The Annals of Statistics
1988, Vol. 16, No. 2, 713–732

APPROXIMATION OF LEAST SQUARES REGRESSION ON NESTED SUBSPACES[1]

BY DENNIS D. COX

University of Illinois

For a regression model $y_i = \theta(x_i) + \varepsilon_i$, the unknown function θ is estimated by least squares on a subspace $\Lambda_m = \text{span}\{\psi_1, \psi, \dots, \psi_m\}$, where the basis functions ψ_i are predetermined and m is varied. Assuming that the design is suitably approximated by an asymptotic design measure, a general method is presented for approximating the bias and variance in a scale of Hilbertian norms natural to the problem. The general theory is illustrated with two examples: truncated Fourier series regression and polynomial regression. For these examples, we give rates of convergence of derivative estimates in (weighted) L_2 norms and establish consistency in supremum norm.

1. Introduction. Suppose we observe

$$y_{ni} = \theta_*(x_{ni}) + \varepsilon_{ni}, \qquad i \le i \le n, \tag{1.1}$$

where $x_{n1}, x_{n2}, \dots, x_{nn}$ are known d-dimensional design points, $\varepsilon_{n1}, \varepsilon_{n2}, \dots, \varepsilon_{nn}$ are uncorrelated mean zero random q-dimensional errors and θ_* is an unknown regression function mapping $\mathbb{R}^d \to \mathbb{R}^q$, which is to be estimated. Perhaps the most common approach to estimating θ_* in such a setup is to apply "parametric" least squares with several models from a predetermined class and then somehow select a model from the class. For instance, one may successively regress on linear, quadratic, cubic, etc., polynomials and then use model selection criterion functions and/or diagnostics to select which degree of polynomial to use in the end. We refer to this as variable degree polynomial regression. Of course, the true regression function is probably not a polynomial of any degree. While one is naturally interested in the behavior of such a procedure, there seem to be no general answers to such basic questions as "is the method consistent and if so, what is the rate of convergence?" It is hoped that this article will help to close that gap between theory and practice. We present here a general purpose mathematical machinery for answering such questions and apply it to obtain what are apparently the first results on consistency with rates for variable degree polynomial regression.

Many readers will recognize this as a nonparametric regression problem and may perhaps advocate the use of kernels, smoothing splines or some other standard nonparametric regression method. Our aim here is not to suggest a new methodology but rather to analyze an old and widely used one. However, variable degree polynomial regression turns out to have some advantages as a

---

Received October 1986; revised June 1987.

[1]Work supported by National Science Foundation Grant DMS-86-03083.

AMS 1980 *subject classifications*. Primary 62J05; secondary 62F12, 41A10.

Key words and phrases. Regression, nonparametric regression, bias approximation, polynomial regression, model selection, rates of convergence, orthogonal polynomials.

nonparametric regression method. For one, since the model dimension (the degree of the polynomial $+1$) will usually be much smaller than the sample size, the estimate is "simpler" than the original data set. Kernel and smoothing spline estimates are at least as complex as the data set. Other advantages of variable degree polynomial regression will emerge such as improved rates of convergence with the existence of more derivatives of θ_*. No boundary conditions are required for these improved rates, either.

The general framework will now be described. Consider nested models given as follows. Let $\{\psi_1, \psi_2, \dots\}$ be a sequence of functions and put

$$\Lambda_m = \operatorname{span}\{\psi_1, \dots, \psi_m\}.$$

Let θ_{mn} be the least squares estimate of θ_* obtained by regressing the data on Λ_m. For the example of variable degree polynomial regression with $d = 1$, of course one may use $\psi_\nu(x) = x^{(\nu-1)}$, and θ_{mn} is the least squares polynomial of degree $m - 1$. Our goal here is to analyze $E\|\theta_* - \theta_{mn}\|^2$ as $n \to \infty$ and (possibly) $m \to \infty$, where $\|\cdot\|$ denotes a function space norm. It is specifically allowed that θ_* may not belong to any of the parametric models Λ_m.

For many analyses, one can be very vague about the asymptotic behavior of the design. Huber (1973) is an example. Note that there is no consideration of bias in that paper, and only local results are given. In order to analyze the bias and to obtain global results (i.e., bounds on the norm of the error), it is necessary to impose more structure on the problem. The additional structure needed is obtained by positing an asymptotic design. Let

$$P_X^{(n)}(A) := n^{-1} \sum_{i=1}^{n} 1_A(x_{ni})$$

be the design measure, i.e., the "empirical" distribution of $x_{n1}, x_{n2}, \dots, x_{nn}$. (Here, $A \subset \mathbb{R}^d$ is a Borel set and 1_A its indicator.) We will assume that $P_X^{(n)}$ tends to some Borel probability measure P_X as $n \to \infty$ in a sense specified in Assumption 2. P_X is called the *asymptotic design measure*. We write F_n and F to denote the distribution functions of $P_x^{(n)}$ and P_x, respectively. The theory presented here will apply to many choices of the basis functions $\{\psi_\nu\}$ and asymptotic designs as illustrated by the examples given later on.

Given basis functions $\{\psi_\nu: \nu = 1, 2, \dots\}$ and an asymptotic design P_X, it is shown in Section 2 that there is an associated natural family of Hilbert norms $\|\cdot\|_\tau$, $\tau \in \mathbb{R}$, with inner products $\langle \cdot, \cdot \rangle_\tau$. This family is called a scale of norms since increasing the parameter τ increases the strength of the norm. In all the examples we have investigated, these norms for $\tau > 0$ are related to weighted L_2 norms on derivatives. While one may justifiably be interested in other norms (e.g., supremum norm), it is particularly easy to analyze the estimation error in these natural norms. Further, by relating these natural norms to a given norm, it is possible to obtain results about the given norm.

We now describe the general form of the results. Because the norms are Hilbertian, there is a decomposition of the mean squared norm error:

$$E\|\theta_* - \theta_{nm}\|_\tau^2 = \|\theta_* - E\theta_{nm}\|_\tau^2 + E\|\theta_{nm} - E\theta_{nm}\|_\tau^2.$$

The first term on the r.h.s. is called the bias squared and the second term the variance. The behavior of the bias depends on the true θ_* and can be characterized to a certain extent by the norms $\|\cdot\|_\rho$ in the scale for which $\|\theta_*\|_\rho < \infty$ (see Theorem 2.2). Note that the scale of norms is used not only to measure the estimation error, but also to measure properties of the true regression function θ_* which determine the rate of convergence of the bias. When the norms are used in this latter role, we will typically designate them with ρ. For both bias and variance, we give results in Section 2 which show how to approximate them by "continuous" analogs (determined by P_X rather than the discrete measure $P_X^{(n)}$). The continuous analogs turn out to be easier to analyze than their discrete counterparts. The approximations will be valid uniformly in the model dimension m provided $m \le M_n$, where $M_n \to \infty$, but not too fast. The rate at which M_n may tend to ∞ depends on the structure of the particular limiting problem and the rate at which $P_X^{(n)}$ approximates P_X.

The proofs of the claims made in the following examples are given in Section 3.

