Abstract

Extended Linear Models, Multivariate Splines and ANOVA

by

Mark Henry Hansen

Doctor of Philosophy in Statistics

University of California at Berkeley

Professor Charles J. Stone, Chair

In this dissertation, we pursue a theoretical investigation into several aspects of multivariate function estimation. In general, we confine ourselves to estimators that are smooth, piecewise polynomial functions, or splines. In the last decade, a considerable body of literature on multivariate spline spaces has been amassed by approximation theorists, numerical analysts and computer scientists, and we hope to demonstrate the practicality of these tools for statistical applications. Initially, we consider estimating a regression function through ordinary least squares projections into certain spaces of splines. In order to tame the curse of dimensionality, we consider ANOVA decompositions of various function spaces. Finally, to accomodate more general estimation problems, we introduce the notion of an extended linear model and corresponding ANOVA decompositions. In the contexts of both regression and the extended linear model, the emphasis here is on rates of convergence.

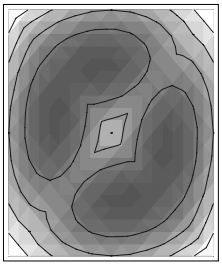
Contents

List of Figures from Chapter 2

Acknowledgments

In so many ways, this dissertation is dedicated to my parents, who have never failed to provide me with unquestioning support, even when it was not entirely within their means to do so. I would also like to take this opportunity to thank my adviser, my friend, Professor Charles J. Stone. If it weren't for his unfailing encouragement and guidance, I certainly would not have made it through the program. I must also thank another member of the department, Professor David Freedman, who opened so many doors, taking an interest in me both as a person and as a student. Finally, I would like to acknowledge the members of my committee, Professors John Rice and Paul Ruud.

Introduction



We have had too much consecration,
too little affirmation,
too much: but this, this, this
has been proved heretical,
too little: I know, I feel
the meaning that words hide;
they are anagrams, cryptograms,
little boxes, conditioned
to hatch butterflies...

from Trilogy by H.D.

In the pages that follow, we pursue a theoretical investigation into several aspects of multivariate function estimation. In general, we confine ourselves to estimators that are smooth, piecewise polynomial functions, or splines. In the last decade, a considerable body of literature on multivariate spline spaces has been amassed by approximation theorists, numerical analysts and computer scientists. Through the next three chapters, we hope to demonstrate the practicality of these tools for statistical applications. Initially, we consider estimating a regression function through ordinary least squares projections into certain spaces of splines. However, as we will see, there are a wide variety of other estimation problems that are also amenable to this approach.

Many of the results derived in this dissertation rely heavily on the pioneering work of Stone (1985, 1986, 1991ab, 1994). In each of these papers, univariate splines or their tensor products were used as the fundamental building blocks for function estimation. We have extended these results to include (virtually) arbitrary spaces of multivariate splines. In addition, we introduce the concept of an extended linear model, which allows us to consider these estimation problems simultaneously. In addition, by removing the dependence on univariate splines, we are able to discern which properties exhibited by these spaces are essential for

statistical applications. In the remainder of this section, we set the stage for the theoretical investigation that follows. In addtion, we comment on the existing theoretical and methodological literature where appropriate.

Regression

Let Y be a real-valued random variable and $\mathbf{X}$ a vector of random covariates taking values in a compact set $\mathcal{X}$, and let $\mu(\mathbf{x})$ denote the regression function

$$\mu(\mathbf{x}) \;=\; E(Y \,|\, \mathbf{X} = \mathbf{x}), \qquad \mathbf{x} \in \mathcal{X}.$$

Let $\mathbb{G}$ be a finite dimensional linear space. Now, given a simple random sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ from the joint distribution of $(\mathbf{X}, Y)$, we estimate $\mu(\mathbf{x})$ by finding a function in $\mathbb{G}$ satisfying

$$\hat{\mu} \;=\; \underset{g \in \mathbb{G}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \big(Y_i - g(\mathbf{X}_i) \big)^2 \right].$$

In the first two sections of Chapter 1, the space $\mathbb{G}$ is taken to be a collection of piecewise polynomial functions, which we will denote $\mathbb{PP}$. Roughly speaking, we divide the set $\mathcal{X}$ into a number of disjoint subsets, and let $\mathbb{PP}$ consist of those functions that are polynomials of a fixed degree in each subset. We envision allowing the sets in the partition of $\mathcal{X}$ to shrink as our sample size grows. This increases the flexibility of the space $\mathbb{PP}$, thereby improving its approximation power. To keep control of the variance of our estimate $\hat{\mu}$, however, it is clear that we cannot shrink these sets too quickly. With this in mind, it is reasonable to expect that if the partition of $\mathcal{X}$ is refined at the right rate, $\hat{\mu}$ should converge in some sense to μ as n tends to infinity.

In Section 1.1, we develop a number of extremely useful properties of piecewise polynomial functions. In particular, we derive relationships between various norms that can be associated with the space $\mathbb{PP}$. Along the way, we are able to demonstrate that with probability approaching one as n tends to infinity, the space $\mathbb{PP}$ is identifiable and hence the estimate $\hat{\mu}$ is unique. For the most part, this discussion follows closely the development in Stone (1994). In particular, we make use of a clever rescaling trick presented by Buja (1994) in his discussion of Stone's paper. In Section 1.2, we derive the L_2 rate of convergence of $\hat{\mu}$ to μ. We include this derivation mainly for pedagogical reasons, since the more complicated results in Section 1.4 follow in a similar fashion. Using the arguments in Stone (1982) we can demonstrate that in this case the derived rate is optimal.

When each of the variables x_l, $1 \leq l \leq M$, is one-dimensional, this rate is derived in Koo (1988). Actually, Koo establishes the L_2 rate of convergence for tensor products of univariate splines, and as a special case obtains the rate for piecewise polynomial functions. In principle, the elegant arguments in Mo (1991) can be adapted to derive these results as well. Korostelev and Tsybakov (1993) provide a particularly lucid treatment of related rates, but in the context of minimax image restoration. Finally, when each of the variables x_l, $1 \leq l \leq M$, are one-dimensional, Chen (1988) and Bruman and Chen (1989) demonstrate data-driven techniques for determining the number of sets into which $\mathcal{X}$ should be divided.

Observe that the function $\mu(\mathbf{x})$ is itself obtained as the solution to a minimization problem; that is,

$$\mu \;=\; \underset{f \in L_2(\mathcal{X})}{\operatorname{argmin}}\, E\big[\,(Y - f(\mathbf{X}))^2\,\big]\,.$$

In this context, we say that μ and $\hat{\mu}$ are defined as the solution to optimization problems over saturated spaces of functions. As we will see, the rate at which $\hat{\mu}$ approaches μ depends in part on the dimension of the vector $\mathbf{X}$, so that the larger this dimension the slower the rate. This has been referred to as the curse of dimensionality. To help reduce this effect, we make use of unsaturated spaces of functions. Suppose, for example, that instead of looking at all square-integrable functions defined on $\mathcal{X}$, we consider an additive model for μ:

$$\mu(\mathbf{x}) \;=\; \mu_0 + \mu_1(x_1) + \cdots + \mu_M(x_M)\,,$$

where $\mathbf{x} = (x_1, \ldots, x_M)$ and each x_l may be a vector. Here, μ_0 is a constant. For obvious reasons, the space of all functions in this form is said to be unsaturated. Now, we can either assume that μ is in this space or define

$$\mu^*(\mathbf{x}) \;=\; \mu_0^* + \mu_1^*(x_1) + \cdots + \mu_M^*(x_M)$$

to be the best approximation to μ of this form. We will have more to say on this approximation in Chapter 1. To estimate such a function, we consider

$$\hat{\mu}(\mathbf{x}) \;=\; \hat{\mu}_0 + \hat{\mu}_1(x_1) + \cdots + \hat{\mu}_M(x_M)\,,$$

where $\hat{\mu}_l$ is a (possibly smooth) piecewise polynomial function depending only on the variable x_l and $\hat{\mu}_0$ is a constant. Then, subject to some identifiability constraints, we might expect that the functions $\hat{\mu}_l$ converge to μ_l^* as n tends to infinity. As we will see, this is in fact the case and the rate depends in part on the largest dimension of the vectors $x_1, \ldots, x_M$ rather than on the dimension of

$\mathbf{x}$. Therefore, by considering an additive model, we have managed to mitigate the curse of dimensionality. This is essentially the approach followed by Stone (1985). The reader is referred to Hastie and Tibshirani (1990) for a more general discussion of additive modelling.

Taking this approach one step farther, let $\mathcal{S}$ denote a hierarchical collection of subsets of $\{1, \ldots, M\}$. By hierarchical we mean that if $s \in \mathcal{S}$, then $r \in \mathcal{S}$ for $r \subset s$. By analogy with treatment above, we consider a model of the form

$$\mu \;=\; \sum_{s \in \mathcal{S}} \mu_s \;,$$

where μ_s is a square-integrable function depending only on the variables x_l, $l \in s$. Here, we take $\mu_\varnothing$ to be a constant. Again, we can either assume that μ is in this unsaturated space, or define

$$\mu^* \;=\; \sum_{s \in \mathcal{S}} \mu_s^* \tag{1}$$

to be the best approximation to μ of this form. To estimate such a function, we consider

$$\hat{\mu} \;=\; \sum_{s \in \mathcal{S}} \hat{\mu}_s \;, \tag{2}$$

where $\hat{\mu}_s$ is a suitable (possibly smooth) piecewise polynomial function depending only on the varibles x_l, $l \in s$, and $\hat{\mu}_\varnothing$ is a constant. As we will see, subject to certain identifiability constraints, the functions $\hat{\mu}_s$ converge to μ_s^* as n tends to infinity.

For each $s \in \mathcal{S}$, we calculate the dimension of the component μ_s^* by summing the dimensions of the vectors x_l, $l \in s$. The rate of convergence of $\hat{\mu}$ to μ^* now depends in part on largest dimension of the components μ_s^*, $s \in \mathcal{S}$. In addition, the functions $\hat{\mu}_s$ tend to μ_s^* at this same rate. Observe that if $\{1, \ldots, M\} \notin \mathcal{S}$, then we have again reduced the curse of dimensionality. This rate is established and discussed in considerable detail in Sections 1.3 and 1.4. As with the saturated spaces, the arguments follow closely the approach in Stone (1994). In fact, when each of the variables x_l, $1 \leq l \leq M$, are one-dimensional we obtain Stone's results in the regression context. Similar results based on tensor products of univariate splines were also obtained by Newey (1991).

In order to make sense of these convergence results, we must impose identifiability constraints on the expansions given above. One technique proposed by Hastie and Tibshirani (1990) and used in Stone (1994) involves forcing higher order components to be orthogonal to corresponding lower order components. Toward this

end, let f_1, f_2 be any two square-integrable functions defined on $\mathcal{X}$. Then, we define

$$\langle f_1, f_2 \rangle \;=\; E\big[\, f_1(\mathbf{X})\, f_2(\mathbf{X}) \,\big]$$

and insist that if μ^* is given by (1) then μ^*_s be orthogonal to μ^*_r with respect to this inner-product for every proper subset r of s. To enforce similar conditions on the components of the expansion in (2), we introduce the inner product

$$\langle f_1, f_2 \rangle_n \;=\; E_n\big[\, f_1(\mathbf{X})\, f_2(\mathbf{X}) \,\big] \;=\; \frac{1}{n} \sum_{i=1}^{n} f_1(\mathbf{X}_i)\, f_2(\mathbf{X}_i)\,.$$

In Section 1.3, we derive a number of properties of the expansions (1) and (2) subject to the indicated identifiability constraints. In particular, we demonstrate that these conditions are in fact sufficient to guarantee a unique expansion of the form (1) and that with probability approaching one as n tends to infinity, the same uniqueness holds for the expansion in (2). In general, we refer to (1) and (2) as ANOVA decompositions.

In the context of ordinary least squares, this approach seems quite natural. However, other inner products could be used to guarantee identifiability in the expansions (1) and (2). In particular, for $1 \le l \le M$, let f_{X_l} denote the marginal density of the random vector X_l, and let $f^*_{\mathbf{X}}$ denote the product of these marginals. Then, if we let E^* denote the expectation operator with respect to $f^*_{\mathbf{X}}$, we can set

$${}^*\langle f_1, f_2 \rangle \;=\; E^*\big[\, f_1(\mathbf{X})\, f_2(\mathbf{X}) \,\big]$$

and insist that if μ^* is given by (1) then μ^*_s is orthogonal to μ^*_r with respect to this new inner-product for every proper subset r of s. Similarly, we can use the inner-product

$$\begin{aligned} {}^*\langle f_1, f_2 \rangle_n \;&=\; E^*_n\big[\, f_1(\mathbf{X})\, f_2(\mathbf{X}) \,\big] \\ &=\; \frac{1}{n^M} \sum_{i_1=1}^{n} \cdots \sum_{i_M=1}^{n} f_1(X_{1i_1}, \ldots, X_{Mi_M})\, f_2(X_{1i_1}, \ldots, X_{Mi_M}) \end{aligned}$$

to enforce the identifiability conditions on the expansion in (2). Not surprisingly, the same rate of convergence for $\hat{\mu}_s$ to μ^*_s applies no matter which set of inner-products are used in conjunction with the expansions in (1) and (2). These issues are discussed in detail in Section 1.5.

Observe that, unlike the treatment in Stone (1994), we allow the vector of covariates $\mathbf{X} = (X_1, \ldots, X_M)$ to be comprised of the vectors X_l, $1 \le l \le M$, where each X_l takes values in a compact set $\mathcal{X}_l$. Additionally, we assume that $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_M$. In this way, we can easily accomodate covariates that are

genuinely spatial, for example. More generally, this framework allows us naturally to include covariates whose domain of definition is not the Cartesian product of a number of one-dimensional sets. A similar approach is followed by Gu and Wahba (1993) in the context of smoothing splines. Moreover, under stronger conditions both on the design points and on the smoothness of the regression function, Chen (1991) derives corresponding convergence rates for estimates based on smoothing splines when each variable x_l, $1 \leq l \leq M$, is one-dimensional. For a general discussion of smoothing splines, the reader is referred to Wahba (1990).

Extended Linear Models

In Chapter 3, we consider using ANOVA decompositions involving multivariate splines for more general estimation problems than ordinary regression. Changing notation slightly, we let $\mathbf{W}$ be a random vector taking values in a set $\mathcal{W}$. Now, for each bounded function f define the functional $l(f, \mathbf{w})$, $\mathbf{w} \in \mathcal{W}$. In this context, we let the unknown function ϕ be the solution of the following optimization problem

$$\phi \;=\; \underset{f \in L_\infty(\mathcal{W})}{\operatorname{argmax}} \; E\big[\, l(f, \mathbf{W}) \,\big] \,. \tag{4}$$

Similarly, given a (not necessarily hierarchical) collection $\mathcal{S}$ of subsets of $\{1, \ldots, M\}$, we consider a model for ϕ of the form

$$\phi \;=\; \sum_{s \in \mathcal{S}} \phi_s \,, \tag{5}$$

where ϕ is a bounded function depending only on the variables x_l, $l \in s$ and $\phi_\varnothing$ is a constant. As in the regression context, we can either assume that ϕ is in this unsaturated space or define

$$\phi^* \;=\; \sum_{s \in \mathcal{S}} \phi_s^* \tag{6}$$

to be the function maximizing the expression in (4) over the space of functions of the form (5). To estimate ϕ^*, we consider

$$\hat{\phi} \;=\; \sum_{s \in \mathcal{S}} \hat{\phi}_s \,, \tag{7}$$

where $\hat{\phi}_s$ is a (possibly smooth) piecewise polynomial function depending only on the variables x_l, $l \in s$. Now, given a simple random sample $\mathbf{W}_1, \ldots, \mathbf{W}_n$ from the distribution of $\mathbf{W}$, we take $\hat{\phi}$ to be that function in the form (7) maximizing

$$E_n\big[\, l(f, \mathbf{W}) \,\big] \;=\; \frac{1}{n} \sum_{i=1}^{n} l(f, \mathbf{W}_i) \,.$$

In order to guarantee the uniqueness of the expansions in (6) and (7), we impose orthogonality constraints on the components ϕ_s^* and $\hat{\phi}_s$, $s \in \mathcal{S}$. As we will see, the inner products defined in the regression context might have to be altered to accomodate the possibility of censoring. In addition, special consideration must be given to the case when $\mathcal{S}$ is not hierarchical. In spirit, however, the approach is the same as that used in the regression context.

The goal of Chapter 3 is to determine under which conditions $\hat{\phi}_s$ converges to ϕ_s^*, $s \in \mathcal{S}$, as n tends to infinity. In his rejoinder, Stone (1994) observes that

> Thus, distinct but closely related theories have been or are being developed for regression, logistic and Poisson regression, polychotomous regression, hazard regression and the estimation of hazard, density, conditional density and spectral density functions. It would be worthwhile to synthesize this theoretical work.

In Section 3.1, we achieve this synthesis by introducing a general form for the functional $l(f, \mathbf{w})$. Typically, $l(f, \mathbf{w})$ will arise as a log-likelihood function, and hence we say that the functionals in our general form as specify an extended linear model. Throughout Chapter 3, we treat five examples of extended linear models explicitly: regression, generalized regression, censored regression, polychotomous regression, and density estimation. The treatment of these cases follows an approach first used by Stone (1986) to derive the rate of convergence for additive models in the context of generalized regression. These techniques also appear in Stone (1994) where more general ANOVA models based on tensor products of univariate splines are considered. In the final section of Chapter 3, we derive the rate of convergence for conditional density estimation. In this context, there is an apparent need for the alternate inner products described at the end of the regression section above to define the associated ANOVA decompositions. Finally, we observe that, in addition to the six examples discussed in Chapter 3, hazard regression and logspline spectral estimation can also be treated in this manner. The interested reader is referred to Kooperberg, Stone and Truong (1993a,1993c).

Methodological Considerations

As mentioned earlier, in the last decade approximation theorists, numerical analysts and computer scientists have made considerable progress in defining the notion of a multivariate spline. For the most part, however, these tools have not found their way into the statistical literature. Therefore, we hope that the theoretical investigation pursued in this dissertation will give rise to practically useful

methodology, which would undoubtedly be highly adaptive in character. In Chapter 2 of this dissertation, we present a brief overview of two important classes of multivariate splines. The first is generated by scaled, integer translates of a fixed spline function. The regularity of these spaces makes them ideal for discussing the larger issue of approximation rates. The second class of splines is based on the so called Bezier- or B-net representation and finds its roots in finite element applications. Because these spaces can be defined for virtually any triangulation, they promise to be powerful tools in the construction of adaptive function estimation routines.

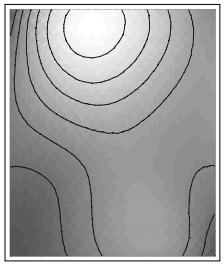
To date, univariate splines and their tensor products are the basis for many successful applications. It was Smith (1982) who first developed a stepwise procedure for adaptive knot deletion in the context of univariate regression. Since that point, adaptively selected univariate spline spaces have found there way into

- additive regression [Friedman and Silverman (1989) and Breiman (1993)],
- multivariate regression [Friedman (1991) and Breiman (1991)],
- density estimation [Kooperberg and Stone (1991, 1992)],
- conditional density estimation [Masse and Truong (1992)],
- hazard regression [Kooperberg, Stone and Truong (1994)], and
- spectral density estimation [Kooperberg, Stone and Truong (1993b)].

Hopefully, adaptive methodologies based on multivariate splines will also enjoy the success that Smith's original techniques have seen.

Chapter 1

Regression



1.1 Piecewise Polynomials

Notation

Let $\mathbb{R}^d$ denote d-dimensional Euclidean space and $\mathbb{N}^d$ denote the set of d-dimensional vectors of nonnegative integers. In defining spaces of polynomials over $\mathbb{R}^d$, we rely on the following standard notation. Given the vectors $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and $j = (j_1, \ldots, j_d) \in \mathbb{N}^d$, set

$$x^j \;=\; x_1^{j_1} \cdots x_d^{j_d} \qquad \text{and} \qquad [\,j\,] \;=\; j_1 + \cdots + j_d\,, \tag{1.1}$$

with $x^j = 1$ if $[\,j\,] = 0$. Next, for any two vectors $x, y \in \mathbb{R}^d$ and any nonzero scalar $h \in \mathbb{R}$, we define $(x + y)\,/\,h$ to be the vector

$$(\,x + y\,)\,/\,h \;=\; \big(\,(\,x_1 + y_1\,)\,/\,h, \ldots, (\,x_d + y_d\,)\,/\,h\,\big)\,.$$

In addition, we write

$$x \;\le\; y \qquad \text{if } x_i \le y_i \text{ for } 1 \le i \le d\,,$$

with similar expressions holding for equality and strict inequality. If, on the other hand, $j \in \mathbb{N}^d$ but $m \in \mathbb{N}$, then we write

$$j \;\le\; m \qquad \text{if } [j] \le m, \tag{1.2}$$

with similar expressions holding for equality and strict inequality. In terms of this notation, for a fixed $m \in \mathbb{N}$, a polynomial of total degree m in $x \in \mathbb{R}^d$ is defined to be any function of the form

$$p(x) \;=\; \sum_{j \le m} b_j \, x^j \tag{1.3}$$

where $j \in \mathbb{N}^d$ and $b_j \in \mathbb{R}$ for $j \le m$. Finally, for any $x \in \mathbb{R}^d$, let

$$| \, x \, |^2 \;=\; x_1^2 + \cdots x_d^2 \tag{1.4}$$

define the usual Euclidean norm of x. Note that in (1.1)–(1.4) the vectors x and j do not appear as boldfaced quantities. In general, we consider predictor variables whose natural domain of definition is a subset of $\mathbb{R}^d$ for some d. Rather than decomposing these domains into Cartesian products of one-dimensional sets, we view these variables as vectors in $\mathbb{R}^d$.

Vectors built from these primitive variables, however, will appear in boldface. To be more precise, let x_l be a vector in $\mathbb{R}^{d_l}$ and j_l be a vector in $\mathbb{N}^{d_l}$ for $1 \le l \le M$. Also, set

$$\mathbf{x} \;=\; (\, x_1, \ldots, x_M \,) \qquad \text{and} \qquad \mathbf{j} \;=\; (\, j_1, \ldots, j_M \,) \,. \tag{1.5}$$

We can now lift the definitions (1.1) to the vectors $\mathbf{x}$ and $\mathbf{j}$ by setting

$$\mathbf{x}^{\mathbf{j}} \;=\; x_1^{j_1} \cdots x_M^{j_M} \qquad \text{and} \qquad [\mathbf{j}] \;=\; [\, j_1 \,] + \cdots + [\, j_M \,] \,, \tag{1.6}$$

with $\mathbf{x}^{\mathbf{j}} = 1$ if $[\mathbf{j}] = 0$. Given two vectors $\mathbf{x}$ and $\mathbf{y}$ each in the form (1.5), for any nonzero vector $h \in \mathbb{R}^M$, we define $(\mathbf{x} + \mathbf{y}) \,/\, h$ to be the vector

$$(\, \mathbf{x} + \mathbf{y} \,) \,/\, h \;=\; \big((\, x_1 + y_1 \,) \,/\, h_1, \ldots, (\, x_M + y_M \,) \,/\, h_M \,\big) \,.$$

Furthermore, if $m = (m_1, \ldots, m_M)$ is a vector in $\mathbb{N}^M$, then we write

$$\mathbf{j} \;\le\; m \qquad \text{if } [\, j_l \,] \le m_l \;\text{ for } 1 \le l \le M. \tag{1.7}$$

Using this notation, for a fixed vector $m \in \mathbb{N}^M$, a polynomial of coordinate degree $m \in \mathbb{N}$ in the variable $\mathbf{x}$ is any function of the form

$$p(\mathbf{x}) \;=\; \sum_{\mathbf{j} \le m} b_{\mathbf{j}} \, \mathbf{x}^{\mathbf{j}} \,, \tag{1.8}$$

where $b_{\mathbf{j}} \in \mathbb{R}$ for $\mathbf{j} \le m$. Referring to such polynomials as having coordinate degree m is consistent with our view of $\mathbf{x}$ being comprised of the vectors x_l, $1 \le l \le M$, but is somewhat nonstandard from a numerical analysis point of view. In any event, observe that the space of all such polynomials is simply the tensor product

of the spaces of polynomials of total degree m_l in the variables x_l for $1 \leq l \leq M$, as given in (1.3). Finally, by extending (1.4) we obtain

$$| \mathbf{x} |^2 \;=\; | x_1 |^2 + \cdots + | x_M |^2 \,, \tag{1.9}$$

which we recognize as the usual Euclidean norm of $\mathbf{x}$ when it is viewed as an element of $(d_1 + \cdots + d_M)$-dimensional space.

Properties of Piecewise Polynomials

For the moment, let m and d be fixed positive integers. We begin this section by establishing a useful fact about polynomials of total degree m in the variable $x \in \mathbb{R}^d$ as given by (1.3). Let M_1 be a positive constant, and let δ be any bounded subset of $\mathbb{R}^d$ having positive volume such that

$$\frac{(\operatorname{diam}\delta)^d}{\operatorname{vol}\delta} \;\leq\; M_1 \,, \tag{1.10}$$

where $\operatorname{diam}\delta$ is defined to be

$$\operatorname{diam}\delta \;=\; \sup \left\{ | x_1 - x_2 | \,:\, x_1, x_2 \in \delta \right\} .$$

This constraint enforces a certain amount of regularity on the set δ and is common in finite element applications [see Oden and Reddy (1976)]. We are now in a position to prove the following lemma.

Lemma 1.1 *Let m and d be positive integers and let $\delta \subset \mathbb{R}^d$ be any set with unit diameter that contains the origin. Then, if* (1.10) *holds, there exists a positive constant M_2 that depends only on m, d and M_1 such that*

$$\frac{1}{M_2} \sum_{j \leq m} | b_j |^2 \;\leq\; \int_\delta p^2(x)\, dx \;\leq\; M_2 \sum_{j \leq m} | b_j |^2 \,, \tag{1.11}$$

for all polynomials p of total degree m given by (1.3).

Proof Dividing both inequalities in (1.11) by $\sum_j | b_j |^2$, we find that it is sufficient to consider only those polynomials for which $\sum_j | b_j |^2 = 1$. Suppose we cannot find a constant M_2 such that

$$\frac{1}{M_2} \;\leq\; \int_\delta \Big(\sum_j b_j x^j \Big)^2 dx \tag{1.12}$$

for all sets δ with unit diameter that contain the origin, and all coefficient vectors $\{ b_j \}$ for which $\sum_j | b_j |^2 = 1$. Then we can find a sequence of sets δ_i and coefficient

vectors $\{b_{j\,i}\}$ such that

$$\int_{\delta_i} \Big(\sum_j b_{j\,i}\, x^j \Big)^2 dx \;\to\; 0 \qquad \text{as} \qquad i \to \infty\,,$$

where $\mathrm{diam}\,\delta_i = 1$ and $\sum_j |\, b_{j\,i}\,|^2 = 1$ for all i. Hence, we can extract a convergent subsequence of the vectors, $\{b_{j\,i'}\}$, the limit of which can be used to construct a polynomial p. Clearly, along this subsequence we also have that

$$\int_{\delta_{i'}} p^2(\,x\,)\,dx \;\to\; 0 \qquad \text{as} \qquad i' \to \infty. \tag{1.13}$$

Next, because $\mathrm{vol}\,\{\,x : |\,x\,| \le 1,\ \text{and}\ p^2(\,x\,) = 0\,\}$ is zero (by induction on d), we can find an $\epsilon > 0$ such that

$$\mathrm{vol}\,\{\,x : |\,x\,| \le 1,\ \text{and}\ p^2(\,x\,) < \epsilon\,\} \;<\; \frac{1}{2\,M_1},$$

which implies that for any δ contained in the unit ball in $\mathbb{R}^d$ having unit diameter and satisfying (1.10),

$$\int_{\delta} p^2(\,x\,)\,dx \;\ge\; \frac{\epsilon}{2\,M_1}.$$

However, by travelling out far enough along the subsequence in (1.13), we can find such a set $\delta_{i'}$ for which $\int_{\delta_{i'}} p^2(\,x\,)\,dx$ is strictly smaller than $\epsilon/(2M_1)$, yielding a contradiction. This establishes the existence of a constant M_2 satisfying the first inequality in (1.11).

Redefining M_2 if necessary, we find that the second inequality in (1.11) follows from the Schwartz inequality since x is restricted to the unit ball and there are not more than $\big(\,m+1\,\big)^d$ functions of the form x^j, where $j \le m$. □

Let $X_1, \ldots, X_M$ be random vectors taking values in the sets $\mathcal{X}_1, \ldots, \mathcal{X}_M$, respectively. We assume that $\mathcal{X}_l$ is a compact subset of $\mathbb{R}^{d_l}$ having unit volume for $1 \le l \le M$. Moreover, we divide each set $\mathcal{X}_l$ into a number of smaller sets which will play the role of δ in Lemma 1.1 when generalizing that result to the polynomials of coordinate degree m defined in (1.8).

Definition 1 *A collection of subsets $\triangle_l = \{\delta_{li} \subset \mathcal{X}_l, 1 \le i \le J_l\}$ forms a partition of $\mathcal{X}_l$ if*

(a) $\mathrm{vol}\,\delta_{li} > 0$,

(b) $\cup\, \delta_{li} = \mathcal{X}_l$,

(c) $\delta_{li} \cap \delta_{lj} = \varnothing$ *for* $i \ne j$.

As mentioned above, with each set $\mathcal{X}_l$ we associate a partition $\triangle_l$. Technically, we envision a sequence of such partitions indexed by sample size. That is, as we collect more and more data, we consider partitions made up of smaller and smaller sets. For the moment, we do not make this connection explicit. Independently of how we choose to refine the partitions $\triangle_l$, however, we require that each element of $\triangle_l$, $1 \leq l \leq M$, satisfies a slightly stronger stability constraint than that given in (1.10).

Condition 1 *Let $\triangle_l$ be given as in Definition 1 for $1 \leq l \leq M$. Assume that there exists a constant M_1 such that for $1 \leq l \leq M$ and $\delta \in \triangle_l$, there exists a ball $B_\delta \subset \delta$ such that*

$$(\operatorname{diam} \delta)^{d_l} \, / \, (\operatorname{vol} B_\delta) \;\; < \;\; M_1 \, .$$

Set $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_M$, $\mathbf{X} = (X_1, \ldots, X_M)$, and $\operatorname{Dim} \mathcal{X} = (d_1, \ldots, d_M)$. Observe that the partitions on each of the $\mathcal{X}_l$ induce a partition $\triangle = \triangle_1 \times \cdots \times \triangle_M$ on $\mathcal{X}$ which, by Definition 1, is made up of elements of the form

$$\delta = \delta_{1 i_1} \times \cdots \times \delta_{M i_M}, \qquad \text{where} \;\; 1 \leq i_l \leq J_l \;\; \text{for} \;\; 1 \leq l \leq M \, .$$

For any such element, define

$$\operatorname{Diam} \delta \;\; = \;\; \big(\operatorname{diam} \delta_{1 i_1}, \ldots, \operatorname{diam} \delta_{M i_M} \big).$$

Therefore, if we set $d = \operatorname{Dim} \mathcal{X}$ then the volume of $\delta \in \triangle$ is at most $(\operatorname{Diam} \delta)^d$. The ability to bound volumes of tensor product sets cleanly is the motivation behind the definition of $\operatorname{Dim} \mathcal{X}$ as a vector in $\mathbb{N}^M$.

The remainder of this section is devoted to recording facts about polynomials defined over certain subsets of $\mathcal{X}$. Throughout, we view m as a fixed vector in $\mathbb{N}^M$, and consider polynomials of coordinate degree m in the variable $\mathbf{x}$. The following result generalizes Lemma 1.1. It is important to note, however, that this extension, while worded in terms of sets $\delta \in \triangle$, does not depend on our choice of partitions provided that Condition 1 holds.

Lemma 1.2 *Set $d = \operatorname{Dim} \mathcal{X}$ and fix $m \in N^M$. If Condition 1 holds, there exists a positive constant M_2 that depends only on m, d, and M_1 such that for all $\delta \in \triangle$,*

$$\frac{1}{M_2} \sum_{\mathbf{j} \leq m} \big| \, b_{\mathbf{j}} \, \big|^2 \;\; \leq \;\; \frac{1}{h^d} \int_\delta p^2 \big[\, (\, \mathbf{x} - \mathbf{x}_0 \,) \, / \, h \, \big] \, d\mathbf{x} \;\; \leq \;\; M_2 \sum_{\mathbf{j} \leq m} \big| \, b_{\mathbf{j}} \, \big|^2 \, ,$$

where $h = \operatorname{Diam} \delta$, $\mathbf{x}_0$ is any point in δ, and p is any polynomial of coordinate degree m given by (1.8).

Proof For the moment, let δ be any set of the form $\delta_1 \times \cdots \times \delta_M$ where each $\delta_l \subset \mathbb{R}^{d_l}$ satisfies the conditions of Lemma 1.1. Clearly, if we can find a constant M_2 such that

$$\frac{1}{M_2} \sum_{\mathbf{j} \le m} \left| b_{\mathbf{j}} \right|^2 \;\le\; \int_\delta p^2(\mathbf{x})\, d\mathbf{x} \;\le\; M_2 \sum_{\mathbf{j} \le m} \left| b_{\mathbf{j}} \right|^2 , \tag{1.14}$$

for all such δ and all polynomials p given by (1.8), then the desired inequalities are obtained by a simple change of variables.

We will first verify (1.14) for the case $M = 2$. Toward this end, set $\mathbf{x} = (x_1, x_2)$ and $\mathbf{j} = (j_1, j_2)$, where x_1, j_1 are vectors d_1-dimensional vectors, and x_2, j_2 are d_2-dimensional vectors. Next, let p be any polynomial of total degree $m = (m_1, m_2) \in \mathbb{N}^2$ given by (1.8), and for each $j_1 \le m_1$, set

$$p_{j_1}(x_2) \;=\; \sum_{j_2 \le m_2} b_{(j_1, j_2)}\, x_2^{j_2} ,$$

so that

$$p(\mathbf{x}) \;=\; p(x_1, x_2) \;=\; \sum_{j_1 \le m_1} p_{j_1}(x_2)\, x_1^{j_1} .$$

Then,

$$\int_\delta p^2(\mathbf{x})\, d\mathbf{x} \;=\; \int_{\delta_2} \int_{\delta_1} \Big(\sum_{j_1 \le m_1} p_{j_1}(x_2)\, x_1^{j_1} \Big)^2 dx_1\, dx_2 ,$$

where $\delta = \delta_1 \times \delta_2$. Therefore, by Lemma 1.1, we know that there exists a constant M_2 depending only on d_1, m_1, and M_1 such that

$$\frac{1}{M_2} \sum_{j_1 \le m_1} \int_{\delta_2} p_{j_1}^2(x_2)\, dx_2 \;\le\; \int_\delta p^2(\mathbf{x})\, d\mathbf{x} \;\le\; M_2 \sum_{j_1 \le m_1} \int_{\delta_2} p_{j_1}^2(x_2)\, dx_2 .$$

Applying Lemma 1.1 again to the remaining integrals and redefining M_2 if necessary, we obtain the desired result. In a similar fashion, we establish (1.14) for general M by induction. □

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ denote a random sample of size n from the distribution of the random vector $\mathbf{X} = (X_1, \ldots, X_M)$, where again each X_l is assumed to range over the set $\mathcal{X}_l \subset \mathbb{R}^{d_l}$ for $1 \le l \le M$. For any function $f(\,\cdot\,)$ defined on $\mathcal{X}$, set

$$E_n\big[\, f(\mathbf{X}) \,\big] = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{X}_i) \quad \text{and} \quad \Big| E_n - E \Big| \big[\, f(\mathbf{X}) \,\big] = \Big| E_n\big[\, f(\mathbf{X}) \,\big] - E\big[\, f(\mathbf{X}) \,\big] \Big| .$$

We impose the following condition on the distribution of $\mathbf{X}$ to ensure the regular behaviour of the empirical quantities $E_n[\,\cdot\,]$.

Condition 2 *Let* $\mathbf{X} = (X_1, \ldots, X_M)$ *have a density function* f *such that*

$$\frac{1}{M_3} \;\leq\; f(\mathbf{x}) \;\leq\; M_3\,, \qquad \mathbf{x} \in \mathcal{X}.$$

Let $\delta \in \triangle$, set $h = \mathrm{Diam}\,\delta$, and choose $\mathbf{x}_0 \in \delta$. In the following lemma, we consider random variables of the form

$$\mathbf{W} = \left[\,(\,\mathbf{X} - \mathbf{x}_0\,)\,/\,h\,\right] I_\delta(\,\mathbf{X}\,)\,. \tag{1.15}$$

Given a binary random variable I and a nonnegative integer j, we define I^j to be zero if $I = 0$, even if $j = 0$. With this convention, given any polynomial p in the form (1.8), $p(\mathbf{W})$ is zero if $\mathbf{X}$ does not lie in δ. We are now in a position to state and prove the following lemma.

Lemma 1.3 *Suppose Conditions* 1 *and* 2 *hold, choose* $\delta \subset \triangle$, *and set* $h = \mathrm{Diam}\,\delta$ *and* $d = \mathrm{Dim}\,\mathcal{X}$. *For any point* $\mathbf{x}_0$ *in* δ, *define* $\mathbf{W}$ *as in* (1.15). *Then there exists a positive constant* M_4 *such that for any* $t > 0$, *the inequality*

$$\left|\,E_n - E\,\right|\left[\,p_1(\mathbf{W})\,p_2(\mathbf{W})\,\right] \;\leq\; t\,\sqrt{E\left[\,p_1^2(\mathbf{W})\,\right]}\,\sqrt{E\left[\,p_2^2(\mathbf{W})\,\right]} \tag{1.16}$$

holds simultaneously for all polynomials p_1, p_2 *of the form* (1.8), *except on an event having probability at most*

$$M_4\,\exp(\,-\eta\,/\,M_4\,)\,, \qquad \textit{where} \qquad \eta \;=\; n\,h^d\,t^2\,/\,(\,t+1\,)\,. \tag{1.17}$$

Proof For $\mathbf{j}_1$ and $\mathbf{j}_2$ defined as in (1.5), we consider random variables of the form $\mathbf{W}^{\mathbf{j}_1+\mathbf{j}_2}$, $\mathbf{j}_1, \mathbf{j}_2 \leq m$. By construction, these variables are bounded above in absolute value by one and

$$\mathrm{var}\left[\,\mathbf{W}^{\mathbf{j}_1+\mathbf{j}_2}\,\right] \;\leq\; M_3\,\mathrm{vol}\,\delta\,, \qquad \mathbf{j}_1, \mathbf{j}_2 \leq m.$$

Therefore, applying Bernstein's inequality [see Hoeffding (1963)] we find that for a single choice of $\mathbf{j}_1, \mathbf{j}_2$,

$$\left|\,E_n - E\,\right|\left[\,\mathbf{W}^{\mathbf{j}_1+\mathbf{j}_2}\,\right] \leq t \tag{1.18}$$

except on an event having probability at most

$$2\exp\left(\,-\,n\,\frac{t^2}{2\,(\,M_3\,\mathrm{vol}\,\delta\;+\;2\,t\,/\,3\,)}\,\right). \tag{1.19}$$

Since $\mathbf{j}_1$ and $\mathbf{j}_2$ are each bounded above by m, we certainly have that the sum

$\mathbf{j}_1 + \mathbf{j}_2 \leq 2m$. Therefore, we can inflate the probability in (1.19) by $\big(2m+1\big)^d$ so that the inequality in (1.18) is guaranteed to hold for all $\mathbf{j}_1, \mathbf{j}_2 \leq m$.

Now, let $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ be polynomials in the form (1.8). Then,

$$p_1(\mathbf{x}) = \sum_{\mathbf{j}_1 \leq m} b_{1\,\mathbf{j}_1}\, \mathbf{x}^{\mathbf{j}_1} \qquad \text{and} \qquad p_2(\mathbf{x}) = \sum_{\mathbf{j}_2 \leq m} b_{2\,\mathbf{j}_2}\, \mathbf{x}^{\mathbf{j}_2},$$

where $b_{1\,\mathbf{j}_1}, b_{2\,\mathbf{j}_2} \in \mathbb{R}$ for $\mathbf{j}_1, \mathbf{j}_2 \leq m$. On the event that (1.18) holds for all $\mathbf{j}_1, \mathbf{j}_2 \leq m$, we have that

$$\big|\, E_n - E\,\big|\big[\, p_1(\mathbf{W})\, p_2(\mathbf{W})\,\big] \;\leq\; t \sum_{\mathbf{j}_1 \leq m} \big|\, b_{1\,\mathbf{j}_1}\,\big| \sum_{\mathbf{j}_2 \leq m} \big|\, b_{2\,\mathbf{j}_2}\,\big|. \tag{1.20}$$

In addition, by Conditions 1 and 2 and Lemma 1.2,

$$\begin{aligned} M_3\, E\big[\, p_1^2(\mathbf{W})\,\big] \;&\geq\; \int_\delta p_1^2\big[\,(\,\mathbf{x} - \mathbf{x}_0\,)\,/\,h\,\big]\, d\mathbf{x} \\ &\geq\; \big(\,h^d/M_2\,\big) \sum_{\mathbf{j}_1 \leq m} \big|\, b_{1\,\mathbf{j}_1}\,\big|^2. \end{aligned}$$

Combining this with a similar result for $p_2(\mathbf{W})$ and applying the Schwartz inequality, we find that (1.20) becomes

$$\big|\, E_n - E\,\big|\big[\, p_1(\mathbf{W})\, p_2(\mathbf{W})\,\big] \;\leq\; \frac{t}{h^d}\, M_2\, M_3\, \big(\,m+1\,\big)^d \sqrt{E\big[\, p_1^2(\mathbf{W})\,\big]}\, \sqrt{E\big[\, p_2^2(\mathbf{W})\,\big]}\,.$$

Replacing t by $T_1\, t\, h^d$ for some positive constant T_1 and choosing M_4 sufficiently large, we arrive at the expression in (1.17). □

The definition in (1.15) will be our starting point for constructing the space of piecewise polynomials on $\mathcal{X}$. For the moment, fix a set $\delta \in \triangle$. Choose a point $\mathbf{x}_0$ in δ and set $h = \mathrm{Diam}\,\delta$. Then, for $\mathbf{j} \leq m$, define

$$p_{\mathbf{j}}(\mathbf{x}) \;=\; \big[\,(\,\mathbf{x} - \mathbf{x}_0\,)\,/\,h\,\big]^{\,\mathbf{j}}\, I_\delta(\mathbf{x})\,, \qquad \mathbf{x} \in \mathcal{X}\,. \tag{1.21a}$$

As mentioned above, these functions are nonzero only for $\mathbf{x} \in \delta$, and they are bounded above by one in absolute value. In addition, these functions form a basis for the space of all polynomials of coordinate degree m on δ. Therefore, any such polynomial p_δ can be written as

$$p_\delta(\mathbf{x}) \;=\; \sum_{\mathbf{j} \leq m} b_{\mathbf{j}}\, p_{\mathbf{j}}(\mathbf{x})\,, \qquad b_{\mathbf{j}} \in \mathbb{R} \;\text{ for }\; \mathbf{j} \leq m\,. \tag{1.21b}$$

Repeat this construction process for each $\delta \in \triangle$, creating (possibly) different polynomials p_δ on each set. Finally, we define a piecewise polynomial of coordinate

degree m on $\triangle$ to be any function of the form

$$p(\mathbf{x}) \;=\; \sum_{\delta \in \triangle} p_\delta(\mathbf{x})\,. \tag{1.21c}$$

Let $\mathbb{PP}$ denote the space of all functions in the form (1.21).

Lemma 1.4 *Suppose Conditions* 1 *and* 2 *hold. Set* $\underline{h} = \min\{\,\mathrm{Diam}\,\delta : \delta \in \triangle\,\}$ *and* $d = \mathrm{Dim}\,\mathcal{X}$. *Then, there exists a positive constant* M_4 *such that for any* $t > 0$, *the inequality*

$$\big|\,E_n - E\,\big|\big[\,p_1(\mathbf{X})p_2(\mathbf{X})\,\big] \;\;\le\;\; t\,\sqrt{E\big[\,p_1^2(\mathbf{X})\,\big]}\,\sqrt{E\big[\,p_2^2(\mathbf{X})\,\big]} \tag{1.22}$$

holds simultaneously for all piecewise polynomials $p_1, p_2 \in \mathbb{PP}$, *except on an event having probability at most*

$$M_4\,\underline{h}^{-d}\,\exp(\,-\eta\,/\,M_4\,)\,, \qquad \textit{where} \qquad \eta \;=\; n\,\underline{h}^{\,d}\,t^2\,/\,(\,t+1\,)\,. \tag{1.23}$$

Proof For each $\delta \in \triangle$, let $p_{1\delta}, p_{2\delta}$ be given by (1.21b), and set

$$p_1(\mathbf{x}) \;=\; \sum_{\delta \in \triangle} p_{1\delta}(\mathbf{x}) \qquad \text{and} \qquad p_2(\mathbf{x}) \;=\; \sum_{\delta \in \triangle} p_{2\delta}(\mathbf{x})\,.$$

Set $\underline{v} = \min\{\,\mathrm{vol}\,\delta \,:\, \delta \in \triangle_s\,\}$, and observe that there are at most $(\,\mathrm{vol}\,\mathcal{X}\,/\,\underline{v}\,)$ elements of $\triangle$. The results of Lemma 1.3 can be applied to any particular element of $\delta \in \triangle$, and therefore, by inflating the exceptional probability in (1.17) by

$$\frac{(\,M_1^M\,)\;\mathrm{vol}\,\mathcal{X}}{\underline{h}^{\,d}} \;\ge\; \frac{\mathrm{vol}\,\mathcal{X}}{\underline{v}}\,,$$

the inequality in (1.16) can be made to hold for all sets in the partition $\triangle$. On this event, we have that

$$\begin{aligned}
\big|\,E_n - E\,\big|\big[\,p_1(\mathbf{X})\,p_2(\mathbf{X})\,\big] \;&\le\; \sum_{\delta \in \triangle} \big|\,E_n - E\,\big|\big[\,p_{1\delta}(\mathbf{X})\,p_{2\delta}(\mathbf{X})\,\big] \\
&\le\; t \sum_{\delta \in \triangle} \Big(\,E\big[\,p_{1\delta}^2(\mathbf{X})\,\big]\cdot E\big[\,p_{2\delta}^2(\mathbf{X})\,\big]\,\Big)^{1/2} \\
&\le\; t\,\Big(\sum_{\delta \in \triangle} E\big[\,p_{1\delta}^2(\mathbf{X})\,\big] \sum_{\delta \in \triangle} E\big[\,p_{2\delta}^2(\mathbf{X})\,\big]\,\Big)^{1/2} \\
&=\; t\,\sqrt{E\big[\,p_1^2(\mathbf{X})\,\big]}\,\sqrt{E\big[\,p_2^2(\mathbf{X})\,\big]}\,.
\end{aligned}$$

Redefining M_4 if necessary, we arrive at the desired conclusion. □

In the previous lemma, we derived the relationship between the empirical expected value $E_n[\,p(\mathbf{X})\,]$ and its theoretical counterpart $E[\,p(\mathbf{X})\,]$ for functions $p \in \mathbb{PP}$. We end this section by exploring the relationship between this result and the identifiability of $\mathbb{PP}$. As in the previous lemma, let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ denote a sample of size n from the distribution of $\mathbf{X}$. Then for any two functions $f_1(\,\cdot\,)$ and $f_2(\,\cdot\,)$ defined on $\mathcal{X}$, set

$$\langle\, f_1, f_2 \,\rangle_n \;=\; E_n[\, f_1(\mathbf{X}) f_2(\mathbf{X}) \,] \qquad \text{and} \qquad \|\, f_1 \,\|_n^2 \;=\; E_n[\, f_1^2(\mathbf{X}) \,]. \quad (1.24a)$$

The theoretical versions of these quantities are given by

$$\langle\, f_1, f_2 \,\rangle \;=\; E[\, f_1(\mathbf{X}) f_2(\mathbf{X}) \,] \qquad \text{and} \qquad \|\, f_1 \,\|^2 \;=\; E[\, f_1^2(\mathbf{X}) \,] \qquad (1.24b)$$

Recall that as our sample size increases, we want to consider finer and finer partitions $\triangle_l$, $1 \le l \le M$. Intuitively it is clear that we cannot refine these partitions too quickly and expect to maintain identifiability of the space $\mathbb{PP}$. Condition 3 and Lemma 1.5 make this statement precise.

Condition 3 *Define the quantities*

$$\overline{h} = \max\{\, \mathrm{Diam}\,\delta : \delta \in \triangle \,\}, \quad \underline{h} = \min\{\, \mathrm{Diam}\,\delta : \delta \in \triangle \,\} \quad \textit{and} \quad d = \mathrm{Dim}\,\mathcal{X},$$

and assume that the partitions $\triangle_l$, $1 \le l \le M$, are refined with sample size so that as $n \to \infty$,

$$\overline{h} \to \mathbf{0} \qquad \textit{while} \qquad \log \underline{h}^d \,+\, n\,\underline{h}^d \to \infty\,. \qquad (1.25)$$

Lemma 1.5 *If Conditions 1, 2 and 3 hold, then except on an event whose probability tends to zero with n, $\mathbb{PP}$ is identifiable.*

Proof From Lemma 1.4, if (1.25) holds, then except on an event whose probability tends to zero with n,

$$\tfrac{1}{2}\,\|\,p\,\|^2 \;\le\; \|\,p\,\|_n^2 \;\le\; 2\,\|\,p\,\|^2, \qquad p \in \mathbb{PP}. \qquad (1.26)$$

Suppose we can find a function $p \in \mathbb{PP}$ such that $p(\mathbf{X}_i)$ is zero for $1 \le i \le n$. Then, $\|\,p\,\|_n^2$ equals zero, and hence if (1.26) holds, $\|\,p\,\|^2$ is also zero. By Condition 2, this implies that p is identically zero. Therefore, if (1.26) holds, $\mathbb{PP}$ is identifiable. □

Suppose that the mesh ratio $\overline{h}\,/\,\underline{h}$ remains bounded as n tends to infinity, and that $\underline{h}^{-d}$ is taken to be $o\big(n^{1-\epsilon}\big)$ for some $\epsilon > 0$. Then, the condition in (1.25) is satisfied so that under Conditions 1 and 2, $\mathbb{PP}$ is identifiable.

1.2 Rate of Convergence in Saturated Spaces

We now consider using the spaces discussed in Section 1.1 for prediction. Let $\mathbf{X} = (X_1, \ldots, X_M)$ represent a vector of predictor variables, and let Y represent our response variable. We make the following assumptions about the distribution of $\mathbf{X}$ and Y.

Condition 4 *Let $\mathbf{X}$ and Y have a joint distribution, and let the marginal density of $\mathbf{X}$ satisfy Condition 2. Set*

$$\mu(\mathbf{x}) \;=\; E\big(Y \,|\, \mathbf{X} = \mathbf{x}\big) \qquad \textit{and} \qquad \sigma^2(\mathbf{x}) \;=\; \mathrm{var}\big(Y \,|\, \mathbf{X} = \mathbf{x}\big)\,,$$

and assume that both $\mu(\mathbf{x})$ and $\sigma^2(\mathbf{x})$ are bounded in absolute value on $\mathcal{X}$. In particular, assume that there exists a positive constant M_5 such that $\sigma^2(\mathbf{x}) \le M_5$ for all $\mathbf{x} \in \mathcal{X}$.

Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ represent a random sample of size n from the distribution of $(\mathbf{X}, Y)$. If Condition 2 holds, we know that the design points $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are unique with probability one, and therefore we can find a function defined on $\mathcal{X}$ that interpolates the values $Y_1, \ldots, Y_n$ at these points. Let $Y(\,\cdot\,)$ denote any such function and let $\hat{\mu}(\mathbf{x})$ denote the orthogonal projection of $Y(\,\cdot\,)$ onto $\mathbb{PP}$ relative to the empirical inner product defined in (1.24a). Recall that as n increases, the cells of the partition $\triangle$ are refined according to (1.25), increasing the flexibility of the space $\mathbb{PP}$. Therefore, it is reasonable to expect that $\hat{\mu}$ would approach μ as n tends to infinity. To make this precise, for a sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ we define the mean square error of $\hat{\mu}$ by

$$\big\|\, \hat{\mu} - \mu \,\big\|^2 \;=\; E\big[\, \hat{\mu}(\mathbf{X}) - \mu(\mathbf{X}) \,\big]^2\,,$$

where the expectation is taken with respect to $\mathbf{X}$ holding the sample data fixed. In this section, we derive the rate at which the mean squared error tends to zero in probability, or rather, the L_2 rate of convergence of $\hat{\mu}$ to μ.

The mean squared error can be bounded from above by

$$\big\|\, \hat{\mu} - \mu \,\big\|^2 \;\le\; 2\,\big\|\, E(\,\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\,) - \hat{\mu} \,\big\|^2 + 2\,\big\|\, E(\,\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\,) - \mu \,\big\|^2\,.$$

In this expression, the first and second terms on the right are referred to as variance and bias components, respectively. In Lemmas 2.1 and 2.2 we investigate how quickly the variance component tends to zero in probability with n, while in Lemmas 2.3 through 2.5 we consider the speed at which the bias component converges to zero in probability with n.

Variance

Consider the variance component of the mean squared error given above. Observe that by Condition 3 and the fact that $\hat{\mu}(\mathbf{x})$ is a linear function of the data values $Y_1, \ldots, Y_n$, we can exchange the order of integration and write the conditional expected value

$$E\Big(\| \hat{\mu} - E(\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) \|^2 \,\Big|\, \mathbf{X}_1, \ldots \mathbf{X}_n \Big)$$

as an integral of the conditional variance

$$\int_{\mathcal{X}} \mathrm{var}\left(\hat{\mu}(\mathbf{x}) \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n \right) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \;\leq\; \sup_{\mathbf{x} \in \mathcal{X}} \mathrm{var}\left(\hat{\mu}(\mathbf{x}) \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n \right).$$

Suppose we can show that the quantity on the righthand side of the above expression is $O_P(a_n)$ for some sequence $\{a_n\}$ of positive numbers. Then, the variance component of the mean squared error of $\hat{\mu}$ is also $O_P(a_n)$.

To see this, consider temporarily a sequence of positive random variables V_n and another sequence of random variables X_n such that $E(V_n \,|\, X_n)$ is $O_P(a_n)$. Then, given $t_1, t_2 > 0$, we have that

$$P(V_n > t_1) \;\leq\; P(V_n > t_1,\, E(V_n \,|\, X_n) \leq t_2) + P(E(V_n \,|\, X_n) > t_2).$$

Now, by Markov's inequality we find that

$$\begin{aligned} P(V_n > t_1,\, E(V_n \,|\, X_n) &\leq t_2) \\ &= E\Big(\mathrm{ind}\left(E(V_n \,|\, X_n) \leq t_2 \right) P\left(V_n > t_1 \,|\, X_n \right) \Big) \\ &\leq E\Big(\mathrm{ind}(E(V_n \,|\, X_n) \leq t_2)\, E(V_n \,|\, X_n) \Big) / t_1 \\ &\leq t_2 / t_1 . \end{aligned}$$

Therefore, for $c > 0$,

$$P(V_n > c\, a_n) \;\leq\; t_2 / (c\, a_n) + P(E(V_n \,|\, X_n) > t_2),$$

and so by replacing t_2 with a_n, we find that

$$\lim_{c \to \infty} \limsup_n P(V_n > c\, a_n) \;=\; 0$$

and hence V_n is also $O_P(a_n)$.

Returning to our original problem, we find that in order to determine the size of the variance component of the mean squared error of $\hat{\mu}$, it is sufficient to determine

the size of $\mathrm{var}\left(\hat{\mu}(\mathbf{x}) \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\right)$. Toward this end, we introduce two sets of special linear functionals. First, for $\mathbf{x}' \in \mathcal{X}$, let $q_{\mathbf{x}'} \in \mathbb{PP}$ denote the representor of the linear functional $p \mapsto p(\mathbf{x}')$ on $\mathbb{PP}$ relative to the empirical inner product $\langle \cdot, \cdot \rangle_n$, so that

$$p(\mathbf{x}') \;=\; \langle q_{\mathbf{x}'}, p \rangle_n . \tag{2.1a}$$

Next, let δ be that set in $\triangle$ containing $\mathbf{x}'$. Then, given any function p in $\mathbb{PP}$ defined as in (1.21), let

$$p(\mathbf{x}) \;=\; p_\delta(\mathbf{x}) \;=\; \sum_{\mathbf{j} \le m} b_{\mathbf{j}}\, p_{\mathbf{j}}(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \delta\,, \tag{2.1b}$$

where $b_{\mathbf{j}} \in \mathbb{R}$ for $\mathbf{j} \le m$, and the functions $p_{\mathbf{j}}$ are defined as in (1.21b). Now, for our choice of δ, consider the linear functionals $p \mapsto b_{\mathbf{j}}$ on $\mathbb{PP}$ relative to the inner product $\langle \cdot, \cdot \rangle_n$. That is, for each $\mathbf{j} \le m$, let $q_{\mathbf{j}} \in \mathbb{PP}$ be such that

$$b_{\mathbf{j}} \;=\; \langle q_{\mathbf{j}}, p \rangle_n . \tag{2.2a}$$

Since the sets in $\triangle$ are disjoint, $q_{\mathbf{j}}$ can also be written in the form

$$q_{\mathbf{j}}(\mathbf{x}) \;=\; \sum_{\mathbf{j}' \le m} \gamma_{\mathbf{j}\mathbf{j}'}\, p_{\mathbf{j}'}(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \delta\,. \tag{2.2b}$$

Note that as was the case with the functions $p_{\mathbf{j}}, \mathbf{j} \le m$, the functions $q_{\mathbf{j}}, \mathbf{j} \le m$, are defined with respect to a particular set $\delta \in \triangle$. In Section 1.1, we found that we could derive results about the global nature of functions in $\mathbb{PP}$ by analyzing their behaviour over the individual sets $\delta \in \triangle$. Therefore, by hiding the dependence of $p_{\mathbf{j}}$ and $q_{\mathbf{j}}, \mathbf{j} \le m$, on δ we can simplify our notation considerably without sacrificing clarity. We now record a number of facts about $q_{\mathbf{x}'}$ and the associated functions $q_{\mathbf{j}}, \mathbf{j} \le m$, in the following lemma and its proof.

