The Annals of Statistics
2003, Vol. 31, No. 5, 1600–1635

LOCAL ASYMPTOTICS FOR POLYNOMIAL SPLINE REGRESSION

BY JIANHUA Z. HUANG

University of Pennsylvania

In this paper we develop a general theory of local asymptotics for least squares estimates over polynomial spline spaces in a regression problem. The polynomial spline spaces we consider include univariate splines, tensor product splines, and bivariate or multivariate splines on triangulations. We establish asymptotic normality of the estimate and study the magnitude of the bias due to spline approximation. The asymptotic normality holds uniformly over the points where the regression function is to be estimated and uniformly over a broad class of design densities, error distributions and regression functions. The bias is controlled by the minimum L_∞ norm of the error when the target regression function is approximated by a function in the polynomial spline space that is used to define the estimate. The control of bias relies on the stability in L_∞ norm of L_2 projections onto polynomial spline spaces. Asymptotic normality of least squares estimates over polynomial or trigonometric polynomial spaces is also treated by the general theory. In addition, a preliminary analysis of additive models is provided.

1. Introduction.

The use of polynomial splines provides an effective approach to modern nonparametric modeling. When fitted by the maximum likelihood method, polynomial splines can be applied to a broad range of statistical problems, including least squares regression, density and conditional density estimation, generalized regression such as logistic and Poisson regression, polychotomous regression and hazard regression. The spline based methods are also very convenient for fitting structural models such as additive models in multivariate function estimation. See Stone, Hansen, Kooperberg and Truong (1997) for a recent review of the subject and related references.

The theoretical investigation of the properties of methods based on polynomial splines has been an active area of research for years. Global rates of convergence of spline estimates have been thoroughly studied for various statistical contexts; see Stone (1985, 1986, 1994), Hansen (1994), Kooperberg, Stone and Truong (1995a, b), Huang (1998a, b), Huang and Stone (1998) and Huang, Kooperberg, Stone and Truong (2000). A systematic treatment of global asymptotics of spline estimates is given in Huang (2001). In contrast, the local properties (behavior at a point) of spline estimates are much less studied. See Stone (1990, 1991) and Zhou, Shen and Wolfe (1998) for some available results. The focus of this paper is on local asymptotics. Local asymptotic results are useful for constructing asymptotic

Received July 1999; revised April 2001.

AMS 2000 subject classifications. Primary 62G07; secondary 62G20.

Key words and phrases. Asymptotic normality, least squares, nonparametric regression, polynomial regression, regression spline.

confidence intervals. They also provide theoretical insights about the properties of estimates that cannot be explained by global asymptotic results.

Usually, polynomial splines are fitted by minimizing a global criterion such as the sum of squared errors or the negative of the log-likelihood. The resulting estimate is a polynomial spline that can be totally characterized by values of the coefficients in a basis expansion. One advantage of this approach is that the estimate is “simpler” than the original data set since the number of coefficients, which equals the dimension of the estimation space, is usually much smaller than the sample size. Unfortunately, along with this advantage there is difficulty in analyzing the local properties. The situation is fundamentally different from another nonparametric approach—local polynomial kernel methods, where a polynomial is fitted to the data in a local neighborhood around a given point and hence local properties of resulting estimates can be conveniently obtained [see, e.g., Fan and Gijbels (1996)]. However, the piecewise polynomial nature of polynomial splines suggests that expecting a reasonably good local behavior of polynomial spline methods is not unrealistic.

In this paper we provide a general theory of local asymptotics in the context of regression. Let X represent a vector of predictor variables and Y a real-valued response variable, where X and Y have a joint distribution. We assume that X ranges over a compact subset $\mathcal{X}$ of some Euclidean space. In addition, we assume that the distribution of X is absolutely continuous and that its density function $p_X(\cdot)$, which we refer to as the design density, is bounded away from zero and infinity on $\mathcal{X}$. Set $\mu(x) = E(Y|X = x)$ and $\sigma^2(x) = \mathrm{var}(Y|X = x)$, and assume that the functions $\mu = \mu(\cdot)$ and $\sigma^2 = \sigma^2(\cdot)$ are bounded on $\mathcal{X}$. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample of size n from the distribution of (X, Y). The primary interest is in estimating μ. We consider least squares estimates over polynomial spline spaces and refer to the corresponding estimation procedures as polynomial spline regression.

In this paper a polynomial spline is referred to broadly as any possibly smooth, piecewise polynomial function. To be specific, let Δ be a partition of $\mathcal{X}$ into disjoint sets. An element of Δ can be an interval, a two-dimensional triangle or rectangle, or a high-dimensional simplex or hyperrectangle. By a polynomial spline on $\mathcal{X}$, we mean a function g on $\mathcal{X}$ such that the restriction of g to each set in Δ is a polynomial and g may satisfy certain smoothness conditions across the boundaries. This setup is very general, containing as special cases univariate splines, tensor product splines, and bivariate or multivariate splines on triangulations. See Hansen, Kooperberg and Sardy (1998) for the practicality of multivariate splines for statistical applications.

We now give a brief description of our results. It is convenient to put polynomial spline regression into a general framework. Let $G = G_n$, referred to as the estimation space, be a linear space of bounded functions with finite dimension N_n. The least squares estimate $\hat{\mu}$ of μ in G is defined as the element $g \in G$ that minimizes $\sum_i [g(X_i) - Y_i]^2$. Polynomial spline regression corresponds to G being a space of polynomial splines. Other common choices of G include polynomials and trigonometric polynomials. Usually the true regression function does not belong to G and members of G are used only as approximations to the truth. Therefore, it is natural to let the dimension of the estimation space grow with the sample size in the asymptotic analysis. In the theory developed in this paper, the dimension of G is allowed to grow with n but not required to do so.

Consider a general linear estimation space G. Set $\tilde{\mu} = E(\hat{\mu}|X_1, \dots, X_n)$. We have the decomposition

$$\hat{\mu}(x) - \mu(x) = [\hat{\mu}(x) - \tilde{\mu}(x)] + [\tilde{\mu}(x) - \mu(x)],$$

where $\hat{\mu} - \tilde{\mu}$ and $\tilde{\mu} - \mu$ are referred to as the variance and bias terms, respectively. We will see that these two terms require very different analyses.

In Section 2 we show that $\hat{\mu}$ and $\tilde{\mu}$ can both be viewed as projections. Precisely, $\hat{\mu} = \Pi_n Y$ and $\tilde{\mu} = \Pi_n \mu$, where Π_n is the orthogonal projection onto the estimation space G relative to the empirical inner product defined in Section 2. This geometric viewpoint is fundamental in our study.

Section 3 establishes the asymptotic normality of the variance term $\hat{\mu}(x) - \tilde{\mu}(x)$ for general linear estimation spaces. Applications to constructing asymptotic confidence intervals are also discussed. The results are generally applicable to any type of estimation space, including polynomials, trigonometric polynomials, and polynomial splines. In Section 4, we strengthen the result in Section 3 by showing that asymptotic normality of $\hat{\mu}(x) - \tilde{\mu}(x)$ holds uniformly over $x \in \mathcal{X}$ and over a broad class of design densities, error distributions, and regression functions. This result can be used to construct an asymptotic confidence interval whose coverage probability converges uniformly to its nominal level. This uniform asymptotic normality result is new in the nonparametric regression literature.

Section 5 evaluates the size of the bias $\tilde{\mu}(x) - \mu(x)$. In contrast to previous sections, we focus in this section on polynomial spline regression since special properties of polynomial splines are crucial in controlling the bias. In Section 5.1, it is shown that the bias is bounded above by a multiple of $\inf_{g \in G} \|g - \mu\|_\infty$, the best approximation rate in L_∞ norm to the regression function by a function in the estimation space. This result relies on the stability in L_∞ norm of L_2 projections onto polynomial spline spaces, a property that is not shared by projections onto spaces of polynomials or trigonometric polynomials. We will see in the Appendix that this property is a consequence of the existence of a locally supported basis of polynomial spline spaces. Section 5.2 gives a sufficient condition for the bias term to be negligible compared with the variance term when the estimation space is a univariate spline space or tensor product spline space and the regression function satisfies a commonly used smoothness condition. Section 5.3 discusses an existing result on asymptotic bias expression and shows what new insights we gain from our general results.

Section 6 gives explicit expressions for conditional variances of least squares estimates in terms of a given basis function. Section 7 provides a preliminary analysis of spline estimation in additive models. The Appendix gives a general result on the stability in L_∞ norm of L_2 projections onto spline spaces, which plays a key role in studying the bias of polynomial spline regression in Section 5. This result also has independent interest.

Zhou, Shen and Wolfe (1998) (henceforth ZSW) studied local asymptotics for univariate spline regression. The geometric approach used in this paper is novel and distinguishes our treatment from the previous work. The present approach allows us to obtain a quite general understanding of the local asymptotics of polynomial spline regression. The general result applies to univariate splines, tensor product splines, and bivariate or multivariate splines on triangulations. In this approach, we see precisely how the special properties of polynomial splines are used in the analysis of the bias term, whereas these properties are not needed in the treatment of the asymptotic normality of the variance term. We obtain substantial additional insights even for univariate spline regression. The setup of ZSW (1998) is restricted: the knots are required to be asymptotically equally spaced, the design density is continuous, and the order of the spline equals the assumed order of derivative of the unknown regression function. All these assumptions are relaxed in this paper. Our condition on the allowed rate of growth of the number of knots for the random design case (i.e., $\lim_n J_n \log n/n = 0$ where J_n is the number of knots) is much weaker than that used in ZSW (i.e., $\lim_n J_n^2/n = 0$). We believe our condition is close to the minimal. Actually this condition is compatible with the similar condition for the local polynomial method (i.e., $\lim_n nh_n = \infty$ where h_n is the bandwidth). We think the $\log n$ term in our condition cannot be dropped because, as a global method, the spline estimator deals with all points in the design space $\mathcal{X}$ at the same time, while the local method treats one point a time.

It is common in the literature to assume the continuity of certain partial derivatives of the unknown function in studying local asymptotics; see, for example, Ruppert and Wand (1994) for results on local polynomial regression. ZSW (1998) followed such a tradition and obtained asymptotic results for a degree $p-1$ spline estimator when the regression function μ has a pth order derivative. Their setup is restricted and rules out the use of quadratic or cubic splines if μ has a continuous second derivative. A general, alternative view of point is taken in this paper. Precisely, the asymptotic bias of a spline estimator is described by the approximation power of the spline space to the unknown regression function, which can be obtained explicitly for any given smoothness condition using results from approximation theory. This general view has the advantage that it allows us to study the asymptotic behavior of a specific degree spline estimator under various smoothness conditions and the behavior of spline estimators with different degrees under the same smoothness conditions.

We believe that the theoretical insights provided by this paper and the techniques developed in this paper are useful for understanding the properties of polynomial spline based estimators in other contexts such as generalized regression, density estimation and hazard regression and for structural models such as additive models. Recently, Huang, Wu and Zhou (2000) extended the techniques in this paper to analyze the asymptotic properties of spline based estimators in the context of longitudinal data analysis.

In what follows, for any function f on $\mathcal{X}$, set $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. Given positive numbers a_n and b_n for $n \ge 1$, let $a_n \lesssim b_n$ and $b_n \gtrsim a_n$ mean that a_n/b_n is bounded and let $a_n \asymp b_n$ mean that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Given random variables W_n for $n \ge 1$, let $W_n = O_P(b_n)$ mean that $\lim_{c\to\infty} \limsup_n P(|W_n| \ge cb_n) = 0$ and let $W_n = o_P(b_n)$ mean that $\lim_n P(|W_n| \ge cb_n) = 0$ for all $c > 0$. When a supremum of an expression of a ratio is taken over some set of arguments, we use the convention that the supremum is always taken with respect to the arguments such that the involved denominator is not zero. For example, $\sup_{g\in G} |\|g\|_\infty/\|g\| - 1|$ should read $\sup_{g\in G, \|g\|\neq 0} |\|g\|_\infty/\|g\| - 1|$.

2. Least squares estimator as a projection. In this section we show that, for a general linear estimation space, the least squares estimate is an orthogonal projection relative to an appropriate inner product. We also give sufficient conditions for the least squares estimate to be well defined.

We start by introducing two inner products on the space of square-integrable functions on $\mathcal{X}$. For any integrable function f defined on $\mathcal{X}$, set $E_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and $E(f) = E[f(X)]$. Define the empirical inner product and norm by $\langle f_1, f_2\rangle_n = E_n(f_1 f_2)$ and $\|f_1\|_n^2 = \langle f_1, f_1\rangle_n$ for square-integrable functions f_1 and f_2 on $\mathcal{X}$. The theoretical versions of these quantities are given by $\langle f_1, f_2\rangle = E(f_1 f_2)$ and $\|f_1\|^2 = \langle f_1, f_1\rangle$.

The estimation space G is said to be theoretically identifiable if $g \in G$ and $\|g\| = 0$ together imply that $g = 0$ everywhere on $\mathcal{X}$. When G is theoretical identifiable, it is a Hilbert space equipped with the theoretical inner product. To rule out pathological choices of the estimation space that are not useful in practice, we require throughout the paper that the estimation space G be theoretically identifiable. This space is said to be empirically identifiable (relative to $X_1, \dots, X_n$) if $g \in G$ and $\|g\|_n = 0$ together imply that $g = 0$ everywhere on $\mathcal{X}$. As we shall see, the empirical identifiability of the estimation space G ensures that the least squares estimate is well defined.

Since X has a density with respect to Lebesgue measure, with probability one, the design points $X_1, \dots, X_n$ are distinct and hence we can find a function defined on $\mathcal{X}$ that interpolates the values $Y_1, \dots, Y_n$ at these points. With a slight abuse of notation, let $Y = Y(\cdot)$ denote any such function. The following result is obviously valid.

LEMMA 2.1. *Given a realization of $X_1, \dots, X_n$, suppose G is empirically identifiable. Then G is a Hilbert space equipped with the empirical inner product. The least squares estimate $\hat{\mu}$ is the orthogonal projection of Y onto G relative to the empirical inner product and is uniquely defined.*

The following lemma, which follows easily from the definitions, tells us that if the theoretical norm is close to the empirical norm uniformly over the estimation space, then theoretical identifiability implies empirical identifiability.

LEMMA 2.2. *Suppose G is theoretically identifiable. If* $\sup_{g\in G} |\|g\|_n/\|g\|-1| = o_P(1)$, *then G is empirically identifiable except on an event whose probability tends to zero as* $n \to \infty$.

