BULL. AUSTRAL. MATH. SOC. 65N30, 65L10
VOL. 15 (1976), 329-338.

On the condition number of certain Rayleigh-Ritz-Galerkin matrices

Bernard J. Omodei

Martin H. Schultz [*Bull. Amer. Math. Soc.* 76 (1970), 840-844] has investigated the spectral condition number of the Rayleigh-Ritz-Galerkin matrices that arise when normalized B-spline coordinate functions are used to approximate the solution of a class of linear, self-adjoint, elliptic boundary value problems in one dimension. This paper shows how results analogous to those of Schultz [*op. cit.*] can be established under weaker assumptions. We also extend the results to boundary value problems in higher dimensions.

We consider the following class of linear, self-adjoint, two-point boundary value problems:

$$(1) \quad L[u(x)] \equiv \sum_{j=0}^{n} (-1)^j D^j \left[p_j(x) D^j u(x)\right] = f(x) \ ,$$

$$0 < x < 1 \ , \quad f \in L^2[0, 1] \ , \quad n \geq 1 \ ,$$

with homogeneous Dirichlet boundary conditions

$$(2) \qquad D^k u(0) = D^k u(1) = 0 \ , \quad 0 \leq k \leq n-1 \ .$$

Assume that $p_j(x)$, $0 \leq j \leq n$, are real-valued bounded measurable functions on $[0, 1]$.

Let $W_0^{n,2}[0, 1]$ denote the completion of the set of all $C^\infty[0, 1]$ functions having compact support in $(0, 1)$, with respect to the Sobolev

Received 17 June 1976.

norm

$$\|w\|_{W^{n,2}} \equiv \left\{\sum_{j=0}^{n} \int_0^1 [D^j w(x)]^2 dx\right\}^{\frac{1}{2}} .$$

We assume that there exists a positive constant K such that for all $w \in W_0^{n,2}[0, 1]$,

(3) $$K\|w\|_{L^2}^2 \le \int_0^1 \left\{\sum_{j=0}^{n} p_j(x)\left[D^j w(x)\right]^2\right\}dx .$$

This assumption corresponds to the assumption that the differential operator L is positive definite. Schultz [7] made the stronger assumption that, for all $w \in W_0^{n,2}[0, 1]$,

$$K\|w\|_{W^{n,2}}^2 \le \int_0^1 \left\{\sum_{j=0}^{n} p_j(x)\left[D^j w(x)\right]^2\right\}dx .$$

It can be shown that the problem (1)-(2) has a unique generalized solution and that the Rayleigh-Ritz method is applicable; see Omodei [6]. Let $\{\phi_i(x)\}_{i=1}^m$ be m given linearly independent coordinate functions such that $\phi_i \in W_0^{n,2}[0, 1]$ for all $1 \le i \le m$. Let S_m denote the approximating subspace spanned by $\{\phi_i\}_{i=1}^m$. We claim, without giving the derivation, that the Rayleigh-Ritz-Galerkin matrix $R \equiv (r_{ik})$ for the problem (1)-(2) is given by

(4) $$r_{ik} = \int_0^1 \left\{\sum_{j=0}^{n} p_j(x) D^j\phi_k(x) D^j\phi_i(x)\right\}dx , \quad 1 \le i, k \le m .$$

We now introduce normalized B-spline coordinate functions. Following the construction of de Boor [1], for a positive integer d , the finite set of real numbers

$$\pi : 0 = x_0 < x_1 \le x_2 \le \ldots \le x_N < x_{N+1} = 1$$

is said to be a *$(d+1)$-extended partition* of $[0, 1]$, if and only if $x_k < x_{k+d}$ for all $0 \le k \le N - d + 1$; that is, if f_k denotes the

https://doi.org/10.1017/S0004972700022759 Published online by Cambridge University Press

multiplicity of the knot x_k in π, then $f_k \le d$ for all $1 \le k \le N$. Let $I \equiv \{0 \le k \le N \mid x_k < x_{k+1}\}$, and define

(5) $$\Delta \equiv \max_{k \in I} \left(x_{k+1}-x_k\right) \quad \text{and} \quad \delta \equiv \min_{k \in I} \left(x_{k+1}-x_k\right) .$$

Let $Sp_0(d, \pi)$ denote the *extended spline space* of all extended splines of degree d on π satisfying the boundary conditions (2); that is, $Sp_0(d, \pi)$ consists of those real-valued functions on $[0, 1]$ which satisfy the boundary conditions (2), reduce to a polynomial of degree less than or equal to d on $[x_k, x_{k+1}]$ for all $k \in I$, and have $d - f_k$ continuous derivatives in a neighbourhood of x_k for all $1 \le k \le N$.