EXAMPLE 1. Let $q = d = 1$ and assume the error variance is constant. Let the basis functions be given by

$$\psi_\nu(x) = x^{(\nu-1)}, \qquad \nu = 1, 2, \dots,$$

i.e., variable degree polynomial regression. Suppose that P_X is a beta distribution with parameters a, $b > 0$, i.e., P_X has density

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \qquad 0 < x < 1.$$

Assume also that

$$\sup_{0<x<1} |F_n(x) - F(x)| = O(n^{-p}) \tag{1.2}$$

for some $p > 0$, and necessarily $p \le 1$. Let

$$h := \max\{a, b, 1/2\}.$$

Assume that M_n satisfies

$$M_n = O(n^{2p/(4h+1)-\varepsilon})$$

for some $\varepsilon > 0$ (we may use $\varepsilon = 0$ if $h = 1/2$ and replace the O with o). If τ is a nonnegative even integer, then we can identify the $\|\cdot\|_\tau$ norms, viz.,

$$\|\theta\|_{2k}^2 \approx \int_0^1 \left([\theta(x)]^2 + [D^k\theta(x)]^2[x(1-x)]^k\right) f(x)\,dx, \tag{1.3}$$

where D is the differentiation operator. Here, the symbol $\approx$ means that the r.h.s. of (1.3) can be bounded above and below by constant multiples of the l.h.s. Then for the variance we have

$$E\|\theta_{nm} - E\theta_{nm}\|_{2k}^2 = O(n^{-1}m^{(2k+1)}). \tag{1.4}$$

For the bias, assume θ_* is $(r-1)$ times differentiable with $D^{(r-1)}\theta_*$ absolutely continuous and

$$\int_0^1 [D^r\theta_*(x)]^2[x(1-x)]^r f(x)\,dx < \infty,$$

i.e., $\|\theta_*\|_{2r} < \infty$. Then if k is an integer with $0 \le k < r$,

$$\|\theta_* - E\theta_{nm}\|_{2k}^2 = O(m^{-2(r-k)}). \tag{1.5}$$

If $m \approx n^{1/(2r+1)}$, then the bounds on variance and bias are balanced and one achieves

$$E\|\theta_* - \theta_{nm}\|_{2k}^2 = O(n^{-2(r-k)/(2r+1)}).$$

If θ_* is infinitely differentiable, then we may take r arbitrarily large and achieve a rate of convergence $O(n^{-1+\varepsilon})$ for any $\varepsilon > 0$. The practical utility of this result is unclear. Note that no boundary conditions are required of θ_*. This is similar to the case of regression on splines [see Agarwal and Studden (1980)] although there it is probably not possible to obtain rates of convergence approaching n^{-1} unless the order of the spline is increased with sample size.

EXAMPLE 2. Take $q = d = 1$ and assume again that the errors ε_{ni} have constant variance. Let the basis functions be given by

$$\psi_\nu(x) = \begin{cases} 1, & \text{if } \nu = 1, \\ \cos(2\pi kx), & \text{if } \nu = 2k \ge 2 \text{ is even}, \\ \sin(2\pi kx), & \text{if } \nu = 2k+1 \ge 3 \text{ is odd}. \end{cases}$$

Let the design be given by

$$x_{ni} := (i - 1/2)/n, \qquad 1 \le i \le n,$$

so the asymptotic design is uniform on $[0,1]$. Now let $M_n \to \infty$ in such a way that

$$M_n = o(n^{4/5}). \tag{1.6}$$

Again we can identify the natural scale of norms for nonnegative even integers, viz.,

$$\|\theta\|_{2k}^2 \approx \int_0^1 \left([\theta(x)]^2 + [D^k\theta(x)]^2\right) dx. \tag{1.7}$$

Indeed, for this example one can identify the natural scale of norms as weighted l_2 norms on the Fourier coefficients with weights being a power of the index of the coefficient (which is more or less how the scale of norms is constructed in general). Then the variance behaves as

$$E\|\theta_{nm} - E\theta_{nm}\|_{2k}^2 = O(n^{-1}m^{(2k+1)}), \tag{1.8}$$

uniformly in $m \le M_n$. To describe the result on the bias, we suppose that for some $r \ge 1$, θ_* is $(r-1)$ times continuously differentiable, $D^{r-1}\theta_*$ is absolutely continuous and that

$$D^k\theta_*(0) = D^k\theta_*(1), \qquad 0 \le k < r, \tag{1.9}$$

$$\int_0^1 [D^r\theta_*(x)]^2\,dx < \infty. \tag{1.10}$$

Then we have for $0 \le p < r$,

$$\|\theta_* - E\theta_{nm}\|_{2k}^2 = O(m^{-2(r-k)}). \tag{1.11}$$

Eubank (1988) gives results on $E\|\theta_* - \theta_{nm}\|_0^2$, where $\|\cdot\|_0$ turns out to be ordinary $L_2[0,1]$ norm.

Note that in each case there are difficulties with the endpoints of the intervals. In Example 2, we force periodic boundary conditions on θ_* to get good rates of convergence on the bias. These constraints are not needed in Example 1, but the natural scale of norms downweights the endvalues with the factor $[x(1-x)]^k$. One seems to always encounter difficulties at the boundaries in nonparametric regression; see Rice and Rosenblatt (1983) or Cox (1988).

The natural norms can be used to get results about other norms such as sup norm. For Example 2, sup norm is weaker than $\|\cdot\|_2$ norm, so the bounds in (1.8) and (1.11) with $k = 1$ give upper bounds on rates of convergence in supremum norm. For Example 1, the argument used to derive (3.18) along with (3.12) shows that $\|\cdot\|_{2h+\varepsilon}$ norm is stronger than sup norm for any $\varepsilon > 0$, so if $k > h$, from (1.4), $E[\sup|\theta_{nm} - E\theta_{nm}|^2] = O(n^{-1}m^{(2k+1)})$, and from (1.5), $\sup|\theta_* - E\theta_{nm}|^2 = O(m^{-2(r-k)})$, where the suprema are over $[0,1]$. These bounds are probably not very good, but they do establish consistency in sup norm provided θ_* is sufficiently smooth and $m \to \infty$ slowly enough.

2. Statement of main results. In this section we present a number of notations, state the general assumptions and state the main results. Recall that P_X denotes the asymptotic design measure. The setup here is considerably more general than needed for the examples in Section 1, but we anticipate working out other applications that require this level of generality.

ASSUMPTION 1. (a) The $\mathbb{R}^q$ valued random variables $\varepsilon_{n1}, \varepsilon_{n2}, \dots, \varepsilon_{nn}$ are mean 0, uncorrelated, with the covariance matrix for ε_{ni} being $\sigma^2 V(x_{ni})$ for some constant $\sigma^2 > 0$. It is assumed that $V(x)$ is a known function which takes positive definite values. Let $K(x) := V^{-1}(x)$.

(b) The sequence of positive integers $\{M_n\}$ is such that the set of $n \times q$ matrices $\{(\psi_k(x_{n1}), \psi_k(x_{n2}), \dots, \psi_k(x_{nn}))': 1 \le k \le M_n\}$ is linearly independent for all n.

(c) Define the linear space

$$\Lambda_\infty := \operatorname{span}\{\psi_1, \psi_2, \dots\}.$$

The inner product,

$$\langle \theta, \zeta \rangle_0 := \int \theta(x)' K(x) \zeta(x) P_X(dx),$$

is well defined for all $\theta, \zeta \in \Lambda_\infty$, i.e.,

$$\|\theta\|_0^2 := \langle \theta, \theta \rangle_0 < \infty, \qquad \forall\, \theta \in \Lambda_\infty.$$

The reason for the subscript 0 on the inner product will appear shortly. Let $\varphi_1, \varphi_2, \dots$ denote a sequence in Λ_∞ such that

$$\operatorname{span}\{\varphi_1, \varphi_2, \dots, \varphi_m\} = \Lambda_m, \qquad \langle \varphi_i, \varphi_j \rangle_0 = \delta_{ij},$$

where δ_{ij} denotes the Kronecker delta. One may obtain the φ_i's by the Gram–Schmidt procedure applied to the ψ_i's. For any real number ρ, put

$$\langle\theta,\zeta\rangle_\rho := \sum_{\nu=1}^{\infty} \nu^\rho \langle\theta,\varphi_\nu\rangle_0 \langle\zeta,\varphi_\nu\rangle_0, \qquad \|\theta\|_\rho^2 := \langle\theta,\theta\rangle_\rho,$$

defined for $\theta, \zeta \in \Lambda_\infty$. Let $\Lambda(\rho)$ denote the completion of Λ_∞ in the norm $\|\cdot\|_\rho$. Then $\Lambda(\rho)$ is a Hilbert space under the preceding inner product. Note that $\{\varphi_1, \varphi_2, \dots\}$ is an orthonormal basis for $\Lambda(0)$. Define a discrete analog of the inner product $\langle\cdot,\cdot\rangle_0$ by

$$\begin{aligned}(\theta,\zeta)_n &:= \int \theta(x)' K(x) \zeta(x) P_X^{(n)}(dx) \\ &= n^{-1} \sum_{i=1}^{n} \theta(x_{ni})' K(x_{ni}) \zeta(x_{ni}),\end{aligned}$$

which is a semiinner product. Part (i) of Assumption 2 guarantees it is well defined on some $\Lambda(s) \supset \Lambda_\infty$.

