integrable functions on $\mathcal{U}$ such that if two functions on $\mathcal{U}$ are essentially equal and one of them is in $\mathbb{H}$, then so is the other one. We refer to $\mathbb{H}$ as the *model space* and to $l(h, \boldsymbol{W})$, $h \in \mathbb{H}$, as forming an *extended linear model.* If $\mathbb{H}$ is the space of all square-integrable functions on $\mathcal{U}$ or differs from this space only by the imposition of some identifiability restrictions as in the context of density estimation, we refer to $\mathbb{H}$ as being *saturated.* Otherwise, we refer to this space as being *unsaturated.*

The use of unsaturated spaces allows us to impose structural assumptions on the extended linear model. Suppose $\mathcal{U}$ is the Cartesian product of compact intervals $\mathcal{U}_1, \ldots, \mathcal{U}_L$, each having positive length. We can impose an additive structure by letting $\mathbb{H}$ be the space of functions of the form $h_1(u_1) + \cdots + h_L(u_L)$, where h_l is a square-integrable function on $\mathcal{U}_l$ for $1 \le l \le L$. This and more general ANOVA structures will be considered in Section 11.1.3. Alternatively, we can impose an additive, semilinear structure by letting $\mathbb{H}$ be the space of functions of the form $h_1(u_1) + b_2 u_2 + \cdots + b_L u_L$, where h_1 is a square-integrable function on $\mathcal{U}_1$ and $b_2, \ldots, b_L$ are real numbers.

The extended linear model is said to be *concave* if the following two properties are satisfied: (i) The log-likelihood function is concave; that is, given any two functions $h_1, h_2 \in \mathbb{H}$ whose log-likelihoods are well-defined, $l(\alpha h_1 + (1-\alpha) h_2, \boldsymbol{w}) \ge \alpha l(h_1, \boldsymbol{w}) + (1-\alpha) l(h_2, \boldsymbol{w})$ for $0 < \alpha < 1$ and $\boldsymbol{w} \in \mathcal{W}$. (ii) The expected log-likelihood function is strictly concave; that is, given any two essentially different functions $h_1, h_2 \in \mathbb{H}$ whose expected log-likelihoods are well-defined, $\Lambda(\alpha h_1 + (1-\alpha) h_2) > \alpha \Lambda(h_1) + (1-\alpha)\Lambda(h_2)$ for $0 < \alpha < 1$. Here, we implicitly assume that the set of functions such that $l(h, \boldsymbol{w})$ and $\Lambda(h)$ are well-defined is a convex set.

In the contexts of ordinary regression, hazard regression, and density estimation, the corresponding extended linear model is automatically concave. In the context of generalized regression, a condition on the Ψ and $B(\cdot)$ is required to guarantee the concavity of the extended linear model (see Section 11.4.3).

As mentioned above, the model space $\mathbb{H}$ incorporates structural assumptions (e.g., additivity) on the true function of interest. Such structural assumptions are not necessarily true and are considered rather as approximations. Thus, it is natural to think that any estimation procedure will estimate the best approximation to the true function with the imposed structure. This "best approximation" can be defined formally using the expected log-likelihood. Observe that if $\eta \in \mathbb{H}$, then $\eta = \operatorname{argmax}_{h \in \mathbb{H}} \Lambda(h)$ since η maximizes the expected log-likelihood by assumption. More generally, we think of $\eta^* = \operatorname{argmax}_{h \in \mathbb{H}} \Lambda(h)$ as the "best approximation" in $\mathbb{H}$ to η. Typically, when the expected log-likelihood function is strictly concave, such a best approximation exists and is essentially unique. If $\eta \in \mathbb{H}$, then η^* is essentially equal to η.

In the regression context η^* is the orthogonal projection of η onto $\mathbb{H}$ with respect to the L_2 norm on $\mathbb{H}$ given by $\|h\|^2 = E[h^2(\boldsymbol{X})]$; that is, $\eta^* = \operatorname{argmin}_{h \in \mathbb{H}} \|h - \eta\|^2$. Here, to guarantee the existence of η^*, we need to assume that $\mathbb{H}$ is a Hilbert space; that is, it is closed in the metric corresponding to the indicated norm.

We now turn to estimation. Let $\boldsymbol{W}_1, \dots, \boldsymbol{W}_n$ be a random sample of size n from the distribution of $\boldsymbol{W}$. Let $\mathbb{G} \subset \mathbb{H}$ be a finite-dimensional linear space of bounded functions, whose dimension may depend on the sample size. We estimate η by using maximum likelihood over $\mathbb{G}$, that is, we take $\widehat{\eta} = \operatorname{argmax}_{g \in \mathbb{G}} \ell(g)$, where $\ell(g) = (1/n) \sum_{i=1}^n l(g, \boldsymbol{W}_i)$ is the normalized log-likelihood. Here the space $\mathbb{G}$ should be chosen such that the function of interest η can be approximated well by some function in $\mathbb{G}$. Thus $\mathbb{G}$ will be called the approximation space. Since $\mathbb{G}$ is where the maximum likelihood estimation is carried out, it will also be called the estimation space. In this setup we do not specify the form of $\mathbb{G}$; any linear function space with good approximation properties can be used. When $\mathbb{H}$ has a specific structure, $\mathbb{G}$ should be chosen to have the same structure. For example, if $\mathbb{H}$ consists of all square-integrable additive functions, then $\mathbb{G}$ should not contain any non-additive functions. A detailed discussion of constructing the model and estimation spaces using functional ANOVA decompositions to incorporate structural assumptions will be given in Section 11.1.3. In our application, $\mathbb{G}$ will be chosen as a space built by polynomial splines and their tensor products. That polynomial splines and their tensor products enjoy good approximation power has been extensively studied and documented; see de Boor (1978), Schumaker (1981), and DeVore and Lorentz (1993).

11.1.2 Consistency and Rates of Convergence

In this section we present results on the asymptotic properties of the maximum likelihood estimate $\widehat{\eta}$ in concave extended linear models. As discussed in the previous section, the best approximation η^* in $\mathbb{H}$ to the function η of interest can be thought as a general target of estimation whether or not $\eta \in \mathbb{H}$. The existence of η^* has been established in various contexts; see the references listed in Section 11.2 following Condition 11.2.1. We say that $\widehat{\eta}$ is consistent in estimating η^* if $\|\widehat{\eta} - \eta^*\| \to 0$ in probability for some norm $\|\cdot\|$. We will state conditions that ensure consistency and also determine the rates of convergence of $\widehat{\eta}$ to η^*. In the asymptotic analysis, it is natural to allow the dimension N_n of the estimation space $\mathbb{G}$ grow with the sample size, reflecting the improved approximation power of this space for increasing sample size.

We assume that the log-likelihood $l(h, \boldsymbol{w})$ and expected log-likelihood $\Lambda(h)$ are well-defined and finite for every bounded function h on $\mathcal{U}$. Since the estimation space $\mathbb{G} \subset \mathbb{H}$ is a finite-dimensional linear space of bounded functions, $\ell(h, \boldsymbol{w})$ and $\Lambda(h)$ are well-defined on $\mathbb{G}$.

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

$$\widehat{\eta} - \eta^* = (\bar{\eta} - \eta^*) + (\widehat{\eta} - \bar{\eta}),$$

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

To get mathematically rigorous results, we need some regularity conditions (that is, Conditions 11.2.1, 11.2.2 and 11.2.4). These conditions will be given explicitly and discussed in detail in Section 11.2 and verified as necessary in the contexts of generalized regression, including ordinary regression as a special case, density estimation, and hazard regression in Section 11.4. The main technical condition is that the log-likelihood is suitably concave. We assume these conditions hold throughout this subsection.

Before proceeding further, it is convenient to introduce some additional notation that will be used in this chapter and the next one. Let $\#(B)$ denote the cardinality (number of members) of a set B. Given a function h on $\mathcal{U}$, let $\|h\|_\infty = \sup_{\boldsymbol{u} \in \mathcal{U}} |h(\boldsymbol{u})|$ denote its L_∞ norm. Given positive numbers a_n and b_n for $n \geq 1$, let $a_n \lesssim b_n$ and $b_n \gtrsim a_n$ mean that a_n/b_n is bounded and let $a_n \asymp b_n$ mean that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Given random variables V_n for $n \geq 1$, let $V_n = O_P(b_n)$ mean that $\lim_{c \to \infty} \limsup_n P(|V_n| \geq cb_n) = 0$ and let $V_n = o_P(b_n)$ mean that $\lim_n P(|V_n| \geq cb_n) = 0$ for $c > 0$. For a random variable V, let E_n denote expectation relative to its empirical distribution; that is, $E_n(V) = n^{-1} \sum_i V_i$, where V_i, $1 \leq i \leq n$ is a random sample from the distribution of V.

Let $\|\cdot\|$ be a norm on $\mathbb{H}$ such that $\|h\| < \infty$ and $\|h\| \leq C_0 \|h\|_\infty$ for $h \in \mathbb{H}$, where C_0 is a fixed positive number. Without loss of generality, we assume that $C_0 = 1$ (otherwise, we replace $\|\cdot\|$ by $\|\cdot\|/C_0$). We also assume that if $h \in \mathbb{H}$ and $\|h\| = 0$, then h is essentially equal to zero. In our asymptotic theory we will use $\|\widehat{\eta} - \eta^*\|$ to measure the discrepancy between $\widehat{\eta}$ and η^*. A particular choice of norm could be the normalized L_2 norm relative to

Lebesgue measure on $\mathcal{U}$; that is, $\|h\|_{\text{Leb}} = \{\int_{\mathcal{U}} h^2(\boldsymbol{u})\,d\boldsymbol{u}/\text{vol}(\mathcal{U})\}^{1/2}$. Note that $\|h\|_{\text{Leb}} \le \|h\|_\infty$.

However, sometimes it is more natural to use other norms to measure the discrepancy. In the regression context, for example, one would use $\|h\|^2 = E[h^2(\boldsymbol{X})]$ where the expectation is with respect to the distribution of the covariates $\boldsymbol{X}$. Such a norm is closely related to the mean prediction error. Precisely, the mean prediction error of a candidate h for the regression function η is defined by $\text{PE}(h) = E\{[Y^* - h(\boldsymbol{X}^*)]^2\}$, where $(\boldsymbol{X}^*, Y^*)$ is a pair of observations independent of the observed data and having the same distribution as $(\boldsymbol{X}, Y)$. It is easily seen that

$$\text{PE}(h) = E[\text{var}(Y|\boldsymbol{X})] + \|h - \eta\|^2.$$

It is interesting to note that, under mild conditions (for example, if the density of $\boldsymbol{X}$ is bounded away from zero and infinity), the norm $\|\cdot\|$ is equivalent to the normalized L_2-norm $\|\cdot\|_{\text{Leb}}$. Later on (in Section 11.1.3 and 11.3), we will choose the norms for measuring discrepancy as those induced by the inner products used for defining functional ANOVA decompositions; see Section 11.4 for specification in each particular context.

We first present an asymptotic result that is applicable to general estimation space $\mathbb{G}$ and then specialize to estimation spaces built by polynomial splines. It involves some constants related to the estimation space $\mathbb{G}$. Set

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

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

and

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

The dimension N_n measures the size of the estimation space $\mathbb{G}$. Observe that $1 \le A_n < \infty$. The constant A_n can be thought of as a measure of the irregularity of the estimation space. Its magnitude can be determined by employing results in the approximation theory literature for various commonly used estimation spaces including polynomials, trigonometric polynomials, splines, wavelets, and finite elements. The constant ρ_n is the minimum L_∞ norm of the error when η^* is approximated by a function in $\mathbb{G}$. Through the use of results from approximation theory, the magnitude of ρ_n can be determined for commonly used estimation spaces if a smoothness condition is imposed on η^*. See Huang (1998a) for more discussion on these constants.

Proposition 11.1.1. *Suppose Conditions* 11.2.1, 11.2.2 *and* 11.2.4 *hold. Suppose also that* $\lim_n A_n \rho_n = 0$ *and* $\lim_n A_n^2 N_n / n = 0$. *Then* $\bar{\eta}$ *exists uniquely for* n *sufficiently large and*

$$\|\bar{\eta} - \eta^*\|^2 = O(\rho_n^2).$$

Moreover, $\widehat{\eta}$ exists uniquely except on an event whose probability tends to zero as $n \to \infty$ and

$$\|\widehat{\eta} - \bar{\eta}\|^2 = O_P\Big(\frac{N_n}{n}\Big).$$

Consequently,

$$\|\widehat{\eta} - \eta^*\|^2 = O_P\Big(\frac{N_n}{n} + \rho_n^2\Big).$$

In particular, $\widehat{\eta}$ is consistent in estimating η^*; that is, $\|\widehat{\eta} - \eta^*\| = o_P(1)$.

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

This result is rather general; it does not specify the form of $\mathbb{G}$. As a simple example, consider $\mathbb{H}$ being a finite-dimensional linear space of bounded functions. Take $\mathbb{G} = \mathbb{H}$, which does not depend on the sample size. Then both A_n and N_n are finite and independent of n and $\rho_n = 0$. Consequently, $\|\widehat{\eta} - \eta^*\|^2 = O_P(1/n)$, which is the parametric rate of convergence.

Proposition 11.1.1 is readily applicable to fixed knot spline estimates where the knot positions are prespecified but the number of knots are allowed to increase with the sample size.

Suppose $\mathcal{U}$ is the Cartesian product of compact intervals $\mathcal{U}_1, \dots, \mathcal{U}_L$. Consider the saturated model, in which η is a bounded function and no structural assumptions are imposed on η (that is, $\mathbb{H}$ is essentially the space of all square-integrable functions on $\mathcal{U}$). Consequently, $\eta^* = \eta$. To construct the estimation space, let $\mathbb{G}_l$ be the a linear space of splines with degree $q \geq p - 1$ for $1 \leq l \leq L$ and let $\mathbb{G}$ be the tensor product of $\mathbb{G}_1, \dots, \mathbb{G}_L$. Suppose the knots have bounded mesh ratio (that is, the ratios of the differences between consecutive knots are bounded away from zero and infinity uniformly in n). Let a_n denote the smallest distance between two consecutive knots.

To get the rates of convergence, we introduce a commonly used smoothness condition. Let $0 < \beta \leq 1$. A function h on $\mathcal{U}$ is said to satisfy a Hölder condition with exponent β if there is a positive number γ such that $|h(\boldsymbol{u}) - h(\boldsymbol{u}_0)| \leq \gamma |\boldsymbol{u} - \boldsymbol{u}_0|^\beta$ for $\boldsymbol{u}_0, \boldsymbol{u} \in \mathcal{U}$; here $|\boldsymbol{u}| = (\sum_{l=1}^L u_l^2)^{1/2}$ is the Euclidean norm of $\boldsymbol{u} = (u_1, \dots, u_L) \in \mathcal{U}$. Given an L-tuple $i = (i_1, \dots, i_L)$ of nonnegative integers, set $[i] = i_1 + \cdots + i_L$ and let D^i denote the differential operator defined by

$$D^i = \frac{\partial^{[i]}}{\partial u_1^{i_1} \dots \partial u_L^{i_L}}.$$

Let k be a nonnegative integer and set $p = k + \beta$. A function on $\mathcal{U}$ is said to be *p-smooth* if it is k times continuously differentiable on $\mathcal{U}$ and D^i satisfies a Hölder condition with exponent β for all i with $[i] = k$.