Lemma 2.1 *Suppose Conditions* 1, 2 *and* 3 *hold. Then there exists a positive constant* M_6 *that does not depend on* n *or* $\overline{h}$ *such that, except on a set whose probability tends to zero with* n,

$$q_{\mathbf{x}'}(\mathbf{x}') \;=\; \left\| q_{\mathbf{x}'} \right\|_n^2 \;\le\; M_6\, \underline{h}^{-d}\,, \qquad \mathbf{x}' \in \mathcal{X}\,. \tag{2.3}$$

Proof Select $\mathbf{x}' \in \mathcal{X}$. Let δ be that element of $\triangle$ containing $\mathbf{x}'$ and set $h = \mathrm{Diam}\,\delta$. Let p be any element of $\mathbb{PP}$, and let the functions $q_{\mathbf{x}'}$ and $q_{\mathbf{j}}, \mathbf{j} \le m$, be as in (2.1) and (2.2). Then, if Condition 3 holds, except on an event whose probability tends to zero with n,

$$h^d\, \gamma_{\mathbf{j}\mathbf{j}}^2 \;\le\; h^d \sum_{\mathbf{j}' \le m} \gamma_{\mathbf{j}\mathbf{j}'}^2 \;\le\; M_2\, M_3\, E\left[\, q_{\mathbf{j}}^2(\mathbf{X}) \,\right] \;\le\; 2\, M_2\, M_3 \left\| q_{\mathbf{j}} \right\|_n^2 \;=\; 2\, M_2\, M_3\, \gamma_{\mathbf{j}\mathbf{j}}$$

where the second inequality follows from Condition 2 and Lemma 1.2 applied to $q_{\mathbf{j}}$, and the third inequality follows from (1.26). Therefore, except on a set whose probability tends to zero with n,

$$\sum_{\mathbf{j}' \le m} \gamma^2_{\mathbf{j}\mathbf{j}'} \;\le\; \bigl(2\,M_2\,M_3\bigr)^2\,\underline{h}^{-2d} \qquad \text{for } \mathbf{j} \le m\,, \tag{2.4}$$

where we have used the fact that $\underline{h} \le h$. It follows from (2.1b) and (2.2a) that

$$p(\mathbf{x}') \;=\; \sum_{\mathbf{j}\le m} \langle q_{\mathbf{j}}\,,\,p\rangle_n\; p_{\mathbf{j}}(\mathbf{x}') \;=\; \left\langle \sum_{\mathbf{j}\le m} q_{\mathbf{j}}\,p_{\mathbf{j}}(\mathbf{x}')\,,\,p \right\rangle_n .$$

Therefore, we can write the representor given in (2.1) as the sum

$$q_{\mathbf{x}'}(\mathbf{x}) \;=\; \sum_{\mathbf{j}\le m} q_{\mathbf{j}}(\mathbf{x})\,p_{\mathbf{j}}(\mathbf{x}')\,.$$

From this last expression we find that except on a set whose probability tends to zero with n,

$$\begin{aligned} q_{\mathbf{x}'}(\mathbf{x}') \;&=\; \sum_{\mathbf{j}\le m}\sum_{\mathbf{j}'\le m} \gamma_{\mathbf{j}\mathbf{j}'}\,p_{\mathbf{j}+\mathbf{j}'}(\mathbf{x}') \\ &\le\; \bigl(m+1\bigr)^{3d/2}\;\max_{\mathbf{j}\le m}\left(\sum_{\mathbf{j}'\le m}\gamma^2_{\mathbf{j}\mathbf{j}'}\right)^{1/2} \end{aligned}$$

by the Schwartz inequality. Finally, except on a set whose probability tends to zero with n, the inequalities leading to (2.4) hold simultaneously for all the representors associated with the sets $\delta \in \triangle$, and hence these results hold uniformly for $\mathbf{x}' \in \mathcal{X}$, as desired. □

Lemma 2.2 *Suppose Conditions* 1–4 *hold. Then*

$$\bigl\|\,\hat\mu - E(\hat\mu\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n)\,\bigr\|^2 \;=\; O_P\bigl(\underline{h}^{-d}/n\bigr)\,. \tag{2.5}$$

Proof Assume that $\mathbb{PP}$ is identifiable. Fix $\mathbf{x}' \in \mathcal{X}$ and let δ be the set in $\triangle$ containing $\mathbf{x}'$. Let $q_{\mathbf{x}'}$ be defined as in (2.1). Then, since $q_{\mathbf{x}'} \in \mathbb{PP}$,

$$\hat\mu(\mathbf{x}') \;=\; \langle q_{\mathbf{x}'},\hat\mu\rangle_n \;=\; \bigl\langle q_{\mathbf{x}'},Y(\,\cdot\,)\bigr\rangle_n\,.$$

Thus, by Condition 4,

$$\mathrm{var}\bigl(\hat\mu(\mathbf{x}')\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\bigr) \;=\; \frac{1}{n^2}\sum_{i=1}^n q^2_{\mathbf{x}'}(\mathbf{X}_i)\,\sigma^2(\mathbf{X}_i) \;\le\; \frac{1}{n}\,M_5\,\bigl\|\,q_{\mathbf{x}'}\,\bigr\|_n^2\,.$$

Therefore, by Lemma 2.1, except on a set whose probability tends to zero with n,

$$\mathrm{var}\bigl(\hat\mu(\mathbf{x}')\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\bigr) \;\le\; \frac{1}{n}\,M_5\,M_6\,\underline{h}^{-d}\,, \qquad \mathbf{x}' \in \mathcal{X}\,.$$

Taking the supremum over $\mathbf{x}' \in \mathcal{X}$, the desired result now follows from the discussion at the beginning of this subsection. □

Bias

Before proceeding with bounding the bias component of the mean squared error, we record a number of facts about the function $E(\hat{\mu}\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n)$. Recall that $\hat{\mu}$ is the orthogonal projection of the $Y(\cdot)$ onto $\mathbb{PP}$ relative to the empirical inner product. Similarly, let $\tilde{\mu}$ denote the orthogonal projection of μ onto $\mathbb{PP}$ relative to the empirical inner product. Pick $\mathbf{x}' \in \mathcal{X}$ and define the linear functional $q_{\mathbf{x}'}$ as in (2.1). Therefore, since $\hat{\mu}$ is in $\mathbb{PP}$,

$$\hat{\mu}(\mathbf{x}') \;=\; \langle q_{\mathbf{x}'}, \hat{\mu}\rangle_n \;=\; \big\langle q_{\mathbf{x}'}, Y(\cdot)\big\rangle_n .$$

Taking the conditional expectation given the design points $\mathbf{X}_1,\ldots,\mathbf{X}_n$, we obtain

$$E\big(\hat{\mu}(\mathbf{x}')\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\big) \;=\; E\Big(\big\langle q_{\mathbf{x}'}, Y(\cdot)\big\rangle_n \,\big|\, \mathbf{X}_1,\ldots,\mathbf{X}_n\Big) \;=\; \langle q_{\mathbf{x}'}, \mu\rangle_n .$$

Similarly, by the definition of $\tilde{\mu}$, we know that

$$\tilde{\mu}(\mathbf{x}') \;=\; \langle q_{\mathbf{x}'}, \tilde{\mu}\rangle_n \;=\; \langle q_{\mathbf{x}'}, \mu\rangle_n ,$$

and hence that

$$E\big(\hat{\mu}(\mathbf{x}')\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\big) \;=\; \tilde{\mu}(\mathbf{x}') .$$

Therefore, since $\mathbf{x}' \in \mathcal{X}$ was arbitrary, we conclude that $E(\hat{\mu}\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n)$ is the orthogonal projection of μ onto $\mathbb{PP}$ relative to $\langle\cdot,\cdot\rangle_n$. Since $\tilde{\mu} \in \mathbb{PP}$, we can write

$$E\big(\hat{\mu}_s(\mathbf{x}_s)\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\big) \;=\; \tilde{\mu}(\mathbf{x}) \;=\; \sum_{\delta\in\triangle} \tilde{\mu}_\delta(\mathbf{x}) , \tag{2.6}$$

where, following (1.21c), each function $\tilde{\mu}_\delta$ is a polynomial of coordinate degree m. Select $\delta \in \triangle$ and set $h = \mathrm{Diam}\,\delta$. For this choice of δ, write

$$\tilde{\mu}_\delta(\mathbf{x}) \;=\; \sum_{\mathbf{j}\le m} \tilde{\beta}_{\mathbf{j}}\, p_{\mathbf{j}}(\mathbf{x}) , \tag{2.7}$$

as in (1.21b). While the result of the following lemma is most easily motivated at this point in our discussion, it will not be used explicitly until Section 1.3.

Lemma 2.3 *Suppose Conditions* 1, 2 *and* 3 *hold. Then, there exists a constant* M_7 *such that except on a set whose probability tends to zero with* n, *for* $\delta \in \triangle$ *the function* $\tilde{\mu}_\delta$ *defined in* (2.7) *satisfies*

$$\big\|\,\tilde{\mu}_\delta\,\big\|_n^2 \;\le\; M_7\, h^{-d} \sum_{\mathbf{j}\le m} \langle p_{\mathbf{j}}, \mu\rangle_n^2 , \tag{2.8}$$

where the functions $p_{\mathbf{j}}$, $\mathbf{j} \le m$, *are defined in* (1.21a), *and* $h = \mathrm{diam}\,\delta$.

Proof Assume that $\mathbb{PP}$ is identifiable and choose $\delta \in \triangle$. From Condition 2, the second inequality in (1.26), and Lemma 1.2 applied to the function $\tilde{\mu}_\delta$ defined in (2.7), we find that except on an event whose probability tends to zero with n,

$$\left\| \tilde{\mu}_\delta \right\|_n^2 \;\leq\; 2 \left\| \tilde{\mu}_\delta \right\|^2 \;\leq\; 2\, M_2\, M_3\, h^d \sum_{\mathbf{j} \leq m} \tilde{\beta}_{\mathbf{j}}^2 . \tag{2.9}$$

For our choice of δ, construct the functionals $q_{\mathbf{j}}$, $\mathbf{j} \leq m$, appearing in (2.2) and recall that the coefficients appearing in (2.9) are given by

$$\tilde{\beta}_{\mathbf{j}} \;=\; \langle q_{\mathbf{j}}, \tilde{\mu}_\delta \rangle_n \;=\; \langle q_{\mathbf{j}}, \mu \rangle_n \;=\; \sum_{\mathbf{j}' \leq m} \gamma_{\mathbf{j}\mathbf{j}'} \langle p_{\mathbf{j}'}, \mu \rangle_n .$$

Therefore, from (2.4) and (2.9) we find that except on a set with probability that tends to zero with n

$$\left\| \tilde{\mu}_\delta \right\|_n^2 \;\leq\; M_7\, h^{-d} \sum_{\mathbf{j} \leq m} \langle p_{\mathbf{j}}, \mu \rangle_n^2 ,$$

where

$$M_7 \;=\; (\, 2\, M_2\, M_3 \,)^3 \, (\, m + 1 \,)^d . \tag{2.10}$$

Repeating the arguments at the end of the proof of Lemma 2.1, we observe that this inequality holds simultaneously for all $\delta \in \triangle$ except on a set whose probability tends to zero with n. □

Under suitable conditions on the regularity of the sets $\mathcal{X}_l$, $1 \leq l \leq M$, it is reasonable to assume that we would be able to achieve the optimal approximation rate from $\mathbb{PP}$. In the next section, however, we will deal with ANOVA decompositions, and in particular we will consider subspaces of $\mathbb{PP}$. In that case, the rate of approximation must be settled to a large extent by approximation theorists. Therefore, from the point of view of the present chapter, it is reasonable to express our requirements on the approximation power of the spaces in question in the form of a condition. In what follows, we measure the error in approximating μ by functions in $\mathbb{PP}$ using $\| \cdot \|_\infty$, the sup norm over $\mathcal{X}$.

Condition 5 *Recall that* $\overline{h} = \max \{ \operatorname{Diam} \delta : \delta \in \triangle \}$. *Assume that there exists a function* $\rho(\, \overline{h} \,)$ *such that*

$$\rho(\, \overline{h} \,) \to 0 \quad \textit{as} \quad \overline{h} \to 0 \qquad \textit{and} \qquad \inf_p \left\| p - \mu \right\|_\infty \;=\; O\Big(\rho(\, \overline{h} \,) \Big) ,$$

where the infinum is taken over all $p \in \mathbb{PP}$.

In proving Lemma 2.4, we found that except on a set whose probability tends to zero with sample size, $\| \cdot \|$ and $\| \cdot \|_n$ are equivalent over $\mathbb{PP}$ providing we do not

refine our partitions too quickly. As is shown in the following lemma, Condition 5 allows us to derive part of this relationship for a wider class of functions.

Lemma 2.4 *Suppose Conditions* 1–3 *and* 5 *hold. Then there is a positive constant* M_8 *that does not depend on* n *or* $\overline{h}$ *such that, except on an event whose probability tends to zero with* n,

$$\| p - \mu \|^2 \;\le\; M_8 \left(\rho^2(\overline{h}) + \| p - \mu \|_n^2 \right), \qquad p \in \mathbb{PP}.$$

Proof By Conditions 3 and 5, there is a function $p_1 \in \mathbb{PP}$ and a positive constant T_1 that does not depend on n or $\overline{h}$ such that

$$\| p_1 - \mu \|_\infty^2 \;\le\; T_1 \rho^2(\overline{h}).$$

Therefore,

$$\| p_1 - \mu \|^2 \;\le\; T_1 \rho^2(\overline{h}) \quad \text{and} \quad \| p_1 - \mu \|_n^2 \;\le\; T_1 \rho^2(\overline{h}), \tag{2.11}$$

where we redefine T_1 if necessary to account for the volume of $\mathcal{X}$. From the triangle and Schwartz inequalities we conclude that

$$\| p - \mu \|^2 \;\le\; 2 \| p_1 - p \|^2 + 2 T_1 \rho^2(\overline{h})$$

and

$$\| p_1 - p \|_n^2 \;\le\; 2 \| p - \mu \|_n^2 + 2 T_1 \rho^2(\overline{h}).$$

Thus, from Condition 2 and the first inequality (1.26) we find that, except on a set whose probability tends to zero with n,

$$\| p - \mu \|^2 \;\le\; 4 \| p - p_1 \|_n^2 + 2 T_1 \rho^2(\overline{h}) \;\le\; 8 \| p - \mu \|_n^2 + 10 T_1 \rho^2(\overline{h}),$$

as desired. □

Lemma 2.5 *Suppose Conditions* 1–3 *and* 5 *hold. Then*

$$\| E(\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) - \mu \|_n^2 \;=\; O_P\Big(\rho^2(\overline{h}) \Big).$$

Proof Assume that $\mathbb{PP}$ is identifiable. Since $E(\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n)$ is the orthogonal projection of μ onto $\mathbb{PP}$ relative to the empirical inner product,

$$\| E(\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) - \mu \|_n^2 \;\le\; \| p - \mu \|_n^2 \;\le\; \| p - \mu \|_\infty^2,$$

for all $p \in \mathbb{PP}$. The desired conclusion now follows from Condition 5.

Rate of Convergence

Theorem 2.1 *Suppose Conditions* 1–5 *hold. Then*

$$\left\| \hat{\mu} - \mu \right\|^2 \;=\; O_P\Big(\rho^2(\overline{h}) + \underline{h}^{-d} / n \Big) . \tag{2.12}$$

Proof Recall the variance-bias decomposition introduced at the beginning of this section,

$$\left\| \hat{\mu} - \mu \right\|^2 \;\leq\; 2 \left\| E(\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) - \hat{\mu} \right\|^2 + 2 \left\| E(\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) - \mu \right\|^2 .$$

From Lemma 2.2, we find that the first term in this decomposition is $O_P(\underline{h}^{-d} / n)$. Applying Lemma 2.4 with $p = E(\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathbb{PP}$, we find from Lemma 2.5 that the second term in this decomposition is $O_P(\rho^2(\overline{h}))$. □

Example 1 Suppose that for positive constants $T_1, \ldots, T_M$, the function $\rho^2(\overline{h})$ is given by

$$\rho^2(\overline{h}) \;=\; \sum_{l=1}^{M} T_l \, \overline{h}_l^{\,2p_l} \qquad \text{where } p_l \in \mathbb{N} \text{ for } 1 \leq l \leq M , \tag{2.13}$$

and set

$$d \;=\; (d_1, \ldots, d_M) \qquad \text{and} \qquad p \;=\; (p_1, \ldots, p_M) .$$

Now, suppose that for each $1 \leq l \leq M$, we take $\underline{h}_l \sim \overline{h}_l = O(n^{-\gamma_l})$, where

$$\gamma_l \;=\; \frac{1}{p_l \left(2 + [d / p] \right)} , \tag{2.14}$$

With this choice of $\overline{h}$ and $\underline{h}$, we now consider the size of the two components that comprise the rate of convergence given in Theorem 2.1. Each term appearing in (2.13) and hence $\rho^2(\overline{h})$ itself is $O(n^{-\gamma_b})$ where

$$\gamma_b \;=\; \frac{2}{2 \;+\; [d / p]} . \tag{2.15a}$$

Here, we use the subscript "b" to refer to the bias component of the error appearing in (2.12). In addition, with this choice of γ_l, $1 \leq l \leq M$, we find that h^d is $O(n^{-\gamma})$ where

$$\gamma \;=\; \frac{[d / p]}{2 \;+\; [d / p]} ,$$

and hence $\underline{h}^{-d} / n$ is $O(n^{-\gamma_v})$ where we also find that

$$\gamma_v \;=\; 1 - \frac{[d / p]}{2 \;+\; [d / p]} \;=\; \frac{2}{2 \;+\; [d / p]} . \tag{2.15b}$$

Here, the subscript "v" is used to denote the variance component of the rate appearing in (2.12).

Note for this choice of $\overline{h}$ and $\underline{h}$ the bias and variance components are the same order of magnitude. In fact, this is one way to motivate our prescription for $\overline{h}$ and $\underline{h}$. As it turns out, we would also arrive at the expression in (2.14) if we had decided to choose $\overline{h}$ and $\underline{h}$ so as to minimize the mean squared error (2.12). In this case, we define a new vector $h = (h_1, \ldots, h_M)$, and assume that each h_l is of the form $O(\, n^{-\gamma_l}\,)$. Then, we minimize (2.12) with respect to the γ_l, setting $\overline{h} = \underline{h} = h$ and substituting in the expression for $\rho^2(\,\overline{h}\,)$ in (2.13). Having found the order of magnitude of the entries of h, we can then consider minimizing the same expression, this time letting $\overline{h}$ and $\underline{h}$ increase and decrease, respectively, from h. We observe, however, that increasing $\overline{h}$ only serves to increase the bias component of the mean squared error, while decreasing $\underline{h}$ increases the variance component. In this way, we find that the minimum mean squared error obtainable by letting $\overline{h}$ and $\underline{h}$ vary in magnitude from h cannot be smaller than that attained by setting $\overline{h} \sim \underline{h} \sim h$, and hence this choice is the best possible.

1.3 ANOVA Decompositions

Notation

In Section 1.1, we defined piecewise polynomials of coordinate degree $m \in \mathbb{N}^M$ in the variable $\mathbf{x} = (x_1, \ldots, x_M)$ relative to the partition $\triangle$. With minor modifications, these definitions can be used in an obvious way to construct piecewise polynomials in any subset of the variables $x_1, \ldots, x_M$. To be more precise, let s be a nonempty subset of $\{1, \ldots, M\}$ and, for convenience let $s = \{s_1, \ldots, s_k\}$ where $s_1 < \cdots < s_k$. Then, define the subvectors

$$\mathbf{x}_s \;=\; (\, x_{s_1}, \ldots, x_{s_k}\,) \qquad \text{and} \qquad \mathbf{j}_s \;=\; (\, j_{s_1}, \ldots, j_{s_k}\,). \tag{3.1a}$$

Similarly, given a vector $z \in \mathbb{R}^M$, we define

$$z_s \;=\; (z_{s_1}, \ldots, z_{s_k})\,. \tag{3.1b}$$

Next, we introduce the random variable $\mathbf{X}_s$ that takes values in the set $\mathcal{X}_s$, where

$$\mathbf{X}_s \;=\; (X_{s_1}, \ldots, X_{s_k}) \qquad \text{and} \qquad \mathcal{X}_s \;=\; \mathcal{X}_{s_1} \times \cdots \times \mathcal{X}_{s_k}\,. \tag{3.2}$$

As in Section 1.1, the set $\mathcal{X}_s$ is partitioned by the sets in $\triangle_s = \triangle_{s_1} \times \cdots \times \triangle_{s_k}$. Observe that for each $\delta \in \triangle_s$, the function Diamδ now returns a k-dimensional

vector. Similarly $\mathrm{Dim}\,\mathcal{X}_s$ is k-dimensional, and by using the definition in (3.1b) we find that $\mathrm{Dim}\,\mathcal{X}_s = (\mathrm{Dim}\,\mathcal{X})_s$. If $s = \varnothing$, then $\mathbf{X}_s$ and the quantities in (3.1) are taken to be the constant 1, while $\mathcal{X}_s$ and $\triangle_s$ are each taken to be the empty set.

Fix a set $s \subset \{1, \ldots, M\}$ and a vector $m \in \mathbb{N}^M$. Applying the definitions in (1.6)–(1.9) directly to the vectors in (3.1), we find that a polynomial of coordinate degree at most m_s in the variable $\mathbf{x}_s$ is any function of the form

$$p(\mathbf{x}_s) \;=\; \sum_{\mathbf{j}_s \le m_s} b_{\mathbf{j}_s}\, \mathbf{x}_s^{\mathbf{j}_s}\,, \tag{3.3}$$

where the coefficients $b_{\mathbf{j}_s} \in \mathbb{R}$ for $\mathbf{j}_s \le m_s$. If $s = \varnothing$, then the resulting polynomial p is just the constant 1. Now, assume that s is nonempty and for each $\delta \in \triangle_s$ apply the recipe given in (1.21a)–(1.21c) using the polynomials in (3.3). In this way, we construct the space of piecewise polynomials of coordinate degree m_s relative to the partition $\triangle_s$, which we denote by $\mathbb{PP}_s$. If $s = \{l\}$ for some $1 \le l \le M$, then we write $\mathbb{PP}_l$ for the resulting space of piecewise polynomials, and if $s = \varnothing$, then $\mathbb{PP}_\varnothing$ is the space of constant functions on $\mathcal{X}$.

Unsaturated Spaces of Piecewise Polynomials

In Section 1.1, we established conditions on the rate at which we can refine the partitions $\triangle_l$, $1 \le l \le M$, to guarantee that with probability approaching 1, the space $\mathbb{PP}$ is identifiable. In that section, the symbol $\mathbb{PP}$ represented the space of polynomials of coordinate degree m in the variable $\mathbf{x} = (x_1, \ldots, x_M)$. At this point, we adopt a more general view and reserve the symbol $\mathbb{PP}$ for spaces of piecewise polynomial functions that depend on all M variables, or at very least, spaces that are not entirely contained in $\mathbb{PP}_s$ for some proper subset $s \subset \{1, \ldots, M\}$. To make this precise, let $\mathcal{S}$ denote a collection of subsets of $\{1, \ldots, M\}$ and define

$$\mathbb{PP} \;=\; \left\{ \sum_{s \in \mathcal{S}} p_s \,:\, p_s \in \mathbb{PP}_s \;\text{ for }\; s \in \mathcal{S} \right\}. \tag{3.4a}$$

Note that if $\{1, \ldots, M\}$ is an element of $\mathcal{S}$, the space in (3.4a) is exactly $\mathbb{PP}$ as defined in Section 1.1. In proving the results of this section, we do not allow arbitrary collections $\mathcal{S}$ to be used in defining $\mathbb{PP}$, but instead require that $\mathcal{S}$ be hierarchical. That is, if $s \in \mathcal{S}$ and $r \subset s$, then $r \in \mathcal{S}$ as well.

Our work in Section 1.2 indicates that the piecewise character of the functions $p \in \mathbb{PP}_s$, $s \subset \{1, \ldots, M\}$, makes them attractive from a theoretical point of view. However, it is precisely because of this inherently piecewise character that these discontinuous functions are not often used for developing methodology. Therefore,

in this section we analyze subspaces of smooth functions in $\mathbb{PP}$, also referred to as spline spaces. Toward this end, for each $1 \le l \le M$, let $\mathbb{G}_l$ denote a non-empty subspace of $\mathbb{PP}_l$ that contains the constant functions, and let $g_{l1}, \ldots, g_{lN_l}$ denote a basis of $\mathbb{G}_l$. Then, given any nonempty subset $s \in \mathcal{S}$, where $s = \{s_1, \ldots, s_k\}$ and $s_1 < \cdots < s_k$, let $\mathbb{G}_s$ denote the tensor product space

$$\mathbb{G}_{s_1} \otimes \cdots \otimes \mathbb{G}_{s_k} \;=\; \text{span}\left\{ g_{s_1 j_1} \cdots g_{s_k j_k} \;:\; 1 \le j_l \le N_l \;\text{ for }\; l \in s \right\}.$$

Recall that the functions appearing on the right form a basis for this space, and hence the dimension of $\mathbb{G}_s$ is the product of the dimensions of $\mathbb{G}_l$ for $l \in s$. If $s = \varnothing$, then we take $\mathbb{G}_s$ to be the space of constant functions on $\mathcal{X}$. Therefore, by construction, the spaces $\mathbb{G}_s$ inherit a hierarchical structure from the collection $\mathcal{S}$. From this nested set of spaces, we define

$$\mathbb{G} \;=\; \left\{ \sum_{s \in \mathcal{S}} g_s \;:\; g_s \in \mathbb{G}_s \;\text{ for }\; s \in \mathcal{S} \right\}. \tag{3.4b}$$

Our immediate goal is to extend the identifiability results from Section 1.1 to the spaces $\mathbb{PP}$ and $\mathbb{G}$.

Unfortunately, knowing that either of these spaces is identifiable does not prevent functions in these spaces from having a number of equivalent expressions as sums of functions in $\mathbb{PP}_s$ or $\mathbb{G}_s$ for $s \in \mathcal{S}$. To account for this overspecification in the definitions (3.4), we want to take a moment and introduce the notion of an ANOVA decomposition of the space $\mathbb{G}$ (the development for $\mathbb{PP}$ is identical). For $s \in \mathcal{S}$, let $\mathbb{G}_s^0$ denote the orthogonal complement of $\mathbb{G}_s$ relative to the sum of $\mathbb{G}_r$ as r ranges over proper subsets of s relative to the empirical inner product $\langle \cdot, \cdot \rangle_n$. We refer to the spaces $\mathbb{G}_s^0$, $s \in \mathcal{S}$, as the components of $\mathbb{G}$. Here, $\mathbb{G}_s^0$ for $\#(s) = 1$ are also called main effect components, while $\mathbb{G}_s^0$ for $\#(s) \ge 2$ are also called interaction components. In the next lemma we demonstrate that if the space $\mathbb{G}$ is identifiable, then each function $g \in \mathbb{G}$ can be written uniquely as a sum of the components $g_s \in \mathbb{G}_s^0$, which we refer to as the ANOVA decomposition of g.

Lemma 3.1 *Let $g = \sum_s g_s$ where $g_s \in \mathbb{G}_s^0$ for $s \in \mathcal{S}$. If $\mathbb{G}$ is identifiable and $g = 0$, then each $g_s = 0$ for $s \in \mathcal{S}$.*

Proof Suppose for the moment that $\mathbf{X}$ is uniformly distributed on $\mathcal{X}$ and hence that for all functions $f_1(\cdot)$, $f_2(\cdot)$ defined on $\mathcal{X}$,

$$\langle f_1, f_2 \rangle \;=\; E\big[f_1(\mathbf{X}) f_2(\mathbf{X}) \big] \;=\; \int_{\mathcal{X}} f_1(\mathbf{x}) f_2(\mathbf{x})\, d\mathbf{x}.$$

For each index l, $1 \leq l \leq M$, let $\mathbb{G}_l^1$ denote the space of all functions of $\mathbb{G}_l$ that are orthogonal to the constant function 1 relative to this inner product. For each nonempty set $s = \{s_1, \ldots, s_k\} \in \mathcal{S}$, set

$$\mathbb{G}_s^1 \;=\; \mathbb{G}_{s_1}^1 \otimes \cdots \otimes \mathbb{G}_{s_k}^1$$

where we assume that $s_1 < \cdots < s_k$. If $s = \varnothing$, we again take $\mathbb{G}_\varnothing$ to be the space of constant functions on $\mathcal{X}$. Observe that the spaces $\mathbb{G}_s^1$, $s \in \mathcal{S}$, are orthogonal to each other. To see this, choose any two sets $r, s \in \mathcal{S}$, and consider functions of the form

$$g_r \;=\; g_{r\cap s}\, g_{r\cap s^c} \quad \text{and} \qquad g_s \;=\; g_{s\cap r}\, g_{s\cap r^c}\,,$$

where

$$g_{r\cap s},\, g_{s\cap r} \,\in\, \mathbb{G}_{r\cap s}^1\,, \quad g_{r\cap s^c} \,\in\, \mathbb{G}_{r\cap s^c}^1\,, \quad \text{and} \quad g_{s\cap r^c} \,\in\, \mathbb{G}_{s\cap r^c}^1\,.$$

Then,

$$\begin{aligned} \langle\, g_r, g_s \,\rangle \;&=\; \int_{\mathcal{X}_{r\cap s^c}} g_{r\cap s^c} \int_{\mathcal{X}_{s\cap r^c}} g_{s\cap r^c} \int_{\mathcal{X}_{r\cap s}} g_{r\cap s}\, g_{s\cap r} \\ &=\; \langle\, g_{r\cap s^c}, 1 \,\rangle\, \langle\, 1, g_{s\cap r^c} \,\rangle\, \langle\, g_{r\cap s}\,, g_{s\cap r} \,\rangle\,, \end{aligned}$$

where we define $\mathcal{X}_\varnothing = \varnothing$. If r and s are not equal, then at least one of the sets $r \cap s^c$, $s \cap r^c$ is non-empty, so at least one of the first two inner-products on the right equals zero. Therefore, g_r and g_s are orthogonal, and hence the spaces $\mathbb{G}_r^1$ and $\mathbb{G}_s^1$ are orthogonal.

Furthermore, we will now demonstrate that for any set $s \in \mathcal{S}$, the space $\mathbb{G}_s^1$ is the direct sum of the spaces $\mathbb{G}_r^1$, where r ranges over subsets of s. If for any linear space $\mathbb{G}$ we let the integer $\dim \mathbb{G}$ denote the dimension of the space $\mathbb{G}$, then recalling that for any nonempty set $s \in \mathcal{S}$, the dimension of $\mathbb{G}_s$ is the product of the dimensions of the spaces $\mathbb{G}_l$ for $l \in s$, we find that

$$\dim \mathbb{G}_s \;=\; \prod_{l\in s} \dim \mathbb{G}_l \;=\; \prod_{l\in s} \left(\, 1 + \dim \mathbb{G}_l \,-\, 1 \,\right) \;=\; \sum_{r\subset s} \prod_{l\in r} \left(\, \dim \mathbb{G}_l \,-\, 1 \,\right)$$

and hence that

$$\dim \mathbb{G}_s \;=\; \sum_{r\subset s} \dim \mathbb{G}_r^1\,.$$

Therefore, since the spaces $\mathbb{G}_r^1$, $r \subset s$, are orthogonal, $\mathbb{G}_s^1$ is the direct sum of the spaces $\mathbb{G}_r^1$, $r \subset s$.

With these facts in hand, we now return to the proof of our original result. Recall our assumptions that

$$g \;=\; \sum_{s \in \mathcal{S}} g_s \qquad \text{where } g_s \in \mathbb{G}_s^0 \text{ for } s \in \mathcal{S}$$

and that $\mathbb{G}$ is identifiable. We want to show that if g equals zero, then g_s also equals zero for $s \in \mathcal{S}$. A set $s \in \mathcal{S}$ is said to be maximal if it is not the proper subset of any other set in $\mathcal{S}$. Because by removing a maximal set from $\mathcal{S}$ we are left with another hierarchical collection of subsets of $\{1, \ldots, M\}$, it is sufficient to show that if s is maximal, then g_s is zero. If $g_s \in \mathbb{G}_s$, we can write

$$g_s \;=\; \sum_{r \subset s} g_{sr}, \qquad \text{where } g_{sr} \in \mathbb{G}_r^1 \subset \mathbb{G}_r \text{ for } r \subset s\,.$$

Suppose that $g = 0$ and write

$$0 \;=\; \sum_s g_s \;=\; \sum_s \sum_{r \subset s} g_{sr} \;=\; \sum_r \sum_{s \supset r} g_{sr}\,.$$

Then,

$$0 \;=\; \sum_r \left\| \sum_{s \supset r} g_{sr} \right\|^2 \qquad \text{and hence} \qquad \sum_{s \supset r} g_{sr} \;=\; 0\,,$$

for $r \in \mathcal{S}$. In particular, if s is maximal, then $g_{ss} = 0$ and $g_s = \sum^{(s)} g_{sr}$, where $\sum^{(s)}$ denotes summation over the proper subsets of s.

Finally, let s be maximal and observe that

$$\| g_s \|_n^2 \;=\; \left\langle g_s, \sum\nolimits^{(s)} g_{sr} \right\rangle_n \;=\; 0\,.$$

Therefore, since $\mathbb{G}$ is identifiable, we conclude that $g_s = 0$. □

In Lemmas 3.2 through 3.5, we establish under what conditions the space $\mathbb{PP}$ and hence the space $\mathbb{G}$ are identifiable. In the process, we will obtain a number of extremely useful results about the relationship between the empirical and theoretical norms on these spaces. But first, we introduce a condition on the rate at which the individual partitions $\triangle_l$, $1 \le l \le M$, are refined.

Condition 3′ *Let $\mathcal{S}$ be a hierarchical collection of subsets of $\{1, \ldots, M\}$. Assume that the partitions $\triangle_l$, $1 \le l \le M$, are refined with sample size so that*

$$\overline{h}_s \;\to\; 0 \quad \textit{for all } s \in \mathcal{S}\,,$$

while

$$\log \underline{h}_r^{dr} \;+\; n\, \underline{h}_r^{dr} \;\to\; \infty \quad \textit{as } n \to \infty,$$

for all sets $r = r_1 \cup r_2$, where $r_1, r_2 \in \mathcal{S}$.

From Condition 3′ and the repeated application of Lemma 1.4 to sets of the form $r \cup s$, where $r, s \in \mathcal{S}$, we have the following result.

Lemma 3.2 *Assume that Conditions* 1, 2, *and* 3′ *hold. Then, for* $t > 0$ *except on a set whose probability tends to zero with* n,

$$\left| E_n - E \right| \left[g^2(\mathbf{X}) \right] \;\leq\; t \sum_{s \in \mathcal{S}} E\left[g_s^2(\mathbf{X}_s) \right],$$

for all $g = \sum_s g_s$, *where* $g_s \in \mathbb{G}_s$ *for* $s \in \mathcal{S}$.

In the following lemma, we derive a powerful tool for analyzing ANOVA decompositions. Roughly speaking, it says that the components in an ANOVA decomposition are not too confounded relative to the theoretical norm defined in (1.24b). In what follows, we define $\epsilon_1 = 1 - \left(1 - M_3^{-3}\right)^{1/2}$.

Lemma 3.3 *Suppose Conditions* 1, 2 *and* 3′ *hold and let* $0 < \epsilon_2 < \epsilon_1$. *Then, except on an event whose probability tends to zero with* n,

$$\| g \|^2 \;\geq\; \epsilon_2^{\#(\mathcal{S})-1} \sum_{s \in \mathcal{S}} \| g_s \|^2 \tag{3.5}$$

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

Proof We will verify (3.5) by induction on $\#(\mathcal{S})$. Observe that it is trivially true when $\#(\mathcal{S}) = 1$. Suppose $\#(\mathcal{S}) \geq 2$ and that (3.5) holds whenever $\mathcal{S}$ is replaced by $\mathcal{S}'$ with $\#(\mathcal{S}') < \#(\mathcal{S})$. Choose a maximal $r_1 \in \mathcal{S}$ and fix $0 < t < M_3^{-3}$. Then, we will show that, except on an event whose probability tends to zero with n,

$$\| g \|^2 \;\geq\; M_3^{-3} \| g_{r_1} \|^2 - t \sum_s \| g_s \|^2 . \tag{3.6}$$

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

In order to verify (3.6), we first assume that $r_1 = \{1, \ldots, M\}$. Then, by the definition of $\mathbb{G}_{r_1}^0$,

$$\| g \|_n^2 \;\geq\; \| g_{r_1} \|_n^2. \tag{3.7}$$

Applying Lemma 3.2, we know that except on an event whose probability tends to zero with n,

$$\begin{aligned} \| g \|^2 \;&\geq\; \| g^2 \|_n^2 - \frac{t}{2} \sum_s \| g_s \|^2 \\ &\geq\; \| g_{r_1} \|_n^2 - \frac{t}{2} \sum_s \| g_s \|^2 \end{aligned}$$

$$\geq \left(1 - \frac{t}{2}\right) \| g_{r_1} \|^2 - \frac{t}{2} \sum_s \| g_s \|^2$$

$$\geq \| g_{r_1} \|^2 - t \sum_s \| g_s \|^2 . \tag{3.8}$$

Suppose, instead, that $1 \leq \#(r_1) \leq M-1$ and let r_2 denote the complement of r_1. By a slight abuse of notation, temporarily set $\mathbf{x} = (\mathbf{x}_{r_1}, \mathbf{x}_{r_2})$ and $\mathbf{X} = (\mathbf{X}_{r_1}, \mathbf{X}_{r_2})$. Then, let f, $f_{\mathbf{X}_{r_1}}$, and $f_{\mathbf{X}_{r_2}}$ denote the density functions of $\mathbf{X}$, $\mathbf{X}_{r_1}$, and $\mathbf{X}_{r_2}$, respectively. By Condition 2, $f_{\mathbf{X}_{r_1}}$ and $f_{\mathbf{X}_{r_2}}$ are both bounded above by M_3, and hence

$$f(\mathbf{x}_{r_1}, \mathbf{x}_{r_2}) \geq M_3^{-3} f_{\mathbf{X}_{r_1}}(\mathbf{x}_{r_1}) f_{\mathbf{X}_{r_2}}(\mathbf{x}_{r_2}), \qquad \mathbf{x}_{r_1} \in \mathcal{X}_{r_1}, \mathbf{x}_{r_2} \in \mathcal{X}_{r_2}. \tag{3.9}$$

Therefore,

$$\| g \|^2 \geq M_3^{-3} \int_{\mathcal{X}_{r_2}} \left\| g(\mathbf{X}_{r_1}, \mathbf{x}_{r_2}) \right\|^2 f_{\mathbf{X}_{r_2}}(\mathbf{x}_{r_2}) \, d\mathbf{x}_{r_2}.$$

From the tensor product structure of the spaces $\mathbb{G}_s$, we know that for a fixed $\mathbf{x}_{r_2} \in \mathcal{X}_{r_2}$, the functions $g_s(\mathbf{x}_{r_1}, \mathbf{x}_{r_2})$ are in $\mathbb{G}_{s \cap r_1}$. However, since r_1 is maximal, $s \cap r_1$ is strictly contained in r_1 for all $s \in \mathcal{S}$ with $s \neq r_1$. Therefore, by the definition of $\mathbb{G}^0_{r_1}$,

$$\left\| g(\mathbf{X}_{r_1}, \mathbf{x}_{r_2}) \right\|_n^2 \geq \| g_{r_1} \|_n^2, \qquad \mathbf{x}_{r_2} \in \mathcal{X}_{r_2} .$$

Repeating the argument leading from (3.7) to (3.8), we find that, except on an event whose probability tends to zero with n,

$$\left\| g(\mathbf{X}_{r_1}, \mathbf{x}_{r_2}) \right\|^2 \geq \| g_{r_1} \|^2 - t \left(\| g_{r_1} \|^2 + \sum_{s \neq r_1} \left\| g_s(\mathbf{X}_{r_1}, \mathbf{x}_{r_2}) \right\|^2 \right)$$

for $\mathbf{x}_{r_2} \in \mathcal{X}_{r_2}$, and hence that $\| g \|^2$ is bounded below by

$$M_3^{-3} \| g_{r_1} \|^2 - M_3^{-3} t \left(\| g_{r_1} \|^2 + \int_{\mathcal{X}_{r_2}} \sum_{s \neq r_1} \left\| g_s(\mathbf{X}_{r_1}, \mathbf{x}_{r2}) \right\|^2 f_{\mathbf{X}_{r_2}}(\mathbf{x}_{r_2}) \, d\mathbf{x}_{r_2} \right).$$

From Condition 2, we obtain (3.6).

Writing r for r_1, it follows from (3.6) that, except on an event whose probability tends to zero with n,

$$\left\| g_r - \beta (g - g_r) \right\|^2 \geq \left(M_3^{-3} - t \right) \| g_r \|^2 - \beta^2 t \sum_s \| g_s \|^2$$

for $\beta \in \mathbb{R}$ and all $g = \sum_s g_s$, where $g_s \in \mathbb{G}^0_s$ for $s \in \mathcal{S}$. Expanding this expression in powers of β, we find that

$$\beta^2 \left(\| g - g_r \|^2 + t \sum_s \| g_s \|^2 \right) - 2 \beta \langle g_r, g - g_r \rangle + \epsilon^2(t) \| g_r \|^2 \geq 0, \tag{3.10}$$

where we have defined the function

$$\epsilon(t) \;=\; \sqrt{1 - M_3^{-3} + t}\;.$$

Setting $\beta = \pm\,\epsilon(t)$ in (3.10), we find that

$$\begin{aligned} 2\left|\,\langle\, g_r, g - g_r \,\rangle\,\right| &\leq \epsilon(t)\left(\,\|\,g - g_r\,\|^2 + \|\,g_r\,\|^2 + t\sum_s \|\,g_s\,\|^2\,\right) \\ &\leq \epsilon(t)\left(\,\|\,g - g_r\,\|^2 + \|\,g_r\,\|^2\,\right) + t\sum_s \|\,g_s\,\|^2 \end{aligned}$$

since $\epsilon(t) < 1$. Consequently, by the induction hypothesis, except on an event whose probability tends to zero with n,

$$\begin{aligned} \|\,g\,\|^2 &\geq \big(1 - \epsilon(t)\big)\left(\,\|\,g_r\,\|^2 + \|g - g_r\,\|^2\,\right) - t\sum_s \|\,g_s\,\|^2 \\ &\geq \epsilon_2\left(\,\|\,g_r\,\|^2 + \epsilon_2^{\#(\mathcal{S})-2}\sum_{s\neq r}\|\,g_s\,\|^2\,\right) - t\sum_s \|\,g_s\,\|^2 \\ &\geq \left(\epsilon_2^{\#(\mathcal{S})-1} - t\right)\sum_s \|\,g_s\,\|^2 \end{aligned}$$

provided that $1 - \epsilon(t) \geq \epsilon_2$, which is true for t sufficiently small. Therefore, (3.5) holds for $\mathcal{S}$. □

The next result is an extension of Lemma 1.4 to a larger collection of functions.

Lemma 3.4 *Suppose Conditions* 1, 2 *and* 3′ *hold and let* $t > 0$. *Then, except on an event whose probability tends to zero with* n,

$$\left|\,E_n - E\,\right|\big[\,g_1(\mathbf{X})\,g_2(\mathbf{X})\,\big] \;\leq\; t\sqrt{E\big[\,g_1^2(\mathbf{X})\,\big]}\,\sqrt{E\big[\,g_2^2(\mathbf{X})\,\big]}\;,$$

for all $g_1, g_2 \in \mathbb{G}$.

Proof It follows from Condition 3′ and Lemma 1.4 that, except on an even whose probability tends to zero with n,

$$\left|\,E_n - E\,\right|\big[\,g_{1r}(\mathbf{X})\,g_{2s}(\mathbf{X})\,\big] \;\leq\; \frac{t}{\#(\mathcal{S})}\sqrt{E\big[\,g_{1r}^2(\mathbf{X})\,\big]}\,\sqrt{E\big[\,g_{2s}^2(\mathbf{X})\,\big]} \qquad (3.11)$$

for $r, s \in \mathcal{S}$, $g_{1r} \in \mathbb{G}_r^0$ and $g_{2s} \in \mathbb{G}_s^0$. If (3.11) holds, then

$$\left|\,E_n - E\,\right|\big[\,g_1(\mathbf{X})\,g_2(\mathbf{X})\,\big] \;\leq\; t\sqrt{\sum_r E\big[\,g_{1r}^2(\mathbf{X})\,\big]}\,\sqrt{\sum_s E\big[\,g_{2s}^2(\mathbf{X})\,\big]}\,,$$

where $g_1 = \sum_r g_{1r}$ and $g_2 = \sum_s g_{2s}$ are in $\mathbb{G}$. The desired result now follows from Lemma 3.3. □

In the derivation of Lemmas 3.1 through 3.4, the subspaces $\mathbb{G}_l$, $1 \le l \le M$, were arbitrary. Therefore, if we take $\mathbb{G}_l = \mathbb{PP}_l$ for $1 \le l \le M$, then we obtain an identifiability result for $\mathbb{PP}$ similar to that stated in Lemma 1.5 for the individual spaces $\mathbb{PP}_s$, $s \subset \{1, \ldots, M\}$.

Lemma 3.5 *Suppose Conditions* 1, 2 *and* 3′ *hold. Then, except on an event whose probability tends to zero with* n, $\mathbb{PP}$ *is identifiable.*

Proof From Condition 3′ and Lemma 3.4 applied to the spaces $\mathbb{G}_l = \mathbb{PP}_l$, $1 \le l \le M$, we get that, except on an event whose probability tends to zero with n,

$$\tfrac{1}{2} \| p \|^2 \;\le\; \| p \|_n^2 \;\le\; 2 \| p \|^2, \qquad p \in \mathbb{PP}. \tag{3.12}$$

Suppose we can find a function $p \in \mathbb{PP}$ such that $p(\mathbf{X}_i)$ is zero for $1 \le i \le n$. Then, the norm $\| p \|_n$ must also equal zero, and if (3.12) holds then the norm $\| p \|$ is also zero. But, by Condition 2, this implies that p is identically zero. Thus, if (3.12) holds, then the space $\mathbb{PP}$ is identifiable. □

In Lemma 3.3, we found that the components in an ANOVA decomposition are not too confounded relative to the theoretical norm. We now extend this result to the empirical norm.

Lemma 3.6 *Suppose Conditions* 1, 2 *and* 3′ *hold, and let* $0 < \epsilon_2 < \epsilon_1$. *Then, except on an event whose probability tends to zero with* n,

$$\| g \|_n^2 \;\ge\; \epsilon_2^{\#(\mathcal{S})-1} \sum_{s \in \mathcal{S}} \| g_s \|_n^2$$

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

Proof It follows from Lemma 3.4 that for $t > 0$, except on an event whose probability tends to zero with n,

$$\| g_s \|_n^2 \;\le\; (1+t) \| g_s \|^2,$$

and hence

$$\sum_s \| g_s \|_n^2 \;\le\; (1+t) \sum_s \| g_s \|^2. \tag{3.13}$$

Choose $\epsilon_3 \in (\epsilon_2, \epsilon_1)$. It follows from (3.13) and Lemmas 3.2 and 3.3 that, except on an event whose probability tends to zero with n,

$$\begin{aligned} \| g \|_n^2 &\ge \| g \|^2 - t \sum_s \| g_s \|^2 \\ &\ge \left(\epsilon_3^{\#(\mathcal{S})-1} - t \right) \sum_s \| g_s \|^2 \end{aligned}$$

$$\geq \frac{\epsilon_3^{\#(\mathcal{S})-1} - t}{1+t} \sum_s \| g_s \|_n^2 .$$

Since t can be made arbitrarily small, the desired result holds. □

ANOVA Decompositions of Square Integrable Functions

In Section 1.2 we considered estimating μ by the the orthogonal projection of the function $Y(\,\cdot\,)$ onto a saturated space of piecewise polynomials relative to the empirical inner product defined in (1.24a). In that case, we found that as our sample size n increased, $\hat{\mu}$ tended to the regression function μ. In the next section we will consider estimating μ by the orthogonal projection of $Y(\,\cdot\,)$ onto the unsaturated spaces $\mathbb{PP}$ and $\mathbb{G}$. The natural limit for our estimate as n increases is now somewhat more complicated and deserves special attention.

Toward this end, for each $s \in \mathcal{S}$, let $\mathbb{F}_s$ denote the space of square integrable functions on $\mathcal{X}$ that depend only on the variables x_l, $l \in s$, and set

$$\mathbb{F} = \left\{ \sum_{s \in \mathcal{S}} f_s : f_s \in \mathbb{F}_s \text{ for } s \in \mathcal{S} \right\}. \tag{3.14}$$

Let $\mathbb{F}_s^0$ denote the space of all functions in $\mathbb{F}_s$ that are orthogonal to each function in $\mathbb{F}_r$, $r \subset s$, relative to the theoretical inner product. In Lemma 3.7 we show that each function $f \in \mathbb{F}$ can be written in a unique manner as

$$\sum_{s \in \mathcal{S}} f_s , \qquad \text{where } f_s \in \mathbb{F}_s^0 \text{ for } s \in \mathcal{S} ,$$

which we refer to as the ANOVA decomposition of f. To avoid confusion, for the remainder of this chapter, we will let $f_{\mathbf{X}}$ denote the density function of the random vector $\mathbf{X}$. Recall from the previous section that $\epsilon_1 = 1 - \left(1 - M_3^{-3}\right)^{1/2}$.

Lemma 3.7 *Suppose Condition 1 holds. Then*

$$\| f \|^2 \geq \epsilon_1^{\#(\mathcal{S})-1} \sum_{s \in \mathcal{S}} \| f_s \|^2 , \tag{3.15}$$

for all $f = \sum_s f_s$, where $f_s \in \mathbb{F}_s^0$ for $s \in \mathcal{S}$.

Proof As in the proof of Lemma 3.3, we will verify (3.15) by induction on $\#(\mathcal{S})$. Observe that it is trivially true when $\#(\mathcal{S}) = 1$. Suppose $\#(\mathcal{S}) \geq 2$ and that (3.15) holds whenever $\mathcal{S}$ is replaced by $\mathcal{S}'$ with $\#(\mathcal{S}') < \#(\mathcal{S})$. Choose a maximal $r_1 \in \mathcal{S}$ and fix $0 < t < M_3^{-3}$. Then, we will show that

$$\| f \|^2 \geq M_3^{-3} \| f_{r_1} \|^2 \tag{3.16}$$

for all $f = \sum_s f_s$, where $f_s \in \mathbb{F}_s^0$ for $s \in \mathcal{S}$.

If $r_1 = \{1, \ldots, M\}$, then (3.16) follows directly from the definition of $\mathbb{F}_{r_1}^0$, Suppose instead that $1 \leq \#(r_1) \leq M-1$, and let r_2 denote the complement of r_1. It follows from (3.9) that

$$\| f \|^2 \;\geq\; M_3^{-3} \int_{\mathcal{X}_{r_2}} \left\| f(\mathbf{X}_{r_1}, \mathbf{x}_{r_2}) \right\|^2 f_{\mathbf{X}_{r_2}}(\mathbf{x}_{r_2})\, d\mathbf{x}_{r_2}.$$

For each $s \in \mathcal{S}$, observe that for a fixed $\mathbf{x}_{r_2} \in \mathcal{X}_{r_2}$, the function $f_s(\mathbf{x}_{r_1}, \mathbf{x}_{r_2})$ is in the space $\mathbb{F}_{s \cap r_1}$. However, since r_1 is maximal, $s \cap r_1$ is strictly contained in r_1 for all $s \in \mathcal{S}$ with $s \neq r_1$. Therefore, since $f = \sum_s f_s$, by the definition of $\mathbb{F}_{r_1}^0$ we see that

$$\left\| f(\mathbf{X}_{r_1}, \mathbf{x}_{r_2}) \right\|^2 \;\geq\; \| f_{r_1} \|^2, \qquad \mathbf{x}_{r_2} \in \mathcal{X}_{r_2}\,,$$

once again establishing (3.16).

Writing r for r_1, it follows from (3.16) that

$$\left\| f_r - \beta\,(f - f_r\,) \right\|^2 \;\geq\; M_3^{-3}\, \| f_r \|^2\,,$$

for $\beta \in \mathbb{R}$ and all $f = \sum_s f_s$, where $f_s \in \mathbb{F}_s^0$ for $s \in \mathcal{S}$. Expanding this expression in powers of β, we find that

$$\beta^2 \| f - f_r \|^2 \,+\, 2\,\beta \left\langle f_r, f - f_r \right\rangle \,+\, \left(1 - M_3^{-3}\right) \| f_r \|^2 \;\geq\; 0\,. \tag{3.17}$$

Choosing that value of β to minimize the expression in (3.17), we obtain

$$\left\langle f_r, f - f_r \right\rangle^2 \;\leq\; \left(1 - M_3^{-3}\right) \| f - f_r \|^2 \, \| f_r \|^2\,.$$

Consequently, by the induction hypothesis,

$$\begin{aligned}
\| f \|^2 \;&\geq\; \epsilon_1 \left(\| f_r \|^2 + \| f - f_r \|^2 \right) \\
&\geq\; \epsilon_1 \left(\| f_r \|^2 \,+\, \epsilon_1^{\#(\mathcal{S})-2} \sum_{s \neq r} \| f_s \|^2 \right) \\
&\geq\; \epsilon_1^{\#(\mathcal{S})-1} \sum_s \| f_s \|^2\,,
\end{aligned}$$

and hence (3.15) holds for $\mathcal{S}$. □

Our goal was to specify a natural limit for regression estimates based on the unsaturated spaces $\mathbb{PP}$ or $\mathbb{G}$. In the following lemma, we characterize this limit as the best approximation to the regression function μ by functions in $\mathbb{F}$. In the next section we derive the rate at which our estimates converge to this limit in probability.

Lemma 3.8 *Suppose Condition* 1 *holds. Then, there is an essentially unique function* $\mu^* \in \mathbb{F}$ *satisfying*

$$\| \mu^* - \mu \|^2 \;=\; \min_{f \in \mathbb{F}} \| f - \mu \|^2 .$$

Proof Since each of the spaces $\mathbb{F}_s$, $s \in \mathcal{S}$, is complete, it follows from Lemma 3.7 that $\mathbb{F}$ is complete. Choose $f_n \in \mathbb{F}$ such that

$$\| f_n - \mu \|^2 \;\to\; \inf_{f \in \mathbb{F}} \| f - \mu \|^2 \qquad \text{as} \quad n \to \infty .$$

Then, setting $f_1 = f_m - \mu$ and $f_2 = f_n - \mu$ into the parallelogram law

$$\left\| f_1 - f_2 \right\|^2 + \left\| f_1 + f_2 \right\|^2 \;=\; 2 \left\| f_1 \right\|^2 + 2 \left\| f_2 \right\|^2 ,$$

we get that

$$\left\| f_n - f_m \right\|^2 \;=\; 2 \left\| f_m - \mu \right\|^2 + 2 \left\| f_n - \mu \right\|^2 - 4 \left\| (f_m + f_n) / 2 - \mu \right\|^2 \to 0$$

as $m, n \to \infty$. Therefore, by the completeness the space $\mathbb{F}$, there exists a function $\mu^* \in \mathbb{F}$ such that

$$\left\| f_n - \mu^* \right\|^2 \to 0 \quad \text{as} \quad n \to \infty .$$

Since

$$\left\| f_n - \mu \right\|^2 \;\to\; \left\| \mu^* - \mu \right\|^2 \quad \text{as} \quad n \to \infty ,$$

it is clear that

$$\left\| \mu^* - \mu \right\|^2 \;=\; \min_{f \in \mathbb{F}} \left\| f - \mu \right\|^2 .$$

Suppose also that $\tilde{\mu}^* \in \mathbb{F}$ and that

$$\left\| \tilde{\mu}^* - \mu \right\|^2 \;=\; \min_{f \in \mathbb{F}} \left\| f - \mu \right\|^2 .$$

Then, by another application of the parallelogram law,

$$\left\| \tilde{\mu}^* - \mu^* \right\|^2 \;=\; 4 \left\| \mu^* - \mu \right\|^2 - 4 \left\| (\tilde{\mu}^* + \mu^*) / 2 - \mu \right\|^2 \;\leq\; 0 ,$$

so

$$\left\| \tilde{\mu}^* - \mu^* \right\|^2 \;=\; 0 ,$$

and hence $\tilde{\mu}^* = \mu^*$ almost everywhere. □

1.4 Rate of Convergence in Unsaturated Spaces

Let $\mathcal{S}$ be fixed, and select subspaces $\mathbb{G}_l \subset \mathbb{PP}_l$ for $1 \leq l \leq M$. Let $\mathbb{G}$ be defined as in (3.4b), and let $\hat{\mu}$ denote the orthogonal projection of $Y(\cdot)$ onto $\mathbb{G}$ relative to the empirical inner product. If $\mathbb{G}$ is identifiable, then by Lemma 3.2, $\hat{\mu}$ can be written uniquely as

$$\hat{\mu} \;=\; \sum_{s \in \mathcal{S}} \hat{\mu}_s, \qquad \text{where } \hat{\mu}_s \in \mathbb{G}_s^0 \text{ for } s \in \mathcal{S}. \tag{4.1a}$$

Similarly, from Lemma 3.8, we know that there is a function $\mu^* \in \mathbb{F}$ that is the best theoretical approximation to μ by functions in $\mathbb{F}$. From Lemma 3.7, μ^* can be written uniquely as

$$\mu^* \;=\; \sum_{s \in \mathcal{S}} \mu_s^*, \qquad \text{where } \mu_s^* \in \mathbb{F}_s^0 \text{ for } s \in \mathcal{S}. \tag{4.1b}$$

We now take up the task of deriving the L_2 rate of convergence of $\hat{\mu}$ to μ^* and $\hat{\mu}_s$ to μ_s^* for $s \in \mathcal{S}$. For clarity, we follow the argument in Section 1.2 quite closely. In the end, the rate quoted in that section depended on the minimum and maximum diameters of sets in the partition $\triangle$. Not surprisingly, by combining various subspaces of $\mathbb{PP}_s$ to form an ANOVA decomposition, we find that the resulting rate depends on combinations of the minimum and maximum diameters of the sets in $\triangle_s$, $s \in \mathcal{S}$. Therefore, by recalling (3.1b) and the definitions

$$\overline{h} = \max\{\,\mathrm{Diam}\,\delta : \delta \in \triangle\,\}, \quad \underline{h} = \min\{\,\mathrm{Diam}\,\delta : \delta \in \triangle\,\} \quad \text{and} \quad d = \mathrm{Dim}\,\mathcal{X},$$

our rate will depend on the quantities $\overline{h}_s$, $\underline{h}_s$ and d_s, for $s \in \mathcal{S}$. Once again making use of the variance-bias decomposition introduced in Section 1.2, we first bound the variance component of the mean squared error.

Variance

Recall from Section 1.2 that we bounded the variance component of the mean squared error through the use of two special representors defined in (2.1) and (2.2). Clearly, these representors can be constructed using any subset of the variables $x_1, \ldots, x_M$ and the definitions in (3.1). For convenience, given any set $s \in \mathcal{S}$, we let $q_{\mathbf{j}_s}$ and $q_{\mathbf{x}'_s}$ denote the resulting functions as defined in (2.1) and (2.2) relative to the subvectors $\mathbf{x}_s$ and $\mathbf{j}_s$. We reserve the notation $q_{\mathbf{x}'}$ for the analagous function in $\mathbb{G}$. That is, for $\mathbf{x}' \in \mathcal{X}$, we let $q_{\mathbf{x}'} \in \mathbb{G}$ denote the representor of the linear

functional $g \mapsto g(\mathbf{x}')$ on $\mathbb{G}$ relative to the inner product $\langle \cdot, \cdot \rangle_n$, so that

$$g(\mathbf{x}') \;=\; \langle q_{\mathbf{x}'}, g \rangle_n , \qquad g \in \mathbb{G}.$$

The function $q_{\mathbf{x}'}$ inherits many of its properties from the functions $q_{\mathbf{x}'_s}$, $s \in \mathcal{S}$, defined in (2.1). We now record a number of facts about $q_{\mathbf{x}'}$ in the following lemma and its proof.