ASSUMPTION 2. (i) There is an $s > 0$ such that the evaluation functional $\theta \mapsto \theta(x)$ is a bounded linear functional on $\Lambda(s)$ for all $x \in \Omega :=$ the support of P_X.

(ii) There exist J, a positive integer, $\rho_1, \dots, \rho_J$ and $\tau_1, \dots, \tau_J$ and $\{k_{ni}: 1 \le i \le J, 1 \le n\}$ such that $\forall\, \theta \in \Lambda(s)$ and $\forall\, \nu \ge 1$,

$$|\langle\theta,\varphi_\nu\rangle_0 - (\theta,\varphi_\nu)_n| \le \sum_{i=1}^{J} k_{ni} \nu^{\rho_i/2} \|\theta\|_{\tau_i} \tag{2.1}$$

and

$$\sum_{i=1}^{J} k_{ni} M_n^{(\rho_i+\tau_i+1)/2} \to 0 \quad \text{as } n \to \infty. \tag{2.2}$$

LEMMA 2.1. *A sufficient condition for* (i) *is that* $\Lambda_\infty \subset C_B(\Omega; \mathbb{R}^q)$ = *the set of all bounded continuous functions mapping* $\Omega \to \mathbb{R}^q$, *and there exists* $s > 0$ *such that*

$$\sum_{\nu=1}^{\infty} \nu^{-s} \|\varphi_\nu\|_{C_B}^2 < \infty, \tag{2.3}$$

where $\|\cdot\|_{C_B}$ *denotes the usual supremum norm on* C_B.

PROOF. For any $\theta \in \Lambda(s)$,

$$\begin{aligned}\|\theta(x)\|_{\mathbb{R}^q} &= \left\| \sum_{\nu=1}^{\infty} \langle\theta,\varphi_\nu\rangle_0 \varphi_\nu(x) \right\|_{\mathbb{R}^q} \le \sum_{\nu=1}^{\infty} |\langle\theta,\varphi_\nu\rangle_0| \, \|\varphi_\nu\|_{C_B} \\ &\le \left(\sum_{\nu=1}^{\infty} \nu^s \langle\theta,\varphi_\nu\rangle_0^2 \right)^{1/2} \left(\sum_{\nu=1}^{\infty} \nu^{-s} \|\varphi_\nu\|_{C_B}^2 \right)^{1/2} \le M \|\theta\|_s,\end{aligned}$$

where M is a finite constant. □

Part (i) of Assumption 2 is fairly mild, but (ii) is a more substantial requirement. Note that the l.h.s. of (2.1) is simply a numerical integration error and may be written as

$$\left|\int \theta(x)'K(x)\varphi_\nu(x)[P_X - P_X^{(n)}](dx)\right|. \tag{2.4}$$

Various techniques for obtaining inequalities like (2.1) are indicated in the next section.

Now we consider some linear operators used in the sequel. First, let Π_m: $\Lambda(0) \to \Lambda_m$ denote orthogonal projection, i.e.,

$$\Pi_m\theta := \sum_{\nu=1}^{m} \langle \theta, \varphi_\nu\rangle_0\varphi_\nu.$$

It is easy to check that Π_m is also the orthogonal projection of any $\Lambda(\rho)$ onto Λ_m. By Assumption 2, there is for each $x \in \Omega$ a $\xi(x) \in \Lambda(s)^q$ such that for $1 \le i \le q$ and $\forall\, \theta \in \Lambda(s)$,

$$\langle \xi_i(x), \theta\rangle_s = [\theta(x)]_i,$$

the ith component of the vector $\theta(x)$. Now let T_n: $\Lambda(s) \to Y_n$ be the design operator, i.e., the (i, j) entry of $T_n\theta$ is $[\theta(x_{ni})]_j$. Here, Y_n is the observation Hilbert space. Now $Y_n = \mathbb{R}^{n\times q}$ as sets but with a different inner product than usual, namely,

$$\langle u, v\rangle_{Y_n} := n^{-1}\sum_{i=1}^{n} u_i'K(x_{ni})v_i,$$

where u_i' is the ith row of the $n \times q$ matrix $u \in Y_n$. Then we may write the linear model as

$$y_n = T_n\theta + \varepsilon_n,$$

where

$$E\varepsilon_n = 0, \qquad E[\varepsilon_{ni}\varepsilon_{nj}'] = \sigma^2\delta_{ij}V(x_{ni}).$$

Put

$$T_{nm} := T_n|_{\Lambda_m},$$

the restriction of T_n to Λ_m. Let T_{nm}^*: $Y_n \to \Lambda_m(s)$ denote the adjoint of T_{nm}. Here, $\Lambda_m(s)$ denotes the finite-dimensional Hilbert space consisting of Λ_m equipped with inner product $\langle\,\cdot\,,\cdot\,\rangle_s$, and the adjoint operator is defined through the relation $\langle z, T_{nm}\theta\rangle_{Y_n} = \langle T_{nm}^* z\, \theta\rangle_s$, $\forall\, z \in Y_n$ and $\forall\, \theta \in \Lambda_m(s)$. Using this and properties of projections, one can show that

$$T_n^*z = n^{-1}\sum_{i=1}^{n}\xi(x_{ni})'K(x_{ni})z_i, \qquad T_{nm}^* = \Pi_m T_n^*.$$

Put $U_n := T_n^*T_n$: $\Lambda(s) \to \Lambda(s)$ and $U_{nm} := T_{nm}^*T_{nm} = \Pi_m U_n|_{\Lambda_m}$. Note that

$$U_n\theta = \int \xi(x)'K(x)\theta(x)P_X^{(n)}(dx),$$

so one can expect that it will be well approximated by the operator U given by

$$U\theta := \int \xi(x)'K(x)\theta(x)P_X(dx).$$

Also, a good approximation to U_{nm} is

$$U_{\infty m} := \Pi_m U|_{\Lambda_m}.$$

Define Π_{nm}: $\Lambda(s) \to \Lambda_m$ as the projection in the discrete norm, i.e., $\Pi_{nm}\theta$ is the element ζ of Λ_m which minimizes $\langle(\zeta - \theta), U_n(\zeta - \theta)\rangle_s$. Π_{nm} is a well defined operator by Assumption 1(b).

For Theorem 2.2, it is convenient to have the bias operator

$$B_{nm} := I - \Pi_{nm}.$$

Note that $\theta_* - E\theta_{nm} = B_{nm}\theta_*$, explaining the nomenclature "bias operator." One expects it to be well approximated by the limiting bias operator

$$B_m := I - \Pi_m.$$

To describe the nature of this approximation, we use the operator norms

$$\|A\|_{\rho,\tau} := \sup\{\|A\theta\|_\tau: \theta \in \Lambda(\rho),\ \|\theta\|_\rho = 1\},$$

where A: $\Lambda(\rho) \to \Lambda(\tau)$ is a bounded linear operator.

It will be convenient to have the following asymptotic notation. We write $a_n \cong b_n$ to mean $a_n - b_n = o(b_n)$. Now we state our main theorems. The proofs are given in Section 4.

THEOREM 2.2. *If $\rho > s$ and $\tau \le \rho$, then*

$$\|\Pi_{mn} - \Pi_m\|_{\rho,\tau} \le C\left(\sum_{i=1}^{J} k_{ni} m^{(\rho_i+\tau_i+1)/2}\right)\|B_m\|_{\rho,\tau},$$

for all $m \in \{1, 2, \ldots, M_n\}$ and $n = 1, 2, \ldots$, where the constant C depends only on ρ and τ.

It is easy to see that

$$\|B_m\|_{\rho,\tau} \approx m^{-(\rho-\tau)/2}, \tag{2.5}$$

since the supremum in the definition of $\|B_m\|_{\rho,\tau}$ is achieved at a multiple of φ_{m+1}.