We now give sufficient conditions for the theoretical norm to be close to the empirical norm uniformly over estimation spaces. These conditions together with Lemmas 2.1 and 2.2 yield sufficient conditions for the least squares estimate to be well defined. The discussion is presented in a general form for weighted versions of theoretical and empirical norms. The weighted versions of these norms are useful in discussions of heteroscedastic errors; see Remarks 3.2, 6.1 and 6.2 in the following. For a nonnegative weight function w defined on $\mathcal{X}$, let the theoretical and empirical inner products be defined by $\langle f_1, f_2\rangle_w = E(f_1 f_2 w^2)$ and $\langle f_1, f_2\rangle_{n,w} = E_n(f_1 f_2 w^2)$. Denote the corresponding norms by $\|\cdot\|_w$ and $\|\cdot\|_{n,w}$.

Set $A_n = \sup_{g\in G}(\|g\|_\infty/\|g\|)$. Observe that $1 \le A_n < \infty$. This constant can be understood as a measure of irregularity of the estimation space G. It was used in Huang (1998a) in a general discussion of L_2 rate of convergence for least squares estimation and will appear again in our discussion of asymptotic normality. Since $g \in G$ and $\|g\| = 0$ implies that $\|g\|_\infty = 0$, we see that $\|g\|_\infty \le A_n\|g\|$ for all $g \in G$.

Note that A_n depends on the distribution of X. When the density of X is bounded away from zero, A_n can be bounded above by a constant that does not depend on the distribution of X. Specifically, suppose that there is a constant $c > 0$ such that $\inf_x p_X(x) \ge c$. Let $\|\cdot\|_{L_2}$ denote the L_2 norm relative to the uniform distribution on $\mathcal{X}$; that is, $\|f\|_{L_2}^2 = \int_{\mathcal{X}} f^2(x)\,dx/|\mathcal{X}|$ for any square-integrable function f. Set $\overline{A}_n = \sup_{g\in G}\{\|g\|_\infty/\|g\|_{L_2}\}$. Then $A_n \le \sqrt{|\mathcal{X}|/c}\,\overline{A}_n$.

Further discussion about the constant A_n (or $\overline{A}_n$) can be found in Section 2.2 of Huang (1998a). Here we only cite some examples from that paper. Suppose $\mathcal{X}$ is a compact interval. Let Pol(J), TriPol(J), and Spl(J) denote, respectively, the space of polynomials of degree J or less, the space of trigonometric polynomials of degree J or less, and the space of polynomial splines with fixed degree m and J equally spaced knots. Then, when G equals Pol(J_n), TriPol(J_n), or Spl(J_n), we have, respectively, $\overline{A}_n \lesssim J_n$, $\overline{A}_n \lesssim J_n^{1/2}$ or $\overline{A}_n \lesssim J_n^{1/2}$. For the multidimensional case, suppose that $\mathcal{X}$ is the Cartesian product of compact intervals $\mathcal{X}_1, \dots, \mathcal{X}_d$. Let G_l be a linear space of functions on $\mathcal{X}_l$ for $1 \le l \le d$ and let G be the tensor product of these spaces. Then, when G_l equals Pol(J_n), TriPol(J_n) or Spl(J_n) for $1 \le l \le d$, we have, respectively, $\overline{A}_n \lesssim J_n^d$, $\overline{A}_n \lesssim J_n^{d/2}$ or $\overline{A}_n \lesssim J_n^{d/2}$.

The next lemma gives sufficient conditions for the empirical norm to be close to the theoretical norm uniformly over the estimation spaces. Note that

$\sup_{g\in G}|\|g\|_{n,w}/\|g\|_w-1|=o_P(1)$ is equivalent to $\sup_{g\in G}|\|g\|_{n,w}^2/\|g\|_w^2-1|=o_P(1)$.

LEMMA 2.3. *Suppose that* $0<\inf_{x\in\mathcal{X}}w(x)\le\sup_{x\in\mathcal{X}}w(x)<\infty$.

(i) (General case.) *If* $\lim_n A_n^2N_n/n=0$, *then* $\sup_{g\in G}|\|g\|_{n,w}/\|g\|_w-1|=o_P(1)$.

(ii) (Polynomial splines.) *Suppose* p_X *is bounded away from zero and infinity. Suppose* G *is a space of polynomial splines satisfying Condition* A.2 *in the Appendix. If* $\lim_n N_n\log n/n=0$, *then* $\sup_{g\in G}|\|g\|_{n,w}/\|g\|_w-1|=o_P(1)$.

In particular, letting $w\equiv 1$, *either* (i) *or* (ii) *implies that* $\sup_{g\in G}|\|g\|_n/\|g\|-1|=o_P(1)$.

PROOF. The result for case (i) is a direct consequence of Lemma 10 of Huang (1998a). The result for case (ii) follows from Lemma A.1 in the Appendix. □

We end this section by giving a result relating the conditional mean of $\hat\mu$ to an orthogonal projection. Let Π_n denote the empirical orthogonal projection (i.e., the orthogonal projection relative to the empirical inner product) onto G. Then $\hat\mu=\Pi_nY$.

LEMMA 2.4. $E(\hat\mu|X_1,\ldots,X_n)=\Pi_n\mu$.

This lemma follows easily from the properties of the expectation and orthogonal projection operators and details are omitted.

3. Asymptotic normality of the variance term. In this section we establish the asymptotic normality of least squares estimates for general estimation spaces. For notational simplicity, we first present the result for the homoscedastic error case and then discuss extensions to the heteroscedastic error case and the fixed design case. Let $\Phi(\cdot)$ denote the standard normal distribution function.

3.1. *Homoscedastic error case.* Write $Y=\mu(X)+\varepsilon$ with $\varepsilon=Y-\mu(X)$. We say that the errors are homoscedastic if $\sigma^2(x)=E(\varepsilon^2|X=x)$ does not depend on x.

THEOREM 3.1. *Suppose* $\sigma^2(x)=\sigma^2$ *is a constant and that* $\sup_{g\in G}|\|g\|_n/\|g\|-1|=o_P(1)$. *In addition, assume that*

$$\lim_{\lambda\to\infty}E\big(\varepsilon^2\,\mathrm{ind}\{|\varepsilon|>\lambda\}|X=x\big)=0.$$

If $\lim_n A_n^2/n=0$, *then, for* $x\in\mathcal{X}$,

$$P\Big(\hat\mu(x)-\tilde\mu(x)\le t\sqrt{\mathrm{Var}(\hat\mu(x)|X_1,\ldots,X_n)}|X_1,\ldots,X_n\Big)-\Phi(t)$$
$$=o_P(1),\qquad t\in\mathbb{R};$$

consequently,

$$\mathcal{L}\left(\frac{\hat{\mu}(x)-\tilde{\mu}(x)}{\sqrt{\operatorname{Var}(\hat{\mu}(x)|X_1,\dots,X_n)}}\right)\Rightarrow N(0,1), \qquad n\to\infty.$$

The condition $\lim_n A_n^2/n=0$ is straightforward to verify for commonly used estimation spaces. Suppose that, for example, $\mathcal{X}$ is the Cartesian product of compact intervals $\mathcal{X}_1,\dots,\mathcal{X}_d$. Suppose also that the density of X is bounded away from zero. Let G_l be a linear space of functions on $\mathcal{X}_l$ for $1\le l\le d$ and let G be the tensor product of these spaces. Then, when G_l equals $\mathrm{Pol}(J_n)$, $\mathrm{TriPol}(J_n)$, or $\mathrm{Spl}(J_n)$ for $1\le l\le d$, this condition reduces respectively to $\lim_n N_n^2/n=0$, $\lim_n N_n/n=0$, or $\lim_n N_n/n=0$ (see the discussion above Lemma 2.3). In this theorem it is not required that the design density be continuous, which is usually assumed for proving asymptotic normality of kernel or local polynomial regression estimators; compare with Theorem 4.2.1 of Härdle (1990) and Theorem 5.2 of Fan and Gijbels (1996).

Asymptotic distribution results such as Theorem 3.1 can be used to construct asymptotic confidence intervals; see, for example, the general discussion in Section 3.5 of Hart (1997). One sensible approach is to think of $\tilde{\mu}$ as the estimable part of μ and construct an asymptotically valid confidence interval for $\tilde{\mu}$. Note that $\tilde{\mu}=\Pi_n\mu$ can be interpreted as the best approximation in the estimation space G to μ. When σ^2 is known, $\mathrm{SD}(\hat{\mu}(x)|X_1,\dots,X_n)=\sqrt{\operatorname{Var}(\hat{\mu}(x)|X_1,\dots,X_n)}$ depends only on the data. Note that $\mathrm{SD}(\hat{\mu}(x)|X_1,\dots,X_n)$ can be conveniently calculated by using the formula given in Theorem 6.1 of Section 6. Set $\mu_\alpha^l(x)=\hat{\mu}(x)-z_{1-\alpha/2}\mathrm{SD}(\hat{\mu}(x)|X_1,\dots,X_n)$ and $\mu_\alpha^u(x)=\hat{\mu}(x)+z_{1-\alpha/2}\mathrm{SD}(\hat{\mu}(x)|X_1,\dots,X_n)$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution. Suppose the conclusion of Theorem 3.1 holds. Then $[\mu_\alpha^l(x),\mu_\alpha^u(x)]$ is an asymptotic level $1-\alpha$ confidence interval of $\tilde{\mu}(x)$; that is,

$$\lim_n P\bigl(\mu_\alpha^l(x)\le\tilde{\mu}(x)\le\mu_\alpha^u(x)\bigr)=1-\alpha.$$

In fact, the conditional coverage probability is also close to $1-\alpha$; that is,

$$P\bigl(\mu_\alpha^l(x)\le\tilde{\mu}(x)\le\mu_\alpha^u(x)|X_1,\dots,X_n\bigr)=1-\alpha+o_P(1).$$

To construct an asymptotic confidence interval when the error variance σ^2 is not known, we can simply replace σ^2 by any consistent estimate. Such estimates of σ^2 can be found, for example, in Rice (1984), Gasser, Sroka and Jennen-Steinmetz (1986) and Hall, Kay and Titterington (1990).

Another, perhaps more acceptable, approach is to select the estimation space G so that, asymptotically, the bias term $\tilde{\mu}-\mu$ is of negligible magnitude compared with the variance term $\hat{\mu}-\tilde{\mu}$. Then the above confidence interval for $\tilde{\mu}$ is also an asymptotically valid confidence interval for μ. To this end, however, one needs to study the magnitude of $\tilde{\mu}-\mu$, which will be discussed in Section 5.

The next result gives the size of the asymptotic conditional variance.

COROLLARY 3.1. *Under the conditions of Theorem* 1,

$$\sup_x \operatorname{Var}\big(\hat{\mu}(x)|X_1,\dots,X_n\big) = \frac{A_n^2}{n}\sigma^2\big(1+o_P(1)\big).$$

Consequently, $\hat{\mu}(x) - \tilde{\mu}(x) = O_P(A_n/\sqrt{n})$ *uniformly in* $x \in \mathcal{X}$; *that is,*

$$\lim_{C\to\infty} \limsup_{n\to\infty} \sup_{x\in\mathcal{X}} P\left(|\hat{\mu}(x) - \tilde{\mu}(x)| \geq C\sqrt{\frac{A_n^2}{n}}\right) = 0.$$

In light of Corollary 3.1, Theorem 3.1 can be interpreted as follows: If the supremum over $\mathcal{X}$ of the conditional variance of $\hat{\mu}(x)$ given $X_1,\dots,X_n$ converges to zero, then the asymptotic normality holds for all $x \in \mathcal{X}$. Another interesting consequence of Corollary 3.1 is that, provided the estimation spaces have the same dimensions, in the worst situation the local conditional variance of the least square estimate for polynomial regression is much larger than its counterpart for polynomial spline regression. To be specific, suppose $\mathcal{X}$ is a compact interval and that the density of X is bounded away from zero and infinity. If G is a space of polynomial splines of a fixed degree m and having $J_n = N_n - m - 1$ interior knots with bounded mesh ratio [see (5.3)], then

$$V_1 := \sup_x \operatorname{Var}\big(\hat{\mu}(x)|X_1,\dots,X_n\big) \asymp \frac{N_n}{n}\big(1+o_P(1)\big) \tag{3.1}$$

and, if G is the space of polynomials of degree $J_n = N_n - 1$ or less, then

$$V_2 := \sup_x \operatorname{Var}\big(\hat{\mu}(x)|X_1,\dots,X_n\big) \asymp \frac{N_n^2}{n}\big(1+o_P(1)\big). \tag{3.2}$$

In contrast, we have that for both choices of G,

$$V_3 := \int_{\mathcal{X}} \operatorname{Var}\big(\hat{\mu}(x)|X_1,\dots,X_n\big)\,dx = O_P\left(\frac{N_n}{n}\right). \tag{3.3}$$

Proofs of these results are given in Section 3.3.

3.2. *Extensions to heteroscedastic case and fixed design.*

REMARK 3.1. When the errors are heteroscedastic [i.e., $\sigma^2(x) = \operatorname{Var}(Y|X=x)$ is not a constant], Theorem 1 still holds (with the same proof) if the function $\sigma(\cdot)$ is bounded away from zero and infinity. Moreover,

$$\sup_x \operatorname{Var}\big(\hat{\mu}(x)|X_1,\dots,X_n\big) \leq \frac{A_n^2}{n}\sup_x \sigma^2(x)\big(1+o_P(1)\big),$$

and as a consequence, $\hat{\mu}(x) - \tilde{\mu}(x) = O_P(A_n/\sqrt{n})$ uniformly in $x \in \mathcal{X}$.