Lemma 4.1 *Suppose Conditions* 1, 2 *and* 3′ *hold. Then there exists a positive constant* M_{11} *independent of* n *and* $\overline{h}_s$ *such that, except on a set whose probability tends to zero with* n,

$$q_{\mathbf{x}'}(\mathbf{x}') \;=\; \| q_{\mathbf{x}'} \|_n^2 \;\le\; M_{11} \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} , \qquad \mathbf{x}' \in \mathcal{X}. \tag{4.2}$$

Proof Assume that $\mathbb{PP}$ and hence $\mathbb{G}$ are identifiable. Then, we can write $\mathbb{G}$ uniquely as

$$q_{\mathbf{x}'} \;=\; \sum_{s \in \mathcal{S}} q_s , \qquad \text{where } q_s \in \mathbb{G}_s^0 \text{ for } s \in \mathcal{S}.$$

Now, $\mathbb{G}_s^0$ is a subspace of $\mathbb{PP}_s$ for $s \in \mathcal{S}$, so

$$\| q_{\mathbf{x}'} \|_n^2 \;=\; \langle q_{\mathbf{x}'}, q_{\mathbf{x}'} \rangle_n \;=\; \sum_{s \in \mathcal{S}} \langle q_s, q_{\mathbf{x}'} \rangle_n \;=\; \sum_{s \in \mathcal{S}} q_s(\mathbf{x}') \;=\; \sum_{s \in \mathcal{S}} \langle q_s, q_{\mathbf{x}'_s} \rangle_n ,$$

where the functions $q_{\mathbf{x}'_s}$ are defined in (2.1) using the subvectors $\mathbf{x}_x$ for $s \in \mathcal{S}$. Therefore, except on a set whose probability tends to zero with n,

$$\| q_{\mathbf{x}'} \|_n^4 \;\le\; \sum_{s \in \mathcal{S}} \| q_s \|_n^2 \sum_{s \in \mathcal{S}} \| q_{\mathbf{x}'_s} \|_n^2 .$$

However, from Lemma 3.6 we find that

$$\sum_{s \in \mathcal{S}} \| q_s \|_n^2 \;\le\; \epsilon_2^{1-\#(\mathcal{S})} \| q_{\mathbf{x}'} \|_n^2 ,$$

and hence, except on a set whose probability tends to zero with n,

$$\| q_{\mathbf{x}'} \|_n^2 \;\le\; \epsilon_2^{1-\#(\mathcal{S})} \sum_{s \in \mathcal{S}} \| q_{\mathbf{x}'_s} \|_n^2 .$$

The desired result now follows from Lemmas 2.1 and 3.5. □

Arguing as in Section 1.2, we observe that $E(\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n)$ is the orthogonal projection of μ onto $\mathbb{G}$. More is true, however. Observe that if $\mathbb{G}$ is identifiable, then from Lemma 3.1 we have that $\hat{\mu}_s(\mathbf{x}_s)$ is a linear functional of $\hat{\mu}$, the orthogonal projection of $Y(\cdot)$ onto $\mathbb{G}$, and hence $E(\hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n)$ is finite. Therefore, using (4.1a) we find that if $\mathbb{G}$ is identifiable we can write

$$E(\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) \;=\; \sum_{s \in \mathcal{S}} E(\hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) .$$

Now, for any set $s \in \mathcal{S}$, let r be any proper subset of a set s and let $g_r \in \mathbb{G}_r$. Then

$$\left\langle E(\,\hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\,)\,, g_r \right\rangle_n \;=\; E\big(\,\langle\,\hat{\mu}_s, g_r\,\rangle_n \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\,\big) \;=\; 0\,,$$

and hence $E\big(\,\hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\,\big)$ is in $\mathbb{G}_s^0$. This fact is important in the proof of the following lemma.

Lemma 4.2 *Suppose Conditions* 1, 2, 3′, *and* 4 *hold. Then for* $s \in \mathcal{S}$,

$$\left\| \,\hat{\mu}_s - E(\,\hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\,) \,\right\|^2 \;=\; O_P\!\left(\sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} \,/\, n \right), \tag{4.3a}$$

and

$$\left\| \,\hat{\mu} - E(\,\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\,) \,\right\|^2 \;=\; O_P\!\left(\sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} \,/\, n \right). \tag{4.3b}$$

Proof Assume that $\mathbb{PP}$ and hence $\mathbb{G}$ are identifiable. Fix $\mathbf{x}' \in \mathcal{X}$. Arguing as in the proof of Lemma 2.2, we get that

$$\hat{\mu}(\mathbf{x}') \;=\; \langle\, q_{\mathbf{x}'}, \hat{\mu} \rangle_n \;=\; \langle\, q_{\mathbf{x}'}, Y(\,\cdot\,) \rangle_n$$

and hence by Condition 4 that

$$\mathrm{var}\left(\,\hat{\mu}(\,\mathbf{x}'\,) \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\,\right) \;=\; \frac{1}{n^2} \sum_{i=1}^{n} q_{\mathbf{x}'}^2(\,\mathbf{X}_i\,)\, \sigma^2(\,\mathbf{X}_i\,) \;\le\; \frac{1}{n}\, M_5 \left\| \,q_{\mathbf{x}'}\, \right\|_n^2 .$$

It now follows from Lemma 4.1 that, except on an event whose probability tends to zero with n,

$$\mathrm{var}\left(\,\hat{\mu}(\,\mathbf{x}'\,) \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\,\right) \;\le\; M_5\, M_{11} \sum_{s \in \mathcal{S}} \left(\,\underline{h}_s^{-d_s} \,/\, n\,\right).$$

The relation in (4.3b) follows from the discussion of the variance-bias decomposition in Section 1.2. Finally, from the discussion prior to the statement of this lemma, we know that $E\big(\,\hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\,\big)$ is in $\mathbb{G}_s^0$. Therefore, from (4.3b) and Lemma 3.3, we obtain (4.3a). □

Bias

We begin this subsection with a result that allows us to bound the components of an ANOVA decomposition with respect to the sum of quantities that are more readily obtainable. This fact allows us to translate the rates derived in Section 1.2 for saturated spaces into rates for ANOVA models. The argument has essentially appeared already in the proof of Lemma 4.2.

Lemma 4.3 *Suppose Conditions* 1, 2 *and* 3′ *hold. Given any hierarchical collection* $\mathcal{S}$ *of subsets of* $\{1,\ldots,M\}$, *let* $\mathbb{G}$ *be the corresponding space defined in* (3.4b), *and assume that* $\mathbb{G}$ *is identifiable. Take* f *to be any function defined on* $\mathcal{X}$ *and, relative to the empirical inner product, let* g *denote the orthogonal projection of* f *onto* $\mathbb{G}$ *and let* g_s^0 *denote the orthogonal projection of* f *onto* $\mathbb{G}_s^0$ *for* $s \in \mathcal{S}$. *Then,*

$$\| g \|_n^2 \;\leq\; \epsilon_2^{1-\#(\mathcal{S})} \sum_{s\in\mathcal{S}} \| g_s^0 \|_n^2 , \tag{4.4}$$

except on a set whose probability tends to zero with n.

Proof Assume that $\mathbb{G}$ is identifiable. Then, we can write g uniquely as

$$g \;=\; \sum_s g_s , \qquad \text{where } g_s \in \mathbb{G}_s^0 \text{ for } s \in \mathcal{S}.$$

Note that g_s^0 need not equal g_s, however, since the spaces $\mathbb{G}_s^0$, $s \in \mathcal{S}$, need not be orthogonal. Observe that

$$\| g \|_n^2 \;=\; \sum_{s\in\mathcal{S}} \langle\, g_s , g \,\rangle_n \;=\; \sum_{s\in\mathcal{S}} \langle\, g_s , g_s^0 \,\rangle_n \;\leq\; \sum_{s\in\mathcal{S}} \| g_s \|_n \| g_s^0 \|_n$$

and hence that

$$\| g \|_n^2 \;\leq\; \sum_{s\in\mathcal{S}} \| g_s \|_n^2 \sum_{s\in\mathcal{S}} \| g_s^0 \|_n^2 . \tag{4.5}$$

However, from Lemma 3.6 we obtain the bound

$$\sum_{s\in\mathcal{S}} \| g_s \|_n^2 \;\leq\; \epsilon_2^{1-\#(\mathcal{S})} \| g \|_n^2 ,$$

except on a set whose probability tends to zero with n, and hence the desired result now follows from (4.5). □

Lemma 4.4 *Suppose Conditions* 1, 2 *and* 3′ *hold. If* $\mu^* = 0$, *then*

$$\| E(\,\hat{\mu}_s \,|\, \mathbf{X}_1,\ldots,\mathbf{X}_n\,) \|_n^2 \;=\; O_P\Big(\sum_{s\in\mathcal{S}} \underline{h}_s^{-d_s} / n \Big) , \qquad s \in \mathcal{S} . \tag{4.6}$$

Proof Suppose $\{1,\ldots M\}$ is contained in the collection $\mathcal{S}$. Then, μ^* equals μ, and by our assumption on μ^*, μ is identically zero. Therefore, since $E(\,\hat{\mu}\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\,)$ is the orthogonal projection of μ onto $\mathbb{G}$, each of the functions $E(\,\hat{\mu}_s\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\,)$, $s \in \mathcal{S}$, are also identically zero, and hence (4.6) is trivially satisfied.

Suppose, instead that $\{1,\ldots,M\}$ is not a member of the collection $\mathcal{S}$ and assume that $\mathbb{PP}$ and hence $\mathbb{G}$ are identifiable. For $s \in \mathcal{S}$, let $\tilde{\mu}_s$ denote the orthogonal

projection of μ onto $\mathbb{PP}_s$ relative to the empirical inner product. Now, let $\tilde{\mu}_s^0$ denote the orthogonal projection of $\tilde{\mu}_s$ onto $\mathbb{G}_s^0$ relative to this inner product, which equals the orthogonal projection of μ onto $\mathbb{G}_s^0$ for $s \in \mathcal{S}$. Therefore, we have that

$$\| \tilde{\mu}_s^0 \|_n^2 \;\leq\; \| \tilde{\mu}_s \|_n^2 ,$$

and hence from Lemma 4.3 applied to the function μ, that

$$\| E(\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \| \tilde{\mu}_s \|_n^2 \Big) .$$

Consider now the sum on the righthand side of the above expression. Fix a set $s \in \mathcal{S}$, and as in the discussion prior to Lemma 2.3, write

$$\tilde{\mu}_s(\mathbf{x}_s) \;=\; \sum_{\delta \in \triangle_s} \tilde{\mu}_\delta(\mathbf{x}_s) , \tag{4.7}$$

where, following (1.21c), each function $\tilde{\mu}_\delta$ is a polynomial of coordinate degree m_s in the variable $\mathbf{x}_s$. Therefore, from Lemma 2.3 we find that except on a set whose probability tends to zero with n, for $\delta \in \triangle_s$ the functions $\tilde{\mu}_\delta$ defined in (2.7) satisfy

$$\| \tilde{\mu}_\delta \|_n^2 \;\leq\; M_7 \, \underline{h}_s^{-d_s} \sum_{\mathbf{j}_s \leq m_s} \langle p_{\mathbf{j}_s} , \mu \rangle_n^2 , \tag{4.8}$$

where the constant M_7 is given explicitly in (2.10) and the functions $p_{\mathbf{j}_s}$ depend on our choice of δ. Here we have used the fact that $\underline{h}_s^{d_s} \leq \mathrm{Diam}\, \delta$. Pick $\delta \in \triangle_s$, and observe that since $\mu^* = 0$,

$$E\big[\langle p_{\mathbf{j}_s} , \mu \rangle_n \big] \;=\; E\big[p_{\mathbf{j}_s}(\mathbf{X}_s)\, \mu(\mathbf{X}) \big] \;=\; 0 ,$$

for $\mathbf{j}_s \leq m$, Moreover, by Condition 2 and the boundedness of μ,

$$\begin{aligned} \max_{\mathbf{j}_s \leq m} \mathrm{var}\, \big(\langle p_{\mathbf{j}_s} , \mu \rangle_n \big) \;&=\; \big(1/n \big) \max_{\mathbf{j}_s \leq m} \mathrm{var}\, \big(p_{\mathbf{j}_s}(\mathbf{X}_s)\, \mu(\mathbf{X}) \big) \\ &=\; \big(1/n \big) \max_{\mathbf{j}_s \leq m} E\big[p_{\mathbf{j}_s}^2(\mathbf{X}_s)\; \mu^2(\mathbf{X}) \big] \\ &\leq\; M_3 \, \big(\mathrm{vol}\, \delta \,/\, n \big) \sup_{\mathbf{x} \in \mathcal{X}} \mu^2(\mathbf{x}) . \end{aligned} \tag{4.9}$$

Therefore, combining (4.8) and (4.9), we find that for each $\delta \in \triangle_s$,

$$\| \tilde{\mu}_\delta \|_n^2 \;=\; O_P\big(\underline{h}_s^{-d_s} \, \mathrm{vol}\, \delta \,/\, n \big) .$$

Since the sets $\delta \in \triangle_s$ are disjoint, the functions $\tilde{\mu}_\delta$ appearing in (4.7) have disjoint supports, so they are orthogonal. Therefore, by summing across the sets in $\triangle_s$, we find that

$$\| \tilde{\mu}_s \|_n^2 \;=\; O_P\big(\underline{h}_s^{-d_s} \,/\, n \big) . \tag{4.10}$$

Repeating the argument leading from (4.8) to (4.10) for each $s \in \mathcal{S}$ yields the desired result. □

As noted in Section 1.2, from the point of view of the present chapter, the rate at which we can approximate μ from functions in $\mathbb{G}_s$, $s \in \mathcal{S}$, is appropriately left as a condition. Therefore, for ANOVA decompositions, we modify Condition 5 as follows.

Condition 5′ *Assume that, for each $s \in \mathcal{S}$, there exists a function $\rho_s(\overline{h}_s)$ such that*

$$\rho_s(\overline{h}_s) \to 0 \quad \text{as} \quad \overline{h}_s \to 0 \qquad \text{and} \qquad \inf_g \| g - \mu_s^* \|_\infty \;=\; O\Big(\rho_s(\overline{h}_s)\Big)\,,$$

where the infinum is taken over all $g \in \mathbb{G}_s$.

Lemma 4.5 *Suppose Conditions 1, 2, 3′, 4 and 5′ hold. If $\mu^* = \mu$, then*

$$\| E(\hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) - \mu_s^* \|_n^2 \;=\; O_P\left(\sum_{s \in \mathcal{S}} \rho_s^2(\overline{h}_s) \;+\; \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \right), \quad s \in \mathcal{S}.$$

Proof Suppose that $\mathbb{PP}$ and hence $\mathbb{G}$ are identifiable, and fix $s \in \mathcal{S}$. From Conditions 3′ and 5′, we can find a function $g_s \in \mathbb{G}_s$ and a positive constant M_{12} that does not depend on n or $\overline{h}_s$ such that

$$\| \mu_s^* - g_s \|_\infty \;\le\; M_{12}\, \rho_s(\overline{h}_s)\,.$$

Let r be any proper subset of s and let $\tilde{g}_{sr}$ denote the orthogonal projection of g_s onto $\mathbb{PP}_r$ relative to the empirical inner product. Then, following (1.21c), we can write

$$\tilde{g}_{sr} \;=\; \sum_{\delta \in \triangle_r} \tilde{g}_\delta\,, \tag{4.11}$$

where each function $\tilde{g}_\delta$ is a polynomial of coordinate degree m. Given $\delta \in \triangle_r$ write

$$\tilde{g}_\delta(\mathbf{x}_r) \;=\; \sum_{\mathbf{j}_r \le m} \tilde{\beta}_{\mathbf{j}_r}\, p_{\mathbf{j}_r}(\mathbf{x}_r)\,, \qquad \mathbf{x}_r \in \mathcal{X}_r\,,$$

as in (1.21b). From the discussion prior to Lemma 2.3 and Lemma 2.3 itself we find that except on a set whose probability tends to zero with n, for $\delta \in \triangle_r$ the functions g_δ defined above satisfy

$$\| \tilde{g}_\delta \|_n^2 \;\le\; M_7\, h^{-d_r} \sum_{\mathbf{j}_r \le m} \langle\, p_{\mathbf{j}_r}, g_s \,\rangle_n^2\,, \tag{4.12}$$

where the functions $p_{\mathbf{j}_r}$ depend on our choice of δ and $h = \operatorname{Diam}\delta$. Pick $\delta \in \triangle_r$ and observe that since $\mu_s^* \in \mathbb{F}_s^0$,

$$\big\langle\, p_{\mathbf{j}_r}\,, \mu_s^* \,\big\rangle \;=\; E\big[\, p_{\mathbf{j}_r}(\mathbf{X})\, \mu_s^*(\mathbf{X}) \,\big] \;=\; 0\,,$$

for all $\mathbf{j}_r \le m$, and hence

$$\begin{aligned}\max_{\mathbf{j}_r \le m} \left| E\left[\langle p_{\mathbf{j}_r}, g_s \rangle_n \right] \right| &= \max_{\mathbf{j}_r \le m} \left| E\left[p_{\mathbf{j}_r}(\mathbf{X})\, g_s(\mathbf{X}) \right] \right| \\ &= \max_{\mathbf{j}_r \le m} \left| E\left[p_{\mathbf{j}_r}(\mathbf{X}) \left(g_s(\mathbf{X}) - \mu_s^*(\mathbf{X}) \right) \right] \right| .\end{aligned}$$

Therefore,

$$\max_{\mathbf{j}_r \le m} \left| E\left[\langle p_{\mathbf{j}_r}, g_s \rangle_n \right] \right| \le M_{12} \left(\operatorname{vol} \delta \right) \rho_s(\overline{h}_s) . \tag{4.13}$$

Similarly, by Conditions $3'$ and $5'$ and the boundedness of μ_s^*, for n sufficiently large, there exists a positive constant M_{13} independent of δ such that

$$\operatorname{var}\left(p_{\mathbf{j}_s}(\mathbf{X})\, g_s(\mathbf{X}) \right) \le M_{13} \left(\operatorname{vol} \delta \right), \qquad \text{for } \mathbf{j}_s \le m,$$

and hence

$$\begin{aligned}\max_{\mathbf{j}_r \le m} \operatorname{var}\left(\langle p_{\mathbf{j}_r}, g_s \rangle_n \right) &= (1 / n) \max_{\mathbf{j}_r \le m} \operatorname{var}\left(p_{\mathbf{j}_r}(\mathbf{X})\, g_s(\mathbf{X}) \right) \\ &\le M_{13} \left(\operatorname{vol} \delta / n \right) . \end{aligned} \tag{4.14}$$

Combining (4.12)–(4.14), using the fact that $(\operatorname{vol} \delta)$ is bounded above by h^{d_r}, we find that

$$\begin{aligned}\left\| \tilde{g}_\delta \right\|_n^2 &= h^{-d_r} O_P\left((\operatorname{vol} \delta)^2 \rho_s^2(\overline{h}_s) + (\operatorname{vol} \delta / n) \right) \\ &= (\operatorname{vol} \delta) O_P\left(\rho_s^2(\overline{h}_s) + \left(h_r^{-d_r} / n \right) \right) .\end{aligned}$$

Therefore, arguing as in the proof of (4.10), we find that

$$\| \tilde{g}_{sr} \|_n^2 = O_P\Big(\rho_s^2(\overline{h}_s) + \left(\underline{h}_r^{-d_r} / n \right) \Big) . \tag{4.15}$$

Now, since $g_s \in \mathbb{G}_s \subset \mathbb{G}$, if $\mathbb{G}$ is identifiable we can write

$$g_s = \sum_{r \subset s} g_{sr} , \qquad \text{where } g_{sr} \in \mathbb{G}_r^0 \text{ for } r \subset s .$$

Relative to the empirical inner product, let $\tilde{g}_{sr}^0$ denote the orthogonal projection of $\tilde{g}_{sr}$ onto $\mathbb{G}_r^0$, which equals the orthogonal projection of g_s onto $\mathbb{G}_r^0$ for $r \subset s$. Let r be a proper subset of s. Since $g_{ss} \in \mathbb{G}_s^0$, the orthogonal projection of g_{ss} onto $\mathbb{G}_r^0$ is zero, and hence $\tilde{g}_{sr}^0$ is also the orthogonal projection of $g_s - g_{ss}$ onto $\mathbb{G}_r^0$. Now,

$$\left\| \tilde{g}_{sr}^0 \right\|_n^2 \le \left\| \tilde{g}_{sr} \right\|_n^2 .$$

Thus, using (4.15) and applying Lemma 4.3 to the function $g_s - g_{ss}$ and the collection of all proper subsets of s, we find that

$$\| g_s - g_{ss} \|_n^2 \;=\; O_P\Big(\sum_{r \subset s} \| \tilde{g}^0_{sr} \|_n^2 \Big) \;=\; O_P\Big(\rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{r \subset s} \underline{h}_r^{-d_r} / n \Big)\,,$$

where each sum is taken over proper subsets r of s.

Finally, then, replacing g_s by g_{ss} if necessary, we see that for each $s \in \mathcal{S}$, there exists a function $g_s \in \mathbb{G}_s^0$ such that

$$\| g_s - \mu_s^* \|_n^2 \;=\; O_P\Big(\rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{r \subset s} \underline{h}_r^{-d_r} / n \Big)\,,$$

and hence, setting $g = \sum_s g_s \in \mathbb{G}$,

$$\| g - \mu^* \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big)\,. \tag{4.16}$$

Since $E(\,\hat{\mu}\,|\,\mathbf{X}_1, \ldots, \mathbf{X}_n\,)$ is the orthogonal projection of μ onto $\mathbb{G}$ relative to the empirical inner product, we see that if $\mu^* = \mu$, then

$$\| E(\,\hat{\mu}\,|\,\mathbf{X}_1, \ldots, \mathbf{X}_n\,) - \mu^* \|_n^2 \;\le\; \| g - \mu^* \|_n^2\,.$$

Thus, by (4.16),

$$\| E(\,\hat{\mu}\,|\,\mathbf{X}_1, \ldots, \mathbf{X}_n\,) - \mu^* \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big)\,,$$

and hence

$$\| E(\,\hat{\mu}\,|\,\mathbf{X}_1, \ldots, \mathbf{X}_n\,) - g \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big)\,.$$

Therefore, from Lemma 4.6, we find that

$$\| E(\,\hat{\mu}_s\,|\,\mathbf{X}_1, \ldots, \mathbf{X}_n\,) - g_s \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big)\,,$$

and hence that

$$\| E(\,\hat{\mu}_s\,|\,\mathbf{X}_1, \ldots, \mathbf{X}_n\,) - \mu_s^* \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big)\,,$$

as desired □

The following lemma reformulates Lemma 2.4 in the context of the spaces $\mathbb{G}_s$, $s \in \mathcal{S}$. Its proof is identical to that of Lemma 2.4 and is not repeated here.

Lemma 4.6 *Suppose Conditions* 1, 2, 3′ *and* 5′ *hold. Then there exists a positive number* M_{14} *not depending on* n *or* $\overline{h}_s$ *such that, except on an event whose probability tends to zero with* n,

$$\| g - \mu_s^* \|^2 \;\le\; M_{14}\Big(\rho_s^2(\,\overline{h}_s\,) \;+\; \| g - \mu_s^* \|_n^2 \Big)\,, \qquad g \in \mathbb{G}_s\,.$$

Rate of Convergence

We end this chapter with the L_2 rate of convergence for ANOVA models.

Theorem 4.1 *Suppose Conditions* 1, 2, 3′, 4 *and* 5′ *hold and let* $\mathcal{S}$ *be a hierarchical collection of subsets of* $\{1, \ldots, M\}$. *Then,*

$$\| \hat{\mu}_s - \mu_s^* \|^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\, \overline{h}_s \,) + \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big), \tag{4.17a}$$

and hence

$$\| \hat{\mu} - \mu^* \|^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\, \overline{h}_s \,) + \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big). \tag{4.17b}$$

Proof Recall the variance-bias decomposition introduced in Section 1.2,

$$\| \hat{\mu}_s - \mu_s^* \|^2 \;\le\; 2 \| \hat{\mu}_s - E(\, \hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n \,) \|^2 + 2 \| E(\, \hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n \,) - \mu_s^* \|^2 ,$$

as applied to the components of our ANOVA decomposition. From Lemma 4.2, we find that the first term in this decomposition satisfies

$$\| \hat{\mu}_s - E(\, \hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n \,) \|^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big).$$

As for the second term, let Q denote the orthogonal projection onto $\mathbb{G}$ relative to the empirical inner product defined in (1.24a). Assume that $\mathbb{G}$ is identifiable, and recall that

$$Q\,\mu \;=\; E(\, \hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n \,) \;=\; \sum_{s \in \mathcal{S}} E(\, \hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n \,),$$

and for convenience define the quantity

$$Q\,\mu^* \;=\; \sum_{s \in \mathcal{S}} \tilde{\mu}_s^*, \qquad \text{where } \tilde{\mu}_s^* \in \mathbb{G}_s^0 \text{ for } s \in \mathcal{S}.$$

Therefore, for each $s \in \mathcal{S}$, we find from the triangle and Schwartz inequalities that the quantity

$$\| E(\, \hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n \,) - \mu_s^* \|_n^2$$

is bounded by

$$2 \| E(\, \hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n \,) - \tilde{\mu}_s^* \|_n^2 \;+\; 2 \| \tilde{\mu}_s^* - \mu^* \|_n^2 .$$

By construction, the function $\mu - \mu^*$ satisfies the hypothesis of Lemma 4.4, and hence we find that

$$\left\| E(\,\hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\,) - \tilde{\mu}_s^* \right\|_n^2 \;=\; O_P\!\left(\sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \right).$$

Similarly, we can apply Lemma 4.5 to μ^* to find that

$$\left\| \tilde{\mu}_s^* - \mu_s^* \right\|_n^2 . \;=\; O_P\!\left(\sum_{s \in \mathcal{S}} \rho_s^2(\,\overline{h}_s\,) + \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \right).$$

Therefore, from Lemma 4.6 we have

$$\left\| E(\,\hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\,) - \mu_s^* \right\|^2 \;=\; O_P\!\left(\sum_{s \in \mathcal{S}} \rho_s^2(\,\overline{h}_s\,) + \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \right).$$

Combining this with the bound on the variance component derived above, we find that the relation in (4.17a) holds, and hence the relation in (4.17b) also holds. □

Example 2 Suppose that for positive constants $T_1, \ldots, T_M$, the functions $\rho_s^2(\,\overline{h}_s\,)$, $s \in \mathcal{S}$, are given by

$$\rho_s^2(\,\overline{h}_s\,) \;=\; \sum_{l \in s} T_l\, \overline{h}_l^{\,2p_l}, \qquad \text{where } p_l \in \mathbb{N} \text{ for } l \in s\,. \tag{4.18}$$

Set $d = \mathrm{Dim}\mathcal{X}$ and $p = (p_1, \ldots, p_M)$, and define the quantities d_s and p_s as in (3.1). Suppose that for each $1 \le l \le M$, we take $\underline{h}_l \sim \overline{h}_l$ to be $O(\,n^{-\gamma_l}\,)$ for

$$\gamma_l \;=\; \min_s \; \frac{1}{p_l\,(\,2 + [\,d_s \,/\, p_s\,]\,)}, \tag{4.19}$$

where the minimum is taken over all sets $s \in \mathcal{S}$ that contain l. Since

$$[\,d_r \,/\, p_r] \;\le\; [\,d_s \,/\, p_s] \qquad \text{if } r \subset s \text{ and } r \neq \varnothing\,, \tag{4.20}$$

the minimum in (4.19) can be found by considering only those maximal sets that contain l. In particular, let $\mathcal{S}'$ denote the collection of sets in $\mathcal{S}$ for which $[\,d_s \,/\, p_s]$ is a maximum. With the assignment in (4.19), we find that both the bias and variance components of the error in (4.17) are $O(\,n^{-\gamma^*}\,)$ where

$$\gamma^* \;=\; \max_{s \in \mathcal{S}} \; \frac{2}{2 \;+\; [\,d_s \,/\, p_s\,]}\,.$$

Therefore, using (4.19) to determine the rate at which we refine the partitions $\triangle_l$, $1 \le l \le M$, the rate in (4.17) is found to be $O(\,n^{-\gamma^*}\,)$.

Next, choose some set $s \in \mathcal{S}'$. From (4.20) we know that s must be a maximal set in $\mathcal{S}$. Therefore, using the rule in (4.19), we find that for this choice of s, each h_l, $l \in s$, is set to the value that we would obtain by considering only the variables in s, as was effectively done in Example 1. Therefore, if we ignore the other sets in

$\mathbb{S}$, the best rate obtainable from the saturated subspace associated with s is again $O(n^{-\gamma^*})$. However, since the rate cannot improve by considering the remaining sets in $\mathbb{S}$, this must be the best rate obtainable from the entire space $\mathbb{G}$.

1.5 Alternate ANOVA Decompositions

In this section we consider a different recipe for forming the ANOVA decompositions discussed in Sections 1.1 and 1.3. Essentially, we introduce an alternative to the theoretical and empirical inner-products defined in Section 1.1, and base the construction of the spaces $\mathbb{F}$ and $\mathbb{G}$ on these new quantities. Many of the results in the previous sections can be extended almost immediately to these new inner-products, and so in this section we concentrate on the less trivial extentions.

Let $\mathbf{X} = (X_1, \ldots, X_M)$ be a vector taking values on $\mathcal{X}$, and for each $1 \le l \le M$, let $f_l(x_l)$ denote the marginal density of X_l on $\mathcal{X}_l$. Set $f^*_{\mathbf{X}}$ equal to product of these marginal densities, and let E^* denote the expectation operator with repect to $f^*_{\mathbf{X}}$. Under Condition 2, we know that the function $f^*_{\mathbf{X}}$ is also bounded above and below by a constant; that is, there exists a positive constant M_{15} such that

$$\frac{1}{M_{15}} \;\le\; f^*_{\mathbf{X}}(\,\mathbf{x}\,) \;\le\; M_{15}\,, \qquad \mathbf{x} \in \mathcal{X}\,. \tag{5.1}$$

Next, let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ denote a random sample of size n from the distribution of the random vector $\mathbf{X}$ For any function $f(\,\cdot\,)$ defined on $\mathcal{X}$, set

$$E^*_n\big[\,f(\mathbf{X})\,\big] = \frac{1}{n^M} \sum_{i_1=1}^{n} \cdots \sum_{i_M=1}^{n} f(\mathbf{X}_{1i_1}, \ldots, \mathbf{X}_{Mi_M}) \tag{5.2}$$

and

$$\big|\,E^*_n - E^*\,\big|\big[\,f(\mathbf{X})\,\big] = \big|\,E^*_n\big[\,f(\mathbf{X})\,\big] - E^*\big[\,f(\mathbf{X})\,\big]\,\big|\,.$$

Observe that E^*_n is the expectation operator with respect to the product of the marginal empirical distributions induced by the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Piecewise Polynomials

As in Section 1.1, let $\delta \in \triangle$, set $h = \mathrm{Diam}\,\delta$, and choose $\mathbf{x}_0 \in \delta$. In the following lemma, we consider random variables of the form

$$\mathbf{W} = \big[\,(\,\mathbf{X} - \mathbf{x}_0\,)\,/\,h\,\big]\,I_\delta(\,\mathbf{X}\,)\,. \tag{5.3}$$

Recall that given a binary random variable I and a nonnegative integer j, we define I^j to be zero if $I = 0$, even if $j = 0$. With this convention, given any polynomial p

in the form (1.8), $p(\mathbf{W})$ is zero if $\mathbf{X}$ does not lie in δ. The following lemma extends the results of Lemma 1.3 to the new expectation operators defined above.

Lemma 5.1 *Suppose Conditions* 1 *and* 2 *hold, choose* $\delta \subset \triangle$, *and set* $h = \mathrm{Diam}\,\delta$ *and* $d = \dim \mathcal{X}$. *For any point* $\mathbf{x}_0$ *in* δ, *define* $\mathbf{W}$ *as in* (5.3). *Then there exists a positive constant* M_4 *such that for any* $t > 0$, *the inequality*

$$\left| E_n^* - E^* \right| \left[p_1(\mathbf{W})\, p_2(\mathbf{W}) \right] \;\leq\; t \sqrt{E^* \left[p_1^2(\mathbf{W}) \right]} \sqrt{E^* \left[p_2^2(\mathbf{W}) \right]} \tag{5.4}$$

holds simultaneously for all polynomials p_1, p_2 *of the form* (1.8), *except on an event having probability at most*

$$M_4 \left(\exp(-\eta_1 / M_4) + \cdots + \exp(-\eta_M / M_4) \right), \tag{5.5a}$$

where

$$\eta_l \;=\; n\, h_l^{d_l}\, t^2 / (t+1), \qquad 1 \leq l \leq M. \tag{5.5b}$$

Proof Throughout this proof, let $T_1, T_2, \ldots$ denote suitable positive constants. Recall that by construction, if $\delta \in \triangle$, then $\delta = \delta_1 \times \cdots \times \delta_M$, where $\delta_l \in \triangle_l$. In addition, let $m = (m_1, \ldots, m_M) \in \mathbb{N}^M$ and $\mathbf{W} = (W_1, \ldots, W_M)$. Then, given the positive constants $t_1, \ldots, t_M$, we find from the proof of Lemma 1.3 that for each $1 \leq l \leq M$, the inequality

$$\left| E_n - E \right| \left[W_l^{j_{1l} + j_{2l}} \right] \;\leq\; t_l \tag{5.6}$$

holds simultaneously for all of the monomials $W_l^{j_{1l}+j_{2l}}$, $j_{1l}, j_{2l} \leq m_l$, except on a set having probability not larger than

$$2\,(2m_l + 1)^{d_l} \exp(-\eta_l),$$

where

$$\eta_l \;=\; n\, t_l^2 / (2\, M_3\, \mathrm{vol}\,\delta_l + 4\, t_l / 3). \tag{5.7}$$

Therefore, there exists a positive constant T_1 such that except on a set with probability at most

$$T_1 \left(\exp(-\eta_1) + \cdots + \exp(-\eta_M) \right), \tag{5.8}$$

the inequalitites in (5.6) hold simultaneously for all $j_{1l}, j_{2l} \leq m_l$ and $1 \leq l \leq M$. Recall that by definition $\mathbf{W}^{\mathbf{j}} \;=\; W_1^{j_1} \cdots W_M^{j_M}$ and hence

$$E_n^* \left[\mathbf{W}^{\mathbf{j}} \right] \;=\; E_n \left[W_1^{j_1} \right] \cdots E_n \left[W_M^{j_M} \right]$$

and

$$E^* \left[\mathbf{W}^{\mathbf{j}} \right] \;=\; E \left[W_1^{j_1} \right] \cdots E \left[W_M^{j_M} \right].$$

Let $\mathbf{j} = \mathbf{j}_1 + \mathbf{j}_2$, where $\mathbf{j}_1, \mathbf{j}_2 \le m$. We claim that there exists a positive constant T_2 such that except on a set having probability at most (5.8), the inequality

$$\left| E_n^* - E^* \right| \left[\mathbf{W}^{\mathbf{j}} \right] \;\le\; (t_1 + h_1^{d_1}) \cdots (t_M + h_M^{d_M}) - h^d, \tag{5.9}$$

holds simultaneously for all such vectors $\mathbf{j}$. To see that (5.9) is valid, temporarily let $\mu_l = E\left[W_l^{j_l} \right]$, $1 \le l \le M$, and observe that $|\mu_l| \le h_l^{d_l}$. Therefore, except on a set having probability at most (5.8),

$$\begin{aligned}
&\left| E_n^* - E^* \right| \left[\mathbf{W}^{\mathbf{j}} \right] \\
&\qquad = \;\; \left| E_n\left[W_1^{j_1} \right] \cdots E_n\left[W_M^{j_M} \right] \;-\; \mu_1 \cdots \mu_M \right| \\
&\qquad = \;\; \left| E_n\left[W_1^{j_1} - \mu_1 + \mu_1 \right] \cdots E_n\left[W_M^{j_M} - \mu_M + \mu_M \right] \;-\; \mu_1 \cdots \mu_M \right|. \\
&\qquad \le \;\; (t_1 + h_1^{d_1}) \cdots (t_M + h_M^{d_M}) - h^d,
\end{aligned}$$

where the last inequality follows from (5.6). Now, as was done in the proof of Lemma 2.3, let $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ be polynomials in the form (1.8). Then,

$$p_1(\mathbf{x}) = \sum_{\mathbf{j}_1 \le m} b_{1\mathbf{j}_1} \mathbf{x}^{\mathbf{j}_1} \quad \text{and} \quad p_2(\mathbf{x}) = \sum_{\mathbf{j}_2 \le m} b_{2\mathbf{j}_2} \mathbf{x}^{\mathbf{j}_2},$$

where $b_{1\mathbf{j}_1}, b_{2\mathbf{j}_2} \in \mathbb{R}$ for $\mathbf{j}_1, \mathbf{j}_2 \le m$. On the event that (5.9) holds for all $\mathbf{j}_1, \mathbf{j}_2 \le m$, we have that

$$\begin{aligned}
&\left| E_n^* - E^* \right| \left[p_1(\mathbf{W})\, p_2(\mathbf{W}) \right] \\
&\qquad \le \;\; \left((t_1 + h_1^{d_1}) \cdots (t_M + h_M^{d_M}) - h^d \right) \sum_{\mathbf{j}_1 \le m} \left| b_{1\mathbf{j}_1} \right| \sum_{\mathbf{j}_2 \le m} \left| b_{2\mathbf{j}_2} \right|.
\end{aligned} \tag{5.10}$$

In addition, by (5.1) and Lemma 1.2,

$$M_{15}\, E^*\left[p_1^2(\mathbf{W}) \right] \;\ge\; \int_\delta p_1^2\left[(\mathbf{x} - \mathbf{x}_0)/h \right] d\mathbf{x} \;\ge\; (h^d / M_2) \sum_{\mathbf{j}_1 \le m} \left| b_{1\mathbf{j}_1} \right|^2.$$

Combining this with a similar result for $p_2(\mathbf{W})$ and applying the Schwartz inequality, we find that (5.10) becomes

$$\begin{aligned}
&\left| E_n^* - E^* \right| \left[p_1(\mathbf{W})\, p_2(\mathbf{W}) \right] \\
&\qquad \le \;\; T_2 \left((t_1 h_1^{-d_1} + 1) \cdots (t_M h_M^{-d_M} + 1) - 1 \right) \sqrt{E^*\left[p_1^2(\mathbf{W}) \right]} \, \sqrt{E^*\left[p_2^2(\mathbf{W}) \right]},
\end{aligned}$$

for some suitable positive constant T_2. By rescaling each t_l in (5.6) and choosing M_4 sufficiently large, we arrive at the probability in (5.5). □

Repeating the arguments in the proof of Lemma 1.4, we find that the relationship in (5.4) holds for the more generallly for piecewise polynomials. In this case, however, the exceptional probability is of the form

$$M_4 \underline{h}^{-d} \left(\exp(-\eta_1) + \cdots + \exp(-\eta_M) \right), \tag{5.11a}$$

where

$$\eta_l \;=\; n h_l^{d_l} t^2 / (t+1), \qquad 1 \le l \le M. \tag{5.11b}$$

Suppose that the partitions $\triangle_l$, $1 \le l \le M$, are refined with sample size so that Condition 3 holds. Then, for n sufficiently large,

$$\exp(-n \underline{h}_1^{d_1}) + \cdots + \exp(-n \underline{h}_M^{d_M}) \;\le\; M \exp(-n\underline{h}^d),$$

and hence if Condition 3 holds, the probability in (5.11) tends to zero with n. As we will see, this type of relationship will be most useful in the context of ANOVA decompositions.

Now, for any two functions $f_1(\cdot)$ and $f_2(\cdot)$ defined on $\mathcal{X}$, set

$$^*\langle f_1, f_2 \rangle_n = E_n^*[f_1(\mathbf{X}) f_2(\mathbf{X})] \quad \text{and} \quad {}^*\| f_1 \|_n^2 = E_n^*[f_1^2(\mathbf{X})]. \tag{5.12a}$$

The theoretical versions of these quantities are given by

$$^*\langle f_1, f_2 \rangle = E^*[f_1(\mathbf{X}) f_2(\mathbf{X})] \quad \text{and} \quad {}^*\| f_1 \|^2 = E^*[f_1^2(\mathbf{X})]. \tag{5.12b}$$

Using (5.11) and following the proof of Lemma 1.5 we find that if Contitions 1–3 hold, then except on a set whose probablility tends to zero with n

$$\tfrac{1}{2} {}^*\| p \|^2 \;\le\; {}^*\| p \|_n^2 \;\le\; 2\, {}^*\| p \|^2, \qquad p \in \mathbb{PP}. \tag{5.13}$$

Now, since $f_{\mathbf{X}}$ and $f_{\mathbf{X}}^*$ are each bounded away from zero and infinity on $\mathcal{X}$, we find that $^*\| \cdot \|$ and $\| \cdot \|$ are equivalent. Combining this fact with (5.13) and the proof of the inequality in (1.26), we find that under Conditions 1–3 the inequalities in (5.13) hold for the sample-based quantities $^*\| \cdot \|_n$ and $\| \cdot \|_n$ with overwhelming probability; that is, except on a set whose probability tends to zero with n,

$$\tfrac{1}{2} \| p \|_n^2 \;\le\; {}^*\| p \|_n^2 \;\le\; 2 \| p \|_n^2, \qquad p \in \mathbb{PP}. \tag{5.14}$$

Shortly, we will see that this equivalence holds for the spline spaces $\mathbb{G}_s$, $s \in \mathcal{S}$, as well. Through the use of this equivalence and its generalization, we are able to translate the results of the previous sections into statements about our new inner-products (5.12). First, however, we collect two facts about the inner products defined in (5.12).

Lemma 5.2 *Suppose Conditions 1 and 2 hold, choose $\delta \subset \triangle$, and set $h = \mathrm{Diam}\,\delta$ and $d = \dim \mathcal{X}$. For any point $\mathbf{x}_0 \in \delta$ define $\mathbf{W}$ as in (5.3). Then, there exit positive constants M_{16} and M_{17} depending only on m, d and M such that for all $\mathbf{j}_1, \mathbf{j}_2 \leq m$,*

$$E\left(E_n^* \left[\mathbf{W}^{\mathbf{j}_1+\mathbf{j}_2} \right] \right) \;\leq\; M_{16} \sqrt{\mathrm{vol}\,\delta}\,/\,n \;+\; E^* \left[\mathbf{W}^{\mathbf{j}_1+\mathbf{j}_2} \right]. \tag{5.15}$$

Proof Let $\mathbf{j}$ be some index of the form (1.5), and let

$$W_l \;=\; \left[(\, X_l - x_{0l} \,) \,/\, h_l \right] I_{\delta_l}(x_l)\,, \qquad \text{for } 1 \leq l \leq M\,.$$

Then, as in the previous lemma, we find that

$$E_n^* \left[\mathbf{W}^{\mathbf{j}} \right] \;=\; E_n \left[W_1^{j_1} \right] \cdots E_n \left[W_M^{j_M} \right]$$

and

$$E^* \left[\mathbf{W}^{\mathbf{j}} \right] \;=\; E \left[W_1^{j_1} \right] \cdots E \left[W_M^{j_M} \right].$$

Suppose initially that $M = 2$ and that $\delta = \delta_1 \times \delta_2$. Then, for $\mathbf{j}_1, \mathbf{j}_2 \leq m$, set $\mathbf{j} = (j_1, j_2) = \mathbf{j}_1 + \mathbf{j}_2$ and observe that

$$\begin{aligned} E\left(E_n^* \left[\mathbf{W}^{\mathbf{j}} \right] \right) \;&=\; E\left(E_n \left[W_1^{j_1} \right] E_n \left[W_2^{j_2} \right] \right) \\ &=\; \mathrm{cov}\left(E_n \left[W_1^{j_1} \right], E_n \left[W_2^{j_2} \right] \right) + E \left[W_1^{j_1} \right] E \left[W_2^{j_2} \right] \\ &=\; \mathrm{cov}\left(E_n \left[W_1^{j_1} \right], E_n \left[W_2^{j_2} \right] \right) + E^* \left[\mathbf{W}^{\mathbf{j}} \right] \end{aligned}$$

However, observe that the covariance in the above expression is equal to

$$\frac{1}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \mathrm{cov}\left(W_{1i_1}^{j_1}, W_{2i_2}^{j_2} \right),$$

which by independence and Condition 2 becomes

$$\frac{1}{n}\, \mathrm{cov}\left(W_1^{j_1}, W_2^{j_2} \right) \;\leq\; M_3 \sqrt{\mathrm{vol}\,\delta_1 \cdot \mathrm{vol}\,\delta_2}\,/\,n \;=\; M_3 \sqrt{\mathrm{vol}\,\delta}\,/\,n\,,$$

since W_1 and W_2 take on nonzero values only on the sets δ_1 and δ_2, respectively, and since each of these functions are bounded in absolute value by 1. While we do not present it here, a similar argument can be used to demonstrate the inductive step required to obtain the inequality appearing in (5.15) for general M. □

Lemma 5.3 *Suppose Conditions* 1 *and* 2 *hold, choose* $\delta \subset \triangle$, *and set* $h = \mathrm{Diam}\,\delta$ *and* $d = \dim \mathcal{X}$. *Let* f *be any bounded function whose domain includes* δ. *Then, there exits a positive constant* M_{17} *depending only on* m, d *and* M *such that for* n *sufficiently large,*

$$\mathrm{var}\left(E_n^* \left[f(\mathbf{X})\, I_\delta(\mathbf{X}) \right] \right) \;\leq\; M_{17}\, (\,\mathrm{vol}\,\delta\,/\,n\,) \sup_{\mathbf{x}\in\mathcal{X}} f^2(\,\mathbf{x}\,) \tag{5.16}$$

Proof Expanding the variance in the relation in (5.16), we find that

$$\mathrm{var}\left(E_n^* \left[f(\mathbf{X})\, I_\delta(\mathbf{X}) \right] \right) \;=\; \mathrm{cov}\left(E_n^* \left[f(\mathbf{X})\, I_\delta(\mathbf{X}) \right], E_n^* \left[f(\mathbf{X})\, I_\delta(\mathbf{X}) \right) \right)$$

which can be rewritten as

$$\frac{1}{n^{2M}} \sum_{i_1=1}^{n} \sum_{i_1'=1}^{n} \cdots \sum_{i_M=1}^{n} \sum_{i_M'=1}^{n} \mathrm{cov}\left(f(X_{1i_1}, \cdots, X_{Mi_M}),\, f(X_{1i_1'}, \cdots, X_{Mi_M'}) \right).$$

Observe, however, that if the two groups of indices $i_1, \ldots, i_M$ and $i_1', \ldots, i_M'$ are disjoint, then the covariance appearing above is zero. Therefore, by a simple counting argument, we find there are

$$n^{2M} \;-\; n^M \left(n\,(\,n-1\,) \cdots (\,n-M\,) \right) \tag{5.17}$$

non-zero terms in the sum. Furthermore, by Condition 2, we know that there exists a positive constant T_1 such that each nonzero term above is bounded in absolute value by T_1 times the square of the $L_2(\delta)$-norm of f. Finally, then we find from (5.17) that

$$\mathrm{var}\left(E_n^* \left[f(\mathbf{X})\, I_\delta(\mathbf{X}) \right] \right) \;\leq\; T_1 \left(1 \;-\; \frac{(\,n-M\,)^M}{n^M} \right) \| f \|_{L_2(\delta)}^2,$$

which yields the desired inequality. □

ANOVA Decomposition

As in Section 1.3, let $\mathcal{S}$ denote a hierarchical collection of subsets of $\{1, \ldots, M\}$, and for each $1 \leq l \leq M$, let $\mathbb{G}_l$ denote a non-empty subspace of $\mathbb{PP}_l$. Then, given any nonempty subset $s \in \mathcal{S}$, let $\mathbb{G}_s$ denote the tensor product of the spaces $\mathbb{G}_l$, $l \in s$. If $s = \varnothing$, then we take $\mathbb{G}_s$ to be the space of constant functions on $\mathcal{X}$. From these spaces we construct

$$\mathbb{G} \;=\; \left\{ \sum_{s\in\mathcal{S}} g_s \,:\, g_s \in \mathbb{G}_s \;\text{ for }\; s \in \mathcal{S} \right\}. \tag{5.18}$$

For $s \in \mathcal{S}$, let ${}^*\mathbb{G}_s^0$ denote the orthogonal complement of $\mathbb{G}_s$ relative to the sum of $\mathbb{G}_r$ as r ranges over proper subsets of s relative to the empirical inner product ${}^*\langle \cdot, \cdot \rangle_n$. By repeating the argument leading to Lemma 3.1, we find that if $\mathbb{G}$ is identifiable, then each function $g \in \mathbb{G}$ can be written uniquely as

$$g \;=\; \sum_{s \in \mathcal{S}} g_s \qquad \text{where } g_s \in {}^*\mathbb{G}_s^0 \text{ for } s \in \mathcal{S}\,. \tag{5.19}$$

Actually, from the discussion in the proof of Lemma 3.1, we find in addition that for each $s = \{s_1, \ldots, s_k\} \in \mathcal{S}$,

$${}^*\mathbb{G}_s \;=\; {}^*\mathbb{G}_{s_1}^0 \otimes \cdots \otimes {}^*\mathbb{G}_{s_k}^0\,, \tag{5.20}$$

and that the spaces ${}^*\mathbb{G}_s^0$, $s \in \mathcal{S}$ are orthogonal. As we will see, this property greatly simplifies the derivation of many of the results in Section 1.3.

For example, for each $g \in \mathbb{G}$ given by (5.19), we know from (5.20) that

$${}^*\| g \|_n^2 \;=\; \sum_{s \in \mathcal{S}} {}^*\| g_s \|_n^2\,. \tag{5.21}$$

Now, recall from the previous subsection that given a positive constant t, the inequality

$$\left| E_n^* - E^* \right| \left[p_1(\mathbf{W})\, p_2(\mathbf{W}) \right] \;\le\; T_1 \sqrt{E^*\left[p_1^2(\mathbf{W}) \right]}\, \sqrt{E^*\left[p_2^2(\mathbf{W}) \right]} \tag{5.22}$$

holds simultaneously for all polynomials p_1, p_2 of the form (1.8), except on an event having probability at most

$$M_4\, \underline{h}^{-d} \left(\exp(-\eta_1) + \cdots + \exp(-\eta_M) \right), \tag{5.23}$$

where

$$\eta_l \;=\; n\, h_l^{d_l}\, t^2 / (t+1)\,, \qquad 1 \le l \le M\,.$$

Since $\mathbb{G}$ is a subspace of the complete space of polynomials of total degree m in $\mathbf{x}$, the inequality in (5.22) holds simultaneously for all $g \in \mathbb{G}$ as well. Suppose that the partitions $\triangle_l$, $1 \le l \le M$, are refined with sample size so that Condition $3'$ holds. Then, for n sufficiently large,

$$\sum_{l \in s} \exp(-n\, \underline{h}_l^{d_l}) \;\le\; \#(s)\, \exp(-n \underline{h}_s^{d_s})\,,$$

and hence if Condition $3'$ holds, the probability in (5.23) tends to zero with n. Therefore, combining (5.21) and (5.22), we find that given any positive constant t, except on a set whose probability tends to zero with n,

$$(1+t)\, {}^*\| g \|^2 \;\ge\; {}^*\| g \|_n^2 \;=\; \sum_{s \in \mathcal{S}} {}^*\| g_s \|_n^2 \;\ge\; (1-t) \sum_{s \in \mathcal{S}} {}^*\| g_s \|_n^2\,,$$

and hence given any positive constant $0 < T_1 < 1$, except on a set with probability tending to zero with n,

$$ {}^*\| g \|^2 \;\geq\; T_1 \sum_{s \in \mathcal{S}} {}^*\| g_s \|^2 , \tag{5.24} $$

for all functions g given by (5.19). The relation in (5.24) is essentially the result of Lemma 3.6, which was derived through considerably more complicated arguments in Section 3.

The stability relation in Lemma 3.8 can also be simplified through the use of the inner-product in (5.12b). Toward this end, for each $s \in \mathcal{S}$, let $\mathbb{F}_s$ denote the space of square integrable functions on $\mathcal{X}$ that depend only on the variables x_l, $l \in s$, and set

$$ \mathbb{F} \;=\; \left\{ \sum_{s \in \mathcal{S}} f_s \,:\, f_s \in \mathbb{F}_s \;\text{ for }\; s \in \mathcal{S} \right\} . \tag{5.25} $$

Let ${}^*\mathbb{F}_s^0$ denote the space of all functions in $\mathbb{F}_s$ that are orthogonal to each function in $\mathbb{F}_r$, $r \subset s$, relative to the theoretical inner product ${}^*\langle \cdot\,, \cdot \rangle$. From the observations above leading to (5.20), we find that the spaces ${}^*\mathbb{F}_s^0$, $s \in \mathcal{S}$, are orthogonal and hence if

$$ f \;=\; \sum_{s \in \mathcal{S}} f_s \qquad \text{where } f_s \in {}^*\mathbb{F}_s^0 \;\text{ for }\; s \in \mathcal{S} , \tag{5.26} $$

then

$$ {}^*\| f \|^2 \;=\; \sum_{s \in \mathcal{S}} {}^*\| f_s \|^2 . \tag{5.27} $$

From this last relation, we find that every function $f \in \mathbb{F}$ can be written uniquely in the form (5.26).

Regression in Unsaturated Spaces

Using the representations in (5.19) and (5.26), write

$$ \mu^* \;=\; \sum_{s \in \mathcal{S}} \mu_s^* \qquad \text{where } \mu_s^* \in {}^*\mathbb{F}_s^0 \text{ for } s \in \mathcal{S} , \tag{5.28a} $$

and

$$ \hat{\mu} \;=\; \sum_{s \in \mathcal{S}} \hat{\mu}_s \qquad \text{where } \hat{\mu}_s \in {}^*\mathbb{G}_s^0 \text{ for } s \in \mathcal{S} . \tag{5.28b} $$

In Section 4, we divided the task of deriving the mean-squared error in estimating μ^* and its components by $\hat{\mu}$ and its components into a variance term and a bias term. In particular, we found that the variance component satisfied

$$ \left\| \hat{\mu} - E(\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) \right\|^2 \;=\; O_P\!\left(\sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} \,/\, n \right) . $$

Now, using the fact that the norms $\|\cdot\|$ and ${}^*\|\cdot\|$ are equivalent, we find that

$$ {}^*\Big\| \hat{\mu} - E(\,\hat{\mu}\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\,)\,\Big\|^2 \;=\; O_P\Big(\sum_{s\in\mathcal{S}} \underline{h}_s^{-d_s}/\,n\,\Big)\,, $$

and so using the relation in (5.24), we conclude that

$$ {}^*\Big\| \hat{\mu}_s - E(\,\hat{\mu}_s\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\,)\,\Big\|^2 \;=\; O_P\Big(\sum_{s\in\mathcal{S}} \underline{h}_s^{-d_s}/\,n\,\Big)\,, $$

just as in Section 4. We are able to bound one component of the bias term in a similar fashion. To be precise, from Lemma 4.4 we find that

$$ \Big\| E(\,\hat{\mu}'_s\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\,)\,\Big\|_n^2 \;=\; O_P\Big(\sum_{s\in\mathcal{S}} \underline{h}_s^{-d_s}\Big)\,, \qquad s\in\mathcal{S}\,, $$

where we have take $\hat{\mu}'_s$ to be the component of $\hat{\mu}$ corresponding to the set s according to the ANOVA decomposition in Section 1.3. But, by the inequality in (5.14),

$$ {}^*\Big\| E(\,\hat{\mu}'_s\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\,)\,\Big\|_n^2 \;=\; O_P\Big(\sum_{s\in\mathcal{S}} \underline{h}_s^{-d_s}\Big)\,, \qquad s\in\mathcal{S}\,, $$

and hence

$$ {}^*\Big\| E(\,\hat{\mu}\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\,)\,\Big\|_n^2 \;=\; O_P\Big(\sum_{s\in\mathcal{S}} \underline{h}_s^{-d_s}\Big)\,. $$

Finally, using the representation in (5.28b) and the relation in (5.24), we find that

$$ {}^*\Big\| E(\,\hat{\mu}_s\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\,)\,\Big\|_n^2 \;=\; O_P\Big(\sum_{s\in\mathcal{S}} \underline{h}_s^{-d_s}\Big)\,, \qquad s\in\mathcal{S}\,. $$

Now, all that remains is to extend Lemma 4.5 to the new ANOVA decompositions given in (5.28). To do this, we require the results of Lemmas 5.2 and 5.3. Once this result has been established, we can repeat the proof of Theorem 4.1 to obtain the rates of convergence of the components $\hat{\mu}_s$ to μ_s^* for $s\in\mathcal{S}$.

Lemma 5.4 *Suppose Conditions* 1, 2, 3′, 4 *and* 5′ *hold. If* $\mu^* = \mu$, *then*

$$ {}^*\Big\| E(\,\hat{\mu}_s\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\,) - \mu_s^*\,\Big\|_n^2 \;=\; O_P\Big(\sum_{s\in\mathcal{S}} \rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{s\in\mathcal{S}} \underline{h}_s^{-d_s}/\,n\,\Big)\,. $$

Proof Let $T_1, T_2, \ldots$ denote suitable positive constants. Fix $s\in\mathcal{S}$, and observe that from Conditions 3′ and 5′, we can find a function $g_s\in\mathbb{G}_s$ and a positive constant T_1 that does not depend on n or $\overline{h}_s$ such that

$$ \|\,\mu_s^* - g_s\,\|_\infty \;\le\; T_1\,\rho_s(\,\overline{h}_s\,)\,, \tag{5.29} $$

Since $g_s \in \mathbb{G}_s$, using the representation in (5.21), we find that

$$g_s \;=\; \sum_{r \subset s} g_{sr}\,, \qquad \text{where } g_{sr} \in {}^*\mathbb{G}_r^0 \text{ for } r \subset s\,.$$

Then, using Lemmas 5.2 and 5.3 and a straightforward extention of Lemma 4.3 to cover the inner products and norms in (5.12), and arguing as in the proof of Lemma 4.5, we find that

$${}^*\| g_s - g_{ss} \|_n^2 \;=\; O_P\Big(\rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{r \subset s} \underline{h}_r^{-d_r} / n \Big)\,,$$

where each sum is taken over proper subsets r of s. Using the equivalence in (5.14), we also find that

$$\| g_s - g_{ss} \|_n^2 \;=\; O_P\Big(\rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{r \subset s} \underline{h}_r^{-d_r} / n \Big)\,.$$

Therefore, replacing g_s by g_{ss} if necessary, we find from (5.29) that for each $s \in \mathcal{S}$, there exists a function $g_s \in {}^*\mathbb{G}_s^0$ such that

$${}^*\| g_s - \mu_s^* \|_n^2 \;=\; O_P\Big(\rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{r \subset s} \underline{h}_r^{-d_r} / n \Big)\,, \tag{5.30}$$

and

$$\| g_s - \mu_s^* \|_n^2 \;=\; O_P\Big(\rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{r \subset s} \underline{h}_r^{-d_r} / n \Big)\,.$$

From this last relation, by setting $g = \sum_s g_s \in \mathbb{G}$, we have that

$$\| g - \mu^* \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big)\,. \tag{5.31}$$

Since $E(\,\hat{\mu}\,|\,\mathbf{X}_1, \ldots, \mathbf{X}_n\,)$ is the orthogonal projection of μ onto $\mathbb{G}$ relative to the usual empirical inner product introduced in Section 1.1, we see that if $\mu^* = \mu$, then

$$\| E(\,\hat{\mu}\,|\,\mathbf{X}_1, \ldots, \mathbf{X}_n\,) - \mu^* \|_n^2 \;\le\; \| g - \mu^* \|_n^2\,.$$

Thus, by (5.31),

$$\| E(\,\hat{\mu}\,|\,\mathbf{X}_1, \ldots, \mathbf{X}_n\,) - \mu^* \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big)\,,$$

and hence

$$\| E(\,\hat{\mu}\,|\,\mathbf{X}_1, \ldots, \mathbf{X}_n\,) - g \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big)\,.$$

For the moment, let $\hat{\mu}_s'$, $s \in \mathcal{S}$, denote the components of $\hat{\mu}$ relative to the ANOVA decomposition in Section 1.3. Similarly, since $g \in \mathbb{G}$, we can write it uniquely as $\sum_s g_s'$ where each $g_s' \in \mathbb{G}_s^0$. Therefore, from Lemma 4.6, we find that

$$\| E(\,\hat{\mu}_s'\,|\,\mathbf{X}_1, \ldots, \mathbf{X}_n\,) - g_s' \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\,\overline{h}_s\,) \;+\; \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big)\,,$$

and hence by the equivalence in (5.14),

$$^*\| E(\hat{\mu}'_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) - g'_s \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\overline{h}_s) + \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big).$$

By summing across the sets $s \in \mathcal{S}$, we find that

$$^*\| E(\hat{\mu} \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) - g \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\overline{h}_s) + \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big),$$

so that by (5.21)

$$^*\| E(\hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) - g_s \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\overline{h}_s) + \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big),$$

for each $s \in \mathcal{S}$. Finally, combining this with (5.30), we find that

$$^*\| E(\hat{\mu}_s \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n) - \mu_s^* \|_n^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\overline{h}_s) + \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big),$$

as desired. □

By a straightforward modification of the argument used to derive Lemma 4.6, we find that the conclusion of that lemma also holds in this context. Finally, then, we arrive at the following theorem. Its proof is virtually identical to that given for Theorem 4.1 and is not repeated here.