The next result provides information on the expected mean square of the norm of the random part of the estimation error.

THEOREM 2.3. *If $\tau \ge 0$, then*

$$E\|U_{nm}^{-1}T_{nm}^*\varepsilon_n\|_\tau^2 \cong \sigma^2 n^{-1}\sum_{\nu=1}^{m}\nu^\tau \approx n^{-1}m^{\tau+1}, \tag{2.6}$$

uniformly in $m \in \{1, 2, \ldots, M_n\}$.

Note that the l.h.s. of (2.6) is the same as $E\|\theta_{nm} - E\theta_{nm}\|_\tau^2$.

3. Applications to examples. In this section, we prove the claims made about the examples of Section 1. In each case, it will be necessary to identify the functions $\{\varphi_\nu\}$ which are an orthonormal basis for $\Lambda(0)$, to identify the spaces $\Lambda(\rho)$ for $\rho > 0$ and then to verify Assumption 2. We treat Example 2 first because it is simpler.

PROOF OF EXAMPLE 2. Clearly $\varphi_1 = \psi_1$ and $\varphi_\nu = \sqrt{2}\psi_\nu$ for $\nu \geq 2$. For a function $\theta \in L_2[0,1]$, let

$$\hat{\theta}_\nu := \langle \theta, \varphi_\nu \rangle_0 \tag{3.1}$$

denote its Fourier coefficient. It is easy to see that $\theta \in \Lambda(2r)$ for some nonnegative integer r if and only if θ satisfies (1.9) and (1.10). Furthermore,

$$\int [D^k\theta(x)]^2\,dx = \sum_{r=1}^{\infty} (2\pi r)^{2k}\left(\hat{\theta}_{2r}^2 + \hat{\theta}_{2r+1}^2\right)$$

(all integrals in this proof are from 0 to 1 unless otherwise indicated). We will show the lower bound in (1.7), the upper bound being similar but easier. Now

$$\begin{aligned}
\|\theta\|_{2k}^2 &= \left(\int \theta(x)\,dx\right)^2 + \sum_{r=1}^{\infty}\left((2r)^{2k}\hat{\theta}_{2r}^2 + (2r+1)^{2k}\hat{\theta}_{2r+1}^2\right) \\
&\geq \left(\int \theta(x)\,dx\right)^2 + \left(\frac{2}{3\pi}\right)^{2k}\sum_{r=1}^{\infty}(2\pi r)^{2k}\left(\hat{\theta}_{2r}^2 + \hat{\theta}_{2r+1}^2\right) \\
&\geq \left(\frac{2}{3\pi}\right)^{2k}\frac{(2\pi)^{2k}-1}{(2\pi)^{2k}}\left\{\int \theta^2(x)\,dx + \sum_{r=1}^{\infty}(2\pi r)^{2k}\left(\hat{\theta}_{2r}^2 + \hat{\theta}_{2r+1}^2\right)\right\}.
\end{aligned}$$

From Lemma 2.1, Assumption 2(i) holds if $s > 1$. To verify Assumption 2(ii), we apply integration by parts:

$$\begin{aligned}
&\int \varphi_\nu(x)\theta(x)\left[P_X^{(n)}(dx) - P_X(dx)\right] \\
&\quad = -\int [F_n(x) - F(x)]\,D[\varphi_\nu\theta](x)\,dx \\
&\quad = \int G_n(x) D^2[\varphi_\nu\theta](x)\,dx,
\end{aligned} \tag{3.2}$$

where

$$G_n(x) := \int_0^x [F_n(\xi) - F(\xi)]\,d\xi.$$

Hence, from (3.2) we have

$$\begin{aligned}
&|\langle \varphi_\nu, \theta\rangle_0 - (\varphi_\nu, \theta)_n| \\
&\quad \leq \left(\sup_{0\leq x\leq 1}|G_n(x)|\right)\int\left(\left[D^2\varphi_\nu(x)\theta(x) + 2D\varphi_\nu(x)D\theta(x)\right.\right. \\
&\qquad\qquad \left.\left. + \varphi_\nu(x)D^2\theta(x)\right]\right)dx.
\end{aligned} \tag{3.3}$$

The first factor on the r.h.s. is $1/8n^2$. If Cauchy–Schwarz is applied to each of the three integrals and our identification of the $\|\cdot\|_{2r}$ norms is used, there results

$$(3.4) \qquad |\langle \varphi_\nu, \theta\rangle_0 - (\varphi_\nu, \theta)_n| \le Cn^{-2}\left[\nu^2\|\theta\|_0 + \nu\|\theta\|_2 + \|\theta\|_4\right]$$

for some constant C. Thus, Assumption 2(ii) holds with $J = 3$, $k_{ni} \approx n^{-2}$ and $\rho_i + \tau_i = 4$. The claims made for this example in Section 1 now follow from Theorems 2.2 and 2.3. □

REMARK 3.1. If one is willing to weaken the bound on M_n in (1.6), then more generality can be obtained for the finite sample design. Assume P_X is uniform on $[0, 1]$, but only that

$$(3.5) \qquad \sup_{0\le x\le 1} |F_n(x) - F(x)| = O(n^{-p})$$

for some $p \in (0, 1]$. This allows for designs other than $x_{ni} = (i - 1/2)/n$. If one stops after the first integration by parts in (3.2) and applies the same argument used to derive (3.3), there results

$$(3.6) \qquad |\langle \varphi_\nu, \theta\rangle_0 - (\varphi_\nu, \theta)_n| \le Cn^{-p}\left[\nu\|\theta\|_0 + \|\theta\|_2\right]$$

for some constant C. When this is used in Assumption 2(ii) along with the main theorems, one obtains the conclusions (1.8) and (1.11), but with the proviso

$$M_n = o(n^{2p/3}).$$

For example, if $x_{ni} = i/n$, then we need $M_n = o(n^{2/3})$ rather than (1.6) when $x_{ni} = (i - 1/2)/n$. It is perhaps surprising that such a small change in the design could make much difference, especially since the i/n design can be shifted to give $(i - 1/2)/n$. It makes sense that regression operators based on the continuous uniform measure will give a better approximation to operators based on finite designs when the finite designs are placed symmetrically in the interval.

PROOF OF EXAMPLE 1. The functions $\{\varphi_\nu\colon \nu = 1, 2, \ldots\}$ are polynomials orthonormal w.r.t. the weight function given by the beta(a, b) density. Thus they are given by

$$\varphi_\nu(x) = c_{ab}(\nu) P_{\nu-1}^{(a-1, b-1)}(1 - 2x),$$

where $P_n^{(\alpha,\beta)}$, α, $\beta > -1$, $n = 0, 1, 2, \ldots$, denote the Jacobi polynomials defined in Abramowitz and Stegun (1964), Askey (1975) and Szegö (1975), abbreviated hereafter as (AS), (A) and (Sz), respectively. The normalizing constants c_{ab} are given by

$$(3.7) \qquad c_{ab}^2(\nu) := \frac{(2\nu + a + b - 3)\Gamma(\nu + a + b - 2)\Gamma(\nu)\Gamma(a)\Gamma(b)}{\Gamma(\nu + a - 1)\Gamma(\nu + b - 1)\Gamma(a + b)},$$

where the factor $(2\nu + a + b - 3)\Gamma(\nu + a + b - 2)$ must be replaced by $\Gamma(a + b)$

when $\nu = 1$. See (22.2.1) of (AS), (2.9) of (A) or (4.3.3) of (Sz). Stirling's formula can be used to show that

$$c_{ab}(\nu) \approx \nu^{1/2}. \tag{3.8}$$

Now we identify the norms $\|\cdot\|_\tau$ associated to these basis functions. One can usually anticipate that the norms will have something to do with weighted L_2 norms on derivatives, so we seek a weight function $u(x)$ on the interval $(0,1)$ so that $\{D^k\varphi_\nu: \nu = 1,2,\dots\}$ are orthogonal w.r.t. u. One can show (see the Appendix) that

$$\int D^k\varphi_\nu(x) D^k\varphi_\mu(x)[x(1-x)]^k f(x)\,dx = 0, \qquad \nu \neq \mu, \tag{3.9}$$

so $u(x) := [x(1-x)]^k f(x)$ is the appropriate weight function. To relate this to a specific $\|\cdot\|_\rho$ norm, we need

$$\begin{aligned}&\int [D^k\varphi_\nu(x)]^2[x(1-x)]^k f(x)\,dx\\ &\quad= \frac{\Gamma(\nu+k+a+b-2)\Gamma(\nu)\Gamma(a)\Gamma(b)}{\Gamma(\nu+a+b-2)\Gamma(\nu-k)\Gamma(a+b)}, \quad \text{if } 1 \le k < \nu.\end{aligned} \tag{3.10}$$

See the Appendix. Of course $D^k\varphi_\nu = 0$ when $k \ge \nu$. One can show

$$\int [D^k\varphi_\nu(x)]^2[x(1-x)]^k f(x)\,dx \approx \nu^{2k}, \qquad \nu > k, \tag{3.11}$$

and hence (1.3) follows in a manner similar to (1.7) in the previous proof.