Corollary 11.1.1 (Saturated Model). *Suppose Conditions* 11.2.1, 11.2.2 *and* 11.2.4 *hold. Suppose also that η is p-smooth. Let $\mathbb{G}$ to be the tensor product spline space with degree $q \geq p-1$. Suppose the knots have bounded mesh ratio. If $p > L/2$, $\lim_n a_n = 0$, and $\lim_n na_n^{2L} = \infty$, then*

$$\|\widehat{\eta} - \eta\|^2 = O_P\Big(\frac{1}{na_n^L} + a_n^{2p}\Big).$$

In particular, for $a_n \asymp n^{-1/(2p+L)}$, we have that

$$\|\widehat{\eta} - \eta\|^2 = O_P(n^{-2p/(2p+L)}).$$

Proof. Since η is p-smooth, $\rho_n \asymp a_n^p \asymp N_n^{-p/L}$ [see (13.69) and Theorem 12.8 of Schumaker (1981)]. Note that $A_n \asymp N_n^{1/2} \asymp a_n^{-L/2}$ [see (12.4.11) and (12.4.12) or see Huang (1998a)]. Thus,

$$\lim_n A_n^2 N_n/n = 0 \Longleftrightarrow \lim_n na_n^{2L} = \infty$$

and

$$\lim_n A_n \rho_n = 0 \Longleftrightarrow \lim_n a_n^{p-L/2} = 0.$$

The result follow from Proposition 11.1.1. □

The choice of $a_n \asymp n^{-1/(2p+L)}$ balances the contributions to the error bound from the estimation error and the approximation error, that is, $1/(na_n^L) \asymp a_n^{2p}$. The resulting rate of convergence $n^{-2p/(2p+L)}$ actually is optimal: no estimate has a faster rate of convergence uniformly over the class of p-smooth functions (Stone 1982). The rate of convergence depends on two quantities: the specified smoothness p of the target function and the dimension L of the domain on which the target function is defined. Note the dependence of the rate of convergence on the dimension L: given the smoothness p, the larger the dimension, the slower the rate of convergence; moreover, the rate of convergence tends to zero as the dimension tends to infinity. This provides a mathematical description of a phenomenon commonly known as the "curse of dimensionality."

11.1.3 ANOVA modeling

One way to tame the curse of dimensionality is to impose an additive structure or, more generally, one involving just main effects and selected low-order interactions in a functional ANOVA decomposition. We will see that structural models (or unsaturated models) do yield faster rate of convergence in estimation. This section will give a relatively nontechnical description of such results. More technical discussion of functional ANOVA modeling will be given in Section 11.3, which contains of the proof of the main result of this section (Proposition 11.1.3).

To illustrate the idea of ANOVA modeling, suppose that $\mathcal{U} = \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{U}_3$, where $\mathcal{U}_1$, $\mathcal{U}_2$, and $\mathcal{U}_3$ are compact intervals having positive length. Any square-integrable function on $\mathcal{U}$ can be decomposed as

$$\begin{aligned}\eta(\boldsymbol{u}) = \eta_\emptyset &+ \eta_{\{1\}}(u_1) + \eta_{\{2\}}(u_2) + \eta_{\{3\}}(u_3) + \eta_{\{1,2\}}(u_1, u_2) \\ &+ \eta_{\{1,3\}}(u_1, u_3) + \eta_{\{2,3\}}(u_2, u_3) + \eta_{\{1,2,3\}}(u_1, u_2, u_3).\end{aligned} \tag{11.1.3}$$

For identifiability, we require that each nonconstant component be orthogonal to all possible values of the corresponding lower-order components relative to an appropriate inner product. The expression (11.1.3) can then be viewed as a functional version of analysis of variance (ANOVA) decomposition. Correspondingly, we call $\eta_\emptyset$ the constant component; $\eta_{\{1\}}(u_1), \eta_{\{2\}}(u_2)$, and $\eta_{\{3\}}(u_3)$ the main effect components; $\eta_{\{1,2\}}(u_1, u_2)$, $\eta_{\{1,3\}}(u_1, u_3)$, and $\eta_{\{2,3\}}(u_2, u_3)$ the two-factor interaction components; and $\eta_{\{1,2,3\}}(u_1, u_2, u_3)$ the three-factor interaction. The right side of (11.1.3) is referred to as the ANOVA decomposition of η.

If no structural assumption is imposed on η, we need to consider all the components in the above ANOVA decomposition. The resulting model is saturated. However, the desire to tame the curse of dimensionality leads us to employ unsaturated models, which discard some terms in the ANOVA decomposition. For example, removing all the interaction components in the above ANOVA decomposition of η, we get the additive model

$$\eta(\boldsymbol{u}) = \eta_\emptyset + \eta_{\{1\}}(u_1) + \eta_{\{2\}}(u_2) + \eta_{\{3\}}(u_3). \tag{11.1.4}$$

We can also include some selected interactions in the model and still keep the model manageable. For example, the following model includes just the interaction between u_1 and u_2:

$$\eta(\boldsymbol{u}) = \eta_\emptyset + \eta_{\{1\}}(u_1) + \eta_{\{2\}}(u_2) + \eta_{\{3\}}(u_3) + \eta_{\{1,2\}}(u_1, u_2). \tag{11.1.5}$$

To fit these models using maximum likelihood, it is necessary to choose the estimation space $\mathbb{G}$ to respect the imposed structure on η. As a result the estimate will have the same structure. To see precisely how this can be done, we need the notion of tensor product spaces. Now, for each $l = 1, 2, 3$, let $\mathbb{G}_{\{l\}}$ denote a suitable finite-dimensional linear space of functions of the variable u_l that contains the constant functions. For the additive model (11.1.4), we can take

$$\begin{aligned}\mathbb{G} &= \mathbb{G}_{\{1\}} + \mathbb{G}_{\{2\}} + \mathbb{G}_{\{3\}} \\ &= \{g_{\{1\}} + g_{\{2\}} + g_{\{3\}} : g_{\{l\}} \in \mathbb{G}_{\{l\}}, \text{ for } l = 1, 2, 3\}\end{aligned}$$

and the resulting maximum likelihood estimate has the form

$$\widehat{\eta}(\boldsymbol{u}) = \widehat{\eta}_\emptyset + \widehat{\eta}_{\{1\}}(u_1) + \widehat{\eta}_{\{2\}}(u_2) + \widehat{\eta}_{\{3\}}(u_3).$$

For the model (11.1.5) with a single interaction component, we can take

$$\begin{aligned}\mathbb{G} &= \mathbb{G}_{\{1\}} \otimes \mathbb{G}_{\{2\}} + \mathbb{G}_{\{3\}} \\ &= \{g_{\{1,2\}} + g_{\{3\}} : g_{\{l\}} \in \mathbb{G}_{\{l\}},\ g_{\{1,2\}} \in \mathbb{G}_{\{1\}} \otimes \mathbb{G}_{\{2\}}\},\end{aligned}$$

and the resulting maximum likelihood estimate has the form

$$\widehat{\eta}(\boldsymbol{u}) = \widehat{\eta}_{\emptyset} + \widehat{\eta}_{\{1\}}(u_1) + \widehat{\eta}_{\{2\}}(u_2) + \widehat{\eta}_{\{3\}}(u_3) + \widehat{\eta}_{\{1,2\}}(u_1, u_2).$$

On the other hand, to fit the saturated model, we can use the maximum likelihood estimate with $\mathbb{G}$ being the tensor product space $\mathbb{G}_{\{1,2,3\}} = \mathbb{G}_{\{1\}} \otimes \mathbb{G}_{\{2\}} \otimes \mathbb{G}_{\{3\}}$. Similar to (11.1.3), the resulting estimate should have the ANOVA decomposition

$$\begin{aligned}\widehat{\eta}(\boldsymbol{u}) = \widehat{\eta}_{\emptyset} &+ \widehat{\eta}_{\{1\}}(u_1) + \widehat{\eta}_{\{2\}}(u_2) + \widehat{\eta}_{\{3\}}(u_3) + \widehat{\eta}_{\{1,2\}}(u_1, u_2) \\ &+ \widehat{\eta}_{\{1,3\}}(u_1, u_3) + \widehat{\eta}_{\{2,3\}}(u_2, u_3) + \widehat{\eta}_{\{1,2,3\}}(u_1, u_2, u_3).\end{aligned}$$

In general, suppose that $\mathcal{U} = \mathcal{U}_1 \times \cdots \times \mathcal{U}_L$ for some positive integer L, where each $\mathcal{U}_l$ is a compact subset of some Euclidean space and it has positive volume in that space. If η is square-integrable, we can define its ANOVA decomposition in a similar manner as above. Selecting certain terms in its ANOVA decomposition in the modeling process corresponds to imposing a particular structural assumption on η. Specifically, let $\mathcal{S}$ be a hierarchical collection of subsets of $\{1, \ldots, L\}$. By hierarchical we mean that if $s \in \mathcal{S}$, then $r \in \mathcal{S}$ for $r \subset s$. Consider a model of the form

$$\eta = \sum_{s \in \mathcal{S}} \eta_s, \tag{11.1.6}$$

where η_s is a square-integrable function depending only on the variables $u_l, l \in s$, and $\eta_\emptyset$ is a constant. Note that the set $\mathcal{S}$ describes precisely which interaction terms are included in the model. For example, the additive model (11.1.4) and the model (11.1.5) with a single interaction component correspond to $\mathcal{S} = \{\emptyset, \{1\}, \{2\}, \{3\}\}$ and $\mathcal{S} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}\}$ respectively.

Again, the maximum likelihood method can be used to do the estimation and the fitting space $\mathbb{G}$ can be chosen to take the form

$$\mathbb{G} = \left\{ \sum_{s \in \mathcal{S}} g_s : g_s \in \mathbb{G}_s \text{ for } s \in \mathcal{S} \right\}, \tag{11.1.7}$$

where $\mathbb{G}_s$ is the tensor product space of $\mathbb{G}_l, l \in s$, and for each $1 \le l \le L$, $\mathbb{G}_l$ is an appropriate finite-dimensional space of functions of u_l that contains all constant functions. The resulting estimate should have the form

$$\widehat{\eta} = \sum_{s \in \mathcal{S}} \widehat{\eta}_s, \tag{11.1.8}$$

where $\widehat{\eta}_s$ is a member of $\mathbb{G}_s$ for $s \in \mathcal{S}$ and $\widehat{\eta}_\emptyset$ is a constant.

When the structural assumption is correctly specified, it is reasonable to expect that, if the dimension of $\mathbb{G}$ grows at the right rate while respecting the structural assumption, $\widehat{\eta}$ should be consistent in estimating η, that is, it should converge to η when the sample size tends to infinity. But what happens when the structural assumption is violated, or, what is $\widehat{\eta}$ estimating under model misspecification? This is a sensible question since in practice the postulated structural assumption is at best an approximation to reality.

Let $\mathbb{H}$ be the closed subspace of $L_2(\mathcal{U})$ that consists of all functions of the form (11.1.6), that is,

$$\mathbb{H} = \left\{ \sum_{s \in \mathcal{S}} h_s : h_s \in \mathbb{H}_s \text{ for } s \in \mathcal{S} \right\}, \tag{11.1.9}$$

where $\mathbb{H}_s$ is the space of square-integrable functions that depends only on u_l, $l \in s$. As discussed previously, the best approximation η^* in $\mathbb{H}$ to η is a sensible target no matter the structural assumption (that is, $\eta \in \mathbb{H}$) is true or not.

Proposition 11.1.1 is readily applicable to this situation. However, it is desirable to replace conditions with A_n and ρ_n by those with quantities that are more straightforward to determine. Consider the ANOVA decomposition

$$\eta^* = \sum_{s \in \mathcal{S}} \eta^*_s \tag{11.1.10}$$

of the best approximation η^* in $\mathbb{H}$ to η, where η^*_s is a member of $\mathbb{H}_s$ for $s \in \mathcal{S}$ and $\eta^*_\emptyset$ is a constant. Set

$$N_s = N_{sn}(\mathbb{G}_s) = \dim(\mathbb{G}_s), \qquad s \in \mathcal{S},$$

$$A_s = A_{sn}(\mathbb{G}_s) = \sup_{g \in \mathbb{G}_s} \left(\frac{\|g\|_\infty}{\|g\|} \right), \qquad s \in \mathcal{S},$$

and

$$\rho_s = \rho_{sn}(\eta^*_s, \mathbb{G}_s) = \inf_{g \in \mathbb{G}_s} \|g - \eta^*_s\|_\infty, \qquad s \in \mathcal{S}.$$

The constants A_s and ρ_s, which are analogs of the constants A_n and ρ_n, are defined on the tensor product spaces that constitute the estimation space $\mathbb{G}$. The proof of the next result will be given at the end of Section 11.3.

Proposition 11.1.2. *Suppose Conditions* 11.2.1, 11.2.2 *and* 11.2.4 *hold. Suppose also that* $\lim_n A_s \rho_{s'} = 0$ *and* $\lim_n A_s^2 N_{s'}/n = 0$ *for each pair* $s, s' \in \mathcal{S}$. *Then*

$$\|\widehat{\eta} - \eta^*\|^2 = O_P\left(\sum_{s \in \mathcal{S}} \left(\frac{N_s}{n} + \rho_s^2 \right) \right).$$

In particular, $\widehat{\eta}$ *is consistent in estimating* η^**; that is,* $\|\widehat{\eta} - \eta^*\| = o_P(1)$.

Since in practice one expects that examination of the components of $\widehat{\eta}$ should shed light on the shape of η^*, it is desirable that the components of $\widehat{\eta}$ are consistent in estimating the corresponding components of η^*. Specifically, suppose the target and the estimate have the ANOVA decomposition

$$\eta^* = \sum_{s \in \mathcal{S}} \eta_s^* \qquad \text{and} \qquad \widehat{\eta} = \sum_{s \in \mathcal{S}} \widehat{\eta}_s. \tag{11.1.11}$$

To impose identifiability constraints on the terms in each expansion to make it uniquely defined, we force each nonconstant component to be orthogonal to all possible values of the corresponding lower-order components relative to an appropriate inner product. (This will be more precisely described in Section 11.3.) The question is: under what conditions will $\widehat{\eta}_s$ are consistent is estimating η_s^* for $s \in \mathcal{S}$?

In imposing the identifiability constraints for ANOVA decompositions, we need use some inner products. Usually, one uses a theoretical inner product to decompose η^* and an empirical inner product to decompose $\widehat{\eta}$. For example, in the regression case, it is natural to define the theoretical and empirical inner products by $\langle h_1, h_2 \rangle = E[h_1(\boldsymbol{X})h_2(\boldsymbol{X})]$ and $\langle h_1, h_2 \rangle_n = (1/n)\sum_i [h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i)]$. The reason for using different inner products is that the theoretical inner product is often defined in terms of the data-generating distribution, and hence depends on unknown quantities, while the empirical inner product needs to be totally determined by the data since it will be used to decompose the estimate. Although using different inner products causes some technical difficulty, we still have the desired result as follows.

Proposition 11.1.3. *Suppose that in* (11.1.11) *the target function η^* and the estimate $\widehat{\eta}$ are decomposed according to a theoretical inner product and an empirical inner product respectively. Suppose Conditions* 11.2.1*,* 11.2.2*,* 11.2.4 *and* 11.3.1–11.3.3 *hold. Then*

$$\|\widehat{\eta}_s - \eta_s^*\|^2 = O_P\biggl(\sum_{s \in \mathcal{S}} \Bigl(\frac{N_s}{n} + \rho_s^2\Bigr)\biggr), \qquad s \in \mathcal{S}.$$

In particular, $\widehat{\eta}_s$ are consistent in estimating η_s^*, $s \in \mathcal{S}$.

This proposition follows from Proposition 11.1.2 and Theorem 11.3.3.

Propositions 11.1.2 and 11.1.3 are applicable to general estimation space $\mathbb{G}$. In particular, the estimation space $\mathbb{G}$ can be built by polynomial splines and their tensor products. We have the following result.