Theorem 5.1 *Suppose Conditions* 1, 2, 3′, 4 *and* 5′ *hold and let* $\mathcal{S}$ *be a hierarchical collection of subsets of* $\{1, \ldots, M\}$. *Then,*

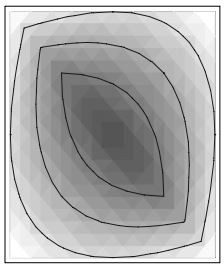
$$^*\| \hat{\mu}_s - \mu_s^* \|^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\overline{h}_s) + \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big),$$

and hence

$$^*\| \hat{\mu} - \mu^* \|^2 \;=\; O_P\Big(\sum_{s \in \mathcal{S}} \rho_s^2(\overline{h}_s) + \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} / n \Big).$$

Chapter 2

Spline Spaces



2.1 Box Splines and Univariate B-splines

Definition and Basic Properties of Box Splines

Throughout the last decade, numerical analysts developed a number of multivariate spline spaces built from functions that are, to a statistician, probability density functions. In this section, we consider a simple but extremely important subclass of these spaces.

Definition 1 *Let $\mathcal{T} = \{t_1, \ldots, t_J\}$ be a collection of vectors in $\mathbb{Z}^d \setminus \{0\}$ that contains a basis of $\mathbb{R}^d$, let $Z_1, \ldots, Z_J$ be independent random variables each distrbuted uniformly on the interval $[0,1]$, and Let $B(x \mid \mathcal{T})$ denote the density of the random variable*

$$X = Z_1 t_1 + \cdots + Z_J t_J \in \mathbb{R}^d.$$

Then, $B(x \mid \mathcal{T})$ is referred to as a box spline with direction vectors $\mathcal{T}$.

For a given collection of vectors $\mathcal{T} = \{t_1, \ldots, t_J\} \subset \mathbb{Z}^d \setminus \{0\}$ that contains a basis of $\mathbb{R}^d$, we can infer a number of properties of $B(x \mid \mathcal{T})$ immediately. First, it is clear that since $B(x \mid \mathcal{T})$ is a density function, it is nonnegative and integrates

to one. Further, the support of $B(\,x\,|\,\mathcal{T})$ is given by the affine cube

$$[\mathcal{T}] \;=\; \{\,\alpha_1 t_1 + \cdots + \alpha_J t_J \;:\; 0 \le \alpha_j \le 1 \text{ for } 1 \le j \le J\,\}\,. \tag{1.1}$$

Before proceeding with a list of the less trivial facts about these functions, we consider a few examples. For clarity, we will restrict ourselves to bivariate box splines and we let e_1 and e_2 denote the unit vectors $(1,0)$ and $(0,1)$, respectively.

Example 1 Suppose that $\mathcal{T}$ is made up of the vector e_1 repeated i_1 times and e_2 repeated i_2 times. Then, by Definition 1, $B(\,x\,|\,\mathcal{T})$ is the density function of the vector

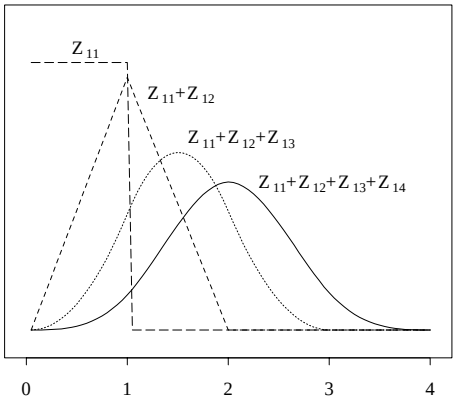
$$X \;=\; (\,Z_{11} + \cdots + Z_{1i_1}, Z_{21} + \cdots + Z_{2i_2}\,) \tag{1.2}$$

where $Z_{11}, \ldots, Z_{1i_1}, Z_{21}, \ldots, Z_{2i_2}$ are independent random variables, each uniformly distributed on the interval $[0,1]$. By independence, $B(\,x\,|\,\mathcal{T})$ is just the tensor product of the marginal densities of X_1 and X_2, where

$$X_1 \;=\; Z_{11} + \cdots + Z_{1i_1}$$

and

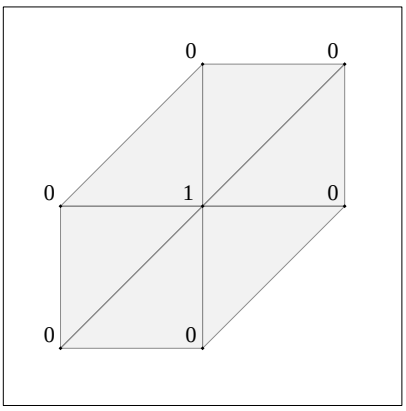
$$X_2 \;=\; Z_{21} + \cdots + Z_{2i_2}\,.$$



In the accompanying figure, we present the density of X_1 for different values of i_1. The resulting functions are referred to as cardinal B-splines. For a given value of i_1, these functions are polynomials of degree $(\,i_1 - 1\,)$ in each of the intervals $(\,k, k+1\,)$ for $k = 0, \ldots, i_1 - 1$. Furthermore, they are $(\,i_1 - 2\,)$ times continuously differentiable, with jumps in the $(\,i_1 - 1\,)$th derivative occuring at the integers $0, \ldots, i_1$.

Since similar statements hold for the density of X_2 and the random variables X_1, X_2 are independent, we conclude from (1.2) that $B(\,x\,|\,\mathcal{T})$ is the tensor product of two cardinal B-splines having degrees $(\,i_1 - 1\,)$ and $(\,i_2 - 1\,)$. In this way, we can view box splines as an extension of tensor product splines. We present a more complete treatment of univariate splines at the end of this section. □

Example 2 Let $\mathcal{T} = \{e_1, e_2, e_1 + e_2\}$. Then, by Definition 1, the function $B(\,x \,|\, \mathcal{T})$ is the density of the vector



$$X = (\,Z_1 + Z_3\,, Z_2 + Z_3\,)\,,$$

where Z_1, Z_2, Z_3 each have the uniform distribution on the interval $[0, 1]$. After a little calculus we find that $B(\,x \,|\, \mathcal{T})$ is a piecewise linear function in x that interpolates the value one at the point (1,1) and zero at every other point in $\mathbb{Z}^2$. While this function is continuous on $\mathbb{R}^2$, it has jumps in its first partial derivatives along the grid lines indicated in the accompanying figure. This function is often called the Courant finite element. □

One of the most studied examples of bivariate box splines is the collection of those for which $\mathcal{T}$ is built from the vectors e_1, e_2, and $e_1 + e_2$. We refer to the partition or grid induced by this choice as the three directional mesh. It is constructed by drawing straight lines through each integer point in the plane in the directions e_1, e_2, and $e_1 + e_2$. For positive integers i_1, i_2, and i_3, let $B_{i_1 i_2 i_3}(x)$ denote the box spline $B(\,x \,|\, \mathcal{T})$ where $\mathcal{T}$ contains the vector e_1 repeated i_1 times, e_2 repeated i_2 times, and $e_1 + e_2$ repeated i_3 times.

Lemma 1.1 *For positive integers i_1, i_2, and i_3, the function $B_{i_1 i_2 i_3}(x)$ is a piecewise polynomial of degree at most $m = i_1 + i_2 + i_3 - 2$ relative to the three directional mesh. Furthermore, $B_{i_1 i_2 i_3}(x)$ is $(m - i_1)$ times continuously differentiable across each horizontal line in the mesh; $(m - i_2)$ times continuously differentiable across each vertical line; and $(m - i_3)$ times continuously differentiable across the each diagonal line.*

Example 3 Consider first the box spline $B_{112}(x)$ associated with the direction vectors

$$\mathcal{T} \;=\; \{\, e_1, e_2, e_1 + e_2, e_1 + e_2 \,\}\,.$$

In principle, we can work out an analytic expression for this function as we did for the Courant finite element in Example 1. A perspective plot of the function $B_{112}(x)$ is given in the left panel of Figure 1. Its support relative to the underlying

three directional mesh is indicated by the shaded triangles in the lower plot in the left panel. This region is just the affine cube $[\mathcal{T}]$ defined in (1.1). While it is not immediately obvious from this presentation, we know from Lemma 1.1 that $B_{112}(x)$ is a quadratic polynomial in each triangle. Furthermore, in addition to being globally continuous, $B_{112}(x)$ is continuously differentiable across each vertical and horizonal line of its grid. This fact can be verified by inspection of the contour lines in the lower plot of the left panel of Figure 1.

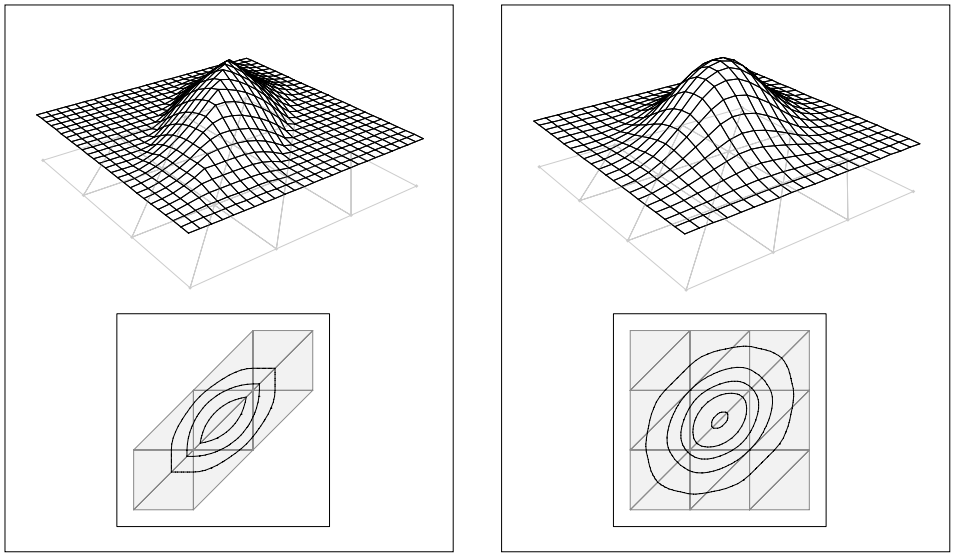


Figure 1. The box splines $B_{112}(x)$ and $B_{221}(x)$.

In the right panel of Figure 1, we present the same pair of plots, but for the function $B_{221}(x)$, the box spline associated with the the direction vectors

$$\mathcal{T} \;=\; \{\, e_1, e_1, e_2, e_2, e_1 + e_2 \,\}\,.$$

According to Lemma 1.1, the function $B_{221}(x)$ is a piecewise cubic polynomial and it is continuously differentiable. Additionally, across each of the diagonal lines in the mesh, it is twice continuously differentiable. This fact is difficult to verify by inspecting the contour lines in the lower righthand plot in Figure 1, so the adventurous reader is invited to derive the analytic expression for $B_{221}(x)$ and verify these facts directly. □

In Example 2, we considered quadratic and cubic box splines with direction sets built from the vectors e_1, e_2 and $e_1 + e_2$. As we will see in the following example,

by adding the vector $e_1 - e_2 = (1, -1)$ to our direction set, we can increase the smoothness of the resulting box splines without greatly increasing their degree. The resulting grid is referred to as a four directional mesh. It is obtained by drawing straight lines at each integer point in the plane in the directions e_1, e_2, $e_1 + e_2$, and $e_1 - e_2$. Furthermore, by analogy with the three directional case, given the positive integers i_1, i_2, i_3, and i_4, let $B_{i_1 i_2 i_3 i_4}(x)$ denote the box spline $B(\, x \,|\, \mathcal{T})$ where $\mathcal{T}$ contains the vector e_1 repeated i_1 times, e_2 repeated i_2 times, $e_1 + e_2$ repeated i_3 times, and $e_1 - e_2$ repeated i_4 times.

Lemma 1.2 *For positive integers i_1, i_2, i_3, and i_4, the function $B_{i_1 i_2 i_3 i_4}(x)$ is a piecewise polynomial of degree at most $m = i_1 + i_2 + i_3 + i_4 - 2$ relative to the four directional mesh. Furthermore, $B_{i_1 i_2 i_3 i_4}(x)$ is $(m - i_1)$ times continuously differentiable across each horizontal line in the mesh; $(m - i_2)$ times continuously differentiable across each vertical line; $(m - i_3)$ times continuously differentiable across the each diagonal line with positive slope; and $(m - i_4)$ times continuously differentiable across the each diagonal line with negative slope.*

Example 4 In the left panel of Figure 2 we present perspective and contour plots of the function $B_{1111}(x)$, the box spline associated with the vectors

$$\mathcal{T} \;=\; \{\, e_1, e_2, e_1 + e_2, e_1 - e_2 \,\}\,.$$

In the finite element literature, this function is referred to as the Zwart element. It is piecewise quadratic and continuously differentiable. Observe that while this function is smoother than the piecewise quadratic $B_{112}(x)$ studied in the previous example, its support is considerably larger. We will encounter this type of tradeoff frequently throughout this chapter.

In the right panel of Figure 2, we present the perspective and contour plots of $B_{2111}(x)$, the box spline associated with the direction vectors

$$\mathcal{T} \;=\; \{\, e_1, e_1, e_2, e_1 + e_2, e_1 - e_2 \,\}\,.$$

This function is piecewise cubic and is again continuously differentiable. Comparing the left and right panels of Figure 2, we see directly the effect that adding a single direction vector to a set $\mathcal{T}$ has on the resulting box spline. □

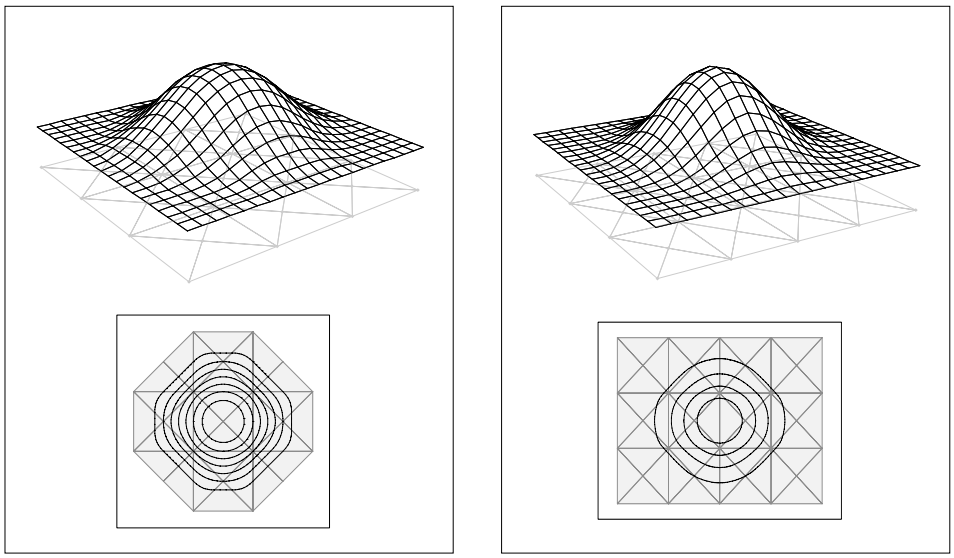


Figure 2. The box splines $B_{1111}(x)$ and $B_{2111}(x)$.

In the following lemma, we collect a number of facts about the functions $B(\,x\,|\,\mathcal{T}\,)$ that place the results of Examples 1-4 and Lemmas 1.1 and 1.2 in a broader mathematical context. Proofs of most of these statements can be found in the original paper by de Boor and Höllig (1982). Other excellent sources of information include the monograph by Chui (1988), the survey paper of Dæhlen and Lyche (1991), and the book by de Boor, Höllig, and Riemenschneider (1993).

Theorem 1.1 *Let $\mathcal{T}$ be a collection of vectors described in Definition 1.*

(a) *The function $B(\,x\,|\,\mathcal{T}\,)$ is a piecewise polynomial of total degree $\#(\,\mathcal{T}\,)-d$ relative to the triangulation whose grid is given by*

$$[\,\mathcal{T}'\,] \;+\; t^* \qquad \textit{for all } \mathcal{T}' \subset \mathcal{T} \textit{ with } \#(\,\mathcal{T}'\,) = d-1\,, \tag{1.3}$$

where t^* is the sum of any collection of the elements in $\mathcal{T}\setminus\mathcal{T}'$.

(b) *The function $B(\,x\,|\,\mathcal{T}\,)$ has $r(\,\mathcal{T}\,)$ continuous derivatives on $\mathbb{R}^d$, where*

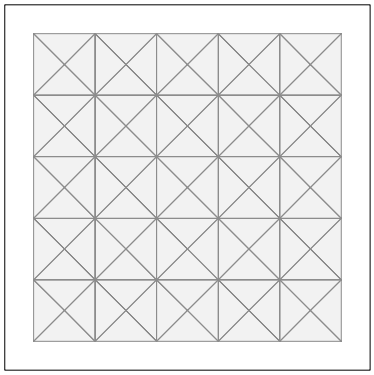
$$r(\mathcal{T}) \;=\; \min\,\big\{\,\#(\mathcal{T}') \,:\, \mathcal{T}'\subset\mathcal{T},\ \textit{the vectors in } \mathcal{T}\setminus\mathcal{T}' \textit{ do not span } \mathbb{R}^d\,\big\} - 2\,.$$

(c) *The functions $B(\,x-i\,|\,\mathcal{T}\,)$, $i\in\mathbb{Z}^d$, form a partition of unity. That is,*

$$\sum_{i\in\mathbb{Z}^d} B(\,x-i\,|\,\mathcal{T}\,) \;=\; 1\,, \qquad x\in\mathbb{R}^d\,.$$

(d) *The functions* $B(x-i\,|\,\mathcal{T})$, $i \in \mathbb{Z}^d$, *are linearly independent if and only if* $|\det \mathcal{T}'|$ *equals one for all subsets* $\mathcal{T}' \subset \mathcal{T}$ *whose vectors form a basis of* $\mathbb{R}^d$.

Example 5 Consider the square region $[0,5]\times[0.5]$ depicted in the accompanying figure. According to Theorem 1.1(d), the set of all functions of the form $B_{1111}(x-i)$, $i \in \mathbb{Z}^2$, that are nonzero zero over some portion of the interior of this region are linearly dependent. In fact, Chui and Wang (1984) demonstrate that for all points (x_1, x_2) in the square, the sum



$$\sum_{k_1,k_2} (-1)^{k_1+k_2} B_{1111}(x_1-k_1, x_2-k_2)$$

is identically zero. However, they also prove that if any one of these functions is removed, the resulting set is linearly independent, and provides a basis for the space of continuously differentiable, piecewise quadratic functions on the region. □

Rate of Approximation

Let $\mathcal{X}$ denote a compact region in $\mathbb{R}^d$ with a piecewise smooth boundary and let $\mathcal{T} \subset \mathbb{Z}^d$ denote a collection of direction vectors that contain a basis of $\mathbb{R}^d$. Then, from Theorem 1.1 we know that for each $i \in \mathbb{Z}^d$, $B(x-i\,|\,\mathcal{T})$ is actually a piecewise polynomial function of total degree at most $\#(\mathcal{T})-d$ relative to the partition defined in (1.3). Let $\triangle$ denote the collection of all elements in this partition whose closure has a non-empty intersection with $\mathcal{X}$. In the left panel of Figure 3, we present $\triangle$ for the four directional mesh and a circular region. Therefore, the span of the functions $B(x-i\,|\,\mathcal{T})$, $i \in \mathbb{Z}^d$, whose support contains some element of $\triangle$ forms a subspace of $\mathbb{PP}$, the collection of all piecewise polynomial functions of total degree $\#(\mathcal{T})-d$ relative to $\triangle$. Next, for any positive real number h, define the scaling operator $\sigma_h f$ which maps functions $f(x)$ into functions $f(x/h)$, and consider integer translates of the function $\sigma_h B(x\,|\,\mathcal{T})$. Observe that $B(x/h-i\,|\,\mathcal{T})$, $i \in \mathbb{Z}^d$, is again a piecewise polynomial function of total degree at most $\#(\mathcal{T})-d$. The underlying partition is obtained by applying the transformation $x \mapsto hx$ to the original grid specified in (1.3). In the right panel of Figure 3, we present $\triangle$ for the four directional mesh with $h = 1/2$ and the same circular region. Let $\triangle$ denote the collection of all elements in this scaled partition whose closure has a

non-empty intersection with $\mathcal{X}$. As was the case in Chapter 1, the dependence of $\triangle$ on h is not made explicit. Finally, the span of the functions $B(\,x/h - i\,|\,\mathcal{T})$, $i \in \mathbb{Z}^d$, whose support contains some element of $\triangle$ forms a subspace of $\mathbb{PP}$, the piecewise polynomial functions of total degree $\#(\,\mathcal{T}) - d$ relative to $\triangle$.

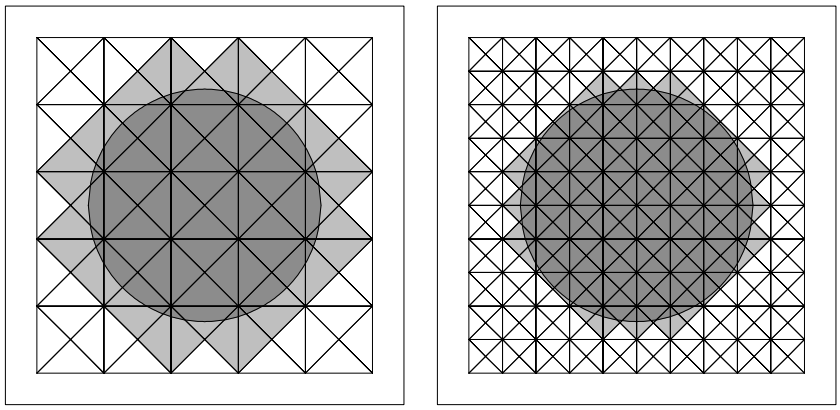


Figure 3. The partition $\triangle$ for a fixed circular region with $h = 1$ and $h = 1/2$.

Using Taylor expansions, it is plausible that the rate at which functions in $\mathbb{PP}$ can approximate a smooth funcition f defined on an open set containing $\mathcal{X}$, should be $r = \#(\,\mathcal{T}) - d + 1$. That is,

$$\inf_{p} \|\, f - p \,\|_{L_\infty(\mathcal{X})} \;=\; O\big(\, h^r \,\big)\,,$$

where the infinum is taken over all $p \in \mathbb{PP}$. Let $\mathbb{G}$ denote the space of functions $B(\,x/h - i\,|\,\mathcal{T})$, $i \in \mathbb{Z}^d$, whose support contains some element of $\triangle$. We now take up the task of determining when $\mathbb{G}$ achieves the same approximation rate as $\mathbb{PP}$. This is accomplished by exhibiting an approximation scheme that delivers the required rate. Abstractly, an approximation scheme is merely a linear operator that maps the space $L_\infty(\mathcal{X})$ of bounded functions on $\mathcal{X}$ into $\mathbb{G}$.

We say that a linear operator Q is bounded if the norm of Q is finite relative to the sup-norm on $\mathbb{R}^d$. We use the term local linear operator to mean a bounded linear operator Q whose domain and range are function spaces on $\mathbb{R}^d$ having the property that $(Qf)(x)$, depends only on the values of f in an open ball centered at x and having a fixed, finite radius independent of x. Likewise, a local linear functional λ has the property that λf depends only on the values of f in an open ball of finite radius centered at the origin. Finally, if a linear operator Q is such that $Qp = p$ for all polynomials of total degree $r - 1$ or less in the variable $x \in \mathbb{R}^d$, then we say that Q has polynomial order r.

The following definition and theorem are taken from the book by de Boor, Höllig, and Riemenschneider (1993). In general, they apply to operators that map into any space that is the span of functions of the form $\phi(\,x/\,h - i\,)$ for $i \in \mathbb{Z}^d$, where $\phi(\,\cdot\,)$ is a compactly supported, piecewise continuous function. For the moment, we restrict our attention to the case when $\phi(x) = B(\,x\,|\,\mathcal{T})$.

Definition 2 *A quasi-interpolant Q of order r for the space $\mathbb{G}$ is a local linear operator that maps $L_\infty(\mathbb{R}^d)$ into $\mathbb{G}$ and has polynomial order r.*

Given a set $\mathcal{T} \subset \mathbb{Z}^d$ that contains a basis of $\mathbb{R}^d$, we restrict our attention to quasi-interpolants of the form

$$Q_\lambda f \;=\; \sum_i B(\,\cdot - i\,|\,\mathcal{T})\;\lambda\, f(\,\cdot - i) \qquad \text{for all } x \in \mathcal{X}\,, \tag{1.4a}$$

where λ is some bounded linear functional and the sum is taken over all $i \in \mathbb{Z}^d$. For any positive real number h, we define the scaled approximation to f based on the operator $Q = Q_\lambda$ to be

$$\sigma_h\, Q\, \sigma_{1/h}\, f \;=\; \sum_i B(\,\cdot/h - i\,|\,\mathcal{T})\;\lambda\, f\big(\,h(\,\cdot - i)\,\big) \qquad \text{for all } x \in \mathcal{X}\,. \tag{1.4b}$$

By considering quasi-interpolants for which λ is allowed to vary with i, Chui and Diamond (1990) have produced operators that can adapt to the boundary of the set $\mathcal{X}$ over which f is defined. For simplicity, however, we restrict our attention to maps of the form (1.4). From a practical standpoint, this means that the function f we are approximating has to be defined on some open set containing $\mathcal{X}$.

Before proceeding with the next approximation result, we have to introduce more notation. First, let $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and $j = (j_1, \ldots, j_d) \in \mathbb{N}^d$ be as in Section 1.1, and given any smooth function $f = f(x)$, define

$$D^j f \;=\; \frac{\partial^{[\,j\,]}}{\partial x_1^{j_1} \cdots \partial x_d^{j_d}} f(\,x\,)\,.$$

Given a nonnegative integer r, we say that a function f is r-times differentiable if the quantity above exists for all vectors $j \le r$. Similarly, let $\mathbf{x} = (x_1, \ldots, x_M)$ and $\mathbf{j} = (j_1, \ldots, j_M)$ be as in (1.5) of Chapter 1, and given any smooth function $f = f(\mathbf{x})$ define

$$D^{\mathbf{j}} f \;=\; \frac{\partial^{[\,\mathbf{j}\,]}}{\partial x_1^{j_1} \cdots \partial x_M^{j_M}} f(\,\mathbf{x}\,)\,.$$

Given a vector $r \in \mathbb{N}^M$, we say that a function $f(\mathbf{x})$ is r-times differentiable if the quantity above exists for all vectors $\mathbf{j}$ satisfying $\mathbf{j} \le r$.

Theorem 1.2 *Suppose that λ is a local linear functional and that the associated quasi-interpolant $Q = Q_\lambda$ defined in* (1.4a) *has polynomial order r. Then for any smooth function f defined on an open set containing $\mathcal{X}$, there exists a constant T_1 that does not depend on f or h such that*

$$\left\| \sigma_h \, Q \, \sigma_{1/h} \, f - f \right\|_{L_\infty(\mathcal{X})} \;\leq\; T_1 \, h^r \sum_{[j]=r} \left\| D^j f \right\|_{L_\infty(\mathcal{X})} .$$

Theoretically, the most attractive functionals λ are those which involve only function evaluations. In the following example, we introduce one such quasi-interpolant.

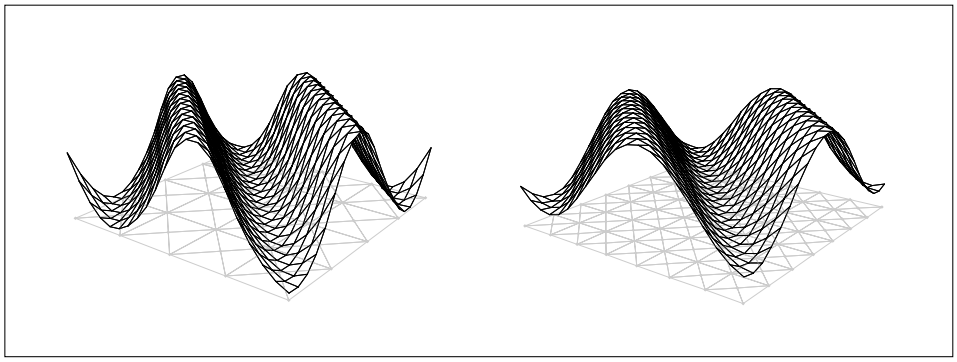
Example 6 Consider again the Zwart element $B_{1111}(x)$, the box spline with direction set

$$\mathcal{T} = \{ e_1, e_2, e_1 + e_2, e_1 - e_2 \} .$$

Recall that it is piecewise quadratic and continuously differentiable. It can be shown that the quasi-interpolant $Q = Q_\lambda$ associated with the linear functional

$$\lambda f \;=\; 2 f(3/2, 3/2) \;-\; \big(f(1,1) + f(2,1) + f(1,2) + f(2,2) \big) / 4$$

reproduces all polynomials of total degree 2, and hence has polynomial order 3. Therefore, from Theorem 1.2, we find that the space spanned by the functions $B_{111}(x/h - i)$, $i \in \mathbb{Z}^d$, can approximate smooth functions at the rate h^3. Since $B(x \,|\, \mathcal{T})$ is piecewise quadratic, the associated piecewise polynomial space $\mathbb{PP}$ also approximates smooth functions at the rate h^3. In the following figure, we present a perspective plot of the approximation Qf for $f(x_1, x_2) = \sin\big(10x_1 - 10x_2\big)$ taking taking $h = 1$ and $h = 1/2$.. In each case, we have included the underlying 4 directional grid.



It is possible that the spaces $\mathbb{G}$ and $\mathbb{PP}$ do not attain the same approximation rate for smooth functions. In the following theorem, we characterize the approximation power of the scaled translates of the function $B(x \mid \mathcal{T})$ in terms of its direction set $\mathcal{T}$. The abstract construction of quasi-interpolants for any space of box splines can be found in de Boor and Höllig (1982).

Theorem 1.3 *Let $\mathcal{T} \subset \mathbb{Z}^d$ contain a basis for $\mathbb{R}^d$, and set $r = r(\mathcal{T}) + 2$. Then, there exists a quasi-interpolant Q such that for any function f defined on an open set containing $\mathcal{X}$ and possessing r continuous derivatives,*

$$\left\| \sigma_h \, Q \, \sigma_{1/h} \, f - f \right\|_{L_\infty(\mathcal{X})} \;=\; O\big(h^r\big) \, .$$

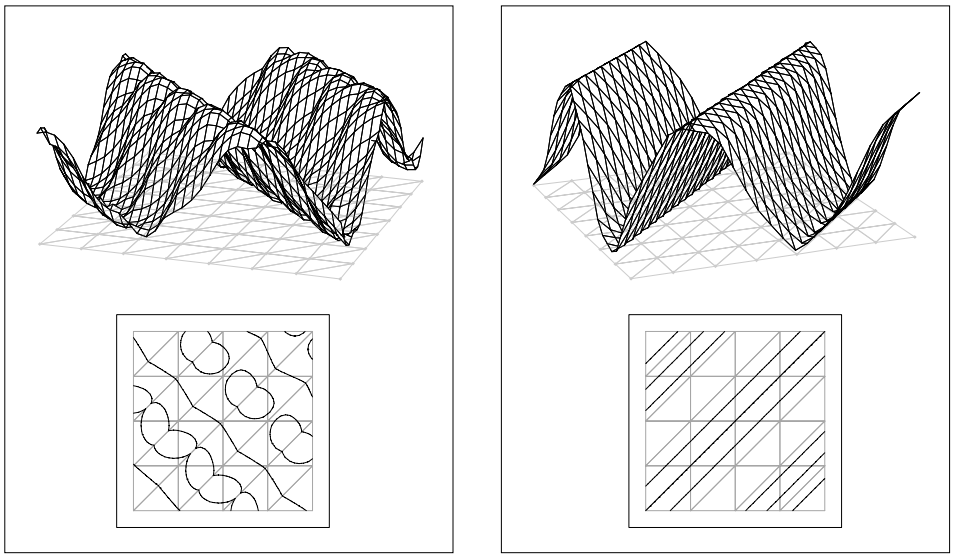
Example 7 Consider the function $B_{112}(x)$, the box spline with direction set

$$\mathcal{T} \;=\; \{ \, e_1, e_2, e_1 + e_2 \, \}$$

introduced in Example 3. Like the Zwart element, it is piecewise quadratic, but it has jumps in its first derivative across the diagonal lines in the underlying three directional mesh. In addition, for this choice of $\mathcal{T}$, $r(\mathcal{T}) = 0$, and hence the space of scaled integer translates of $B_{112}(x)$ can approximate smooth functions only at the rate $r = 2$, the same rate obtainable for piecewise linear functions. We can construct a quasi-interpolant $Q = Q_\lambda$ that reproduces all linear functions by choosing

$$\lambda f \;=\; 3 f(1,1) / 2 \;-\; f(0,0) / 2 \, .$$

In the left panel of the following figure, we present both a perspective and a partial contour plot of the approximation Qf for $f = \sin(10x + 10y)$. In each plot, we have included the underlying three directional mesh for reference. Because f varies considerably along the diagonal mesh lines, the approximation Qf exhibits a number of spurious peaks along the crests of f. In the right panel of this figure, we present similar plots for the function Qf where $f = \sin(10x - 10y)$. Not surprisingly, the approximation Qf has improved dramatically. The visible discontinuities in the first derivative of Qf along the crests of f occur because the underlying function $B_{112}(x)$ has a jump in its first derivative across the diagonal mesh lines.



Tensor Products of Box Splines

In this section we return to the notation of Chapter 1. For each $1 \leq l \leq M$, let $\mathcal{X}_l$ denote a compact set in $\mathbb{R}^{d_1}$. Next, let $\mathcal{T}_l$ denote a collection of vectors in $\mathbb{Z}^{d_l}$ that contains a basis of $\mathbb{R}^{d_l}$. As we have seen, the collections $\mathcal{T}_l$ define a box spline $B(\, x_l \,|\, \mathcal{T}_l \,)$, which is a piecewise polynomial function of total degree $\#(\, \mathcal{T}_l \,) - d_l$ in the variable $x_l \in \mathbb{R}^{d_l}$. In addition, recall from the discussion prior to Figure 3 that, for a positive real number h_l, each set of direction vectors $\mathcal{T}_l$ induces a partition $\triangle_l$ of the set $\mathcal{X}_l$. Finally, let $\mathbb{G}_l$ denote the span of the functions $B(\, x_l / h_l - i_l \,|\, \mathcal{T}_l \,)$, $i_l \in \mathbb{Z}^{d_l}$, whose support contains at least one of the elements of $\triangle_l$, and set

$$\mathbb{G} \;\; = \;\; \mathbb{G}_1 \otimes \cdots \otimes \mathbb{G}_M \, .$$

The functions in $\mathbb{G}$ are defined on the set $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_M$. Recall from Chapter 1 that the individual partitions $\triangle_l$ induce a partition $\triangle = \triangle_1 \times \cdots \triangle_M$ on $\mathcal{X}$. Therefore, the members of $\mathbb{G}$ are contained in the space of all piecewise polynomial functions of coordinate degree

$$m \;\; = \;\; \big(\#(\, \mathcal{T}_1 \,) - d_1, \ldots, \#(\, \mathcal{T}_M \,) - d_M \,\big)$$

in the variable $\mathbf{x} = (x_1, \ldots, x_M)$. It is worth noting that the functions

$$B(\, x_1 \,|\, \mathcal{T}_1 \,) \cdots B(\, x_M \,|\, \mathcal{T}_M \,)$$

that span $\mathbb{G}$ are themselves box splines. To see this, for each $1 \leq l \leq M$, let $\mathcal{T}_l$ be given by $\{t_{l1}, \ldots, t_{lJ_l}\}$ and set

$$\mathbf{t}_{lj} \;=\; \big(0, \ldots, t_{lj}, \ldots, 0\big) \qquad \text{for } 1 \leq j \leq J_l\,.$$

That is $\mathbf{t}_{lj} = (t_1, \ldots, t_M)$ is a vector in $[d]$-dimensional space whose only non-zero component is the vector $t_l \in \mathbb{Z}^{d_l}$ which is equal to t_{lj}. Then, if we let $\mathcal{T}$ denote the collection of all such direction vectors, we find from Definition 1 that $B(\,\mathbf{x}\,|\,\mathcal{T})$ is the density of the random variable

$$\mathbf{X} \;=\; \mathbf{t}_1 Z_1 + \cdots + \mathbf{t}_J Z_J$$

where $J = J_1 + \cdots + J_M$, and each $Z_1, \ldots, Z_J$ are independent and identically distributed uniformly over the interval $[0, 1]$. (It is important that the reader does not confuse this random variable with the predictor variable $\mathbf{X}$ discussed in Chapter 1.) Arguing by induction, it is clear from the construction of $\mathcal{T}$ and Definition 1 that

$$B(\,\mathbf{x}\,|\,\mathcal{T}) \;=\; B(\,x_1\,|\,\mathcal{T}_1\,) \cdots B(\,x_M\,|\,\mathcal{T}_M\,)\,,$$

and that

$$r(\,\mathcal{T}) \;=\; \min\big\{\, r(\,\mathcal{T}_1\,), \ldots, r(\,\mathcal{T}_M\,)\,\big\}.$$

Therefore, from Theorem 1.3 we know that generically the rate of approximation obtainable by the span of the functions $B(\,\mathbf{x}/h - \mathbf{i}\,|\,\mathcal{T})$, where $\mathbf{i} = (i_1, \ldots, i_M)$ and each $i_l \in \mathbb{Z}^{d_l}$, is $r(\,\mathcal{T}) + 2$, or rather, the minimum of the rates of the spaces $\mathbb{G}_1, \ldots, \mathbb{G}_M$ when we take $h_l = h$ for some positive scalar h and all $1 \leq l \leq M$. By letting the scale h_l vary for each space $\mathbb{G}_l$, however, we can obtain a slightly more informative bound.

Toward this end, for each space $\mathbb{G}_l$ let Q_l denote a quasi-interpolant that maps bounded functions defined on $\mathcal{X}_l$ into the space $\mathbb{G}_l$ as described in Theorem 1.3. Then, if f is a function defined on $\mathcal{X}$, we interpret $Q_l f$ to mean that Q_l is applied to the function f, holding the remaining variables $x_{l'}$, $l' \neq l$, fixed. Furthermore, for $h = (h_1, \ldots, h_M)$, define

$$\sigma_h f(\mathbf{x}) \;=\; f\big(\,x_1/h_1, \ldots, x_M/h_M\,\big)\,.$$

Finally, by setting $Q = Q_1 \times \cdots \times Q_M$ we have the following lemma.

Lemma 1.3 *Let f be any smooth function defined on an open set containing $\mathcal{X}$. Then there exists a constant T_1 that does not depend on f or h such that*

$$\big\|\,\sigma_h\, Q\, \sigma_{1/h}\, f - f\,\big\|_{L_\infty(\mathcal{X})} \;\leq\; T_1 \sum_{l=1}^{M} \big\|\,\sigma_{h_l}\, Q_l\, \sigma_{1/h_l}\, f - f\,\big\|_{L_\infty(\mathcal{X})}\,.$$

Proof Suppose initially that $M = 2$ and $h = (1, 1)$. Then, we observe that

$$\begin{aligned} Qf - f &= Q_1 Q_2 f - f \\ &= Q_1 Q_2 f - Q_1 f + Q_1 f - f \\ &= Q_1 \big(Q_2 f - f \big) + \big(Q_1 f - f \big) . \end{aligned}$$

Since f and hence $Q_2 f$ are bounded for all $x_1 \in \mathcal{X}_1$, we find that

$$\begin{aligned} \big| Qf - f \big| &\le \big| Q_1 f - f \big| + \sup_{x_1 \in \mathcal{X}_1} \Big| Q_1 \big(Q_2 f - f \big) \Big| \\ &\le \big| Q_1 f - f \big| + \big\| Q_1 \big\| \sup_{x_1 \in \mathcal{X}_1} \big| Q_2 f - f \big| , \end{aligned}$$

so that

$$\big\| Qf - f \big\|_\infty \le \big\| Q_1 f - f \big\|_\infty + \big\| Q_1 \big\| \big\| Q_2 f - f \big\|_\infty .$$

The argument for general $h = (h_1, h_2)$ is similar, and the proof for general M follows by induction. □

Therefore, if we set $r_l = r(\mathcal{T}_l) + 2$ for $1 \le l \le M$, then by combining Theorem 1.3 and Lemma 1.3, we find that

$$\big\| \sigma_h \, Q \, \sigma_{1/h} \, f - f \big\|_{L_\infty(\mathcal{X})} \le O\big(h_1^{r_1} + \cdots + h_M^{r_M} \big) .$$

Note that if we take $h_1 = \cdots = h_M$, then the approximation rate from functions in $\mathbb{G}$ is just the minimum of $r_1, \ldots, r_M$, as we observed previously. To translate these results into the terminology of Chapter 1, let $s = \{s_1, \ldots, s_k\}$ be some subset of $\{1, \ldots, M\}$, and recall the definitions

$$\mathcal{X}_s = \mathcal{X}_{s_1} \times \cdots \times \mathcal{X}_{s_k} \quad \text{and} \quad \mathbb{G}_s = \mathbb{G}_{s_1} \otimes \cdots \otimes \mathbb{G}_{s_k}$$

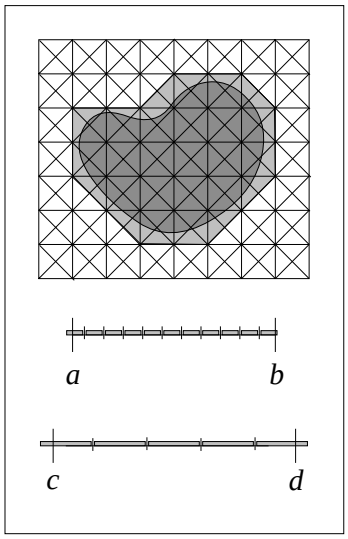
where for convenience we assume that $s_1 < \cdots < s_k$. Then, by applying Theorem 1.3 and Lemma 1.3 to the space $\mathbb{G}_s$, we have that

$$\inf_g \big\| g - f \big\|_{L_\infty(\mathcal{X}_s)} = O\big(h_{s_1}^{r_{s_1}} + \cdots + h_{s_k}^{r_{s_k}} \big) ,$$

where the infinum is taken over all functions $g \in \mathbb{G}_s$. Therefore, by setting $h_s = (h_{s_1}, \ldots, h_{s_k})$, the function $\rho_s(\cdot)$ appearing in Condition 5′ is given explicitly by

$$\rho_s(h_s) = h_{s_1}^{r_{s_1}} + \cdots + h_{s_k}^{r_{s_k}} .$$

Example 8 Suppose we have three covariates X_1, X_2, and X_3. We assume that X_1 ranges over some compact region in $\mathbb{R}^2$ having a smooth boundary, and that X_2 and X_3 range over the intervals $[a,b]$ and $[c,d]$ respectively. For some scalar $h_1 > 0$, let $\mathbb{G}_1$ denote the span of the functions $B_{1111}(\, x_1/h_1 - i_1\,)$, $i_1 \in \mathbb{Z}^2$, whose support has a non-empty intersection with the interior of $\mathcal{X}_1$. Let $B_2(\,\cdot\,)$ and $B_3(\,\cdot\,)$ denote the univariate cardinal B-splines of degree 2 and 3, respectively, introduced in Example 1. Then, for some scalar $h_2 > 0$, let $\mathbb{G}_2$ be the span of the functions $B_2(\, x_2/h_2 - i_2\,)$, $i_2 \in \mathbb{Z}$, whose support has a non-empty intersection with $\mathcal{X}_2 = [a,b]$; and let $\mathbb{G}_3$ be the span of the functions $B_3(\, x_3/h_3 - i_3\,)$, $i_3 \in \mathbb{Z}$, whose support has a non-empty intersection with $\mathcal{X}_3 = [c,d]$. In the accompanying figure, we present $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{X}_3$ as well as the underlying partitions $\triangle_1$, $\triangle_2$ and $\triangle_3$ for some value of $h = (h_1, h_2, h_3)$.



Suppose that we are interested in modeling the main effects of each covariate as well as an interaction between X_1 and X_2. Using the notation of Chapter 1, for this model the hierarchical collection $\mathcal{S}$ of subsets of $\{1,2,3\}$ is given by

$$\mathcal{S} = \{\, \varnothing, \{\,1\,\}, \{\,2\,\}, \{\,3\,\}, \{\,1,2\,\}\,\}\,.$$

If $s = \varnothing$, then $\rho_{\varnothing}(\,\cdot\,)$ is identically zero, while for any non-empty set $s \in \mathcal{S}$, the function $\rho_s(\, h_s\,)$ is given by

$$\rho_s(\, h_s\,) \;=\; \sum_{l \in s} h_l^{r_l}\,, \qquad \text{where } r_1 = r_2 = 3, \text{ and } r_3 = 4\,.$$

Setting $d = (2,1,1)$, we know from Theorem 5.1 that if $h^{-d} = h_1^{-2} h_2^{-1} h_3^{-1}$ is $o(\, n^{1-\epsilon}\,)$ for some $\epsilon > 0$, then

$$\| \,\hat{\mu} - \mu^*\,\|^2 \;=\; O_P\Big(\, h_1^3 + h_2^3 + h_3^4 + h_1^{-2} h_2^{-1}/n + h_3^{-1}/n\,\Big)\,,$$

where $\hat{\mu}$ is the orthogonal projection onto the space

$$\mathbb{G} \;=\; \left\{\, \sum_{s \in \mathcal{S}} g_s \,:\, g_s \in \mathbb{G}_s \,\right\}\,,$$

relative to the empirical inner product (2.15b).

Univariate B-splines

Suppose that $\mathcal{X}$ is the interval $[0,1]$, and let $\triangle$ denote the collection of intervals

$$\triangle \;=\; \Big\{ \big(k \,/\, K, (k+1) \,/\, K \big) : k = 0, \ldots, K-1 \Big\}$$

where K is a positive integer. According to Example 1, if we let $B(x)$ denote the density of the random variable

$$X = Z_1 + \cdots Z_{m+1}$$

where the random variables $Z_1, \ldots, Z_{m+1}$ are independent and identically distributed uniformly over the interval $[0,1]$, then $B(x)$ is a piecewise polynomial of degree m and is globally $m-1$ times continuously differentiable. Furthermore, the only discontinuities in its $(m-1)$st derivative occur at the points $0, \ldots, m$.

Suppose instead that the intervals in $\triangle$ are given by

$$\triangle \;=\; \Big\{ \big(t_k, t_{k+1} \big) : k = 0, \ldots, K-1 \Big\}$$

where $t_k < t_{k+1}$ for $0 \leq k \leq K-1$. Consider the space of piecewise polynomials of degree at most m that possess $r_k < m$ continuous derivatives at the point t_k, $1 \leq k \leq K-2$. Using a construction first discussed by Curry and Schoenberg (1966) and later extended by de Boor (1976), we can again derive a basis for this space from marginal density functions. In this case, however, we construct vectors that are uniformy distributed in high dimensional simplexes rather than boxes. The resulting functions are referred to as univariate B-splines. Ironically, the observation that univariate B-splines can be obtained in this manner led to considerable advances in multivariate splines in the mid-1980's.

The approximation rate obtainable from spaces of univariate splines and their tensor products again depends on the existence of quasi-interpolants, and can be derived using the techniques of the previous subsection. The interested reader is referred to the book by Schumaker (1981) for more details. When quoting approximation results, Stone (1985, 1986, 1991ab, 1994) uses a more refined form of the approximation rate that involves the modulus of continuity. This rate is established using an equivalence derived in Johnen and Scherer (1977), and once again the interested reader is referred to Schumaker (1981). Because of this more elegant rate, Stone's smoothness conditions are worded in terms of Hölder conditions on various derivatives of the function being estimated.

2.2 Finite Elements and B-nets

In Section 2.1, we considered a class of splines that were obtained as integer translates of a (multivariate) density function. As we have seen, the resulting partition $\triangle$ is extremely regular and is in fact determined by the structure of the density function used to generate the space. In this section, we take a slightly different approach and consider spline spaces that are defined relative to a given partition $\triangle$. Unlike box splines, we will be able to specify to a certain extent both the degree and the global smoothness of the resulting spline functions.

Piecewise Polynomials

By first deriving a generalized Taylor's expansion, Dupont and Scott (1980) establish a very general result concerning the rate of approximation obtainable from spaces of polynomials defined over compact subsets of $\mathbb{R}^d$, where d is some positive integer. To apply their results, we must strenghten Condition 1 slightly. To be more precise, for $1 \leq l \leq M$, assume that each set $\delta \in \triangle_l$ contains an open ball such that any line connecting a point in the ball to another point in δ lies entirely within δ. A set satisfying this property is said to be star-shaped. Now, using the partitions $\triangle_1, \ldots, \triangle_M$, define $\triangle$ as usual and choose $\delta \in \triangle$. Temporarily, set $h = \text{diam}\,\delta = (h_1, \ldots, h_M)$. Then, using the techniques of Dupont and Scott, we can show that given $m \in \mathbb{N}^M$ and a sufficiently smooth function $f = f(\mathbf{x})$ defined on $\mathcal{X}$, there exists a positive constant T_1 independent of h such that

$$\inf_p \left\| f - p \right\|_{L_\infty(\delta)} \;\leq\; T_1 \left(h_1^{m_1+1} + \cdots + h_M^{m_M+1} \right),$$

where the infinum is taken over all polynomials of coordinate degree $m \in \mathbb{N}^M$. Here, a sufficiently smooth function f is one possessing bounded derivatives of the form $D^{\mathbf{j}_l}$ where for $1 \leq l \leq M$, the only nonzero entries of the vector $\mathbf{j}_l = (j_1, \ldots, j_M)$ occur in j_l, and these satisfy $[j_l] = m_l + 1$. Unfortunately, the constant T_1 above depends on the geometry of the set δ. For most finite element applications, however, each element of the partition $\triangle$, is the affine transformation of some reference set. In this case, the dependence of T_1 on the geometry of δ is unimportant, and the approximation result above can be made to hold uniformly across the sets in $\triangle$. For further discussion of these issues, the reader is referred to Dupont and Scott (1980) and chapter 13 of the book by Schumaker (1981). We now take up a more detailed discussion of finite element spaces.

Finite Elements

For the remainder of this chapter, we restrict our attention to spline spaces defined over a compact region $\mathcal{X}$ of $\mathbb{R}^2$, and let $z = (x, y)$ denote an arbitrary point in $\mathbb{R}^2$. Of course, the results derived in this section can be extended to higher dimensions, but only with a corresponding increase in notational complexity. As mentioned in the previous subsection, the partitions $\triangle$ of $\mathcal{X}$ used in finite element applications are typically generated by affine transformations of a single reference set. To further simplify our presentation, we will confine our discussion to partitions made up of triangles. In this case, we say that $\triangle$ is a triangulation of $\mathcal{X}$. Throughout our discussion, we insist that $\triangle$ consist of triangles such that the vertices of each triangle are not contained in the interior of an edge of any other triangle. Such a triangulation is said to be conforming.

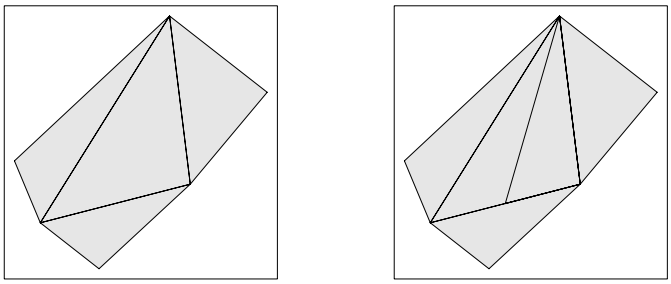


Figure 1. Conforming and non-conforming triangulations of a polyhedral region.

To quote Oden and Reddy (1976), "The finite-element method is, first, a systematic and very powerful method of interpolation." For ease of exposition, we consider here only Lagrange interpolation over triangles. Let m be a positive integer and let δ be a triangle in the plane. We say that a set of points in δ is m-unisolvent if a polynomial of total degree m is uniquely determined by its values on the set. In the Figure 2, we illustrate a key correspondence between the total degree of a polynomial and a set of unisolvent points in a triangle. The exact placement of these nodes will be discussed in the next subsection.

For a fixed positive integer m, define the interpolation operator that takes any function f defined on δ into the unique polynomial p that agrees with f at the nodal points described in Figure 2. Then, using the generalized Taylor's expansion discussed in the previous subsection, we can derive the approximation rate obtainable from the space of polynomials of total degree m on δ. Repeating this process for each $\delta \in \triangle$, we can derive approximation rates similar to those described in the

context of integer translates of a single box spline. Furthermore, by considering the more general spaces of polynomials introduced in Section 1.1, we can repeat the analysis given in the previous section for tensor products of box splines and establish rates of approximation for models involving spaces of Lagrange finite elements over triangles. Having indicated how these results could be obtained, we now leave the approximation properties of finite element interpolations and instead consider aspects of these spaces that make them attractive from a methodological standpoint.

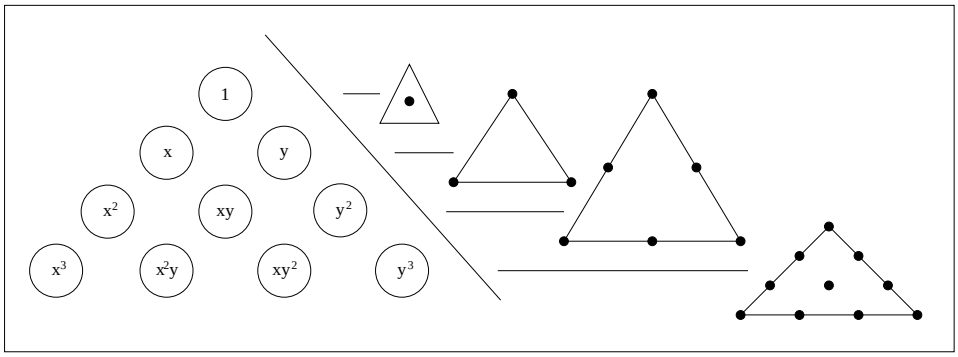
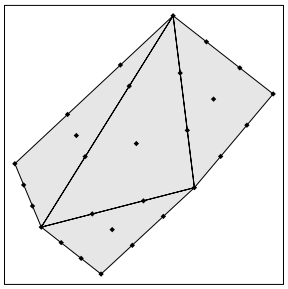


Figure 2. Pascal's triangle and associated unisolvent sets.

In the accompanying figure, we present a triangulation $\triangle$ of a polyhedral region $\mathcal{X}$ along with the set of 3-unisolvent node points associated with each triangle. Observe that by construction, there are exactly 4 nodes on each edge of the triangles in $\triangle$. Hence, the nodes appearing on a given edge in the triangulation are 3-unisolvent for the univariate polynomial defined on that edge. Therefore, we find that given any function f defined on $\mathcal{X}$, the interpolant that is constructed by matching the values of f at the node points in each triangle in $\triangle$ is in fact a continuous function on $\mathcal{X}$. It is possible to create interpolants having greater smoothness by incorporating information about the derivatives of f at certain node points. For a more complete discussion of these issues, the reader is referred to Oden and Reddy (1976), Ciarlet (1978) and Schwarz (1980).



We can use the Lagrange interpolation scheme described above to construct a basis for the space of continuous piecewise polynomial functions of total degree

m on $\mathcal{X}$. Each such basis function corresponds to interpolating the value of one at a single node, and zero at the remaining nodes. This process is repeated for each node in the triangulation $\triangle$. In the next subsection, we discuss an alternate basis for this space that has become quite popular among approximation theorists and computer scientists. As we will see, part of the success of this alternate basis stems from the fact that smoothness constraints across edges in $\triangle$ can be described simply and in a geometrically significant manner.

B-nets and Multivariate Splines

Let $\delta \in \triangle$ be a triangle in the plane with verticies $v_1 = (x_1, y_1)$, $v_2 = (x_2, y_2)$, and $v_3 = (x_3, y_3)$. For the moment, let $z = (x, y)$ denote an arbitrary point in $\mathbb{R}^2$, and define the functions $\phi_1(\,z\,)$, $\phi_2(\,z\,)$, and $\phi_3(\,z\,)$ such that

$$x \;=\; x_1\,\phi_1(\,z\,) \;+\; x_2\,\phi_2(\,z\,) \;+\; x_3\,\phi_3(\,z\,)\,,$$

$$y \;=\; y_1\,\phi_1(\,z\,) \;+\; y_2\,\phi_2(\,z\,) \;+\; y_3\,\phi_3(\,z\,)\,,$$

and

$$1 \;=\; \phi_1(\,z\,) \;+\; \phi_2(\,z\,) \;+\; \phi_3(\,z\,)\,,$$

for all $z \in \mathbb{R}^2$. We refer to triple $\big(\phi_1(\,z\,), \phi_2(\,z\,), \phi_3(\,z\,)\big)$ as the barycentric coordinate of z with respect to the triangle δ.

It is not hard to demonstrate that, for $1 \le i \le 3$, the function $\phi_i(\,z\,)$ is in fact that linear function in x and y which takes on the value one at the point v_i and zero at the two remaining verticies of δ. Observe that if $z \in \delta$, then each of these funcitons is nonnegative. Moreover, it can be shown that, for $j = (j_1, j_2, j_3)$, the functions

$$\phi^j(\,z\,) \;=\; \frac{(j_1 \,+\, j_2 \,+\, j_3)!}{j_1!\,j_2!\,j_3!}\;\phi_1^{j_1}(\,z\,)\;\phi_2^{j_2}(\,z\,)\;\phi_3^{j_3}(\,z\,)\,, \qquad [\,j\,] \,=\, m \;,\, z \in \delta\,, \quad (2.1)$$

are a basis for the space of polynomials of total degree m on δ. By construction these basis functions are naturally associated with points in δ; that is, given a vector $j = (j_1, j_2, j_3)$ satisfying $[\,j\,] = m$, the point having barycentric coordinates $(j_1/m, j_2/m, j_3/m)$ lies in δ. These points are in fact exactly the m-unisolvent points given in Figure 2 above. Therefore, we arrive once again at a correspondence between polynomials of total order m and the set of points given in Figure 2. In this case however, we represent the correspondence through the use of the so called Bezier- or B-net. To be more precise, given any polynomial p of total degree m on

δ, we can write

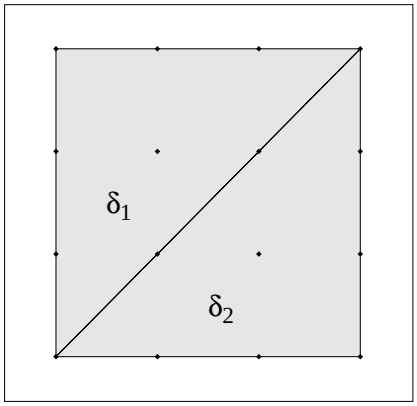
$$p(z) \;=\; \sum_{[\,j\,]\,=\,m} \beta_j\, \phi^j(z)\,, \qquad \text{for } z \in \mathbb{R}^2\,. \tag{2.2}$$

We then refer to the collection of points

$$\big(\, j_1/m,\, j_2/m,\, j_2/m,\, \beta_j \,\big)\,, \qquad \text{for } j = (j_1, j_2, j_3) \text{ with } [\,j\,] \;=\; m\,, \tag{2.3}$$

as the B-net associated with the polynomial p. While we have derived this representation with respect to a single triangle $\delta \in \triangle$, we can easily extend these definitions to construct the B-net associated with a piecewise polynomial function defined on $\mathcal{X}$ relative to the triangulation $\triangle$.

In fact, more is possible. Observe that if two triangles $\delta_1, \delta_2 \in \triangle$ share a common edge, then for any positive integer m, the m-unisolvent points in each triangle agree on the common edge. The following figure demonstrates the case when $m = 3$.



As with a single triangle, each point in the above figure corresponds to a basis function. If a point lies on an edge of $\triangle$ separating δ_1 and δ_2 then we take the associated basis function to be given by (2.1) with respect to δ_1, for $z \in \delta_1$, and by (2.1) with respect to δ_2, for $z \in \delta_2$. Arguing as in the subsection on finite element interpolations, we find that the resulting basis functions are in fact continuous on $\mathcal{X}$. Therefore, by assigning values to the coefficients represented in the B-net above, we obtain a continuous, piecewise polynomial function on $\mathcal{X}$.

One of the major reasons for introducing the B-net of a triangulation is the ease with which higher order smoothness constraints can be enforced across the edges of $\triangle$. It can be shown that these smoothness constraints are simple linear contrasts of the B-net coefficients. For an excellent account of these and many other properties of this representation, the reader is referred to Farin (1987), Goodman (1987ab), de Boor (1987), and Chui (1988).