Next we turn to verification of Assumption 2. Put

$$h := \max\{a, b, 1/2\}.$$

It follows from known results on the maximum of Jacobi polynomials on $[-1,1]$ [see (22.14.1) of (AS) or (7.32.2) of (Sz)] and (3.8) that

$$\|\varphi_\nu\|_{C_B} \approx \nu^{h-1/2}. \tag{3.12}$$

Hence by Lemma 2.1, any $s > 2h$ will work for part (i) of Assumption 2.

Part (ii) of Assumption 2 is naturally more involved. The basic idea is similar to the derivation of (3.6), but it will be necessary to do some surgery near the endpoints because $u(x) \to 0$ as $x \to 0, 1$. We consider four cases.

CASE 1. Both $a, b > 1/2$. Let $\{\delta_{n0}\}$ and $\{\delta_{n1}\}$ be sequences with values in $[0, 1/2)$ to be determined. In the integral version of the l.h.s. of (2.1) [i.e., the l.h.s. of (3.2)] break the interval $(0,1)$ into three regions, viz.,

$$\begin{aligned}&\int_0^1 \theta(x)\varphi_\nu(x)[P_X^{(n)}(dx) - P_X(dx)]\\ &\quad= \left(\int_0^{\delta_{n0}} + \int_{1-\delta_{n1}}^1 + \int_{\delta_{n0}}^{1-\delta_{n1}}\right)\theta(x)\varphi_\nu(x)[P_X^{(n)}(dx) - P_X(dx)]\\ &\quad:= I_0 + I_1 + I_2, \quad \text{say.}\end{aligned} \tag{3.13}$$

Now the last term can be handled by integration by parts, viz.,

$$I_2 = [F_n(1-\delta_{n1}) - F(1-\delta_{n1})]\theta(1-\delta_{n1})\varphi_\nu(1-\delta_{n1}) + [F(\delta_{n0}) - F_n(\delta_{n0})]\theta(\delta_{n0})\varphi_\nu(\delta_{n0}) + \int_{\delta_{n0}}^{1-\delta_{n1}}[F(x) - F_n(x)]D[\theta\varphi_\nu](x)\,dx := I_{21} + I_{20} + I_{22}, \text{ say.} \tag{3.14}$$

Now,

$$|I_{22}| \le \left(\sup_{\delta_{n0}\le x\le 1-\delta_{n1}} \frac{|F_n(x) - F(x)|}{[x(1-x)]^{1/2}f(x)}\right) \times \int_0^1 [|D\varphi_\nu(x)|\,|\theta(x)| + |\varphi_\nu(x)|\,|D\theta(x)|][x(1-x)]^{1/2}f(x)\,dx \le Cn^{-p}\max\{\delta_{n0}^{-(a-1/2)}, \delta_{n1}^{-(b-1/2)}\}[\nu\|\theta\|_0 + \|\theta\|_2] \tag{3.15}$$

for some constant C. In this last expression, the first factor results from the hypothesized bound (1.2) on $|F_n - F|$, the max results from the denominator of the sup in the previous line and the quantity in brackets is obtained by Cauchy–Schwarz and (3.11) similarly to (3.6). Next, we have

$$|I_0| + |I_{20}| \le C[n^{-p} + F(\delta_{n0})]\left(\sup_{0<x\le\delta_{n0}} |\varphi_\nu(x)|\right) \times \left(\sup_{0<x\le\delta_{n0}} |\theta(x)|\right). \tag{3.16}$$

The argument on pages 169–170 of (Sz) shows that

$$\sup_{0<x\le\delta_{n0}} |\varphi_\nu(x)| = |\varphi_\nu(0)| \approx \nu^{a-1/2} \tag{3.17}$$

and hence also

$$\sup_{0<x\le\delta_{n0}} |\theta(x)| \le \sum_{\mu=1}^{\infty} |\hat{\theta}_\mu| \sup_{0<x\le\delta_{n0}} |\varphi_\mu(x)| \le C\sum_\mu \mu^{(a-1/2)}|\hat{\theta}_\mu| \le C\left(\sum_\mu \mu^{-(1+\varepsilon)}\right)^{1/2}\left(\sum_\mu \mu^{(2a+\varepsilon)}\hat{\theta}_\mu^2\right)^{1/2} \approx \|\theta\|_{2a+\varepsilon} \tag{3.18}$$

for any $\varepsilon > 0$. Here, $\hat{\theta}_\mu$ is a generalized Fourier coefficient as in (3.1). An easy calculation shows

$$F(\delta_{n0}) \approx \delta_{n0}^a. \tag{3.19}$$

The same techniques produce analogous bounds for $|I_1| + |I_{21}|$. Putting together the results from (3.13)–(3.19), we obtain

$$|\langle\theta, \varphi_\nu\rangle_0 - (\theta, \varphi_\nu)_n| \le C\{n^{-p}\max\{\delta_{n0}^{-(a-1/2)}, \delta_{n1}^{-(b-1/2)}\}[\nu\|\theta\|_0 + \|\theta\|_2] + [n^{-p} + \delta_{n0}^a]\nu^{(2a-1)/2}\|\theta\|_{2a+\varepsilon} + [n^{-p} + \delta_{n1}^b]\nu^{(2b-1)/2}\|\theta\|_{2b+\varepsilon}\} \tag{3.20}$$

for some constant C, which is of the form required by Assumption 2(ii). In order to make these estimates usable, we need to specify the δ_{ni}'s. One can check that the best possible choices (obtained by solving a small linear programming problem involving each δ_{ni} separately of the form $\delta_{ni} = n^{-\kappa}$) are $\delta_{n0} = n^{-2p/(4a+5)}$ and $\delta_{n1} = n^{-2p/(4b+5)}$, which gives

$$k_{n1} = k_{n2} \approx n^{-7p/(4h+5)}, \qquad k_{n3} = n^{-4ap/(4a+5)}, \qquad k_{n4} = n^{-4bp/(4b+5)},$$
$$\rho_1 = 2, \qquad \rho_2 = 0, \qquad \rho_3 = 2a - 1, \qquad \rho_4 = 2b - 1,$$
$$\tau_1 = 0, \qquad \tau_2 = 2, \qquad \tau_3 = 2a + \varepsilon, \qquad \tau_4 = 2b + \varepsilon.$$

When these are plugged into Assumption 2 and Theorems 2.2 and 2.3, the claims in Example 1 for this case are obtained.

CASE 2. Both $a, b \le 1/2$. In this case, it is not necessary to cut away the ends of the intervals in (3.13) as the denominator of the sup in (3.15) is bounded away from 0. One obtains immediately, as in (3.15),

$$|\langle \theta, \varphi_\nu \rangle_0 - (\theta, \varphi_\nu)_n| \le Cn^{-p}[\nu\|\theta\|_0 + \|\theta\|_2], \tag{3.21}$$

which will give the desired results.

CASE 3. $a > 1/2$ but $b \le 1/2$. One proceeds as in Case 1, but does not cut away the end at $x = 1$ (i.e., set $\delta_{n1} = 0$). There will be no term with k_{n4} in the final bound.