REMARK 3.2. In the case of heteroscedastic errors, if $\sigma^2(x)$ is known, we can also consider the weighted least squares estimate $\hat{\mu}^w$, which is defined as the minimizer in G of $\sum_i[(Y_i - g(X_i))^2/\sigma^2(X_i)]$. Redefine the theoretical and empirical inner products by $\langle f_1, f_2\rangle_{1/\sigma} = E[f_1(X)f_2(X)/\sigma^2(X)]$ and $\langle f_1, f_2\rangle_{n,1/\sigma} = E_n[f_1(X)f_2(X)/\sigma^2(X)]$. The corresponding norms are denoted by $\|\cdot\|_{1/\sigma}$ and $\|\cdot\|_{n,1/\sigma}$. Let Π_n^w denote the orthogonal projection onto G relative to the above modified empirical inner product. Observe that $\hat{\mu}^w = \Pi_n^w Y$. Set $\tilde{\mu}^w = E(\hat{\mu}^w|X_1, \dots, X_n)$. Then $\tilde{\mu}^w = \Pi_n^w \mu$. Suppose $\sigma(\cdot)$ is bounded away from zero and infinity and that $\sup_{g\in G}|\|g\|_{n,1/\sigma}/\|g\|_{1/\sigma} - 1| = o_P(1)$. The same argument as in the proof of Theorem 1 gives that if $\lim_n A_n^2/n = 0$, then

$$P\Big(\hat{\mu}^w(x) - \tilde{\mu}^w(x) \le t\sqrt{\operatorname{Var}\big(\hat{\mu}^w(x)|X_1, \dots, X_n\big)}\,|X_1, \dots, X_n\Big) - \Phi(t)$$
$$= o_P(1), \qquad t \in \mathbb{R},$$

and

$$\mathcal{L}\bigg(\frac{\hat{\mu}^w(x) - \tilde{\mu}^w(x)}{\sqrt{\operatorname{Var}(\hat{\mu}^w(x)|X_1, \dots, X_n)}}\bigg) \Rightarrow N(0, 1), \qquad n \to \infty.$$

Moreover,

$$\sup_x \operatorname{Var}\big(\hat{\mu}^w(x)|X_1, \dots, X_n\big) = \frac{1}{n}\sup_{g\in G}\frac{\|g\|_\infty^2}{\|g\|_{n,1/\sigma}^2} \le \frac{A_n^2}{n}\sup_x \sigma^2(x)\big(1 + o_P(1)\big),$$

and consequently, $\hat{\mu}^w(x) - \tilde{\mu}^w(x) = O_P(A_n/\sqrt{n})$ uniformly in $x \in \mathcal{X}$.

REMARK 3.3. The discussion for the random design case carries over to the fixed design case. We need only replace expectations conditional on $X_1, \dots, X_n$ by unconditional expectations. The definitions of empirical inner product, empirical norm, and empirical projection carry over to the fixed design case in an obvious manner. Let $Y_i = \mu(x_i) + \varepsilon_{i,n}$, $i = 1, \dots, n$, where $x_1, \dots, x_n$ are fixed design points in $\mathcal{X}$ and $\varepsilon_{1,n}, \dots, \varepsilon_{n,n}$ are independent errors with mean 0 and variances $\sigma_1^2, \dots, \sigma_n^2$. Suppose there are constants C_1 and C_2 with $0 < C_1 \le C_2 < \infty$ such that $C_1 \le \sigma_i^2 \le C_2$ for $i = 1, \dots, n$. Moreover, assume that

$$\lim_{\lambda\to\infty}\sup_n \sup_{1\le i\le n} E\big(\varepsilon_{i,n}^2 \operatorname{ind}\{|\varepsilon_{i,n}| > \lambda\}\big) = 0.$$

Let $\hat{\mu}$ and $\hat{\mu}^w$ be the ordinary least squares estimate and the weighted least squares estimate defined above. Set $\widetilde{A}_n = \sup_{g\in G}(\|g\|_\infty/\|g\|_n)$. If $\lim_n \widetilde{A}_n^2/n = 0$, then

$$\mathcal{L}\bigg(\frac{\hat{\mu}(x) - E[\hat{\mu}(x)]}{\sqrt{\operatorname{Var}(\hat{\mu}(x))}}\bigg) \Rightarrow N(0, 1), \qquad n \to \infty,$$

and

$$\mathcal{L}\bigg(\frac{\hat{\mu}^w(x) - E[\hat{\mu}^w(x)]}{\sqrt{\operatorname{Var}(\hat{\mu}^w(x))}}\bigg) \Rightarrow N(0, 1), \qquad n \to \infty.$$

The conditions on the design points are implicit in the condition that $\lim_n \widetilde{A}_n^2/n = 0$.

3.3. *Proofs.*

PROOF OF THEOREM 3.1. Let $\{\phi_j, 1 \le j \le N_n\}$ be an orthonormal basis of G relative to the empirical inner product. Since

$$\hat{\mu}(x) = \sum_j \langle Y, \phi_j \rangle_n \phi_j(x)$$

and

$$\tilde{\mu}(x) = (\Pi_n \mu)(x) = \sum_j \langle \mu, \phi_j \rangle_n \phi_j(x),$$

we have that

$$\hat{\mu}(x) - \tilde{\mu}(x) = \left\langle Y - \mu, \sum_j \phi_j(x)\phi_j \right\rangle_n = \sum_i a_i \varepsilon_i,$$

where $a_i = a_i(x; X_1, \dots, X_n) = \sum_j \phi_j(x)\phi_j(X_i)/n$, $\varepsilon_i = Y_i - \mu(X_i)$, and $\sum_i$ is summation is over $1 \le i \le n$. Consequently,

$$\operatorname{Var}\bigl(\hat{\mu}(x) | X_1, \dots, X_n\bigr) = \sum_i a_i^2 \sigma^2.$$

We need the following lemma, which can be proved easily by checking the Lindeberg condition.

LEMMA 3.1. *Suppose* $\xi_{i,n}$ *are independent with mean* 0 *and variance* 1. *In addition, assume that*

$$\lim_{\lambda \to \infty} \sup_n \sup_{1 \le i \le n} E\bigl(\xi_{i,n}^2 \operatorname{ind}\{|\xi_{i,n}| > \lambda\}\bigr) = 0.$$

If $\max_i \alpha_i^2 / \sum_i \alpha_i^2 \to 0$, *then*

$$\frac{\sum_i \alpha_i \xi_{i,n}}{\sqrt{\sum_i \alpha_i^2}} \Rightarrow N(0, 1).$$

Note that

$$\sum_i a_i^2 = \frac{1}{n^2} \sum_i \left(\sum_j \phi_j(x)\phi_j(X_i) \right)^2 = \frac{1}{n} \left\langle \sum_j \phi_j(x)\phi_j, \sum_j \phi_j(x)\phi_j \right\rangle_n.$$

Since $\{\phi_j\}$ is orthonormal,

$$\sum_i a_i^2 = \frac{1}{n} \sum_j \phi_j^2(x) \|\phi_j\|_n^2 = \frac{1}{n} \sum_j \phi_j^2(x).$$

By the Cauchy–Schwarz inequality,

$$a_i^2 \le \frac{1}{n^2} \sum_j \phi_j^2(x) \sum_j \phi_j^2(X_i).$$

Thus

$$\frac{a_i^2}{\sum_i a_i^2} \le \frac{1}{n} \sum_j \phi_j^2(X_i) \le \frac{1}{n} \sup_x \sum_j \phi_j^2(x).$$

Observe that

$$\begin{aligned} \sup_x \sqrt{\sum_j \phi_j^2(x)} &= \sup_x \sup_{(b_j)} \frac{|\sum_j b_j \phi_j(x)|}{\sqrt{\sum_j b_j^2}} \\ &\le \sup_{(b_j)} \frac{\sup_x |\sum_j b_j \phi_j(x)|}{\sqrt{\sum_j b_j^2}} \qquad (3.4) \\ &= \sup_{g \in G} \frac{\|g\|_\infty}{\|g\|_n}. \end{aligned}$$

[In fact, equality holds, since $\sup_x |\sum_j b_j \phi_j(x)| \le \sqrt{\sum_j b_j^2} \sup_x \sqrt{\sum_j \phi_j^2(x)}$ by the Cauchy–Schwarz inequality.] Hence,

$$\max_i \frac{a_i^2}{\sum_i a_i^2} \le \frac{1}{n} \sup_{g \in G} \frac{\|g\|_\infty^2}{\|g\|_n^2} = \frac{1}{n} \sup_{g \in G} \frac{\|g\|_\infty^2}{\|g\|^2} (1 + o_P(1)) = \frac{A_n^2}{n} (1 + o_P(1)).$$

Consequently, there is a set Ω_n with $P(\Omega_n) \to 0$ such that, $\max_i a_i^2 / \sum_i a_i^2 \le 2A_n^2/n$ on Ω_n. On the other hand, according to Lemma 3.1, if $\max_i a_i^2 / \sum_i a_i^2 \to 0$, then for any $\eta > 0$,

$$\Big| P\Big(\hat{\mu}(x) - \tilde{\mu}(x) \le t\sqrt{\operatorname{Var}(\hat{\mu}(x)|X_1, \dots, X_n)} \Big| X_1, \dots, X_n\Big) - \Phi(t)\Big| < \eta$$

for n sufficiently large. The first conclusion follows. The second conclusion then follows by the dominated convergence theorem. □

PROOF OF COROLLARY 3.1. By the proof of Theorem 1,

$$\operatorname{Var}(\hat{\mu}(x)|X_1, \dots, X_n) = \frac{\sigma^2}{n} \sum_j \phi_j^2(x).$$

It follows from (3.4) and its parenthetical remark that

$$\sup_x \sum_j \phi_j^2(x) = \left(\sup_{g \in G} \frac{\|g\|_\infty}{\|g\|_n} \right)^2.$$

Since $\sup_{g\in G} |\|g\|_n/\|g\| - 1| = o_P(1)$,

$$\sup_x \operatorname{Var}\bigl(\hat{\mu}(x)|X_1,\dots,X_n\bigr) = \frac{\sigma^2}{n}\biggl(\sup_{g\in G}\frac{\|g\|_\infty}{\|g\|}\biggr)^2 (1+o_P(1))$$
$$= \frac{A_n^2}{n}\sigma^2\bigl(1+o_P(1)\bigr).$$

Thus, there is a set Ω_n with $P(\Omega_n)\to 1$ such that, on Ω_n,

$$\sup_x \operatorname{Var}\bigl(\hat{\mu}(x)|X_1,\dots,X_n\bigr) \le \frac{2A_n^2\sigma^2}{n}.$$

Hence, by conditioning and using Chebyshev's inequality, we get that

$$\sup_x P\Biggl(|\hat{\mu}(x)-\tilde{\mu}(x)| \ge C\sqrt{\frac{A_n^2}{n}}\Biggr)$$
$$\le P(\Omega_n^c) + \sup_x E\Biggl[I_\Omega P\Biggl(|\hat{\mu}(x)-\tilde{\mu}(x)| \ge C\sqrt{\frac{A_n^2}{n}}\Bigg|X_1,\dots,X_n\Biggr)\Biggr]$$
$$\le P(\Omega_n^c) + \frac{2\sigma^2}{C^2}.$$

The desired result follows. □

PROOFS OF (3.1)–(3.3). Since $A_n \asymp \overline{A}_n = \sup_{g\in G}(\|g\|_\infty/\|g\|_{L_2})$, (3.1) follows from the fact that $\overline{A}_n \asymp J_n^{1/2}$ for polynomial splines [see Theorem 5.4.2 of DeVore and Lorentz (1993)]. To prove (3.2), without loss of generality, suppose $\mathcal{X} = [-1,1]$. We need only prove that $\overline{A}_n \asymp J_n$. It follows from Theorem 4.2.6 of DeVore and Lorentz (1993) that $\overline{A}_n \lesssim J_n$. To see that $\overline{A}_n \gtrsim J_n$, consider the Legendre polynomials p_j, $j=0,\dots,J_n$, which are special cases of Jacobi polynomials; see Section 4.1 of Szegö (1975). We have that $p_j(1)=1$, $j=1,\dots,J_n$, and

$$\int_{-1}^{1} p_j(x)p_{j'}(x)\,dx = \frac{2}{2j+1}\delta_{jj'}, \qquad j=0,\dots,J_n,$$

where $\delta_{jj'}$ is the Kronecker delta; see pages 58 and 68 of Szegö (1975). Set $\tilde{p}_j(x) = \sqrt{2j+1}\,p_j(x)$, $j=1,\dots,J_n$. Then $\{\tilde{p}_j, j=1,\dots,J_n\}$ is an orthonormal basis of G relative to the inner product induced by the uniform density on $[-1,1]$. Thus

$$\overline{A}_n^2 = \sup_{g\in G}\biggl(\frac{\|g\|_\infty}{\|g\|_{L_2}}\biggr)^2 \ge \sup_x \sum_{j=0}^{J_n}\tilde{p}_j^2(x) \ge \sum_{j=0}^{J_n}\tilde{p}_j^2(1) \ge \sum_{j=0}^{J_n}(2j+1) = (J_n+1)^2,$$

as desired; here the first "$\geq$" is obtained using the same argument as in (3.4). Finally let us prove (3.3). Let Ω_n with $P(\Omega_n) \to 1$ be the event that $\sup_{g \in G} | \|g\|_n / \|g\| - 1| < 1/2$ (see Lemma 2.3 for the existence of such a Ω_n). Note that

$$V_3 := E\left\{ \int_{\mathcal{X}} (\hat{\mu}(x) - \tilde{\mu}(x))^2 \, dx \,\middle|\, X_1, \dots, X_n \right\}.$$

Then $E(V_3 I_{\Omega_n}) \asymp E[\|\hat{\mu} - \tilde{\mu}\|^2 I_\Omega] \lesssim E[\|\hat{\mu} - \tilde{\mu}\|_n^2] = O(N_n/n)$ [see the proof of Theorem 1 of Huang (1998a)]. Consequently, $V_3 = O_P(N_n/n)$. □

4. Uniform asymptotic normality. We have proved in the last section that $\hat{\mu}(x) - \tilde{\mu}(x)$ is asymptotically normally distributed for general linear estimation spaces. In this section we show that a stronger result holds, namely that the asymptotic normality holds uniformly over the points $x \in \mathcal{X}$ where the regression function is to be estimated and uniformly over a broad class of design densities, error distributions, and regression functions. This type of uniformity is of interest in constructing asymptotic confidence intervals.

4.1. *Homoscedastic error case.* Consider now the regression model $Y = \mu(X) + \varepsilon$, where X and ε are independent, $E(\varepsilon) = 0$, and $\mathrm{Var}(\varepsilon) = \sigma^2 < \infty$. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample from the joint distribution of (X, Y). Let $\mathcal{P}_X$ be a class of possible distributions of X, and let $\mathcal{P}_\varepsilon$ be a class of possible distributions of ε. Let $\mathcal{F}$ be a function class in which μ resides.

Recall that the constant A_n defined in the previous section depends on the design distribution $\mathcal{L}(X)$. To specify this dependence, write $A_n = A_n(\mathcal{L}(X))$. The uniform asymptotic normality is given in the following theorem, whose proof is postponed to the end of this section.