Assuming that $n \le d$, we add $2(d-n)$ extra knots to π to form the partition

$$\tilde{\pi} : x_{-d+n} = \ldots = x_{-1} = x_0 < x_1 \le \ldots \le x_N < x_{N+1} = x_{N+2} = \ldots = x_{N+d+1-n} .$$

We now define the classical *B-splines* for the partition $\tilde{\pi}$ (see [4]):

$$M_k(x) \equiv (d+1)g\left(x_k, x_{k+1}, \ldots, x_{k+d+1}; x\right) , \quad -d + n \le k \le N - n ,$$

is $(d+1)$ times the $(d+1)$-th divided difference in y of the function $g(y; x) \equiv (y-x)_+^d$ based on the points $x_k, x_{k+1}, \ldots, x_{k+d+1}$. The *normalized B-splines* are defined by

(6) $$\psi_k(x) \equiv \frac{x_{k+d+1}-x_k}{d+1} M_k(x) , \quad -d + n \le k \le N - n .$$

It can be shown that $\{\psi_k(x)\}_{k=-d+n}^{N-n}$ form a basis for $Sp_0(d, \pi)$ (see [4]).

The following lemma is a simple consequence of a theorem in [2].

LEMMA 1. *For an arbitrary $(d+1)$-extended partition π, there exists a positive constant D depending on d but not on π such that*

(7) $$\left\| \sum_{k=-d+n}^{N-n} a_{k+d+1-n} \left(\frac{d+1}{x_{k+d+1}-x_k}\right)^{\frac{1}{2}} \psi_k \right\|_{L^2} \ge D\|a\|_2$$

for all $a \in R^{N+d+1-2n}$ *where* $\|a\|_2 \equiv \left(\sum_{i=1}^{N+d+1-2n} a_i^2\right)^{\frac{1}{2}}$.

We consider the case where the approximating subspace $S_m \equiv Sp_0(d, \pi)$, $m = N + d + 1 - 2n$, and the coordinate functions $\phi_i(x) \equiv \psi_{i+n-d-1}(x)$, $i = 1, 2, \ldots, m$. Assume that $f_k \le d + 1 - n$ for all $i \le k \le N$ to ensure that $Sp_0(d, \pi) \subset W_0^{n,2}[0, 1]$. The *spectral condition number* of the Rayleigh-Ritz-Galerkin matrix R is defined by

$$\kappa(R) \equiv \|R\|_2 \|R^{-1}\|_2 \quad \text{where} \quad \|R\|_2 \equiv \sup_{a \in R^m} \|Ra\|_2 / \|a\|_2 \ .$$

Using (3), it can easily be shown that R is positive definite and symmetric, and hence $\kappa(R) = \lambda^{-1}\Lambda$ where λ and Λ are the minimum and maximum eigenvalues, respectively, of R . The following theorem is analogous to that of Schultz [7].

THEOREM 1. *If* (3) *holds and* π *is an arbitrary* $(d+1)$*-extended partition of* $[0, 1]$ *such that* $f_k \le d + 1 - n$ *for all* $1 \le k \le N$ *, then there exists a positive constant* C *depending on* d *but not on* π *such that*

$$\kappa(R) \le C(\Delta/\delta)\delta^{-2n} \ . \tag{8}$$

Proof. From (4) and (3), we obtain for all $a \in R^m$,

$$a^T R a = \int_0^1 \left\{\sum_{j=0}^{n} p_j(x)\left[D^j \sum_{i=1}^{m} a_i \psi_{i+n-d-1}(x)\right]^2\right\}dx \ge K\left\|\sum_{i=1}^{m} a_i \psi_{i+n-d-1}\right\|_{L^2}^2$$

which, by Lemma 1, yields

$$\begin{aligned} a^T R a &\ge KD^2 \sum_{i=1}^{m} a_i^2 \frac{(x_{i+n} - x_{i+n-d-1})}{d+1} \\ &\ge KD^2 (d+1)^{-1} \delta \|a\|_2^2 \ , \end{aligned}$$