Corollary 11.1.2 (Unsaturated Model). *Suppose Conditions* 11.2.1*,* 11.2.2 *and* 11.2.4 *hold. Suppose each η_s^*, $s \in \mathcal{S}$, in* (11.1.10) *is p-smooth. Let the estimation space be given by* (11.1.7) *with each $\mathbb{G}_l$ being a linear space of degree q splines on $\mathcal{U}_l$ with $q \ge p - 1$. Suppose the knots have bounded mesh ratio. Set $d = \max_{s \in \mathcal{S}} \#(s)$. If $p > d/2$, $\lim_n a_n = 0$, and*

$\lim_n na_n^{2d} = \infty$, *then*

$$\|\widehat{\eta}_s - \eta_s^*\|^2 = O_P\Big(\frac{1}{na_n^d} + a_n^{2p}\Big), \qquad s \in \mathcal{S},$$

and

$$\|\widehat{\eta} - \eta^*\|^2 = O_P\Big(\frac{1}{na_n^d} + a_n^{2p}\Big).$$

In particular, for $a_n \asymp n^{-1/(2p+d)}$, *we have that*

$$\|\widehat{\eta}_s - \eta_s^*\|^2 = O_P(n^{-2p/(2p+d)}), \qquad s \in \mathcal{S},$$

and

$$\|\widehat{\eta} - \eta^*\|^2 = O_P(n^{-2p/(2p+d)}).$$

Proof. Similar to Corollary 11.1.1, we have that $A_s \asymp a_n^{-\#(s)/2}, N_s \asymp a_n^{-\#(s)}$, and $\rho_s \asymp a_n^p$ for $s \in \mathcal{S}$. Thus,

$$\lim_n A_s^2 N_{s'}/n = 0 \Longleftrightarrow \lim_n na_n^{\#(s)+\#(s')} = \infty$$

and

$$\lim_n A_s \rho_{s'} = 0 \Longleftrightarrow \lim_n a_n^{p-\#(s)/2} = 0.$$

The results follow from Propositions 11.1.2 and 11.1.3. □

The rate of convergence $n^{-2p/(2p+d)}$ for an unsaturated model should be compared with the rate $n^{-2p/(2p+L)}$ for the saturated model. Note here that d is the maximum order of interaction between the components of the argument variable $\boldsymbol{u}$. For the additive model, we have $d = 1$, so the rate is $n^{-2p/(2p+1)}$, which is the same as that for estimating a one-dimensional target function. For models with interaction of order 2, we have $d = 2$ with corresponding rate of convergence $n^{-2p/(2p+2)}$, the same rate for estimating a two-dimensional target function. Hence, for large L, we can achieve faster rates of convergence by considering structural models involving only low-order interactions in the ANOVA decomposition of the target function and thereby tame the curse of dimensionality.

11.2 The main result on rates of convergence

In this section we prove the main result on convergence rates (Proposition 11.1.1) and give precise statements of the necessary technical conditions. As noted previously, the error $\widehat{\eta} - \eta^*$ can be decomposed into two terms: the approximation error and the estimation error. It is insightful to treat them separately. The two theorems in this section, one for handling each term, together yield Proposition 11.1.1. The results are established under general conditions that synthesize the common features of various estimation problems that can be treated in the framework of concave extended linear models. The verification of some of these conditions in specific contexts will be given in Section 11.4.

11.2.1 Approximation Error.

Condition 11.2.1. *The best approximation η^* in $\mathbb{H}$ to η exists and there is a positive constant K_0 such that $\|\eta^*\|_\infty \le K_0$.*

Since we require $\Lambda(h)$ to be a strictly concave function of η, η^* is essentially uniquely defined. If $\eta \in \mathbb{H}$, this condition is just the requirement that η be bounded. In the regression context, if $\mathbb{H}$ is a Hilbert space, then η^* is just the orthogonal projection of η onto $\mathbb{H}$ relative to a certain inner product, which obviously exists (Huang 1998a). In general, the existence of η^* can be established by taking into account the specific properties of the log-likelihood; see, for example, Theorems 4.1 and 5.1 of Stone (1994), Theorem 1 of Kooperberg, Stone and Truong (1995a), Theorem 2.1 of Huang and Stone (1998), and Theorem 1 of Huang, Kooperberg, Stone and Truong (2000).

Condition 11.2.2. *For each pair of bounded functions $h_1, h_2 \in \mathbb{H}$, $\Lambda(h_1 + \alpha(h_2 - h_1))$ is twice continuously differentiable with respect to α. For any positive constant K, there are positive numbers M_1 and M_2 such that*

$$-M_1\|h_2 - h_1\|^2 \le \frac{d^2}{d\alpha^2}\Lambda(h_1 + \alpha(h_2 - h_1)) \le -M_2\|h_2 - h_1\|^2 \qquad (11.2.1)$$

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

It follows from Condition 11.2.2 that the restriction of $\Lambda(\cdot)$ to the bounded functions in $\mathbb{H}$ is strictly concave.

Lemma 11.2.1. *Suppose Conditions* 11.2.1 *and* 11.2.2 *hold. Let K_1 be a positive constant such that $K_1 > K_0$ with K_0 as in Condition* 11.2.1. *Then there are positive numbers M_3 and M_4 such that*

$$-M_3\|h - \eta^*\|^2 \le \Lambda(h) - \Lambda(\eta^*) \le -M_4\|h - \eta^*\|^2$$

for all $h \in \mathbb{H}$ with $\|h\|_\infty \le K_1$.

Proof. Let $h \in \mathbb{H}$ with $\|h\|_\infty \le K_1$. Since η^* maximizes $\Lambda(\cdot)$,

$$\frac{d}{d\alpha}\Lambda((1-\alpha)\eta^* + \alpha h)\bigg|_{\alpha=0} = 0.$$

Integrating by parts, we get that

$$\Lambda(h) - \Lambda(\eta^*) = \int_0^1 (1-\alpha) \frac{d^2}{d\alpha^2} \Lambda((1-\alpha)\eta^* + \alpha h)\, d\alpha.$$

The desired result now follows from Condition 11.2.2. □

Theorem 11.2.1 (Approximation Error). *Suppose Conditions* 11.2.1 *and* 11.2.2 *hold and that* $\lim_n A_n \rho_n = 0$. *Let* K_1 *be a positive constant such that* $K_1 > K_0$ *with* K_0 *as in Condition* 11.2.1. *Then,* $\bar{\eta}$ *exists uniquely and* $\|\bar{\eta}\|_\infty \le K_1$ *for* n *sufficiently large. Moreover,* $\|\bar{\eta} - \eta^*\|^2 = O(\rho_n^2)$.

Proof. Since $\mathbb{G}$ is finite-dimensional, it follows by a compactness argument that there is a function $g^* \in \mathbb{G}$ such that $\|g^* - \eta^*\|_\infty = \rho_n$. Let $a > 1$ denote a positive constant (to be determined later). Choose $g \in \mathbb{G}$ with $\|g - \eta^*\| \le a\rho_n$. Then, by the definition of A_n,

$$\|g - g^*\|_\infty \le A_n \|g - g^*\| \le A_n(\|g - \eta^*\| + \|\eta^* - g^*\|) \le A_n \rho_n (a+1),$$

so

$$\|g\|_\infty \le \|g - g^*\|_\infty + \|g^* - \eta^*\|_\infty + \|\eta^*\|_\infty \le A_n\rho_n(a+1) + \rho_n + \|\eta^*\|_\infty.$$

Since $\lim_n A_n \rho_n = 0$, we obtain that, for n sufficiently large, $\|g\|_\infty \le K_1$ for all $g \in \mathbb{G}$ with $\|g - \eta^*\| \le a\rho_n$. It now follows from Lemma 11.2.1 that, for n sufficiently large,

$$\Lambda(g) - \Lambda(\eta^*) \le -M_4 a^2 \rho_n^2 \quad \text{for all } g \in \mathbb{G} \text{ with } \|g - \eta^*\| = a\rho_n \tag{11.2.2}$$

and

$$\Lambda(g^*) - \Lambda(\eta^*) \ge -M_3 \rho_n^2. \tag{11.2.3}$$

Let a be chosen such that $a > \sqrt{M_3/M_4}$. Then it follows from (11.2.2) and (11.2.3) that, for n sufficiently large,

$$\Lambda(g) < \Lambda(g^*) \quad \text{for all } g \in \mathbb{G} \text{ with } \|g - \eta^*\| = a\rho_n.$$

Observe that $\|g^* - \eta^*\| < a\rho_n$. Therefore, we conclude from the definition of $\bar{\eta}$ and the strict concavity of $\Lambda(\cdot)$ that $\bar{\eta}$ exists uniquely and $\|\bar{\eta} - \eta^*\| < a\rho_n$ for n sufficiently large. Hence $\|\bar{\eta}\|_\infty \le K_1$ and $\|\bar{\eta} - \eta^*\|^2 = O(\rho_n^2)$. □

11.2.2 Estimation Error.

Condition 11.2.3. *There is a positive constant* K_0 *such that, for* n *sufficiently large, the best approximation* $\bar{\eta}$ *in* $\mathbb{G}$ *to* η *exists uniquely and* $\|\bar{\eta}\|_\infty \le K_0$.

Although Condition 11.2.3 is a consequence of Theorem 11.2.1, 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 11.2.2 below.

Condition 11.2.4. *For any pair of functions $g_1, g_2 \in \mathbb{G}$, $\ell(g_1+\alpha(g_2-g_1))$ is twice continuously differentiable with respect to α. Moreover, (i)*

$$\sup_{g\in\mathbb{G}} \frac{\left|\frac{d}{d\alpha}\ell(\bar{\eta}+\alpha g)\Big|_{\alpha=0}\right|}{\|g\|} = O_P\left(\left(\frac{N_n}{n}\right)^{1/2}\right);$$

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

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

for any $g_1, g_2 \in \mathbb{G}$ with $\|g_1\|_\infty \le K$ and $\|g_2\|_\infty \le K$, except on an event whose probability tends to zero as $n\to\infty$.

It follows from Condition 11.2.4(ii) that $\ell(\cdot)$ is strictly concave on $\mathbb{G}$.

Remark 11.2.1. *We give a sufficient condition for Condition 11.2.4(i) when the norm $\|\cdot\|$ is associated with an inner product $\langle\cdot,\cdot\rangle$ defined on $\mathbb{G}$. Let $\{\phi_j : 1\le j\le N_n\}$ be an orthonormal basis for $\mathbb{G}$ with respect to $\langle\cdot,\cdot\rangle$. Then each function $g\in\mathbb{G}$ can be represented uniquely as $g=\sum_j \beta_j\phi_j$, where $\beta_j=\langle g,\phi_j\rangle$ for $j=1,\cdots,N_n$. Let $\boldsymbol{\beta}$ denote the N_n-dimensional column vector with entries β_j. To indicate the dependence of g on $\boldsymbol{\beta}$, write $g(\cdot)=g(\cdot;\boldsymbol{\beta})$ and $\ell(g(\cdot;\boldsymbol{\beta}))=\ell(\boldsymbol{\beta})$. Let $\boldsymbol{S}(\boldsymbol{\beta})=\partial\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ denote the score at $\boldsymbol{\beta}$, that is, the N_n-dimensional column vector having entries $\partial\ell(\boldsymbol{\beta})/\partial\beta_j$. Write $\bar{\eta}=\sum_j\bar{\beta}_j\phi_j$ with $\bar{\beta}_j=\langle\bar{\eta},\phi_j\rangle$. Let $\bar{\boldsymbol{\beta}}$ be the column vector with entries $\bar{\beta}_j$. Then*

$$\frac{d}{d\alpha}\ell(\bar{\eta}+\alpha g)\Big|_{\alpha=0} = [\boldsymbol{S}(\bar{\boldsymbol{\beta}})]^T\boldsymbol{\beta}$$

and hence

$$\sup_{g\in\mathbb{G}} \frac{\left|\frac{d}{d\alpha}\ell(\bar{\eta}+\alpha g)\Big|_{\alpha=0}\right|}{\|g\|} \le |\boldsymbol{S}(\bar{\boldsymbol{\beta}})|.$$

Consequently, a sufficient condition for Condition 11.2.4(i) is that $|\boldsymbol{S}(\bar{\boldsymbol{\beta}})| = O_P(N_n/n)^{1/2}$.

Theorem 11.2.2 (Estimation Error). *Suppose Conditions* 11.2.3 *and* 11.2.4 *hold and that* $\lim_n A_n^2 N_n/n = 0$. *Let K_1 be a positive constant such that $K_1 > K_0$ with K_0 as in Condition* 11.2.3. *Then $\widehat{\eta}$ exists uniquely and $\|\widehat{\eta}\|_\infty \le K_1$, except on an event whose probability tends to zero as $n\to\infty$. Moreover, $\|\widehat{\eta}-\bar{\eta}\|^2 = O_P(N_n/n)$.*

Proof. Integrating by parts, we get that

$$\begin{aligned}\ell(g) = \ell(\bar\eta) &+ \frac{d}{d\alpha}\ell(\bar\eta + \alpha(g-\bar\eta))\Big|_{\alpha=0} \\ &+ \int_0^1 (1-\alpha)\frac{d^2}{d\alpha^2}\ell(\bar\eta+\alpha(g-\bar\eta))\,d\alpha, \qquad g \in \mathbb{G}.\end{aligned} \tag{11.2.4}$$

Let a be a positive number (to be determined later). Choose $g \in \mathbb{G}$ such that $\|g - \bar\eta\| \le a(N_n/n)^{1/2}$. Then, by the definition of A_n, $\|g-\bar\eta\|_\infty \le A_n\|g-\bar\eta\| \le a(A_n^2 N_n/n)^{1/2} = o(1)$. Thus, for n sufficiently large, $\|g\|_\infty \le K_1$ for all $g \in \mathbb{G}$ with $\|g - \bar\eta\| \le a(N_n/n)^{1/2}$. Consequently, it follows from Condition 11.2.4(ii) that, except on an event whose probability tends to zero as $n \to \infty$,

$$\int_0^1 (1-\alpha)\frac{d^2}{d\alpha^2}\ell(\bar\eta+\alpha(g-\bar\eta))\,d\alpha \le -\frac{M}{2}a^2\Big(\frac{N_n}{n}\Big) \tag{11.2.5}$$

for all $g \in \mathbb{G}$ with $\|g-\bar\eta\| = a(N_n/n)^{1/2}$. Fix an arbitrary positive constant ϵ. By Condition 11.2.4(i), we can choose a sufficiently large such that, except on an event whose probability is less than ϵ,

$$\left|\frac{d}{d\alpha}\ell(\bar\eta+\alpha(g-\bar\eta))\Big|_{\alpha=0}\right| < \frac{M}{2}a^2\Big(\frac{N_n}{n}\Big) \tag{11.2.6}$$

for all $g \in \mathbb{G}$ with $\|g-\bar\eta\| = a(N_n/n)^{1/2}$. Suppose (11.2.5) and (11.2.6) hold. Then, by (11.2.4), $\ell(g) < \ell(\bar\eta)$ for all $g \in \mathbb{G}$ with $\|g-\bar\eta\| = a(N_n/n)^{1/2}$. Hence, by the strict concavity of $\ell(\cdot)$, $\widehat\eta$ exists uniquely and $\|\widehat\eta - \bar\eta\| \le a(N_n/n)^{1/2}$. Since ϵ is arbitrary, the conclusions of the theorem follow. □

11.3 Functional ANOVA

In this section we elaborate on functional ANOVA decompositions and provide proofs of results in Section 11.1.3. As in Section 11.1.3, we assume that $\mathcal{U}$ is the Cartesian product of $\mathcal{U}_1, \ldots, \mathcal{U}_L$, where each $\mathcal{U}_l$ is a compact subset of some Euclidean space and it has positive volume in that space.

11.3.1 Functional ANOVA Decompositions

To introduce the notion of an ANOVA decomposition of functions, we need an inner product on the space of square-integrable functions on $\mathcal{U}$. Let ψ_l be a probability measure on $\mathcal{U}_l$ for $1 \le l \le L$ and let ψ be the corresponding product measure on $\mathcal{U}$. For example, if ψ_l is the uniform distribution on $\mathcal{U}_l$ for $1 \le l \le L$, then ψ is the uniform distribution on $\mathcal{U}$. The inner product induced by ψ is defined as

$$\begin{aligned}(f_1, f_2)_\psi &= \int_{\mathcal{U}} f_1 f_2 \, d\psi \\ &= \int_{\mathcal{U}_1} \cdots \int_{\mathcal{U}_L} f_1(u_1, \ldots, u_L) f_2(u_1, \ldots, u_L) \, d\psi_1(u_1) \ldots d\psi_L(u_L)\end{aligned}$$

with associated norm $\|f\|_\psi^2 = (f, f)_\psi$ for functions f, f_1 and f_2 on $\mathcal{U}$ that are square-integrable with respect to ψ.