THEOREM 4.1. *Suppose*

$$\lim_{n\to\infty} \sup\{A_n^2(\mathcal{L}(X)) : \mathcal{L}(X) \in \mathcal{P}_X\}/n = 0, \tag{4.1}$$

$$\limsup_{\lambda\to\infty} \sup\left\{ \frac{E[\varepsilon^2 \operatorname{ind}\{\varepsilon^2 \geq \lambda E(\varepsilon^2)\}]}{E(\varepsilon^2)} : \mathcal{L}(\varepsilon) \in \mathcal{P}_\varepsilon \right\} = 0, \tag{4.2}$$

and

$$\sup\left\{ P\left(\sup_{g \in G} \left| \frac{\|g\|_n}{\|g\|} - 1 \right| > \eta \right) : \mathcal{L}(X) \in \mathcal{P}_X \right\} = o(1), \qquad \eta > 0. \tag{4.3}$$

Set

$$\Delta_n = \sup_t \left| P\left(\frac{\hat{\mu}(x) - \tilde{\mu}(x)}{\sqrt{\mathrm{Var}(\hat{\mu}(x) | X_1, \dots, X_n)}} \leq t \,\middle|\, X_1, \dots, X_n \right) - \Phi(t) \right|.$$

Then

$$\sup\{P(\Delta_n > \eta) : x \in \mathcal{X}, \mathcal{L}(X) \in \mathcal{P}_X, \mathcal{L}(\varepsilon) \in \mathcal{P}_\varepsilon, \mu \in \mathcal{F}\} = o(1), \qquad \eta > 0.$$

Consequently,

$$\sup\Big\{\Big|P\Big(\tilde{\mu}(x) \ge \hat{\mu}(x) - t\sqrt{\operatorname{Var}(\hat{\mu}(x)|X_1, \dots, X_n)}\Big) - \Phi(t)\Big|, \\ t \in \mathbb{R},\ x \in \mathcal{X},\ \mathcal{L}(X) \in \mathcal{P}_X, \mathcal{L}(\varepsilon) \in \mathcal{P}_\varepsilon, \mu \in \mathcal{F}\Big\} = o(1).$$

When the density of X is bounded away from zero uniformly over $\mathcal{P}_X$, there is a simple upper bound for the quantity $\sup\{A_n^2(\mathcal{L}(X)) : \mathcal{L}(X) \in \mathcal{P}_X\}$ and the assumption (4.1) can be simplified accordingly. To be precise, suppose there is a constant $c > 0$ such that $\inf_x p_X(x) \ge c$ for $\mathcal{L}_X \in \mathcal{P}_X$. Recall that $\overline{A}_n = \sup_{g \in G}\{\|g\|_\infty / \|g\|_{L_2}\}$, where $\|\cdot\|_{L_2}$ is the L_2 norm relative to the uniform distribution on $\mathcal{X}$. We have that $A_n(\mathcal{L}(X)) \le \sqrt{|\mathcal{X}|/c}\,\overline{A}_n$ and hence that $\lim_n \overline{A}_n^2/n = 0$ is sufficient for (4.1).

The assumption (4.2) requires the class of standardized error distributions $\{\varepsilon/[E(\varepsilon^2)]^{1/2} : \mathcal{L}(\varepsilon) \in \mathcal{P}_\varepsilon\}$ to be uniformly integrable. It is satisfied when the standardized error distributions $\varepsilon/[E(\varepsilon^2)]^{1/2}$ for ε in the class $\mathcal{P}_\varepsilon$ possess uniformly bounded moments of order $2 + \delta$ for some $\delta > 0$, that is, $\sup\{[E(|\varepsilon|^{2+\delta})]^{1/(2+\delta)}/[E(\varepsilon^2)]^{1/2} : \mathcal{L}(\varepsilon) \in \mathcal{P}_\varepsilon\} < \infty$.

The assumption (4.3) requires that the empirical and theoretical norms be close uniformly over the estimation space when the sample size is large and that this should also hold uniformly over the class $\mathcal{P}_X$ of design densities. It follows from the proof of Lemma 10 of Huang (1998a) that (4.3) is satisfied if $\lim_n \sup\{A_n^2(\mathcal{L}(X)) : \mathcal{L}(X) \in \mathcal{P}_X\} N_n/n = 0$, so a sufficient condition for (4.3) is that $\lim_n \overline{A}_n^2 N_n/n = 0$ when the density of X is bounded away from zero uniformly over $\mathcal{P}_X$. If G is a space of polynomial splines satisfying Condition A.2 in the Appendix, then $\lim_n N_n \log n/n = 0$ is sufficient for (4.3), provided that the density of X is bounded away from zero and infinity uniformly over $\mathcal{P}_X$.

We now discuss the implication of this theorem. Let $\mu_\alpha^l(x)$ and $\mu_\alpha^u(x)$ be defined as in the previous section. Suppose the conclusion of Theorem 4.1 holds. Then

$$\lim_n \sup\{|P(\mu_\alpha^l(x) \le \tilde{\mu}(x) \le \mu_\alpha^u(x)) - (1-\alpha)| : \\ x \in \mathcal{X}, \mathcal{L}(X) \in \mathcal{P}_X, \mathcal{L}(\varepsilon) \in \mathcal{P}_\varepsilon, \mu \in \mathcal{F}\} = 0.$$

This says that the probability that the confidence interval $[\mu_\alpha^l(x), \mu_\alpha^u(x)]$ contains $\tilde{\mu}(x)$ is arbitrarily close to $1 - \alpha$ when n is sufficiently large; moreover, this closeness holds uniformly over the entire domain of the predictor and over a broad range of design densities, error distributions and regression functions. To be more specific, for any $\delta > 0$, there is a positive integer n_δ such that the coverage probability of the confidence interval differs from $1-\alpha$ by less than δ when $n > n_\delta$, where n_δ can be chosen to work simultaneously for $x \in \mathcal{X}$, $\mathcal{L}(X) \in \mathcal{P}_X$, $\mathcal{L}(\varepsilon) \in \mathcal{P}_\varepsilon$ and $\mu \in \mathcal{F}$. Thus $[\mu_\alpha^u(x), \mu_\alpha^l(x)]$ is an asymptotic level $(1-\alpha)$ confidence interval for $\tilde{\mu}(x)$ in the strong sense of Lehmann [(1999), page 222]. The uniform convergence of the coverage probability of a confidence interval to the nominal level is called uniform robustness by Lehmann and Loh (1990). Results in Section 5 can be used to determine when the bias term is of negligible size compared with the variance term; see Remark 5.2.

4.2. *Extensions.*

REMARK 4.1. Theorem 4.1 can be extended to handle heteroscedastic errors. Consider the model $Y = \mu(X) + \sigma(X)\varepsilon$, where X and ε are independent, $E(\varepsilon) = 0$, and $\mathrm{Var}(\varepsilon) < \infty$. For $0 < C_1 \le C_2 < \infty$, set $\Sigma = \{\sigma(\cdot): C_1 \le \sigma(x) \le C_2, x \in \mathcal{X}\}$. Under the conditions of Theorem 4.1, the least squares estimate $\hat{\mu}$, standardized by its conditional mean and conditional standard deviation, is asymptotically $N(0,1)$ uniformly in $x \in \mathcal{X}$, $\mathcal{L}(X) \in \mathcal{P}_X$, $\mathcal{L}(\varepsilon) \in \mathcal{P}_\varepsilon$, $\sigma \in \Sigma$ and $\mu \in \mathcal{F}$. The same result holds for the weighted least squares estimate $\hat{\mu}^w$ defined in Remark 3.2.

REMARK 4.2. Theorem 4.1 (and Remark 4.1) can also be extended to the general case when the error is not independent of the design random variable. Suppose we observe a random sample from the joint distribution of X and Y. Set $\varepsilon = Y - E(Y|X)$, which need not be independent of X. Theorem 4.1 remains valid if (4.2) is replaced by

$$\limsup_{\lambda\to\infty} \sup\left\{\frac{E[\varepsilon^2 \operatorname{ind}\{\varepsilon^2 \ge \lambda E(\varepsilon^2|X=x)\}|X=x]}{E(\varepsilon^2|X=x)}: \right.$$
$$\left. x \in \mathcal{X},\ \mathcal{L}(\varepsilon|X=x) \in \mathcal{P}_\varepsilon\right\} = 0,$$

where $\mathcal{P}_\varepsilon$ is now a class of possible conditional distributions of ε given $X = x$. This condition is satisfied if $\sup\{[E(|\varepsilon|^{2+\delta}|X=x)]^{1/(2+\delta)}/[E(\varepsilon^2|X=x)]^{1/2}: x \in \mathcal{X},\ \mathcal{L}(\varepsilon|X=x) \in \mathcal{P}_\varepsilon\} < \infty$ for some $\delta > 0$.

REMARK 4.3. The discussions in Theorem 4.1 and Remark 4.1 carry over to the fixed design case in an obvious manner as explained in Remark 3.3.

4.3. *Proof.*

PROOF OF THEOREM 4.1. From the proof of Theorem 1,

$$\frac{\hat{\mu}(x) - \tilde{\mu}(x)}{\sqrt{\mathrm{Var}(\hat{\mu}(x)|X_1,\dots,X_n)}} = \frac{\sum_i a_i \varepsilon_i}{\sqrt{\sum_i a_i^2}},$$

where a_i's are defined as in the proof of Theorem 1. It follows from Theorem V.8 of Petrov (1975) that there is an absolute constant C such that, for $\delta > 0$,

$$\begin{aligned}
&\sup_t \left| P\left(\frac{\sum_i a_i \varepsilon_i}{\sigma\sqrt{\sum_i a_i^2}} \le t \middle| X_1, \dots, X_n \right) - \Phi(t) \right| \\
&\quad \le C\left\{ \delta + \sum_i E\left[\frac{a_i^2 \varepsilon_i^2}{\sum_i a_i^2 \sigma^2} \operatorname{ind}\left(\frac{a_i^2 \varepsilon_i^2}{\sum_i a_i^2 \sigma^2} > \delta^2 \right) \middle| X_1, \dots, X_n \right] \right\} \\
&\quad \le C\left\{ \delta + \sum_i \frac{a_i^2}{\sum_i a_i^2} E\left[\frac{\varepsilon_i^2}{\sigma^2} \operatorname{ind}\left(\frac{\max_i a_i^2}{\sum_i a_i^2 \sigma^2} \varepsilon_i^2 > \delta^2 \right) \middle| X_1, \dots, X_n \right] \right\}.
\end{aligned}$$

Since $\max_i a_i^2 / (\sum_i a_i^2) \le (1/n) \sup_{g \in G} \{ \|g\|_\infty^2 / \|g\|_n^2 \}$,

$$\begin{aligned}
&\sup_t \left| P\left(\frac{\sum_i a_i \varepsilon_i}{\sigma\sqrt{\sum_i a_i^2}} \le t \middle| X_1, \dots, X_n \right) - \Phi(t) \right| \\
&\quad \le C\left\{ \delta + \max_i E\left[\frac{\varepsilon_i^2}{\sigma^2} \operatorname{ind}\left(\frac{1}{n} \sup_{g \in G} \left\{ \frac{\|g\|_\infty^2}{\|g\|_n^2} \right\} \varepsilon_i^2 > \sigma^2 \delta^2 \right) \middle| X_1, \dots, X_n \right] \right\}.
\end{aligned}$$

Note that the right-hand side of the above inequality does not depend on μ or $x \in \mathcal{X}$. By (4.2), we can choose ξ small enough so that $E[\varepsilon^2/\sigma^2 \operatorname{ind}(\varepsilon^2 > \sigma^2\delta^2/\xi)] < \delta$ uniformly for all ε with $\mathcal{L}(\varepsilon) \in \mathcal{P}_\varepsilon$. Thus,

$$\begin{aligned}
&\left\{ \frac{1}{n} \sup_{g \in G} \frac{\|g\|_\infty^2}{\|g\|_n^2} \le \xi \right\} \\
&\qquad \subset \left\{ \sup_t \left| P\left(\frac{\sum_i a_i \varepsilon_i}{\sigma\sqrt{\sum_i a_i^2}} \le t \middle| X_1, \dots, X_n \right) - \Phi(t) \right| \le 2C\delta \right\}.
\end{aligned} \tag{4.4}$$

On the other hand, (4.3) implies that

$$\sup\left\{ \left| P\left(\frac{1}{n} \sup_{g \in G} \left\{ \frac{\|g\|_\infty^2}{\|g\|_n^2} \right\} \le 2 \frac{1}{n} \sup_{g \in G} \left\{ \frac{\|g\|_\infty^2}{\|g\|^2} \right\} \right) - 1 \right| : \mathcal{L}(X) \in \mathcal{P}_X \right\} = o(1). \tag{4.5}$$

Moreover,

$$\frac{1}{n} \sup_{g \in G} \left\{ \frac{\|g\|_\infty^2}{\|g\|^2} \right\} \le \frac{1}{n} \sup\{ A_n^2(\mathcal{L}(X)) : \mathcal{L}(X) \in \mathcal{P}_X \} = o(1).$$

The first conclusion then follows from (4.1), (4.4) and (4.5). Note that

$$\begin{aligned}
&\left| P\left(\tilde{\mu}(x) \ge \hat{\mu}(x) - t\sqrt{\operatorname{Var}(\hat{\mu}(x) | X_1, \dots, X_n)} \right) - \Phi(t) \right| \\
&\qquad \le E|\Delta_n| \le \eta + P(\Delta_n > \eta), \qquad \eta > 0.
\end{aligned}$$

The second conclusion is a simple consequence of the first one. □

5. Size of the bias in polynomial spline regression. We show in Section 5.1 that the bias $\tilde{\mu}(x) - \mu(x)$ in polynomial spline regression is controlled by the minimum L_∞ norm of the error when the target regression function μ is approximated by a function in the estimation space. In contrast to the asymptotic normality result, the special properties of polynomial splines now play a crucial role. In Section 5.2, we will provide a condition for the bias to be asymptotically negligible compared with the variance term. Some discussion on obtaining the asymptotic expression of the bias is given in Section 5.3.

5.1. *Bias bound.* The control of the bias term relies on the stability in L_∞ norm of L_2 projections onto polynomial spline spaces. Specifically, let $\mathbb{G} = \mathbb{G}_n$ be a sequence of linear spaces and let $P = P_n$ denote the L_2 orthogonal projection onto $\mathbb{G}$ relative to an inner product $(\cdot,\cdot)_n$. Denote the norm corresponding to $(\cdot,\cdot)_n$ by $|||\cdot|||_n$. Since P is an orthogonal projection, it follows immediately that $|||Pf|||_n \le |||f|||_n$. If $\mathbb{G}$ is a sequence of polynomial spline spaces, we have the following much stronger result: under some regularity conditions, there is an absolute constant C which does not depend on n such that $\|Pf\|_\infty \le C\|f\|_\infty$ for any function f. The precise statement of such a result and its proof, along with the regularity conditions, will be given in the Appendix. The importance of this stability property of polynomial spline spaces can be seen from the following lemma.