Let m and r nonnegative integers, and let $\triangle$ be a conforming triangulation of a polyhedral region $\mathcal{X}$. Using the smoothness constraints described above, many authors have constructed locally supported basis functions for the space of piecewise polynomials of total degree m possessing r continuous derivatives on $\mathcal{X}$. Typically, r must be taken much smaller than m for these constructions to work. For example, Chui and Lai (1985, 1987, 1990) assume that m is not smaller than $4r+1$, while by considerably more complicated arguments, de Boor and Höllig (1988) and de Boor (1989) are able to reduce this lower bound to $3r+2$. For additional constructions, the reader is referred to Alfeld, Piper, and Schumaker (1987ab), Schumaker (1989) and Ibrahim and Schumaker (1991).

The restrictions imposed on the relationship between m and r by these constructions are fairly severe from a modelling standpoint. It appears that we are forced to consider high degree polynomials and fairly high dimensional spaces. Fortunately, there is a way around these problems. By subdividing each triangle $\triangle$ in a symmetric way and applying the smoothness constraints discussed above, Powell and Sabin (1977) and later Chui and He (1990) were able to construct smooth, low degree piecewise polynomial functions relative to a triangulation $\triangle$ of $\mathcal{X}$. In these cases, the term "piecewise" refers to the subdivided triangulation. In Figure 3, we present an original triangulation $\triangle$ and its refinement following the scheme of Chui and He (1990).

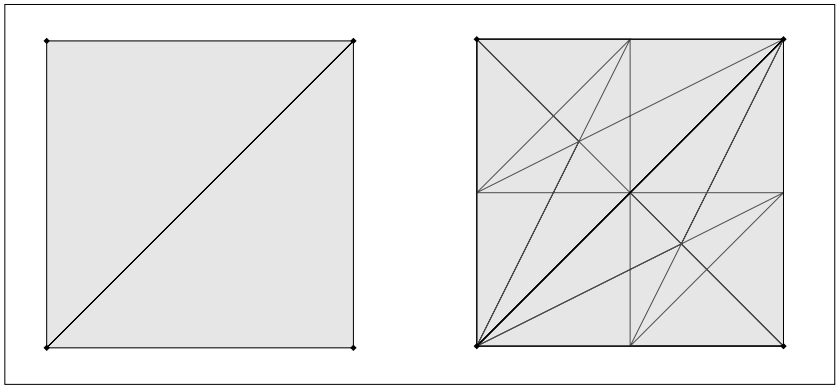


Figure 3. A triangulation and its symmetric refinement.

The spline spaces built from various symmetric subdivisions like the one presented above are powerful tools from which practically useful methodology can be built. Researchers have only just begun to investigate the potential of these spaces. For example, Dierckx, Van Leemput, and Vermeire (1992) present a sur-

face fitting routine based on adaptive triangulations and the splines of Powell and Sabin (1977). For the most part, however, these investigations are being conducted by numerical analysists. Given out experience with non-parametric procedures, we feel that statisticians can make significant contributions to this growing area.

2.3 Finite Elements and B-nets

In Section 2.1, we considered a class of splines that were obtained as integer translates of a (multivariate) density function. As we have seen, the resulting partition $\triangle$ is extremely regular and is in fact determined by the structure of the density function used to generate the space. In this section, we take a slightly different approach and consider spline spaces that are defined relative to a given partition $\triangle$. Unlike box splines, we will be able to specify to a certain extent both the degree and the global smoothness of the resulting spline functions.

Piecewise Polynomials

By first deriving a generalized Taylor's expansion, Dupont and Scott (1980) establish a very general result concerning the rate of approximation obtainable from spaces of polynomials defined over compact subsets of $\mathbb{R}^d$, where d is some positive integer. To apply their results, we must strenghten Condition 1 slightly. To be more precise, for $1 \leq l \leq M$, assume that each set $\delta \in \triangle_l$ contains an open ball such that any line connecting a point in the ball to another point in δ lies entirely within δ. A set satisfying this property is said to be star-shaped. Now, using the partitions $\triangle_1, \ldots, \triangle_M$, define $\triangle$ as usual and choose $\delta \in \triangle$. Temporarily, set $h = \text{diam}\,\delta = (h_1, \ldots, h_M)$. Then, using the techniques of Dupont and Scott, we can show that given $m \in \mathbb{N}^M$ and a sufficiently smooth function $f = f(\mathbf{x})$ defined on $\mathcal{X}$, there exists a positive constant T_1 independent of h such that

$$\inf_p \left\| f - p \right\|_{L_\infty(\delta)} \leq T_1 \left(h_1^{m_1+1} + \cdots + h_M^{m_M+1} \right),$$

where the infinum is taken over all polynomials of coordinate degree $m \in \mathbb{N}^M$. Here, a sufficiently smooth function f is one possessing bounded derivatives of the form $D^{\mathbf{j}_l}$ where for $1 \leq l \leq M$, the only nonzero entries of the vector $\mathbf{j}_l = (j_1, \ldots, j_M)$ occur in j_l, and these satisfy $[j_l] = m_l + 1$. Unfortunately, the constant T_1 above depends on the geometry of the set δ. For most finite element applications, however, each element of the partition $\triangle$, is the affine transformation of some reference set.

In this case, the dependence of T_1 on the geometry of δ is unimportant, and the approximation result above can be made to hold uniformly across the sets in $\triangle$. For further discussion of these issues, the reader is referred to Dupont and Scott (1980) and chapter 13 of the book by Schumaker (1981). We now take up a more detailed discussion of finite element spaces.

Finite Elements

For the remainder of this chapter, we restrict our attention to spline spaces defined over a compact region $\mathcal{X}$ of $\mathbb{R}^2$, and let $z = (x, y)$ denote an arbitrary point in $\mathbb{R}^2$. Of course, the results derived in this section can be extended to higher dimensions, but only with a corresponding increase in notational complexity. As mentioned in the previous subsection, the partitions $\triangle$ of $\mathcal{X}$ used in finite element applications are typically generated by affine transformations of a single reference set. To further simplify our presentation, we will confine our discussion to partitions made up of triangles. In this case, we say that $\triangle$ is a triangulation of $\mathcal{X}$. Throughout our discussion, we insist that $\triangle$ consist of triangles such that the vertices of each triangle are not contained in the interior of an edge of any other triangle. Such a triangulation is said to be conforming.

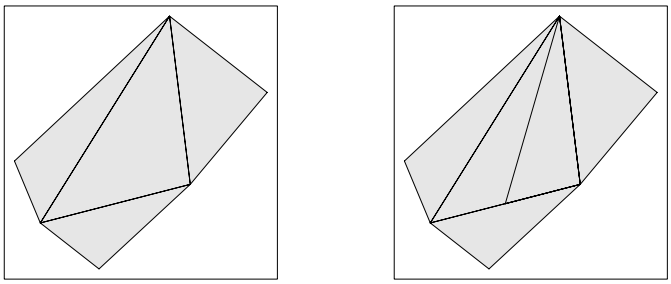


Figure 1. Conforming and non-conforming triangulations of a polyhedral region.

To quote Oden and Reddy (1976), "The finite-element method is, first, a systematic and very powerful method of interpolation." For ease of exposition, we consider here only Lagrange interpolation over triangles. Let m be a positive integer and let δ be a triangle in the plane. We say that a set of points in δ is m-unisolvent if a polynomial of total degree m is uniquely determined by its values on the set. In the Figure 2, we illustrate a key correspondence between the total degree of a polynomial and a set of unisolvent points in a triangle. The exact placement of these nodes will be discussed in the next subsection.

For a fixed positive integer m, define the interpolation operator that takes any function f defined on δ into the unique polynomial p that agrees with f at the nodal points described in Figure 2. Then, using the generalized Taylor's expansion discussed in the previous subsection, we can derive the approximation rate obtainable from the space of polynomials of total degree m on δ. Repeating this process for each $\delta \in \triangle$, we can derive approximation rates similar to those described in the context of integer translates of a single box spline. Furthermore, by considering the more general spaces of polynomials introduced in Section 1.1, we can repeat the analysis given in the previous section for tensor products of box splines and establish rates of approximation for models involving spaces of Lagrange finite elements over triangles. Having indicated how these results could be obtained, we now leave the approximation properties of finite element interpolations and instead consider aspects of these spaces that make them attractive from a methodological standpoint.

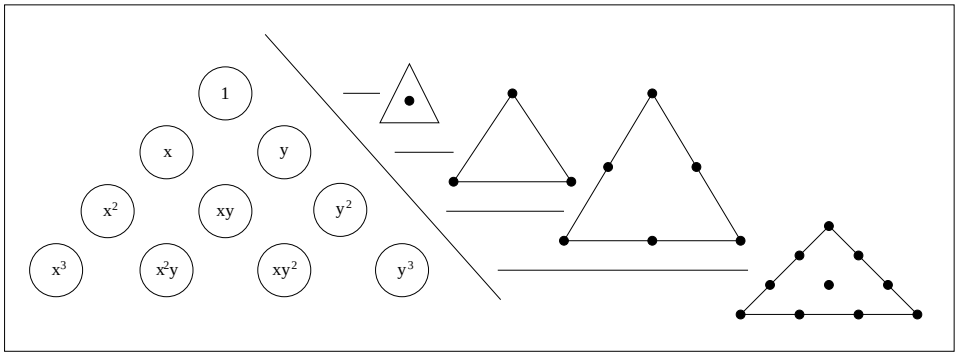
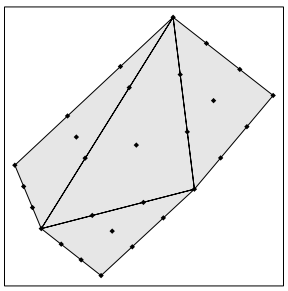


Figure 2. Pascal's triangle and associated unisolvent sets.

In the accompanying figure, we present a triangulation $\triangle$ of a polyhedral region $\mathcal{X}$ along with the set of 3-unisolvent node points associated with each triangle. Observe that by construction, there are exactly 4 nodes on each edge of the triangles in $\triangle$. Hence, the nodes appearing on a given edge in the triangulation are 3-unisolvent for the univariate polynomial defined on that edge. Therefore, we find that given any function f defined on $\mathcal{X}$, the interpolant that is



constructed by matching the values of f at the node points in each triangle in $\triangle$ is in fact a continuous function on $\mathcal{X}$. It is possible to create interpolants having greater smoothness by incorporating information about the derivatives of f at certain node points. For a more complete discussion of these issues, the reader is referred to Oden and Reddy (1976), Ciarlet (1978) and Schwarz (1980).

We can use the Lagrange interpolation scheme described above to construct a basis for the space of continuous piecewise polynomial functions of total degree m on $\mathcal{X}$. Each such basis function corresponds to interpolating the value of one at a single node, and zero at the remaining nodes. This process is repeated for each node in the triangulation $\triangle$. In the next subsection, we discuss an alternate basis for this space that has become quite popular among approximation theorists and computer scientists. As we will see, part of the success of this alternate basis stems from the fact that smoothness constraints across edges in $\triangle$ can be described simply and in a geometrically significant manner.

B-nets and Multivariate Splines

Let $\delta \in \triangle$ be a triangle in the plane with verticies $v_1 = (x_1, y_1)$, $v_2 = (x_2, y_2)$, and $v_3 = (x_3, y_3)$. For the moment, let $z = (x, y)$ denote an arbitrary point in $\mathbb{R}^2$, and define the functions $\phi_1(z)$, $\phi_2(z)$, and $\phi_3(z)$ such that

$$x = x_1 \phi_1(z) + x_2 \phi_2(z) + x_3 \phi_3(z),$$

$$y = y_1 \phi_1(z) + y_2 \phi_2(z) + y_3 \phi_3(z),$$

and

$$1 = \phi_1(z) + \phi_2(z) + \phi_3(z),$$

for all $z \in \mathbb{R}^2$. We refer to triple $\big(\phi_1(z), \phi_2(z), \phi_3(z)\big)$ as the barycentric coordinate of z with respect to the triangle δ.

It is not hard to demonstrate that, for $1 \leq i \leq 3$, the function $\phi_i(z)$ is in fact that linear function in x and y which takes on the value one at the point v_i and zero at the two remaining verticies of δ. Observe that if $z \in \delta$, then each of these funcitons is nonnegative. Moreover, it can be shown that, for $j = (j_1, j_2, j_3)$, the functions

$$\phi^j(z) = \frac{(j_1 + j_2 + j_3)!}{j_1!\, j_2!\, j_3!}\, \phi_1^{j_1}(z)\, \phi_2^{j_2}(z)\, \phi_3^{j_3}(z), \qquad [j] = m \ , z \in \delta \,, \quad (2.1)$$

are a basis for the space of polynomials of total degree m on δ. By construction these basis functions are naturally associated with points in δ; that is, given a

vector $j = (j_1, j_2, j_3)$ satisfying $[\,j\,] = m$, the point having barycentric coordinates $(j_1/m, j_2/m, j_3/m)$ lies in δ. These points are in fact exactly the m-unisolvent points given in Figure 2 above. Therefore, we arrive once again at a correspondence between polynomials of total order m and the set of points given in Figure 2. In this case however, we represent the correspondence through the use of the so called Bezier- or B-net. To be more precise, given any polynomial p of total degree m on δ, we can write

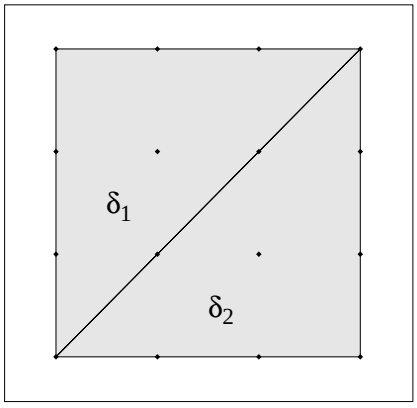
$$p(z) \;=\; \sum_{[\,j\,]\,=\,m} \beta_j \, \phi^j(z)\,, \qquad \text{for } z \in \mathbb{R}^2\,. \tag{2.2}$$

We then refer to the collection of points

$$\big(\, j_1/m,\, j_2/m,\, j_2/m,\, \beta_j \,\big)\,, \qquad \text{for } j = (j_1, j_2, j_3) \text{ with } [\,j\,] \;=\; m\,, \tag{2.3}$$

as the B-net associated with the polynomial p. While we have derived this representation with respect to a single triangle $\delta \in \triangle$, we can easily extend these definitions to construct the B-net associated with a piecewise polynomial function defined on $\mathcal{X}$ relative to the triangulation $\triangle$.

In fact, more is possible. Observe that if two triangles $\delta_1, \delta_2 \in \triangle$ share a common edge, then for any positive integer m, the m-unisolvent points in each triangle agree on the common edge. The following figure demonstrates the case when $m = 3$.



As with a single triangle, each point in the above figure corresponds to a basis function. If a point lies on an edge of $\triangle$ separating δ_1 and δ_2 then we take the associated basis function to be given by (2.1) with respect to δ_1, for $z \in \delta_1$, and by (2.1) with respect to δ_2, for $z \in \delta_2$. Arguing as in the subsection on finite element interpolations, we find that the resulting basis functions are in fact continuous on $\mathcal{X}$. Therefore, by assigning values to the coefficients represented in the B-net above, we obtain a continuous, piecewise polynomial function on $\mathcal{X}$.

One of the major reasons for introducing the B-net of a triangulation is the ease with which higher order smoothness constraints can be enforced across the edges of $\triangle$. It can be shown that these smoothness constraints are simple linear contrasts of the B-net coefficients. For an excellent account of these and many other properties of this representation, the reader is referred to Farin (1987), Goodman (1987ab), de Boor (1987), and Chui (1988).

Let m and r nonnegative integers, and let $\triangle$ be a conforming triangulation of a polyhedral region $\mathcal{X}$. Using the smoothness constraints described above, many authors have constructed locally supported basis functions for the space of piecewise polynomials of total degree m possessing r continuous derivatives on $\mathcal{X}$. Typically, r must be taken much smaller than m for these constructions to work. For example, Chui and Lai (1985, 1987, 1990) assume that m is not smaller than $4r+1$, while by considerably more complicated arguments, de Boor and Höllig (1988) and de Boor (1989) are able to reduce this lower bound to $3r+2$. For additional constructions, the reader is referred to Alfeld, Piper, and Schumaker (1987ab), Schumaker (1989) and Ibrahim and Schumaker (1991).

The restrictions imposed on the relationship between m and r by these constructions are fairly severe from a modelling standpoint. It appears that we are forced to consider high degree polynomials and fairly high dimensional spaces. Fortunately, there is a way around these problems. By subdividing each triangle $\triangle$ in a symmetric way and applying the smoothness constraints discussed above, Powell and Sabin (1977) and later Chui and He (1990) were able to construct smooth, low degree piecewise polynomial functions relative to a triangulation $\triangle$ of $\mathcal{X}$. In these cases, the term "piecewise" refers to the subdivided triangulation. In Figure 3, we present an original triangulation $\triangle$ and its refinement following the scheme of Chui and He (1990).

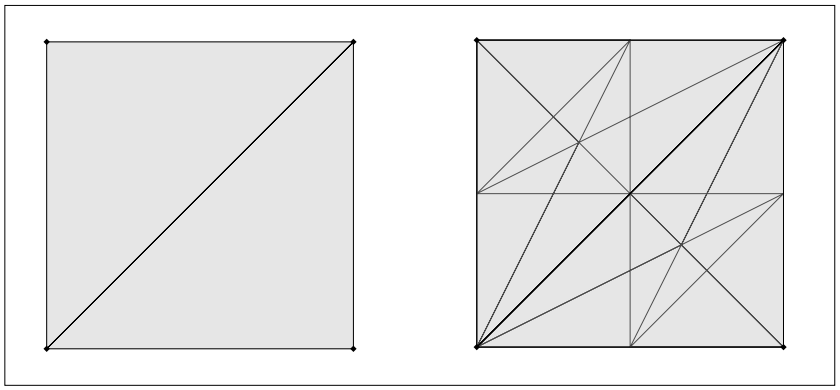
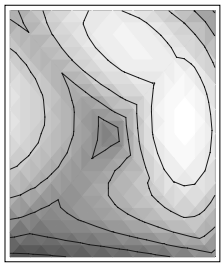


Figure 3. A triangulation and its symmetric refinement.

The spline spaces built from various symmetric subdivisions like the one presented above are powerful tools from which practically useful methodology can be built. Researchers have only just begun to investigate the potential of these spaces. For example, Dierckx, Van Leemput, and Vermeire (1992) present a surface fitting routine based on adaptive triangulations and the splines of Powell and Sabin (1977). For the most part, however, these investigations are being conducted by numerical analysists. Given out experience with non-parametric procedures, we feel that statisticians can make significant contributions to this growing area.

Chapter 3

Maximum Likelihood Estimation



3.1 Extended Linear Models

In Chapter 1, we considered estimating an unknown function $\mu(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ by ordinary least squares based on a suitably defined space of polynomial splines. With each covariate X_l, $1 \leq l \leq M$, we associated a space of smooth piecewise polynomial functions $\mathbb{G}_l$ adapted to a collection $\triangle_l$ of subsets that partitioned the range of X_l. Given a hierarchical collection of subsets of $\{1, \ldots, M\}$, we constructed the unsaturated spaces $\mathbb{G}$ and $\mathbb{F}$ and defined ANOVA decompositions for functions in these spaces. By restricting our attention to estimates in $\mathbb{G}$, we found that we could improve the rate of convergence of $\hat{\mu}$ at the expense of estimating only an approximation $\mu^* \in \mathbb{F}$ to μ. In this chapter, we explore the use of the spaces $\mathbb{G}$ and $\mathbb{F}$, as well as their associated ANOVA decompositions, in the context of maximum likelihood estimation. More generally, we consider estimates that are obtained as solutions to a particular class of optimization problems.

Notation

Let $\mathbf{W}$ be a random vector that takes values in some subset $\mathcal{W}$ of $\mathbb{R}^d$, $d > 0$, and let ϕ denote an unknown, possibly vector-valued function that depends on the

distribution of $\mathbf{W}$. As in Chapter 1, we assume that the set $\mathcal{W}$ is of the form $\mathcal{W} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_N$, where each set $\mathcal{W}_l$ is a subset of $\mathbb{R}^{d_l}$ for some positive integer d_l, and let $\mathbf{w} = (w_1, \ldots, w_N)$ denote a typical point in $\mathcal{W}$. Next, we suppose that the function ϕ depends only on the M variables $w_1, \ldots, w_M$, where $M \leq N$. Then we set $\mathcal{W} = \mathcal{U} \times \mathcal{V}$, where

$$\mathcal{U} \;=\; \mathcal{W}_1 \times \cdots \times \mathcal{W}_M \;=\; \mathcal{U}_1 \times \cdots \times \mathcal{U}_M$$

and

$$\mathcal{V} \;=\; \mathcal{W}_{M+1} \times \cdots \times \mathcal{W}_N \;=\; \mathcal{V}_1 \times \cdots \times \mathcal{V}_{N-M}\,,$$

and we partition $\mathbf{w} = (\mathbf{u}, \mathbf{v})$ and $\mathbf{W} = (\mathbf{U}, \mathbf{V})$ in a similar manner. In addition, for each $1 \leq l \leq M$, we assume that $\mathcal{U}_l$ is a compact subset of $\mathbb{R}^{d_l}$, $d_l > 0$, having unit volume.

Given a real-valued function f_0 whose domain includes $\mathcal{U}$, let $\| f_0 \|_\infty$ denote the sup-norm of f_0 on $\mathcal{U}$, and let $\| f_0 \|_2$ denote the L_2-norm of f_0 on $\mathcal{U}$. Then, given the real-valued functions $f_1, \ldots, f_J$, set $f(\mathbf{u}) = \big(f_1(\mathbf{u}), \ldots, f_J(\mathbf{u})\big)$, $\mathbf{u} \in \mathcal{U}$, and define

$$\| f \|_\infty \;=\; \max_j \left\{ \| f_1 \|_\infty, \ldots, \| f_J \|_\infty \right\} \tag{1.1a}$$

and

$$\| f \|_2^2 \;=\; \| f_1 \|_2^2 + \cdots + \| f_J \|_2^2\,. \tag{1.1b}$$

A straightforward argument demonstrates that the quantities in (1.1) are in fact norms. A vector-valued function is said to be bounded on $\mathcal{U}$ if its sup-norm given in (1.1a) is finite. Similarly, f is said to be square-integrable on $\mathcal{U}$ if the quantity in (1.1b) is finite. For the most part, this notational convention will allow us to ignore the distinction between vector-valued and real-valued functions on $\mathcal{U}$. In the next section, we introduce versions of the empirical and theoretical norms defined in Chapter 1 for real-valued functions. Using the definition in (1.1b) we will extend these definitions to vector-valued functions defined on $\mathcal{U}$.

Extended Linear Models and Saturated Spaces

In this chapter, we view ϕ as the solution to an optimization problem over a saturated space of functions whose domain contains $\mathcal{U}$. To make this precise, we introduce a functional $l(f, \mathbf{w})$ and set

$$\Lambda(f) \;=\; E\big[\, l(f, \mathbf{W}) \,\big]$$

for all bounded functions f defined on $\mathcal{U}$. It is assumed that

$$\phi \;=\; \underset{f \in L_\infty(\mathcal{U})}{\operatorname{argmax}}\; \Lambda(\, f\,),$$

where $\Lambda(\phi)$ is finite. Suppose, for example, that ϕ is some unknown component of the distribution of $\mathbf{W}$ and that $l(f, \mathbf{w})$ denotes the log-likelihood of a single observation of $\mathbf{W}$ under the assumption that $\phi = f$. Then, by the information inequality, $\Lambda(\,\cdot\,)$ is maximized by the true underlying function ϕ. Of course, this framework allows for more than simple maximum likelihood estimation.

In general, we consider functionals $l(f, \mathbf{w})$ of the form

$$l(\, f, \mathbf{w}\,) \;=\; \sum_{k=1}^{K} b_k(\, f, \mathbf{u}\,)\, B_k(\, \mathbf{w}\,), \qquad \mathbf{w} = (\mathbf{u}, \mathbf{v}) \in \mathcal{W}, \tag{1.2}$$

where the functionals b_k are subject to certain regularity conditions to be discussed in the next section. We say that functionals in the form (1.2) specify an extended linear model. The appeal of this form is that we are able to separate quantities that depend on $\mathbf{v}$ from quantities that depend on f. As we will see, this property is useful in establishing stochastic bounds for $l(f, \mathbf{W})$. In the next example, we illustrate the variety of models that fall into form (1.2). As these models will be used continually throughout this chapter, we enumerate them as

[R] Regression,
[GR] Generalized Regression,
[CR] Censored Regression,
[PR] Polychotomous Regression,
[D] Density Estimation.

In many of these cases, it is clear that ϕ is more naturally obtained as the maximum over a larger class of functions than $L_\infty(\mathcal{U})$. However, in each of the cases we consider, it will be necessary to assume that ϕ is bounded, so it is sufficient to view ϕ as being defined by (1.1).

Example 1 **[R]** Let V be a real-valued random variable, and consider choosing ϕ so as to maximize the expression $-E[V - f(\,\mathbf{U}\,)]^2$ over all square-integrable functions f defined on $\mathcal{U}$. In order to put this in the form of (1.2), ignore the term that does not involve f and set

$$l(f, \mathbf{w}) \;=\; -f^2(\mathbf{u}) \;+\; 2\, f(\mathbf{u})\, v\,.$$

If the conditional variance function $\sigma(\mathbf{u}) = \mathrm{var}(V \,|\, \mathbf{U} = \mathbf{u})$ is bounded over $\mathbf{u} \in \mathcal{U}$, then $\phi(\mathbf{u})$ is the regression function $E(V \,|\, \mathbf{U} = \mathbf{u})$. In Chapter 1,

however, we were forced to assume that ϕ is bounded on $\mathcal{U}$, and hence, under this assumption, ϕ can be obtained by maximizing $\Lambda(\,\cdot\,)$ over all bounded functions on $\mathcal{U}$.

[GR] Consider an exponential family of probability distributions of the form

$$\exp\big(\,b_1(\,\eta\,)\,v + b_2(\,\eta\,)\,\big)\,\nu(dv)\,,$$

where η ranges over $\mathbb{R}$ and ν is a nonzero measure on $\mathbb{R}$ that is not concentrated at a single point. It is assumed that b_1 is a twice continuously differentiable function on $\mathbb{R}$. It is also assumed that b_1' is strictly positive on $\mathbb{R}$ and hence that b_1 is strictly increasing on $\mathbb{R}$. Then the mean of this distribution is given by the expression

$$b_3(\eta) \;\;=\;\; -b_2'(\eta)\,/\,b_1'(\eta)\,,$$

where b_3 is referred to as the link function. Set $\mathbf{W} = (\mathbf{U}, V)$, where V is a real random variable, and assume that the conditional distribution of V given that $\mathbf{U} = \mathbf{u}$ belongs to the exponential family given above with $\eta = \phi(\mathbf{u})$. Then,

$$E(\,V\,|\,\mathbf{U} = \mathbf{u}\,) \;\;=\;\; b_3\big(\,\phi(\mathbf{u})\,\big)\,. \tag{1.3}$$

Under suitable conditions, if we set

$$l(f, \mathbf{w}) \;\;=\;\; b_1\big(\,f(\mathbf{u})\,\big)\,v + b_2\big(\,f(\mathbf{u})\,\big)\,, \tag{1.4}$$

where $\mathbf{w} = (\mathbf{u}, v)$, then by the information inequality, $\Lambda(\,\cdot\,)$ is maximized by the true underlying function ϕ. Observe that $l(f, \mathbf{w})$ is of the form (1.2).

More generally, suppose that (1.3) holds, but that the conditional distribution of V given $\mathbf{U} = \mathbf{u}$ does not necessarily belong to the indicated exponential family. If we define $l(f, \mathbf{w})$ as in (1.4), then the function ϕ that maximizes $\Lambda(\,\cdot\,)$ is still given by (1.3).

[CR] Suppose that, given $\mathbf{U} = \mathbf{u}$, the real random variable Z has a normal distribution with conditional mean $\phi(\mathbf{u})$ and unit variance. Let I be the indicator random variable for the event that Z is greater than zero, and set

$$\mathbf{V} \;\;=\;\; (\,V_1, V_2\,) \;\;=\;\; (\,IZ, I\,)\,.$$

Here, $I = 1$ only if V_1 is an observation from the distribution of Z. Otherwise, V_1 is said to have been censored. Let $l(f, \mathbf{w})$ be the log-likelihood corresponding to a single observation of the random vector $\mathbf{W} = (\mathbf{U}, \mathbf{V})$ assuming that

$\phi = f$. Then, the likelihood corresponding to a single observation $\mathbf{W}$ is given by

$$\Big[\,\varphi\big(\,v_1 - f(\,\mathbf{u}\,)\big)\,\Big]^{v_2}\;\Big[\,1-\Phi\big(\,f(\,\mathbf{u}\,)\big)\,\Big]^{1-v_2},$$

where $\varphi(\,\cdot\,)$ and $\Phi(\,\cdot\,)$ are the standard normal density and distribution functions, respectively, and $\mathbf{w} = (\mathbf{u}, \mathbf{v}) = (\mathbf{u}, v_1, v_2)$. Ignoring the term in the log-likelihood that does not involve f, we set

$$l(f, \mathbf{w}) \;=\; -v_2\,\big(\,f^2(\mathbf{u}) \;-\; 2\,f(\mathbf{u})\,v_1\,\big)\,/\,2 \;+\; \big(\,1-v_2\,\big)\,\log\Big(\,1-\Phi\big(\,f(\,\mathbf{u}\,)\big)\,\Big),$$

which can easily be put in the form (1.2). By the information inequality, we find that $\Lambda(\,\cdot\,)$ is maximized at the true underlying function ϕ.

[PR] Consider a discrete random variable V taking values in the set $\mathcal{V} = \{0, \ldots, J\}$, where J is some positive integer. Assume that $(\mathbf{U}, V)$ have a joint distribution and that $P(\,V = v\,|\,\mathbf{U} = \mathbf{u}) > 0$ for all $\mathbf{w} = (\mathbf{u}, v) \in \mathcal{W}$. In this case, we let f be the vector-valued function $f = (f_1, \ldots, f_J)$, where each f_j is a real-valued function whose domain includes $\mathcal{U}$. Now, define

$$\begin{aligned} l(f, \mathbf{w}) \;&=\; f_1(\mathbf{u})\,I_1(v) \;+\; \cdots \;+\; f_J(\mathbf{u})\,I_J(v) \\ &\qquad -\;\log\Big(\,1 \;+\; \exp\big(\,f_1(\mathbf{u})\big) \;+\; \cdots \;+\; \exp\big(\,f_J(\mathbf{u})\big)\,\Big), \end{aligned}$$

where $I_j(v) = 1$ if $v = j$ and is zero otherwise. Clearly, this functional is in the form (1.2). Now, by a straight calculation, we find that the function f maximizing $\Lambda(\,\cdot\,)$ is comprised of the functions

$$f_j(\mathbf{u}) \;=\; \log\big(\,P(\,V = j\,|\,\mathbf{U} = \mathbf{u}\,)\,/\,P(\,V = 0\,|\,\mathbf{U} = \mathbf{u}\,)\big),$$

for $\mathbf{u} \in \mathcal{U}$ and $1 \le j \le J$. By our assumption that $P(\,V = v\,|\,\mathbf{U} = \mathbf{u}) > 0$ for all $\mathbf{w} = (\mathbf{u}, v) \in \mathcal{W}$, these functions are bounded over $\mathcal{U}$.

[D] Suppose $\mathbf{w} = \mathbf{u}$, and let f be any bounded function defined on $\mathcal{W}$. Set

$$\exp(\,c(f)\,) \;=\; \int_{\mathcal{U}} \exp\,(\,f(\mathbf{u})\,)\,d\mathbf{u}\,,$$

and observe that $\exp(\,f - c(f)\,)$ is a density function on $\mathcal{U}$. In this context, we set $l(f, \mathbf{u}) = f(\mathbf{u}) - c(f)$, which is in the form (1.2). Under suitable conditions, we know from the information inequality that $\phi(\mathbf{u})$ is the log of the density of $\mathbf{U}$. However, given any bounded function f and any constant $a \in \mathbb{R}$, $\Lambda(\,f + a\,) = \Lambda(\,f\,)$ and hence $\Lambda(\,\cdot\,)$ is also maximized by any function of the form $\phi(\mathbf{u}) + a$, $a \in \mathbb{R}$. Imposing the constraint that $\phi(\mathbf{u})$ integrates to one is sufficient to specify $\phi(\mathbf{u})$ in an essentially unique manner.

A comment is in order about our choice of examples. Regression is included because it is the simplest conceptually, and because it has been treated in considerable depth in Chapter 1. We include the generalized regression example because it serves as a model for the form of the objective function $l(f, \mathbf{w})$ given in (1.2). As its name might indicate, censored regression affords us the opportunity to discuss censoring and in particular consider the effect that this has on the underlying norms used to generate our ANOVA decompositions. By treating the polychotomous regression context, we are forced to consider vector-valued functions ϕ. Observe that in each of these regression contexts, however, the functional $l(f, \mathbf{w})$ depends on f only through the quantity $f(\mathbf{u})$. Density estimation provides us with an example in which the dependence of $l(f, \mathbf{w})$ on f is more complicated. For the most part, we will be able to treat all of these cases simultaneously. Of course there are some minor notational conventions that have to be observed to make this possible.

We have considered one such convention already when we defined the norms in (1.1). Another issue that has to be addressed comes from the fact that in the density estimation case, any reasonable estimate of ϕ should integrate to one. As shown in the example, this can be accomplished by subtracting from f the constant $c(f)$. Therefore, in this case we consider estimates of ϕ that are of the form $f - c(f)$, where f is any bounded function. To make this notation compatible with the regression contexts, we define $c(\,\cdot\,)$ to be identically zero if we not estimating a density function. In the polychotomous regression case, this means that $c(\,\cdot\,)$ is in fact a vector of zeros. While it may appear awkward, by introducing this normalizing factor we are able to examine theoretically each of the cases listed in Example 1 simultaneously.

All that we require from the map $c(\,\cdot\,)$ is that it be Lipshitz continuous in some neighborhood of zero. In other words, it is assumed that for every positive constant T_1, there is a positive constant T_2 such that, for all pairs of real-valued functions f_1, f_2 defined on $\mathcal{U}$ satisfying

$$\| f_1 \|_\infty \;\leq\; T_1 \qquad \text{and} \qquad \| f_2 \|_\infty \;\leq\; T_1 \,, \tag{1.5a}$$

we have that

$$\big| c(f_1) \,-\, c(f_2) \big| \;\;\leq\;\; T_2 \, \| f_1 - f_2 \| \,. \tag{1.5b}$$

A direct computation indicates that this holds for the functional $c(\,\cdot\,)$ defined in the context of density estimation. In the remaining cases of Example 1, the relation in (1.5b) holds trivially.

ANOVA Decompositions (Real-Valued ϕ)

As we have seen in Chapter 1, the heart of an ANOVA decomposition is the underlying inner-product. When selecting either a theoretical or an empirical inner-product in the context of maximum likelihood estimation, it is important to accommodate the possibility that observations may be censored. In general, suppose that among the variables that comprise the vector $\mathbf{V}$, there exists a random variable I that is the indicator for the event that an observation has been censored. Let $f_{\mathbf{U}|I}(\,\mathbf{u}\,|\,i\,)$ denote the conditional density of $\mathbf{U}$ given that $I = i \in \{0, 1\}$, and assume that the following generalization of Condition 2 holds.

Condition 2′ *If censoring is present, assume that $P(I = 1) > 0$ and that there exists a positive constant M_3 such that the density of $\mathbf{U}$ given that $I = 1$ satisfies*

$$\frac{1}{M_3} \;\leq\; f_{\mathbf{U}|I}(\,\mathbf{u}\,|\,1\,) \;\leq\; M_3\,, \qquad \mathbf{u} \in \mathcal{U}.$$

Otherwise, assume that the above inequalities apply to the density $f_{\mathbf{U}}$ of $\mathbf{U}$ itself.

For any two bounded, real-valued functions f_1, f_2 whose domain includes $\mathcal{U}$, set

$$\langle\, f_1, f_2 \,\rangle \;=\; E\big[\, I\, f_1(\mathbf{U})\, f_2(\mathbf{U})\,\big] \quad \text{and} \quad \|\, f_1 \,\|^2 \;=\; E\big[\, I\, f_1^2(\mathbf{U})\,\big]\,. \qquad (1.6a)$$

Suppose that we are given a simple random sample $\mathbf{W}_1, \ldots, \mathbf{W}_n$ from the distribution of $\mathbf{W}$, and extract from $\mathbf{V}_1, \ldots, \mathbf{V}_n$ the associated sample of indicator random variables $I_1, \ldots, I_n$. Then, given any pair of functions f_1, f_2 whose domain includes $\mathcal{U}$, we define the empirical inner-product and norm by

$$\langle\, f_1, f_2 \,\rangle_n \;=\; E_n\big[\, I\, f_1(\mathbf{U})\, f_2(\mathbf{U})\,\big] \quad \text{and} \quad \|\, f_1 \,\|_n^2 \;=\; E_n\big[\, I\, f_1^2(\mathbf{U})\,\big]\,, \qquad (1.6b)$$

where we recall from Chapter 1 that, for a function $f(\mathbf{w})$ defined for all $\mathbf{w} \in \mathcal{W}$, the empirical expectation E_n of $f(\mathbf{W})$ is defined by

$$E_n\big[\, f(\mathbf{W})\,\big] \;=\; \frac{1}{n} \sum_{i=1}^{n} f(\,\mathbf{W}_i\,)\,.$$

Observe that only in the censored regression context do the inner-products and norms defined in (1.6) differ from their definition in Chapter 1.

Now, let $\mathcal{S}_1$ denote a hierarchical collection of subsets of $\{1, \ldots, M\}$ and for each $s \in \mathcal{S}_1$, let $\mathbb{F}_s$ denote the space of square-integrable functions on $\mathcal{U}$ that depend only on the variables u_l for $l \in s$. From these spaces we construct

$$\mathbb{F}_1 \;=\; \left\{ \sum_{s \in \mathcal{S}_1} f_s \,:\, f_s \in \mathbb{F}_s \;\text{ for } s \in \mathcal{S}_1 \right\}. \qquad (1.7a)$$

For each $s \in \mathcal{S}_1$, let $\mathbb{F}_s^0$ denote the space of functions in $\mathbb{F}_s$ that are orthogonal to $\mathbb{F}_r$, $r \subset s$, relative to the theoretical inner-product in (1.6a). If we are estimating one of the real-valued regression functions listed in Example 1, set $\mathcal{S} = \mathcal{S}_1$. On the other hand, if we are estimating a density function, set $\mathcal{S} = \mathcal{S}_1 \setminus \{\varnothing\}$. Then, define

$$\mathbb{F} \;=\; \left\{ \sum_{s \in \mathcal{S}} f_s \,:\, f_s \in \mathbb{F}_s^0 \;\text{ for } s \in \mathcal{S} \right\}. \tag{1.7b}$$

Therefore, in the real-valued regression contexts, $\mathbb{F} = \mathbb{F}_1$, while in the density estimation context, $\mathbb{F}$ is the space of functions in $\mathbb{F}_1$ that are orthogonal to the constant functions relative to the theoretical inner-product in (1.6a). Now, if the function $f_{\mathbf{U}\,|\,I}(\,\mathbf{u}\,|\,1\,)$ is bounded away from zero and infinity on $\mathcal{U}$, then we know from Lemma 3.7 in Chapter 1 that each function $f \in \mathbb{F}$ can be represented uniquely as the sum

$$f \;=\; \sum_{s \in \mathcal{S}} f_s\,, \qquad \text{where } f_s \in \mathbb{F}_s^0 \;\text{ for }\; s \in \mathcal{S}\,, \tag{1.8}$$

Recall that for each $1 \le l \le M$, we have assumed that the set $\mathcal{U}_l$ is a compact subset of $\mathbb{R}^{d_l}$, $d_l > 0$, having unit volume. Proceeding as in (3.1) of Chapter 1, given $s = \{s_1, \ldots, s_k\} \in \mathcal{S}_1$, set

$$\mathcal{U}_s \;=\; \mathcal{U}_{s_1} \times \cdots \times \mathcal{U}_{s_k} \qquad \text{and} \qquad \mathbf{u}_s \;=\; (u_{s_1}, \ldots, u_{s_k})\,,$$

assuming that $s_1 < \cdots < s_k$. Next, let $\triangle_1, \ldots, \triangle_M$ denote partitions of the sets $\mathcal{U}_1, \ldots, \mathcal{U}_M$, respectively, and for each $s \in \mathcal{S}_1$, define $\triangle_s$ as was done prior to (3.2) in Chapter 1. Then, for a given vector $m \in \mathbb{N}^M$, let $\mathbb{PP}_s$ denote the space of piecewise polynomials of total degree m_s in the variables $\mathbf{u}_s$. Additionally, for $1 \le l \le M$, let $\mathbb{G}_l$ be a space of spline functions defined relative to $\triangle_l$. Then, for each subset $s = \{s_1, \ldots, s_k\}$ in $\mathcal{S}_1$ set

$$\mathbb{G}_s \;=\; \mathbb{G}_{s_1} \otimes \cdots \otimes \mathbb{G}_{s_k}\,,$$

and construct

$$\mathbb{G}_1 \;=\; \left\{ \sum_{s \in \mathcal{S}_1} g_s \,:\, g_s \in \mathbb{G}_s \;\text{ for } s \in \mathcal{S}_1 \right\}. \tag{1.9a}$$

(Recall the definition of a tensor product space given in Chapter 1.) Observe that the space $\mathbb{G}_1$ is a subspace of

$$\mathbb{PP} \;=\; \left\{ \sum_{s \in \mathcal{S}_1} p_s \,:\, p_s \in \mathbb{PP}_s \;\text{ for } s \in \mathcal{S}_1 \right\}.$$

Now, for each $s \in \mathcal{S}_1$, let $\mathbb{G}_s^0$ denote the space of functions in $\mathbb{G}_s$ that are orthogonal to $\mathbb{G}_r$ for $r \subset s$ relative to the empirical inner-product defined in (1.6b). By analogy with (1.7b), set

$$\mathbb{G} \;=\; \left\{ \sum_{s \in \mathcal{S}} g_s \,:\, g_s \in \mathbb{G}_s^0 \;\text{ for } s \in \mathcal{S} \right\}. \tag{1.9b}$$

Recall from Chapter 1 that the space $\mathbb{G}_1$ is said to be identifiable if we cannot find a nonzero function $g \in \mathbb{G}_1$ that is zero on the design points. In this context, however, we must account for the possibility that some of the observations in our sample have been censored. Therefore, throughout this chapter, we say that the space $\mathbb{G}_1$ is identifiable if we cannot find a nonzero function $g \in \mathbb{G}_1$ that is zero on the subset of $\mathbf{U}_1, \ldots, \mathbf{U}_n$ corresponding to the uncensored observations. This definition explains the form of the empirical inner-product and its associated norm given in (1.6) when censoring is present. Finally, as we have seen in Chapter 1, if $\mathbb{G}_1$ is identifiable then each function $g \in \mathbb{G}_1$ can be represented uniquely as the sum

$$g \;=\; \sum_{s \in \mathcal{S}} g_s \,, \qquad \text{where } g_s \in \mathbb{G}_s^0 \;\text{ for }\; s \in \mathcal{S}\,. \tag{1.10}$$

At this point, a comment is in order about the difference between the norms appearing in (1.6) and the related quantities defined in (1.24) of Chapter 1. Let n' denote the number of uncensored observations; that is, the number of observations for which $I = 1$. Since the probability that $I = 1$ is positive, from the law of large numbers we know that $n' / n - P(I = 1)$ is $o_P(1)$. Therefore, as n tends to infinity, n' also tends to infinity except on a set whose probability tends to zero with n. Now, recall from Chapter 1 that the space $\mathbb{G}_1$ is identifiable with respect to a simple random sample of size n subject to Condition 2, a boundedness constraint on the density of the underlying random vector, and Condition $3'$ which controls the rate at which we can refine the triangulations from which $\mathbb{G}_1$ is constructed. If we apply this result only to the uncensored observations, then Conditions $2'$ and $3'$ are sufficient to guarantee that the probability that $\mathbb{G}_1$ is nonidentifiable tends to zero with n. Moreover, recall from that chapter that the rate at which the ordinary least squares estimates based on a simple random sample of size n approached a suitably defined theoretical limit was of the form $O_P(\, n^{-\gamma}\,)$ for some $\gamma < 1$. Suppose we apply this result directly to the uncensored observations, replacing n with n' in (1.6b). Then the rate is given by a power of $1/n'$. However, since $n' / n - P(I = 1)$ is $o_P(1)$, we again obtain the rate $O_P(\, n^{-\gamma}\,)$. This fact will be important in the proof of Lemma 2.3 in the next section.

ANOVA Decompositions (Vector-Valued ϕ)

If the function ϕ is in fact vector-valued, we extend the definitions given above in an obvious manner. First, suppose that $f_1, \ldots, f_J$ are real-valued functions on $\mathcal{U}$, and set $f = (f_1, \ldots, f_J)$. Then, define

$$\| f \|^2 = \sum_{j=1}^{J} E\left[I f_j^2(\mathbf{U}) \right] . \tag{1.11a}$$

Similarly, given a simple random sample $\mathbf{W}_1, \ldots, \mathbf{W}_n$ from the distribution of $\mathbf{W}$, extract from $\mathbf{V}_1, \ldots, \mathbf{V}_n$ the associated sample of indicator random variables $I_1, \ldots, I_n$, and define

$$\| f \|_n^2 = \sum_{j=1}^{J} E_n\left[I f_j^2(\mathbf{U}) \right] . \tag{1.11b}$$

In (1.8) we used the theoretical norm and inner-product (1.7a) to construct an ANOVA decomposition for real-valued functions. Similarly, in (1.10), we used the empirical version of these quantities given in (1.7b) to generate an ANOVA decomposition for certain piecewise polynomials. We now extend these notions to include vector-valued functions.

Toward this end, let $\mathbb{F}$ be any linear space of real-valued functions, and given $f_1, \ldots, f_J \in \mathbb{F}$, consider vector-valued functions of the form $f = (f_1, \ldots, f_J)$. The resulting collection of functions f is again a linear space. Using this device, we can construct vector-valued analogs of each of the spaces defined in the previous subsection. In addition, if $f = (f_1, \ldots, f_J)$ where each $f_j \in \mathbb{F}_1$, we can apply the ANOVA decomposition in (1.8) and write

$$f = \sum_{s \in \mathcal{S}_1} f_s , \qquad \text{where } f_s = (f_{1s}, \ldots, f_{Js}) \text{ and } f_{js} \in \mathbb{F}_s^0 \text{ for } s \in \mathcal{S} .$$

A similar expression holds for vector-valued functions $g = (g_1, \ldots, g_J)$ where each $g_j \in \mathbb{G}$. For convenience, we do not introduce special notation to differentiate between the spaces defined in the previous subsection and their vector-valued analogs defined here. It will always be clear which definition is appropriate. For example, when we are estimating a univariate function ϕ, the space $\mathbb{F}_1$ is always the collection given in (1.7a). When ϕ is vector-valued we use the definition introduced in this subsection instead.

Extended Linear Models and Unsaturated Spaces

Recall the form of the extended linear model given in (1.2). We have already assumed that there exists a unique bounded function ϕ that maximizes the functional

$\Lambda(f) = E[\,l(f, \mathbf{W})\,]$ over all $f \in L_\infty(\mathcal{U})$. We now consider maximizing $\Lambda(\,\cdot\,)$, this time over the unsaturated space $\mathbb{F}$. In this context, we restrict our attention to functionals $l(f, \mathbf{w})$ for which $\Lambda(\,\cdot\,)$ is strictly concave over bounded functions in $\mathbb{F}$. That is, given any two essentially different bounded functions $f_1, f_2 \in \mathbb{F}$ we have that

$$\Lambda\big(\,f_1 + \alpha(\,f_2 - f_1\,)\,\big) \;>\; \alpha\,\Lambda\big(\,f_1\,\big) + (1-\alpha)\,\Lambda\big(\,f_2\,\big)\,, \qquad \alpha \in (0,1)\,. \tag{1.12}$$

Here, f_1 and f_2 are said to be essentially different if the norm of their difference is not equal to zero. This property is shared by the functionals associated with each case treated in Example 1, a fact that will be discussed at length in Example 2. In addition, for each of these cases, the function

$$\Lambda\big(\,f_1 + \alpha(f_2 - f_1)\,\big)\,, \qquad \alpha \in (0,1)\,,$$

is twice continuously differentiable in α. Therefore, adding this differentiability condition to the list of assumptions about $\Lambda(f)$ we find that the strict concavity property in (1.12) is satisfied if, given any two essentially different bounded functions $f_1, f_2 \in \mathbb{F}$,

$$\frac{d^2}{d\alpha^2}\,\Lambda\big(\,f_1 + \alpha(f_2 - f_1)\,\big) \;<\; 0\,, \qquad \alpha \in (0,1)\,. \tag{1.13}$$

Finally, assume that there exists an essentially unique bounded function $\phi^* \in \mathbb{F}$ satisfying

$$\phi^* \;=\; \underset{f \in \mathbb{F}}{\operatorname{argmax}}\;\Lambda(\,f\,)\,,$$

where $\Lambda(\phi^*)$ is finite. We refer to $\phi^* - c(\phi^*)$ as the best theoretical approximation to ϕ in $\mathbb{F}$. Using the representation in (1.8), we write that

$$\phi^* \;=\; \sum_{s \in \mathcal{S}} \phi^*_s\,, \qquad \text{where } \phi^*_s \in \mathbb{F}^0_s \;\text{ for }\; s \in \mathcal{S}\,,$$

and we refer to the functions ϕ^*_s, $s \in \mathcal{S}$, as the components of the ANOVA decomposition of ϕ^*. The function ϕ^* represents the natural limit for our sample-based estimates to be defined shortly.

In addition to the existence of ϕ^*, we assume that Condition $5'$ holds for the components ϕ^*_s, restated here for convenience. The reader should keep in mind that if we are estimating a vector-valued function ϕ, then ϕ^* and its components are also vector-valued. In this case, the spline space $\mathbb{G}$ referred to in this restatement of Condition $5'$ is the vector-valued version introduced in the previous subsection. Suppose, for example, that the components of ϕ^* are given by

$$\phi^*_s \;=\; (\phi^*_{1s}, \ldots, \phi^*_{Js})\,, \qquad \text{where } \phi^*_{js} \in \mathbb{F}^0_s \;\text{ for }\; 1 \le j \le J \;\text{ and }\; s \in \mathcal{S}\,.$$

If for a fixed $s \in \mathcal{S}$, each real-valued function ϕ_{js}^* can be approximated at the same rate by elements in the space $\mathbb{G}$ defined in (1.9b), then ϕ_s^* can be approximated by elements in the vector-valued version of $\mathbb{G}$ at this same rate.

Condition 5′ *Assume that, for each $s \in \mathcal{S}$, there exists a function $\rho_s(\overline{h}_s)$ such that*

$$\rho_s(\overline{h}_s) \to 0 \quad as \quad \overline{h}_s \to 0 \qquad and \qquad \inf_g \left\| g - \phi_s^* \right\|_\infty = O\Big(\rho_s(\overline{h}_s)\Big),$$

where the infinum is taken over all $g \in \mathbb{G}_s$.

In each of the regression contexts, the concavity condition in (1.13) also guarantees that $\Lambda(\cdot)$ is strictly concave over $\mathbb{G}$. In the density estimation context, however, this is not the case. Therefore, we also assume that $l(g, \mathbf{w})$ is such that $\Lambda(\cdot)$ is strictly concave in $g \in \mathbb{G}$. That is, we require that (1.13) hold for any two different functions $g_1, g_2 \in \mathbb{G}$. In addition, we assume that there exists a unique function $\phi_n^* \in \mathbb{G}$ satisfying

$$\phi_n^* = \operatorname*{argmax}_{g \in \mathbb{G}} \Lambda(g),$$

where $\Lambda(\phi_n^*)$ is finite. We can then approximate $\phi^* - c(\phi^*)$ by $\phi_n^* - c(\phi_n^*)$. Using the representation in (1.10), we find that except on a set whose probability tends to zero with n, we can write

$$\phi_n^* = \sum_{s \in \mathcal{S}} \phi_{ns}^*, \qquad \text{where } \phi_{ns}^* \in \mathbb{G}_s^0 \text{ for } s \in \mathcal{S}.$$

In the next section we derive the rate at which ϕ_n^* and its components approach ϕ^* and its components as n tends to infinity. As we will see, the quantity $\| \phi^* - \phi_n^* \|^2$ plays the role of the bias component of the mean squared error given in Chapter 1.

So far, we have only considered maximizing the functional $\Lambda(\cdot)$ over the spaces $\mathbb{F}$ and $\mathbb{G}$. Suppose now that $\mathbf{W}_1, \ldots, \mathbf{W}_n$ is a simple random sample from the distribution of $\mathbf{W}$. Then, as in the case of ordinary maximum likelihood estimation, we set

$$\ell(f) = E_n\big[l(f, \mathbf{W})\big] = \frac{1}{n} \sum_{i=1}^n l(f, \mathbf{W}_i)$$

for all bounded functions f whose domain contains $\mathcal{U}$ and consider the sample-based estimate

$$\hat{\phi} = \operatorname*{argmax}_{g \in \mathbb{G}} \ell(g).$$

As is the case with $\Lambda(\cdot)$, we assume that except on a set whose probability tends to zero with n, $\ell(\cdot)$ is strictly concave on $\mathbb{G}$. Then, assuming a solution $\hat{\phi}$ exists,

we let $\hat{\phi} - c(\hat{\phi})$ denote the associated estimate of $\phi^* - c(\phi^*)$, the best theoretical approximation of ϕ in $\mathbb{F}$. In the next section, we will demonstrate that if each of the triangulations $\triangle_l$, $1, \leq l \leq M$, are refined with sample size so that a strengthening of Condition 3′ holds, then except on an event whose probability tends to zero with n, the estimate $\hat{\phi}$ exists and is unique. Combining this with results from Chapter 1, we find that with probability approaching one, the estimate $\hat{\phi}$ exists and admits the ANOVA decomposition

$$\hat{\phi} \;=\; \sum_{s \in \mathcal{S}} \hat{\phi}_s \,, \qquad \text{where } \hat{\phi}_s \in \mathbb{G}_s^0 \text{ for } s \in \mathcal{S} \,.$$

At the end of the next section, we derive the rate at which $\hat{\phi}$ and its components tend to ϕ_n^* and its components as n tends to infinity. As we will see, the quantity $\| \hat{\phi} - \phi_n^* \|^2$ plays the role of the variance component of the mean squared error introduced in Chapter 1.

Example 2 For each of the cases introduced in Example 1, we investigate the concavity properties of $l(f, \mathbf{w})$. Toward this end, let f_1, f_2 denote any two bounded functions in $\mathbb{F}$ and let g_1, g_2 be any two functions in $\mathbb{G}$. Then, for $\alpha \in (0,1)$ set

$$f_\alpha \;=\; f_1 + \alpha\big(f_2 - f_1\big) \qquad \text{and} \qquad g_\alpha \;=\; g_1 + \alpha\big(g_2 - g_1\big) \,.$$

In addition, we briefly discuss the existence and uniqueness of the functions $\phi^* \in \mathbb{F}$ and $\phi_n^* \in \mathbb{G}$ mentioned above.

[R] Let $\mathcal{S}$ be a hierarchical collection of subsets of $\{1, \ldots, M\}$. Differentiating $\Lambda(f_\alpha)$ twice with respect to α yields

$$\begin{aligned} \frac{d^2}{d\alpha^2} E\big[\, l(f_\alpha, \mathbf{W}) \,\big] \;&=\; -\frac{d^2}{d\alpha^2} E\Big[\, \big(f_\alpha(\mathbf{U}) - V\big)^2 \,\Big] \\ &=\; -E\Big[\, \big(f_1(\mathbf{U}) - f_2(\mathbf{U})\big)^2 \,\Big] \end{aligned}$$

which is strictly less than zero unless f_1 and f_2 are essentially equal. A similar argument shows that this result holds for $\Lambda(g_\alpha)$, and hence $\Lambda(\cdot)$ is strictly concave in the required sense. Replacing $\Lambda(\cdot)$ with $\ell(\cdot)$ and proceeding as above, we find that

$$\frac{d^2}{d\alpha^2} E_n\big[\, l(g_\alpha, \mathbf{W}) \,\big] \;=\; -E_n\Big[\, \big(g_1(\mathbf{U}) - g_2(\mathbf{U})\big)^2 \,\Big] \,.$$

Therefore, if $\mathbb{G}$ is identifiable, $\ell(\cdot)$ is strictly concave in the required sense. Furthermore, from Lemma 3.8 of Chapter 1 we know that there exists an

essentially unique function $\phi^* \in \mathbb{F}$ that maximizes $\Lambda(\,\cdot\,)$ over $\mathbb{F}$. By a simplification of that argument, it can be shown that there exists a unique function $\phi_n^* \in \mathbb{G}$ that maximizes $\Lambda(\,\cdot\,)$ over $\mathbb{G}$.

[GR] Again, let $\mathcal{S}$ be a hierarchical collection of subsets of $\{1, \ldots, M\}$. Recalling the definition of $l(f, \mathbf{w})$ in (1.4), we observe that the strict concavity condition is satisfied if

$$\begin{aligned} \frac{d^2}{d\alpha^2} E\big[\, l(\, f_\alpha, \mathbf{W}\,)\,\big] &= \frac{d^2}{d\alpha^2} E\Big[\, b_1\big(\, f_\alpha(\mathbf{U})\,\big)V + b_2\big(\, f_\alpha(\mathbf{U})\,\big)\,\Big] \\ &= -E\Big[\, \big(\, f_1(\mathbf{U}) - f_2(\mathbf{U})\,\big)^2 B(\, f_\alpha, \mathbf{U}\,)\,\Big] \qquad (1.14) \end{aligned}$$

where

$$B(\, f_\alpha, \mathbf{u}\,) = -\Big(\, b_1''\big(\, f_\alpha(\,\mathbf{u}\,)\,\big)\, b_3\big(\, \phi(\,\mathbf{u}\,)\,\big) + b_2''\big(\, f_\alpha(\,\mathbf{u}\,)\,\big)\,\Big)\,.$$

Now, assume that $\mathcal{V}$ is some interval such that $\nu(\,\mathcal{V}^c\,) = 0$, where ν is the measure appearing in the definition of this exponential family, and that

$$b_1''(\,\eta\,)\, v + b_2''(\,\eta\,) \;<\; 0 \qquad \text{for } \eta \in \mathbb{R} \text{ and } v \in \mathcal{V}\,. \qquad (1.15)$$

Therefore, the last expression in (1.14) is strictly negative unless f_1 and f_2 are essentially the same function. Repeating the above argument for g_α, we find that $\Lambda(\,\cdot\,)$ is strictly concave in the required sense. Replacing $\Lambda(\,\cdot\,)$ with $\ell(\,\cdot\,)$ and proceeding as above, we find that

$$\frac{d^2}{d\alpha^2} E_n\big[\, l(\, g_\alpha, \mathbf{W}\,)\,\big] = -E_n\Big[\, \big(\, g_1(\mathbf{U}) - g_2(\mathbf{U})\,\big)^2 B(\, g_\alpha, \mathbf{U}\,)\,\Big]\,.$$

Therefore, if $\mathbb{G}$ is identifiable, the functional $\ell(\,\cdot\,)$ is strictly concave in the required sense. Additionally, in Stone (1994) it is shown that there exists an essentially unique function $\phi^* \in \mathbb{F}$ that maximizes $\Lambda(\,\cdot\,)$ over $\mathbb{F}$. By a simplification of that argument, it can be shown that there exists a unique function $\phi_n^* \in \mathbb{G}$ that maximizes $\Lambda(\,\cdot\,)$ over $\mathbb{G}$.