Let $\mathbb{F}_\emptyset$ denote the space of constant functions on $\mathcal{U}$. Given $1 \le l \le L$, let $\mathbb{F}_l \supset \mathbb{F}_\emptyset$ denote a closed linear subspace of the space of functions on $\mathcal{U}_l$ that are square-integrable with respect to ψ_l. Given a nonempty subset $s = \{l_1, \ldots, l_k\}$ of $\{1, \ldots, L\}$, let $\mathbb{F}_s$ denote the tensor product of $\mathbb{F}_{l_1}, \ldots, \mathbb{F}_{l_k}$, which is the closure (relative to $\|\cdot\|_\psi$) of the space of functions on $\mathcal{U}$ spanned by functions of the form

$$f(\boldsymbol{u}) = \prod_{i=1}^{k} f_{l_i}(u_{l_i}), \qquad f_{l_i} \in \mathbb{F}_{l_i} \text{ and } \boldsymbol{u} = (u_1, \ldots, u_L) \in \mathcal{U}.$$

Let $\mathcal{S}$ denote a hierarchical collection of subsets of $\{1, \ldots, L\}$. Here, as in the previous section, hierarchical means that if s is a member of $\mathcal{S}$ and r is a subset of s, then r is a member of $\mathcal{S}$. Set

$$\mathbb{F} = \left\{ \sum_{s \in \mathcal{S}} f_s : f_s \in \mathbb{F}_s \text{ for } s \in \mathcal{S} \right\}.$$

It follows from Lemma 11.3.2 below that $\mathbb{F}$ is a closed subspace of the space of functions on $\mathcal{U}$ that are square-integrable with respect to ψ and hence that $\mathbb{F}$ is a Hilbert space equipped with the inner product $(\cdot, \cdot)_\psi$.

Set $\mathbb{F}_\emptyset^0 = \mathbb{F}_\emptyset$ and, for each nonempty set $s \in \mathcal{S}$, let $\tilde{\mathbb{F}}_s^0$ denote the space of functions in $\mathbb{F}_s$ that are orthogonal (relative to $(\cdot, \cdot)_\psi$) to each function in $\mathbb{F}_r$ for every proper subset r of s. Then the spaces $\tilde{\mathbb{F}}_s^0$, $s \in \mathcal{S}$, are orthogonal to each other and $\tilde{\mathbb{F}}_r^0$, $r \subset s$, are orthogonal spaces whose direct sum is $\mathbb{F}_s$ [see Takemura (1983)]. Consequently, we have the following result.

Lemma 11.3.2. *Each function $f \in \mathbb{F}$ has a unique decomposition $f = \sum_{s \in \mathcal{S}} f_s$, where $f_s \in \tilde{\mathbb{F}}_s^0$ for $s \in \mathcal{S}$. Moreover, $\|f\|_\psi^2 = \sum_{s \in \mathcal{S}} \|f\|_\psi^2$.*

We refer to $\sum_{s \in \mathcal{S}} f_s$, $f_s \in \tilde{\mathbb{F}}_s^0$, as the ANOVA decomposition of f relative to the inner product $(\cdot, \cdot)_\psi$, and to f_s, $s \in \mathcal{S}$, as the components of f in this decomposition. The component f_s is referred to as the constant component if $\#(s) = 0$, as a main effect component if $\#(s) = 1$, and as an interaction component if $\#(s) \ge 2$; here $\#(s)$ is the number of members of s.

Now let us consider functional ANOVA decompositions relative to an inner product that is not induced by a product measure. Let $(\cdot, \cdot)$ be an arbitrary inner product on $\mathbb{F}$ and denote the associated norm by $\|\cdot\|$.

Suppose the norms $\|\cdot\|$ and $\|\cdot\|_\psi$ are equivalent, that is, there are positive numbers M_1 and $M_2 \ge M_1$ such that

$$M_1 \|f\|_\psi \le \|f\| \le M_2 \|f\|_\psi, \qquad f \in \mathbb{F}. \tag{11.3.1}$$

Under (11.3.1), $\mathbb{F}$ is a Hilbert space equipped with the inner product $(\cdot,\cdot)$ and each $\mathbb{F}_s$, $s \in \mathcal{S}$, is closed relative to $\|\cdot\|$. For $s \in \mathcal{S}$, let $\mathbb{F}_s^0$ denote the space of functions in $\mathbb{F}_s$ that are orthogonal (relative to $(\cdot,\cdot)$) to each function in $\mathbb{F}_r$ for every proper subset r of s.

Lemma 11.3.3. *Suppose* (11.3.1) *holds. Set* $\epsilon = 1 - \sqrt{1-(M_1/M_2)^2} \in (0,1]$. *Then*

$$\|f\|^2 \ge \epsilon^{\#(\mathcal{S})-1} \sum_{s\in\mathcal{S}} \|f_s\|^2, \qquad f = \sum_s f_s, \quad \textit{where } f_s \in \mathbb{F}_s^0 \textit{ for } s \in \mathcal{S}.$$

It follows from this lemma that each function $f \in \mathbb{F}$ has an essentially unique decomposition $f = \sum_{s\in\mathcal{S}} f_s$, where $f_s \in \mathbb{F}_s^0$ for $s \in \mathcal{S}$. Such a decomposition is called the ANOVA decomposition relative to the inner product $(\cdot,\cdot)$.

Since the inner product $(\cdot,\cdot)$ is not necessarily induced by a product measure, the spaces $\mathbb{F}_s^0$ need not be orthogonal to each other. Thus we do not have the Pythagorean identity for the squared norms as in Lemma 11.3.2. Instead, it follows from the triangle inequality that $\|f\| \le \sum_{s\in\mathcal{S}} \|f_s\|$. The above lemma gives an inequality in the opposite direction.

The requirement (11.3.1) is not needed to define the ANOVA decomposition. It is used only to establish the relationship between the norm of a function and the norms of its ANOVA components. We require only that each $\mathbb{F}_s$, $s \in \mathcal{S}$, be closed relative to $\|\cdot\|$ for the above ANOVA decomposition to be well-defined.

Proof of Lemma 11.3.3. This proof is taken from the proof of Lemma 3.1 of Stone (1994). We proceed by induction on $\#(\mathcal{S})$. Observe that the lemma is trivially true when $\#(\mathcal{S}) = 1$. Suppose $\#(\mathcal{S}) \ge 2$ and that the desired result holds whenever $\mathcal{S}$ is replaced by $\mathcal{S}'$ with $\#(\mathcal{S}') < \#(\mathcal{S})$. Choose a "maximal" $r \in \mathcal{S}$ (that is, such that r is not a proper subset of any set s in $\mathcal{S}$). We first verify that

$$\Big\| \sum_s f_s \Big\|^2 \ge (M_1/M_2)^2 \|f_r\|^2. \tag{11.3.2}$$

If $\#(r) = L$, then (11.3.2) follows immediately from the definition of $\mathbb{F}_r^0$. Suppose, instead, that $1 \le \#(r) \le L-1$. For every $\boldsymbol{u} = (u_1, \dots, u_L) \in \mathcal{U}$, write $\boldsymbol{u} = (\boldsymbol{u}_r, \boldsymbol{u}_{r^c})$, where $\boldsymbol{u}_r$ consists of u_l, $l \in r$, and $\boldsymbol{u}_{r^c}$ consists of u_l, $l \notin r$. Then $\boldsymbol{u}_r$ is $\mathcal{U}_r$-valued and $\boldsymbol{u}_{r^c}$ is $\mathcal{U}_{r^c}$-valued. Let ψ_r denote the product of $\psi_l, l \in r$, and let ψ_{r^c} denote the product of $\psi_l, l \in r^c$. Observe that

$$\begin{aligned}\Big\|\!\Big\| \sum_s f_s \Big\|\!\Big\|^2 &\ge M_1^2 \int_{\mathcal{U}_{r^c}} \left[\int_{\mathcal{U}_r} \Big(f_r(\boldsymbol{u}_r) + \sum_{s \ne r} f_s(\boldsymbol{u}_r, \boldsymbol{u}_{r^c}) \Big)^2 \psi_r(d\boldsymbol{u}_r) \right] \psi_{r^c}(d\boldsymbol{u}_{r^c}) \\ &\ge (M_1/M_2)^2 \int_{\mathcal{U}_{r^c}} \Big\|\!\Big\| f_r + \sum_{s \ne r} f_s(\cdot, \boldsymbol{u}_{r^c}) \Big\|\!\Big\|^2 \psi_{r^c}(d\boldsymbol{u}_{r^c}).\end{aligned}$$

Now $\|\!| f_r + \sum_{s \ne r} f_s(\cdot, \boldsymbol{u}_{r^c}) \|\!|^2 \ge \|\!| f_r \|\!|^2$ for $\boldsymbol{u}_{r^c} \in \mathcal{U}_{r^c}$ by the maximality of r and the definition of $\mathbb{F}_r^0$, so (11.3.2) again holds. It follows from (11.3.2) that $\|\!| f_r - \beta \sum_{s \ne r} f_s \|\!|^2 \ge (M_1/M_2)^2 \|\!| f_r \|\!|^2$ for $\beta \in \mathbb{R}$. Setting

$$\beta = \frac{(f_r, \sum_{s \ne r} f_s)}{\|\!| \sum_{s \ne r} f_s \|\!|^2},$$

we get that

$$\Big| \Big(f_r, \sum_{s \ne r} f_s \Big) \Big|^2 \le [1 - (M_1/M_2)^2] \|\!| f_r \|\!|^2 \Big\|\!\Big\| \sum_{s \ne r} f_s \Big\|\!\Big\|^2.$$

Thus, by the induction hypothesis,

$$\begin{aligned}\Big\|\!\Big\| \sum_s f_s \Big\|\!\Big\|^2 &\ge \{1 - [1 - (M_1/M_2)^2]^{1/2}\} \Big(\|\!| f_r \|\!|^2 + \Big\|\!\Big\| \sum_{s \ne r} f_s \Big\|\!\Big\|^2 \Big) \\ &\ge \epsilon \Big(\|\!| f_r \|\!|^2 + \epsilon^{\#(\mathcal{S})-2} \sum_{s \ne r} \|\!| f_s \|\!|^2 \Big) \\ &\ge \epsilon^{\#(\mathcal{S})-1} \sum_s \|\!| f_s \|\!|^2.\end{aligned}$$

This completes the proof. □

11.3.2 Construction of Model Space and Estimation Spaces using Functional ANOVA

In this subsection we discuss construction of model and estimation spaces in extended linear modeling to incorporate structural assumptions using functional ANOVA decompositions.

Model Space

Let $\mathbb{H}_\emptyset$ denote the space of constant functions on $\mathcal{U}$. Given $1 \le l \le L$, let $\mathbb{H}_l \supset \mathbb{H}_\emptyset$ denote a finite- or infinite-dimensional, closed linear subspace of the space of Lebesgue square-integrable functions on $\mathcal{U}_l$. Given a nonempty subset $s = \{l_1, \ldots, l_k\}$ of $\{1, \ldots, L\}$, let $\mathbb{H}_s$ denote the tensor product of $\mathbb{H}_{l_1}, \ldots, \mathbb{H}_{l_k}$. Given a hierarchical collection $\mathcal{S}$ of subsets of $\{1, \ldots, L\}$, define the corresponding model space $\mathbb{H}$ by

$$\mathbb{H} = \left\{ \sum_{s \in \mathcal{S}} h_s : h_s \in \mathbb{H}_s \text{ for } s \in \mathcal{S} \right\}.$$

Let $\langle \cdot, \cdot \rangle$ be a theoretical inner product defined on the space of Lebesgue square-integrable functions on $\mathcal{U}$, which may depend on the data-generating distribution, and let $\| \cdot \|$ denote the associated norm.

Condition 11.3.1. *For the uniform distribution ψ on $\mathcal{U}$, there are positive numbers M_1 and $M_2 \geq M_1$ such that $M_1 \|h\|_\psi \leq \|h\| \leq M_2 \|h\|_\psi$ for $h \in \mathbb{H}$.*

Remark 11.3.1. *In the context of generalized regression and density estimation, $\| \cdot \|$ is the L_2 norm relative to a density function $f_{\boldsymbol{U}}$ on $\mathcal{U}$ that is bounded away from zero and infinity on this set (see Sections 11.4.3 and 11.4.4). Thus Condition 11.3.1 automatically holds with M_1 and M_2 being chosen so that $M_1^2/\mathrm{vol}(\mathcal{U}) \leq f_{\boldsymbol{U}} \leq M_2^2/\mathrm{vol}(\mathcal{U})$ on $\mathcal{U}$. A definition of theoretical inner product and norm and verification of Condition 11.3.1 will be given in Section 11.4.5.*

Set $\mathbb{H}_\emptyset^0 = \mathbb{H}_\emptyset$ and, for each nonempty set $s \in \mathcal{S}$, let $\mathbb{H}_s^0$ denote the space of functions in $\mathbb{H}_s$ that are orthogonal (relative to the theoretical inner product) to each function in $\mathbb{H}_r$ for every proper subset r of s. It follows from Lemma 11.3.3 that, under Condition 11.3.1, each function $h \in \mathbb{H}$ can be written essentially uniquely in the form $\sum_{s \in \mathcal{S}} h_s$, where $h_s \in \mathbb{H}_s^0$ for $s \in \mathcal{S}$. We refer to $\sum_{s \in \mathcal{S}} h_s$ as the *theoretical ANOVA decomposition* of h and to h_s, $s \in \mathcal{S}$, as the components of h in this decomposition. According to Lemma 11.3.3, under Condition 11.3.1,

$$\sum_{s \in \mathcal{S}} \|h_s\| \lesssim \|h\| \leq \sum_{s \in \mathcal{S}} \|h_s\|, \qquad h = \sum_{s \in \mathcal{S}} h_s, \text{ where } h_s \in \mathbb{H}_s^0 \text{ for } s \in \mathcal{S}.$$

Estimation Spaces

Let $\mathbb{G}_\emptyset$ denote the space of constant functions on $\mathcal{U}$, which has dimension $N_\emptyset = 1$. Given $1 \leq l \leq L$, let $\mathbb{G}_l$ ($\mathbb{G}_\emptyset \subset \mathbb{G}_l \subset \mathbb{H}_l$) denote a linear space of bounded, real-valued functions on $\mathcal{U}_l$, which may vary with the sample size and has finite, positive dimension $N_l = N_{ln}$. It is also supposed that if $g_l \in \mathbb{G}_l$ and $\|g_l\|_{\psi_l} = 0$, then $g_l = 0$ on $\mathcal{U}_l$. Given a nonempty subset $s = \{l_1, \ldots, l_k\}$ of $\{1, \ldots, L\}$, let $\mathbb{G}_s$ be the tensor product of $\mathbb{G}_{l_1}, \ldots, \mathbb{G}_{l_k}$. The estimation space $\mathbb{G}$ is defined by

$$\mathbb{G} = \left\{ \sum_{s \in \mathcal{S}} g_s : g_s \in \mathbb{G}_s \text{ for } s \in \mathcal{S} \right\}.$$

Note that the dimension of $\mathbb{G}_s$ is given by $N_s = \prod_{i=1}^k N_{l_i}$ and that the dimension N_n of $\mathbb{G}$ satisfies $N_n \asymp \sum_{s \in \mathcal{S}} N_s$. If $\mathbb{H}_l$ is finite-dimensional, we can choose $\mathbb{G}_l = \mathbb{H}_l$.

The theoretical inner product often involves some unknown quantities, for example, the probability distribution of the observations. However, to define ANOVA decomposition of an estimate, it is necessary to use an inner product that does not depend on unknown quantities. We assume that an inner product totally determined by the data is defined on the estimation space and refer to it as the empirical inner product. Let $\langle \cdot, \cdot \rangle_n$ and $\| \cdot \|_n$ denote the empirical inner product and the associated norm.