LEMMA 5.1. *If there is an absolute constant C such that $\|Pf\|_\infty \le C\|f\|_\infty$ for any function f, then $\|P\mu - \mu\|_\infty \le (C+1)\inf_{g\in\mathbb{G}}\|\mu - g\|_\infty$.*

PROOF. Since $\mathbb{G}$ is finite-dimensional, by a compactness argument there is a $g^* \in \mathbb{G}$ such that $\|\mu - g^*\|_\infty = \inf_{g\in\mathbb{G}}\|\mu - g\|_\infty$. Note that $P\mu - g^* = P(\mu - g^*)$. So $\|P\mu - g^*\|_\infty \le C\|\mu - g^*\|_\infty$. Hence, by the triangle inequality,

$$\|P\mu - \mu\|_\infty \le \|P\mu - g^*\|_\infty + \|\mu - g^*\|_\infty \le (C+1)\|\mu - g^*\|_\infty. \qquad \square$$

The stability in L_∞ norm of L_2 projections enjoyed by polynomial spline spaces is not shared by other linear spaces such as polynomial or trigonometric polynomial spaces. In fact, if $\mathbb{G}$ is the space of polynomials (or trigonometric polynomials) of degree J_n or less on a compact interval, then there is a constant C that does not depend on n such that, $\sup_f\{\|Pf\|_\infty/\|f\|_\infty\} \ge C\log J_n$, where the supremum is taken over all continuous functions; see Corollaries 5.2 and 5.4 in Chapter 9 of DeVore and Lorentz (1993).

In applications to polynomial spline regression, we take $(\cdot,\cdot)_n$ to be the empirical inner product and $\mathbb{G} = G$ in the above general discussion. The desired stability property is satisfied according to the general result in the Appendix. The main result of this section is the following theorem. Set $\rho_n = \inf_{g\in G}\|\mu - g\|_\infty$.

THEOREM 5.1. *Suppose the sequence of estimation spaces G satisfies Conditions* A.2 *and* A.3 *in the Appendix*, *the density of X is bounded away from zero and infinity*, *and* $\lim_n N_n \log n/n = 0$. *Then there is an absolute constant C such that*, *except on an event whose probability tends to zero as $n \to \infty$*, $\|\Pi_n \mu - \mu\|_\infty \le C\rho_n$ [*i.e.*, $\sup_x |\tilde{\mu}(x) - \mu(x)| \le C\rho_n$].

The desired result follows from Corollary A.1 in the Appendix and Lemma 5.1.

5.2. *Univariate splines and tensor product splines.* If μ satisfies a suitable smoothness condition, results in approximation theory can be used to quantify ρ_n in Theorem 5.1. Under reasonable conditions, it can be shown that $\rho_n \lesssim N_n^{-p/d}$, where p typically corresponds to the number of bounded or continuous derivatives of μ. In the following we will give a precise statement of such a result in the case of univariate splines and tensor product splines. Results for bivariate or multivariate splines on triangulations are much more complicated. We refer readers to the approximation theory literature; see, for example, Chui (1988), de Boor, Höllig and Riemenschneider (1993) and Oswald (1994).

We first describe a smoothness condition commonly used in the nonparametric estimation literature and give the magnitude of ρ_n under such a condition. To this end, assume that $\mathcal{X}$ is the Cartesian product of compact intervals $\mathcal{X}_1, \dots, \mathcal{X}_d$. Let $0 < \beta \le 1$. A function h on $\mathcal{X}$ is said to satisfy a Hölder condition with exponent β if there is a positive number γ such that $|h(x_2) - h(x_1)| \le \gamma |x_2 - x_1|^\beta$ for x_1, $x_2 \in \mathcal{X}$; here $|x| = \left(\sum_{l=1}^d x_l^2\right)^{1/2}$ is the Euclidean norm of $x = (x_1, \dots, x_d) \in \mathcal{X}$. Given a d-tuple $\alpha = (\alpha_1, \dots, \alpha_d)$ of nonnegative integers, set $[\alpha] = \alpha_1 + \cdots + \alpha_d$ and let D^α denote the differential operator defined by

$$D^\alpha = \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

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

Given a set of real numbers $a = t_0 < t_1 < \cdots < t_J < t_{J+1} = b$, a function on $[a, b]$ is a polynomial spline with degree m and J interior knots $\{t_j, 1 \le j \le J\}$ if it is a polynomial of degree m in the intervals $[t_j, t_{j+1}]$, $0 \le j \le J$, and globally has $m - 1$ continuous derivatives. Let G_l, $1 \le l \le d$, be a space of polynomial splines on $\mathcal{X}_l$ with degree $m \ge p - 1$ and J_n interior knots. Suppose the knots have bounded mesh ratio (that is, the ratios of the differences between consecutive knots are bounded away from zero and infinity uniformly in n). Let G be the tensor product of $G_1, \dots, G_d$. (For $d = 1$, $G = G_1$, which is a univariate spline space.) If μ is p-smooth, then $\rho_n \lesssim J_n^{-p} = N_n^{-p/d}$ [see (13.69) and Theorem 12.8 of Schumaker (1981)]. Consequently, by Theorem 5.1, the bias is bounded above by a constant multiple of $N_n^{-p/d}$ except on an event with probability tending to zero.

THEOREM 5.2. *Suppose p_X is bounded away from zero and infinity. In addition, assume that $\sigma(\cdot)$ is bounded away from* 0. *Under the above setup, if* $\lim_n N_n/n^{d/(2p+d)} = \infty$ *and* $\lim_n N_n \log n/n = 0$, *then*

$$\sup_{x\in\mathcal{X}}\left|\frac{\tilde{\mu}(x)-\mu(x)}{\sqrt{\operatorname{Var}(\hat{\mu}(x)|X_1,\dots,X_n)}}\right| = o_P(1).$$

PROOF. Let $\{B_j, 1 \le j \le N_n\}$ be the tensor product B-spline basis of G [see Chapter 12 of Schumaker (1981)]. Then Conditions A.2 and A.3 are satisfied. It follows from Theorem 5.1 that

$$\sup_x |\tilde{\mu}(x)-\mu(x)| = O_P(\rho_n) = O_P(N_n^{-p/d}). \tag{5.1}$$

By Lemma 2.3, $\lim_n N_n \log n/n = 0$ ensures that $\sup_{g\in G} |\|g\|_n/\|g\| - 1| = o_P(1)$. Let $\{\phi_j, 1 \le j \le N_n\}$ be an orthonormal basis of G relative to the empirical inner product. For any $g \in G$, write $g = \sum_j b_j\phi_j$ with $b_j = \langle g, \phi_j\rangle_n$. By the proof of Theorem 3.1 and the Cauchy–Schwarz inequality,

$$\begin{aligned}\operatorname{Var}(\hat{\mu}(x)|X_1,\dots,X_n) &\lesssim \frac{1}{n}\sum_j \phi_j^2(x) \ge \frac{1}{n}\frac{|\sum_j b_j\phi_j(x)|^2}{\sum_j b_j^2}\\ &= \frac{1}{n}\frac{|g(x)|^2}{\|g\|_n^2} = \frac{1}{n}\frac{|g(x)|^2}{\|g\|^2}(1+o_P(1)).\end{aligned} \tag{5.2}$$

Set $g_x = \sum_{j\in I_x} B_j$, where I_x denotes the set of indices of the basis functions whose support contains x. Then, by (A.2) in the Appendix, $\|g_x\|^2 \lesssim h^d\#(I_x) \lesssim h^d \asymp N_n^{-1}$. Moreover, $g_x(x) = 1$ since $\sum_j B_j(x) = 1$ for all x. Taking $g = g_x$ in (5.2), we obtain that

$$\operatorname{Var}(\hat{\mu}(x)|X_1,\dots,X_n) \ge \frac{1}{n}\frac{|g_x(x)|^2}{\|g_x\|^2}(1+o_P(1)) \ge C\frac{N_n}{n}(1+o_P(1))$$

for some constant C that can be made independent of x. This together with (5.1) yields the desired result. □

The above theorem determines when the bias term is of negligible magnitude compared with the variance term. For example, if μ is a univariate function with bounded second derivative, then a sufficient condition for being able to "ignore" the bias asymptotically in constructing a confidence interval is that $\lim_n N_n/n^{1/5} = \infty$. Note that the variance is of order N_n/n. Balancing the squared bias and variance, that is, letting $N_n^{-2p/d} \asymp N_n/n$ or equivalently $N_n \asymp n^{d/(2p+d)}$, yields the optimal rate of convergence $n^{-2p/(2p+d)}$ [see Stone (1982)]. The required condition that $\lim_n N_n/n^{d/(2p+d)} = \infty$ for making the bias asymptotically negligible simply means that one need use a larger number of knots than what is needed for achieving the optimal rate of convergence ("undersmoothing").

REMARK 5.1. Consider G being a space of univariate splines on a compact interval $[a, b]$ with degree m and knots $a = t_0 < t_1 < \cdots < t_J < t_{J+1} = b$, where $J = J_n$. It is required in Theorems 5.1 and 5.2 that the knot sequence have bounded mesh ratio, that is,

$$\frac{\max_{0\le j\le J}(t_{j+1} - t_j)}{\min_{0\le j\le J}(t_{j+1} - t_j)} \le \gamma \tag{5.3}$$

for some positive constant γ. In light of the work of de Boor (1976), this condition can be weakened to

$$\frac{\max_{0\le j\le J-m}(t_{j+m+1} - t_j)}{\min_{0\le j\le J-m}(t_{j+m+1} - t_j)} \le \gamma$$

for some positive constant γ. In ZSW (1998) a stronger condition was used; it is required that the knots be asymptotically equally spaced, namely,

$$\max_{1\le j\le J}\left|\frac{(t_{j+1} - t_j)}{(t_j - t_{j-1})} - 1\right| = o(1). \tag{5.4}$$

REMARK 5.2. Theorems 5.1 and 5.2 also hold when the weighted least squares estimate is used, that is, when Π_n is replaced by Π_n^w as defined in Remark 3.2. [We need to assume that the variance function $\sigma(\cdot)$ is bounded away from zero and infinity.] Moreover, this theorem extends to the fixed design case in an obvious manner. Note that Π_n depends on the design points $X_1, \dots, X_n$ and hence on the design density p_X. The result in Theorems 5.1 and 5.2 can be made to hold uniformly over a class of design densities. Specifically, let C_1 and C_2 be constants such that $0 < C_1 \le C_2 < \infty$. Set $\mathcal{P}_X = \{p_X : C_1 \le p_X(x) \le C_2$ for $x \in \mathcal{X}\}$. Then, under the conditions in Theorem 5.1, there is an absolute constant C such that

$$\lim_n \sup\{P(\|\Pi_n\mu - \mu\|_\infty \le C\rho_n) : \mathcal{L}(X) \in \mathcal{P}_X\} = 1.$$

Similarly, under the conditions in Theorem 5.2,

$$\sup\left\{P\left(\sup_{x\in\mathcal{X}}\left|\frac{\tilde{\mu}(x) - \mu(x)}{\sqrt{\operatorname{Var}(\hat{\mu}(x)|X_1, \dots, X_n)}}\right| > \eta\right) : \mathcal{L}(X) \in \mathcal{P}_X\right\} = o(1), \qquad \eta > 0.$$

5.3. *Asymptotic bias expression.* In the previous section, we derived an upper bound for the bias term in polynomial spline regression. It is tantalizing to obtain precise asymptotic bias expressions for our spline estimators in the general setup of this paper. We found that this is a very difficult task, however. Recently, ZSW (1998) provided formulas of local asymptotic bias for univariate spline regression assuming that the regression function μ is in C^p (i.e., μ has a continuous pth derivative). In the following we will discuss what additional insights we can gain using our general results. We found that, surprisingly, the leading term in the asymptotic bias expression disappears if one uses splines with higher degree than that in ZSW (1998).

Consider estimating a univariate regression function $\mu(x) = E(Y|X = x)$ based on an i.i.d. sample from the joint distribution of (X, Y), where $X \in \mathcal{X}$, $Y \in \mathbb{R}$ with $\mathcal{X}$ a compact interval on $\mathbb{R}$. Let the estimation space $G = G_n$ be the space of polynomial splines on $\mathcal{X} = [a, b]$ with degree m and knots $a = t_0 < t_1 < \cdots < t_J < t_{J+1} = b$. Denote $h_j = t_{j+1} - t_j$ and $h_n = \max_j h_j$. Suppose that $\mu \in C^p$ for an integer $p > 0$. ZSW (1998) obtain that, under some regularity conditions, if the degree of splines satisfies $m = p - 1$ and the knots are asymptotically equally spaced [see (5.4)], then

$$\begin{aligned} &E\big(\hat{\mu}(x)|X_1, \dots, X_n\big) - \mu(x) \\ &\qquad = -\frac{f^{(p)}(x)h_i^p}{p!} B_p\left(\frac{x - t_i}{h_i}\right) + o(h_n^p), \qquad t_i < x \le t_{i+1}, \end{aligned} \tag{5.5}$$

where B_p is the pth Bernoulli polynomial [see Barrow and Smith (1978)]. This provides the first asymptotic bias expression for polynomial spline regression. However, the condition on the knot sequence is stringent and one would like to know if it can be relaxed. Inspecting the proofs reveals that the condition on the knots is a critical one. A key step in the argument uses a result of Barrow and Smith (1978), which relies crucially on the assumption that the knots be asymptotically equally spaced. Moreover, the requirement that the degree of the spline must satisfy $m = p - 1$ may significantly limit the scope of application of the result. For example, if μ has a continuous second derivative, (5.5) only gives the asymptotic bias for linear spline estimates. One wonders what we can say about the bias for quadratic or cubic spline estimates. Indeed, quadratic or cubic splines are more commonly used than linear splines in practice because of their smooth appearance.

Can our general results shed some light? Let the degree of splines satisfy $m \ge p$. Since $\mu \in C^p$, it follows from Theorem 6.27 of Schumaker (1981) that $\rho_n = \inf_{g \in G} \|\mu - g\|_\infty = o(h_n^p)$. [If $m = p - 1$, we only have $\rho_n = O(h_n^p)$.] Suppose the knot sequence has bounded mesh ratio [see (5.3)]. By Theorem 5.1, we have that $\sup_x |E(\hat{\mu}(x)|X_1, \dots, X_n) - \mu(x)| \le C\rho_n = o(h_n^p)$. Hence, interestingly enough, if one were to increase the degree of spline from $m = p - 1$ to any integer $m \ge p$, then the leading term in the bias expression (5.5) would disappear. To be specific, suppose that μ has a continuous second derivative. Then if one uses linear splines, the asymptotic bias is given by (5.5) according to ZSW (1998). On the other hand, if one uses quadratic or cubic splines, then the asymptotic bias is of a smaller order. Hence, for constructing asymptotic confidence intervals, use of quadratic or cubic splines will make the bias asymptotically negligible and thus avoid the additional burden of estimating the second derivative in (5.5) for linear splines. Note that increasing the degree of splines by a fixed amount will not change the asymptotic order of the variance term (see Corollary 3.1 and the proof of Theorem 5.2). The above discussion could be viewed as an asymptotic argument for promoting the use of quadratic or cubic splines instead of linear splines. Of course, one could prefer quadratic or cubic splines to linear splines just because they provide estimates with smoother visual appearance.