[CR] As in the two previous regression contexts, let $\mathcal{S}$ be a hierarchical collection of subsets of $\{1, \ldots, M\}$. For convenience, set

$$\varphi_\alpha(\,\mathbf{u}\,) = \varphi\big(\, f_\alpha(\,\mathbf{u}\,)\,\big) \qquad \text{and} \qquad \Phi_\alpha(\,\mathbf{u}\,) = \Phi\big(\, f_\alpha(\,\mathbf{u}\,)\,\big)\,.$$

Then, by a straightforward computation, we find that

$$\frac{d^2}{d\alpha^2} E\big[\, l(\, f_\alpha, \mathbf{W}\,)\,\big] = -E\Big[\, \big(\, f_1(\mathbf{U}) - f_2(\,\mathbf{u}\,)\,\big)^2\, B\big(\, f_\alpha, \mathbf{W}\,\big)\,\Big]\,, \qquad (1.16)$$

where

$$B\big(f_\alpha, \mathbf{w}\big) \;=\; v_2 \;+\; (1-v_2)\,\frac{\varphi_\alpha(\,\mathbf{u}\,)}{1\,-\,\Phi_\alpha(\,\mathbf{u}\,)}\left[\,-\,f_\alpha(\,\mathbf{u}\,)\;+\;\frac{\varphi_\alpha(\,\mathbf{u}\,)}{1\,-\,\Phi_\alpha(\,\mathbf{u}\,)}\,\right].$$

We claim that the quantity in brackets on the right is strictly positive. To see this, observe that if Z has a standard normal distribution, then for $a \in \mathbb{R}$,

$$E\big[Z\,\big|\,Z > a\,\big] \;=\; \frac{\varphi(\,a\,)}{1\,-\,\Phi(\,a\,)}\,,$$

which is certainly larger that a. Therefore, we find that

$$-a \;+\; \frac{\varphi(\,a\,)}{1\,-\,\Phi(\,a\,)} \;>\; 0\,.$$

Replacing $f_\alpha(\,\mathbf{u}\,)$ for a in the above expression, we find that the right side of (1.16) is strictly negative unless f_1 and f_2 are essentially the same function. Repeating this argument for g_α, we find that $\Lambda(\,\cdot\,)$ is strictly concave in the required sense. Replacing $\Lambda(\,\cdot\,)$ with $\ell(\,\cdot\,)$ and proceeding as above, we find that

$$\frac{d^2}{d\alpha^2}\,E_n\big[\,l(\,g_\alpha\,,\mathbf{W}\,)\,\big] \;=\; -E_n\Big[\,\big(\,g_1(\mathbf{U}) - g_2(\,\mathbf{U}\,)\,\big)^2\,B\big(\,g_\alpha,\mathbf{W}\,\big)\,\Big]\,.$$

Therefore, if $\mathbb{G}$ is identifiable, the functional $\ell(\,\cdot\,)$ is concave in the required sense. Furthermore, following an argument similar to that in Stone (1994), it can be shown that there is an essentially unique function $\phi^* \in \mathbb{F}$ that maximizes $\Lambda(\,\cdot\,)$ over $\mathbb{F}$ in this context. Similarly, there exists a unique function $\phi_n^* \in \mathbb{G}$ maximizing $\Lambda(\,\cdot\,)$ in $\mathbb{G}$.

[PR] As in the two previous regression contexts, let $\mathcal{S}$ be a hierarchical collection of subsets of $\{1,\ldots,M\}$. Applying the definition of f_α to the case when f_1 and f_2 are vector-valued functions, we find that if

$$f_1 \;=\; (f_{11},\ldots,f_{1J}) \qquad \text{and} \qquad f_2 \;=\; (f_{21},\ldots,f_{2J})\,,$$

then

$$f_\alpha \;=\; \big(\,f_{11} \,+\, \alpha\,(f_{21} \,-\, f_{11}\,),\ldots,f_{1J} \,+\, \alpha\,(f_{2J} \,-\, f_{1J}\,)\,\big)\,.$$

For each fixed $\mathbf{u} \in \mathcal{U}$, define the quantity $\pi_0(\,f_\alpha\,) = \pi_0\big(\,f_\alpha(\mathbf{u})\,\big)$ to be

$$\Big(\,1 \,+\, \exp\big(\,f_{11} + \alpha(\,f_{21} - f_{11}\,)\,\big) \,+\, \cdots \,+\, \exp\big(\,f_{1J} + \alpha(\,f_{2J} - f_{1J}\,)\,\big)\,\Big)^{-1}$$

and for each $1 \le j \le J$, set

$$\pi_j(\,f_\alpha\,) \;=\; \pi_0(\,f_\alpha\,)\,\exp\big(\,f_{11} + \alpha(\,f_{21} - f_{11}\,)\,\big)\,,$$

where $\pi_j(f_\alpha) = \pi_j\big(f_\alpha(\mathbf{u})\big)$ for $0 \le j \le J$. Now, by a straight application of the chain rule, we find that the quantity

$$\frac{d^2}{d\alpha^2} E\big[\, l(f_\alpha, \mathbf{W}) \,|\, \mathbf{U} = \mathbf{u}\big] \tag{1.17a}$$

is given by

$$-\sum_{j=1}^{J} \pi_j(f_\alpha)(f_{2j} - f_{1j})^2 + \left(\sum_{j=1}^{J} \pi_j(f_\alpha)(f_{2j} - f_{1j})\right)^2. \tag{1.17b}$$

However, by the Schwartz inequality, we find that the expression on the right is bounded by

$$\begin{aligned}\left(\sum_{j=1}^{J} \pi_j(f_\alpha)(f_{2j} - f_{1j})\right)^2 &\le \sum_{j=1}^{J} \pi_j(f_\alpha) \sum_{j=1}^{J} \pi_j(f_\alpha)(f_{2j} - f_{1j})^2 \\ &= \big(1 - \pi_0(f_\alpha)\big) \sum_{j=1}^{J} \pi_j(f_\alpha)(f_{2j} - f_{1j})^2\end{aligned}$$

and hence (1.17) yields

$$\frac{d^2}{d\alpha^2} E\big[\, l(f_\alpha, \mathbf{W}) \,|\, \mathbf{U} = \mathbf{u}\,\big] \;\le\; -\pi_0(f_\alpha) \sum_{j=1}^{J} \pi_j(f_\alpha)(f_{2j} - f_{1j})^2. \tag{1.18}$$

Since f_1, f_2 are bounded in absolute value, each of the functions $\pi_j\big(f_\alpha(\mathbf{u})\big)$, $0 \le j \le J$, are bounded away from zero and 1. Therefore, taking an expectation with respect to the distribution of $\mathbf{U}$ in (1.18), we find that unless f_{1j} and f_{2j} are essentially equal for all $1 \le j \le J$, the function $\Lambda(f_\alpha)$ is strictly concave in α. Repeating the same argument for g_α, we find that $\Lambda(\cdot)$ is concave in the required sense. Replacing $\Lambda(\cdot)$ with $\ell(\cdot)$ and repeating the above argument, we find that

$$\frac{d^2}{d\alpha^2} E_n\big[\, l(g_\alpha, \mathbf{W})\,\big]$$

is bounded from above by

$$-E_n\left[\pi_0\big(g_\alpha(\mathbf{U})\big) \sum_{j=1}^{J} \pi_j\big(g_\alpha(\mathbf{U})\big)\big(g_{2j}(\mathbf{U}) - g_{1j}(\mathbf{U})\big)^2\right].$$

Therefore, if $\mathbb{G}$ is identifiable, then the functional $\ell(\cdot)$ is strictly concave in the required sense. Furthermore, following an argument similar to that in Stone (1994), we find that there exists an essentially unique function $\phi^* \in \mathbb{F}$ that maximizes $\Lambda(\cdot)$ over $\mathbb{F}$. Similarly, there exists a unique function $\phi_n^* \in \mathbb{G}$ that maximizes $\Lambda(\cdot)$ over $\mathbb{G}$.

[D] Let $\mathcal{S}_1$ be any hierarchical collection of sets of $\{1,\ldots,M\}$ and set $\mathcal{S} = \mathcal{S}_1 \setminus \{\varnothing\}$. Recall the definition of the functional $c(\,\cdot\,)$ in this context, and observe that for any pair of bounded functions $f_1, f_2 \in \mathbb{F}$,

$$\begin{aligned} \frac{d^2}{d\alpha^2} E\big[\, l(\, f_\alpha, \mathbf{W}\,)\,\big] &= \frac{d^2}{d\alpha^2} E\big[\, f_\alpha(\mathbf{U}) \,-\, c(f_\alpha)\,\big] \\ &= -\,\mathrm{var}\big(\, f_1(\,\mathbf{U}_\alpha\,) \,-\, f_2(\,\mathbf{U}_\alpha\,)\,\big)\,, \qquad (1.19) \end{aligned}$$

where the random vector $\mathbf{U}_\alpha$ has the density on $\mathcal{U}$ given by

$$f_{\mathbf{U}_\alpha} = \exp\Big(\, f_1 + \alpha(f_2 - f_1) - c\big(\, f_1 + \alpha(f_2 - f_1)\,\big)\,\Big)\,. \qquad (1.20)$$

Therefore, setting $f_0 = f_1 - f_2$, we find that the last expression in (1.19) is strictly less than zero unless f_0 is essentially a constant function in $\mathbf{u}$. However, since f_1 and f_2 are each in $\mathbb{F}$, if $f_0 = a$ for some $a \in \mathbb{R}$, then

$$a^2 = \langle\, f_1 - f_2, f_0\,\rangle = \langle\, f_1, a\,\rangle - \langle\, f_2, a\,\rangle = 0$$

by the definition of $\mathbb{F}$. Arguing similarly for g_α, we again find that the second derivative of $\Lambda(\, g_\alpha\,)$ with respect to α is strictly less than zero unless the function $g_0 = g_1 - g_2$ is essentially a constant function in $\mathbf{u}$. However, since g_1 and g_2 are each in $\mathbb{G}$, if $g_0 = a$ for some $a \in \mathbb{R}$, then $a = 0$ by the definition of $\mathbb{G}$. Therefore, $\Lambda(\,\cdot\,)$ is strictly concave in the required sense. Replacing $\Lambda(\,\cdot\,)$ with $\ell(\,\cdot\,)$ and arguing as above, we find that

$$\frac{d^2}{d\alpha^2} E_n\big[\, l(\, g_\alpha, \mathbf{W}\,)\,\big] = -\,\mathrm{var}\big(\, g_1(\,\mathbf{U}_\alpha\,) \,-\, g_2(\,\mathbf{U}_\alpha\,)\,\big)\,.$$

The righthand side of this expression is the same as that appearing in the second equality of (1.19). Therefore, repeating the arguments above we conclude $\ell(\,\cdot\,)$ is concave in the required sense. Furthermore, in Stone (1994), it is shown that there exists an essentially unique function $\phi^* \in \mathbb{F}$ that maximizes $\Lambda(\,\cdot\,)$ over $\mathbb{F}$. By a simplification of that argument, it can be shown that there exists a unique function $\phi_n^* \in \mathbb{G}$ maximizing $\Lambda(\,\cdot\,)$ over $\mathbb{G}$.

In the next section we derive the rate of convergence of the sample based estimate $\hat{\phi}$ and its components to ϕ^* and its components in the context of extended linear models. As in Chapter 1, our goal will be to bound separately the variance and bias components of the mean squared error. In this context, the variance term is of the form $\|\,\hat{\phi} - \phi_n^*\,\|^2$ while the bias term is given by $\|\,\phi_n^* - \phi^*\,\|^2$.

3.2 Rate of Convergence

Suppose that the triangulations $\triangle_l$, $1 \leq l \leq M$, are refined with sample size so that Condition 3′ holds. Then we know that with probability approaching one, the space $\mathbb{G}_1$ and hence the space $\mathbb{G}$ are identifiable. In addition, in the case of ordinary least squares estimation, this condition is sufficient to guarantee the rate of convergence for $\hat{\mu}$ and its components established in Theorem 4.1 of Chapter 1. Unfortunately, we need to strengthen this assumption slightly when we consider the extended linear models discussed in the previous section. Toward this end, recall the definition of $\overline{h}_s$ and $\underline{h}_s$ from Chapter 1, and for any collection of subsets $\mathcal{S}$ of $\{1, \ldots, M\}$, set

$$\rho^2(\overline{h}, \mathcal{S}) \;=\; \sum_{s \in \mathcal{S}} \rho_s^2(\overline{h}_s) \qquad \text{and} \qquad \omega^2(\underline{h}, \mathcal{S}) \;=\; \sum_{s \in \mathcal{S}} \underline{h}^{-2d_s},$$

which we recognize as ingredients in the rate of convergence of $\hat{\mu}$ to μ^* if $\mathcal{S}$ is hierarchical. We require the following condition on the quantities $\rho(\overline{h}, \mathcal{S})$ and $\omega(\underline{h}, \mathcal{S})$.

Condition 3″ *Assume that the triangulations $\triangle_l$, $1 \leq l \leq M$, are refined with sample size so that*

$$\omega^2(\underline{h}, \mathcal{S}) \;=\; o\big(n^{1-\epsilon}\big) \qquad \textit{and} \qquad \omega(\underline{h}, \mathcal{S})\, \rho^2(\overline{h}, \mathcal{S}) \;=\; o\big(1\big),$$

where ϵ is some positive constant.

At first glance this condition seems somewhat difficult to verify in practice. In truth, however, these expressions lead to a very elegant condition on the vectors $\overline{h}_s$ and $\underline{h}_s$, $s \in \mathcal{S}_1$. Given positive constants $T_1, \ldots, T_M$, assume that for each $s \in \mathcal{S}_1$, the function $\rho_s^2(\overline{h}_s)$ is of the form

$$\rho_s^2(\overline{h}_s) \;=\; \sum_{l \in s} T_l\, \overline{h}_l^{\,2p_l}, \qquad \text{where } p_l \in \mathbb{N} \text{ for } 1 \leq l \leq M.$$

Accordingly, recall from Example 2 in Chapter 1 that if the quantities $\overline{h}_l$ and $\underline{h}_l$ are chosen according to the prescription in (4.19) of that chapter, then $\rho^2(\overline{h}, \mathcal{S})$ and $\omega(\underline{h}, \mathcal{S}) / n$ are each $O(n^{-\gamma})$ where γ is given by

$$\gamma \;=\; \frac{2}{2 + [\, d_s / p_s \,]},$$

for some maximal $s \in \mathcal{S}$. Therefore, we have that $\omega^2(\underline{h}, \mathcal{S})$ is $O(n^{2(1-\gamma)})$, and hence the first relation in Condition 3″ holds if

$$[\, d_s / p_s \,] \;<\; 2. \tag{2.1}$$

In addition, we find that the quantity $\omega(\underline{h}, \mathcal{S})\,\rho^2(\overline{h}, \mathcal{S})$ is $O(n^{1-2\gamma})$, and hence that under the prescription in (4.19) of Chapter 1, (2.1) is also sufficient to guarantee the second relation in Condition 3″.

Population Model (Bias)

Suppose $l(f, \mathbf{w})$ satisfies the smoothness assumptions described above and that $\Lambda(\cdot)$ is strictly concave in the required sense. Let f_1, f_2 be two essentially different bounded functions in $\mathbb{F}$, and let g_1, g_2 be any two different functions in $\mathbb{G}$. For $\alpha \in [0, 1]$, set

$$f_\alpha \;=\; f_1 + \alpha(f_2 - f_1) \qquad \text{and} \qquad g_\alpha \;=\; g_1 + \alpha(g_2 - g_1)\,. \tag{2.2}$$

Then, using Taylor's theorem with integral remainder, we find that

$$\Lambda(f_2) \;=\; \Lambda(f_1) \;+\; \frac{d}{d\alpha}\Lambda(f_\alpha)\Big|_{\alpha=0} \;+\; \int_0^1 (1-\alpha)\,\frac{d^2}{d\alpha^2}\Lambda(f_\alpha)\,d\alpha\,. \tag{2.3a}$$

Recall that ϕ^* is the essentially unique bounded function in $\mathbb{F}$ that maximizes the functional $\Lambda(\cdot)$ over $\mathbb{F}$. Therefore, applying the above equality for $f_1 = \phi^*$ and f_2 any other bounded function in $\mathbb{F}$, we find that

$$\Lambda(\phi^*) \;-\; \Lambda(f) \;=\; -\int_0^1 (1-\alpha)\,\frac{d^2}{d\alpha^2}\Lambda\big(\phi^* + \alpha(f - \phi^*)\big)\,d\alpha\,. \tag{2.3b}$$

Recalling that ϕ_n^* is the function that maximizes $\Lambda(\cdot)$ over $\mathbb{G}$, we can derive similar expressions for ϕ_n^* and any other function $g \in \mathbb{G}$.

Observe that by the strict concavity condition (1.13), the quantity appearing on righthand side of equation (2.3b) is strictly positive. In the following lemma, we present a pair of inequalities based on this expression that bound the size of the difference $\Lambda(\phi^*) - \Lambda(f)$ for $f \in \mathbb{F}$. An equivalent expression for ϕ_n^* and any $g \in \mathbb{G}$ can be derived in the same way. For convenience, given any bounded function $f \in \mathbb{F}_1$, we let $\sigma(f)$ denote the function given by $\sigma(f) = f - c(f)$.

Lemma 2.1 *Assume that Condition 2′ holds. Then, for each positive constant T_1, there is a positive constant T_2 such that*

$$\frac{1}{T_2}\,\big\|\,\sigma(\phi^*) - \sigma(f)\,\big\|^2 \;\le\; \Lambda(\phi^*) - \Lambda(f) \;\le\; T_2\,\big\|\,\sigma(\phi^*) - \sigma(f)\,\big\|^2 \tag{2.4}$$

for all functions $f \in \mathbb{F}_1$ such that $\sigma(f)$ is bounded in absolute value by T_1.

Proof Let $T_3, T_4, \ldots$ denote suitable positive constants. Consider first the real-valued regression contexts discussed in Examples 1 and 2 and for two essentially different bounded functions $f_1, f_2 \in \mathbb{F}_1$, define f_α as in (2.2). In each of these cases, $\sigma(\,\cdot\,)$ is just the identity operator, and the functional $\Lambda(f)$ depends on f only through the quantity $f(\mathbf{U})$. Then, as we have seen,

$$\frac{d^2}{d\alpha^2}\,\Lambda(f_\alpha) \;=\; -E\Big[\big(f_1(\mathbf{U}) - f_2(\mathbf{U})\big)^2\, B\big(f_\alpha, \mathbf{W}\big)\Big],$$

for some appropriate function $B(\,\cdot\,,\cdot\,)$. Observe, however, that in each of these cases,

$$E\Big[B\big(f_\alpha, \mathbf{W}\big)\,\Big|\,\mathbf{U} = \mathbf{u}\Big]$$

is strictly positive and bounded away from zero and infinity uniformly over all $\alpha \in [0,1]$, $\mathbf{u} \in \mathcal{U}$ and $f \in \mathbb{F}$ such that $\|f\|_\infty \le T_1$. The desired inequalities (2.4) now follow from these observations and the boundedness assumptions on ϕ^* and f.

In the case of polychotomous regression, observe that from (1.17b),

$$\frac{d^2}{d\alpha^2}\,E\big[\,l(f_\alpha, \mathbf{W})\,|\,\mathbf{U} = \mathbf{u}\big] \;\ge\; -\sum_{j=1}^{J} \pi_j\big(f_\alpha(\mathbf{u})\big)\,\big(f_{2j}(\mathbf{u}) - f_{1j}(\mathbf{u})\big)^2.$$

The second inequality in (2.4) now follows in this context by the definition of $\|\cdot\|^2$ for vector valued functions, the continuity of the strictly positive functions $\pi_j(f_\alpha)$, and the boundedness assumptions on $f = (f_1, \ldots, f_J)$ and ϕ^*. Similarly, the first inequality in (2.4) follows from the relation

$$\frac{d^2}{d\alpha^2}\,E\big[\,l(f_\alpha, \mathbf{W})\,|\,\mathbf{U} = \mathbf{u}\,\big] \;\le\; -\pi_0\big(f_\alpha(\mathbf{u})\big)\sum_{j=1}^{J} \pi_j\big(f_\alpha(\mathbf{u})\big)\,\big(f_{2j}(\mathbf{u}) - f_{1j}(\mathbf{u})\big)^2.$$

which is equation (1.18) in Example 2.

Finally, consider the case of density estimation. In what follows, let $f_1 = \sigma(\phi^*)$ and $f_2 = \sigma(f)$ for any function $f \in \mathbb{F}_1$ satisfying $\|f\|_\infty \le T_1$. Then, defining f_α as in (2.2), we find from Example 2 that

$$\frac{d^2}{d\alpha^2}\,E\big[\,l(f_\alpha, \mathbf{W})\,\big] \;=\; -\mathrm{var}\big(\,f_1(\mathbf{U}_\alpha) \;-\; f_2(\mathbf{U}_\alpha)\,\big), \tag{2.5}$$

where the random vector $\mathbf{U}_\alpha$ has a density $f_{\mathbf{U}_\alpha}$ on $\mathcal{U}$ given by (1.17). Now, observe that independently of our choice of $f \in \mathbb{F}_1$ used to define f_2 subject to the constraint that $\|f\|_\infty \le T_1$, there exists a positive constant T_3 such that

$$\frac{1}{T_3} \;\le\; f_{\mathbf{U}_\alpha}(\mathbf{u}) \;\le\; T_3, \qquad \alpha \in [0,1] \text{ and } \mathbf{u} \in \mathcal{U}.$$

As an immediate consequence of this fact, we find from Condition 2′ that

$$\begin{aligned} \mathrm{var}\left(f_1(\mathbf{U}_\alpha) - f_2(\mathbf{U}_\alpha) \right) &\leq E\Big[\left(f_1(\mathbf{U}_\alpha) - f_2(\mathbf{U}_\alpha) \right)^2 \Big] \\ &\leq M_3 T_3 \left\| f_1(\mathbf{U}) - f_2(\mathbf{U}) \right\|^2 . \end{aligned}$$

Combining this result with (2.5) and (2.3b) and redefining T_3 if necessary, we obtain the second inequality in (2.4).

Next, from Condition 2′ and the bound on $f_{\mathbf{U}_\alpha}$ we also find that

$$\mathrm{var}\left(f_1(\mathbf{U}) - f_2(\mathbf{U}) \right) \leq M_3 T_3 \, \mathrm{var}\left(f_1(\mathbf{U}_\alpha) - f_2(\mathbf{U}_\alpha) \right).$$

Therefore, to establish the second relation in (2.4), it is sufficient to show that there exists a constant $\epsilon > 0$ such that

$$\mathrm{var}\left(f_1(\mathbf{U}) - f_2(\mathbf{U}) \right) \geq \epsilon E\Big[\left(f_1(\mathbf{U}) - f_2(\mathbf{U}) \right)^2 \Big],$$

for all functions of the form $f_2 = \sigma(f)$, where $f \in \mathbb{F}_1$ and $\| f \|_\infty \leq T_1$. Toward this end, suppose we cannot find such a constant and construct a sequence of bounded functions f_i, $i \geq 1$, from $\mathbb{F}_1$ such that each is bounded in absolute value by T_1 and

$$\mathrm{var}\left(f_1(\mathbf{U}) - f_i(\mathbf{U}) + c(f_i) \right) < \frac{1}{i} E\Big[\left(f_1(\mathbf{U}) - f_i(\mathbf{U}) + c(f_i) \right)^2 \Big]. \quad (2.7)$$

Suppose, additionally, that

$$\liminf_{i \to \infty} E\Big[\left(f_1(\mathbf{U}) - f_i(\mathbf{U}) + c(f_i) \right)^2 \Big] > 0. \quad (2.8)$$

Note that

$$\lim_{i \to \infty} \mathrm{var}\left(f_1(\mathbf{U}) - f_i(\mathbf{U}) + c(f_i) \right) = 0, \quad (2.9)$$

and hence that

$$\lim_{i \to \infty} E\Big[f_1(\mathbf{U}) - f_i(\mathbf{U}) + c(f_i) \Big] = 0. \quad (2.10)$$

To see this last equality, recall that by construction, the functions f_1 and $f_i - c(f_i)$ are log-densities over $\mathcal{U}$, and hence

$$\begin{aligned} 1 &= \int_{\mathcal{U}} \exp\left(f_i(\mathbf{u}) - c(f_i) \right) d\mathbf{u} \\ &= E\Big[\exp\left(f_i(\mathbf{U}_0) - c(f_i) - f_1(\mathbf{U}_0) \right) \Big] \end{aligned} \quad (2.11)$$

where $\mathbf{U}_0$ denotes the random vector on $\mathcal{U}$ having log-density f_1 corresponding to the choice of $\alpha = 0$ in (2.2). According to (2.9) and the boundedness assumptions

on the functions ϕ^* and f_i, we know that in the limit, the function $f_1 - f_i + c(f_i)$ is essentially a constant. However, we see from (2.11) that this constant must be zero. Finally, combining (2.10) with (2.7), we arrive at a contradiction to (2.8). Suppose instead that

$$\lim_{i\to\infty} E\Big[\big(f_1(\mathbf{U}) - f_2(\mathbf{U}) + c(f_i) \big)^2 \Big] = 0 .$$

By the bound on the density $f_{\mathbf{U}_\alpha}$ and Condition 2′, we have that

$$\lim_{i\to\infty} E\Big[\big(f_1(\mathbf{U}_0) - f_2(\mathbf{U}_0) + c(f_i) \big)^2 \Big] = 0 . \tag{2.12}$$

Hence, from (2.11) and Taylor's theorem, we find that there exists a bounded function $B(\mathbf{u})$ such that for all i sufficiently large

$$-E\Big[f_i(\mathbf{U}_0) - c(f_i) - f_1(\mathbf{U}_0) \Big] = E\Big[B(\mathbf{U}_0) \big(f_i(\mathbf{U}_0) - c(f_i) - f_1(\mathbf{U}_0) \big)^2 \Big] .$$

However, from Condition 2′ and the bound on $f_{\mathbf{U}_0}$, it follows from (2.7) that

$$\mathrm{var}\big(f_1(\mathbf{U}_0) - f_i(\mathbf{U}_0) + c(f_i) \big) < M_3^2 T_3^2 \frac{1}{i} E\Big[\big(f_1(\mathbf{U}_0) - f_i(\mathbf{U}_0) + c(f_i) \big)^2 \Big] .$$

which yields a contradiction to the statement in (2.12). □

Lemma 2.2 *Assume that Conditions* 1, 2′, 3″ *and* 5′ *hold. Then,*

$$\big\| \sigma(\phi_n^*) - \sigma(\phi^*) \big\|^2 = O\Big(\rho^2(\overline{h}, \mathcal{S}) \Big) ,$$

and

$$\big\| \sigma(\phi_n^*) - \sigma(\phi^*) \big\|_\infty^2 = O\Big(\omega(\underline{h}, \mathcal{S})\, \rho^2(\overline{h}, \mathcal{S}) \Big) .$$

Proof Throughout this proof, let $T_1, T_2, \ldots$ denote suitable positive constants. According to Condition 5′, there exists a positive constant T_1, a function a $g_n \in \mathbb{G}$ and a constant $a \in \mathbb{R}$ such that

$$\big\| g_n + a - \phi^* \big\|_\infty^2 \le T_1\, \rho^2(\overline{h}, \mathcal{S}) ,$$

where in the each of the regression contexts, we take $a = 0$. Therefore, by the continuity of the operator $\sigma(\cdot)$ we find that there exists a positive constant T_2 such that

$$\big\| \sigma(g_n) - \sigma(\phi^*) \big\|_\infty^2 = \big\| \sigma(g_n + a) - \sigma(\phi^*) \big\|_\infty^2 \le T_2 \big\| g_n + a - \phi^* \big\|_\infty^2$$

and hence

$$\big\| \sigma(g_n) - \sigma(\phi^*) \big\|^2 \le \big\| \sigma(g_n) - \sigma(\phi^*) \big\|_\infty^2 \le T_2\, \rho^2(\overline{h}, \mathcal{S}) , \tag{2.13}$$

redefining T_2 if necessary. Thus, by the second inequality in Lemma 2.1 there exists a positive constant T_3 such that

$$\Lambda(\phi^*) - \Lambda(g_n) \leq T_3 \rho^2(\overline{h}, \mathcal{S}), \tag{2.14}$$

Next, let $b \in \mathbb{R}$ denote a large positive constant, and choose $g \in \mathbb{G}$ so that

$$\| \sigma(g) - \sigma(\phi^*) \|^2 = b \rho^2(\overline{h}, \mathcal{S}). \tag{2.15}$$

Then, by the triangle and Schwartz inequalities, we find that

$$\begin{aligned} \| \sigma(g_n) - \sigma(g) \|^2 &\leq 2 \left(\| \sigma(g_n) - \sigma(\phi^*) \|^2 + \| \sigma(g) - \sigma(\phi^*) \|^2 \right) \\ &\leq 2 (b + T_2) \rho^2(\overline{h}, \mathcal{S}), \end{aligned}$$

and hence by Lemma 1 in the Appendix there exists a positive constant T_4 such that

$$\| \sigma(g_n) - \sigma(g) \|_\infty^2 \leq 2 T_4 (b + T_2) \omega(\underline{h}, \mathcal{S}) \rho^2(\overline{h}, \mathcal{S}).$$

Therefore, from Condition 3'' we find that for n sufficiently large,

$$\begin{aligned} \| \sigma(g) \|_\infty &\leq \| \sigma(g_n) - \sigma(g) \|_\infty + \| \sigma(g_n) - \sigma(\phi^*) \|_\infty + \| \sigma(\phi^*) \|_\infty \\ &\leq 1 + \| \sigma(\phi^*) \|_\infty. \end{aligned}$$

Applying the first inequality in Lemma 2.1, we find that there exists a positive constant T_5 such that for all $g \in \mathbb{G}$ satisfying (2.15) and all n sufficiently large,

$$\Lambda(\phi^*) - \Lambda(g) \geq b T_5 \rho^2(\overline{h}, \mathcal{S}). \tag{2.16}$$

Choosing $b \in \mathbb{R}$ large enough so that $b > T_2$ and $b T_5 > T_3$, we find from the inequalities (2.14) and (2.16) that for all $g \in \mathbb{G}$ satisfying (2.15) and all n sufficiently large,

$$\Lambda(g) < \Lambda(g_n). \tag{2.17}$$

Suppose now that there exist arbitrary large values of n such that

$$\| \sigma(\phi_n^*) - \sigma(\phi^*) \|^2 > b \rho^2(\overline{h}, \mathcal{S}).$$

Then, for $\alpha \in [0, 1]$, define the function $g_\alpha = g_n + \alpha(\phi_n^* - g_n)$ and observe that by continuity, for some $\alpha' \in [0, 1]$, the function $\sigma(g_{\alpha'})$ satisfies (2.15). Hence, by (2.17), we have that

$$\Lambda(g_{\alpha'}) < \Lambda(g_n),$$

which is impossible since $\Lambda(\,\cdot\,)$ is a strictly concave function on $\mathbb{G}$. Therefore, we conclude that for all n sufficiently large,

$$\left\| \sigma(\,\phi_n^*\,) - \sigma(\,\phi^*\,) \right\|^2 \;\leq\; b\,\rho^2(\,\overline{h}, \mathcal{S}\,)\,, \tag{2.18}$$

which is the first relation in the statment of the lemma. As for the second, by the triangle and Schwartz inequalities,

$$\left\| \sigma(\,g_n\,) - \sigma(\,\phi_n^*\,) \right\|^2 \;\leq\; 2\left\| \sigma(\,g_n\,) - \sigma(\,\phi^*\,) \right\|^2 \;+\; 2\left\| \sigma(\,\phi_n^*\,) - \sigma(\,\phi^*\,) \right\|^2,$$

so that from (2.13) and (2.18),

$$\left\| \sigma(\,g_n\,) - \sigma(\,\phi_n^*\,) \right\|^2 \;=\; O\big(\,\rho^2(\,\overline{h}, \mathcal{S}\,)\,\big)\,.$$

Combining the last inequality with the result of Lemma 1 in the Appendix yields the desired result. □

Lemma 2.3 *Suppose that Conditions* 1, 2′, 3″, *and* 5′ *hold. Then,*

$$\left\| \phi_{ns}^* \;-\; \phi_s^* \right\|^2 \;=\; O_P\big(\,\rho^2(\,\overline{h}, \mathcal{S}\,) \;+\; \omega(\,\underline{h}, \mathcal{S}\,)/\,n\,\big)\,, \qquad s \in \mathcal{S}\,.$$

Proof Throughout this proof, let $T_1, T_2, \ldots$ denote suitable positive constants. Consider first the case when ϕ is real-valued. Assume that the space $\mathbb{PP}$ and hence the space $\mathbb{G}$ are identifiable, and let $\tilde{g}_n$ denote the orthogonal projection of $\sigma(\,\phi^*\,)$ onto $\mathbb{G}_1$ relative to the empirical inner product defined in (1.6b). Then, we can express $\tilde{g}_n$ uniquely as

$$\tilde{g}_n \;=\; \sum_{s \in \mathcal{S}_1} \tilde{g}_{ns}\,, \qquad \text{where } \tilde{g}_{ns} \in \mathbb{G}_s^0 \text{ for } s \in \mathcal{S}_1\,.$$

It follows from Theorem 4.1 in Chapter 1 that

$$\left\| \tilde{g}_n \;-\; \sigma(\,\phi^*\,) \right\|^2 \;=\; O_P\big(\,\rho^2(\,\overline{h}, \mathcal{S}\,) \;+\; \omega(\,\underline{h}, \mathcal{S}\,)/\,n\,\big)\,, \tag{2.19a}$$

and that for each set $s \in \mathcal{S}$,

$$\left\| \tilde{g}_{ns} \;-\; \phi_s^* \right\|^2 \;=\; O_P\big(\,\rho^2(\,\overline{h}, \mathcal{S}\,) \;+\; \omega(\,\underline{h}, \mathcal{S}\,)/\,n\,\big)\,. \tag{2.19b}$$

Suppose in addition that the size of $\sigma(\,\tilde{g}_n\,) - \sigma(\,\phi^*\,)$ is also given by

$$\left\| \sigma(\,\tilde{g}_n\,) \;-\; \sigma(\,\phi^*\,) \right\|^2 \;=\; O_P\big(\,\rho^2(\,\overline{h}, \mathcal{S}\,) \;+\; \omega(\,\underline{h}, \mathcal{S}\,)/\,n\,\big)\,. \tag{2.20}$$

Then, by the triangle and Schwartz inequalities we find that

$$\left\| \sigma(\,\tilde{g}_n\,) \;-\; \sigma(\,\phi_n^*\,) \right\|^2 \;\leq\; 2\left\| \sigma(\,\tilde{g}_n\,) \;-\; \sigma(\,\phi^*\,) \right\|^2 \;+\; 2\left\| \sigma(\,\phi_n^*\,) \;-\; \sigma(\,\phi^*\,) \right\|^2,$$

and hence from (2.20) and the first expression in Lemma 2.2,

$$\left\| \sigma(\tilde{g}_n) - \sigma(\phi_n^*) \right\|^2 = O_P\left(\rho^2(\overline{h}, \mathcal{S}) + \omega(\underline{h}, \mathcal{S})/n \right). \tag{2.21}$$

However, by the definition of $\sigma(\cdot)$, we can write

$$\sigma(\tilde{g}_n) - \sigma(\phi_n^*) = \sum_{s \in \mathcal{S}_1} \left(\tilde{g}_{ns} - \phi_{ns}^* \right) - \left(c(\tilde{g}_n) - c(\phi_n^*) \right),$$

where $\phi_{ns}^* = 0$ if $s = \varnothing$ in the density estimation context. Then, from (2.21) and Lemma 4.3 in Chapter 1, we find that for each $s \in \mathcal{S}$,

$$\left\| \tilde{g}_{ns} - \phi_{ns}^* \right\|^2 = O_P\left(\rho^2(\overline{h}, \mathcal{S}) + \omega(\underline{h}, \mathcal{S})/n \right). \tag{2.22}$$

The desired relation now follows immediately from (2.19b) and (2.22). All that remains is to demonstrate the relationship (2.20).

In the regression contexts, since $\sigma(\cdot)$ is the identity operator, (2.20) holds automatically if (2.19a) holds. In the density estimation case, observe that by (2.19a) and Lemma 1 in the Appendix,

$$\left\| \tilde{g}_n - \sigma(\phi^*) \right\|_\infty^2 = O_P\left(\omega(\underline{h}, \mathcal{S}) \left(\rho^2(\overline{h}, \mathcal{S}) + \omega(\underline{h}, \mathcal{S})/n \right) \right),$$

which we know to be $o_P(1)$ by Condition 3''. Therefore, by the continuity condition (1.5), there exists a constant T_1 such that except on a set with probability tending to zero with n,

$$\left| c(\tilde{g}_n) \right|^2 = \left| c(\tilde{g}_n) - c\left(\sigma(\phi^*) \right) \right|^2 \leq T_1 \left\| \tilde{g}_n - \sigma(\phi^*) \right\|^2.$$

Therefore, by one final application of the triangle and Schwartz inequalities, the relation in (2.20) now follows from (2.19a).

The case when ϕ is vector-valued follows from essentially the same argument given above. In this case however, since the function $\phi^* = (\phi_1^*, \ldots, \phi_J^*)$, we take $\tilde{g}_n = (\tilde{g}_{1n}, \ldots, \tilde{g}_{Jn})$, where for $1 \leq j \leq J$, the function $\tilde{g}_{jn}$ is the orthogonal projection of ϕ_j^* onto the space of real-valued functions $\mathbb{G}$ given in (1.9b). Aside from this minor change, the rest of the argument carries through using the properties of the norms for vector-valued functions defined in (1.11). □

Sample-Based Estimates (Variance)

Recall that the general form of the functional $l(f, \mathbf{w})$ corresponding to an extended linear model is given by

$$l(f, \mathbf{w}) = \sum_{k=1}^{K} b_k(f, \mathbf{u}) B_k(\mathbf{w}), \qquad \mathbf{w} = (\mathbf{u}, \mathbf{v}) \in \mathcal{W}. \tag{2.23}$$

In addition to the concavity requirements mentioned in the previous section, we must also impose regularity conditions on both the functionals b_k and the random variables $B_k(\mathbf{W})$. To be precise, we assume that the functionals b_k, $1 \le k \le K$, satisfy a Lipshitz continuity condition in f. Specifically, it is assumed that for each positive constant T_1 there exist positive constants T_2 and T_3 such that for all pairs of functions f_1, f_2 defined on $\mathcal{U}$ satisfying (1.5a), we have that

$$\left\| b_k(f_1, \cdot) - b_k(f_2, \cdot) \right\| \;\le\; T_2 \left\| \sigma(f_1) - \sigma(f_2) \right\| \tag{2.24a}$$

and that

$$\left\| b_k(f_1, \cdot) - b_k(f_2, \cdot) \right\|_\infty \;\le\; T_3 \left\| \sigma(f_1) - \sigma(f_2) \right\|_\infty . \tag{2.24b}$$

In order to satisfy the strict concavity condition (1.13), we also must require that for all bounded functions $f_1, f_2 \in \mathbb{F}_1$ the functionals

$$b_k\big(f_1 + \alpha(f_2 - f_1), \mathbf{u}\big), \qquad 1 \le k \le K,$$

are twice continuously differentiable in α over $(0,1)$ for all $\mathbf{u} \in \mathcal{U}$. If the functionals b_k depend on f only through the values $f(\mathbf{u})$, then these properties follow immediately if the b_k are sufficiently smooth.

As for the functions B_k, assume that there exists positive constants M_4, M_5, and M_6 such that

$$E\big[\, B_k(\mathbf{W}) \,|\, \mathbf{U} = \mathbf{u} \,\big] \;<\; M_4 \qquad \mathbf{u} \in \mathcal{U}, \tag{2.25a}$$

and that for all real numbers $t_1, \ldots, t_K$ bounded in absolute value by M_5,

$$E\Big[\, \exp\big(t_1 B_1(\mathbf{W}) + \cdots + t_K B_K(\mathbf{W})\big) \,\Big|\, \mathbf{U} = \mathbf{u} \Big] \;\le\; M_6 \tag{2.25b}$$

for all $\mathbf{u} \in \mathcal{U}$. Put more simply, for each $1 \le k \le K$, let $Y_k = B_k(\mathbf{W})$. Then, (2.25a) is equivalent to assuming that for each $1 \le k \le K$, the conditional mean $E(Y_k \,|\, \mathbf{U} = \mathbf{u})$ is bounded uniformly on $\mathcal{U}$. The boundedness condition in (2.25b) is equivalent to requiring that given $\mathbf{U} = \mathbf{u}$, the the conditional moment generating function of $(Y_1, \ldots, Y_K)$ exists and is bounded in a neighborhood of the origin uniformly on $\mathcal{U}$.

Example 3 **[R]** Recall from Example 2 that in the regression context

$$l(f, \mathbf{w}) \;=\; -f^2(\mathbf{u}) \;+\; 2\, f(\mathbf{u})\, v\,.$$

In Chapter 1, the requirement (2.25a) was also necessary, however the expression in (2.25b) implies that the conditional distribution of V given $\mathbf{U} = \mathbf{u}$

has bounded moments of all orders. This represents a strenghtening of the assumption that $\sigma^2(\mathbf{u}) = \mathrm{var}(V \mid \mathbf{U} = \mathbf{u})$ is bounded over $\mathbf{u} \in \mathcal{U}$ considered in Chapter 1.

[GR] In the context of generalized regression,

$$l(f, \mathbf{w}) \;=\; b_1\big(f(\mathbf{u})\big)\, v + b_2\big(f(\mathbf{u})\big)\,,$$

where $\mathbf{w} = (\mathbf{u}, v)$. By the assumed smoothness of the functions b_1 and b_2, the continuity requirements in (2.24) are clearly satisfied. Furthermore, the condition in (2.25a) holds by our assumptions that $E(V \mid \mathbf{U} = \mathbf{u}) = b_3^{-1}(\,\phi(\mathbf{u})\,)$ and that ϕ is bounded. Finally, the expression in (2.25b) is equivalent to requiring that for some positive constants M_5 and M_6,

$$E\big[\,\exp(\,tV\,)\,\big|\,\mathbf{U} = \mathbf{u}\,\big] \;\leq\; M_6 \qquad \text{for all } |\,t\,| \leq M_5 \;\text{ and }\; \mathbf{u} \in \mathcal{U}\,.$$

[CR] Recall that in the censored regression context

$$l(f, \mathbf{w}) \;=\; -v_2\,\big(\,f^2(\mathbf{u}) \;+\; 2\,f(\mathbf{u})\,v_1\,\big)\,/\,2 \;+\; \big(\,1 - v_2\,\big)\,\log\Big(\,1 - \Phi\big(\,f(\,\mathbf{u}\,)\,\big)\,\Big)\,.$$

where $\mathbf{w} = (\mathbf{u}, \mathbf{v}) = (\mathbf{u}, v_1, v_2)$. By the continuity of $\Phi(\,\cdot\,)$, the conditions in (2.24) is clearly satisfied. Furthermore, by our assumption that ϕ is bounded, (2.25a) holds. Finally, since V is assumed to be normally distributed with unit variance, the exponential bound in (2.25b) is also satisfied.

[PR] In the polychotomous regression context,

$$\begin{aligned} l(f, \mathbf{w}) \;=\;& f_1(\mathbf{u})\, I_1(v) \;+\; \cdots \;+\; f_J(\mathbf{u})\, I_J(v) \\ & -\; \log\Big(\,1 \;+\; \exp\big(\,f_1(\mathbf{u})\,\big) \;+\; \cdots \;+\; \exp\big(\,f_J(\mathbf{u})\,\big)\,\Big)\,, \end{aligned}$$

where $\mathbf{w} = (\mathbf{u}, v)$, with v taking values in the finite set $\{0, \ldots, J\}$. In this case, we can take $K = J + 1$ and set $B_k(\mathbf{W}) = I_k(V)$ for $1 \leq k \leq K - 1$, and $B_K = 1$. Using this prescription, we find that the continuity conditions in (2.24) are clearly satisfied. In addition, the requirements in (2.25) are trivially satisfied since the functions $B_k(\mathbf{W})$ are either indicator functions or the constant one.

[D] In this context, we have that

$$l(f, \mathbf{w}) \;=\; l(f, \mathbf{u}) \;=\; f(\mathbf{u}) \;-\; c(f)\,,$$

where $\mathbf{w} = \mathbf{u}$. Taking $K = 1$, we find that the continuity conditions (2.24) have been discussed previously in connection with the function $c(\,\cdot\,)$. In

addition, since we can take B_1 to be the constant 1, the conditions in (2.25) are trivially satisfied.

Assuming that Condition 3″ holds, let $\tau \in (0, 1/2)$ be such that

$$\omega(\underline{h}, \mathcal{S})\, n^{-2\tau} \;=\; O(1) \qquad \text{and} \qquad \omega(\underline{h}, \mathcal{S})\, \log n \;=\; o\big(n^{1-2\tau}\big)\,. \tag{2.26}$$

If we let $\omega^2(\underline{h}, \mathcal{S}) = o(n^{1-\epsilon})$ for some $\epsilon > 0$ as in Condition 3″, observe that by choosing the constant τ such that

$$(1-\epsilon)/4 \;\le\; \tau \;<\; (1+\epsilon)/4\,,$$

then both relations in (2.26) hold. We use this constant in establishing an exponential bound on the difference between $\ell(g)$ and $\Lambda(g)$ for certain functions $g \in \mathbb{G}$. Recall from the definition of these quantities that

$$\big[E_n - E\big]\big[l(f, \mathbf{W})\big] \;=\; \ell(f) - \Lambda(f)\,.$$

With this in mind, we state and prove the following lemma.

Lemma 2.4 *Assume Conditions* 1, 2′ *and* 3″ *hold, that* $l(f, \mathbf{w})$ *is in the form* (2.23)*, and that* (2.24) *and* (2.25) *hold. Then, given positive constants* t_1 *and* T_1*, there is a positive constant* T_2 *such that for all* n *sufficiently large, except on a set having probability at most* $2\exp(-T_2\, n^{1-2\tau})$*, the inequality*

$$\big|E_n - E\big|\big[l(g, \mathbf{W}) - l(\phi_n^*, \mathbf{W})\big] \;\le\; T_1\, n^{-2\tau}$$

holds for all functions $g \in \mathbb{G}$ *satisfying* $\|\sigma(g) - \sigma(\phi_n^*)\| \;\le\; t_1\, n^{-\tau}$.

Proof Throughout this proof, let $T_3, T_4, \ldots$ denote suitable positive constants. Now, for each index $1 \le k \le K$, set

$$b_k^*(g, \mathbf{U}) \;=\; b_k(g, \mathbf{U}) \;-\; b_k(\phi_n^*, \mathbf{U})$$

and

$$B_k^*(\mathbf{W}) \;=\; B_k(\mathbf{W}) \;-\; E\big(B_k(\mathbf{W}) \,|\, \mathbf{U}\big)\,.$$

From (2.25) and Taylor's theorem, we know that if each t_k is bounded in absolute value by M_5, then there exists a positive constant T_3 such that

$$E\left[\exp\left(\sum_{k=1}^{K} t_k\, B_k^*(\mathbf{W})\right) \,\middle|\, \mathbf{U} = \mathbf{u}\right] \;\le\; 1 + T_3 \sum_{k=1}^{K} t_k^2\,,$$

and hence if $|\, t_k\, b_k^*(g, \mathbf{U})\,| < M_5$ for each $1 \le k \le K$, then

$$E\left[\, \exp\left(\, \sum_{k=1}^{K}\, t_k\, b_k^*(g, \mathbf{U})\; B_k^*(\mathbf{W})\, \right) \;\middle|\; \mathbf{U} = \mathbf{u} \right] \;\le\; 1 \;+\; T_3 \sum_{k=1}^{K} \left(\, t_k\, b_k^*(\, g, \mathbf{u}\,)\, \right)^2 .$$

Now, for any sufficiently small scalar $t > 0$, take each $t_k = t$, and set

$$F_1(\, g, \mathbf{w}\,) \;=\; \sum_{k=1}^{K} b_k^*(g, \mathbf{u})\; B_k^*(\mathbf{w})$$

and

$$F_3(\, g, \mathbf{u}\,) \;=\; \left(\, \sum_{k=1}^{K} \left(\, b_k^*(\, g, \mathbf{u}\,)\, \right)^2 \right)^{1\,/\,2},$$

so that the last inequality becomes

$$E\Big[\, \exp\left(\, tF_1(\, g, \mathbf{W}\,)\, \right) \,\Big|\, \mathbf{U} = \mathbf{u} \Big] \;\le\; 1 \;+\; T_3\, t^2\, F_3^2(\, g, \mathbf{u}\,)\,. \tag{2.27}$$

Further, by defining the function

$$F_2(\, g, \mathbf{u}\,) \;=\; \sum_{k=1}^{K} b_k^*(g, \mathbf{u})\, E\big(\, B_k(\mathbf{W}) \,|\, \mathbf{U} = \mathbf{u}\,\big) \;-\; \Lambda(\, g\,) \;+\; \Lambda(\, \phi_n^*\,)\,,$$

we have from (2.27) that the quantitiy

$$E\Big[\, \exp\left(\, tF_1(\, g, \mathbf{W}\,) \;+\; t\, F_2(\, g, \mathbf{U}\,)\, \right) \,\Big|\, \mathbf{U} = \mathbf{u} \Big] \tag{2.28}$$

is not larger than

$$\exp\left(\, t\, F_2(g, \mathbf{u})\, \right) \left(\, 1 \;+\; T_3\, t^2\, F_3^2(\, g, \mathbf{u}\,)\, \right). \tag{2.29}$$

Therefore, given a positive constant T_4 there is a positive constant T_5 such that if

$$t^2 \left(\, F_3^2(\, g, \mathbf{u}\,) + \;F_2^2(\, g, \mathbf{u}\,)\, \right) \;\le\; T_4\,,$$

then the quantity in (2.29) is bounded above by

$$1 \;+\; t\, F_2(\, g, \mathbf{u}\,) \;+\; T_5\, t^2 \left(\, F_3^2(\, g, \mathbf{u}\,) \;+\; F_2^2(\, g, \mathbf{u}\,)\, \right). \tag{2.30}$$

By taking expectations in (2.28) and (2.30), we find that if

$$t^2 \left(\, \left\|\, F_3(\, g, \cdot\,)\, \right\|_\infty^2 \;+\; \left\|\, F_2(\, g, \cdot\,)\, \right\|_\infty^2 \right) \;\le\; T_4 \tag{2.31a}$$

then, since $E\big[\, F_2(\, g, \mathbf{U}\,)\, \big] = 0$, the quantity

$$E\Big[\, \exp\left(\, tF_1(\, g, \mathbf{W}\,) \;+\; t\, F_2(\, g, \mathbf{U}\,)\, \right) \Big] \tag{2.31b}$$

is bounded above by

$$1 \;+\; T_5\, t^2 \left(\left\| F_3(\, g, \cdot\,) \right\|^2 \;+\; \left\| F_2(\, g, \cdot\,) \right\|^2 \right),$$

which is certainly not larger than

$$\exp\left(T_5\, t^2 \left(\left\| F_3(\, g, \cdot\,) \right\|^2 \;+\; \left\| F_2(\, g, \cdot\,) \right\|^2 \right) \right). \tag{2.31c}$$

Observe that by the definition of the functions F_1 and F_2,

$$E_n\big[\, F_1(\, g, \mathbf{W}\,) \;+\; F_2(\, g, \mathbf{U}\,)\,\big] \;\;=\;\; \big[E_n - E\big]\big[\, l(g, \mathbf{W}) - l(\phi_n^*, \mathbf{W})\,\big]\,,$$

and hence from (2.31), if

$$t^2 \left(\left\| F_3(\, g, \cdot\,) \right\|_\infty^2 \;+\; \left\| F_2(\, g, \cdot\,) \right\|_\infty^2 \right) \;\;\leq\;\; T_4\, n^2\,, \tag{2.32a}$$

then the quantity

$$E\left[\exp\left(t\, \big[E_n - E\big]\big[\, l(g, \mathbf{W}) - l(\phi_n^*, \mathbf{W})\,\big] \right) \right] \tag{2.32b}$$

is bounded above by

$$\exp\left(T_5\, n^{-1}\, t^2 \left(\left\| F_3(\, g, \cdot\,) \right\|^2 \;+\; \left\| F_2(\, g, \cdot\,) \right\|^2 \right) \right). \tag{2.32c}$$

Now, let g be any function in $\mathbb{G}$ satisfying

$$\left\| \sigma(\, g\,) \;-\; \sigma(\, \phi_n^*\,) \right\| \;\;\leq\;\; T_1\, n^{-\tau}\,.$$

Observe that by Lemma 1 in the Appendix,

$$\left\| \sigma(\, g\,) \;-\; \sigma(\, \phi_n^*\,) \right\|_\infty^2 \;\;\leq\;\; T_1\, \omega(\, \underline{h}, \mathcal{S}\,)\, n^{-2\tau}\,,$$

and hence by Condition $3''$, $\sigma(\, g\,) - \sigma(\, \phi_n^*\,)$ is bounded in absolute value. Next, since $g - \phi_n^* \in \mathbb{G}$ and

$$\sigma(\, g\,) \;-\; \sigma(\, \phi_n^*\,) \;\;=\;\; g - \phi_n^* \;-\; c(\, g\,) \;+\; c(\, \phi_n^*\,)$$

we have that

$$\left| c(\, g\,) - c(\, \phi_n^*\,) \right| \;\;=\;\; \left| \left\langle\, g - \phi_n^* - c(\, g\,) + c(\, \phi_n^*\,), 1 \right\rangle_n \right| \;\;\leq\;\; \left\| \sigma(\, g\,) \;-\; \sigma(\, \phi_n^*\,) \right\|_\infty.$$

Therefore,

$$\begin{aligned} \left\| g \;-\; \phi_n^* \right\|_\infty \;\;&\leq\;\; \left\| \sigma(\, g\,) \;-\; \sigma(\, \phi_n^*\,) \right\|_\infty \;+\; \left| c(\, g\,) - c(\, \phi_n^*\,) \right| \\ &\leq\;\; 2 \left\| \sigma(\, g\,) \;-\; \sigma(\, \phi_n^*\,) \right\|_\infty \end{aligned}$$

so that

$$\left\| g - \phi_n^* \right\|_\infty^2 \leq 2\, T_1\, \omega(\underline{h}, \mathcal{S})\, n^{-2\tau}.$$

Furthermore, by the continuity condition (2.24), there exists positive constants T_7 and T_8 such that

$$\left(\left\| F_3(g, \cdot) \right\|_\infty^2 + \left\| F_2(g, \cdot) \right\|_\infty^2 \right) \leq T_7\, \omega(\underline{h}, \mathcal{S})\, n^{-2\tau}$$

and

$$\left(\left\| F_3(g, \cdot) \right\|^2 + \left\| F_2(g, \cdot) \right\|^2 \right) \leq T_8\, n^{-2\tau}.$$

Finally, if for some positive constant T_9,

$$|t| \leq T_9\, n^{1+\tau} / \sqrt{\omega(\underline{h}, \mathcal{S})},$$

and hence if there is another positive constant T_{10} such that $|t| \leq T_{10}\, n$, then by (2.32)

$$E\left[\exp\left(t\, [E_n - E]\, [\, l(g, \mathbf{W}) - l(\phi_n^*, \mathbf{W})\,] \right) \right] \leq \exp\left(T_{11}\, t^2\, n^{-1-2\tau} \right)$$

for some suitable positive constant T_{11}. If we take t such that $t = \pm T_{12}\, n$ for some positive constant T_{12} satisfying

$$0 < T_{12} \leq \min\left\{ T_{10}, T_1 / (2\, T_{11}) \right\},$$

then from Markov's inequality we find that

$$\left| E_n - E \right| \left[l(g, \mathbf{W}) - l(\phi^*, \mathbf{W}) \right] \leq T_1\, n^{-2\tau}$$

except on a set with probability at most $2 \exp(-T_2\, n^{1-2\tau})$, where we have chosen T_2 so that $T_2 = T_1\, T_{12} / 2$. □

Lemma 2.5 *Assume that Conditions* 1, $2'$ *and* $3''$ *hold, that* $l(f, \mathbf{w})$ *is of the form* (2.23), *and that* (2.24) *and* (2.25) *hold. Then, given positive constants* T_1 *and* T_2, *there is a positive constant* t_2 *such that, except on a set with probability tending to zero with* n, *the inequality*

$$\left| E_n - E \right| \left[l(g_1, \mathbf{W}) - l(g_2, \mathbf{W}) \right] \leq T_1\, n^{-2\tau}$$

holds for all functions $g_1, g_2 \in \mathbb{G}$ *satisfying*

$$\| \sigma(g_1) \|_\infty \leq T_2, \quad \| \sigma(g_2) \|_\infty \leq T_2 \quad \textit{and} \quad \| \sigma(g_1) - \sigma(g_2) \|_\infty \leq t_2\, n^{-2\tau},$$

where τ *is defined as in* (2.26).

Proof Throughout this proof, let $T_3, T_4, \ldots$ denote suitable constants. For notational convenience, assume initially that $\sigma(\,\cdot\,)$ is the identity operator. Let $g_1, g_2 \in \mathbb{G}$ satisfy the conditions of the lemma. Then, as was done in the proof of the previous lemma, introduce the quantities

$$b_k^*(g_1, g_2, \mathbf{U}) \;=\; b_k(g_1, \mathbf{U}) \;-\; b_k(g_2, \mathbf{U})$$

and

$$B_k^*(\mathbf{W}) \;=\; B_k(\mathbf{W}) \;-\; E\big(\, B_k(\mathbf{W}) \,|\, \mathbf{U} \,\big)\,,$$

and define the function

$$F_1(\, g_1, g_2, \mathbf{w}\,) \;=\; \sum_{k=1}^{K} b_k^*(g_1, g_2, \mathbf{u})\; B_k^*(\mathbf{w})\,.$$

Now, by the continuity condition (2.24), there exists a postive constant T_3 that does not depend on our choice of g_1, g_2 such that

$$\big\|\, b_k^*(g_1, g_2, \,\cdot\,) \,\big\|_\infty \;\leq\; T_3\, n^{-2\tau}\,.$$

Further, the conditional mean $E(\, B_k(\mathbf{W}) \,|\, \mathbf{U} = \mathbf{u}\,)$ is bounded, and the variables $E_n|\, B_k^*(\mathbf{W})\,|$ are bounded in probability. Therefore, there exists a positive constant T_4 such that except on a set whose probability tends to zero with n,

$$E_n\big|F_1(g_1, g_2, \mathbf{W}\,)\,\big| \;\leq\; T_4\, n^{-2\tau}\,. \tag{2.33a}$$

Next, define the function

$$F_2(\, g_1, g_2, \mathbf{u}\,) \;=\; \sum_{k=1}^{K} b_k^*(g_1, g_2, \mathbf{u})\, E\big(\, B_k(\mathbf{W}) \,|\, \mathbf{U} = \mathbf{u}\,\big) \;-\; \Lambda(\, g_1\,) \;+\; \Lambda(\, g_2\,)\,,$$

and observe that by (2.24) and (2.25), there exists a constant T_5 such that except on a set whose probability tends to zero with n

$$E_n\big|F_2(g_1, g_2, \mathbf{U}\,)\,\big| \;\leq\; T_5\, n^{-2\tau}\,. \tag{2.33b}$$

Finally, since

$$\big|\, E_n \;-\; E\,\big|\big[\, l(g_1, \mathbf{W}) \;-\; l(g_2, \mathbf{W})\,\big] \;=\; E_n\big|\, F_1(\, g_1, g_2, \mathbf{W}\,) \;+\; F_2(\, g_1, g_2, \mathbf{U}\,)\,\big|\,,$$

the desired conclusion follows from (2.33). Combining the observations after (2.32c) with the same argument above, substituting $\sigma(\, g_1\,), \sigma(\, g_2\,)$ for g_1, g_2, and noting that

$$l\big(\, \sigma(\, g\,), \mathbf{w}\,\big) \;=\; l(\, g, \mathbf{w}\,)\,,$$

for all $g \in \mathbb{G}$, we obtain the desired conclusion for the case when $\sigma(\cdot)$ is not the identity operator. □

Given a set of functions $\mathcal{G}$ of $\mathbb{G}$, define the diameter of $\mathcal{G}$ by

$$\operatorname{diam}\mathcal{G} \;=\; \sup\left\{ \| g_1 - g_2 \|_\infty \,:\, g_1, g_2 \in \mathcal{G} \right\}.$$

We are now in a position to state and prove the following lemma.