Condition 11.3.2.

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

A sufficient condition for Condition 11.3.2 will be given in Lemma 11.4.7.

Set $\mathbb{G}_\emptyset^0 = \mathbb{G}_\emptyset$ and, for each nonempty set $s \in \mathcal{S}$, let $\mathbb{G}_s^0$ denote the space of functions in $\mathbb{G}_s$ that are orthogonal (relative to the empirical inner product) to each function in $\mathbb{G}_r$ for every proper subset r of s. Suppose Conditions 11.3.1 and 11.3.2 hold. Then, except on an event whose probability tends to zero as $n \to \infty$, $\mathbb{G}$ is a Hilbert space equipped with the empirical inner product; hence, by Lemma 11.3.3, each function $g \in \mathbb{G}$ can be written uniquely in the form $\sum_{s \in \mathcal{S}} g_s$, where $g_s \in \mathbb{G}_s^0$ for $s \in \mathcal{S}$. Correspondingly, we refer to $\sum_{s \in \mathcal{S}} g_s$ as the *empirical ANOVA decomposition* of g, and we refer to g_s, $s \in \mathcal{S}$, as the components of g.

Let Q_s denote the empirical orthogonal projection (that is, the orthogonal projection relative to the empirical inner product) onto $\mathbb{G}_s$ for $s \in \mathcal{S}$.

Lemma 11.3.4. *Suppose Conditions* 11.3.1 *and* 11.3.2 *hold. Let* $0 < \epsilon < 1 - \sqrt{1 - (M_1/M_2)^2}$ *with* M_1, M_2 *given in Condition* 11.3.1. *Then, except on an event whose probability tends to zero as* $n \to \infty$,

$$\|g\|^2 \geq \epsilon^{\#(\mathcal{S})-1} \sum_{s \in \mathcal{S}} \|g_s\|^2 \tag{11.3.3}$$

and

$$\|g\|_n^2 \geq \epsilon^{\#(\mathcal{S})-1} \sum_{s \in \mathcal{S}} \|g_s\|_n^2 \tag{11.3.4}$$

for $g = \sum_{s \in \mathcal{S}} g_s$, *where* $g_s \in \mathbb{G}_s^0$ *for* $s \in \mathcal{S}$, *and*

$$\|g\|_n^2 \leq \epsilon^{1-\#(\mathcal{S})} \sum_{s \in \mathcal{S}} \|Q_s g\|_n^2 \tag{11.3.5}$$

for $g \in \mathbb{G}$.

Proof. By Conditions 11.3.1 and 11.3.2, there are positive constants M_1' and $M_2' \geq M_1'$ that are arbitrarily close to M_1 and M_2, respectively, and such that, except on an event whose probability tends to zero as $n \to \infty$, $M_1' \|g\|_\psi \leq \|g\|_n \leq M_2' \|g\|_\psi$ for $g \in \mathbb{G}$. Applying Lemma 11.3.3 to $\mathbb{F} = \mathbb{G}$ and $\| \cdot \| = \| \cdot \|_n$, we obtain (11.3.4). Then (11.3.3) follows from (11.3.4) and Condition 11.3.2. Write $g = \sum_{s \in \mathcal{S}} g_s$, where $g_s \in \mathbb{G}_s^0$ for $s \in \mathcal{S}$. Then, by the Cauchy–Schwarz inequality,

$$\begin{aligned}\|g\|_n^2 = \sum_{s \in \mathcal{S}} \langle g_s, g\rangle_n = \sum_{s \in \mathcal{S}} \langle g_s, Q_s g\rangle_n &\le \sum_{s \in \mathcal{S}} \|g_s\|_n \|Q_s g\|_n \\ &\le \Big(\sum_{s \in \mathcal{S}} \|g_s\|_n^2\Big)^{1/2} \Big(\sum_{s \in \mathcal{S}} \|Q_s g\|_n^2\Big)^{1/2},\end{aligned}$$

and (11.3.5) now follows from (11.3.4). □

11.3.3 Rates of Convergence of ANOVA Components

In this section we provide proof of Proposition 11.1.2. In particular we will prove that, under suitable conditions, the components of the maximum likelihood estimate in an appropriate ANOVA decomposition will converge to the corresponding components of the target function.

Condition 11.3.3. *Fix any subspace $\widetilde{\mathbb{G}}$ of $\mathbb{G}$ with dimension $\widetilde{N}_n$. Then, for any fixed sequence h_n, $n \ge 1$, of uniformly bounded functions on $\mathcal{U}$,*

$$\sup_{g \in \widetilde{\mathbb{G}}} \frac{|\langle h_n, g\rangle_n - \langle h_n, g\rangle|}{\|g\|} = O_P\Big(\Big(\frac{\widetilde{N}_n}{n}\Big)^{1/2}\Big).$$

A sufficient condition for Condition 11.3.3 will be given in Lemma 11.4.8.

The next theorem says that if $h \in \mathbb{H}$ and $g \in \mathbb{G}$ are close, then their components in appropriate ANOVA decompositions are also close.

Theorem 11.3.3. *Suppose that Conditions 11.3.1–11.3.3 hold. Let $\mathcal{S}_0$ be a (not necessarily hierarchical) subset of $\mathcal{S}$. Suppose that $h = \sum_{s \in \mathcal{S}_0} h_s$, where $h_s \in \mathbb{H}_s^0$ for $s \in \mathcal{S}_0$. Set $\gamma_s = \gamma_{sn} = \inf_{g \in \mathbb{G}_s} \|g - h_s\|_\infty$. Then*

$$\|g_s - h_s\|^2 = O_P\bigg(\|g - h\|^2 + \sum_{s \in \mathcal{S}_0} \Big(\frac{N_s}{n} + \gamma_s^2\Big)\bigg)$$

and

$$\|g_s - h_s\|_n^2 = O_P\bigg(\|g - h\|_n^2 + \sum_{s \in \mathcal{S}_0} \Big(\frac{N_s}{n} + \gamma_s^2\Big)\bigg)$$

uniformly for $g = \sum_{s \in \mathcal{S}_0} g_s$, where $g_s \in \mathbb{G}_s^0$ for $s \in \mathcal{S}$.

Suppose $\widehat{\eta} \in \mathbb{G}$ is an estimate of the target function η^* and consider the ANOVA decompositions $\eta^* = \sum_{s \in \mathcal{S}} \eta_s^*$ and $\widehat{\eta} = \sum_{s \in \mathcal{S}} \widehat{\eta}_s$, where $\eta_s^* \in \mathbb{H}_s^0$ and $\widehat{\eta}_s \in \mathbb{G}_s^0$ for $s \in \mathcal{S}$. Applying the above theorem to $h = \eta^*$ and $g = \widehat{\eta}$, we see that, that if the distance between $\widehat{\eta}$ and η^* is small, then the differences $\widehat{\eta}_s - \eta_s^*$ of the components of $\widehat{\eta}$ and η^* should also be small.

To prove Theorem 11.3.3, we need the following result.

Lemma 11.3.5. *Under the same setup as Theorem* 11.3.3, *for* $s \in \mathcal{S}$, *there is a function* $\tilde{g}_s \in \mathbb{G}_s^0$ *such that*

$$\|\tilde{g}_s - h_s\|^2 = O_P\Big(\frac{N_s}{n} + \gamma_s^2\Big) \tag{11.3.6}$$

and

$$\|\tilde{g}_s - h_s\|_n^2 = O_P\Big(\frac{N_s}{n} + \gamma_s^2\Big). \tag{11.3.7}$$

Proof. Without loss of generality, assume that $\gamma_s < \infty$. Then h_s is bounded and, by a compactness argument, there is a function $g_s^* \in \mathbb{G}_s$ such that $\|g_s^* - h_s\|_\infty = \inf_{g \in \mathbb{G}_s} \|g - h_s\|_\infty = \gamma_s$. Moreover, $\|g_s^*\|_\infty \le \gamma_s + \|h_s\|_\infty \le 2\|h_s\|_\infty$. Write $g_s^* = \tilde{g}_s + (g_s^* - \tilde{g}_s)$, where $\tilde{g}_s \in \mathbb{G}_s^0$ and $g_s^* - \tilde{g}_s \in \sum_{r \subset s, r \ne s} \mathbb{G}_r$ is the empirical orthogonal projection of g_s^* onto $\mathbb{G}' := \sum_{r \subset s, r \ne s} \mathbb{G}_r$. We will verify that $\tilde{g}_s$ has the desired property.

It follows from Lemma 11.3.4 applied to $\mathbb{G}'$ that, except on an event whose probability tends to zero as $n \to \infty$, $\|g_s^* - \tilde{g}_s\|_n^2 \lesssim \sum_{r \subset s, r \ne s} \|Q_r g_s^*\|_n^2$. It follows from the triangle inequality and the fact that Q_r is an orthogonal projection that, for each proper subset r of s,

$$\|Q_r g_s^*\|_n \le \|Q_r(g_s^* - h_s)\|_n + \|Q_r h_s\|_n \le \|g_s^* - h_s\|_n + \|Q_r h_s\|_n.$$

Since $E[\|g_s^* - h_s\|_n^2] = \|g_s^* - h_s\|^2 \le \gamma_s^2$, we obtain that $\|g_s^* - h_s\|_n = O_P(\gamma_s)$. On the other hand,

$$\begin{aligned} \|Q_r h_s\|_n &= \sup_{g \in \mathbb{G}_r} \frac{|\langle Q_r h_s, g\rangle_n|}{\|g\|_n} \\ &= \sup_{g \in \mathbb{G}_r} \frac{|\langle h_s, g\rangle_n|}{\|g\|_n} \\ &= \sup_{g \in \mathbb{G}_r} \frac{|\langle h_s, g\rangle_n - \langle h_s, g\rangle|}{\|g\|_n}, \end{aligned}$$

the last step being valid since $h_s \in \mathbb{H}_s^0$. Therefore, $\|Q_r h_s\|_n = O_P((N_r/n)^{1/2})$ by Condition 11.3.3. Consequently,

$$\|g_s^* - \tilde{g}_s\|_n^2 = O_P\Big(\sum_{r \subset s, r \ne s} \frac{N_r}{n} + \gamma_s^2\Big) = O_P\Big(\frac{N_s}{n} + \gamma_s^2\Big).$$

Thus, by Condition 11.3.2,

$$\|g_s^* - \tilde{g}_s\|^2 = O_P\Big(\frac{N_s}{n} + \gamma_s^2\Big).$$

The desired results now follow from the triangle inequality. □

Proof of Theorem 11.3.3. We prove only the first result, the proof of the second result being similar. By Lemma 11.3.5, for each $s \in \mathcal{S}_0$, there is a function $\tilde{g}_s \in \mathbb{G}_s^0$ such that (11.3.6) holds. Set $\tilde{g} = \sum_{s \in \mathcal{S}_0} \tilde{g}_s$. Then

$$\|\tilde{g} - h\|^2 = O_P\biggl(\sum_{s \in \mathcal{S}_0} \Bigl(\frac{N_s}{n} + \gamma_s^2\Bigr)\biggr).$$

It follows from the triangle inequality that

$$\|g - \tilde{g}\|^2 \le 2\|g - h\|^2 + O_P\biggl(\sum_{s \in \mathcal{S}_0} \Bigl(\frac{N_s}{n} + \gamma_s^2\Bigr)\biggr).$$

The first result now follows from Lemma 11.3.4, (11.3.6), and the triangle inequality. □

Remark 11.3.2. *A much simpler version of Theorem 11.4.1 and its proof is available when the spaces $\mathbb{H}_s^0$ and $\mathbb{G}_s^0$, $s \in \mathcal{S}$, are defined according to the same inner product that is induced by a product measure. In this case, $\mathbb{G}_s^0 \subset \mathbb{H}_s^0$ for $s \in \mathcal{S}$ and the spaces $\mathbb{H}_s^0$, $s \in \mathcal{S}$, are orthogonal to each other. Let $\mathcal{S}_0$ be a (not necessarily hierarchical) subset of $\mathcal{S}$. Suppose that $h = \sum_{s \in \mathcal{S}_0} h_s$ and $g = \sum_{s \in \mathcal{S}_0} g_s$, where $h_s \in \mathbb{H}_s^0$ and $g_s \in \mathbb{G}_s^0$ for $s \in \mathcal{S}_0$. Then $h_s - g_s \in \mathbb{H}_s^0$ and hence*

$$\|h - g\|^2 = \sum_{s \in \mathcal{S}_0} \|h_s - g_s\|^2. \tag{11.3.8}$$

It follows from (11.3.8) *that if g and h are close, then their components are also close. As an application, suppose $\widehat{\eta}$ is an estimate of the target η^* and consider the ANOVA decompositions $\eta^* = \sum_{s \in \mathcal{S}_0} \eta_s^*$ and $\widehat{\eta} = \sum_{s \in \mathcal{S}_0} \widehat{\eta}_s$, where $\eta_s^* \in \mathbb{H}_s^0$ and $\widehat{\eta}_s \in \mathbb{G}_s^0$ for $s \in \mathcal{S}_0$. Applying* (11.3.8) *to $h = \eta^*$ and $g = \widehat{\eta}$, we obtain that $\|\widehat{\eta}_s - \eta_s^*\| \le \|\widehat{\eta} - \eta^*\|$ for $s \in \mathcal{S}_0$. In particular, the convergence of $\widehat{\eta}$ to η^* guarantees the convergence of the components of $\widehat{\eta}$ to those of η^*.*

Proof of Proposition 11.1.2. For each $g \in \mathbb{G}$, write $g = \sum_{s \in \mathcal{S}} g_s$ where $g_s \in \mathbb{G}_s$ and $g_s \perp \mathbb{G}_r$ for all proper subsets r of s. It follows from Lemma 11.3.3 that $\sum_{s \in \mathcal{S}} \|g_s\|^2 \le \epsilon^{1-\#(\mathcal{S})}\|g\|^2$, where $\epsilon = 1 - \sqrt{1 - (M_1/M_2)^2}$ with M_1, M_2 as in Condition 11.3.1. By the definition of A_s and the Cauchy–Schwarz inequality, we get that

$$\|g\|_\infty \le \sum_{s \in \mathcal{S}} \|g_s\|_\infty \le \sum_{s \in \mathcal{S}} A_s \|g_s\| \le \biggl(\sum_{s \in \mathcal{S}} A_s^2\biggr)^{1/2} \biggl(\sum_{s \in \mathcal{S}} \|g_s\|^2\biggr)^{1/2}.$$

Thus

$$\|g\|_\infty \le \biggl(\sum_{s \in \mathcal{S}} A_s^2\biggr)^{1/2} [\epsilon^{1-\#(\mathcal{S})}\|g\|^2]^{1/2}.$$

Consequently, $A_n \le [\epsilon^{1-\#(\mathcal{S})} \sum_{s\in\mathcal{S}} A_s^2]^{1/2}$. Observe that $N_n \le \sum_{s\in\mathcal{S}} N_s$ and $\rho_n \le \sum_{s\in\mathcal{S}} \rho_s$. Hence, it follows from the assumption $\lim_n A_s\rho_{s'} = 0$ for $s, s' \in \mathcal{S}$ that $\lim_n A_n\rho_n = 0$. Furthermore, it follows from the assumption $\lim_n A_s^2 N_{s'}/n = 0$ for $s, s' \in \mathcal{S}$ that $\lim_n A_n^2 N_n/n = 0$. The desired result then follows from Proposition 11.1.1. □

11.4 Verification of Technical Conditions

In this section we will verify the technical conditions (namely, Conditions 11.2.2, 11.2.4, 11.3.1–11.3.3) required in Propositions 11.1.1, 11.1.2 and 11.1.3. General sufficient conditions for Conditions 11.3.2 and 11.3.3 will be given in Section 11.4.2. Sections 11.4.3–11.4.5 will consider respectively the contexts of generalized regression, density estimation and hazard regression.