and thus

$$\lambda \ge KD^2(d+1)^{-1}\delta \ . \tag{9}$$

Conversely, since $p_j(x)$, $0 \le j \le n$, are bounded on $[0, 1]$, there exists a positive constant P such that, for all $\mathrm{a} \in R^m$,

$$\mathrm{a}^T R\mathrm{a} \le P \sum_{j=0}^{n} \int_0^1 \left[\sum_{i=1}^{m} a_i D^j \psi_{i+n-d-1}(x)\right]^2 dx$$

$$\le P \sum_{j=0}^{n} (2d+1) \sum_{i=1}^{m} a_i^2 \int_0^1 \left[D^j \psi_{i+n-d-1}(x)\right]^2 dx ,$$

since $\psi_{i+n-d-1}(x)$, $1 \le i \le m$, has support $\left[x_{i+n-d-1}, x_{i+n}\right]$. Thus

$$\mathrm{a}^T R\mathrm{a} \le P(2d+1) \sum_{i=1}^{m} a_i^2 \left(x_{i+n}-x_{i+n-d-1}\right) \sum_{j=0}^{n} \left\|D^j \psi_{i+n-d-1}\right\|_{L^\infty}^2 .$$

Using Lemma 3.1 of [3], it can be shown that there exists a positive constant E depending on d but not on π such that

$$\sum_{j=0}^{n} \left\|D^j \psi_{i+n-d-1}\right\|_{L^\infty}^2 \le E\delta^{-2n} \quad \text{for all} \quad 1 \le i \le m .$$

Hence

$$\mathrm{a}^T R\mathrm{a} \le PE(2d+1)(d+1)\Delta\delta^{-2n}\|\mathrm{a}\|_2^2 ,$$

and thus

$$\Lambda \le PE(2d+1)(d+1)\Delta\delta^{-2n} . \tag{10}$$

Combining (9) and (10), we obtain the desired result with $C = PE(2d+1)(d+1)^2 K^{-1} D^{-2}$. //

A corollary analogous to the Corollary of [7] is clearly valid.

Extension to higher dimensions

We consider the following class of linear, self-adjoint, boundary value problems defined on an M-dimensional hypercube $\Omega \equiv \mathop{\times}\limits_{j=1}^{M} [0, 1]$ with boundary $\partial\Omega$:

$$L[u(\mathrm{x})] = f(\mathrm{x}) , \quad \mathrm{x} \in \Omega , \quad f \in L^2(\Omega) , \tag{11}$$

https://doi.org/10.1017/S0004972700022759 Published online by Cambridge University Press

with homogeneous Dirichlet boundary conditions

(12) $$D^{\alpha}u(x) = 0 \ , \quad x \in \partial\Omega \ , \quad 0 \le |\alpha| \le n - 1 \ , \quad n \ge 1 \ ,$$

where the linear differential operator L is defined by

(13) $$L[u(x)] \equiv \sum_{0 \le |\alpha|, |\beta| \le n} (-1)^{|\alpha|} D^{\alpha}\left[p_{\alpha\beta}(x) D^{\beta} u(x)\right] .$$

We are using the usual multi-index notation, see [5]. Assume that all the coefficients $p_{\alpha\beta}(x)$ are bounded measurable functions in Ω and that $p_{\alpha\beta} = p_{\beta\alpha}$ for all $0 \le |\alpha|, |\beta| \le n$.

Let $W_0^{n,2}(\Omega)$ denote the completion of the set of all $C^{\infty}(\bar{\Omega})$ functions having compact support in Ω., with respect to the Sobolev norm

$$\|w\|_{W^{n,2}} \equiv \left\{\sum_{0 \le |\alpha| \le n} \int_{\Omega} [D^{\alpha}w(x)]^2 dx\right\}^{\frac{1}{2}} .$$

We assume that there exists a positive constant K such that for all $w \in W_0^{n,2}(\Omega)$,

(14) $$K\|w\|_{L^2}^2 \le \int_{\Omega} \left\{\sum_{0 \le |\alpha|, |\beta| \le n} p_{\alpha\beta}(x) D^{\alpha}w(x) D^{\beta}w(x)\right\} dx .$$

It can be shown that the problem (11)-(13) has a unique generalized solution and that the Rayleigh-Ritz method is applicable, see [6]. Let $\{\phi_i(x)\}_{i=1}^{m}$ be m linearly independent coordinate functions such that $\phi_i \in W_0^{n,2}(\Omega)$ for all $1 \le i \le m$. The Rayleigh-Ritz-Galerkin matrix $R \equiv (r_{ik})$ for the problem (11)-(13) is given by