It is interesting to compare the results in the previous paragraph with those in Section 5.2. If μ has bounded (not necessarily continuous) pth derivative, then $\sup_x |E(\hat{\mu}(x)|X_1,\dots,X_n)-\mu(x)| = O(N_n^{-p}) = O(h_n^p)$. Taking $h_n \asymp n^{-1/(2p+1)}$ (or, equivalently, $N_n \asymp n^{2p+1}$), which balances the order of squared bias and variance (Section 5.2), we obtain the optimal rate of convergence $n^{-2p/(2p+1)}$ [see Stone (1982)]. On the other hand, if μ has continuous pth derivative, for $h_n \asymp n^{-(2p+1)}$, the squared bias is bounded by $o(h_n^{2p}) = o(n^{-2p/(2p+1)})$ while the variance is of order $n^{-2p/(2p+1)}$. Hence, if one would like to assume the continuity of the pth derivative of μ, then the bias is asymptotically negligible when the number of knots is chosen for the estimate to achieve the optimal rate of convergence.

To generalize the above discussion to tensor product splines, let $\mathcal{X}$ be the Cartesian product of compact intervals $\mathcal{X}_1,\dots,\mathcal{X}_d$, and as in Theorem 5.2, let the estimation space G be the tensor product of spaces $G_1,\dots,G_d$, of univariate splines with degree m defined on $\mathcal{X}_1,\dots,\mathcal{X}_d$, respectively. The proof of the next result is similar to that of Theorem 5.2 and is omitted.

THEOREM 5.3. *Suppose p_X is bounded away from zero and infinity and that μ has all continuous partial derivatives of order $p>0$. If $m \ge p$ and* $\lim_n N_n \log n/n = 0$, *then $\tilde{\mu}(x)-\mu(x) = o_P(N_n^{-p/d})$ and* $\mathrm{Var}(\hat{\mu}(x)|X_1,\dots,X_n) \asymp N_n/n$. *Consequently, if $N_n \asymp n^{d/(2p+d)}$, then*

$$\sup_{x\in\mathcal{X}} \left| \frac{\tilde{\mu}(x)-\mu(x)}{\sqrt{\mathrm{Var}(\hat{\mu}(x)|X_1,\dots,X_n)}} \right| = o_P(1).$$

According to this theorem, if one assumes the continuity of all order $p>0$ partial derivatives of μ, then the leading term in the asymptotic bias of a tensor product spline estimate (with degree $m \ge p$) is zero. We believe that the leading term will not be zero if we assume only boundedness of all the partial derivatives of μ. However, finding the precise asymptotic bias expression for spline estimates under this assumption is difficult. As a comparison, Ruppert and Wand (1994) derive the asymptotic bias expression for multivariate local linear and quadratic regression assuming the continuity of partial derivatives of the regression function. (They also require that the density of X be continuously differentiable, which is not needed for our results.) No result is available for multivariate local polynomial regression under boundedness conditions on the partial derivatives of the regression function.

6. Expressions of the conditional variance. In applications, a convenient basis of the linear estimation space is usually employed and consequently the least squares estimate is represented as a linear combination of the basis functions. For example, the B-spline basis is often used if polynomial splines are used to construct the estimation space. In this section we give expressions for the conditional variance and asymptotic conditional variance of the least squares estimate in terms of a basis of the estimation space. These expressions help in evaluating the variability of the least squares estimate. They also tell us how the variance of the least squares estimate at a point depends on the design densities and the location of this point. The results in this section apply to general estimation spaces.

6.1. *Homoscedastic error case.* Let $\{B_j, 1 \le j \le N_n\}$ be a basis of G and let $\mathbf{B}(x)$ denote the column vector with entries $B_j(x), 1 \le j \le N_n$. Then the matrix $E_n[\mathbf{B}(X)\mathbf{B}^t(X)]$ is nonnegative definite. When G is empirically identifiable, $E_n[\mathbf{B}(X)\mathbf{B}^t(X)]$ is positive definite. In fact, $\beta^t E_n[\mathbf{B}(X)\mathbf{B}^t(X)]\beta = 0$ implies that $E_n[(\mathbf{B}^t(X)\beta)^2] = 0$. By the empirical identifiability of G, $\mathbf{B}^t(x)\beta = 0$ for all $x \in \mathcal{X}$ and hence $\beta = 0$.

THEOREM 6.1. *Suppose $\sigma^2(x) = \sigma^2$ is a constant. If G is empirically identifiable, then*

$$\mathrm{Var}(\hat{\mu}(x)|X_1, \dots, X_n) = \frac{1}{n}\mathbf{B}^t(x)\{E_n[\mathbf{B}(X)\mathbf{B}^t(X)]\}^{-1}\mathbf{B}(x)\sigma^2.$$

Moreover, if $\sup_{g\in G} |\|g\|_n/\|g\| - 1| = o_P(1)$, *then*

$$\mathrm{Var}(\hat{\mu}(x)|X_1, \dots, X_n) = \frac{1}{n}\mathbf{B}^t(x)\{E[\mathbf{B}(X)\mathbf{B}^t(X)]\}^{-1}\mathbf{B}(x)\sigma^2(1 + o_P(1)).$$

PROOF. The first conclusion of Theorem 6.1 follows from standard linear regression theory. To prove the second conclusion, we need the following lemma, whose proof is simple and thus omitted.

LEMMA 6.1. *For positive definite symmetric matrices A and B, set*

$$\varepsilon_n = \sup_u \left| \frac{u^t A u}{u^t B u} - 1 \right|.$$

Then

$$\sup_u \left| \frac{u^t A^{-1} u}{u^t B^{-1} u} - 1 \right| \le \frac{\varepsilon_n^2}{1 - \varepsilon_n} + 2\frac{\varepsilon_n}{\sqrt{1 - \varepsilon_n}}.$$

Consequently, if $A = A_n$ and $B = B_n$, then

$$\sup_u \left| \frac{u^t A u}{u^t B u} - 1 \right| = o(1) \quad \Longleftrightarrow \quad \sup_u \left| \frac{u^t A^{-1} u}{u^t B^{-1} u} - 1 \right| = o(1).$$

Note that

$$\begin{aligned}
&\sup_{\beta}\left|\frac{\beta^t E_n[\mathbf{B}(X)\mathbf{B}^t(X)]\beta}{\beta^t E[\mathbf{B}(X)\mathbf{B}^t(X)]\beta}-1\right| \\
&\qquad = \sup_{\beta}\left|\frac{E_n[(\beta^t\mathbf{B}(X))^2]}{E[(\beta^t\mathbf{B}(X))^2]}-1\right| = \sup_{g\in G}\left|\frac{\|g\|_n^2}{\|g\|^2}-1\right| = o_P(1).
\end{aligned}$$

It follows from Lemma 6.1 that

$$\sup_{x}\left|\frac{\mathbf{B}^t(x)\{E_n[\mathbf{B}(X)\mathbf{B}^t(X)]\}^{-1}\mathbf{B}(x)}{\mathbf{B}^t(x)\{E[\mathbf{B}(X)\mathbf{B}^t(X)]\}^{-1}\mathbf{B}(x)}-1\right| = o_P(1),$$

which yields the second conclusion of Theorem 6.1. □

6.2. *Extensions to heteroscedastic case and fixed design.*

REMARK 6.1. When the errors are heteroscedastic, expressions for conditional variance of the least squares estimate can be obtained similarly. Indeed,

$$\begin{aligned}
&\operatorname{Var}(\hat{\mu}(x)|X_1,\dots,X_n) \\
&\qquad = \frac{1}{n}\mathbf{B}^t(x)\{E_n[\mathbf{B}(X)\mathbf{B}^t(X)]\}^{-1}E_n[\sigma^2(X)\mathbf{B}(X)\mathbf{B}^t(X)] \\
&\qquad\quad \times \{E_n[\mathbf{B}(X)\mathbf{B}^t(X)]\}^{-1}\mathbf{B}(x).
\end{aligned} \tag{6.1}$$

If $\sup_{g\in G}|\|g\|_n/\|g\|-1| = o_P(1)$ and $\sup_{g\in G}|\|g\|_{n,\sigma}/\|g\|_\sigma - 1| = o_P(1)$, then

$$\begin{aligned}
&\operatorname{Var}(\hat{\mu}(x)|X_1,\dots,X_n) \\
&\qquad = \frac{1}{n}\mathbf{B}^t(x)\{E[\mathbf{B}(X)\mathbf{B}^t(X)]\}^{-1}E[\sigma^2(X)\mathbf{B}(X)\mathbf{B}^t(X)] \\
&\qquad\quad \times \{E[\mathbf{B}(X)\mathbf{B}^t(X)]\}^{-1}\mathbf{B}(x)(1+o_P(1)).
\end{aligned} \tag{6.2}$$

The argument in the proof of Theorem 6.1 can be modified to prove these results.

REMARK 6.2. When the errors are heteroscedastic, if the variance function $\sigma(\cdot)$ is known, the weighted least squares estimate $\hat{\mu}^w$ in Remark 3.2 can be used. Suppose $g \in G$ and $\|g\|_{n,1/\sigma} = 0$ together imply that $g = 0$ everywhere on $\mathcal{X}$. Then $E_n[\mathbf{B}(X)\mathbf{B}^t(X)/\sigma^2(X)]$ is positive definite. The same argument as in the proof of Theorem 6.1 yields that

$$\operatorname{Var}(\hat{\mu}^w(x)|X_1,\dots,X_n) = \frac{1}{n}\mathbf{B}^t(x)\{E_n[\mathbf{B}(X)\mathbf{B}^t(X)/\sigma^2(X)]\}^{-1}\mathbf{B}(x).$$

Moreover, if $\sup_{g\in G}|\|g\|_{n,1/\sigma}/\|g\|_{1/\sigma}-1| = o_P(1)$, then

$$\operatorname{Var}(\hat{\mu}^w(x)|X_1,\dots,X_n) = \frac{1}{n}\mathbf{B}^t(x)\{E[\mathbf{B}(X)\mathbf{B}^t(X)/\sigma^2(X)]\}^{-1}\mathbf{B}(x)(1+o_P(1)).$$

REMARK 6.3. Theorem 6.1 and Remarks 6.1 and 6.2 carry over to the fixed design case. We need only replace conditional expectations by unconditional expectations. Obviously, the empirical inner products are interpreted as nonrandom quantities and conditions such as $\sup_{g\in G} |\|g\|_n/\|g\| - 1| = o_P(1)$ should be replaced by $\sup_{g\in G} |\|g\|_n/\|g\| - 1| = o(1)$ and so forth.

7. Additive models. In this section we give a preliminary analysis of additive models. For tractability, we focus on the special case of tensor product designs. Such a design was used in Chen (1991) when discussing rates of convergence of interaction spline models, and can be viewed as a first step towards a general theory.

Let $\mathcal{X}$ be the Cartesian product of compact intervals $\mathcal{X}_1, \dots, \mathcal{X}_d$. Consider the additive model

$$\mu(x) = \mu_1(x_1) + \mu_2(x_2) + \cdots + \mu_d(x_d), \qquad x_l \in \mathcal{X}_l, 1 \le l \le d.$$

To construct an appropriate estimation space G, let G_l, $1 \le l \le d$, be a space of polynomial splines on $\mathcal{X}_l$ of a fixed degree $m > 0$ and having $J_n = N_n - m - 1$ interior knots with bounded mesh ratio [see (5.3)]. Here the number of interior knots is chosen to be the same for all G_l for notational simplicity. Set $G = G_1 + \cdots + G_d = \{g_1 + \cdots + g_d : g_l \in G_l, 1 \le l \le d\}$. The asymptotic normality results in Sections 3 and 4 are established for general estimation spaces and thus are applicable to the current situation to deal with the variance term. However, since G does not have a locally supported basis, the argument in Section 5 cannot be used to handle the bias term. In fact, a basis of G consists of basis functions of $G_1, \dots, G_d$. For each $l = 1, \dots, d$, any basis function of G_l, viewed as a function on $\mathcal{X}$, will be supported on the whole range of $\mathcal{X}_k$ for all $k \ne l$.

In the following we will restrict our attention to the special case of fixed tensor product design where we can get a good handle on the bias term. To be specific, suppose the observed covariates are $\{(x_{i_1}^{(1)}, \dots, x_{i_d}^{(d)}), x_{i_l}^{(l)} \in \mathcal{X}_l, 1 \le i_l \le n_l, 1 \le l \le d\}$ so that the sample size is $n = \prod_{l=1}^d n_l$. We consider the asymptotics when $n \to \infty$ and $n_l \to \infty$, $1 \le l \le k$. Note that in this setup,

$$E_n(f) = \frac{1}{n_1 \cdots n_d} \sum_{i_1,\dots,i_d} f(x_{i_1}^{(1)}, \dots, x_{i_d}^{(d)}).$$

As in previous sections, let Π_n be the orthogonal projection onto G relative to the empirical inner product. According to Lemma 2.4, the bias is $E(\hat{\mu}) - \mu = \Pi_n\mu - \mu$. We need to know how to handle the projection operator Π_n. Let $\Pi_{n,0}$ and $\Pi_{n,l}$, $1 \le l \le d$, be orthogonal projections onto the space of constant functions and onto G_l, respectively. Because of the tensor product design, for any function f on $\mathcal{X}$, $\Pi_{n,0}(f) = E_n(f)$ and $\Pi_n f - \Pi_{n,0} f = \sum_{1\le l\le d}(\Pi_{n,l} f - \Pi_{n,0} f)$. Hence the projection onto G can be decomposed as the summation of the projections onto component spaces G_l. This turns out to be important, as the projection onto the individual spline space G_l can then be handled as in Section 5.1 using results in the Appendix.

Set $\rho_n = \inf_{g \in G} \|\mu - g\|_\infty$. We present results only for the homoscedastic error case. An extension to the heteroscedastic error case is straightforward.