11.4.1 Preliminaries

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

$$|E\xi_i^m| \le \frac{m!}{2} b_i^2 H^{m-2}, \qquad m \ge 2, \tag{11.4.1}$$

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

$$P(|\bar{\xi}| \ge t) \le 2\exp\Big(-\frac{nt^2}{2(B_n^2 + tH)}\Big) \tag{11.4.2}$$

for $t > 0$. Suppose, in particular, that $E\xi_i = 0$, $\operatorname{var}(\xi_i) \le \sigma^2$, and $P(|\xi_i| \le b) = 1$ for $1 \le i \le n$, where $b > 0$. Then (11.4.1) and hence (11.4.2) hold with $b_i = \sigma$ for $1 \le i \le n$, $B_n = \sigma$, and H replaced by b. In this case, however, (11.4.2) 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 $|E\xi_i| \le b$ and hence $|\xi_i - E\xi_i| \le 2b$.) It follows easily from (11.4.2) that

$$P\big(|\bar{\xi}| \ge tH^{-1}[B_n(N_n/n)^{1/2} + N_n/n]\big) \le 2\exp\Big(-\frac{tH^{-2}N_n}{2}\Big) \tag{11.4.3}$$

for $t \ge 1$.

In the proofs of Lemmas 11.4.7, 12.3.7, and 12.3.11, we will use a "chaining argument" that is well known in the empirical process theory literature; see, for example, Pollard (1984). For convenience, we summarize a portion of this argument in the form of the following result.

Lemma 11.4.6 (Chaining argument). *Let $\mathbb{S}$ be a nonempty subset of $\widetilde{\mathbb{S}}$; let V_s, $s \in \widetilde{\mathbb{S}}$, be random variables; let K be a positive integer; let $\mathbb{S}_k$ be a finite, nonempty subset of $\widetilde{\mathbb{S}}$ for $0 \le k \le K$ such that $V_s = 0$ for $s \in \mathbb{S}_0$; let $C_1, \ldots, C_6$ be positive numbers; let $0 < \delta \le 1/4$; and let Ω be an event. Suppose that*

$$P\Big(\sup_{s \in \mathbb{S}} \min_{\widetilde{s} \in \mathbb{S}_K} |V_s - V_{\widetilde{s}}| > C_1; \Omega\Big) \le C_2; \tag{11.4.4}$$

$$\#(\mathbb{S}_k) \le C_3 \exp(C_4 k), \qquad 1 \le k \le K; \tag{11.4.5}$$

and

$$\begin{aligned} \max_{s \in \mathbb{S}_k} \min_{\widetilde{s} \in \mathbb{S}_{k-1}} & P(|V_s - V_{\widetilde{s}}| > 2^{-(k-1)} C_5; \Omega) \\ & \le C_6 \exp\big(-2C_4 (2\delta)^{-(k-1)}\big), \qquad 1 \le k \le K. \end{aligned} \tag{11.4.6}$$

Then

$$\begin{aligned} P\Big(\sup_{s \in \mathbb{S}} |V_s| > C_1 + 2C_5\Big) & \le C_2 + \frac{C_3 C_6}{\exp(C_4) - 1} + P(\Omega^c) \\ & \le C_2 + \frac{C_3 C_6}{C_4} + P(\Omega^c). \end{aligned}$$

Proof. Observe that

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

So, in light of (11.4.4), it suffices to verify that

$$P\Big(\max_{s \in \mathbb{S}_K} |V_s| > 2C_5; \Omega\Big) \le C_3 C_6 \frac{\exp(-C_4)}{1 - \exp(-C_4)}. \tag{11.4.7}$$

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

$$\begin{aligned} & P(|V_s - V_{\sigma_{k-1}(s)}| \ge 2^{-(k-1)} C_5; \Omega) \\ & \qquad \le C_6 \exp\big(-2C_4 (2\delta)^{-(k-1)}\big), \qquad 1 \le k \le K \text{ and } s \in \mathbb{S}_k; \end{aligned}$$

the existence of σ_{k-1} follows from (11.4.6). Then, by (11.4.5),

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

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

$$\begin{aligned} C_3 C_6 \sum_{k=1}^{K} \exp\big(-C_4 (2\delta)^{-(k-1)}\big) & \le C_3 C_6 \sum_{k=1}^{K} \exp(-C_4 k) \\ & \le C_3 C_6 \frac{\exp(-C_4)}{1 - \exp(-C_4)}. \end{aligned}$$

Suppose that $|V_s - V_{\sigma_{k-1}(s)}| \le 2^{-(k-1)}C_5$ for $1 \le k \le K$ and $s \in \mathbb{S}_k$. Choose $s \in \mathbb{S}_K$ and set $s_K = s$, $s_{K-1} = \sigma_{K-1}(s_K)$, …, $s_0 = \sigma_0(s_1)$. (We refer to $s_K, \dots, s_0$ as forming a "chain" from the point $s \in \mathbb{S}_K$ to a point $s_0 \in \mathbb{S}_0$.) Then $V_{s_0} = 0$ and $|V_{s_k} - V_{s_{k-1}}| < 2^{-(k-1)}C_5$ for $1 \le k \le K$, so

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

Consequently,

$$\begin{aligned}&P\Big(\max_{s\in\mathbb{S}_K}|V_s| > 2C_5;\Omega\Big)\\ &\le P\big(|V_s - V_{\sigma_{k-1}(s)}| > 2^{-(k-1)}C_5 \text{ for some } k \in \{1,\dots,K\} \text{ and } s \in \mathbb{S}_k;\Omega\big).\end{aligned}$$

Thus (11.4.7) holds as desired. □

11.4.2 Theoretical and Empirical Inner Products

Consider a $\mathcal{W}$-valued random variable $\boldsymbol{W}$, where $\mathcal{W}$ is an arbitrary set. Let $\boldsymbol{W}_1, \dots, \boldsymbol{W}_n$ be a random sample of size n from the distribution of $\boldsymbol{W}$. For any function f on $\mathcal{W}$, set $E(f) = E[f(\boldsymbol{W})]$ and $E_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{W}_i)$. Let $\mathcal{U}$ be another arbitrary set. We consider a real-valued functional $\Psi(f,g;\boldsymbol{w})$ defined on $\boldsymbol{w} \in \mathcal{W}$ and functions f,g on $\mathcal{U}$. For fixed functions f and g on $\mathcal{U}$, $\Psi(f,g;\boldsymbol{w})$ is a function on $\mathcal{W}$. For notational simplicity, write $\Psi(f,g) = \Psi(f,g;\boldsymbol{w})$. We assume that Ψ is symmetric and bilinear in its first two arguments: given functions $f, \widetilde{f}, g$ on $\mathcal{U}$, $\Psi(f,g) = \Psi(g,f)$ and $\Psi(af + b\widetilde{f}, g) = a\Psi(f,g) + b\Psi(\tilde{f}, g)$ for $a, b \in \mathbb{R}$. We also assume that $\Psi(f,f)$ is nonnegative.

Throughout this subsection, let the empirical inner product and norm be defined by

$$\langle f,g\rangle_n = E_n\big[\Psi(f,g)\big] \quad \text{and} \quad \|f\|_n^2 = \langle f,f\rangle_n,$$

and let the theoretical versions of these quantities be defined by

$$\langle f,g\rangle = E\big[\Psi(f,g)\big] \quad \text{and} \quad \|f\|^2 = \langle f,f\rangle.$$

In particular, this more general definition of the theoretical norm is now used in the definition of the constant A_n. We assume that

$$\Psi(f,f) \le M\|f\|_\infty^2 \tag{11.4.8}$$

for some positive constant M and

$$\Psi^2(f,g) \le \Psi(f,f)\Psi(g,g). \tag{11.4.9}$$

It follows from (11.4.8) and (11.4.9) that

$$|\Psi(f,g)| \le M\|f\|_\infty \|g\|_\infty \tag{11.4.10}$$

and

$$\operatorname{var}(\Psi(f,g)) \le M\|f\|^2\|g\|_\infty^2. \tag{11.4.11}$$

In the context of generalized regression and density estimation in Sections 11.4.3 and 11.4.4, we choose $\Psi(f,g) = f(\boldsymbol{U})g(\boldsymbol{U})$. See Section 11.4.5 for the choice of Ψ in the context of hazard regression.

The next result says that the empirical inner product is uniformly close to the theoretical inner product on the estimation space $\mathbb{G} = \mathbb{G}_n$ with probability tending to one. It provides a sufficient condition for Condition 11.3.2.

Lemma 11.4.7. *Suppose that* $\lim_n A_n^2 N_n/n = 0$ *and let* $t > 0$*. Then, for n sufficient large,*

$$P\Bigg(\sup_{f,g\in\mathbb{G},\|f\|\neq 0,\|g\|\neq 0} \frac{|\langle f,g\rangle_n - \langle f,g\rangle|}{\|f\|\,\|g\|} > t\Bigg) \le \frac{2}{\exp\Big(\frac{t^2}{16M(9+3t)}\frac{n}{A_n^2}\Big) - 1},$$

where M is as in (11.4.8)*. Consequently, except on an event whose probability tends to zero as* $n \to \infty$*,*

$$|\langle f,g\rangle_n - \langle f,g\rangle| \le t\,\|f\|\,\|g\|, \qquad f,g\in\mathbb{G}.$$

Proof. Set $\mathbb{B} := \{g \in \mathbb{G} : \|g\| \le 1\}$. Then

$$\sup_{f,g\in\mathbb{G},\|f\|\neq 0,\|g\|\neq 0} \frac{|\langle f,g\rangle_n - \langle f,g\rangle|}{\|f\|\,\|g\|} = \sup_{f,g\in\mathbb{B}} |\langle f,g\rangle_n - \langle f,g\rangle|.$$

Let $0 < \delta \le 1/4$, and let $\{0\} = \mathbb{B}_0 \subset \mathbb{B}_1 \subset \cdots$ be a sequence of subsets of $\mathbb{B}$ with the following property: for $k \ge 1$, $\mathbb{B}_k$ is a maximal superset of $\mathbb{B}_{k-1}$ such that each pair of functions in $\mathbb{B}_k$ is at least δ^k apart in the norm $\|\cdot\|$. Then $\min_{\widetilde{g}\in\mathbb{B}_k} \|g - \widetilde{g}\| \le \delta^k$ for $k \ge 0$ and $g \in \mathbb{B}$. Moreover,

$$\#(\mathbb{B}_k) \le \Big(\frac{1+\delta^k/2}{\delta^k/2}\Big)^{N_n} \le (3\delta^{-k})^{N_n}, \qquad k \ge 1. \tag{11.4.12}$$

(Observe that there are $\#(\mathbb{B}_k)$ disjoint balls each with radius $\delta^k/2$ that together can be covered by a ball with radius $1 + \delta^k/2$.)

Let $K = K_n$ be a positive integer such that $2MA_n^2\delta^K \le t$. We will apply Lemma 11.4.6 with $s = (f,g)$, $V_s = \langle f,g\rangle_n - \langle f,g\rangle = (E_n - E)\Psi(f,g)$, $\mathbb{S} = \{(f,g) : f,g \in \mathbb{B}\}$, $\mathbb{S}_k = \{(f,g) : f,g\in\mathbb{B}_k\}$ for $0 \le k \le K$, and $\Omega^c = \emptyset$. It follows from (11.4.12) that $\#(\mathbb{S}_k) \le (3\delta^{-k})^{2N_n}$ for $1 \le k \le K$ and hence that (11.4.5) holds with $C_3 = 1$ and any $C_4 \ge 2N_n \log(3/\delta)$.

Let k be a positive integer, and let $f, \widetilde{f}, g, \widetilde{g} \in \mathbb{B}$, where $\|f - \widetilde{f}\| \le \delta^{k-1}$ and $\|g - \widetilde{g}\| \le \delta^{k-1}$. Then, by (11.4.10) and the triangle inequality,

$$\begin{aligned}|\Psi(f,g) - \Psi(\widetilde{f},\widetilde{g})| &\le |\Psi(f-\widetilde{f},g)| + |\Psi(\widetilde{f}, g-\widetilde{g})| \\ &\le M\|f-\widetilde{f}\|_\infty \|g\|_\infty + M\|\widetilde{f}\|_\infty \|g-\widetilde{g}\|_\infty \\ &\le MA_n^2 \|f-\widetilde{f}\|\,\|g\| + MA_n^2\|\widetilde{f}\|\,\|g-\widetilde{g}\| \\ &\le 2MA_n^2\delta^{k-1};\end{aligned}$$

and, by (11.4.11)

$$\begin{aligned}&\operatorname{var}\big[\Psi(f,g) - \Psi(\widetilde{f},\widetilde{g})\big] \\ &\quad\le 2\operatorname{var}\big[\Psi(f-\widetilde{f},g)\big] + 2\operatorname{var}\big[\Psi(\widetilde{f},g-\widetilde{g})\big] \\ &\quad\le 2M\|g\|_\infty^2\|f-\widetilde{f}\|^2 + 2M\|\widetilde{f}\|_\infty^2\|g-\widetilde{g}\|^2 \\ &\quad\le 2MA_n^2(\|g\|^2\|f-\widetilde{f}\|^2 + \|\widetilde{f}\|^2\|g-\widetilde{g}\|^2) \\ &\quad\le 4MA_n^2\delta^{2(k-1)}.\end{aligned}$$

Applying (11.4.2) and noting that $0 < 2\delta \le 1$, we get that

$$\begin{aligned}&P\Big(\big|(E_n - E)\big(\Psi(f,g) - \Psi(\widetilde{f},\widetilde{g})\big)\big| > t2^{-(k-1)}\Big) \\ &\qquad \le 2\exp\Big(-\frac{(n/A_n^2)t^2(2\delta)^{-(k-1)}}{8M(1+t)}\Big).\end{aligned} \tag{11.4.13}$$

For any $f, g \in \mathbb{B}$, there exist $\widetilde{f}, \widetilde{g} \in \mathbb{B}_K$ such that $\|f - \widetilde{f}\| \le \delta^K$ and $\|g - \widetilde{g}\| \le \delta^K$. Then $|\Psi(f,g) - \Psi(\widetilde{f},\widetilde{g})| \le 2MA_n^2\delta^K < t$. Consequently, (11.4.4) holds with $C_1 = t$ and $C_2 = 0$.

Let $1 \le k \le K$. For any $f, g \in \mathbb{B}_k$, there are $\widetilde{f}, \widetilde{g} \in \mathbb{B}_{k-1}$ such that $\|f - \widetilde{f}\| \le \delta^{k-1}$ and $\|g - \widetilde{g}\| \le \delta^{k-1}$. Since $\lim_n A_n^2 N_n/n = 0$, we now conclude from (11.4.13) that (11.4.6) holds with $C_5 = t$, $C_6 = 2$, and

$$C_4 = \frac{(n/A_n^2)t^2}{16M(1+t)} \ge 2N_n \log\frac{3}{\delta}$$

for n sufficiently large. It now follows from Lemma 11.4.6 that, for n sufficiently large,

$$P\Big(\sup_{f,g\in\mathbb{B}} |\langle f,g\rangle_n - \langle f,g\rangle| > 3t\Big) \le \frac{2}{\exp\Big(\frac{t^2}{16M(1+t)}\frac{n}{A_n^2}\Big) - 1},$$

which tends to zero as $n \to \infty$. Thus the desired conclusion is valid. □

The next result provides a sufficient condition for Condition 11.3.3.

Lemma 11.4.8. *Let $\widetilde{\mathbb{G}} = \widetilde{\mathbb{G}}_n$ be a linear space of functions with finite dimension $\widetilde{N}_n$ for $n \geq 1$. Let M be a positive constant. Let $\{h_n\}$ be a sequence of functions on $\mathcal{X}$ such that $\|h_n\|_\infty \leq \widetilde{M}$ for $n \geq 1$. Then*

$$\sup_{g \in \widetilde{\mathbb{G}}} \frac{|\langle h_n, g\rangle_n - \langle h_n, g\rangle|}{\|g\|} = O_P\Big(\Big(\frac{\widetilde{N}_n}{n}\Big)^{1/2}\Big).$$

Proof. Let $\{\phi_j\}$ be an orthonormal basis of $\widetilde{\mathbb{G}}$ relative to the theoretical inner product. For each function $g \in \widetilde{\mathbb{G}}$, we have the expansion $g = \sum_j b_j \phi_j$ and $\|g\|^2 = \sum_j b_j^2$. Thus

$$\begin{aligned} \big|\langle h_n, g\rangle_n - \langle h_n, g\rangle\big| &= \Big|\sum_j b_j \big(\langle h_n, \phi_j\rangle_n - \langle h_n, \phi_j\rangle\big)\Big| \\ &\leq \Big\{\sum_j b_j^2\Big\}^{1/2} \Big\{\sum_j \big(\langle h_n, \phi_j\rangle_n - \langle h_n, \phi_j\rangle\big)^2\Big\}^{1/2}, \end{aligned}$$

so

$$\sup_{g \in \widetilde{\mathbb{G}}} \Big|\frac{\langle h_n, g\rangle_n - \langle h_n, g\rangle}{\|g\|}\Big| \leq \Big\{\sum_j \big(\langle h_n, \phi_j\rangle_n - \langle h_n, \phi_j\rangle\big)^2\Big\}^{1/2}.$$

Since $E(\langle h_n, \phi_j\rangle_n) = \langle h_n, \phi_j\rangle$, we conclude that

$$E\big[(\langle h_n, \phi_j\rangle_n - \langle h_n, \phi_j\rangle)^2\big] = \mathrm{var}\big(\langle h_n, \phi_j\rangle_n\big) = \frac{1}{n}\mathrm{var}\big(\varPsi(h_n, \phi_j)\big).$$

By (11.4.11),

$$\mathrm{var}\big(\varPsi(h_n, \phi_j)\big) \leq M\|h_n\|_\infty^2 \|\phi_j\|^2 \leq \widetilde{M}^2 M,$$

so

$$E\Big(\sum_j \big(\langle h_n, \phi_j\rangle_n - \langle h_n, \phi_j\rangle\big)^2\Big) \leq \widetilde{M}^2 M \frac{\widetilde{N}_n}{n}.$$

The desired conclusion now follows from the Markov inequality. □

11.4.3 Generalized Regression

Recall the generalized regression setup from Section 11.1. Here, we will verify the technical conditions required in Propositions 11.1.1, 11.1.2 and 11.1.3 under five auxiliary assumptions.

Assumption 11.4.1. $B(\cdot)$ is twice continuously differentiable and its first derivative $B'(\cdot)$ is strictly positive on $\mathbb{R}$. There is a subinterval S of $\mathbb{R}$ such that Ψ is concentrated on S and

$$B''(\xi)y - C''(\xi) < 0, \qquad -\infty < \xi < \infty, \tag{11.4.14}$$

for all $y \in \overset{\circ}{S}$, where $\overset{\circ}{S}$ denotes the interior of S. If S is bounded, (11.4.14) holds for at least one of its endpoints.

Note that $A(\eta) \in \overset{\circ}{S}$ for $-\infty < \eta < \infty$. Thus by Assumption 11.4.1,

$$B''(\xi)A(\eta) - C''(\xi) < 0, \qquad -\infty < \xi, \eta < \infty. \tag{11.4.15}$$

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 11.4.1 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 11.4.2. *$P(Y \in S) = 1$ and $E(Y|\boldsymbol{X} = \boldsymbol{x}) = A(\eta(\boldsymbol{x}))$ for $\boldsymbol{x} \in \mathcal{X}$.*

Observe that Assumption 11.4.2 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 (11.1.1).

Assumption 11.4.3. The response function $\eta(\cdot)$ is bounded.

Assumption 11.4.4. There is a positive constant D such that $\operatorname{var}(Y|\boldsymbol{X} = \boldsymbol{x}) \le D$ for $\boldsymbol{x} \in \mathcal{X}$.

Assumption 11.4.5. The distribution of $\boldsymbol{X}$ is absolutely continuous and its density function $f_{\boldsymbol{X}}$ is bounded away from zero and infinity on $\mathcal{X}$.

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 $\|h\|^2 = \langle h, h\rangle$. Then Condition 11.3.1 is an immediate consequence of Assumption 11.4.5. Conditions 11.3.2 and 11.3.3 follow from Lemmas 11.4.7 and 11.4.8.

Verification of Condition 11.2.2.

Let $h_1, h_2 \in \mathbb{H}$ be a pair of bounded functions on $\mathcal{Y}$. Set $h_\alpha = h_1 + \alpha(h_2 - h_1)$ for $0 \le \alpha \le 1$. Then

$$\frac{d}{d\alpha} l(h_\alpha; \boldsymbol{X}, Y) = (h_2(\boldsymbol{X}) - h_1(\boldsymbol{X}))[B'(h_\alpha(\boldsymbol{X}))Y - C'(h_\alpha(\boldsymbol{X}))]$$

and

$$\frac{d^2}{d\alpha^2} l(h_\alpha; \boldsymbol{X}, Y) = (h_2(\boldsymbol{X}) - h_1(\boldsymbol{X}))^2[B''(h_\alpha(\boldsymbol{X}))Y - C''(h_\alpha(\boldsymbol{X}))].$$

Thus, by (11.1.2),

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

Condition 11.2.2 now follows from (11.4.15), the boundedness of $\eta(\cdot)$, and the continuity of $A(\cdot)$, $B''(\cdot)$, and $C''(\cdot)$.

Verification of Condition 11.2.4(i).

We assume that for large n, $\bar{\eta}$ exists uniquely and $\|\bar{\eta}\|_\infty \le K_0$ for some constant K_0. (This follows from Theorem 11.2.1 when relevant conditions are satisfied.) We need only verify the sufficient condition given in Remark 11.2.1. In the present context the score $\boldsymbol{S}(\boldsymbol{\beta})$ has entries

$$\frac{\partial}{\partial \beta_j}\ell(\boldsymbol{\beta}) = \frac{1}{n}\sum_i \phi_j(\boldsymbol{X}_i)\big[B'(g(\boldsymbol{X}_i;\boldsymbol{\beta}))Y_i - C'(g(\boldsymbol{X}_i;\boldsymbol{\beta}))\big], \qquad 1 \le j \le N_n.$$

Since $\bar{\boldsymbol{\beta}}$ maximizes $\Lambda\big(g(\cdot;\boldsymbol{\beta})\big) = E\big[B(g(\boldsymbol{X};\boldsymbol{\beta}))Y - C(g(\boldsymbol{X};\boldsymbol{\beta}))\big]$, we see that

$$\frac{d}{d\boldsymbol{\beta}}\Lambda\big(g(\cdot;\boldsymbol{\beta})\big)\bigg|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} = 0.$$

This implies that

$$E\big[\phi_j(\boldsymbol{X})\big(B'(\bar{\eta}(\boldsymbol{X}))Y - C'(\bar{\eta}(\boldsymbol{X}))\big)\big] = 0, \qquad 1 \le j \le N_n.$$

Thus

$$\begin{aligned} E\Big(\big|\mathbf{S}(\bar{\boldsymbol{\beta}})\big|^2\Big) &= \sum_j E\Big[\frac{\partial}{\partial \beta_j}\ell(\bar{\boldsymbol{\beta}})\Big]^2 \\ &= \frac{1}{n}\sum_j \operatorname{var}\big[\phi_j(\boldsymbol{X})\big(B'(\bar{\eta}(\boldsymbol{X}))Y - C'(\bar{\eta}(\boldsymbol{X}))\big)\big]. \end{aligned}$$

Note that

$$\begin{aligned} &\operatorname{var}\big[\phi_j(\boldsymbol{X})\big(B'(\bar{\eta}(\boldsymbol{X})))Y - C'(\bar{\eta}(\boldsymbol{X}))\big)\big] \\ &\quad = E\Big[\operatorname{var}\big[\phi_j(\boldsymbol{X})\big(B'(\bar{\eta}(\boldsymbol{X})))Y - C'(\bar{\eta}(\boldsymbol{X}))\big)\big|\boldsymbol{X}\big]\Big] \\ &\qquad + \operatorname{var}\Big[E\big[\phi_j(\boldsymbol{X})\big(B'(\bar{\eta}(\boldsymbol{X}))Y - C'(\bar{\eta}(\boldsymbol{X}))\big)\big|\boldsymbol{X}\big]\Big] \\ &\quad = E\big[\phi_j^2(\boldsymbol{X})\big(B'(\bar{\eta}(\boldsymbol{X}))\big)^2\operatorname{var}(Y|\boldsymbol{X})\big] \\ &\qquad + \operatorname{var}\big[\phi_j(\boldsymbol{X})\big(B'(\bar{\eta}(\boldsymbol{X}))A(\eta(\boldsymbol{X})) - C'(\bar{\eta}(\boldsymbol{X}))\big)\big] \end{aligned}$$

Hence, it follows from the boundedness of $\bar{\eta}$, Assumptions 11.4.4, (11.4.15), and the continuity of $B'(\cdot)$, $C'(\cdot)$, and $A(\cdot)$ that, for some positive constant M,

$$\operatorname{var}[\phi_j(\boldsymbol{X})\big(B'(\bar{\eta}(\boldsymbol{X})))Y - C'(\bar{\eta}(\boldsymbol{X}))\big)] \le ME[\phi_j^2(\boldsymbol{X}_i)].$$

Consequently,

$$E\Big(\big|\mathbf{S}(\bar{\boldsymbol{\beta}})\big|^2\Big) \le \frac{M}{n}\sum_j E\big[\phi_j^2(\boldsymbol{X}_i)\big] = \frac{M}{n}\sum_j \|\phi_j\|^2 = M\frac{N_n}{n}.$$

Condition 11.2.4(i) now follows from Remark 11.2.1.

Verification of Condition 11.2.4(ii).

We need the following result (to be proved later):

Claim 1. Under Assumptions 11.4.1–11.4.4, there exist a positive constant δ_1 and a compact subinterval S_0 of S such that $P(Y \in S_0 | \boldsymbol{X} = \boldsymbol{x}) \geq \delta_1$ for $\boldsymbol{x} \in \mathcal{X}$ and $B''(\xi)y - C''(\xi) < 0$ for $-\infty < \xi < \infty$ and $y \in S_0$.

By Claim 1 and the continuity of B'' and C'', there is a positive constant δ_2 such that

$$B''(\xi)y - C''(\xi) \leq -\delta_2, \qquad -K \leq \xi \leq K \text{ and } y \in S_0. \tag{11.4.16}$$

Set $\mathcal{I}_n = \{i : 1 \leq i \leq n \text{ and } Y_i \in S_0\}$. By (11.4.14) and (11.4.16), except on an event whose probability tends to zero as $n \to \infty$,

$$\begin{aligned} &\frac{d^2}{d\alpha^2}\ell\big(g_1 + \alpha(g_2 - g_1)\big) \\ &\quad = \frac{1}{n}\sum_i \Big\{ \big[g_2(\boldsymbol{X}_i) - g_1(\boldsymbol{X}_i)\big]^2 \\ &\qquad\quad \times \Big[B''\big([g_1 + \alpha(g_2 - g_1)](\boldsymbol{X}_i)\big)Y_i - C''\big([g_1 + \alpha(g_2 - g_1)](\boldsymbol{X}_i)\big)\Big]\Big\} \\ &\quad \leq -\frac{\delta_2}{n}\sum_{i \in \mathcal{I}_n} \big[g_2(\boldsymbol{X}_i) - g_1(\boldsymbol{X}_i)\big]^2 \end{aligned} \tag{11.4.17}$$

for $0 \leq \alpha \leq 1$ and all $g_1, g_2 \in \mathbb{G}$ with $\|g_1\|_\infty \leq K$ and $\|g_2\|_\infty \leq K$. Set $I_n = \#(\mathcal{I}_n)$. Then $\lim_n P(I_n \geq \delta_1 n/2) = 1$. Observe that, given $\mathcal{I}_n = \{i_1, \ldots, i_{I_n}\}$, the random vectors $\boldsymbol{X}_i, i \in \mathcal{I}_n$, are independent and have the common density

$$f\big(\boldsymbol{x} \big| Y \in S_0\big) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x})P\big(Y \in S_0 \big| \boldsymbol{X} = \boldsymbol{x}\big)}{P\big(Y \in S_0\big)}.$$

Note that $\delta_1 f_{\boldsymbol{X}}(\boldsymbol{x}) \leq f\big(\boldsymbol{x}\big|Y \in S_0\big) \leq (1/\delta_1) f_{\boldsymbol{X}}(\boldsymbol{x})$. Therefore, it follows from Lemma 11.4.7 that

$$\frac{\delta_2}{n}\sum_{i \in \mathcal{I}_n}\big[g_2(\boldsymbol{X}_i) - g_1(\boldsymbol{X}_i)\big]^2 \geq \frac{\delta_1\delta_2}{2}\|g_2 - g_1\|^2 \tag{11.4.18}$$

for all $g_1, g_2 \in \mathbb{G}$ with $\|g_1\|_\infty \leq K$ and $\|g_2\|_\infty \leq K$, except on an event whose probability tends to zero as $n \to \infty$. Condition 11.2.4(ii) now follows from (11.4.17) and (11.4.18).

Proof of Claim 1. Since $A(\cdot)$ is continuous and strictly increasing and $\eta(\cdot)$ is bounded, $E(Y|\boldsymbol{X} = \boldsymbol{x}) = A(\eta(x))$ ranges over a compact subinterval $S_1 = [c_1, c_2]$ of $\overset{\circ}{S}$. We consider three cases.

Case I. $S = \mathbb{R}$. By Chebyshev's inequality and Assumption 11.4.4,

$$\begin{aligned} &P\big(\big|Y - E(Y|\boldsymbol{X} = \boldsymbol{x})\big| \le \sqrt{2D}\big|\boldsymbol{X} = \boldsymbol{x}\big) \\ &\qquad \ge 1 - \frac{\operatorname{var}(Y|\boldsymbol{X} = \boldsymbol{x})}{2D} \ge \frac{1}{2}, \qquad \boldsymbol{x} \in \mathcal{X}. \end{aligned}$$

Therefore, Claim 1 holds with $S_0 = [c_1 - \sqrt{2D}, c_2 + \sqrt{2D}]$ and $\delta_1 = 1/2$.

Case II. $\overset{\circ}{S} = (-\infty, a)$ or (a, ∞) for some $a \in \mathbb{R}$. Without loss of generality, suppose that $\overset{\circ}{S} = (0, \infty)$. Otherwise, we can replace Y by $-Y + a$ or $Y - a$. Thus $0 < c_1 < c_2$. By Assumption 11.4.4,

$$E(Y^2|\boldsymbol{X} = \boldsymbol{x}) = \operatorname{var}(Y|\boldsymbol{X} = \boldsymbol{x}) + \big[E(Y|\boldsymbol{X} = \boldsymbol{x})\big]^2 \le D + c_2^2.$$

By an obvious modification of Markov's inequality, for any $M > 0$,

$$E\big[Y\operatorname{ind}(Y > M)|\boldsymbol{X} = \boldsymbol{x}\big] \le \frac{E(Y^2|\boldsymbol{X} = \boldsymbol{x})}{M} \le \frac{D + c_2^2}{M};$$

here $\operatorname{ind}(A)$ denotes the indicator function of the set A. Hence, for any $\delta, M \in \mathbb{R}$ with $M > \delta > 0$,

$$\begin{aligned} c_1 &\le E(Y|\boldsymbol{X} = \boldsymbol{x}) \\ &= E\big(Y\operatorname{ind}(Y < \delta)\big|\boldsymbol{X} = \boldsymbol{x}\big) \\ &\qquad + E\big(Y\operatorname{ind}(\delta \le Y \le M)\big|\boldsymbol{X} = \boldsymbol{x}\big) + E\big(Y\operatorname{ind}(Y > M)\big|\boldsymbol{X} = \boldsymbol{x}\big) \\ &\le \delta + MP\big(\delta \le Y \le M|\boldsymbol{X} = \boldsymbol{x}\big) + \frac{D + c_2^2}{M}. \end{aligned}$$

This implies that

$$P\big(\delta \le Y \le M\big|\boldsymbol{X} = \boldsymbol{x}\big) \ge \frac{c_1 - \delta - (D + c_2^2)/M}{M}.$$

Letting $\delta = c_1/3$ and $M = 3(D + c_2^2)/c_1$, we get that

$$P\big(\delta \le Y \le M\big|\boldsymbol{X} = \boldsymbol{x}\big) \ge \frac{c_1^2}{9(D + c_2^2)} > 0.$$

Therefore, Claim 1 holds with $S_0 = [c_1/3, 3(D+c_2^2)/c_1]$ and $\delta_1 = c_1^2/(9(D+c_2^2))$.