(15) $$r_{ik} = \int_{\Omega} \left\{\sum_{0 \le |\alpha|, |\beta| \le n} p_{\alpha\beta}(x) D^{\alpha}\phi_k(x) D^{\beta}\phi_i(x)\right\} dx \ , \quad 1 \le i, k \le m .$$

For each j, $1 \le j \le M$, let π_j be a $(d+1)$-extended partition of $[0, 1]$ in the j-th dimension:

$$\pi_j : 0 = x_0^{(j)} < x_1^{(j)} \le x_2^{(j)} \le \ldots \le x_{N_j}^{(j)} < x_{N_j+1}^{(j)} = 1 ,$$

https://doi.org/10.1017/S0004972700022759 Published online by Cambridge University Press

and let Δ_j and δ_j be defined as in (5). Using expression (6), we construct the normalized B-spline basis $\left\{\psi_k\left(x^{(j)}\right)\right\}_{k=-d+n}^{N_j-n}$ for $Sp_0\left(d, \pi_j\right)$, $j = 1, 2, \ldots, M$. Let $\overline{\pi} \equiv \mathop{\times}\limits_{j=1}^{M} \pi_j$ be a $(d+1)$*-extended product partition* of Ω and let

$$\overline{\Delta} \equiv \max_{1 \le j \le M} \Delta_j \quad \text{and} \quad \overline{\delta} \equiv \min_{1 \le j \le M} \delta_j \ .$$

The *extended multivariate spline space* $Sp_0(d, \overline{\pi})$ is defined to be the tensor product $\bigotimes_{j=1}^{M} Sp_0\left(d, \pi_j\right)$. It can be shown that $Sp_0(d, \overline{\pi})$ is the linear span of all the *normalized multivariate* B*-splines*

$$\psi_{k_1}\left(x^{(1)}\right)\psi_{k_2}\left(x^{(2)}\right) \ldots \psi_{k_M}\left(x^{(M)}\right) , \quad -d + n \le k_j \le N_j - n ,$$
$$j = 1, 2, \ldots, M ,$$

which we rename as $\{B_i(\mathrm{x})\}_{i=1}^{m}$, where

$$\mathrm{x} = \left(x^{(1)}, x^{(2)}, \ldots, x^{(M)}\right) \quad \text{and} \quad m = \prod_{j=1}^{M} \left(N_j+d+1-2n\right) .$$

Using Lemma 1, it is straightforward to prove the following:

LEMMA 2. *For an arbitrary* $(d+1)$*-extended product partition* $\overline{\pi}$ *of* Ω *, there exists a positive constant* $\overline{D}$ *depending only on* d *such that for all* $\mathrm{a} \in R^m$,

$$\left\|\sum_{i=1}^{m} a_i B_i\right\|_{L^2} \ge \overline{D}\overline{\delta}^{M/2}\|\mathrm{a}\|_2 . \tag{16}$$

In applying the Rayleigh-Ritz method, let the approximating subspace $S_m \equiv Sp_0(d, \overline{\pi})$ and let the coordinate functions $\phi_i(\mathrm{x}) \equiv B_i(\mathrm{x})$, $i = 1, 2, \ldots, m$. Assuming that the maximum multiplicity of the interior knots of π_j is less than or equal to $d + 1 - n$, for all $1 \le j \le M$, then it can be shown that $Sp_0(d, \overline{\pi}) \subset W_0^{n,2}(\Omega)$, see [6].

THEOREM 2. *If* (14) *holds and* $\overline{\pi}$ *is an arbitrary* $(d+1)$*-extended product partition of* Ω *such that the multiplicity assumption above is valid, then there exists a positive constant* $\overline{C}$ *depending only on* d *such that*

(17)
$$\kappa(R) \le \overline{C}(\overline{\Delta}/\overline{\delta})^{M}\overline{\delta}^{-2n} .$$