CONDITION 7.1. For each l, $1 \le l \le d$, there is a probability cumulative distribution function $F^{(l)}$ which has a density that is bounded away from 0 and infinity on $\mathcal{X}_l$ such that

$$\sup_{x \in \mathcal{X}_l} |F_n^{(l)}(x) - F^{(l)}(x)| = o\left(\frac{1}{J_n}\right), \qquad 1 \le l \le d,$$

where $F_n^{(l)}(x) = (1/n_l)\sum_{j=1}^{n_l} \operatorname{ind}(x_j^{(l)} \le x)$ is the empirical cumulative distribution of $x_1^{(l)}, \dots, x_{n_l}^{(l)}$.

THEOREM 7.1. *Suppose* $\sigma^2(x) = \sigma^2$ *is a constant and Condition* 7.1 *holds. Under the above setup, if* $\lim_n J_n/n = 0$, *then*

$$\mathcal{L}\left(\frac{\hat{\mu}(x) - E(\hat{\mu}(x))}{\sqrt{\operatorname{Var}(\hat{\mu}(x))}}\right) \Rightarrow N(0, 1), \qquad n \to \infty.$$

Moreover, there is an absolute constant C such that $\sup_{x \in \mathcal{X}} |E(\hat{\mu}(x)) - \mu(x)| \le C\rho_n$.

We should point out that the above theorem only deals with a special case of additive models. Local asymptotics for general random design additive models are unknown. It is also of interest to study the behavior of the components of the estimates in additive models. Such issues are of both theoretical and practical importance and deserve substantial further development.

PROOF OF THEOREM 7.1. The first part of the theorem follows from Theorem 3.1 (see Remark 3.3). To check the required conditions, one need only note that

$$\tilde{A}_n \lesssim \sum_{1 \le l \le d} \sup_{g_l \in G_l} \frac{\|g_l\|_\infty}{\|g_l\|_n} \lesssim J_n^{1/2},$$

which follows from the nature of the tensor product design, the properties of polynomial splines and Lemma 7.1 in the following. It remains to prove the second part of the theorem. By a compactness argument, there is a μ^* in G such that $\|\mu^* - \mu\|_\infty = \rho_n$. Set $\mu_0 = E_n(\mu)$ and $\mu_0^* = E_n(\mu)$. Since

$$\Pi_n((\mu - \mu_0) - (\mu^* - \mu_0^*)) = \sum_j \Pi_{n,j}((\mu - \mu_0) - (\mu^* - \mu_0^*)),$$

we have that, except on an event whose probability tends to zero,

$$\begin{aligned}\|\Pi_n\big((\mu-\mu_0)-(\mu^*-\mu_0^*)\big)\|_\infty &\le \sum_j \|\Pi_{n,j}\big((\mu-\mu_0)-(\mu^*-\mu_0^*)\big)\| \\ &\lesssim \sum_j \|(\mu-\mu_0)-(\mu^*-\mu_0^*)\|_\infty \lesssim \|\mu-\mu^*\|_\infty;\end{aligned}$$

here we used Theorem A.1, Lemma 7.1, and the fact that $\|\mu_0-\mu_0^*\|_\infty \le \|\mu-\mu^*\|_\infty$. Consequently, $\|\Pi_n\mu-\mu^*\|_\infty \lesssim \|\mu-\mu^*\|_\infty = \rho_n$. The desired result then follows from the triangle inequality. □

LEMMA 7.1. *Let G_l be defined as at the beginning of this section and denote by $t_0 < t_1 < \cdots < t_J < t_{J+1}$ the knot sequence of splines in G_l. Suppose Condition* 7.1 *holds. Then*

$$\sup_{g\in G_l}\left|\frac{\int_{t_j}^{t_{j+1}} g^2(x)\,dF_n^{(l)}(x)}{\int_{t_j}^{t_{j+1}} g^2(x)\,dF^{(l)}(x)}-1\right| = o(1), \qquad 0\le j\le J.$$

PROOF. Denote $h_n=\max_j(t_{j+1}-t_j)$. Integration by parts gives

$$\begin{aligned}&\int_{t_j}^{t_{j+1}} g^2(x)\,dF_n^{(l)}(x)-\int_{t_j}^{t_{j+1}} g^2(x)\,dF^{(l)}(x)\\ &\qquad = -\int_{t_j}^{t_{j+1}} \{F_n^{(l)}(x)-F^{(l)}(x)\}g(x)g'(x)\,dx.\end{aligned}$$

By Theorem 2.7 of Chapter 4 in DeVore and Lorentz (1993),

$$\left\{\int_{t_j}^{t_{j+1}} \{g'(x)\}^2\,dx\right\}^{1/2} \le h_n^{-1}\left\{\int_{t_j}^{t_{j+1}} g^2(x)\,dx\right\}^{1/2}.$$

Thus,

$$\begin{aligned}&\left|\int_{t_j}^{t_{j+1}} g^2(x)\,dF_n^{(l)}(x)-\int_{t_j}^{t_{j+1}} g^2(x)\,dF^{(l)}(x)\right|\\ &\quad\le \sup_{x\in\mathcal{X}_l}|F_n^{(l)}(x)-F^{(l)}(x)|\left\{\int_{t_j}^{t_{j+1}} g^2(x)\,dx\right\}^{1/2}\left\{\int_{t_j}^{t_{j+1}} \{g'(x)\}^2\,dx\right\}^{1/2}\\ &\quad\le o(h_n)h_n^{-1}\int_{t_j}^{t_{j+1}} g^2(x)\,dx.\end{aligned}$$

The desired result follows. □

APPENDIX: THE STABILITY IN L_∞ NORM OF L_2 PROJECTIONS ONTO POLYNOMIAL SPLINE SPACES

In this Appendix we prove a result on the stability in L_∞ norm of L_2 projections onto polynomial spline spaces, which plays a key role in controlling the bias for polynomial spline regression. This result is general enough to handle polynomial spline spaces on regions of arbitrary dimension. In particular, the polynomial spline space we considered can be a tensor product of an arbitrary number of univariate spline spaces or a space of splines constructed directly on triangles or high-dimensional simplices. Similar results were established by Douglas, Dupont and Wahlbin (1975) and de Boor (1976) for univariate spline spaces; see also Section 13.4 of DeVore and Lorentz (1993). A result for the tensor product of two univariate spline spaces was obtained by Stone (1989).

Let $\mathcal{X}$ be a closed, bounded subset of $\mathbb{R}^d$. Suppose that $\mathcal{X}$ is polyhedral, that is, representable by a finite partition into nondegenerating simplices. Consider a sequence of partitions $\Delta_n = \{\delta : \delta \subset \mathcal{X}\}$ of $\mathcal{X}$. We require that each $\delta \in \Delta_n$ be polyhedral. This includes as special cases simplicial (triangular in $\mathbb{R}^2$, tetrahedral in $\mathbb{R}^3$) or rectangular partitions of $\mathcal{X}$ (if $\mathcal{X}$ itself is composed of $\mathbb{R}^d$-rectangles). As n grows, the elements in Δ_n are required to be shrinking in size and increasing in number. In statistical applications the index n usually corresponds to sample size.

Consider a space $\mathbb{G}_n$ of piecewise polynomials (polynomial splines) over the partition Δ_n of $\mathcal{X}$ with the degree of each polynomial piece bounded above by a common constant. Specifically, let m be a fixed integer. Every $g \in \mathbb{G}_n$ is a polynomial of degree m or less when restricted to each $\delta \in \Delta_n$. In our discussion the polynomial pieces may or may not join together smoothly.

Let ν_n be a measure on $\mathcal{X}$. Define a corresponding inner product by $(f_1, f_2)_n = \int_{\mathcal{X}} f_1 f_2 \, d\nu_n$ for any functions f_1 and f_2 on $\mathcal{X}$ such that the indicated integral is well defined. Denote the induced L_2 norm by $\|\cdot\|_n$. For later usage, we define the local versions of this L_2 norm by $\|f\|_{n,\delta} = (\int_\delta f^2 \, d\nu_n)^{1/2}$ for $\delta \in \Delta_n$. For any function f on $\mathcal{X}$, let $P_n f$ denote the orthogonal projection onto $\mathbb{G}$ relative to $(\cdot,\cdot)_n$. The main result of this appendix is concerned with bounding the supreme norm of $P_n f$ by the supreme norm of f under suitable conditions.

The dependence on n of the measure ν_n and hence of the projection operator P_n is important in the formulation, since ν_n will be taken as the empirical distribution in our application of the result. This formulation is different from those in the mathematics literature, where ν_n is usually taken as Lebesgue measure. For notational convenience, we suppress from now on the subscript n in Δ_n, $\mathbb{G}_n$, ν_n, $(\cdot,\cdot)_n$, $\|\cdot\|_n$, $\|f\|_{n,\delta}$ and P_n.

Let $\|\cdot\|^*$ denote the L_2 norm induced by Lebesgue measure, that is, $\|f\|^{*2} = \int_{\mathcal{X}} f^2(x)\, dx$. Given $\delta \in \Delta$, we define the supreme norm and the L_2 norm induced by Lebesgue measure by $\|f\|_{\infty,\delta} = \sup\{|f(x)| : x \in \delta\}$ and $|||f|||_\delta^* = (\int_\delta f^2\,dx)^{1/2}$ for a function f on δ.

CONDITION A.1. There are absolute constants γ_1 and γ_2 that do not depend on n such that

$$\gamma_1 |||f|||_\delta^{*2} \le |||f|||_\delta^2 \le \gamma_2 |||f|||_\delta^{*2}, \qquad \delta \in \Delta, f \in \mathbb{G} = \mathbb{G}_n.$$

We now state some regularity conditions on $\mathbb{G}$. The first set of conditions is on the partition Δ. Define the diameter of a set δ to be $\operatorname{diam}(\delta) = \sup\{|x_1 - x_2| : x_1, x_2 \in \delta\}$.

CONDITION A.2. (i) Given any distinct $\delta, \delta' \in \Delta$, the closures of δ and δ' are disjoint or intersect in a common vertex, edge, face and so on (no mixture allowed).

(ii) There is a constant $\gamma_3 > 0$ (independent of n) such that the ratio of the sizes of inscribed and circumscribed balls of each $\delta \in \Delta$ is bounded from below by γ_3.

(iii) The partition is quasi-uniform; that is, there is a constant $\gamma_4 < \infty$ (independent of n) such that

$$\frac{\max\{\operatorname{diam}(\delta) : \delta \in \Delta\}}{\min\{\operatorname{diam}(\delta) : \delta \in \Delta\}} \le \gamma_4.$$

These mild conditions are commonly used in the literature. For the univariate case, Condition A.2 reduces to the requirement of bounded mesh ratio, which was used in Douglas, Dupont and Wahlbin (1975). Set $h = h(\Delta) = \max\{\operatorname{diam}(\delta) : \delta \in \Delta\}$, which is usually called the (maximal) mesh size of Δ in the approximation theory literature. Under Condition A.2, h can be used as a universal measure of size for elements of Δ.

For $\delta \in \Delta$ and a function f on $\mathcal{X}$, we say that f is *active* on δ if it is not identically zero on the interior of δ. The following condition says that there is a locally supported basis of $\mathbb{G}$ and the basis has some special properties.

CONDITION A.3. There is a basis $\{B_i\}$ of $\mathbb{G}$ satisfying the following requirements.

(i) For each basis function B_i, the union of the elements of Δ on which B_i is active is a connected set. In addition, there is a constant $\gamma_5 < \infty$ (independent of n) such that, for each B_i, the number of elements of Δ on which B_i is active is bounded by γ_5.

(ii) Let I_δ denotes the collection of indices i whose corresponding basis function B_i is active on δ. There are positive constants γ_6 and γ_7 (independent of n) such that

$$\gamma_6 h^d \sum_{i \in I_\delta} \alpha_i^2 \le \left|\left|\left| \sum_{i \in I_\delta} \alpha_i B_i \right|\right|\right|_\delta^{*2} \le \gamma_7 h^d \sum_{i \in I_\delta} \alpha_i^2, \qquad \delta \in \Delta.$$

This condition is satisfied for commonly used finite element spaces [see Chapter 2 of Oswald (1994)]. For a univariate spline space, it is sufficient to use a B-spline basis to satisfy this condition [see Section 4.4 of DeVore and Lorentz (1993)]. Tensor products of B-splines can be used for a tensor product spline space.

Condition A.1 implies that the norms $|||\cdot||| = |||\cdot|||_n$ are equivalent to $|||\cdot|||^*$; that is,

$$\gamma_1 |||f|||^{*2} \le |||f|||^2 \le \gamma_2 |||f|||^{*2}, \qquad f \in \mathbb{G} = \mathbb{G}_n.$$

It follows from Condition A.1 and Condition A.3(ii) that

$$\gamma_1\gamma_6 h^d \sum_{i\in I_\delta} \alpha_i^2 \le \Big|\Big|\Big| \sum_{i\in I_\delta} \alpha_i B_i \Big|\Big|\Big|_\delta^2 \le \gamma_2\gamma_7 h^d \sum_{i\in I_\delta} \alpha_i^2, \qquad \delta \in \Delta. \tag{A.1}$$

This, together with Condition A.3(i), implies that

$$\gamma_1\gamma_6 h^d \sum_{i} \alpha_i^2 \le \Big|\Big|\Big| \sum_{i} \alpha_i B_i \Big|\Big|\Big|^2 \le \gamma_2\gamma_5\gamma_7 h^d \sum_{i} \alpha_i^2, \qquad \delta \in \Delta. \tag{A.2}$$

The following theorem is the main result of this appendix.

THEOREM A.1. *Suppose Conditions* A.1–A.3 *hold. Then there is a constant C that depends on $\gamma_1, \dots, \gamma_7$ but not on n such that $\|Pu\|_\infty \le C\|u\|_\infty$ for any function u on $\mathcal{X}$.*

The proof of this theorem, which extends the ideas of Douglas, Dupont and Wahlbin (1975), will be given shortly. One can also use the result of Descloux (1972) on finite element matrices to establish a similar result.

In application of the above result to polynomial spline regression (Theorem 5.1, Section 5), $\nu = \nu_n$ is chosen to be the empirical measure. Recall that the empirical and theoretical norms are defined by $\|f\|_n^2 = E_n[f^2(X)]$ and $\|f\|^2 = E[f^2(X)]$. Define their local versions by $\|f\|_{n,\delta}^2 = E_n[f^2(X)\mathbb{1}_\delta(X)]$ and $\|f\|_\delta^2 = E[f^2(X)\mathbb{1}_\delta(X)]$ for $\delta \in \Delta$. The following result is from Huang (1999).