Lemma 2.6 *Assume that Condition 1 holds. Then, given $s \in \mathcal{S}$ and the constants $T_1 > T_2 > 0$, there exists a constant T_3 depending only on M_1, m, $\dim \mathcal{U}_s$ and the dimension of ϕ, such that*

$$\mathcal{G} \;=\; \left\{ g \in \mathbb{G}_s \,:\, \| g \|_\infty \le T_1 \right\},$$

can be covered by

$$\left(T_3 \left(T_1 / T_2 \right) + 2 \right)^{T_3 \underline{h}_s^{-d_s}}$$

sets having diameter at most T_2.

Proof Let $T_4, T_5, \ldots$ denote suitable positive constants. Consider initially a very simple example. For a fixed positive integer d, let $\delta_0 \subset \delta_1 \subset \mathbb{R}^d$ be two balls each centered at the origin. In order to apply Condition 1 easily, we let δ_1 have unit diameter and δ_0 have volume $1/(2M_1)$. For a positive integer m, let $x_1, \cdots, x_L$ be a collection of points in δ_0, and assume that the space of polynomials of total degree m on $\mathbb{R}^d$ is identifiable with respect to this collection. For example, we might take $L = (m+1)^d$, and choose $x_1, \ldots, x_L$ to be arranged uniformly inside a d-dimensional cube that is contained in δ_0. Then, there exists a positive constant T_4 depending only on d, m, M_1, and our choice of points such that

$$\frac{1}{T_4} \| p \|_\infty \;\le\; \sup_j \left| p(x_j) \right| \tag{2.34}$$

for all polynomials p of degree m (here, we temporarily let $\| \cdot \|_\infty$ denote the sup-norm on δ_1). Indeed, arguing as in the proof of Lemma 1.1 in Chapter 1, we observe that it is sufficient to consider only those functions p having unit sup-norm, and assume that there does not exist one constant T_4 satisfying (2.34). Then, we can find a sequence of polynomials p_i such that $\| p_i \|_\infty = 1$ and

$$\sup_j \left| p_i(x_j) \right| \;\to\; 0 \qquad \text{as} \qquad i \;\to\; \infty\,. \tag{2.35}$$

Associated with each polynomial p_i is a vector of coefficients having sup-norm bounded above by a fixed constant determined in (1.11) of Chapter 1 depending only on m and d. Therefore, we can find a subsequence of these coefficient vectors

that converge to a vector that we in turn use to construct a polynomial p^*. Observe that while the polynomial p^* has unit sup-norm on δ_0, the relation in (2.35) implies that $\sup_j |\, p^*(x_j)\,|$ must equal zero. This contradicts our assumption that the space of polynomials of total degree m is identifiable with respect to the points $x_1, \ldots, x_L$.

Suppose next that for positive integers $d_1, \ldots, d_M$,

$$\delta_0 = \delta_{01} \times \cdots \times \delta_{0M} \qquad \text{and} \qquad \delta_1 = \delta_{11} \times \cdots \times \delta_{1M}\,,$$

where the sets $\delta_{0l} \subset \delta_{1l} \subset \mathbb{R}^{d_l}$. As in the previous paragraph, we assume that the sets δ_{0l} have volume $1\,/\,(2M_1)$, while the sets δ_{1l} each have unit diameter. Next, choose $m \in \mathbb{N}^M$, and for $1 \le l \le M$, let $x_{l1}, \ldots, x_{lL_l}$ denote a set of points in δ_{0l} with respect to which the space of polynomials of total degree m_l over $\mathbb{R}^{d_l}$ is identifiable. Let $\mathbf{x}_1, \ldots, \mathbf{x}_L$ denote the Cartesian product of these collections. Then, by repeated applications of (2.34) and redefining T_4 if necessary, we find that in fact (2.34) holds for all polynomials of coordinate degree m where now we take $\|\cdot\|_\infty$ to be the sup-norm on the product set δ_1. More precisely, there exists a postitive constant T_4 such that for all polynomials p of coordinate degree m,

$$\frac{1}{T_4}\,\|\,p\,\|_\infty \;\le\; \sup_l \big|\,p(\mathbf{x}_l)\,\big|\,, \tag{2.36}$$

where $T_4 \ge 1$ depends only on d, m, M_1, and our choice of points in each set δ_{0l}.

Using the definitions in the previous paragraph, we now consider covering the collection of polynomials of coordinate degree at most m satisfying

$$\mathcal{A} \;=\; \Big\{\,\|\,p\,\|_\infty \;\le\; T_1\,\Big\}\,,$$

by sets with diameter at most T_2, where T_1 and T_2 are given in the statment of the lemma. Introduce the constant $T_5 = T_2\,/\,T_4$, choose $T_6 \in \mathbb{N}$ such that

$$T_6 \;\ge\; T_1\,/\,T_5 \,-\, 1\,/\,2\,,$$

and for each $1 \le l \le L$, let $\mathcal{A}_{li}$ denote the set of polynomials p of coordinate degree m such that

$$\big|\,p(\mathbf{x}_l) \,-\, i\,T_5\,\big| \;\le\; T_5\,/\,2\,,$$

where i ranges from $-T_6$ to T_6 and $\mathbf{x}_l$ is one point of the collection $\mathbf{x}_1, \ldots, \mathbf{x}_L$ defined in the previous paragraph. Then, for each point $\mathbf{x}_l$, any polynomial in $\mathcal{A}$ must fall into one of the sets $\mathcal{A}_{li}$. Conversely, given the integers $i_1, \ldots, i_L$, where each i_l is between $-T_6$ and T_6, the set

$$\mathcal{A}_{1i_1} \cap \cdots \cap \mathcal{A}_{Li_L} \tag{2.37}$$

specifies a collection of polynomials whose sup-norm is not larger than

$$T_4\,T_5\,\left(\,T_6\,+\,1\,/\,2\,\right)\;\;\geq\;\;T_4\,T_5\,\left(\,T_1\,/\,T_5\,\right)\;\;=\;\;T_4\,T_1\;\;\geq\;\;T_1\,.$$

Here we have used the expression in (2.36) to obtain the leftmost quantity above. Therefore, by considering all possible sets of the form (2.37) we obtain a cover of the set $\mathcal{A}$ having cardinality at most $(\,2T_6\,+\,1\,)^L$. In addition, suppose that for some choice of the integers $i_1,\ldots,i_M$, the polynomials $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ are both in the set (2.37). Then, by definition, we have that for $1\leq l\leq L$,

$$\left|\,p_1(\mathbf{x}_l)\,-\,p_2(\mathbf{x}_l)\,\right|\;\;\leq\;\;\left|\,p_1(\mathbf{x}_l)\,-\,i_l\,T_5\right|\,+\,\left|\,p_2(\mathbf{x}_l)\,-\,i_l\,T_5\right|\;\;\leq\;\;T_5\,,$$

and hence by (2.36),

$$\left\|\,p_1-p_2\,\right\|_\infty\;\;\leq\;\;T_4\,T_5\;\;=\;\;T_2\,.$$

Therefore, each of the sets in (2.37) have diameter at most T_2.

Returning to our original problem, suppose that ϕ is real-valued and that $\mathbb{G}_s=\mathbb{PP}_s$, the space of piecewise polynomials of coordinate degree m on $\mathcal{U}_s$ relative to the triangulation $\triangle_s$, where $s=\{s_1,\ldots,s_k\}$. By Condition 1 and the arguments given above, we find that there exists a positive constant T_7 such that for any choice of $\delta\in\triangle_s$, the space of polynomials in $\mathbb{PP}_s$ that are in $\mathcal{G}$ can be covered by $(\,2\,T_7\,+\,1\,)^{L_s}$ sets having diameter at most T_2, where $L_s=L_{s_1}+\cdots+L_{s_k}$. Then, repeating this construction for each $\delta\in\triangle_s$, we find that the set $\mathcal{G}$ can be covered by

$$\left(\,2\,T_6\,+\,1\,\right)^{T_8\,L_s\,\underline{h}_s^{-d_s}}$$

sets having diameter at most T_2. Here, we defined the constant $T_8=M_1^{\#(\,s\,)}$ and used the fact that there are at most $T_8\,/\,\underline{h}_s^{d_s}$ sets $\delta\in\triangle_s$ by Condition 1. Therefore, the desired result is valid for $\mathbb{G}_s=\mathbb{PP}_s$. Note however, for any proper subspace $\mathbb{G}_s$, the result now automatically holds since the set $\mathcal{G}$ corresponding to $\mathbb{G}$ is a subset of the space we have just examined.

Finally, if ϕ is a vector-valued function, the space $\mathbb{G}_s$ appearing in the definition of $\mathbb{G}$ is the vector-valued version defined in the previous section; that is, the functions $g\in\mathbb{G}$ are of the form $g=(g_1,\ldots,g_J)$, where each g_j is in the space of real-valued functions defined in (1.9b). By the definition of the sup-norm given in (1.1a) we find that a vector-valued function satisfies $\|\,g\,\|_\infty<T_1$ if and only if $\|\,g_j\,\|_\infty<T_1$ for $1\leq j\leq J$. Hence, repeating the arguments above for each function g_j, $1\leq j\leq J$, we find that the set $\mathcal{G}$ can be covered by

$$\left(\,2\,T_6\,+\,1\,\right)^{J\,T_8\,L_s\,\underline{h}_s^{-d_s}}$$

sets having diameter at most T_2. □

Lemma 2.7 *Suppose that Conditions 1, 2′ and 3″ holds. Then, given positive constants t_1 and t_2, there is a positive constant T_1 such that the set*

$$\mathcal{G} \;=\; \Big\{ g \in \mathbb{G} \,:\, \big\| \sigma(g) - \sigma(\phi_n^*) \big\| \;\le\; t_1 n^{-\tau} \Big\} \tag{2.37}$$

can be covered by

$$O\Big(\exp\big(T_1 \, \omega(\underline{h}, \mathcal{S}) \, \log n \big) \Big)$$

subsets each having diameter at most $t_2 n^{-2\tau}$.

Proof Throughout this proof, we let $T_2, T_3, \ldots$ denote suitable constants. For the moment, assume that ϕ is real-valued and that $\mathbb{G} = \mathbb{PP}$. Let p denote any function in $\mathbb{PP}$ such that

$$\big\| \sigma(p) - \sigma(\phi_n^*) \big\| \;\le\; t_1 n^{-\tau} .$$

Given two square-integrable functions f_1, f_2 on $\mathcal{U}$, let $\langle \cdot, \cdot \rangle$ temporarily denote the inner product defined by

$$\langle f_1, f_2 \rangle \;=\; \int_{\mathcal{U}} f_1(\mathbf{u}) \, f_2(\mathbf{u}) \, d\mathbf{u} .$$

For each $s \in \mathcal{S}_1$, let $\mathbb{PP}_s^1$ denote the space of functions in $\mathbb{PP}_s$ that are orthogonal to each of the spaces $\mathbb{PP}_r$, for subsets $r \subset s$. Then, $\mathbb{PP}_s$ is the direct sum of the spaces $\mathbb{PP}_r^1$ for $r \subset s$, and we can uniquely express

$$\sigma(p) - \sigma(\phi_n^*) \;=\; \sum_{s \in \mathcal{S}_1} p_s \qquad \text{where } p_s \in \mathbb{PP}_s^1 \text{ for } s \in \mathcal{S}_1 .$$

Now, given $r \in \mathcal{S}$, choose $\mathbf{u}_r' \in \mathcal{U}_r$ and define the representor of the point-evaluation functional $q_{\mathbf{u}_r'} \in \mathbb{PP}_r$ as in (2.1) of Chapter 1, but this time with respect to the inner-product defined above. Then, arguing as in the proof of Lemma 2.1 in that chapter, we find that there exists a postive constant T_2 independent of $\underline{h}_r$ and our choice of $\mathbf{u}_r'$ such that

$$q_{\mathbf{u}_r'}(\mathbf{u}_r') \;=\; \big\| q_{\mathbf{u}_r'} \big\|^2 \;\le\; T_2 \, \underline{h}_r^{-d_r} , \qquad \mathbf{u}_r' \in \mathcal{U}_r . \tag{2.38}$$

Observe that by the inclusion-exclusion formula for orthogonal projections, the component p_s corresponding to some set $s \in \mathcal{S}_1$ is obtained as a linear combination having weights ± 1 of projections into the spaces $\mathbb{PP}_r$ for $r \subset s$, $r \neq s$. Select one such r and let $\tilde{p}_{sr}$ denote the orthogonal projection of $\sigma(p) - \sigma(\phi^*)$ onto $\mathbb{PP}_r$. Then from (2.38), for any point $\mathbf{u}_r \in \mathcal{U}_r$,

$$\tilde{p}_{sr}(\mathbf{u}_r) \;=\; \big\langle q_{\mathbf{u}_r'}, \tilde{p}_{sr}(\mathbf{u}_r) \big\rangle \;=\; \big\langle q_{\mathbf{u}_r'}, \sigma(p) - \sigma(\phi^*) \big\rangle ,$$

and hence from the Schwartz inequality,

$$\tilde{p}_{sr}^2(\mathbf{u}_r) \;\leq\; T_2^2\, \underline{h}_r^{-d_r}\, \left\| \sigma(p) - \sigma(\phi^*) \right\|^2 .$$

Therefore,

$$\left\| \tilde{p}_{sr} \right\|_\infty^2 \;\leq\; T_2^2\, t_1\, \underline{h}_r^{-d_r}\, n^{-2\tau} ,$$

so by the inclusion-exclusion formula for orthogonal projections,

$$\left\| p_s \right\|_\infty^2 \;\leq\; T_2^2\, t_1\, \omega(\underline{h}, \mathcal{S})\, n^{-2\tau} .$$

Consequently, from Lemma 2.6, there exists a positive constant T_3 such that the set $\mathcal{G}$ given in (2.37) can be covered by

$$\left(T_3\, \omega(\underline{h}, \mathcal{S})\, n^{2\tau} + 2 \right)^{T_3\, \omega(\underline{h}, \mathcal{S})}$$

subsets each having diameter at most $t_2\, n^{-2\tau}$. Since

$$\log\left(T_3\, \omega(\underline{h}, \mathcal{S})\, n^{2\tau} + 2 \right) \;=\; O\left(\log n \right),$$

the desired result is valid for $\mathbb{G} = \mathbb{G}_1 = \mathbb{PP}$. As argued at the end of the proof of Lemma 2.6, for any proper subspace $\mathbb{G}$, the result now automatically holds since the set in (2.37) corresponding to $\mathbb{G}$ is a subset of (2.37) when $\mathbb{G} = \mathbb{G}_1 = \mathbb{PP}$. Finally, if ϕ is vector-valued, then $\mathcal{G}$ is made up of functions of the form $g = (g_1, \ldots, g_J)$, and the conclusion of the lemma is obtained by repeating the above arguments for the functions g_j, $1 \leq j \leq J$, separately. □

Lemma 2.8 *Assume that Conditions* 1, 2′, 3″, *and* 5′ *hold. Then, given a positive constant* t_1, *except on an event whose probability tends to zero with* n,

$$\ell(g) \;<\; \ell(\phi_n^*),$$

for all $g \in \mathbb{G}$ *satisfying*

$$\left\| \sigma(g) - \sigma(\phi_n^*) \right\| \;=\; t_1\, n^{-\tau} .$$

Proof Throughout this proof, let $T_1, T_2, \ldots$ denote suitable constants. Given t_1, define the set

$$\mathcal{G} \;=\; \left\{ \sigma(g) : g \in G \text{ and } \left\| \sigma(g) - \sigma(\phi_n^*) \right\| \;\leq\; t_1\, n^{-\tau} \right\}.$$

For any $g \in \mathbb{G}$ we find from Lemma 1 in the Appendix that there exists a positive constant T_1 such that

$$\left\| \sigma(g) - \sigma(\phi_n^*) \right\|_\infty \;\leq\; T_1 \sqrt{\omega(\underline{h}, \mathcal{S})}\, t_1\, n^{-\tau}$$

and hence by (2.26), Condition 3'' and Lemma 2.2, for all n sufficiently large, g is bounded above in absolute value by some positive constant T_2. Furthermore, by repeating the argument in Lemma 2.1 replacing $\mathbb{G}$ for $\mathbb{F}$ and ϕ_n^* for ϕ^*, we know that there exists a positive constant T_3 such that

$$\Lambda(g) - \Lambda(\phi_n^*) \leq -\frac{1}{T_3} \left\| \sigma(g) - \sigma(\phi_n^*) \right\|^2 . \tag{2.39}$$

Now, fix $T_4 > 0$ and observe from Lemma 2.5 that given T_4 and T_2, there exists a positive constant t_2 such that except on a set whose probability tends to zero with n,

$$\left| E_n - E \right| \left[l(g_1, \mathbf{W}) - l(g_2, \mathbf{W}) \right] \leq T_4 \, n^{-2\tau} \tag{2.40}$$

for all functions g_1, g_2 satisfying

$$\left\| \sigma(g_1) \right\|_\infty \leq T_2 , \quad \left\| \sigma(g_2) \right\|_\infty \leq T_2 \quad \text{and} \quad \left\| \sigma(g_1) - \sigma(g_2) \right\|_\infty \leq t_2 \, n^{-2\tau} .$$

Given t_1 and t_2 we know from Lemma 2.7 that there exists a positive constant T_5 such that the set $\mathcal{G}$ can be covered by

$$O\Big(\exp\big(T_5 \, \omega(\underline{h}, \mathcal{S}) \, \log n \big) \Big) \tag{2.41}$$

subsets having diameter at most $t_2 \, n^{-2\tau}$. Let $\mathcal{G}_1, \ldots, \mathcal{G}_L$ denote one such covering of $\mathcal{G}$, and with each of these sets, associate an element $g_l \in \mathcal{G}_l$. Now, fix $T_6 > 0$ and observe from Lemma 2.4 that given t_1 and T_6, there exists a positive constant T_7 such that the following inequality

$$\left| E_n - E \right| \left[l(g_l, \mathbf{W}) - l(\phi_n^*, \mathbf{W}) \right] \leq T_6 \, n^{-2\tau} \tag{2.42}$$

holds simultaneously for all of the functions g_l, $1 \leq l \leq L$, except on a set having probability at most

$$T_8 \exp\big(-T_7 \, n^{1-2\tau} + T_5 \, \omega(\underline{h}, \mathcal{S}) \, \log n \big) , \tag{2.43}$$

where T_8 is some positive constant. However, by (2.26), we know that the quantity in (2.43) tends to zero with n for any choice of t_1 and T_6.

Now, choose $g \in \mathcal{G}$ such that

$$\left\| \sigma(g) - \sigma(\phi_n^*) \right\| = t_1 \, n^{-\tau} .$$

Since $g \in \mathcal{G}$, it must be contained in at least one of the sets $\mathcal{G}_1, \ldots, \mathcal{G}_L$. Suppose that $g \in \mathcal{G}_l$, and observe that from (2.40)

$$\left| E_n - E \right| \left[l(g, \mathbf{W}) - l(g_l, \mathbf{W}) \right] \leq T_4 \, n^{-2\tau} . \tag{2.44}$$

However, combining (2.42) and (2.44), we find that except on a set whose probability tends to zero with n,

$$\begin{aligned} E_n\big[\, l(g, \mathbf{W}) \,-\, l(\phi_n^*, \mathbf{W}) \,\big] &\leq \big(\, T_4 \,+\, T_6 \,\big)\, n^{-2\tau} + E\big[\, l(g, \mathbf{W}) \,-\, l(\phi_n^*, \mathbf{W}) \,\big] \\ &= \big(\, T_4 \,+\, T_6 \,\big)\, n^{-2\tau} + \Lambda(\, g \,) \,-\, \Lambda(\, \phi_n^* \,) \end{aligned}$$

and hence from (2.39),

$$\begin{aligned} E_n\big[\, l(g, \mathbf{W}) \,-\, l(\phi_n^*, \mathbf{W}) \,\big] &\leq \big(\, T_4 \,+\, T_6 \,\big)\, n^{-2\tau} - \frac{1}{T_3} \big\|\, \sigma(\, g \,) \,-\, \sigma(\, \phi_n^* \,) \,\big\|_\infty \\ &\leq \big(\, T_4 \,+\, T_6 \,\big)\, n^{-2\tau} - \frac{t_1}{T_3}\, n^{-2\tau} \,. \end{aligned}$$

Therefore, since we are free to choose T_4 and T_6 as small as possible, the desired inequality holds. □

Lemma 2.9 *Assume that Conditions* 1, 2′, 3″ *and* 5′ *hold. Then, the maximum likelihood estimate* $\sigma(\,\hat{\phi}\,)$ *exists and is unique except on an event whose probability tends to zero with* n*. In addition,*

$$\big\|\, \sigma(\,\hat{\phi}\,) \,-\, \sigma(\,\phi_n^*\,) \,\big\|_\infty \;=\; o_P(\,1\,)\,. \tag{2.45}$$

Proof Using Lemma 2.8 and arguing as in the proof of Lemma 2.2, we find from the strict concavity of $\ell(\,\cdot\,)$ demonstrated in Example 2 that $\hat{\phi}$ exists and is unique except on a set whose probablility tends to zero with n and that

$$\big\|\, \sigma(\,\hat{\phi}\,) \,-\, \sigma(\,\phi_n^*\,) \,\big\| \;=\; o_P\big(\, n^{-2\tau} \,\big)\,.$$

The relation (2.45) follows by another application of Lemma 1 in the Appendix and equation (2.26). □

Rate of Convergence

We close this chapter by examining the rate of convergence obtainable by our sample-based estimate $\hat{\phi}$ in the context of extended linear models. In order to obtain this rate, however, we must impose regularity conditions on the underlying (real-valued) spline spaces $\mathbb{G}_l$, $1 \leq l \leq M$. In Chapter 1, we found that the relationship between the size of the coefficients of a piecewise polynomial function and its L_2-norm was extremely important in deriving the rate of convergence of our regression estimates. In particular, Lemma 1.2 of that chapter was pivotal in establishing not only the identifiability of spaces of piecewise polynomials, but

also in determining the stochastic order of various projections into these spaces. In that context, to obtain the rate of convergence for estimates built from the spline spaces $\mathbb{G}_l$, $1 \le l \le M$, it was sufficient to know that each $\mathbb{G}_l$ was contained in a space of piecewise polynomial functions, the elements of which satisfied the relationship in Lemma 1.2. As we will see, however, the nonlinearity inherent in maximum likelihood estimation forces us to consider the structure of these spline spaces more carefully.

Toward this end, for each index $1 \le l \le M$, let $\triangle_l$ denote a partition of $\mathcal{U}_l$ satisfying Condition 1, and define the quantities $\overline{h}$ and $\underline{h}$ as in Chapter 1. For the remainder of this chapter, we assume that $\overline{h} \sim \underline{h}$. Recall that this condition was automatically satisfied when we carried out the rate computations for Theorems 2.1 and 4.1 of Chapter 1. In addition, let $g_{l1}, \ldots, g_{lN_l}$ denote a basis of the spline space $\mathbb{G}_l$, $1 \le l \le M$. Given a function g_l defined on $\mathcal{U}_l$, we let $\operatorname{supp} g_l$ denote the support of g_l relative to $\triangle_l$; that is, the collection of sets $\delta \in \triangle_l$ such that g_l is not identically zero on the interior of δ.

Condition 6 *Assume that there exists positive constants M_7, M_8, and M_9 that are independent of $\overline{h}$, $\underline{h}$ and n such that, for $1 \le l \le M$,*

(a) *the functions $g_{l1}, \ldots, g_{lN_l}$ are each bounded in absolute value by one;*

(b) *the diameters of the supports of $g_{l1}, \ldots, g_{lN_l}$ are bounded above by $M_7 \overline{h}_l$;*

(c) *the number of functions $g_{l1}, \ldots, g_{lN_l}$ that are nonzero on any set $\delta \in \triangle_l$ is bounded above by M_8;*

(d) *if $g_l = \sum_k \beta_{lk} g_{lk}$ then*

$$\left| \beta_{lk} \right| \;\le\; M_9 \sup \left\{ \left| g_l(u_l) \right| : u_l \in \operatorname{supp} g_{lk} \right\} \qquad \text{for } 1 \le k \le N_l \,.$$

A comment is in order as to the reasonableness of Condition 6. As it turns out, these assumptions along with Condition 1 from Chapter 1 are familiar in the finite element literature. Douglas, Dupont and Wahbin (1975) imposed virtually identical conditions when considering the L_∞ stability of orthogonal projections onto finite element spaces, and observed that many popular elements satisfy these properties. In particular, they mention the Lagrange elements from Chapter 2, Hermite elements over simplexes and parallelopipeds and the Clough-Tocher elements, early precursors to the generalized vertex splines. Moreover, it has been shown that univariate B-splines satisfy this property, see de Boor (1978). In addition, the regularity of the box spline basis implies that these properties are satisfied for this basis as well.

For the moment, consider the case when ϕ is real-valued. Given a set $s \in \mathcal{S}$, recall that the space of real-valued functions $\mathbb{G}_s$ is defined to be the tensor product of the spaces $\mathbb{G}_l$ for $l \in s$. Therefore, since the functions $g_{l1}, \ldots, g_{lN_l}$ form a basis for $\mathbb{G}_l$, $1 \leq l \leq M$, the space $\mathbb{G}_s$ is spanned by functions of the form

$$\prod_{l \in s} g_{lk_l}(u_l) \qquad \text{where } 1 \leq k_l \leq N_l \text{ for } l \in s\,. \tag{2.46}$$

Let N_s denote the dimension $\mathbb{G}_s$, which is simply the product of N_l for $l \in s$, and given some ordering of the functions in (2.46), let $g_{s1}, \ldots, g_{sN_s}$ be a basis of $\mathbb{G}_s$. Then, any function $g_s \in \mathbb{G}_s$ can be expressed uniquely as

$$g_s \;=\; \sum_{k=1}^{N_s} \beta_{sk}\, g_{sk}\,. \tag{2.47a}$$

Let $\beta_s = (\beta_{s1}, \ldots, \beta_{sN_s})$ represent the N_s-dimensional vector of coefficients in this expansion. If, on the other hand, ϕ is vector-valued, we recall that the space $\mathbb{G}_s$ is comprised of functions $g_s = (g_{1s}, \ldots, g_{Js})$, where each g_{js} is of the form (2.47); that is,

$$g_{js} \;=\; \sum_{k=1}^{N_s} \beta_{jsk}\, g_{sk}\,, \qquad \text{for } 1 \leq j \leq J\,. \tag{2.47b}$$

In this case, we let $\beta_s = (\beta_{1s}, \ldots, \beta_{Js})$ denote the (JN_s)-dimensional vector of coefficients such that for $1 \leq j \leq J$, the N_s-dimensional vector β_{js} represents the coefficients in the expansion in (2.47b). From Conditions 1 and 6 we have the following lemma.

Lemma 2.10 *Suppose that Conditions* 1 *and* 6 *hold and let* $s \in \mathcal{S}$. *Then, there exists a positive constant* M_8 *such that*

$$\left|\, \beta_s \,\right|^2 \;=\; \sum_{k=1}^{N_s} \beta_{sk}^2 \;\leq\; M_8\, \underline{h}_s^{-d_s} \left\|\, g_s \,\right\|^2,$$

for all functions g_s *given by* (2.47).

Proof Throughout this proof, let $T_2, T_3 \ldots$ denote suitable positive constants. To begin with, assume that ϕ is a real-valued function. Fix $l \in s$ and consider functions of the form

$$g_l \;=\; \sum_{k=1}^{N_l} \beta_{lk}\, g_{lk}\,, \tag{2.48}$$

where we recall that $g_{l1}, \ldots, g_{lN_l}$ form a basis of $\mathbb{G}_l$. Let $\delta_{lk} = \operatorname{supp} g_{lk}$ so that from Condition 6(d) there exists a positive constant M_9 such that,

$$\left|\, \beta_{lk} \,\right| \;\leq\; M_9 \left\|\, g_l \,\right\|_{L_\infty(\delta_{lk})} \qquad \text{for } 1 \leq k \leq N_l\,.$$

Using the relationship between the sup-norm and $\| \cdot \|$ derived in the proof of Lemma 1 in the Appendix, we find that there exists a positive constant T_2 such that

$$\beta_{lk}^2 \;\leq\; M_9^2\, T_2\, \underline{h}_l^{-d_l} \left\| g_l \right\|_{L_2(\delta_{lk})}^2 \qquad \text{for } 1 \leq k \leq N_l \,.$$

Summing over the N_s values of k, we find that

$$\begin{aligned} \sum_{k=1}^{N_s} \beta_{lk}^2 &\leq M_9^2\, T_2\, \underline{h}_l^{-d_l} \sum_{k=1}^{N_s} \left\| g_l \right\|_{L_2(\delta_{lk})}^2 \\ &\leq M_8\, M_9^2\, T_2\, \underline{h}_l^{-d_l} \sum_{\delta \in \Delta_l} \left\| g_l \right\|_{L_2(\delta)}^2 \\ &= M_8\, M_9^2\, T_2\, \underline{h}_l^{-d_l} \left\| g_l \right\|_{L_2}^2 , \end{aligned}$$

where the second inequality follows from Condition 6(c). Therefore, there exists a positive constant T_3 such that

$$\sum_{k=1}^{N_s} \beta_{lk}^2 \;\leq\; T_3\, \underline{h}_l^{-d_l} \left\| g_l \right\|_{L_2}^2 , \qquad \text{for } 1 \leq l \leq M \,, \tag{2.49}$$

where g_l is given by (2.48). We are now in a position to derive the relation in the statement of the lemma under the current assumption that ϕ is real-valued. Assume for simplicity that $\#(\, s\,) = 2$, and that in fact $s = \{1, 2\}$. Recall from (2.46) that the basis functions $g_{s1}, \ldots, g_{sN_s}$ of $\mathbb{G}_s$ were obtained by placing some order on the functions

$$g_{1k_1}(\, x_1\,)\, g_{2k_2}(\, x_2\,) \,, \qquad \text{for } 1 \leq k_1 \leq N_1 \text{ and } 1 \leq k_2 \leq N_2 \,.$$

For the moment, set $k = (k_1, k_2)$ and observe that any function given by (2.47) can be written as

$$g_s \;=\; \sum_{k_1=1}^{N_1} \sum_{k_2=1}^{N_2} \beta_k\, g_{1k_1} g_{2k_2} \,,$$

or, for some fixed $u_2 \in \mathcal{U}_2$,

$$g_s(\,\cdot\,, u_2\,) \;=\; \sum_{k_1=1}^{N_1} \left(\sum_{k_2=1}^{N_2} \beta_k\, g_{2k_2}(\, u_2\,) \right) g_{1k_1} \,.$$

With this representation, we find from (2.49),

$$\sum_{k_1=1}^{N_1} \left(\sum_{k_2=1}^{N_2} \beta_k g_{2k_2}(\, u_2\,) \right)^2 \;\leq\; T_3\, \underline{h}_1^{-d_1} \int_{\mathcal{U}_1} g_s^2(\, u_1, u_2\,)\, du_1 \,.$$

Integrating the above expression with respect to u_2 and applying (2.49) again, we

find that

$$\sum_{k_1=1}^{N_1} \sum_{k_2=1}^{N_2} \beta_k^2 \;\leq\; T_3^2 \, \underline{h}_1^{-d_1} \, \underline{h}_2^{-d_2} \, \| \, g_s \, \|_{L_2}^2$$

The desired inequality now follows from Condition 2′ for the case when $\#(\, s \,) = 2$. Proceeding by induction, assume that the inequality holds for $\#(\, s \,) = M' - 1$, and for simplicity take $s = \{1, \ldots, M'\}$. Then, as was done above, temporarily set $k = (k_1, \ldots, k_{M'})$ and observe that any function given by (2.47) can be written as

$$g_s \;=\; \sum_{k_1=1}^{N_1} \cdots \sum_{k_{M'}=1}^{N_{M'}} \beta_k \, g_{1k_1} \cdots g_{M'k_{M'}} \, .$$

Again, with this representation, we find that

$$\| \, g_s \, \|_{L_2}^2 \;=\; \int_{\mathcal{U}_1} \cdots \int_{\mathcal{U}_{M'}} \left(\sum_{k_1=1}^{N_1} \cdots \sum_{k_{M'}=1}^{N_{M'}} \beta_k \, g_{1k_1} \cdots g_{M'k_{M'}} \right)^2 du_{M'} \cdots du_1$$

which by (2.49) is bounded below by

$$T_3^{-1} \, \underline{h}_1^{d_1} \sum_{k_1=1}^{N_1} \int_{\mathcal{U}_2} \cdots \int_{\mathcal{U}_{M'}} \left(\sum_{k_2=1}^{N_2} \cdots \sum_{k_{M'}=1}^{N_{M'}} \beta_k \, g_{2k_2} \cdots g_{M'k_{M'}} \right)^2 du_{M'} \cdots du_2 \, .$$

Therefore, by induction, we find that there exists a positive constant T_4 such that

$$\sum_{k_1=1}^{N_1} \cdots \sum_{k_{M'}=M'}^{N_{M'}} \beta_k^2 \;\leq\; T_4 \, \underline{h}_s^{-d_s} \, \| \, g_s \, \|_{L_2}^2 \, .$$

The desired conclusion for the case when ϕ is a real-valued function now follows by Condition 2′.

Suppose, on the other hand, that ϕ is a vector-valued function. Then, the functions in $\mathbb{G}_s$ are of the form $g_s = (g_{1s}, \ldots, g_{Js})$, where each g_{js} is given by (2.47b). In this context we have that $\beta_s = (\beta_{1s}, \ldots, \beta_{Js})$ and hence

$$| \, \beta_s \, |^2 \;=\; \sum_{j=1}^{J} | \, \beta_{js} \, |^2 \;=\; \sum_{j=1}^{J} \sum_{k=1}^{N_s} \beta_{jsk}^2 \, .$$

In addition, by the definition in (1.11a), we know that

$$\| \, g_s \, \|^2 \;=\; \sum_{j=1}^{J} \| \, g_{js} \, \|^2 \, .$$

Therefore, by applying to each function g_{js} the result derived above for the case when ϕ is real-valued, we obtain the desired inequality when ϕ is a vector-valued function. □

Once again, assume that ϕ is a real-valued function. Then, given some ordering of the sets $s \in \mathcal{S}$, let $\boldsymbol{\beta}$ denote the column vector $\boldsymbol{\beta} = (\beta_s)$ and let $N = \sum_{s \in \mathcal{S}} N_s$ denote the length of $\boldsymbol{\beta}$. Using this construction, given any function $g \in \mathbb{G}$, there exists a coefficient vector $\boldsymbol{\beta}$ such that

$$g \;=\; \sum_{s \in \mathcal{S}} \sum_{k=1}^{N_s} \beta_{sk}\, g_{sk}\,. \tag{2.50}$$

For convenience, we summarize the relationship in (2.50) by writing $g = g(\,\cdot\,;\boldsymbol{\beta})$. Similarly, the functional $\ell(\,g\,)$ depends only on the coefficients $\boldsymbol{\beta}$ in (2.50). Therefore, we will also write $\ell(\,g\,)$ as $\ell(\,\boldsymbol{\beta}\,)$ for $g = g(\,\cdot\,;\boldsymbol{\beta})$. Let $\boldsymbol{\beta}^*$ be the coefficient vector corresponding to the function ϕ_n^* such that $\phi_n^* = \sum_s \phi_{ns}^*$, where

$$\phi_{ns}^* \;=\; \sum_{k=1}^{N_s} \beta_{sk}^*\, g_{sk} \;\;\in\;\; \mathbb{G}_s^0\,, \qquad s \in \mathcal{S}\,. \tag{2.51a}$$

Similarly, let $\widehat{\boldsymbol{\beta}}$ be the coefficient vector corresponding to the maximum likelihood estimate $\hat{\phi}$ such that $\hat{\phi} = \sum_s \hat{\phi}_s$, where

$$\hat{\phi}_s \;=\; \sum_{k=1}^{N_s} \hat{\beta}_{sk}\, g_{sk} \;\;\in\;\; \mathbb{G}_s^0\,, \qquad s \in \mathcal{S}\,. \tag{2.51b}$$

Next, assume that ϕ is a vector-valued function. Then, $\mathbb{G}$ is made up of functions of the form $g = (g_1, \ldots, g_J)$, where each g_j is given by (2.50); that is,

$$g_j \;=\; \sum_{s \in \mathcal{S}} \sum_{k=1}^{N_s} \beta_{j\,sk}\, g_{sk} \qquad \text{for } 1 \le j \le J\,. \tag{2.52}$$

For $1 \le j \le J$, let $\boldsymbol{\beta}_j$ denote the coefficient vector corresponding to the above expansion for g_j. Then, set $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$ and write $g = g(\,\cdot\,;\boldsymbol{\beta})$ to summarize the relationship given in (2.52). Similarly, the functional $\ell(\,g\,)$ depends only on the coefficients $\boldsymbol{\beta}$ in (2.52), and as was done above, we write $\ell(\,g\,)$ as $\ell(\,\boldsymbol{\beta}\,)$ for $g = g(\,\cdot\,;\boldsymbol{\beta})$. Let $\boldsymbol{\beta}^*$ be the coefficients corresponding to the function $\phi_n^* = (\phi_{1n}^*, \ldots, \phi_{Jn}^*)$ such that each function $\phi_{jn}^* = \sum_s \phi_{jns}^*$, where

$$\phi_{j\,ns}^* \;=\; \sum_{k=1}^{N_s} \beta_{j\,sk}^*\, g_{sk} \;\;\in\;\; \mathbb{G}_s^0\,, \qquad s \in \mathcal{S} \text{ and } 1 \le j \le J\,. \tag{2.53a}$$

Similarly, let $\widehat{\boldsymbol{\beta}}$ be the coefficient vector corresponding to the maximum likelihood estimate $\hat{\phi} = (\hat{\phi}_1, \ldots, \hat{\phi}_J)$ such that each function $\hat{\phi}_j = \sum_s \phi^* js$, where

$$\hat{\phi}_{j\,s} \;=\; \sum_{k=1}^{N_s} \hat{\beta}_{j\,sk}\, g_{sk} \;\;\in\;\; \mathbb{G}_s^0\,, \qquad s \in \mathcal{S} \text{ and } 1 \le j \le J\,. \tag{2.53b}$$

When ϕ is a real-valued function, for each $s \in \mathcal{S}$, let $\mathbf{S}$ denote the N-dimensional column vector having entries

$$\frac{\partial}{\partial \beta_{sk}} \ell(\boldsymbol{\beta}) .$$

Since $\widehat{\boldsymbol{\beta}}$ is a maximum likelihood estimate of $\boldsymbol{\beta}$ we have that $\mathbf{S}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, which we can rewrite as

$$\int_0^1 \frac{d}{d\alpha} \mathbf{S}\big(\boldsymbol{\beta}^* + \alpha(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\big)\, d\alpha = -\mathbf{S}(\boldsymbol{\beta}^*) . \tag{2.54}$$

Furthermore, if we let $\mathbf{H}(\boldsymbol{\beta})$ denote the N-by-N dimensional matrix having entries

$$\frac{\partial^2}{\partial \beta_{s_1 k_1} \partial \beta_{s_2 k_2}} \ell(\boldsymbol{\beta})$$

and set

$$\mathbf{D} = \int_0^1 \mathbf{H}\big(\boldsymbol{\beta}^* + \alpha(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\big)\, d\alpha ,$$

then we can rewrite (2.54) as

$$\mathbf{D}\,(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = -\mathbf{S}(\boldsymbol{\beta}^*) ,$$

which implies that

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^t\, \mathbf{D}\,(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = -(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^t\, \mathbf{S}(\boldsymbol{\beta}^*) . \tag{2.55}$$

If ϕ is a vector-valued function, then we extend the above definitions in an obvious way. In the following lemma, we present conditions under which we achieve the proper rate of convergence.

Lemma 2.11 *Assume that Conditions* 1, 2′, 3″ *and* 6 *hold. In addition, assume that*

$$\big|\, \mathbf{S}(\boldsymbol{\beta}^*) \,\big|^2 = O_P(1/n) , \tag{2.56a}$$

and that there exists a positive constant M_9 *such that*

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^t\, \mathbf{D}\,(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \leq -M_8 \,\|\, \hat{\phi} - \phi_n^* \,\|^2 , \tag{2.56b}$$

except on a set with probability tending to zero with n. *Then,*

$$\|\, \hat{\phi}_s - \phi_{ns}^* \,\|^2 = O_P(\omega(\underline{h}, \mathcal{S})/n) , \qquad s \in \mathcal{S} , \tag{2.57a}$$

and

$$\|\, \sigma(\hat{\phi}) - \sigma(\phi_n^*) \,\|^2 = O_P(\omega(\underline{h}, \mathcal{S})/n) . \tag{2.57b}$$

Proof Let $T_1, T_2 \ldots$ denote suitable positive constants. Assume initially that ϕ is a real-valued function. From (2.55) and (2.56b) we find that except on a set whose probability tends to zero with n,

$$\begin{aligned} \| \phi_n^* - \hat{\phi} \|^2 &\leq -(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^t \mathbf{D} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) / M_8 \\ &= (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^t \mathbf{S}(\boldsymbol{\beta}^*) / M_8 \\ &\leq |\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*| |\mathbf{S}(\boldsymbol{\beta}^*)| / M_8 . \end{aligned} \tag{2.58}$$

By Lemma 2.10 we find that there exists a positive constant M_8 such that

$$|\hat{\beta}_s - \beta_s^*|^2 \leq M_8 \underline{h}_s^{-d_s} \| \hat{\phi}_s - \phi_{ns}^* \|^2 , \qquad s \in \mathcal{S} ,$$

and hence that

$$|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|^2 = \sum_{s \in \mathcal{S}} |\hat{\beta}_s - \beta_s^*|^2 \leq M_8 \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} \| \hat{\phi}_s - \phi_{ns}^* \|^2 . \tag{2.59}$$

Next, observe that by the Schwartz inequality

$$\sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} \leq \left(\#(\mathcal{S}) \sum_{s \in \mathcal{S}} \underline{h}_s^{-2d_s} \right)^{1/2} = \sqrt{\#(\mathcal{S})}\, \omega(\underline{h}, \mathcal{S}) ,$$

so that from (2.59) we have that except on a set whose probability tends to zero with n,

$$|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|^2 \leq M_8 \sqrt{\#(\mathcal{S})}\, \omega(\underline{h}, \mathcal{S}) \sum_{s \in \mathcal{S}} \| \hat{\phi}_s - \phi_{ns}^* \|^2 .$$

Therefore, since $\hat{\phi}_s - \phi_{ns}^* \in \mathbb{G}_s$ for $s \in \mathcal{S}$, we find from Lemma 3.3 in Chapter 1 that there exists a positive constant T_2 such that except on a set whose probability tends to zero with n,

$$|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|^2 \leq T_2\, \omega(\underline{h}, \mathcal{S}) \| \hat{\phi} - \phi_n^* \|^2 .$$

Combining this result with the relation in (2.58), we find that except on a set whose probability tends to zero with n,

$$\| \hat{\phi} - \phi_n^* \|^2 \leq \sqrt{T_2\, \omega(\underline{h}, \mathcal{S})}\, \| \hat{\phi} - \phi_n^* \| \, |\mathbf{S}(\boldsymbol{\beta}^*)| / M_8 ,$$

or, by setting $T_3 = T_2 / M_8^2$,

$$\| \hat{\phi} - \phi_n^* \|^2 \leq T_3\, \omega(\underline{h}, \mathcal{S}) |\mathbf{S}(\boldsymbol{\beta}^*)|^2 . \tag{2.60}$$

Combining (2.60) with the relation in (2.56a) we find from another application of Lemma 3.3 in Chapter 1 that the relation in (2.57a) holds. In addition, applying

the result of Lemma 1 in the Appendix and Condition 3″, we find that

$$\left\| \hat{\phi} - \phi_n^* \right\|_\infty^2 \;=\; O_P\big(\omega^2(\underline{h}, \mathcal{S})/n\big) \;=\; o_P(1). \tag{2.61}$$

Now, recall from Lemma 2.2 and Condition 3″ that

$$\left\| \sigma(\phi_n^*) - \sigma(\phi^*) \right\|_\infty \;=\; o(1),$$

and hence that

$$\left\| \sigma(\phi_n^*) \right\|_\infty \;=\; O(1).$$

Therefore, arguing as after (2.32c), we find that $\| \phi_n^* \|_\infty = O(1)$. Finally, from (2.57a), (2.61) and the continuity condition (1.5), we obtain the relation in (2.57b).

On the other hand, suppose that the function ϕ is vector-valued. Then, the arguments above can essentially be applied to the components $\hat{\phi}_j - \phi_{j\,s}^*$ for $1 \le j \le J$, to obtain the desired result. □

We now take up the task of verifying the conditions in (2.56) for the cases we considered in Examples 1–3. After this example, we present a general theorem stating the rate of convergence of the sample based estimate $\hat{\phi}$ and its components to ϕ^* and its components in each of these extended linear models.

Example 4 **[R]** Let $g = g(\,\cdot\,;\boldsymbol{\beta}) \in \mathbb{G}$ be given by (2.50). Recall that in the regression context

$$l(g, \mathbf{w}) \;=\; -g^2(\,\mathbf{u}\,;\boldsymbol{\beta}) \;+\; 2\,g(\,\mathbf{u}\,;\boldsymbol{\beta})\,v\,,$$

and hence that

$$\frac{\partial}{\partial \beta_{sk}}\, l\big(\,g(\mathbf{u};\boldsymbol{\beta}), \mathbf{w}\,\big) \;=\; -2\,g_{sk}(\,\mathbf{u}\,)\,\big(\,g(\,\mathbf{u}\,;\boldsymbol{\beta}) \;-\; v\,\big).$$

Therefore, by the definition of $\boldsymbol{\beta}^*$, we have that

$$E\Big[\, -2\,g_{sk}(\,\mathbf{U}\,)\,\big(\,\phi_n^*(\,\mathbf{U}\,) \;-\; V\,\big)\,\Big] \;=\; 0\,. \tag{2.62}$$

From Lemma 2.2 and Condition 3″, we know that ϕ_n^* is bounded. Combining this with our assumption (2.25) on the distribution of V (see Example 3), and (2.62), we find that there exists a positive constant T_1 such that

$$\begin{aligned} E\Big[\,\big|\,\mathbf{S}(\boldsymbol{\beta}^*)\,\big|^2\,\Big] \;&=\; (1/n) \sum_{s \in \mathcal{S}} \sum_{k=1}^{N_s} 4\ \mathrm{var}\Big(\,g_{sk}(\,\mathbf{U}\,)\,\big(\,\phi_n^*(\,\mathbf{U}\,) \;-\; V\,\big)\,\Big) \\ &\le\; (T_1/n) \sum_{s \in \mathcal{S}} \sum_{k=1}^{N_s} E\big[\,g_{sk}^2(\,\mathbf{U}\,)\,\big] \\ &=\; O(1/n)\,, \end{aligned}$$

where the last relation follows from Conditions 6(a) and 6(c). This establishes (2.56a). As for (2.56b), let $g_\alpha = \phi_n^* + \alpha(\hat{\phi} - \phi_n^*)$, and observe that

$$\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)^t \mathbf{D} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = -2 \int_0^1 \| g_\alpha \|_n^2 \, d\alpha .$$

The desired relation now follows immediately from (3.12) in Chapter 1.

[GR] Again, let $g = g(\cdot\,;\boldsymbol{\beta}) \in \mathbb{G}$, and recall that in the generalized regression context,

$$l(g, \mathbf{w}) = b_1\big(g(\mathbf{u}\,;\boldsymbol{\beta})\big)\, v + b_2\big(g(\mathbf{u}\,;\boldsymbol{\beta})\big) ,$$

and hence that

$$\frac{\partial}{\partial \beta_{sk}} l\big(g(\mathbf{u}\,;\boldsymbol{\beta}), \mathbf{w}\big) = g_{sk}(\mathbf{u}) \left(b_1'\big(g(\mathbf{u}\,;\boldsymbol{\beta})\big)\, v + b_2'\big(g(\mathbf{u}\,;\boldsymbol{\beta})\big)\right).$$

Therefore, proceeding as in the regression context, we find that from the definition of $\boldsymbol{\beta}^*$,

$$E\Big[g_{sk}(\mathbf{U}) \left(b_1'\big(\phi_n^*(\mathbf{U})\big)\, V + b_2'\big(\phi_n^*(\mathbf{U})\big)\right)\Big] = 0 . \tag{2.63}$$

From Lemma 2.2 and Condition 3″, we know that ϕ_n^* is bounded. Combining this with our assumption (2.25) on the distribution of V (see Example 3), and (2.63), we find that there exists a positive constant T_1 such that

$$\begin{aligned} & E\Big[\left| \mathbf{S}(\boldsymbol{\beta}^*) \right|^2 \Big] \\ & \quad = (1/n) \sum_{s \in \mathcal{S}} \sum_{k=1}^{N_s} \operatorname{var}\Big(g_{sk}(\mathbf{U}) \left(b_1'\big(\phi_n^*(\mathbf{U})\big)\, V + b_2'\big(\phi_n^*(\mathbf{U})\big)\right)\Big) \\ & \quad \le (T_1/n) \sum_{s \in \mathcal{S}} \sum_{k=1}^{N_s} E\big[g_{sk}^2(\mathbf{U}) \big] \\ & \quad = O(1/n) , \end{aligned}$$

where the last relation follows from Conditions 6(a) and 6(c). This establishes the condition in (2.56a).

As for (2.56b), observe that from Lemma 1 in the Appendix, Condition 3″ and Lemma 2.9, we find that there exists a positive constant T_2 such that

$$\lim_{n \to \infty} P\big(\| \phi_n^* \|_\infty \le T_2 \text{ and } \| \hat{\phi} \|_\infty \le T_2 \big) = 1 . \tag{2.64}$$

Now, given the positive constant T_3, set

$$\mathcal{V}_0 = \big\{ v \in \mathcal{V} : b_1''(\eta)\, v - b_2''(\eta) \le -T_3 \text{ for } |\eta| \le T_2 \big\} .$$

Therefore, by (1.15) we can choose T_3 so small that

$$P\big(V \in \mathcal{V}_0 \,|\, \mathbf{U} = \mathbf{u}\big) \;\geq\; T_3 \qquad \text{for all } \mathbf{u} \in \mathcal{U}\,.$$

For each $1 \leq i \leq n$, let I_i denote the indicator random variable such that $I_i = 1$ if $V_i \in \mathcal{V}_0$ and zero otherwise. Note that the conditional density of $\mathbf{U}$ given that $V \in \mathcal{V}_0$ is bounded away from zero and infinity on $\mathcal{U}$. Temporarily let $\| \cdot \|_n$ denote the empirical norm defined in (1.6b), using the $I_1, \ldots, I_n$ as censoring variables. Then, from (1.15), (2.64) and the definition of $\mathbf{D}$, we find that except on a set whose probability tends to zero with n,

$$-(\,\widehat{\boldsymbol{\beta}} \;-\; \boldsymbol{\beta}^*\,)^t\,\mathbf{D}\,(\,\widehat{\boldsymbol{\beta}} \;-\; \boldsymbol{\beta}^*\,) \;\geq\; T_4 \,\big\|\,\hat{\phi} - \phi_n^*\,\big\|_n^2\,, \tag{2.65}$$

for some suitable positive constant T_4. Now, since ϕ_n^* and $\hat{\phi}$ are each in $\mathbb{G}$, we have from (3.12) in Chapter 1 that except on a set whose probability tends to zero with n,

$$2\,\big\|\,\hat{\phi} - \phi^*\,\big\|_n^2 \;\geq\; \big\|\,\hat{\phi} - \phi^*\,\big\|^2\,,$$

where $\| \cdot \|$ denotes the theoretical norm defined in (1.6a), using the indicator I of the event that $V \in \mathcal{V}_0$ as a censoring variable. The desired result now follows from Condition $2'$ and our earlier observations about the distribution of $\mathbf{U}$ given that $V \in \mathcal{V}_0$.

[CR] Let $g = g(\,\cdot\,;\boldsymbol{\beta}\,) \in \mathbb{G}$, and recall that in the context of censored regression,

$$\begin{aligned} l(g, \mathbf{w}) \;\;=\;\; & -v_2\,\big(\,g^2(\,\mathbf{u}\,;\boldsymbol{\beta}\,) \;-\; 2\,g(\,\mathbf{u}\,;\boldsymbol{\beta}\,)\,v_1\,\big)\,/\,2 \\ & +\;\big(\,1 - v_2\,\big)\,\log\Big(\,1 - \Phi\big(\,g(\,\mathbf{u}\,;\boldsymbol{\beta}\,)\,\big)\,\Big)\,, \end{aligned}$$

and hence that

$$\begin{aligned} \frac{\partial}{\partial \beta_{sk}}\, l\big(\,g(\,\mathbf{u}\,;\boldsymbol{\beta}\,) \;\;=\;\; & -v_2\,g_{sk}(\,\mathbf{u}\,)\,\big(\,g(\,\mathbf{u}\,;\boldsymbol{\beta}\,) \;-\; v_1\,\big) \\ & -\;\big(\,1 - v_2\,\big)\,g_{sk}(\,\mathbf{u}\,)\,\frac{\varphi\big(\,g(\,\mathbf{u}\,;\boldsymbol{\beta}\,)\,\big)}{\Big(\,1 - \Phi\big(\,g(\,\mathbf{u}\,;\boldsymbol{\beta}\,)\,\big)\,\Big)}\,. \end{aligned}$$

As in the previous two cases, we have from the definition of $\boldsymbol{\beta}^*$ that

$$E\left[\,\frac{\partial}{\partial \beta_{sk}}\, l\big(\,g(\,\mathbf{U}\,;\boldsymbol{\beta}^*\,)\,\right] \;\;=\;\; 0\,. \tag{2.66}$$

From Lemma 2.2 and Condition $3''$, we know that ϕ_n^* is bounded. Combining this with our assumption (2.25) on the distribution of V (see Example 3),

and (2.66),we find that there exists a positive constant T_1 such that

$$\begin{aligned} E\Big[\,\big|\,\mathbf{S}(\,\boldsymbol{\beta}^*\,)\,\big|^2\,\Big] &= (\,1/n\,)\sum_{s\in\mathcal{S}}\sum_{k=1}^{N_s}\mathrm{var}\left(\frac{\partial}{\partial\beta_{sk}}\,l\big(\,g(\,\mathbf{U}\,;\boldsymbol{\beta}^*\,)\,\big)\right) \\ &\le (\,T_1/n\,)\sum_{s\in\mathcal{S}}\sum_{k=1}^{N_s}E\big[\,g_{sk}^2(\,\mathbf{U}\,)\,\big] \\ &= O(\,1/n\,)\,, \end{aligned}$$

where the last relation follows from Conditions 6(a) and 6(c). This establishes (2.56a).

As for (2.56b), observe that from Lemma 1 in the Appendix, Condition $3''$ and Lemma 2.9, we find that there exists a positive constant T_2 such that

$$\lim_{n\to\infty} P\big(\,\|\,\phi_n^*\,\|_\infty \le T_2 \text{ and } \|\,\hat{\phi}\,\|_\infty \le T_2\,\big) = 1\,. \tag{2.67}$$

Set $\boldsymbol{\beta}_\alpha = \boldsymbol{\beta}^* + \alpha(\,\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\,)$ and by arguing as in Example 2, we find that

$$-(\,\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\,)^t\,\mathbf{H}(\,\boldsymbol{\beta}_\alpha\,)\,(\,\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\,)$$

is given by

$$E_n\Big[\,\big(\,\hat{\phi}(\,\mathbf{U}\,) - \phi_n^*(\,\mathbf{U}\,)\,\big)^2\,B\big(\,g(\,\mathbf{U}\,;\boldsymbol{\beta}_\alpha\,),\mathbf{W}\,\big)\,\Big]\,,$$

where the strictly positive function $B(\,\cdot\,,\mathbf{w})$ is defined in Example 2. Therefore, by (2.67) and the continuity of $B(\,\cdot\,,\mathbf{w})$, we find that there exists a positive constant T_3 such that except on a set whose probability tends to zero with n,

$$-(\,\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\,)^t\,\mathbf{H}(\,\boldsymbol{\beta}_\alpha\,)\,(\,\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\,) \ge T_3\,\big\|\,\hat{\phi} - \phi_n^*\,\big\|_n^2\,. \tag{2.68}$$

Now, since $\hat{\phi}$ and ϕ_n^* are each in $\mathbb{G}$, we have from (3.12) in Chapter 1 that except on a set whose probability tends to zero with n,

$$2\,\big\|\,\hat{\phi} - \phi^*\,\big\|_n^2 \ge \big\|\,\hat{\phi} - \phi^*\,\big\|^2\,. \tag{2.69}$$

The desired conclusion now follows from (2.68) and (2.69).

[PR] In this context, ϕ is vector valued, and hence we have that each function $g \in \mathbb{G}$ is of the form

$$g = g(\,\cdot\,;\boldsymbol{\beta}) = \big(\,g_1(\,\cdot\,;\boldsymbol{\beta}_1),\ldots,g_J(\,\cdot\,;\boldsymbol{\beta}_J\,)\,\big)\,.$$

Recall that in this context

$$l(g, \mathbf{w}) = g_1(\mathbf{u};\boldsymbol{\beta}_1)\, I_1(v) + \cdots + g_J(\mathbf{u};\boldsymbol{\beta}_J)\, I_J(v) - \log\Big(1 + \exp\big(g_1(\mathbf{u};\boldsymbol{\beta}_1)\big) + \cdots + \exp\big(g_J(\mathbf{u};\boldsymbol{\beta}_J)\big)\Big),$$

and hence,

$$\frac{\partial}{\partial \beta_{jsk}}\, l\big(g(\mathbf{u};\boldsymbol{\beta})\,\mathbf{w}\big) = g_{jsk}(\mathbf{u})\Big(I_j(v) - \pi_j\big(g(\mathbf{u};\boldsymbol{\beta})\big)\Big),$$

where for $1 \le j \le J$, we temporarily set

$$\pi_j\big(g(\mathbf{u};\boldsymbol{\beta})\big) = \frac{\exp\big(g_j(\mathbf{u};\boldsymbol{\beta}_j)\big)}{1 + \exp\big(g_1(\mathbf{u};\boldsymbol{\beta}_1)\big) + \cdots + \exp\big(g_J(\mathbf{u};\boldsymbol{\beta}_J)\big)}.$$

As in the previous examples, we have from the definition of $\boldsymbol{\beta}^*$ that

$$E\Big[\, g_{jsk}(\mathbf{U})\Big(I_j(V) - \pi_j\big(g(\mathbf{U};\boldsymbol{\beta})\big)\Big)\,\Big] = 0. \tag{2.70}$$

From Lemma 2.2 and Condition 3″, we know that ϕ_n^* is bounded. Combining this with our assumption (2.25) on the distribution of V (see Example 3), and (2.63), we find that there exists a positive constant T_1 such that

$$\begin{aligned} E\Big[\,\big|\mathbf{S}(\boldsymbol{\beta}^*)\big|^2\,\Big] &= (1/n)\sum_{j=1}^{J}\sum_{s\in\mathcal{S}}\sum_{k=1}^{N_s} \operatorname{var}\Big(\, g_{jsk}(\mathbf{U})\Big(I_j(V) - \pi_j\big(g(\mathbf{U};\boldsymbol{\beta})\big)\Big)\,\Big) \\ &\le (T_1/n)\sum_{j=1}^{J}\sum_{s\in\mathcal{S}}\sum_{k=1}^{N_s} E\big[\, g_{jsk}^2(\mathbf{U})\,\big] \\ &= O(1/n), \end{aligned}$$

where the last relation follows from Conditions 6(a) and 6(c). This establishes the condition in (2.56a).