Case III. $\overset{\circ}{S} = (a_1, a_2)$ for $a_1, a_2 \in \mathbb{R}$ and (11.4.14) holds at $y = a_1$ or $y = a_2$. Without loss of generality, suppose that $\overset{\circ}{S} = (0, 1)$ and (11.4.14) holds at $y = 1$. Otherwise, we can replace Y by $(Y - a_1)/(a_2 - a_1)$ or $(-Y + a_2)/(a_2 - a_1)$. Thus $Y \le 1$ and $c_1 > 0$. Note that, for $\delta > 0$,

$$c_1 \le E(Y|\boldsymbol{X} = \boldsymbol{x}) \le \delta + P(Y \ge \delta|\boldsymbol{X} = \boldsymbol{x}), \qquad \boldsymbol{x} \in \mathcal{X}.$$

Let $\delta = c_1/2$. Then $P(Y \ge c_1/2|\boldsymbol{X} = \boldsymbol{x}) \ge c_1/2$ for $\boldsymbol{x} \in \mathcal{X}$. Therefore, Claim 1 holds with $S_0 = [c_1/2, 1]$ and $\delta_1 = c_1/2$. □

11.4.4 Density Estimation

We now verify the technical conditions in the context of the density estimation setup from Section 11.1.1.

Let $\mathbb{H}_1$ be a linear space of functions on $\mathcal{Y}$ 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 $c(\phi) = \log \int_{\mathcal{Y}} \exp \phi(\boldsymbol{y})\, d\boldsymbol{y} = 0$. It is convenient to write $\phi = \eta - c(\eta)$ such that η satisfies a linear constraint. To this end, set $\mathbb{H} = \{h \in \mathbb{H}_1 : \int_{\mathcal{Y}} h(\boldsymbol{y})\, d\boldsymbol{y} = 0\}$. 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 11.4.6. The density $f_{\boldsymbol{Y}}$ is bounded away from zero and infinity on $\mathcal{Y}$.

This assumption is equivalent to the assumption that η is bounded.

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 $\|h\|^2 = \langle h, h \rangle$. Then Condition 11.3.1 is an immediate consequence of Assumption 11.4.6. Conditions 11.3.2 and 11.3.3 follow from Lemmas 11.4.7 and 11.4.8, as in the context of generalized regression.

Let $h_1, h_2 \in \mathbb{H}$ be a pair of bounded functions on $\mathcal{Y}$. For $\alpha \in (0,1)$, set $h_\alpha = h_1 + \alpha(h_2 - h_1)$. Then

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

and

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

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

Verification of Condition 11.2.2.

Note that

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

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

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

here, we use the fact that $\int_{\mathcal{Y}} h_2(\boldsymbol{y})\,d\boldsymbol{y} = \int_{\mathcal{Y}} h_2(\boldsymbol{y})\,d\boldsymbol{y} = 0$. 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 11.2.4.

Note that, for $g \in \mathbb{G}$,

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

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

$$\frac{d}{d\alpha}\Lambda(\bar{\eta} + \alpha g)\Big|_{\alpha=0} = 0, \qquad g \in \mathbb{G},$$

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

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

Condition 11.2.4(i) now follows from Lemma 11.4.8.

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

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

Condition 11.2.4(ii) now follows from Condition 11.2.2.

11.4.5 Hazard Regression

Consider, finally, the hazard regression setup from Section 11.1.1. We verify various technical conditions under the following assumptions.

Assumption 11.4.7. (i) The vector $\boldsymbol{X}$ of covariates has a density function $f_{\boldsymbol{X}}$ that is bounded away from zero and infinity on $\mathcal{X}$; (ii) $P(C = \tau | \boldsymbol{X} = \boldsymbol{x})$ is bounded away from zero as $\boldsymbol{x}$ ranges over $\mathcal{X}$; (iii) the log-hazard function $\eta(\cdot, \cdot)$ is bounded on $\mathcal{U} = \mathcal{X} \times [0, \tau]$.

Define the empirical inner product and empirical norm by

$$\langle h_1, h_2 \rangle_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^{Y_i} h_1(\boldsymbol{X}_i, t) h_2(\boldsymbol{X}_i, t)\,dt$$

and $\|h\|_n^2 = \langle h, h \rangle_n$. The corresponding theoretical inner product and theoretical norm are defined by

$$\langle h_1, h_2 \rangle = E\int_0^{Y} h_1(\boldsymbol{X}, t) h_2(\boldsymbol{X}, t)\,dt$$

and $\|h\|^2 = \langle h, h \rangle$.

Verification of Condition 11.3.1.

By conditioning,

$$\begin{aligned}\|h\|^2 &= E\int_0^Y h^2(\boldsymbol{X},t)\,dt\\ &= E\int_0^\tau h^2(\boldsymbol{X},t)\mathrm{ind}(Y \ge t)\,dt\\ &= E\int_0^\tau h^2(\boldsymbol{X},t)P(Y \ge t|\boldsymbol{X})\,dt.\end{aligned}$$

Since T and C are conditionally independent given $\boldsymbol{X}$,

$$P(T \ge t|\boldsymbol{X})P(C \ge t|\boldsymbol{X}) = P(T \ge t|\boldsymbol{X})P(C \ge t|\boldsymbol{X}).$$

Hence,

$$\begin{aligned}\|h\|^2 &= E\int_0^\tau h^2(\boldsymbol{X},t)\exp\Big(-\int_0^t \exp\eta(\boldsymbol{X},u)\,du\Big)P(C \ge t|\boldsymbol{X})\,dt\\ &= \int_{\mathcal{X}}\int_0^\tau h^2(\boldsymbol{x},t)\Phi(\boldsymbol{x},t)\,dt\,d\boldsymbol{x},\end{aligned}$$

where

$$\Phi(\boldsymbol{x},t) = \exp\Big(-\int_0^t \exp\eta(\boldsymbol{x},u)\,du\Big)P(C \ge t|\boldsymbol{X}=\boldsymbol{x})f_{\boldsymbol{X}}(\boldsymbol{x})$$

for $(\boldsymbol{x},t) \in \mathcal{U}$. By Assumption 11.4.7, $\Phi(\boldsymbol{x},t)$ is bounded away from zero and infinity and thus Condition 11.3.1 is valid.

Verification of Conditions 11.3.2 and 11.3.3.

Note that the empirical and theoretical inner products defined above have the form given in Section 11.4.2 with $\Psi(h_1,h_2) = \int_0^Y h_1(\boldsymbol{X},t)h_2(\boldsymbol{X},t)\,dt$. Since $Y \le \tau$, (11.4.8) holds with $M = \tau$. The Cauchy–Schwarz inequality yields (11.4.9). Conditions 11.3.2 and 11.3.3 then follow from Lemmas 11.4.7 and 11.4.8 respectively.

Let $h_1, h_2 \in \mathbb{H}$ be a pair of bounded functions on $\mathcal{Y}$. Set $h_\alpha = h_1 + \alpha(h_2 - h_1)$ for $0 \le \alpha \le 1$. Then

$$\begin{aligned}\frac{d}{d\alpha}l(h_\alpha) &= \delta[h_2(\boldsymbol{X},Y) - h_1(\boldsymbol{X},Y)]\\ &\qquad - \int_0^Y [h_2(\boldsymbol{X},t) - h_1(\boldsymbol{X},t)]\exp h_\alpha(\boldsymbol{X},t)\,dt\end{aligned} \tag{11.4.19}$$

and

$$\frac{d^2}{d\alpha^2}l(h_\alpha) = -\int_0^Y [h_2(\boldsymbol{X},t) - h_1(\boldsymbol{X},t)]^2\exp h_\alpha(\boldsymbol{X},t)\,dt. \tag{11.4.20}$$

Verification of Condition 11.2.2.

Suppose $\|h_1\|_\infty \le K$ and $\|h_1\|_\infty \le K$ for some positive constant K. It follows from (11.4.20) that

$$\frac{d^2}{d\alpha^2}\Lambda(h_1 + \alpha(h_2 - h_1)) = -E\int_0^Y [h_2(\boldsymbol{X}, t) - h_1(\boldsymbol{X}, t)]^2 \exp h_\alpha(\boldsymbol{X}, t)\, dt.$$

Since $\|h_\alpha\|_\infty \le K$, the above right side is bounded above and below by multiples of

$$-E\int_0^Y [h_2(\boldsymbol{X}, t) - h_1(\boldsymbol{X}, t)]^2\, dt = \|h_2 - h_1\|^2.$$

The desired result follows.

Verification of Condition 11.2.4.

We assume that for large n, $\bar{\eta}$ exists uniquely and $\|\bar{\eta}\|_\infty \le K_0$ for some constant K_0. (This follows from Theorem 11.2.1 when relevant conditions are satisfied.) Let $g \in \mathbb{G}$. Since

$$\frac{d}{d\alpha}\Lambda(\bar{\eta} + \alpha g)\Big|_{\alpha=0} = 0,$$

we conclude from (11.4.19) that

$$\frac{d}{d\alpha}\ell(\bar{\eta} + \alpha g)\Big|_{\alpha=0} = (E_n - E)(\delta g) - (\langle \exp\bar{\eta}, g\rangle_n - \langle \exp\bar{\eta}, g\rangle).$$

It follows from Lemma 11.4.8 that

$$\sup_{g\in\mathbb{G}} \frac{|\langle \exp\bar{\eta}, g\rangle_n - \langle \exp\bar{\eta}, g\rangle|}{\|g\|} = O_P\Big(\Big(\frac{N_n}{n}\Big)^{1/2}\Big),$$

and it follows by arguing as in the proof of Lemma 11.4.8 that

$$\sup_{g\in\mathbb{G}} \frac{|(E_n - E)(\delta g)|}{[E(g^2)]^{1/2}} = O_P\Big(\Big(\frac{N_n}{n}\Big)^{1/2}\Big).$$

On the other hand, Assumption 11.4.7(i) and Condition 11.3.1 together imply that $E(g^2) \asymp \|g\|^2$ uniformly in $g \in \mathbb{G}$. Thus Condition 11.2.4(i) is valid.

For $g_1, g_2 \in \mathbb{G}$, set $g_\alpha = g_1 + \alpha(g_2 - g_2)$ for $0 \le \alpha \le 1$. It follows from (11.4.20) that

$$\frac{d^2}{d\alpha^2}\ell(g_\alpha) = -E_n\int_0^Y [g_2(\boldsymbol{X}, t) - g_1(\boldsymbol{X}, t)]^2 \exp g_\alpha(\boldsymbol{X}, t)\, dt.$$

Suppose $\|g_1\|_\infty \le K$ and $\|g_2\|_\infty \le K$ for some positive constant K. Then $\|g_\alpha\|_\infty \le K$ and thus the right side of the above display is bounded above and below by multiples of

$$-E_n \int_0^Y [g_2(\boldsymbol{X}, t) - g_1(\boldsymbol{X}, t)]^2 \, dt = \|g_2 - g_1\|_n^2.$$

Condition 11.2.4(ii) then follows from Lemma 11.4.7.

11.5 Notes

This chapter is mainly based on Huang (2001), which is a synthesis of the theoretical development of various subsets of authors of this book during the last two decades. The motivation of this line of research can go back to Stone (1980, 1982), where it was shown that $n^{-2p/(2p+L)}$ is the optimal pointwise or L_2 rate of convergence in functional estimation, where the unknown function η of a L-dimensional vector $\boldsymbol{u} = (u_1, \ldots, u_L)$ has a p bounded derivatives and and an estimate $\widehat{\eta}$ of η is based on a random sample of size n. Stone (1982) raised the possibility that if η is the sum of functions of individual variables u_l, then the optimal L_2 rate of convergence would be $n^{-2p/(2p+1)}$. This possibility was verified in Stone (1985) in the context of additive regression and in Stone (1986) in the context of logistic regression, Poisson regression and other generalized additive models. In these papers, the estimate $\widehat{\eta}$ (of η if η is additive or, more generally, of the best additive approximation η^* of η) that was shown to achieve the optimal L_2 rate of convergence has the form of a nonadaptive sum of polynomial splines in the individual variables u_l.

In Stone (1985) it was suggested that if η is the sum of functions of specified subsets of the variables u_l having at most d variables in each such subset, the the optimal L_2 rate of convergence would be $n^{-2p/(2p+d)}$. This result was verified in Stone (1994) in the context of regression, generalized regression and density estimation, in Kooperberg, Stone and Truong (1995b) in the context of hazard regression, and in Kooperberg, Stone and Truong (1995c) in the context of spectral density estimation (where $d = L = 1$). In these papers the estimate $\widehat{\eta}$ (of the best approximation η^* to η having the specified form) that was shown to achieve the optimal L_2 rate of convergence has the form of a nonadaptive sum of tensor products of polynomial splines in the specified subsets of variables. The L_2 rate of convergence result for unsaturated models demonstrated the potential of the corresponding estimation procedure for ameliorating the curse of dimensionality.

In his Ph. D. thesis, Hansen (1994) extended the previous theoretical results involving L_2 rates of convergence for estimates based on polynomial splines and selected tensor products to include bivariate and more general multivariate polynomial splines as well as univariate splines. He also showed that the theories for regression, generalized regression, multiple logistic regression, density estimation, conditional density estimation, and so forth could be treated within a common framework, referred to as an extended linear model. This is where the name extended linear model was coined.

In an attempt to better understand the mathematical structural of functional ANOVA modeling and the role of low order functional ANOVA models (including additive models) in overcoming the curse of dimensionality, Huang (1998a) realized that the use of polynomial splines is not essential and virtually arbitrary linear spaces and their tensor products can be used to build the estimation spaces. The result of Huang (1998a) was established in the context of regression, and was extended to the generalized regression context in Huang (1998b). The new theoretical approach of Huang (1998a, 1998b) simplified much of the theoretical research on extended linear modeling and was then used in Huang and Stone (1998) to extend the theory for hazard regression in Kooperberg, Stone and Truong (1995b) to event history analysis involving repeated events of multiple kinds and time-dependent covariates, and in Huang, Kooperberg, Stone and Truong (2000) to study functional ANOVA modeling in proportional hazards regression. A fresh synthesis on the theory of extended linear models was achieved in Huang (2001) to unify the previous efforts. This synthesis provides a convenient framework to investigate the theoretical properties of extended linear modeling with free knot splines in Stone and Huang (2001a, 2001b), which will be summarized in the next chapter.

We use maximum likelihood estimation over a finite-dimensional estimation space in fitting an extended linear model. This approach can be thought of as a special case of the method of sieves (Grenander 1981). Rates of convergence for the general method of sieves have been developed using the theory of empirical processes; see, for example, Shen and Wong (1994), Wong and Shen (1995), van de Geer (1995), Birgé and Massart (1998). None of these papers considered functional ANOVA modeling.

An alternative approach that is also very convenient in incorporating functional ANOVA structure in functional estimation is the penalized likelihood method (or smoothing spline ANOVA). See Wahba (1990), Wahba, Wang, Gu, Klein and Klein (1995) and the references therein for general discussions. The theoretical papers on the penalized likelihood method for the saturated models include Silverman (1982), Cox and O'Sullivian (1990), Zucker and Karr (1990), Chen (1991), O'Sullivan (1993), Gu and Qiu (1994), and Gu (1995, 1996). The rate of convergence of smoothing spline ANOVA in the context of regression is studied recently in Lin (2000).