LEMMA A.1. *Suppose Condition* A.2 *is satisfied and that the density p_X of X is bounded away from zero and infinity on $\mathcal{X}$. If* $\lim_n \dim(\mathbb{G}) \log n/n = 0$, *then* $\sup_{\delta\in\Delta} \sup_{g\in\mathbb{G}} |\|g\|_{n,\delta}/\|g\|_\delta - 1| = o_P(1)$.

If the design density is bounded away from zero and infinity, then the theoretical norm is equivalent to the L_2 norm induced by Lebesgue measure and so are their local versions. Thus, if the conclusion of the above lemma holds, then Condition A.1 is satisfied with $\nu = \nu_n$ being the empirical measure, except on an event whose probability tends to 0 as $n \to \infty$. Let Q denote the empirical projection onto $\mathbb{G}$. The following is a direct consequence of Theorem A.1.

COROLLARY A.1. *Suppose Conditions* A.2 *and* A.3 *hold, that the density of X is bounded away from zero and infinity, and that* $\lim_n \dim(\mathbb{G}) \log n/n = 0$. *Then there is an absolute constant C such that $\|Qf\|_\infty \le C\|f\|_\infty$ except on an event whose probability tends to zero as $n \to \infty$.*

In the proof of Theorem A.1 we need a distance measure between elements of Δ defined as follows. Given $\delta, \delta' \in \Delta$, set $d(\delta,\delta) = 0$ and for a positive integer l, set $d(\delta,\delta') = l$ if there is a connected path of line segments $V_1 V_2 \cdots V_l$ with V_1 a vertex of δ, V_l a vertex of δ' and $V_i V_{i+1}$ an edge of some $\delta \in \Delta$ for $1 \le i < l$, and no shorter such path. Note that $d(\delta,\delta')$ is symmetric in δ and δ'. Under Condition A.2, there is a constant $C = C(\gamma_3,\gamma_4)$ such that $\#\{\delta' : d(\delta',\delta) = l\} \le Cl^d$ for $\delta \in \Delta$. For sequences of numbers a_n and b_n, let $a_n \lesssim b_n$ mean that $a_n \le Cb_n$ for some constant C which does not depend on n but may depend on the constants $\gamma_1,\dots,\gamma_7$ above.

PROOF OF THEOREM A.1. Write $u = \sum_{\delta'} u_{\delta'}$, where $u_{\delta'}(x) = u(x) \times \operatorname{ind}(x \in \delta')$ for $\delta' \in \Delta$. Note that the supreme norm and the L_2 norm are equivalent on a space of polynomials of bounded degree in a bounded region of $\mathbb{R}^d$. Since Pu is a polynomial on $\delta \in \Delta$, by affine transforming each δ to a set circumscribing the unit ball we obtain that $\|Pu\|_{\infty,\delta} \le Ch^{-d/2}|||Pu|||_\delta^*$ for $\delta \in \Delta$, where the constant C can be chosen independently of n by Condition A.2. Thus it follows from Condition A.1 that

$$\|Pu\|_{\infty,\delta} \lesssim h^{-d/2}|||Pu|||_\delta \le h^{-d/2}\sum_{\delta'} |||Pu_{\delta'}|||_\delta, \qquad \delta \in \Delta. \tag{A.3}$$

We need the following lemma, whose proof will be given shortly.

LEMMA A.2. *Suppose Conditions* A.2 *and* A.3 *hold. There is a constant $\lambda \in (0,1)$ that depends on $\gamma_1,\dots,\gamma_7$ but not on n such that, for any $\delta_0 \in \Delta$ and any function v supported on δ_0 [i.e., $v(x) = 0$ for $x \notin \delta_0$],*

$$\sum_{\delta : d(\delta,\delta_0)=l} |||Pv|||_\delta^2 \le \lambda^{l-1}|||Pv|||^2, \qquad l \ge 1.$$

It follows from Conditions A.1 and A.2 that the ν-measure $\nu(\delta)$ of δ satisfies $\nu(\delta) \lesssim h^{d/2}$ for $\delta \in \Delta$. Since P is an orthogonal projection,

$$|||Pu_{\delta'}||| \le |||u_{\delta'}||| \le (\nu(\delta'))^{1/2}\|u_{\delta'}\|_\infty \lesssim h^{d/2}\|u_{\delta'}\|_\infty \le h^{d/2}\|u\|_\infty.$$

By Lemma A.2, for $\delta \in \Delta$ such that $d(\delta,\delta') = l \ge 1$,

$$|||Pu_{\delta'}|||_\delta^2 \le \sum_{\delta'' : d(\delta'',\delta')=l} |||Pu_{\delta'}|||_{\delta''}^2 \le \lambda^{l-1}|||Pu_{\delta'}|||^2.$$

Moreover, $|||Pu_\delta|||_\delta \le |||Pu_\delta|||$. Hence, by (A.3),

$$\begin{aligned}\|Pu\|_{\infty,\delta} &\lesssim h^{-d/2}|||Pu_\delta|||_\delta + h^{-d/2}\sum_{\delta':\,\delta'\neq\delta}|||Pu_{\delta'}|||_\delta\\ &\lesssim \|u\|_\infty + h^{-d/2}\sum_{l\ge 1}\#\{\delta':d(\delta',\delta)=l\}\lambda^{(l-1)/2}h^{d/2}\|u\|_\infty.\end{aligned}$$

Under Condition A.2, $\#\{\delta':d(\delta',\delta)=l\}\le Cl^d$ for $l\ge 1$. Consequently,

$$\|Pu\|_{\infty,\delta}\lesssim \|u\|_\infty + \|u\|_\infty\sum_{l\ge 1}\lambda^{(l-1)/2}l^d \lesssim \|u\|_\infty. \qquad \square$$

PROOF OF LEMMA A.2. Write $Pv=\sum_i \alpha_i B_i$, where $\{B_i\}$ is the locally supported basis specified in Condition A.3. For a nonnegative integer l, set

$$\tilde v_l = \sum_{i\in I_l}\alpha_i B_i, \tag{A.4}$$

where I_l denotes the collection of indices i whose corresponding basis function B_i is active on a $\delta\in\Delta$ such that $d(\delta,\delta_0)\le l$.

Since Pv is an orthogonal projection onto $\mathbb{G}$, $|||Pv-v|||^2\le|||\tilde v_l - v|||^2$. Note that $Pv=\tilde v_l$ on those $\delta\in\Delta$ with $d(\delta,\delta_0)\le l$. Moreover, $v=0$ on $\delta\neq\delta_0$. Hence

$$\sum_{\delta:\,d(\delta,\delta_0)>l}|||Pv|||_\delta^2 \le \sum_{\delta:\,d(\delta,\delta_0)>l}|||\tilde v_l|||_\delta^2. \tag{A.5}$$

On the other hand, let $\delta\in\Delta$ be such that $d(\delta,\delta_0)>l$ and suppose that B_i appears in the expansion (A.4) of $\tilde v_l$ and is active on δ. Then there is a $\delta'\in\Delta$ such that $d(\delta',\delta_0)=l$ and B_i is active on δ', since the union of the elements of Δ on which B_i is active is a connected set. Thus

$$\tilde v_l(x)=\sum_{i\in\tilde I_l}\alpha_i B_i(x),\qquad x\in\delta, d(\delta,\delta_0)>l,$$

where $\tilde I_l$ denotes the collection of indices i whose corresponding function B_i is active on a $\delta\in\Delta$ such that $d(\delta,\delta_0)=l$. Consequently,

$$\sum_{\delta:\,d(\delta,\delta_0)>l}|||\tilde v_l|||_\delta^2 = \sum_{\delta:\,d(\delta,\delta_0)>l}\left|\left|\left|\sum_{i\in\tilde I_l}\alpha_i B_i\right|\right|\right|_\delta^2 \le \left|\left|\left|\sum_{i\in\tilde I_l}\alpha_i B_i\right|\right|\right|^2. \tag{A.6}$$

Recall that I_δ is the collection of indices i whose corresponding basis function B_i is active on δ. It follows from (A.1) and (A.2) that

$$\begin{aligned}\left|\left|\left|\sum_{i\in\tilde I_l}\alpha_i B_i\right|\right|\right|^2 &\lesssim h^d\sum_{i\in\tilde I_l}\alpha_i^2\\ &\le h^d\sum_{\delta:\,d(\delta,\delta_0)=l}\sum_{i\in I_\delta}\alpha_i^2 \lesssim \sum_{\delta:\,d(\delta,\delta_0)=l}\left|\left|\left|\sum_{i\in I_\delta}\alpha_i B_i\right|\right|\right|_\delta^2.\end{aligned} \tag{A.7}$$

Combining (A.5)–(A.7), we obtain that

$$\sum_{\delta : d(\delta,\delta_0)>l} |||Pv|||_\delta^2 \lesssim \sum_{\delta : d(\delta,\delta_0)=l} \left|\left|\left| \sum_{i\in I_\delta} \alpha_i B_i \right|\right|\right|_\delta^2 . \tag{A.8}$$

Since $Pv = \tilde{v}_l = \sum_{i\in I_\delta} \alpha_i B_i$ on each δ with $d(\delta, \delta_0) = l$,

$$\sum_{\delta : d(\delta,\delta_0)=l} \left|\left|\left| \sum_{i\in I_\delta} \alpha_i B_i \right|\right|\right|_\delta^2 = \sum_{\delta : d(\delta,\delta_0)=l} |||Pv|||_\delta^2 . \tag{A.9}$$

Set $a_l = \sum_{\delta : d(\delta,\delta_0)=l} |||Pv|||_\delta^2$. Then it follows from (A.8) and (A.9) that $\sum_{k>l} a_k \le ca_l$, $l \ge 0$, for some constant c. Set $s_l = \sum_{k>l} a_k$. Then $s_l \le c(s_{l-1} - s_l)$, which implies that $s_l \le [c/(c+1)]^l s_0$ for $l \ge 0$. Hence

$$a_l \le s_{l-1} \le \left(\frac{c}{c+1}\right)^{l-1} s_0 \le \left(\frac{c}{c+1}\right)^{l-1} \sum_{l\ge 0} a_l, \qquad l \ge 1;$$

equivalently,

$$\sum_{\delta : d(\delta,\delta_0)=l} |||Pv|||_\delta^2 \le \left(\frac{c}{c+1}\right)^{l-1} |||Pv|||^2, \qquad l \ge 1.$$

The proof of Lemma A.2 is complete. □

Acknowledgments. The author thanks Chuck Stone for many helpful comments. I also wish thank Xiaotong Shen for helpful discussions. Comments from the Editor, Hans R. Künsch, an Associate Editor and two referees encouraged me to rethink some issues and resulted in the inclusion of Sections 5.3 and 7.

REFERENCES

BARROW, D. L. and SMITH, P. W. (1978). Asymptotic properties of best $L_2[0, 1]$ approximation by splines with variable knots. *Quart. Appl. Math.* **36** 293–304.

CHEN, Z. (1991). Interaction spline models and their convergence rates. *Ann. Statist.* **19** 1855–1868.

CHUI, C. K. (1988). *Multivariate Splines.* SIAM, Philadelphia.

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

DE BOOR, C., HÖLLIG, K. and RIEMENSCHNEIDER, S. (1993). *Box Splines.* Springer, New York.

DESCLOUX, J. (1972). On finite element matrices. *SIAM J. Numer. Anal.* **9** 260–265.

DEVORE, R. A. and LORENTZ, G. G. (1993). *Constructive Approximation.* Springer, Berlin.

DOUGLAS, J., DUPONT, T. and WAHLBIN, L. (1975). Optimal L_∞ error estimates for Galerkin approximations to solutions of two-point boundary value problems. *Math. Comp.* **29** 475–483.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

GASSER, T., SROKA, L. and JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633.

HALL, P., KAY, J. W. and TITTERINGTON, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528.

HANSEN, M. (1994). Extended linear models, multivariate splines, and ANOVA. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

HANSEN, M., KOOPERBERG, C. and SARDY, S. (1998). Triogram models. *J. Amer. Statist. Assoc.* **93** 101–119.

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press.

HART, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer, New York.

HUANG, J. Z. (1998a). Projection estimation for multiple regression with application to functional ANOVA models. *Ann. Statist.* **26** 242–272.

HUANG, J. Z. (1998b). Functional ANOVA models for generalized regression. *J. Multivariate Anal.* **67** 49–71.

HUANG, J. Z. (1999). Asymptotics for polynomial spline regression under weak conditions. Unpublished manuscript.

HUANG, J. Z. (2001). Concave extended linear modeling: A theoretical synthesis. *Statist. Sinica* **11** 173–197.

HUANG, J. Z., KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (2000). Functional ANOVA modeling for proportional hazards regression. *Ann. Statist.* **28** 961–999.

HUANG, J. Z. and STONE, C. J. (1998). The L_2 rate of convergence for event history regression with time-dependent covariates. *Scand. J. Statist.* **25** 603–620.

HUANG, J. Z., WU, C. O. and ZHOU, L. (2000). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica*. To appear.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995a). The L_2 rate of convergence for hazard regression. *Scand. J. Statist.* **22** 143–157.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995b). Rate of convergence for logspline spectral density estimation. *J. Time Ser. Anal.* **16** 389–401.

LEHMANN, E. L. (1999). *Elements of Large-Sample Theory*. Springer, New York.

LEHMANN, E. L. and LOH, W.-Y. (1990). Pointwise versus uniform robustness of some large-sample tests and confidence intervals. *Scand. J. Statist.* **17** 177–187.

OSWALD, P. (1994). *Multilevel Finite Element Approximations*: *Theory and Applications.* Teubner, Stuttgart.

PETROV, V. V. (1975). *Sums of Independent Random Variables*. Springer, New York.

RUPPERT, D. and WAND, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22** 1346–1370.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230.

SCHUMAKER, L. (1981). *Spline Functions*: *Basic Theory*. Wiley, New York.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

STONE, C. J. (1989). Uniform error bounds involving logspline models. In *Probability, Statistics and Mathematics*: *Papers in Honor of Samuel Karlin* (T. W. Anderson, K. B. Athreya and D. L. Iglehart, eds.) 335–355. Academic Press, New York.

STONE, C. J. (1990). Large-sample inference for logspline models. *Ann. Statist.* **18** 717–741.

STONE, C. J. (1991). Asymptotics for doubly flexible logspline response models. *Ann. Statist.* **19** 1832–1854.

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184.

STONE, C. J., HANSEN, M., KOOPERBERG, C. and TRUONG, Y. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25** 1371–1470.

SZEGÖ, G. (1975). *Orthogonal Polynomials*, 4th ed. Amer. Math. Soc., Providence, RI.

ZHOU, S., SHEN, X. and WOLFE, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.* **26** 1760–1782.

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6340
E-MAIL: jianhua@wharton.upenn.edu