As for (2.56b), observe that from Lemma 1 in the Appendix, Condition 3″ and Lemma 2.9, we find that there exists a positive constant T_2 such that

$$\lim_{n\to\infty} P\big(\,\|\phi_n^*\|_\infty \le T_2 \text{ and } \|\hat{\phi}\|_\infty \le T_2\,\big) = 1. \tag{2.71}$$

Set $\boldsymbol{\beta}_\alpha = \boldsymbol{\beta}^* + \alpha(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ and by arguing as in Example 2, we find that

$$-(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^t\, \mathbf{H}(\boldsymbol{\beta}_\alpha)\,(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$$

is bounded below by

$$E_n\left[\pi_0\big(g_\alpha(\mathbf{U})\big)\sum_{j=1}^{J}\pi_j\big(g_\alpha(\mathbf{U})\big)\big(\hat{\phi}_j(\mathbf{U})-\phi^*_{jn}(\mathbf{U})\big)^2\right], \tag{2.72}$$

where the functions $\pi_j(g_\alpha,\mathbf{w})$, $0\le j\le J$, are defined in Example 2. Therefore, by (2.71) and the continuity of the functions π_j, $0\le j\le J$, we find that there exists a positive constant T_3 such that except on a set whose probability tends to zero with n,

$$-(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^t\,\mathbf{H}(\boldsymbol{\beta}_\alpha)\,(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\;\ge\;T_3\,\|\hat{\phi}-\phi^*_n\|_n^2. \tag{2.73}$$

Now, since $\hat{\phi}$ and ϕ^*_n are each in $\mathbb{G}$, we have from (3.12) in Chapter 1 that except on a set whose probability tends to zero with n,

$$2\,\|\hat{\phi}-\phi^*\|_n^2\;\ge\;\|\hat{\phi}-\phi^*\|^2. \tag{2.74}$$

The desired conclusion now follows from (2.73) and (2.74).

[D] In the density estimation context, we have that

$$\frac{\partial}{\partial\beta_{sk}}\,l\big(g(\mathbf{u}\,;\boldsymbol{\beta}),\mathbf{w}\big)\;=\;g_{sk}(\mathbf{u})-\int_{\mathcal{U}}g_{sk}(\mathbf{u})\exp\big(\sigma(\phi^*_n)\big)\,d\mathbf{u},$$

and hence that

$$E\Big[g_{sk}(\mathbf{U})-\int_{\mathcal{U}}g_{sk}(\mathbf{u})\,\exp\big(\sigma(\phi^*_n)\big)\,d\mathbf{u}\Big]\;=\;0.$$

Therefore, as seen in the previous example, from Condition 6(a) and (c), and the boundedness of ϕ^*_n,

$$E\Big[\,\big|\mathbf{S}(\boldsymbol{\beta}^*)\big|^2\Big]\;=\;(1/n)\sum_{s\in\mathcal{S}}\sum_{k=1}^{N_s}\mathrm{var}\big(g_{sk}(\mathbf{U})\big)\;=\;O(1/n).$$

Now, set $\boldsymbol{\beta}_\alpha=\boldsymbol{\beta}^*+\alpha(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)$ so that

$$-(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^t\,\mathbf{H}(\boldsymbol{\beta}_\alpha)\,(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\;=\;\mathrm{var}\big(\phi^*(\mathbf{U}_\alpha)-\hat{\phi}(\mathbf{U}_\alpha)\big), \tag{2.75}$$

where $\mathbf{U}_\alpha$ is a random vector on $\mathcal{U}$ having log-density given by $\sigma\big(g(\cdot\,:\boldsymbol{\beta}_\alpha)\big)$. As in the previous examples, we find that there exists a positive constant T_1 such that

$$\lim_{n\to\infty}P\big(\|\phi^*_n\|_\infty\le T_1\text{ and }\|\hat{\phi}\|_\infty\le T_1\big)\;=\;1.$$

Therefore, there exists a positive constant T_2 such that except on a set whose probability tends to zero with n, the density of $\mathbf{U}_\alpha$ satisfies

$$\frac{1}{T_2}\;\le\;f_{\mathbf{U}_\alpha}(\mathbf{u})\;\le\;T_2\qquad\text{for }\mathbf{u}\in\mathcal{U}. \tag{2.76}$$

Now, from Lemma 3.4 in Chapter 1, given any positive constant T_3, except on a set whose probability tends to zero with n,

$$\left| \langle g, 1 \rangle_n - \langle g, 1 \rangle \right| \leq T_3 \left\| g \right\|, \qquad g \in \mathbb{G}_1 .$$

However, since $\mathbb{G}$ is the space of functions in $\mathbb{G}_1$ that are orthogonal to the constant function 1, we find that except on a set whose probability tends to zero with n,

$$\Big(E\big[g(\mathbf{U}) \big] \Big)^2 \leq T_3^2 \left\| g \right\|^2, \qquad \text{for } g \in \mathbb{G},$$

and hence

$$\text{var}\big(g(\mathbf{U}) \big) = \left\| g \right\|^2 - \Big(E\big[g(\mathbf{U}) \big] \Big)^2 \geq \big(1 - T_3^2 \big) \left\| g \right\|^2 . \tag{2.77}$$

Combining (2.75)–(2.77) with the definition of $\mathbf{D}$ and Condition $2'$, we arrive at the desired relation.

Theorem 2.1 *Assume that Conditions* 1, $2'$, $3''$, $5'$ *and* 6 *hold, and that the relations in* (2.56) *are satisfied. Then,*

$$\left\| \sigma(\hat{\phi}) - \sigma(\phi^*) \right\| = O_P\big(\rho_s^2(\overline{h}_s) + \omega(\underline{h}, \mathcal{S}) / n \big)$$

and for each $s \in \mathcal{S}$,

$$\left\| \hat{\phi}_s - \phi_s^* \right\| = O_P\big(\rho_s^2(\overline{h}_s) + \omega(\underline{h}, \mathcal{S}) / n \big) .$$

Proof The theorem follows directly from Lemmas 2.3 and 2.11, Condition $3''$ and the continuity of $\sigma(\cdot)$. □

3.3 Conditional Density Estimation

In Sections 3.1 and 3.2, we introduced the notion of an extended linear model and discussed five examples in considerable depth. In this section we again treat a special case of an extended linear model, but one that is sufficiently different to warrant a separate discussion. To be precise, in the context of conditional density estimation, we make use of the alternate ANOVA decompositions introduced in Section 1.5. The apparent need for this alternate decomposition will be discussed at various places throughout this section. In general, however, we will follow the outline of the arguments in Sections 3.1 and 3.2 quite closely.

Preliminaries

Recall that ϕ is the solution to an optimization problem over a saturated space of functions whose domain contains $\mathcal{U}$; that is, given a functional $l(f, \mathbf{w})$ of the form

$$l(\, f, \mathbf{w}\,) \;=\; \sum_{k=1}^{K} b_k(\, f, \mathbf{u}\,)\, B_k(\, \mathbf{w}\,), \qquad \mathbf{w} = (\mathbf{u}, \mathbf{v}) \in \mathcal{W},$$

we assume that

$$\phi \;=\; \underset{f \in L_\infty(\mathcal{U})}{\operatorname{argmax}}\; \Lambda(\, f\,),$$

where $\Lambda(\phi)$ is finite. In the context of conditional density estimation, we take $\mathbf{w} = \mathbf{u}$, where $\mathbf{u}$ is in turn partitioned into $(\mathbf{u}_1, \mathbf{u}_2)$. Write $\mathcal{U} = \mathcal{U}_1' \times \mathcal{U}_2'$ and set

$$\exp(\, c(\, \mathbf{u}_1, f\,)\,) \;=\; \int_{\mathcal{U}_2'} \exp(\, f(\mathbf{u}_1, \mathbf{u}_2)\,)\, d\mathbf{u}_2\,.$$

Clearly, the function

$$\exp\,\big(\, f(\mathbf{u}_1, \mathbf{u}_2) - c(\, \mathbf{u}_1, f\,)\,\big)$$

integrates to one over $\mathcal{U}_2'$, and hence if we set

$$l(f, \mathbf{w}) \;=\; l(f, \mathbf{u}_1, \mathbf{u}_2) \;=\; f(\mathbf{u}_1, \mathbf{u}_2) - c(\, \mathbf{u}_1, f\,),$$

it follows from the information inequality that $\phi(\mathbf{u}_1, \mathbf{u}_2)$ is the log of the conditional density of $\mathbf{U}_2$ given that $\mathbf{U}_1 = \mathbf{u}_1$. As in the density estimation context treated in the previous two sections, the functional $\Lambda(\,\cdot\,)$ does not have a unique maximum over $L_\infty(\mathcal{U})$. Any function of the form $\phi(\mathbf{u}_1, \mathbf{u}_2) + f(\mathbf{u}_1)$, where f is a bounded function depending only on the variable $\mathbf{u}_1$ maximizes $\Lambda(\,\cdot\,)$. However, by restricting our search to only valid conditional density functions, $\phi(\mathbf{u}_1, \mathbf{u}_2)$ is obtained as the essentially unique function maximizing $\Lambda(\,\cdot\,)$.

As in the density estmation context, we accomodate this restriction by considering estimates of ϕ that are of the form $f - c(\,\cdot\,; f)$, where f is any bounded function. Again, all that we require from the map $c(\,\cdot\,)$ is that it be Lipshitz continuous in some neighborhood of zero. In other words, it is assumed that for every positive constant T_1, there is are positive constants T_2 and T_3 such that, for all pairs of real-valued functions f_1, f_2 defined on $\mathcal{U}$ satisfying

$$\|\, f_1\,\|_\infty \;\le\; T_1 \qquad \text{and} \qquad \|\, f_2\,\|_\infty \;\le\; T_1\,, \tag{3.1a}$$

we have that

$$\big\|\, c(\,\cdot\,; f_1) \,-\, c(\,\cdot\,; f_2)\,\big\|_\infty \;\le\; T_2\,\big\|\, f_1 - f_2\,\big\|_\infty\,, \tag{3.1b}$$

and that

$$\big\|\, c(\,\cdot\,; f_1) \,-\, c(\,\cdot\,; f_2)\,\big\| \;\le\; {T_3}^*\big\|\, f_1 - f_2\,\big\|\,, \tag{3.1c}$$

where $^*\|\cdot\|$ is equivalent to the L_2-norm on $\mathcal{U}$ and will be given explicitly in the next subsection. A direct computation indicates that the relations in (3.1) hold for the functional $c(\cdot)$ as defined above.

ANOVA Decompositions

Recall that the random vector $\mathbf{U} = (U_1, \ldots, U_M)$ takes values in $\mathcal{U} = \mathcal{U}_1 \times \cdots \times \mathcal{U}_M$, and for each $1 \leq l \leq M$ let $f_l(u_l)$ denote the marginal density of U_l on $\mathcal{U}_l$. Set $f^*_{\mathbf{U}}$ equal to product of these marginal densities, and let E^* denote the expectation operator with repect to $f^*_{\mathbf{U}}$. In this context, we do not allow for censoring, and hence under Condition $2'$ we know that the function $f^*_{\mathbf{U}}$ is also bounded above and below by a constant. For convenience, throughout this section we will take these constants to be M_3 and $1/M_3$, respectively. Next, let $\mathbf{U}_1, \ldots, \mathbf{U}_n$ denote a random sample of size n from the distribution of the random vector $\mathbf{U}$. For any function $f(\cdot)$ defined on $\mathcal{U}$, set

$$E^*_n\big[\, f(\mathbf{U}) \,\big] = \frac{1}{n^M} \sum_{i_1=1}^{n} \cdots \sum_{i_M=1}^{n} f(\mathbf{U}_{1i_1}, \ldots, \mathbf{U}_{Mi_M}) \,.$$

Observe that E^*_n is the expectation operator with respect to the product of the marginal empirical distributions induced by the sample $\mathbf{U}_1, \ldots, \mathbf{U}_n$. For any two bounded, real-valued functions f_1, f_2 whose domain includes $\mathcal{U}$, set

$$^*\big\langle\, f_1, f_2 \,\big\rangle \;=\; E^*\big[\, f_1(\mathbf{U})\, f_2(\mathbf{U}) \,\big] \quad \text{and} \quad ^*\big\|\, f_1 \,\big\|^2 \;=\; E^*\big[\, f_1^2(\mathbf{U}) \,\big] \,. \tag{3.2a}$$

Similarly, we define the empirical inner-product and norm by

$$^*\big\langle\, f_1, f_2 \,\big\rangle_n \;=\; E^*_n\big[\, f_1(\mathbf{U})\, f_2(\mathbf{U}) \,\big] \quad \text{and} \quad ^*\big\|\, f_1 \,\big\|^2_n \;=\; E^*_n\big[\, f_1^2(\mathbf{U}) \,\big] \,. \tag{3.2b}$$

As mentioned above, in the context of conditional density estimation, we use these inner products and norms to generate the various ANOVA decompositions introduced in Section 3.1.

Toward this end, let $\mathcal{S}_1$ denote a hierarchical collection of subsets of $\{1, \ldots, M\}$ as usual. Then, let $\mathcal{S}$ denote all those sets in $\mathcal{S}_1$ that contain at least one index corresponding to a variable in the subvector $\mathbf{u}_2$. Clearly, there are many possible collections $\mathcal{S}_1$ that give rise to the same set $\mathcal{S}$, and so without loss of generality we assume that $\mathcal{S}_1$ is the smallest (in terms of cardinality) such collection possible. Then, we set $\mathcal{S}_0 = \mathcal{S}_1 \setminus \mathcal{S}$. Observe that by constructing $\mathcal{S}_0$ in this manner, the sets in $\mathcal{S}_0$ do not contain indices corresponding the variables in $\mathbf{u}_2$. As in the previous sections, for each $s \in \mathcal{S}_1$, let $\mathbb{F}_s$ denote the space of square-integrable functions on $\mathcal{U}$ that depend only on the variables u_l for $l \in s$. From these spaces we construct

$$\mathbb{F}_1 \;=\; \left\{ \sum_{s \in \mathcal{S}_1} f_s : f_s \in \mathbb{F}_s \;\text{ for } s \in \mathcal{S}_1 \right\}, \tag{3.3a}$$

and

$$\mathbb{F}_0 \;=\; \left\{ \sum_{s \in \mathcal{S}_0} f_s : f_s \in \mathbb{F}_s \;\text{ for } s \in \mathcal{S}_0 \right\}. \tag{3.3b}$$

For each $s \in \mathcal{S}_1$, let ${}^*\mathbb{F}_s^0$ denote the space of functions in $\mathbb{F}_s$ that are orthogonal to $\mathbb{F}_r$ relative to the theoretical inner-product in (3.2a), where r a proper subset of s. Then, define the space

$${}^*\mathbb{F} \;=\; \left\{ \sum_{s \in \mathcal{S}} f_s : f_s \in {}^*\mathbb{F}_s^0 \;\text{ for } s \in \mathcal{S} \right\}. \tag{3.3c}$$

By construction, $\mathbb{F}_1$ is the direct sum of the spaces ${}^*\mathbb{F}$ and $\mathbb{F}_0$, and under Condition 2′ we know from the analog of Lemma 3.7 in Chapter 1 that was established for the alternate ANOVA decomposition in Section 1.5, that each function $f_1 \in \mathbb{F}_1$ can be represented uniquely as the sum of the functions $f \in {}^*\mathbb{F}$ and $f_0 \in \mathbb{F}_0$, which can in turn be written uniquely as

$$f \;=\; \sum_{s \in \mathcal{S}} f_s\,, \qquad \text{where } f_s \in {}^*\mathbb{F}_s^0 \;\text{ for } s \in \mathcal{S}\,, \tag{3.4a}$$

and

$$f_0 \;=\; \sum_{s \in \mathcal{S}} f_{0s}\,, \qquad \text{where } f_{0s} \in {}^*\mathbb{F}_s^0 \;\text{ for } s \in \mathcal{S}_0\,. \tag{3.4b}$$

Furthermore, since the functions in $\mathbb{F}_0$ depend only on the variables in the subvector $\mathbf{u}_1$, we have that $\Lambda(f + f_0) = \Lambda(f)$ for all $f \in \mathbb{F}$ and $f_0 \in \mathbb{F}_0$. This fact motivates our choice of notation presented in this paragraph.

Recall from the previous section that $\mathbb{G}_l$ is a space of spline functions defined relative to $\triangle_l$, for $1 \le l \le M$, and for each $s \in \mathcal{S}_1$ define the spaces $\mathbb{G}_s$ as usual. By analogy with the definition in (3.3a), set

$$\mathbb{G}_1 \;=\; \left\{ \sum_{s \in \mathcal{S}_1} g_s : g_s \in \mathbb{G}_s \;\text{ for } s \in \mathcal{S}_1 \right\}, \tag{3.5a}$$

and

$$\mathbb{G}_0 \;=\; \left\{ \sum_{s \in \mathcal{S}_0} g_s : g_s \in \mathbb{G}_s \;\text{ for } s \in \mathcal{S}_0 \right\}. \tag{3.5b}$$

Now, for each $s \in \mathcal{S}_1$, let ${}^*\mathbb{G}_s^0$ denote the space of functions in $\mathbb{G}_s$ that are orthogonal to $\mathbb{G}_r$ relative to the empirical inner-product defined in (3.2b), where r is a proper subset of s. Then, set

$$ {}^*\mathbb{G} \;=\; \left\{ \sum_{s \in \mathcal{S}} g_s \,:\, g_s \in {}^*\mathbb{G}_s^0 \;\text{ for } s \in \mathcal{S} \right\}. \tag{3.5c} $$

By construction, if $\mathbb{G}_1$ is identifiable, then $\mathbb{G}_1$ is the direct sum of the spaces ${}^*\mathbb{G}$ and $\mathbb{G}_0$. Therefore, under Condition 2′, except on a set whose probability tends to zero with n, each function $g_1 \in \mathbb{G}_1$ can be represented uniquely as the sum of the functions $g \in {}^*\mathbb{G}$ and $g_0 \in \mathbb{G}_0$, which can in turn be written uniquely as

$$ g \;=\; \sum_{s \in \mathcal{S}} g_s \,, \qquad \text{where } g_s \in {}^*\mathbb{G}_s^0 \;\text{ for } s \in \mathcal{S}\,, \tag{3.6a} $$

and

$$ g_0 \;=\; \sum_{s \in \mathcal{S}_0} g_{0s} \,, \qquad \text{where } g_{0s} \in {}^*\mathbb{G}_s^0 \;\text{ for } s \in \mathcal{S}_0\,. \tag{3.6b} $$

Furthermore, since the functions in $\mathbb{G}_0$ depend only on the variables in the subvector $\mathbf{u}_1$, we have that $\Lambda(\, g + g_0 \,) = \Lambda(\, g \,)$ for all $g \in \mathbb{G}$ and $g_0 \in \mathbb{G}_0$.

Recall from construction of the collection $\mathcal{S}$ and the discussion in the proof of Lemma 3.1 of Chapter 1, that for each $s = \{s_1, \ldots, s_k\} \in \mathcal{S}$,

$$ {}^*\mathbb{G}_s \;=\; {}^*\mathbb{G}_{s_1}^0 \otimes \cdots \otimes {}^*\mathbb{G}_{s_k}^0 \tag{3.7} $$

where s contains at least one index corresponding to a variable in the subvector $\mathbf{u}_2$. Consider for the moment a set of the form $s = r_1 \cup r_2$, where r_1 is a colleciton of indices corresponding to variables in the subvector $\mathbf{u}_1$ and r_2 is a nonempty colleciton of indices corresponding to variables in the subvector $\mathbf{u}_2$. Then, for any function $g = g_1(\, \mathbf{u}_1 \,)\, g_2(\, \mathbf{u}_2 \,)$, where $g_1 \in {}^*\mathbb{G}_{r_1}^0$ and $g_2 \in {}^*\mathbb{G}_{r_2}^0$, and any other function $f_0(\, \mathbf{u}_1 \,)$, we find that

$$ {}^*\big\langle\, g, f_0 \,\big\rangle_n \;=\; {}^*\big\langle\, g_1\, g_2, f_0 \,\big\rangle_n \;=\; {}^*\big\langle\, g_1, f_0 \,\big\rangle_n \, {}^*\big\langle\, g_2, 1 \,\big\rangle_n \;=\; 0\,. \tag{3.8} $$

Combining (3.7) and (3.8), we find that

$$ {}^*\big\langle\, g, f_0 \,\big\rangle_n \;=\; 0 \qquad \text{for } g \in {}^*\mathbb{G} \;\text{ and }\; f_0 = f_0(\, \mathbf{u}_1 \,)\,. \tag{3.9a} $$

As an important special case of this fact, we observe that the above relation holds for $g \in {}^*\mathbb{G}$ and $f = c(\,\cdot\,;\, g\,)$. Repeating the arguments leading from (3.7) to (3.9), substituting in the spaces ${}^*\mathbb{F}_s$ for ${}^*\mathbb{G}_s$, we find in addition that

$$ {}^*\big\langle\, f, f_0 \,\big\rangle \;=\; 0 \qquad \text{for } f \in {}^*\mathbb{F} \;\text{ and }\; f_0 = f_0(\, \mathbf{u}_1 \,)\,. \tag{3.9b} $$

The fact that (3.9) holds is one of our primary motivations for choosing the alternate ANOVA decompositions presented in Section 1.5.

Conditional Density Estimation and Unsaturated Spaces

We now investigate the concavity properites of $l(f, \mathbf{u})$ in the context of conditional density estimation. Toward this end, let f_1, f_2 denote any two bounded functions in $\mathbb{F}$ and let g_1, g_2 be any two functions in $\mathbb{G}$. Then, for $\alpha \in (0,1)$ set

$$f_\alpha \;=\; f_1 + \alpha\big(f_2 - f_1\big) \qquad \text{and} \qquad g_\alpha \;=\; g_1 + \alpha\big(g_2 - g_1\big). \tag{3.10}$$

Observe that

$$\begin{aligned}\frac{d^2}{d\alpha^2} E\big[\, l(f_\alpha, \mathbf{U})\,\big] \;&=\; \frac{d^2}{d\alpha^2} E\big[\, f_\alpha(\mathbf{U}) \,-\, c(\mathbf{U}_1 f_\alpha)\,\big] \\ &=\; -E\Big[\, \mathrm{var}\big(\, f_1(\mathbf{U}_\alpha) \;-\; f_2(\mathbf{U}_\alpha) \mid \mathbf{U}_1 \,\big)\Big], \end{aligned} \tag{3.11}$$

where given that $\mathbf{U}_1 = \mathbf{u}_1$, the random vector $\mathbf{U}_\alpha$ has the density $f_{\mathbf{U}_\alpha}$ on $\mathcal{U}_2'$

$$\begin{aligned}\log\big(f_{\mathbf{U}_\alpha}(\mathbf{u}_2)\big) \;&=\; f_1(\mathbf{u}_1, \mathbf{u}_2) + \alpha\big(f_2(\mathbf{u}_1, \mathbf{u}_2) - f_1(\mathbf{u}_1, \mathbf{u}_2)\big) \\ &\qquad - c\Big(f_1(\mathbf{u}_1, \mathbf{u}_2) + \alpha\big(f_2(\mathbf{u}_1, \mathbf{u}_2) - f_1(\mathbf{u}_1, \mathbf{u}_2)\big)\Big). \end{aligned}$$

Therefore, setting $f_0 = f_1 - f_2$, we find that the last expression in (3.10) is strictly less than zero unless f_0 is essentially a function of the variables in $\mathbf{u}_1$. However, since f_1 and f_2 are each in $\mathbb{F}$, if $f_0 = f_0(\mathbf{u}_1) \in \mathbb{F}_0$, then

$$^*\big\|\, f_0 \,\big\|^2 \;=\; {}^*\langle\, f_1 - f_2, f_0 \,\rangle \;=\; {}^*\langle\, f_1, f_0 \,\rangle - {}^*\langle\, f_2, f_0 \,\rangle \;=\; 0$$

by (3.9b). Arguing similarly for g_α, we again find that the second derivative of $\Lambda(g_\alpha)$ with respect to α is strictly less than zero unless the function $g_0 = g_1 - g_2$ depends only on the variables in $\mathbf{u}_1$. However, since g_1 and g_2 are each in $\mathbb{G}$, if $g_0 = g_0(\mathbf{u}_1) \in \mathbb{G}_0$, then g_0 is identically zero by the definition of $\mathbb{G}$. Therefore, $\Lambda(\cdot)$ is strictly concave in the sense outlined in Section 3.1. Replacing $\Lambda(\cdot)$ with $\ell(\cdot)$ and arguing as above, we conclude that

$$\frac{d^2}{d\alpha^2} E_n\big[\, l(g_\alpha, \mathbf{W})\,\big] \;=\; -E_n\Big[\, \mathrm{var}\big(\, g_1(\mathbf{U}_\alpha) \;-\; g_2(\mathbf{U}_\alpha) \mid \mathbf{U}_1 \,\big)\Big],$$

and hence $\ell(\cdot)$ is also concave in the required sense. Furthermore, in Stone (1991), it is shown that there exists an essentially unique function $\phi^* \in \mathbb{F}$ that maximizes $\Lambda(\cdot)$ over $\mathbb{F}$. By a simplification of that argument, it can be shown that there exists a unique function $\phi_n^* \in \mathbb{G}$ maximizing $\Lambda(\cdot)$ over $\mathbb{G}$.

Population Model (Bias)

As in Section 3.2, we find from the definitions in (3.10) and Taylor's theorem with integral remainder that

$$\Lambda(f_2) \;=\; \Lambda(f_1) \;+\; \frac{d}{d\alpha}\Lambda\big(f_\alpha\big)\Big|_{\alpha=0} \;+\; \int_0^1 (1-\alpha)\,\frac{d^2}{d\alpha^2}\Lambda\big(f_\alpha\big)\,d\alpha\,. \quad (3.12a)$$

In particular, applying the above equality for $f_1 = \phi^*$ and f_2 any other bounded function in $\mathbb{F}$, we find that

$$\Lambda(\phi^*) \;-\; \Lambda(f) \;=\; -\int_0^1 (1-\alpha)\,\frac{d^2}{d\alpha^2}\Lambda\big(\phi^* + \alpha(f-\phi^*)\big)\,d\alpha\,, \quad (3.12b)$$

since ϕ^* is the essentially unique function in $\mathbb{F}$ maximizing $\Lambda(\cdot)$ over $\mathbb{F}$. Observe that by the strict concavity condition (1.13) established in the previous subsection, the quantity appearing on righthand side of equation (3.12b) is strictly positive. In fact, by setting $\sigma(f) = f - c(\cdot, f)$, we find that Lemma 2.1 holds in this context, wiht the usual norm $\|\cdot\|$ replaced by ${}^*\|\cdot\|$. The proof is virtually identical to that given in the context of ordinary density estimation and is not repeated here. In a similar way, the arguments presented in the proof of Lemma 2.2 establish that result in the context of conditional density estimation as well. Lemma 2.3, however, requires special treatment.

Lemma 2.3 *Suppose that Conditions* 1, 2′, 3″, *and* 5′ *hold. Then,*

$$^*\big\|\,\phi^*_{ns} \;-\; \phi^*_s\,\big\|^2 \;=\; O_P\big(\rho^2(\overline{h},\mathcal{S}) \;+\; \omega(\underline{h},\mathcal{S})/n\big)\,, \qquad s \in \mathcal{S}\,.$$

Proof Throughout this proof, let $T_1, T_2, \ldots$ denote suitable positive constants. Assume that the space $\mathbb{G}$ is identifiable, and let $\tilde{g}_n$ denote the orthogonal projection of ϕ^* onto $\mathbb{G}_1$ relative to the empirical inner product defined in (3.2b). Then, we can express $\tilde{g}_n$ uniquely as

$$\tilde{g}_n \;=\; \sum_{s\in\mathcal{S}_1} \tilde{g}_{ns}\,, \qquad \text{where } \tilde{g}_{ns} \in \mathbb{G}^0_s \text{ for } s \in \mathcal{S}_1\,.$$

It follows from Theorem 5.1 in Chapter 1 that

$$^*\big\|\,\tilde{g}_n \;-\; \phi^*\,\big\|^2 \;=\; O_P\big(\rho^2(\overline{h},\mathcal{S}) \;+\; \omega(\underline{h},\mathcal{S})/n\big)\,, \quad (3.13a)$$

and that for each set $s \in \mathcal{S}$,

$$^*\big\|\,\tilde{g}_{ns} \;-\; \phi^*_s\,\big\|^2 \;=\; O_P\big(\rho^2(\overline{h},\mathcal{S}) \;+\; \omega(\underline{h},\mathcal{S})/n\big)\,. \quad (3.13b)$$

Arguing as at the end of Lemma 2.3 in Section 3.2, we find from the continuity of $c(\cdot)$ and Condition 3″ that

$$^*\big\|\,\sigma(\tilde{g}_n) \;-\; \sigma(\phi^*)\,\big\|^2 \;=\; O_P\big(\rho^2(\overline{h},\mathcal{S}) \;+\; \omega(\underline{h},\mathcal{S})/n\big)\,. \quad (3.14)$$

Then, by the triangle and Schwartz inequalities,

$$^*\| \sigma(\tilde{g}_n) - \sigma(\phi_n^*) \|^2 \;\le\; 2\,^*\| \sigma(\tilde{g}_n) - \sigma(\phi^*) \|^2 \;+\; 2\,^*\| \sigma(\phi_n^*) - \sigma(\phi^*) \|^2 ,$$

and hence from (3.14) and the first expression in Lemma 2.2,

$$^*\| \sigma(\tilde{g}_n) - \sigma(\phi_n^*) \|^2 \;=\; O_P\big(\rho^2(\overline{h}, \mathcal{S}) \;+\; \omega(\underline{h}, \mathcal{S})/n \big) . \tag{3.15}$$

For each $s \in \mathcal{S}_1$, we temporarily define the spaces $^*\mathbb{G}_s^1$ to be those functions in $\mathbb{G}_s$ that are orthogonal to all the functions in $\mathbb{G}_r$ relative to the theoretical inner-product in (3.2a), where r is a proper subset of s. Set

$$^*\mathbb{G}^1 \;=\; \left\{ \sum_{s \in \mathcal{S}} g_s : g_s \in {}^*\mathbb{G}_s^1 \;\text{ for } s \in \mathcal{S} \right\},$$

and observe that $\mathbb{G}_1$ is a direct sum of the spaces $\mathbb{G}_0$ and $^*\mathbb{G}^1$. Furthermore, by construction, we have that $\mathbb{G}_0$ and $^*\mathbb{G}^1$ are orthogonal with respect to the innner-product in (3.2a). Now, let $P_{*_{\mathbb{G}^1}}$ and $P_{\mathbb{G}_0}$ denote the projection operators onto $^*\mathbb{G}^1$ and $\mathbb{G}_0$, respecitvely. Then, by orthogonality,

$$P_{*_{\mathbb{G}^1}} \tilde{g}_n \;=\; \tilde{g}_n \;-\; P_{\mathbb{G}_0} g . \tag{3.16}$$

Therefore, setting $\tilde{f}_n = P_{*_{\mathbb{G}^1}} \tilde{g}_n$, we find that $\tilde{f}_n - \tilde{g}_n \in \mathbb{G}_0$, and that

$$\tilde{g}_n \;-\; c(\,\cdot\,; \tilde{g}_n) \;=\; \tilde{f}_n \;-\; c(\,\cdot\,; \tilde{f}_n) \tag{3.17a}$$

Similarly, by setting $\tilde{\phi}_n = P_{*_{\mathbb{G}^1}} \phi_n^*$, we find that $\tilde{\phi}_n - \phi_n^* \in \mathbb{G}_0$, and that

$$\phi_n^* \;-\; c(\,\cdot\,; \phi_n^*) \;=\; \tilde{\phi}_n \;-\; c(\,\cdot\,; \tilde{\phi}_n) \tag{3.17b}$$

Arguing as in the proof of (3.9), we find that

$$\begin{aligned} ^*\| \sigma(\tilde{g}_n) - \sigma(\phi_n^*) \|^2 \;&=\; ^*\| \sigma(\tilde{f}_n) - \sigma(\tilde{\phi}_n) \|^2 \\ &=\; ^*\| \tilde{f}_n - \tilde{\phi}_n \|^2 \;+\; ^*\| c(\,\cdot\,; \tilde{f}_n) - c(\,\cdot\,; \tilde{\phi}_n) \|^2 \end{aligned}$$

and hence from (3.15),

$$^*\| \tilde{f}_n - \tilde{\phi}_n \|^2 \;=\; O_P\big(\rho^2(\overline{h}, \mathcal{S}) \;+\; \omega(\underline{h}, \mathcal{S})/n \big) . \tag{3.18}$$

Because $\tilde{f}_n, \tilde{\phi}_n \in \mathbb{G}_1$, we find from the analog of (3.12) in Chapter 1 established for the alternate ANOVA decompositions in Section 1.5, that

$$^*\| \tilde{f}_n - \tilde{\phi}_n \|_n^2 \;=\; O_P\big(\rho^2(\overline{h}, \mathcal{S}) \;+\; \omega(\underline{h}, \mathcal{S})/n \big) . \tag{3.19}$$

However, since

$$
\begin{aligned}
{}^*\| \tilde{f}_n - \tilde{\phi}_n \|_n^2 &= {}^*\| \tilde{g}_n - \phi_n^* + (\tilde{f}_n - \tilde{g}_n + \phi_n^* - \tilde{\phi}_n) \|_n^2 \\
&= {}^*\| \tilde{g}_n - \phi_n^* \|_n^2 + {}^*\| \tilde{f}_n - \tilde{g}_n + \phi_n^* - \tilde{\phi}_n \|_n^2 ,
\end{aligned}
$$

the relation in (3.19) implies that

$$
{}^*\| \tilde{g}_n - \phi_n^* \|_n^2 = O_P(\rho^2(\overline{h}, \mathcal{S}) + \omega(\underline{h}, \mathcal{S}) / n) .
$$

Combining this last relation with the analog of Lemma 3.7 in Chapter 1 established for the alternate ANOVA decompositions in Section 1.5, we conclude that

$$
{}^*\| \tilde{g}_{ns} - \phi_{ns}^* \|_n^2 = O_P(\rho^2(\overline{h}, \mathcal{S}) + \omega(\underline{h}, \mathcal{S}) / n) , \qquad s \in \mathcal{S} . \tag{3.20}
$$

The desired relation now follows from (3.20), another application of analog of (3.12) in Chapter 1 mentioned above, and (3.13b). □

Sample-Based Estimates (Variance)

Observe that the proof of Lemma 2.? in Chapter 1 carries over to establish this result for the alternate ANOVA decompositions in Section 1.5. Now, arguing as in the proof of Lemma 5 in Stone (1990), we find from this extention of Lemma 2.? in Chapter 1 and Bernstein's inequality that Lemma 2.4 holds in the context of conditional density estimation. In addition, the proof of Lemma 2.5 in this context can be greatly simplified. The reader is referred to the proof of Lemma 8 in Stone (1991). In addition, Lemmas 2.7–2.9 follow exactly as they did in Section 3.2, and so the proofs are not repeated here. This leaves Lemma 2.6. To establish this result in this context requires some extra effort.

Toward this end, define the spaces $\mathbb{G}_s^1$, $s \in \mathcal{S}_1$ as in the proof of Lemma 2.3 in the present context given above. Then, for $g \in \mathbb{G}$, let $\tilde{g} = P_{*\mathbb{G}^1} \tilde{g}$, so that from (3.17),

$$
g - c(\,\cdot\,; g) = \tilde{g} - c(\,\cdot\,; \tilde{g}) \qquad \text{and} \qquad \phi_n^* - c(\,\cdot\,; \phi_n^*) = \tilde{\phi}_n - c(\,\cdot\,; \tilde{\phi}_n) .
$$

Therefore, from (3.18), we find that

$$
{}^*\| \tilde{g} - \tilde{\phi}_n \|^2 \le {}^*\| \sigma(g) - \sigma(\phi_n^*) \|^2 ,
$$

and hence the set

$$
\mathcal{G} = \left\{ g \in \mathbb{G} : \| \sigma(g) - \sigma(\phi_n^*) \| \le t_1 n^{-\tau} \right\}
$$

is certainly contained in the set

$$\widetilde{\mathcal{G}} \;=\; \Big\{\, \tilde{g} \in \mathbb{G}_1 \;:\; \big\|\, \tilde{g} - \tilde{\phi}_n \,\big\| \;\leq\; t_1\, n^{-\tau} \,\Big\}.$$

Arguing as in the proof of Lemma 2.6 in Section 3.2, we find that this lemma is also true in the context of conditional density estimation.

Rate of Convergence

Recall from the previous section, that the functions $g_{l1}, \ldots, g_{lN_l}$ form a basis for $\mathbb{G}_l$, $1 \leq l \leq M$, and that the space $\mathbb{G}_s$, $s = \{s_1, \ldots, s_k\} \in \mathcal{S}_1$, is spanned by functions of the form

$$\prod_{l \in s} g_{lk_l}(\, u_l \,) \qquad \text{where } \; 1 \leq k_l \leq N_l \;\text{ for }\; l \in s\,. \tag{3.21}$$

Let N_s denote the dimension $\mathbb{G}_s$, which is simply the product of N_l for $l \in s$, and given some ordering of the functions in (3.21), let $g_{s\,1}, \ldots, g_{s\,N_s}$ be a basis of $\mathbb{G}_s$. Then, any function $g_s \in \mathbb{G}_s$ can be expressed uniquely as

$$g_s \;=\; \sum_{k=1}^{N_s} \beta_{sk}\, g_{sk}\,. \tag{2.47a}$$

Let $\beta_s = (\beta_{s\,1}, \ldots, \beta_{s\,N_s})$ represent the N_s-dimensional vector of coefficients in this expansion. Now, given some ordering of the sets $s \in \mathcal{S}$, let $\boldsymbol{\beta}$ denote the column vector $\boldsymbol{\beta} = (\,\beta_s\,)$ and let $N = \sum_{s \in \mathcal{S}} N_s$ denote the length of $\boldsymbol{\beta}$. Using this construction, given any function $g \in \mathbb{G}$, there exists a coefficient vector $\boldsymbol{\beta}$ such that

$$g \;=\; \sum_{s \in \mathcal{S}} \sum_{k=1}^{N_s} \beta_{sk}\, g_{sk}\,. \tag{3.22}$$

Recall the definitions of $g = g(\,\cdot\,;\boldsymbol{\beta})$. and $\ell(\,\boldsymbol{\beta}\,)$ given in Section 3.2. Let $\boldsymbol{\beta}^*$ be the coefficient vector corresponding to the function ϕ_n^* such that $\phi_n^* = \sum_s \phi_{ns}^*$, where

$$\phi_{ns}^* \;=\; \sum_{k=1}^{N_s} \beta_{sk}^*\, g_{sk} \;\;\in\;\; {}^*\mathbb{G}_s^0\,, \qquad s \in \mathcal{S}\,. \tag{3.23a}$$

Similarly, let $\widehat{\boldsymbol{\beta}}$ be the coefficient vector corresponding to the maximum likelihood estimate $\hat{\phi}$ such that $\hat{\phi} = \sum_s \hat{\phi}_s$, where

$$\hat{\phi}_s \;=\; \sum_{k=1}^{N_s} \hat{\beta}_{sk}\, g_{sk} \;\;\in\;\; {}^*\mathbb{G}_s^0\,, \qquad s \in \mathcal{S}\,. \tag{3.23b}$$

Observe that Lemma 2.10 holds in this new context because $\|\cdot\|$ and ${}^*\|\cdot\|$

are equivalent norms. Furthermore, since the analog of Lemma 3.3 in Chapter 1 applies to the alternate ANOVA decompostions described in Section 1.5, Lemma 2.11 holds in this context as well. Therefore, in order to establish the rate of convergence for $\hat{\phi}$ and its componentes to ϕ^* and its components, all we have to do is verify properties (2.56) in Section 3.2; that is

$$\left| \mathbf{S}(\boldsymbol{\beta}^*) \right|^2 \;=\; O_P\big(1/n\big)\,, \tag{3.24a}$$

and there exists a positive constant M_9 such that

$$\big(\widehat{\boldsymbol{\beta}} \;-\; \boldsymbol{\beta}^*\big)^t\, \mathbf{D}\, \big(\widehat{\boldsymbol{\beta}} \;-\; \boldsymbol{\beta}^*\big) \;\leq\; -M_8{}^* \big\| \hat{\phi} \;-\; \phi_n^* \big\|^2\,, \tag{3.24b}$$

except on a set with probability tending to zero with n. Toward this end, let $g = g(\,\cdot\,;\boldsymbol{\beta}\,) \in \mathbb{G}$ and recall that in this context,

$$l(g,\mathbf{w}) \;=\; l(g,\mathbf{u}) \;=\; g(\,\mathbf{u}_1,\mathbf{u}_2\,;\boldsymbol{\beta}\,) - c\big(\,\mathbf{u}_1, g(\,\cdot\,;\boldsymbol{\beta}\,)\big)\,,$$

where $\mathbf{u} = (\mathbf{u}_1,\mathbf{u}_2)$, and hence

$$\frac{\partial}{\partial \beta_{sk}}\, l\big(\,g(\mathbf{u};\boldsymbol{\beta}),\mathbf{u}\,\big)$$

$$\qquad =\quad g_{sk}(\,\mathbf{u}_1,\mathbf{u}_2\,) \;-\; \int_{\mathcal{U}_2'} g_{sk}(\,\mathbf{u}_1,\mathbf{u}_2\,) \exp\Big(\,\sigma\big(\,\phi_n^*(\,\mathbf{u}_1,\mathbf{u}_2\,)\,\big)\Big)\, d\mathbf{u}_2\,.$$

Taking expectations we find that

$$E\bigg[\, g_{sk}(\,\mathbf{U}_1,\mathbf{U}_2\,) \;-\; \int_{\mathcal{U}_2'} g_{sk}(\,\mathbf{U}_1,\mathbf{u}_2\,)\, \exp\Big(\,\sigma\big(\,\phi_n^*(\mathbf{U}_1,\mathbf{u}_2)\,\big)\Big)\, d\mathbf{u}_2 \bigg] \;=\; 0\,.$$

Therefore, as seen in Example 4 in Section 3.2, from Condition 6(a) and (c), and the boundedness of ϕ_n^*,

$$E\Big[\, \left| \mathbf{S}(\boldsymbol{\beta}^*) \right|^2 \Big] \;=\; (\,1/n\,) \sum_{s\,\in\,\mathcal{S}} \sum_{k=1}^{N_s} \mathrm{var}\,\big(\,g_{sk}(\mathbf{U})\,\big) \;=\; O(\,1/n\,)\,.$$

Next, set $\boldsymbol{\beta}_\alpha = \boldsymbol{\beta}^* + \alpha(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ so that

$$-(\,\widehat{\boldsymbol{\beta}} \;-\; \boldsymbol{\beta}^*\,)^t\, \mathbf{H}(\,\boldsymbol{\beta}_\alpha\,)\,(\,\widehat{\boldsymbol{\beta}} \;-\; \boldsymbol{\beta}^*\,) \;=\; E_n\Big[\, \mathrm{var}\,\big(\,\phi^*(\mathbf{U}_\alpha) - \hat{\phi}(\mathbf{U}_\alpha)\,|\,\mathbf{U}_1\,\big)\,\Big]\,, \tag{3.25}$$

where given that $\mathbf{U}_1 = \mathbf{u}_1$, the random vector $\mathbf{U}_\alpha$ has density $f_{\mathbf{U}_\alpha}$ on $\mathcal{U}_2'$,

$$\log\big(\,f_{\mathbf{U}_\alpha}(\,\mathbf{u}_2\,)\big) \;=\; \tilde{\phi}(\,\mathbf{u}_1,\mathbf{u}_2\,) + \alpha\big(\,\phi_n^*(\,\mathbf{u}_1,\mathbf{u}_2\,) - \tilde{\phi}(\,\mathbf{u}_1,\mathbf{u}_2)\,\big)$$

$$\qquad\qquad - c\Big(\,\tilde{\phi}(\,\mathbf{u}_1,\mathbf{u}_2\,) + \alpha\big(\,\phi_n^*(\,\mathbf{u}_1,\mathbf{u}_2\,) - \tilde{\phi}(\,\mathbf{u}_1,\mathbf{u}_2\,)\,\big)\,\Big)\,.$$

As in Example 4, we find that there exists a positive constant T_1 such that

$$\lim_{n\to\infty}\; P\big(\,\|\,\phi_n^*\,\|_\infty \leq T_1 \text{ and } \|\,\hat{\phi}\,\|_\infty \leq T_1\,\big) \;=\; 1\,.$$

Therefore, there exists a positive constant T_2 such that except on a set whose probability tends to zero with n, the density of $\mathbf{U}_\alpha$ satisfies

$$\frac{1}{T_2} \;\leq\; f_{\mathbf{U}_\alpha}(\,\mathbf{u}_2\,) \;\leq\; T_2 \qquad \text{for } \mathbf{u}_2 \in \mathcal{U}_2' \,.$$

From this last expression and (3.25), we find that there exists a postitive constant T_3 such that, except on an event whose probability tends to zero with n,

$$-(\,\widehat{\boldsymbol{\beta}} \;-\; \boldsymbol{\beta}^*\,)^t\,\mathbf{D}\,(\,\widehat{\boldsymbol{\beta}} \;-\; \boldsymbol{\beta}^*\,) \;\geq\; T_3\,E_n\Big[\,\int_{\mathcal{U}_2'} \big(\,g(\,\mathbf{U}_1,\mathbf{u}_2\,) \;-\; g_0(\,\mathbf{U}_1\,)\,\big)^2\,d\mathbf{u}_2\Big]\,, \tag{3.26}$$

where $g = \hat{\phi} - \phi_n^*$ and

$$g_0 \;=\; \int_{\mathcal{U}_2'} g(\,\cdot\,,\mathbf{u}_2)\,d\mathbf{u}_2 \;\in\; \mathbb{G}_0\,. \tag{3.27}$$

Here, the fact that $g_0 \in \mathbb{G}_0$ follows from the definition of $\mathcal{S}$ and the tensor product structure of the spaces $\mathbb{G}_s$, $s \in \mathcal{S}$. From Lemma 3.4 in Chapter 1 we find that for $T_4 > 0$, the expression in (3.26) is bounded from below by

$$T_3\,(\,1 - T_4\,)E\Big[\,\int_{\mathcal{U}_2'} \big(\,g(\,\mathbf{U}_1,\mathbf{u}_2\,) \;-\; g_0(\,\mathbf{U}_1\,)\,\big)^2\,d\mathbf{u}_2\Big]\,.$$

Therefore, from Condition $2'$, we find that there exists a postitive constant T_5 such that except on a set whose probability tends to zero with n,

$$-(\,\widehat{\boldsymbol{\beta}} \;-\; \boldsymbol{\beta}^*\,)^t\,\mathbf{D}\,(\,\widehat{\boldsymbol{\beta}} \;-\; \boldsymbol{\beta}^*\,) \;\geq\; T_3\,(\,1 - T_4\,)\,T_5\;{}^*\!\left\|\,g \;-\; g_0\,\right\|^2. \tag{3.28}$$

Now, from the analog of Lemma 3.4 in Chapter 1 established for the alternate ANOVA decomposition, given any positive constant T_6, except on a set whose probability tends to zero with n,

$$\left|\,{}^*\langle\, g, g_0\,\rangle_n \;-\; {}^*\langle\, g, g_0\,\rangle\,\right| \;\leq\; T_6\;{}^*\!\left\|\,g\,\right\|{}^*\!\left\|\,g_0\,\right\|.$$

However, by (3.9a) this becomes

$$\left|\,{}^*\langle\, g, g_0\,\rangle\,\right| \;\leq\; T_6\;{}^*\!\left\|\,g\,\right\|{}^*\!\left\|\,g_0\,\right\|. \tag{3.29}$$

In addition, by the Schwartz inequality we find that

$$\left|\,{}^*\langle\, g - g_0, g_0\,\rangle\,\right| \;\leq\; {}^*\!\left\|\,g - g_0\,\right\|{}^*\!\left\|\,g_0\,\right\|. \tag{3.30}$$

It follows from (3.27) and (3.29) that except on a set whose probability tends to zero with n,

$${}^*\!\left\|\,g_0\,\right\|^2 \;\leq\; \Big(\,T_6\;{}^*\!\left\|\,g\,\right\| \;+\; {}^*\!\left\|\,g - g_0\,\right\|\Big)\,{}^*\!\left\|\,g_0\,\right\|,$$

and so

$$ {}^{*}\| g_0 \| \;\leq\; T_6 \, {}^{*}\| g \| \,+\, {}^{*}\| g - g_0 \| \,. $$

Therefore, by the triangle inequality, except on a set whose probability tends to zero with n,

$$ {}^{*}\| g - g_0 \|^2 \;\geq\; \frac{\left(1 - T_6\right)^2}{4} \, {}^{*}\| g \|^2 \,. \tag{3.31} $$

Finally, from (3.28) and (3.31) we arrive at the relation in (3.24b) since T_4 and T_6 can be chosen arbitrarily small. In turn, we can derive the following theorem.

Theorem 3.1 *Assume that Conditions* 1, 2′, 3″, 5′ *and* 6 *hold. Then, in the context of conditional density estimation,*

$$ \| \sigma(\hat{\phi}) - \sigma(\phi^*) \| \;=\; O_P\big(\rho_s^2(\overline{h}_s) + \omega(\underline{h}, \mathcal{S})/n\big) $$

and for each $s \in \mathcal{S}$,

$$ \| \hat{\phi}_s - \phi_s^* \| \;=\; O_P\big(\rho_s^2(\overline{h}_s) + \omega(\underline{h}, \mathcal{S})/n\big) \,. $$

Appendix to Chapter 3

Lemma 1 *Assume that Conditions* 1 *and* 2′ *hold. Then there exists a positive constant* T_1 *such that*

$$\| g \|_\infty^2 \;\leq\; T_1\, \omega(\underline{h}, \mathcal{S})\, \| g \|^2 , \qquad g \in \mathbb{G} .$$

Proof Let $T_2, T_3, \ldots$ denote suitable positive constants. Fix some $s \in \mathcal{S}$, and let $g_s \in \mathbb{G}_s$. Then, g_s can be written in the form (2.12) in Chapter 1 with respect to the piecewise polynomial basis relative to $\triangle_s$. Recall, however, that each of these basis functions is bounded in absolute value by one. Therefore, the square of the supremum of g_s over $\mathcal{U}_s$ is not larger than $(m+1)^d$ times the sum of the squares of the coefficients in this expansion for g_s. Using Lemma 2.2 in Chapter 1, however, we find that there exists a positive constant T_2 such that this sum of squares is bounded above by $T_2\, \underline{h}_s^{-d_s} \| g_s \|^2$. We have thus established that there exists a positive constant T_3 such that

$$\| g_s \|_\infty^2 \;\leq\; T_3\, \underline{h}_s^{-d_s} \| g_s \|^2 , \qquad g_s \in \mathbb{G}_s ,$$

where we can choose T_3 independent of s. For each $s \in \mathcal{S}$, let $\mathbb{G}_s^1$ denote the space of functions in $\mathbb{G}_s$ that are orthogonal to each of the spaces $\mathbb{G}_r$, $r \subset s$, relative to the theoretical inner-product $\langle \cdot, \cdot \rangle$. Then, each function $g \in \mathbb{G}$ can be written uniquely as the sum

$$g \;=\; \sum_{s \in \mathcal{S}} g_s , \qquad \text{where } g_s \in \mathbb{G}_s^1 \text{ for } s \in \mathcal{S} . \tag{1.1}$$

Furthermore, arguing as in the proof of Lemma 3.7 in Chapter 1, we find that there exists a positive constant T_4 such that

$$\| g \|^2 \;\geq\; T_4 \sum_{s \in \mathcal{S}} \| g_s \|^2 . \tag{1.2}$$

Finally, combining (1.1) and (1.2), we find that

$$\| g \|_\infty^2 \;\leq\; \sum_{s \in \mathcal{S}} \| g_s \|_\infty^2 \;\leq\; T_3 \sum_{s \in \mathcal{S}} \underline{h}_s^{-d_s} \| g_s \|^2 \;\leq\; (T_3 / T_4)\, \#(\mathcal{S})\, \omega(\underline{h}, \mathcal{S})\, \| g \|^2 ,$$

as desired. □

References

1. G. G. Agarwal and W. J. Studden (1980). Asymptotic integrated mean square error using least squares and bias minimizing spline. *Annals of Statistics* **8** 1307-1325.

2. P. Alfeld, B. Piper, and L. Schumaker (1987a). An explicit basis for C^1 quartic bivariate splines. *SIAM Journal of Numerical Analysis* **24** 891-911.

3. P. Alfeld, B. Piper, and L. Schumaker (1987b). Minimally supported bases for spaces of bivariate piecewise polynomials of smoothness r and degree $d \geq 4r+1$. *Computer Aided Geometric Design* **4** 105-123.

4. P. Alfeld, L. Schumaker, and M. Sirvent (1992). On dimension and existence of local bases for multivariate spline spaces. *Journal of Approximation Theory* **70** 243-264.

5. C. de Boor (1976). Splines as linear combinations of B-splines. *Approximation Theory II* (G. G. Lorentz, C. K. Chui and L. L. Schumaker, Eds.) 1-47.

6. C. de Boor (1987). B-Form basics. *Geometric Modeling: Algorithms and New Trends* (G. Farin, Ed.) 131-148.

7. C. de Boor (1989). A local basis for certain smooth bivariate PP spaces. *Multivariate Approximation Theory IV* (C. Chui, W. Schempp, and K. Zeller, Eds.) 25-30.

8. C. de Boor and K. Höllig (1982). B-splines from parallelepipeds. *Journal d'Analyse Mathematique* **42** 99-115.

9. C. de Boor and K. Höllig (1988). Approximation power of smooth bivariate PP functions. *Mathematische Zeitschrift* **197** 343-363.

10. C. de Boor, K. Höllig, and S. Riemenschneider (1993). *Box Splines*. Springer-Verlag, New York.

11. L. Breiman (1991). The Π method for estimating multivariate functions from noisy data–(with discussion). *Technometrics* **33** 125-143.

12. L. Breiman (1993). Fitting additive models to data. *Computational Statistics and Data Analysis* **15** 13-46.

13. A. Buja (1994). The use of polynomial splines and their tensor products in multivariate function estimation–Discussion. *Annals of Statistics* **22** 171-177.

14. K.-W. Chen (1988). Optimal rates of convergence for the piecewise polynomial estimator with the index chosen by the FPE selection rule. *Communications in Statistics* **17** 2887-2906.

15. Z. Chen (1991). Interaction spline models and their convergence rates. *Annals of Statistics* **19** 1855-1868.

16. C. Chui (1988). *Multivariate Splines*. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia.

17. C. Chui (1990). Vertex splines and their applications to interpolation of discrete data. *Computation of Curves and Surfaces* (W. Dahmen and C. Micchelli, Eds.) 137-181.

18. C. Chui and H. Diamond (1990). A characterization of multivariate quasi-interpolation formulas and its applications. *Numerische Mathematik* **57** 105-121.

19. C. Chui and T. He (1990). Bivariate C^1 quadratic finite elements and vertex splines. *Mathematics of Computation* **54** 169-187.

20. C. Chui and M. Lai (1985). On bivariate vertex splines. *Multivariate Approximation Theory III* (W. Schempp and K. Zeller, Eds.) 84-115.

21. C. Chui and M. Lai (1987). On multivariate vertex splines and applications. *Topics in Multivariate Approximation* (C. Chui, L. Schumaker and F. Uteras, Eds.) 19-36.

22. C. Chui and M. Lai (1990). Multivariate vertex splines and finite elements. *Journal of Approximation Theory* **60** 245-343.

23. C. Chui and R. Wang (1984). Spaces of bivariate cubic and quartic splines on type-1 triangulations. *Journal of Mathematical Analysis and its Applications* **101** 540-554.

24. P. Dierckz, S. Van Leemput, and T. Vermeire (1992). Algorithms for surface fitting using Powell-Sabin splines. *IMA Journal of Numerical Analysis* **12** 271-299.

25. P. G. Ciarlet (1978). *The Finite Element Method for Elliptic Problems.* North Holland, Amsterdam.

26. H. B. Curry and I. J. Schoenberg (1966). On Polya frequency functions IV: The fundamental spline functions and their limits. *Journal d'Analyse Mathematique* **17** 71-107.

27. M. Dæhlen and T. Lyche (1991). Box Splines and Applications. *Geometric Modelling* (H. Hagen and D. Roller Eds.) 35-94.

28. T. Dupont and R. Scott (1980). Polynomial approximation of functions in Sobolev spaces. *Mathematics of Computation* **34** 441-463.

29. G. Farin (1986). Triangular Bernstein-Bezier patches. *Computer Aided Geometric Design* **3** 83-127.

30. J. H. Friedman (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics* **19** 1-141.

31. J. H. Friedman and B. Silverman (1989). Flexibile parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3-39.

32. T. Goodman (1987a). Variation diminishing properties of Bernstein polynomials on triangles. *Journal of Approximation Theory* **50** 111-126.

33. T. Goodman (1987b). Further variation diminishing properties of Bernstein polynomials on triangles. *Constructive Approximation* **3** 297-305.

34. C. Gu and G. Wahba (1993). Semiparametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society, Series B–Methodological* **55** 353-368.

35. T. J. Hastie and R. J. Tibshirani (1990). *Generalized Additive Models.* Chapman and Hall, London.

36. W. Hoeffding (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58** 13-30.

37. A. Ibrahim and L. Schumaker (1991). Super spline spaces of smoothness r and degree $d \geq 3r + 2$. *Constructive Approximation* **7** 401-423.

38. H. Johnen and K. Scherer (1977). On the equivalence of the K-functional and the moduli of continuity and some applications. In *Constructive Theory of Functions of Several Variables*, Lecture Notes in Mathematics 571. Springer, New York. 119-140.

39. J.-Y. Koo (1988). Tensor product splines in the estimation of regression functions, exponential response functions, and multivariate densities. Ph.D. Dissertation, Department of Statistics, University of California at Berkeley.

40. C. Kooperberg and C. J. Stone (1991). A study of logspline density estimation. *Computational Statistics and Data Analysis* **12** 327-347.

41. C. Kooperberg and C. J. Stone (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics* **1** 301-328.

42. C. Kooperberg, C. J. Stone and Y. K. Truong (1993a). The L_2 rate of convergence for hazard regression. Technical Report 390, Department of Statistics, University of California at Berkeley.

43. C. Kooperberg, C. J. Stone and Y. K. Truong (1993b). Logspline estimation of a possibly mixed spectral distribution. Technical Report 395, Department of Statistics, University of California at Berkeley.

44. C. Kooperberg, C. J. Stone and Y. K. Truong (1993c). Rate of convergence for logspline spectral density estimation. Technical Report 396, Department of Statistics, University of California at Berkeley.

45. C. Kooperberg, C. J. Stone and Y. K. Truong (1994). Hazard Regression. Technical Report 389, Department of Statistics, University of California at Berkeley.

46. A. P. Korostelev and A. B. Tsybakov (1993). *Minimax Theory of Image Reconstruction*. Springer-Verlag, New York.

47. B. R. Masse and Y. K. Truong (1992). Conditional logspline models. Unpublished manuscript.

48. P. McCullagh and J. A. Nelder (1989). *Generalized Linear Models*. Chapman and Hall, London.

49. M. Mo (1991). Nonparametric estimation by parametric linear regression (I): Global rate of convergence. Unpublished manuscript.

50. W. K. Newey (1991). Consistency and asymptotic normality of nonparametric projection estimators. Unpublished manuscript.

51. J. D. Oden and J. N. Reddy (1976). *An Introduction to the Mathematical Theory of Finite Elements*. Wiley, New York.

52. M. J. D. Powell and M. A. Sabin (1977). Piecewise quadratic approximation on triangles. *ACM TOMS* **3** 316-325.

53. L. Schumaker (1981). *Spline Functions: Basic Theory.* Wiley, New York.

54. L. Schumaker (1989). On super splines and finite elements. *SIAM Journal of Numerical Analysis* **26** 997-1005.

55. H. R. Schwarz (1980). *Finite Element Methods.* Academic Press Limited, London.

56. P. Smith (1982). Curve fitting and modeling with splines using statistical variable selection techniques. Report NASA 166034, NASA, Langley Research Center, Hampton, VA.

57. C. J. Stone (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* **10** 1040-1053.

58. C. J. Stone (1985). Additive regression and other nonparametric models. *Annals of Statistics* **13** 689-705.

59. C. J. Stone (1986). The dimensionality reduction principle for generalized additive models. *Annals of Statistics* **14** 590-606.

60. C. J. Stone (1990). Large-sample inference for log-spline models. *Annals of Statistics* **18** 717-741.

61. C. J. Stone (1991a). Asymptotics for doubly flexible logspline response models. *Annals of Statistics* **19** 1832-1854.

62. C. J. Stone (1991b). Multivariate logspline conditional models. Technical Report 320, Department of Statistics, University of California at Berkeley.

63. C. J. Stone (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Annals of Statistics* **22** 118-184.

64. G. Wahba (1990). *Spline models for observational data.* CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia.