ON THE NUMERICAL PERFORMANCE OF SPECTRAL METHODS

R.S. Anderssen

ABSTRACT

In essence, the spectral method simply involves: (i) the choice of a more or less arbitrary orthonormal system ϕ_j, $j = 1,2,3,\ldots,$ to define approximations of the form

$$u_n = \sum_{j=1}^{n} a_j^{(n)} \phi_j$$

with unknown (constant) coefficients $a_j^{(n)}$, $j = 1,2,\ldots,n$; and (ii) the choice of n conditions which, in conjunction with the problem being solved, yield a non-singular matrix equation

$$L_n \underset{\sim}{a}^{(n)} = \underset{\sim}{f}^{(n)} , \qquad \underset{\sim}{a}^{(n)} = [a_1^{(n)}, a_2^{(n)}, \ldots, a_n^{(n)}]^T ,$$

for the $a_j^{(n)}$, $j = 1,2,\ldots, n$, where the structure of the matrix L_n and the right-hand-side vector $\underset{\sim}{f}^{(n)}$ will depend on the choice of the ϕ_j, $j = 1,2,\ldots,$ the problem being solved, and the n conditions.

Because of its success, it is often viewed as a standard "ansatz" for the numerical solution of ordinary and partial differential equations as well as integral equations. The key to this success is the choice of the ϕ_j, $j = 1,2,\ldots,$ to be an orthonormal system; not the n conditions of (ii). In this paper, we show how theory developed by Mikhlin for studying the numerical performance of variational methods can be used to identify to what extent an arbitrary choice of an orthonormal system can be justified numerically. In particular, we show for the Ritz-Galerkin and Bubnov-Galerkin counterparts of the spectral method that such an arbitrary choice guarantees numerical

stability, but not the convergence of the residual for ordinary and partial differential equations. The additional conditions necessary to guarantee the latter are also discussed.

§1. INTRODUCTION

Because the concept of a spectral method is quite general and has a natural variational interpretation, we develop its definition within the framework of linear operator equations

(1) $$\underset{\sim}{L}u = f \, , \quad \underset{\sim}{L} : \underline{\underline{D}}(\underset{\sim}{L}) \to \underline{\underline{R}}(\underset{\sim}{L}) \, , \; u = u(x), \; x \in \Omega \subset \mathbb{R}^q \, ,$$

where the domain and range of $\underset{\sim}{L}$, $\underline{\underline{D}}(\underset{\sim}{L})$ and $\underline{\underline{R}}(\underset{\sim}{L})$, are assumed to be dense in some Hilbert space $\underline{\underline{H}}$ with inner product $(\cdot,\cdot)$ and norm $\|\cdot\|$; and Ω is a bounded region in q-dimensional Euclidean space $\mathbb{R}^q$.

Knowledge of the concept of the energy space $\underline{\underline{H}}_{\underset{\sim}{A}}$ associated with a selfadjoint and positive definite operator $\underset{\sim}{A}$ will be assumed (cf.Mikhlin [13], §3). In this paper, $\underset{\sim}{A}$ will always denote a selfadjoint and positive definite operator.

Computationally, the starting point for *spectral methods* is the decision to use approximations of the form

(2) $$u_n(x) = \sum_{j=1}^{n} a_j^{(n)} \phi_j(x) \, ,$$

where the (*coordinate, basis, trial, shape*) *functions* $\phi_j(x)$, $j = 1,2,\ldots,n$, are chosen to be the first n elements of an orthonormal system ϕ_j, $j = 1,2,\ldots$, in $\underline{\underline{H}}$. Clearly, the qualifier "spectral" identifies this particular choice for the coordinate functions. These methods are subclassified in terms of the procedure used to determine the unknowns $a_j^{(n)}$, $j = 1,2,\ldots,n$; i.e. in terms of the n conditions which, in conjunction with (1), yield a non-singular matrix equation

(3) $$L_n \underset{\sim}{a}^{(n)} = \underset{\sim}{f}^{(n)} \, , \quad \underset{\sim}{a}^{(n)} = [a_1^{(n)}, a_2^{(n)}, \ldots, a_n^{(n)}]^T \, ,$$

for the determination of the $a_j^{(n)}$, $j = 1,2,\dots,n$.

In this paper, we limit attention to

1.1 The Ritz-Galerkin (Spectral) Methods

This class corresponds to the situation where $\underset{\sim}{L} = \underset{\sim}{A}$ (cf. Mikhlin [13],§3) and the n conditions are defined by the projection of the residual

(4) $$r_{\underset{\sim}{A}}(u_n) = \underset{\sim}{A}\, u_n - f$$

onto the zero element of $\underline{\underline{H}}_A^{(n)} = \mathrm{span}(\phi_1,\phi_2,\dots,\phi_n)$; viz.

(5) $$(r_{\underset{\sim}{A}}(u_n),\phi_j) = 0\ , \qquad j = 1,2,\dots,n\ .$$

In this situation, L_n becomes the Ritz-matrix

(6) $$R_n = \begin{pmatrix} (\underset{\sim}{A}\,\phi_1,\phi_1) & \dots & (\underset{\sim}{A}\,\phi_n,\phi_1) \\ \dots & \dots & \dots \\ (\underset{\sim}{A}\,\phi_1,\phi_n) & \dots & (\underset{\sim}{A}\,\phi_n,\phi_n) \end{pmatrix}$$

and

(7) $$\underset{\sim}{f}^{(n)} = [(f,\phi_1),(f,\phi_2),\dots,(f,\phi_n)]^T\ .$$

The qualifier "spectral" is invoked when the ϕ_j, $j = 1,2,\dots,$ form an orthonormal system.

1.2 The Bubnov-Galerkin (Spectral) Method

This class corresponds to the situation where

(8) $$\underset{\sim}{L} = \underset{\sim}{A} + \underset{\sim}{B}\ , \quad \underset{\sim}{A}^{-1}\,\underset{\sim}{B} \text{ compact}\ ,$$

and the n conditions are defined by the projection of the residual

(9) $$r_{\underset{\sim}{L}}(u) = \underset{\sim}{L}\, u_n - f$$

onto the zero element of $\underline{\underline{H}}_{\underset{\sim}{A}}^{(n)}$; viz.

(10) $$(r_{\underset{\sim}{L}}(u),\phi_j) = 0\ , \qquad j = 1,2,\dots,n\ .$$

The qualifier "spectral" is invoked when the ϕ_j, $j = 1,2,\dots$ form an orthonormal system.

The rationale for this subclassification is the way in which the theoretical results are usually derived for variational and projection methods (cf. Mikhlin [14]). They are first established for positive definite operators and then extended to linear operator equations of the form (8) by exploiting the underlying second kind integral equation structure.

The motivation for the use of spectral methods is two-fold. (1) The existence of extensive mathematical properties for particular orthonormal systems, such as the Legendre and Chebyshev polynomials, which can be exploited in various ways to manipulate the structure of numerical methods based on the use of orthonormal functions (cf. Delves and Freeman [4]). (2) The knowledge that, in the numerical performance of variational methods, the choice of the coordinate functions $\phi_j(x)$, $j = 1,2,\ldots,n$, appears to play a more crucial role than the n conditions defining (3); and thereby, the heuristic conclusion that in some sense an orthonormal system must be better than a non-orthonormal.

Though the success of spectral methods for the approximate solution of a wide class of practical problems (cf. Gottlieb and Orszag [8] , Hussaini et al. [9] and Peyret and Taylor [17] yields verification for this conclusion, it is well known (cf. Gottlieb and Orszag [8] and Anderssen and Omodei [1]) that the choice of an orthonormal system does not guarantee unconditionally that a spectral method will perform well computationally.

The aim of this paper is to give a more definite characterization of the numerical performance of spectral methods for stationary (i.e. time independent) problems than is contained in the standard texts on the subject (cf. Peyret and Taylor [17] and Fletcher [5]). In particular, we show how theory developed by Mikhlin [14]) for studying the numerical performance of variational methods can be used to identify to what extent an arbitrary choice of an orthonormal system can be justified numerically.

After developing appropriate preliminaries in §2 concerning minimal systems, similar operators and comparison theorems, we discuss in §3 conditions under which an arbitrary choice of an orthonormal system in $\underline{\underline{H}}$ guarantees numerical stability for the Ritz-Galerkin and Bubnov-Galerkin procedures. The fact that such an arbitrary choice does not guarantee the convergence of the residuals of $r_{\underset{\sim}{A}}(u)$ and $r_{\underset{\sim}{L}}(u)$ of (4) and (9) is pursued in §4. In addition, conditions are examined which do in fact guarantee their convergence. Some concluding remarks about time-dependent and eigenvalue problems as well as other aspects are made in §5.

§2. PRELIMINARIES: MINIMAL SYSTEMS, SIMILAR OPERATORS AND COMPARISON THEOREMS

As we shall see in §§3 and 4, the key to the present analysis is in the comparison theorems for systems $\{\phi_j\}_1^\infty = \{\phi_j,\ j = 1,2\ldots\}$, which lie simultaneously in two Hilbert spaces $\underline{\underline{H}}_1$ and $\underline{\underline{H}}_2$. Given that the coordinate system is orthonormal in $\underline{\underline{H}}_2$ and that, in some appropriate sense, $\underline{\underline{H}}_1$ is imbedded in $\underline{\underline{H}}_2$, a comparison theorem determines the properties of the coordinate system in $\underline{\underline{H}}_1$.

For the systems we use the properties of minimal systems, while the imbedding is accomplished using similar operators. The key concepts are:

Minimal Systems

A system $\{\phi_j\}_1^\infty$ which spans $\underline{\underline{H}}$ is said to be *minimal* in $\underline{\underline{H}}$, if the deletion of any element from the system restricts the span of the remaining elements to a proper subspace of $\underline{\underline{H}}$; and *non-minimal* otherwise.

Consider the Gram matix G_n of the first n elements $\{\phi_j\}_1^n$ of the system $\{\phi_j\}_1^\infty$:

$$G_n = \begin{pmatrix} (\phi_1,\phi_1) & \cdots & (\phi_n,\phi_1) \\ \cdots & \cdots & \cdots \\ (\phi_1,\phi_n) & \cdots & (\phi_n,\phi_n) \end{pmatrix} . \tag{11}$$

Because G_n is Hermitian and positive definite, its eigenvalues are positive and can be written in increasing order as

$$0 < \lambda_1^{(n)} \leq \lambda_2^{(n)} \leq \ldots \leq \lambda_n^{(n)} . \tag{12}$$

The interlacing consequences of the minimax principles for such eigenvalues (viz. for all m and n, m ≤ n, $\lambda_m^{(n+1)} \leq \lambda_m^{(n)} \leq \lambda_{m+1}^{(n+1)}$) imply that $\lambda_1^{(n)}$ and $\lambda_n^{(n)}$ can only decrease and increase, respectively, as n increases. As a consequence concepts such as strong minimality and almost orthonormality are important computationally because they potentially limit the growth of the spectral condition number of $G_n, K(G_n) = \lambda_n^{(n)}/\lambda_1^{(n)}$.

The system $\{\phi_j\}_1^\infty$ is said to be *strongly minimal in* $\underline{\underline{H}}$, if

$$\inf \lambda_1^{(n)} = \lim_{n\to\infty} \lambda_1^{(n)} > 0 , \tag{13}$$

and *almost orthonormal in* $\underline{\underline{H}}$, if it is strongly minimal and

$$\sup \lambda_n^{(n)} = \lim_{n\to\infty} \lambda_n^{(n)} < \infty . \tag{14}$$

Remark 2.1 The central role of the Gram Matrix G_n in the formulation of these definitions can be explained in the following way. It defines the matrix L_n which is generated in the construction of best approximations of the form (2) for a given f ; or equivalently, the matrix L_n which the

Ritz-Galerkin method generates when applied to (1) with $\underset{\sim}{L} = \underset{\sim}{I}$, the identity operator. Therefore, when viewed as operators, the Gram matrices G_n define mappings from the elements $f \in \underline{\underline{H}}$ to the elements $\underset{\sim}{a}^{(n)} \in \ell_2$ (the Hilbert space of infinite sequences of elements $\underset{\sim}{\alpha} = (\alpha_1, \alpha_2, \ldots)$ with norm $\|\underset{\sim}{\alpha}\| = \sum_{i=1}^{\infty} \alpha_i^2 < \infty$).

The above conditions which define minimality, strong minimality and almost orthonormality correspond to the conditions which identify special properties of the $\underset{\sim}{a}^{(n)}$ as elements of ℓ_2. A discussion of such properties is contained in Mikhlin [14],§5, though, as indicated there, the original results date back to Lewin [11] and Taldykin [18]. In particular, the minimality of the $\{\phi_j\}_1^{\infty}$ in $\underline{\underline{H}}$ guarangees that, for fixed j, the $a_j^{(n)}$ have limits a_j as $n \to \infty$. However, it is necessary to invoke the strong minimality assumption to ensure that the resulting infinite sequences $(a_1, a_2, a_3, \ldots)$ lie in ℓ_2. In fact, it follows from the definition of strong minimality that

$$\sum_{j=1}^{n} |a_j^{(n)}| \leq \lambda_0^{-1} \|f\| , \quad f \in \underline{\underline{H}} ,$$

and hence that the mappings $G_n : \underline{\underline{H}} \to \ell_2$, $n=1,2,\ldots$, are bounded.

An immediate consequence is the observation that the additional condition which ensures that a strongly minimal system is almost orthonormal guaranteesthat the inverse mappings $G_n^{-1} : \ell_2 \to \underline{\underline{H}}$, $n=1,2,\ldots$ (which exist because of the strong minimality assumption) are bounded. Thus, when the system $\{\phi_j\}_1^{\infty}$ is almost orthonormal, the mappings $G_n : \underline{\underline{H}} \to \ell_2$, $n=1,2,\ldots$, induce an isomorphism between $\underline{\underline{H}}$ and ℓ_2.

The minimality definitions could be based on these properties, but, from a computational point of view, those given above are the more appropriate because of the key role the spectral condition number $\mathbb{K}(G_n) = \lambda_n^{(n)} / \lambda_1^{(n)}$ plays in the numerical analysis of positive definite matrices. #

Similar and Semi-Similar Operators

For the analysis of spectral methods developed below, the key step rests on results which allow the properties of a system $\{\phi_j\}_1^\infty$ in one Hilbert space to be inferred from its properties in a related Hilbert space; in particular, when one space is continuously (densely) imbedded in the other (cf. Gilbarg and Trudinger [7]). In fact, we examine the simplest possible form of continuous imbedding where, for two spaces

$\underline{\underline{H}}_1$ and $\underline{\underline{H}}_2$ with norms $\|\cdot\|_1$ and $\|\cdot\|_2$, $\underline{\underline{H}}_1$ is dense in $\underline{\underline{H}}_2$ and

$$\|u\|_2 \leq K\|u\|_1 , \qquad u \in \underline{\underline{H}}_1 , \qquad K = \text{const.} \tag{15}$$

Since we are principally concerned with selfadjoint and positive definite operators, we require conditions on them which guarantee inequalities of the form of (15). For this we use the concept of similar and semi-similar operators.

Two selfadjoint and positive definite operators $\underset{\sim}{A}$ and $\underset{\sim}{B}$ are *similar*, if $\underline{\underline{D}}(\underset{\sim}{A}) = \underline{\underline{D}}(\underset{\sim}{B})$; and *semi-similar*, if $\underline{\underline{H}}_{\underset{\sim}{A}} = \underline{\underline{H}}_{\underset{\sim}{B}}$.

In particular, one obtains results of the following form (cf. Mikhlin [14], §3):

Theorem 2.1. *Let $\underset{\sim}{A}$ and $\underset{\sim}{B}$ be positive definite operators such that $\underline{\underline{H}}_{\underset{\sim}{A}}$ is contained in $\underline{\underline{H}}_{\underset{\sim}{B}}$. Then there exists a constant* c *such that*

$$\|u\|_{\underset{\sim}{B}} \leq c\|u\|_{\underset{\sim}{A}} , \qquad u \in \underline{\underline{H}}_{\underset{\sim}{A}} . \tag{16}$$

Theorem 2.2. *Let $\underset{\sim}{A}$ and $\underset{\sim}{B}$ be selfadjoint and positive definite operators which are similar. Then there exists positive constants c_1 and c_2 such that*

$$c_1\|\underset{\sim}{B}u\| \leq \|\underset{\sim}{A}u\| \leq c_2\|\underset{\sim}{B}u\| , \qquad u \in \underline{\underline{D}}(\underset{\sim}{A}) . \tag{17}$$

Theorem 2.3. *Let $\underset{\sim}{A}$ and $\underset{\sim}{B}$ be selfadjoint and positive definite operators which are semi-similar. Then there exist positive constants c_1 and c_2 such that*

$$c_1\|u\|_{\underset{\sim}{B}} \le \|u\|_{\underset{\sim}{A}} \le c_2\|u\|_{\underset{\sim}{B}} \tag{18}$$

where $\|u\|_{\underset{\sim}{A}}$ *and* $\|u\|_{\underset{\sim}{B}}$ *denote the energy norms of* $\underline{\underline{H}}_{\underset{\sim}{A}}$ *and* $\underline{\underline{H}}_{\underset{\sim}{B}}$*, respectively.*

Comparison Theorems

The results which allow the properties of a system $\{\phi_j\}_1^\infty$ in one Hilbert space to be inferred from its properties in another Hilbert space are called *comparison theorems*. For the minimality, strong minimality and almost orthonormality concepts defined above, the relevant comparison theorems are

Theorem 2.4. *Let* $\underline{\underline{H}}_1$ *be continuously imbedded in* $\underline{\underline{H}}_2$ *and assume that the coordinate system* $\{\phi_j\}_1^\infty$ *lies in and spans* $\underline{\underline{H}}_1$*. If this system is (strongly) minimal in* $\underline{\underline{H}}_2$*, it is (strongly) minimal in* $\underline{\underline{H}}_1$.

Proof. The denseness of $\underline{\underline{H}}_1$ in $\underline{\underline{H}}_2$ implies that the system $\{\phi_j\}_1^\infty$ also spans $\underline{\underline{H}}_2$. The minimality in $\underline{\underline{H}}_2$ and the continuous imbedding of $\underline{\underline{H}}_1$ in $\underline{\underline{H}}_2$ imply, using a *reductio ad absurdum* argument, minimality in $\underline{\underline{H}}_1$. For the strong minimality case, we let G_n^i, i=1,2, denote the following Gram matrices

$$G_n^{(i)} = \begin{bmatrix} (\phi_1,\phi_1)_i,\ldots,(\phi_n,\phi_1)_i \\ ------------ \\ (\phi_1,\phi_n)_i,\ldots,(\phi_n,\phi_n)_i \end{bmatrix}$$

where $(u,v)_i$ denotes the inner products of $\underline{\underline{H}}_i$, $i=1,2$. We know from Remark 2.1 that strong minimality in $\underline{\underline{H}}_2$ implies that the mappings $G_n^{(2)}:\underline{\underline{H}}_2 \to \ell_2, n=1,2,\ldots,$ are bounded. Because $\underline{\underline{H}}_1$ is continuously imbedded in $\underline{\underline{H}}_2$, it follows from (15) that the mappings $G_n^{(2)}:\underline{\underline{H}}_2 \to \ell_2, n=1,2,\ldots,$ are bounded. This implies the existence of a constant μ_0, greater than zero and independent of n, such that

$$\inf_n \lambda_{1,1}^{(n)} = \mu_0 ,$$

where the $\lambda_{1,1}^{(n)}$ denote the smallest eigenvalues of the matrices $G_n^{(1)}$; and therefore establishes the strong minimality in $\underline{\underline{H}}_1$. #