CASE 4. $a \le 1/2$ but $b > 1/2$. Same argument as Case 3. □

REMARK 3.2. (a) One can use the special design $x_{ni} = (i - 1/2)/n$ as in Example 2 (which implies $a = b = 1$) and obtain some improvement on the rate at which $M_n \to \infty$.

(b) Polynomial regression with other asymptotic designs can also be treated, e.g., P_X Gaussian (which gives the φ_ν's as Hermite polynomials) or gamma (φ_ν's being Laguerre polynomials). A general theory would be desirable.

4. Proof of main theorems. In this section are given the proofs of the theorems stated in Section 2. Assumptions 1 and 2 are in force throughout this section.

PROOF OF THEOREM 2.2. Let $\theta \in \Lambda(\rho)$ and put $\tilde{\theta}_{nm} := \Pi_{nm}\theta$ and $\theta_m := \Pi_m\theta$. Since both $\tilde{\theta}_{nm}$ and θ_m are in Λ_m, we have

$$\|\tilde{\theta}_{nm} - \theta_m\|_\tau^2 = \sum_{\nu=1}^{m} \nu^\tau \langle \tilde{\theta}_{nm} - \theta_m, \varphi_\nu \rangle_0^2$$
$$= \sum \nu^\tau \left[\int (\tilde{\theta}_{nm} - \theta_m)' K \varphi_\nu \, dP_X \right]^2.$$

By the properties of projections, $\int \tilde{\theta}_{nm}' K \zeta \, dP_X^{(n)} = \int \theta' K \zeta \, dP_X^{(n)}$ and $\int \theta_m' K \zeta \, dP_X = \int \theta' K \zeta \, dP_X$ for all $\zeta \in \Lambda_m$, so for all $\nu \le m$,

$$\int (\tilde{\theta}_{nm} - \theta_m)' K \varphi_\nu \, dP_X = \int (\theta - \tilde{\theta}_{nm})' K \varphi_\nu [dP_X^{(n)} - dP_X].$$

Applying this and Assumption 2 to the formula for $\|\tilde{\theta}_{nm} - \theta_m\|_\tau^2$ yields

$$\begin{aligned} \|\tilde{\theta}_{nm} - \theta_m\|_\tau^2 &\le \sum_{\nu=1}^{m} \nu^\tau \left(\sum_{i=1}^{J} k_{ni} \nu^{\rho_i/2} \|\theta - \tilde{\theta}_{nm}\|_{\tau_i} \right)^2 \\ &= \sum_i \sum_j \sum_\nu \nu^{\tau + (\rho_i + \rho_j)/2} k_{ni} k_{nj} \|\theta - \tilde{\theta}_{nm}\|_{\tau_i} \|\theta - \tilde{\theta}_{nm}\|_{\tau_j} \\ &\le m^{\tau+1} \left(\sum_{i=1}^{J} k_{ni} m^{\rho_i/2} \|\theta - \tilde{\theta}_{nm}\|_{\tau_i} \right)^2 . \end{aligned} \tag{4.1}$$

Now we show that $\tilde{\theta}_{nm}$ may be replaced with θ_m in the last expression in (4.1). By Assumption 2(ii), there exists N_1 such that

$$k_{n1} M_n^{(\rho_1 + \tau_1 + 1)/2} \le 1/2, \qquad \forall\, n \ge N_1. \tag{4.2}$$

Set $\tau = \tau_1$ in (4.1) and use $\|\theta - \tilde{\theta}_{nm}\|_{\tau_1} \le \|\theta - \theta_m\|_{\tau_1} + \|\tilde{\theta}_{nm} - \theta_m\|_{\tau_1}$ to obtain

$$\begin{aligned} \|\tilde{\theta}_{nm} - \theta_m\|_{\tau_1} &\le (1/2)\left[\|\tilde{\theta}_{nm} - \theta_m\|_{\tau_1} + \|\theta - \theta_m\|_{\tau_1} \right] \\ &\quad + m^{(\tau_1+1)/2} \sum_{i=2}^{J} k_{ni} m^{\rho_i/2} \|\theta - \tilde{\theta}_{nm}\|_{\tau_i}. \end{aligned} \tag{4.3}$$

Move the term with the factor of $\|\tilde{\theta}_{nm} - \theta_m\|_{\tau_1}$ from the r.h.s. to the l.h.s. of (4.3) to obtain

$$\|\tilde{\theta}_{nm} - \theta_m\|_{\tau_1} \le \|\theta - \theta_m\|_{\tau_1} + 2m^{(\tau_1+1)/2} \sum_{i=2}^{J} k_{ni} m^{\rho_i/2} \|\theta - \tilde{\theta}_{nm}\|_{\tau_i} \tag{4.4}$$

for all $n \ge N_1$. Substituting this back into (4.1) and using (4.2) once more gives

$$\begin{aligned} \|\tilde{\theta}_{nm} - \theta_m\|_\tau &\le 2m^{(\tau+1)/2} k_{n1} m^{\rho_1/2} \|\theta - \theta_m\|_{\tau_1} \\ &\quad + 2m^{(\tau+1)/2} \sum_{i=2}^{J} k_{ni} m^{\rho_i/2} \|\theta - \tilde{\theta}_{nm}\|_{\tau_i}. \end{aligned} \tag{4.5}$$

Now this procedure can be iterated: Put $\tau = \tau_2$ in (4.5), assume N_2 is large enough that (4.2) holds, but with an extra factor of 2 on the l.h.s. and, after some algebra, obtain

$$\begin{aligned} \|\tilde{\theta}_{nm} - \theta_m\|_\tau &\le 4m^{(\tau+1)/2} \sum_{i=1}^{2} k_{ni} m^{\rho_i/2} \|\theta - \theta_m\|_{\tau_i} \\ &\quad + 4m^{(\tau+1)/2} \sum_{i=3}^{J} k_{ni} m^{\rho_i/2} \|\theta - \tilde{\theta}_{nm}\|_{\tau_i}. \end{aligned}$$

After $(J-2)$ more steps, one obtains

$$\|\tilde{\theta}_{nm} - \theta_m\|_\tau \le C m^{(\tau+1)/2} \sum_{i=1}^{J} k_{ni} m^{\rho_i/2} \|\theta - \theta_m\|_{\tau_i}, \tag{4.6}$$

where C does not depend on n, m or θ.

Note that

$$\begin{aligned} \|\theta - \theta_m\|_{\tau_i} &= \|(I - \Pi_m)\theta\|_{\tau_i} \le \|I - \Pi_m\|_{\rho, \tau_i} \|\theta\|_\rho \\ &= (m+1)^{-(\rho - \tau_i)/2} \|\theta\|_\rho. \end{aligned} \tag{4.7}$$

Note that the supremum defining $\|I - \Pi_m\|_{\rho,\tau}$ is achieved at a multiple of φ_{m+1}. Substitute this back into (4.6) to obtain

$$\|(\Pi_{nm} - \Pi_m)\theta\|_\tau \leq C\left(\sum_{i=1}^{J} k_{ni} m^{(\tau_i+\rho_i+1)/2}\right) m^{-(\rho-\tau)/2}\|\theta\|_\rho,$$

which proves the theorem. □

Proof of Theorem 2.3. We have

$$E\|U_{nm}^{-1}T_{nm}^{*}\varepsilon_n\|_\tau^2 = \sum_{\nu=1}^{m} \nu^\tau E\left[\langle U_{nm}^{-1}T_{nm}^{*}\varepsilon_n, \varphi_\nu\rangle_0^2\right]. \tag{4.8}$$

The $\langle \cdot, \cdot\rangle_0$ inner product in this last expression is equal to

$$\begin{aligned}\langle U_{nm}^{-1}T_{nm}^{*}\varepsilon_n, U_{\infty m}\varphi_\nu\rangle_{s,m} &= \langle \varepsilon_n, T_{nm}U_{nm}^{-1}U_{\infty m}\varphi_\nu\rangle_{Y_n}\\ &= n^{-1}\sum_{i=1}^{J}\left[U_{nm}^{-1}U_{\infty m}\varphi_\nu\right](x_{ni})'K(x_{ni})\varepsilon_{ni}.\end{aligned}$$