Proof. From (15) and (14), we obtain for all $\mathrm{a} \in R^m$,

$$\begin{aligned}\mathrm{a}^T R\mathrm{a} &= \int_\Omega \left\{\sum_{0\le|\alpha|,|\beta|\le n} p_{\alpha\beta}(\mathrm{x})D^\alpha\left[\sum_{i=1}^m a_iB_i(\mathrm{x})\right]D^\beta\left[\sum_{i=1}^m a_iB_i(\mathrm{x})\right]\right\}d\mathrm{x} \\ &\ge K\left\|\sum_{i=1}^m a_iB_i\right\|_{L^2}^2 ,\end{aligned}$$

which, by Lemma 2, yields

$$\mathrm{a}^T R\mathrm{a} \ge K\overline{D}^2\overline{\delta}^M\|\mathrm{a}\|_2^2 ,$$

and thus

(18)
$$\lambda \ge K\overline{D}^2\overline{\delta}^M .$$

Conversely, since $p_{\alpha\beta}(\mathrm{x})$, $0 \le |\alpha|, |\beta| \le n$, are bounded in Ω, there exists a positive constant Q such that

$$\mathrm{a}^T R\mathrm{a} \le Q \sum_{0\le|\alpha|\le n} \int_\Omega \left[D^\alpha \sum_{i=1}^m a_iB_i(\mathrm{x})\right]^2 d\mathrm{x} ,$$

and using the minimal support properties of $\{B_i\}_{i=1}^m$, it can be shown that there exists a positive constant F depending only on d such that

$$\begin{aligned}\mathrm{a}^T R\mathrm{a} &\le QF \sum_{0\le|\alpha|\le n} \sum_{i=1}^m a_i^2 \int_\Omega \left[D^\alpha B_i(\mathrm{x})\right]^2 d\mathrm{x} \\ &\le QF \sum_{i=1}^m a_i^2(d+1)^M\overline{\Delta}^M \sum_{0\le|\alpha|\le n} \left\|D^\alpha B_i\right\|_{L^\infty}^2 .\end{aligned}$$

Using Lemma 3.1 of [3], it can be shown that there exists a positive constant $\overline{E}$ depending only on d such that

$$\sum_{0\le|\alpha|\le n} \left\|D^{\alpha}B_i\right\|^2_{L^\infty} \le \bar{E}\bar{\delta}^{-2n} \quad \text{for all} \quad 1 \le i \le m \, .$$

Hence

$$\mathrm{a}^T R \mathrm{a} \le Q F \bar{E}(d+1)^{M}\bar{\Delta}^{-M}\bar{\delta}^{-2n}\|\mathrm{a}\|_2^2 \, ,$$

and thus

$$\Lambda \le Q F \bar{E}(d+1)^{M}\bar{\Delta}^{-M}\bar{\delta}^{-2n} \, . \tag{19}$$

Combining (18) and (19), we obtain the desired result with

$\bar{C} = QF\bar{E}(d+1)^{M}K^{-1}\bar{D}^{-2}$. //

References

[1] Carl de Boor, "On uniform approximation by splines", *J. Approximation Theory* 1 (1968), 219-235.

[2] Carl de Boor, "The quasi-interpolant as a tool in elementary polynomial spline theory", *Approximation theory*, 269-276 (Proc. Internat. Sympos., Univ. Texas, Austin, Texas, 1973. Academic Press, New York, London, 1973).

[3] C. de Boor and G.J. Fix, "Spline approximation by quasiinterpolants", *J. Approximation Theory* 8 (1973), 19-45.

[4] H.B. Curry and I.J. Schoenberg, "On Pólya frequency functions IV: the fundamental spline functions and their limits", *J. Analyse Math.* 17 (1966), 71-107.

[5] С.Г. Михлин, Численная реализация вариационных методов (Izdat. "Nauka", Moscow, 1966).

S.G. Mikhlin, *The numerical performance of variational methods* (translated by R.S. Anderssen. Wolters-Noordhoff, Groningen, 1971).

[6] Bernard J. Omodei, "Stability of the Rayleigh-Ritz-Galerkin procedure for elliptic boundary value problems" (PhD thesis, Australian National University, Canberra, 1976). See also: Abstract, *Bull. Austral. Math. Soc.* 14 (1976), 471-472.

[7] Martin H. Schultz, "The condition number of a class of Rayleigh-Ritz-Galerkin matrices", *Bull. Amer. Math. Soc.* 14 (1970), 840-844.

Department of Mathematics,
University of Manchester,
Manchester,
England.

Present Address:
School of Mathematical Sciences,
Flinders University of South Australia,
Bedford Park,
South Australia.

https://doi.org/10.1017/S0004972700022759 Published online by Cambridge University Press