Theorem 2.5. *Let $\underline{\underline{H}}_1$ and $\underline{\underline{H}}_2$ be continuously imbedded in each other. If the system $\{\phi_j\}_1^\infty$ is almost orthonormal in one of them it is almost orthonormal in the other.*

§3. THE STABILITY OF THE RITZ-GALERKIN AND BUBNOV-GALERKIN METHODS

The starting point of the backwards error analysis approach is the decision to interpret the computed solution $\underset{\sim}{x}_c$ of the matrix equation (cf. Wilkinson [21])

(19) $$A \underset{\sim}{x} = \underset{\sim}{b}$$

as the exact solution of some other matrix equation

(20) $$B \underset{\sim}{x}_c = \underset{\sim}{d}$$

where the choice of B and $\underset{\sim}{d}$ will depend on circumstances under which $\underset{\sim}{x}_c$ was derived from (19) as well as on A and $\underset{\sim}{b}$. Clearly, because $\underset{\sim}{x}_c$ is a specific solution, it must be viewed as the unique solution of (20). This implies that B must be non-singular and $\underset{\sim}{d}$ uniquely defined. Usually, it is assumed that (20) takes the form

(21) $$(A + \delta A)\underset{\sim}{x}_c = \underset{\sim}{b} + \delta\underset{\sim}{b} .$$

Because a standard argument (cf. Wilkinson [21]) yields a bound for $\underset{\sim}{x}_c - \underset{\sim}{x}$ in terms of $\delta\underset{\sim}{b}$ and δA, it follows that, if a definition of stability is required, it must assert the boundedness of $\|\underset{\sim}{x}_c - \underset{\sim}{x}\|$ in terms of $\|\delta\underset{\sim}{b}\|$ and $\|\delta A\|$. This is the essence of the definition of stability introduced by Mikhlin [14]).

Corresponding to the *exact Ritz-Galerkin process*

$$R_n \underset{\sim}{a}^{(n)} = \underset{\sim}{f}^{(n)}, \qquad n = 1,2,3,\ldots, \tag{22}$$

one considers the *perturbed Ritz-Galerkin process*

$$(R_n + \Gamma_n)\underset{\sim}{b}^{(n)} = \underset{\sim}{f}^{(n)} + \underset{\sim}{\delta}^{(n)}, \qquad n = 1,2,3,\ldots, \tag{23}$$

which defines the exact Ritz-Galerkin process for the non-exact Ritz-Galerkin solution $\underset{\sim}{b}^{(n)}$.

<u>Definition 3.1</u>. The Ritz-Galerkin process is said to be *stable*, if there exist constants p, q and r independent of n such that, for $\|\Gamma_n\| \le r$ and arbitrary $\delta^{(n)}$, the matrix $R_n + \Gamma_n$ is non-singular and the following inequality holds

$$\|\underset{\sim}{b}^{(n)} - \underset{\sim}{a}^{(n)}\| \le p\|\Gamma_n\| + q\|\delta^{(n)}\|. \tag{24}$$

The relationship between this and other forms of stability are discussed and examined in Linz [12] (§4.3) and Omodei [15].

The result of Mikhlin [14], which we use to characterize the numerical performance of spectral methods, is contained in his stability theorems. For the Ritz-Galerkin and Bubnov-Galerkin methods introduced in §1, we have:

<u>Theorem 3.1</u>. *A necessary and sufficient condition for the stability of the Ritz-Galerkin process is that its generating system* $\{\phi_j\}_1^\infty$ *be strongly minimal in* $\underset{\sim}{\underline{\underline{H}}}_A$.