By the definition of $K(x_{ni})$ and Assumption 1(a) for $\nu \leq m$, the expected square of this last quantity is equal to

$$\begin{aligned}&E\left(n^{-2}\sum_i\sum_j\left[U_{nm}^{-1}U_{\infty m}\varphi_\nu\right](x_{ni})'K(x_{ni})\varepsilon_{ni}\varepsilon_{nj}'K(x_{nj})\left[U_{nm}^{-1}U_{\infty m}\varphi_\nu\right](x_{nj})\right)\\
&= \sigma^2 n^{-2}\sum_{i=1}^{n}\left[U_{nm}^{-1}U_{\infty m}\varphi_\nu\right](x_{ni})'K(x_{ni})\left[U_{nm}^{-1}U_{\infty m}\varphi_\nu\right](x_{ni})\\
&= \sigma^2 n^{-1}\langle T_{nm}U_{nm}^{-1}U_{\infty m}\varphi_\nu, T_{nm}U_{nm}^{-1}U_{\infty m}\varphi_\nu\rangle_{Y_n}\\
&= \sigma^2 n^{-1}\langle U_{nm}^{-1}U_{\infty m}\varphi_\nu, U_{nm}U_{nm}^{-1}U_{\infty m}\varphi_\nu\rangle_{s,m} = \sigma^2 n^{-1}\langle U_{\infty m}\varphi_\nu, U_{nm}^{-1}U_{\infty m}\varphi_\nu\rangle_s\\
&= \sigma^2 n^{-1}\langle \Pi_m U\varphi_\nu, U_{nm}^{-1}U_{\infty m}\varphi_\nu\rangle_s = \sigma^2 n^{-1}\langle \varphi_\nu, \Pi_m U_{nm}^{-1}U_{\infty m}\varphi_\nu\rangle_0\\
&= \sigma^2 n^{-1}\langle \varphi_\nu, U_{nm}^{-1}U_{\infty m}\varphi_\nu\rangle_0 = \sigma^2 n^{-1}\int \varphi_\nu(x)'K(x)\left[U_{nm}^{-1}U_{\infty m}\varphi_\nu\right](x)P_X(dx).\end{aligned}$$

When this is substituted into (4.8), there results

$$E\|U_{nm}^{-1}T_{nm}^{*}\varepsilon_n\|_\tau^2 = \sigma^2 n^{-1}\sum_{\nu=1}^{m}\nu^\tau\int\varphi_\nu(x)'K(x)\left[U_{nm}^{-1}U_{\infty m}\varphi_\nu\right](x)P_X(dx). \tag{4.9}$$

We will show that $U_{nm}^{-1}U_{\infty m} \to I$ as $n \to \infty$ in some sense, and hence that the integrals in this last expression may be replaced with 1.

For convenience, let

$$\begin{aligned}u_{nm}(\nu,\mu) &:= \langle U_{nm}^{-1}U_{\infty m}\varphi_\nu, \varphi_\mu\rangle_0 = u_{nm}(\mu,\nu), \qquad 1 \leq \nu, \mu \leq m,\\
S_{nmi}^2(\nu) &:= \sum_{\mu=1}^{m}\mu^{\tau_i}u_{nm}^2(\nu,\mu), \qquad 1 \leq \nu \leq m, 1 \leq i \leq J,\\
&= \|U_{nm}^{-1}U_{\infty m}\varphi_\nu\|_{\tau_i}^2.\end{aligned}$$

We will drop the subscripts n and m when there is no danger of confusion. Since $\langle \theta, \zeta \rangle_0 = \langle \theta, U_{\infty m}\zeta \rangle_s$ for $\theta, \xi \in \Lambda_m$, we have by Assumption 2,

$$\begin{aligned} |\delta_{\nu\mu} - u(\nu, \mu)| &= |\langle [U_{nm} - U_{\infty m}]\varphi_\nu, U_{nm}^{-1}U_{\infty m}\varphi_\mu \rangle_s| \\ &\le \sum_{i=1}^{J} k_{ni}\nu^{\rho_i/2}\|U_{nm}^{-1}U_{\infty m}\varphi_\mu\|_{\tau_i} \\ &= \sum_{i=1}^{J} k_{ni}\nu^{\rho_i/2}S_i(\mu), \end{aligned} \tag{4.10}$$

which implies

$$|u(\nu, \mu)| \le \delta_{\nu\mu} + \sum_{i=0}^{J} k_{ni}\nu^{\rho_i/2}S_i(\mu). \tag{4.11}$$

Square the last expression, use the inequality $(a+b)^2 \le 2a^2 + 2b^2$, multiply by ν^{τ_j} and sum from $\nu = 1$ to m to obtain

$$\begin{aligned} S_j^2(\mu) &\le 2\mu^{\tau_j} + 2\sum_{\nu=1}^{m} \nu^{\tau_j}\left(\sum_{i=1}^{J} k_{ni}\nu^{\rho_i/2}S_i(\mu)\right)^2 \\ &\le 2\mu^{\tau_j} + C^2 m^{\tau_j+1}\left(\sum_{i=1}^{J} k_{ni}m^{\rho_i/2}S_i(\mu)\right)^2, \end{aligned} \tag{4.12}$$

where the constant C depends only on the τ_j's and ρ_i's. To see the last inequality, expand the square in the previous expression into a double sum, bring the sum on ν inside and use an elementary estimate on the sum of the $(\tau_j + \rho_i + \rho_k)$-powers of the first m integers. Now take square roots of the extremes of (4.12), use $(a+b)^{1/2} \le a^{1/2} + b^{1/2}$ for $a, b > 0$, multiply by $k_{nj}m^{\rho_j/2}$, sum from $j = 1$ to J and obtain

$$\begin{aligned} \sum_{j=1}^{J} k_{nj}m^{\rho_j/2}S_j(\mu) &\le 2^{1/2}\sum_{j=1}^{J} k_{nj}m^{(\rho_j+\tau_j)/2} \\ &\quad + C\sum_{j=1}^{J} k_{nj}m^{(\rho_j+\tau_j+1)/2}\sum_{i=1}^{J} k_{ni}m^{\rho_i/2}S_i(\mu). \end{aligned} \tag{4.13}$$

Suppose by Assumption 2 that N is large enough that for all $n \ge N$,

$$C\sum_{j=1}^{J} k_{nj}m^{(\rho_j+\tau_j+1)/2} \le 1/2. \tag{4.14}$$

As for the derivation of (4.4), one obtains from (4.13)

$$\sum_{j=1}^{J} k_{nj}m^{\rho_j/2}S_j(\mu) \le 2^{3/2}\sum_{j=1}^{J} k_{nj}m^{(\rho_j+\tau_j)/2}, \qquad \forall\, n \ge N.$$

This last bound in conjunction with (4.10), (4.14) and the hypothesis $\tau \geq 0$, gives

$$\sum_{\nu=1}^{m} \nu^{\tau}|1 - u(\nu,\nu)| \leq 2^{3/2} \sum_{\nu=1}^{m} \nu^{\tau} \sum_{i=1}^{J} k_{ni} m^{(\rho_j+\tau_j)/2} \leq 2^{1/2} m^{-1} \sum_{\nu=1}^{m} \nu^{\tau}. \tag{4.15}$$

Now if we use this last estimate and (4.9), we obtain

$$\left| E\|U_{nm}^{-1} T_{nm}^{*} \varepsilon_n\|_\rho^2 - \sigma^2 n^{-1} \sum_{\nu=1}^{m} \nu^{\rho} \right| \leq \sigma^2 n^{-1} \sum_{\nu=1}^{m} \nu^{\rho} |1 - u(\nu,\nu)| = o\left(n^{-1} \sum_{\nu=1}^{m} \nu^{\rho} \right).$$

This completes the proof. □

REMARK 4.1. (a) We conjecture that it is possible to prove the last result with the weaker hypothesis

$$\sum_{i=1}^{J} k_{ni} m^{(\rho_i+\tau_i)/2} \to 0.$$

(b) In the last proof, (2.2) was needed only for $\theta \in \Lambda(m)$. This can be used to improve the rate $M_n \to \infty$ in some applications of Theorem 2.3. For instance, (3.18) can be replaced with

$$\sup_{0<x\leq\delta_{n0}} |\theta(x)| \leq C \sum_{\mu=1}^{m} |\theta_\mu| \mu^{(a-1/2)} \leq C m^{1/2} \left(\sum_{\mu=1}^{m} \mu^{(2a-1)} \hat{\theta}_\mu^2 \right)^{1/2} \leq C M_n^{1/2} \|\theta\|_{2a-1}. \tag{4.16}$$

A similar result holds near $x = 1$. When this is used in the proof of the claims in Example 1, one obtains that the variance approximation holds if $M_n = o(n^{-2p/(4h+1)})$.

APPENDIX

In this Appendix we give proofs of (3.9) and (3.10), which were needed in the proof of Example 1 in Section 3. Let

$$w(x) := (1-x)^{\alpha}(1+x)^{\beta}, \qquad -1 < x < +1,$$

where $\alpha, \beta > -1$ are fixed throughout this appendix. Let $P_n^{(\alpha,\beta)}(x)$ denote the Jacobi polynomial of nth degree as defined in (AS), (S), (Sz) or (A). These form an orthogonal sequence with respect to the weight function w. Let

$$Q_{n,p}(x) := (1-x^2)^{p/2} D^p P_n^{(\alpha,\beta)}(x),$$

which generalizes Ferrer's functions associated to Legendre polynomials [(AS) and (S)]. Note that $Q_{n,p} = 0$ if $p > n$. In this appendix, all integrals are from -1 to $+1$, unless otherwise stated.

LEMMA A.1. *Let m, n, p be any nonnegative integers. Then*

$$\int Q_{n,p}(x) Q_{m,p}(x) w(x)\, dx = 0 \quad \text{if } m \neq n, \tag{A.1}$$

and if $p \leq n$,

$$\int Q_{n,p}^2(x) w(x)\, dx = \frac{2^{\alpha+\beta+1}\Gamma(n+\alpha+\beta+p+1)\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{(2n+\alpha+\beta+1)\Gamma(n-p+1)\Gamma^2(n+\alpha+\beta+1)}. \tag{A.2}$$

PROOF. The proof basically follows (S), pages 246–248. Letting $u = P_n^{(\alpha,\beta)}$, u satisfies the differential equation (AS), (Sz) and (A)

$$D\left[(1-x^2)w(x)Du\right] + n(n+\alpha+\beta+1)w(x)u = 0 \tag{A.3}$$

or equivalently,

$$(1-x^2)D^2u + [\beta - \alpha - (\alpha+\beta+2)x]Du + n(n+\alpha+\beta+1)u = 0. \tag{A.4}$$

If we substitute

$$D^p x Du = xD^{p+1}u + pD^p u,$$

$$D^p(1-x^2)D^2u = (1-x^2)D^{p+2}u - 2pxD^{p+1}u - p(p-1)D^p u$$

(obtainable from Leibniz' formula) into (A.4) after applying D^p, we obtain

$$(1-x^2)D^{p+2}u + [\beta-\alpha-(\alpha+\beta+2p+2)x]D^{p+1}u + (n-p)(n+\alpha+\beta+p+1)D^p u = 0. \tag{A.5}$$

Using

$$D^p u = (1-x^2)^{-p/2} Q_{n,p},$$

one can obtain expressions for $D^{p+1}u$ and $D^{p+2}u$ which can be substituted into (A.5) to obtain

$$0 = D(1-x^2)w(x)DQ_{n,p} + \left\{n(n+\alpha+\beta+1) + p[(\beta-\alpha)x - (\alpha+\beta+p)] \times (1-x^2)^{-1}\right\} wQ_{n,p}. \tag{A.6}$$

To prove (A.1), multiply (A.6) by $Q_{m,p}$ and the corresponding equations for $Q_{m,p}$ by $Q_{n,p}$ and subtract to obtain

$$D\left\{(1-x^2)w\left[Q_{m,p}DQ_{n,p} - Q_{n,p}DQ_{m,p}\right]\right\} + (n-m)(n+m+\alpha+\beta+1)wQ_{n,p}Q_{m,p} = 0. \tag{A.7}$$

Integrate (A.7) and note that the integral of the first term vanishes to obtain (A.1). (Note that $n + m + \alpha + \beta + 1 > n + m - 1 > 0$ since $n \neq m$ and $n, m \geq 0$.)

To prove (A.2), we have from the definition of $Q_{n,p}$ that

$$\begin{aligned} Q_{n,p+1} &= (1 - x^2)^{(p+1)/2} D(1 - x^2)^{-p/2} Q_{n,p} \\ &= (1 - x^2)^{1/2} DQ_{n,p} + p(1 - x^2)^{-1/2} x Q_{n,p}. \end{aligned}$$

Square the last equation, multiply by $w(x)$ and integrate to obtain

$$\begin{aligned} \int Q_{n,p+1}^2(x) w(x)\, dx = {} & \left[Q_{n,p}(x) w(x)(1 - x^2) DQ_{n,p}(x)\right]_{x=-1}^{1} \\ & - \int Q_{n,p}(x) D\left[w(x)(1 - x^2) DQ_{n,p}(x)\right] dx \\ & + p\left[x w(x) Q_{n,p}^2(x)\right]_{x=-1}^{1} \\ & - p \int Q_{n,p}^2(x) D[x w(x)]\, dx \\ & + p^2 \int x^2 (1 - x^2)^{-1} Q_{n,p}(x) w(x)\, dx, \end{aligned} \tag{A.8}$$

where some obvious integrations by parts were applied. The first boundary term from the integration by parts vanishes since $w(x)(1 - x^2)$ vanishes at $x = \pm 1$. The second boundary term vanishes for the same reason when $p > 0$ [note that $Q_{n,p}^2$ contains a factor of $(1 - x^2)$] and trivially when $p = 0$. Use (A.6) on the first integral on the l.h.s. of (A.8) and apply some elementary calculus to the second and simplify and there results

$$\int Q_{n,p+1}^2(x) w(x)\, dx = (n - p)(n + p + \alpha + \beta + 1) \int Q_{n,p}^2(x) w(x)\, dx. \tag{A.9}$$

Thus,

$$\begin{aligned} \int Q_{n,p}^2(x) w(x)\, dx = {} & (n - p + 1)(n - p + 2) \cdots n \\ & \times (n + p + \alpha + \beta)(n + p - 1 + \alpha + \beta) \cdots \\ & \times (n + \alpha + \beta + 1) \int Q_{n,0}^2(x) w(x)\, dx. \end{aligned} \tag{A.10}$$

Now $Q_{n,0} = P_n^{(\alpha,\beta)}$ and

$$\int \left[P_n^{(\alpha,\beta)}(x)\right]^2 w(x)\, dx = \frac{2^{\alpha+\beta+1}\Gamma(n + \alpha + 1)\Gamma(n + \beta + 1)}{(2n + \alpha + \beta + 1)\Gamma(n + 1)\Gamma(n + \alpha + \beta + 1)};$$

see (22.2.1) of (AS), (2.9) of (A) or (4.3.3) of (Sz). When this is substituted into (A.9) then an equation equivalent to (A.2) is obtained. □

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. (1965). *Handbook of Mathematical Functions*. Dover, New York.

AGARWAL, G. and STUDDEN, W. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.* **8** 1307–1325.

ASKEY, R. (1975). *Orthogonal Polynomials and Special Functions*. SIAM, Philadelphia.

COX, D. (1988). Approximation of method of regularization estimators. *Ann. Statist.* **16** 694–712.

EUBANK, R. (1988). *Splines and Nonparametric Regression*. Dekker, New York.

HUBER, P. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821.

RICE, J. and ROSENBLATT, M. (1983). Smoothing splines: Regression, derivatives and deconvolution. *Ann. Statist.* **11** 141–156.

SANSONE, G. (1959). *Orthogonal Functions*. Wiley, London.

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

DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS
725 SOUTH WRIGHT STREET
CHAMPAIGN, ILLINOIS 61820