<u>Theorem 3.2</u>. *Sufficient conditions for the stability of the Bubnov-*

Galerkin process are that $\underset{\sim}{L}u = f$ *has only one solution and that its generating system* $\{\phi_j\}_1^\infty$ *is strongly minimal in any* $\underline{\underline{H}}_{\underset{\sim}{A}}$ *for which* $\underset{\sim}{L} = \underset{\sim}{A} + \underset{\sim}{B}$ *with* $\underset{\sim}{A}^{-1}\underset{\sim}{B}$ *compact.*

Thus, the task of guaranteeing the stability of the Ritz-Galerkin and Bubnov-Galerkin processes reduces to identifying the properties of $\{\phi_j\}_1^\infty$ in $\underline{\underline{H}}$ which imply the strong minimality of $\{\phi_j\}_1^\infty$ in $\underline{\underline{H}}_{\underset{\sim}{A}}$. In particular, the numerical performance of spectral methods can be characterized in terms of the conditions which must be imposed on the choice of an orthonormal system in $\underline{\underline{H}}$ to guarantee strong minimality in $\underline{\underline{H}}_{\underset{\sim}{A}}$.

In fact, from the results of §2, we obtain

Proposition 3.1. *A system* $\{\phi_j\}_1^\infty$ *which lies in both* $\underline{\underline{H}}$ *and* $\underline{\underline{H}}_{\underset{\sim}{A}}$ *, which is orthonormal in* $\underline{\underline{H}}$ *and which spans* $\underline{\underline{H}}_{\underset{\sim}{A}}$ *, is strongly minimal in* $\underline{\underline{H}}_{\underset{\sim}{A}}$.

Proof. On the strength of Theorem 2.1, $\underline{\underline{H}}_{\underset{\sim}{A}}$ is imbedded in $\underline{\underline{H}}$. However, when $\underline{\underline{H}}$ cannot be imbedded in $\underline{\underline{H}}_{\underset{\sim}{A}}$, Theorem 2.5 cannot be applied. Thus, only the strong minimality of an orthonormal system in $\underline{\underline{H}}$ is preserved in $\underline{\underline{H}}_{\underset{\sim}{A}}$ as shown by Theorem 2.4. #

The proviso of Proposition 3.1, that the orthonormal system be located in both $\underline{\underline{H}}_{\underset{\sim}{A}}$ and $\underline{\underline{H}}$, is needed so that the imbedding assumptions of Theorem 2.4 and 2.5 hold, and is guaranteed by the global-$\underline{\underline{D}}(\underset{\sim}{A})$ and global-$\underline{\underline{H}}_{\underset{\sim}{A}}$ conditions on the system $\{\phi_j\}_1^\infty$ which ensure convergence.

§4. LIMITATIONS ON THE UTILITY OF THE SPECTRAL METHOD

The proposition derived in §3 yields direct verification of the utility of spectral methods. It shows that convergent and stable approximations of the form (2) can be constructed using arbitrary orthonormal systems in $\underline{\underline{H}}$, when the procedures used to determine the unknowns $a_j^{(n)}$, $j = 1,2,\ldots,n$, $n = 1,2,\ldots$, correspond to one of the standard methodologies such as Ritz-Galerkin, Bubnov-Galerkin or Least Squares.

There are however limitations on the utility of taking arbitrary orthonormal systems in $\underline{\underline{H}}$ to construct approximations of the form (2) for (1). The example of Anderssen and Omodei [1] shows that the use of orthonormal systems cannot undo the damage being done by a poor methodology for the construction of the approximations (2). In addition, even using the standard methodologies, an arbitrary choice is unable to guarantee all the desirable numerical properties, such as the existence of a bounded condition number for the Ritz matrices R_n, $n = 1,2,\ldots$, and the convergence of the residuals $\underset{\sim}{A}\, u_n - f$ and $\underset{\sim}{L}\, u_n - f$. It is this aspect which we pursue here using the backwards error analysis for matrix equations developed in §3.

One interpretation of the backwards error analysis representation (21) for the computed solution $\underset{\sim}{x}_c$ of $A\, \underset{\sim}{x} = \underset{\sim}{b}$ is that, except for the errors δA and $\delta \underset{\sim}{b}$ which were introduced during the construction of (20) to yield (21), the matrix equation is solved exactly (without error). Clearly, in this interpretation, the effect of rounding errors is ignored. Even if (21) is interpreted as accounting for the errors arising during both the construction of (20) and the subsequent approximate solution of the matrix equation, the present stability analysis has one crucial defect. It is limited to an analysis of absolute errors.

As the standard texts in numerical analysis indicate, relative errors are usually more appropriate in assessing the effect of rounding errors than absolute. Thus, a definition of stability based on relative errors (i.e. a *relative error stability*) must assert the boundedness of $\|\underset{\sim}{x}_c - \underset{\sim}{x}\|/\|\underset{\sim}{x}\|$ in terms of $\|\delta\underset{\sim}{b}\|/\|\underset{\sim}{b}\|$ and $\|\delta A\|/\|A\|$. However, we know from Wilkinson [21] that

$$(25) \qquad \frac{\|\underset{\sim}{x}_c - \underset{\sim}{x}\|}{\|\underset{\sim}{x}\|} \le \frac{\mathbb{K}(A)}{1 - \mathbb{K}(A)\ \|\delta A\|/\|A\|} \left\{\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta\underset{\sim}{b}\|}{\|\underset{\sim}{b}\|}\right\} ,$$

where $\mathbb{K}(A)$ denotes the condition number of the matrix A

$$(26) \qquad \mathbb{K}(A) = \|A\|\ \|A^{-1}\| \ .$$

Usually, results like this are used to emphasise the importance of the concept of condition number in the analysis of rounding error (cf. Atkinson [2], Forsythe and Moler [6]). In the present context, it shows immediately that demonstration of stability for relative errors reduces immediately to proving the boundedness of $\mathbb{K}(A)$ independent of n.

Recalling the definition of almost orthonormality, we obtain

<u>Proposition 4.1</u>. *For the matrix spectral norm, a sufficient condition for the relative error stability of the Ritz-Galerkin process is that its generating system* $\{\phi_j\}_1^\infty$ *be almost orthonormal in* $\underset{\sim}{\underline{\underline{H}}}_A$.

<u>Proof</u>. Since the spectral norm of a general matrix A corresponds to the positive square root of the largest eigenvalue of A^TA, it follows immediately that the spectral condition number of R_n, $\mathbb{K}_S(R_n)$, is given by

$$(27) \qquad \mathbb{K}_S(R_n) = \lambda_n^{(n)}/\lambda_1^{(n)} \ .$$

If $\{\phi_j\}_1^\infty$ is almost orthonormal in $\underset{\sim}{\underline{\underline{H}}}_A$, then there exist constants λ_0 and Λ_0 such that

$$(28) \qquad 0 < \lambda_0 \le \lambda_m^{(n)} \le \Lambda_0 < \infty \ , \qquad m = 1,2,\ldots,n, \quad n = 1,2,\ldots .$$

This proves that $\mathbb{K}_S(R_n)$ is bounded independently of n by Λ_0/λ_0. #

In passing, we note that, as a direct consequence of the minimax properties of the eigenvalues $\lambda_m^{(n)}$, $m = 1,2,\ldots,n$, $n = 1,2,\ldots$, it follows that $\mathbb{K}_S(R_n)$ is an increasing function of n.

A similar proposition holds for the Bubnov-Galerkin process.

We have already seen in the proof of Proposition 3.1 that, when the Hilbert space $\underline{\underline{H}}$ from which $\underline{\underline{H}}_{\underset{\sim}{A}}$ is formed cannot be imbedded in $\underline{\underline{H}}_{\underset{\sim}{A}}$, an orthonormal system in $\underline{\underline{H}}$ can only be strongly minimal in $\underline{\underline{H}}_{\underset{\sim}{A}}$. It follows from Theorem 2.5 that, if the orthonormal system $\{\phi_j\}_1^\infty$ in $\underline{\underline{H}}$ was also orthonormal or almost orthonormal in a Hilbert space $\hat{\underline{\underline{H}}}$ such that $\hat{\underline{\underline{H}}}$ and $\underline{\underline{H}}_{\underset{\sim}{A}}$ could be imbedded in each other, then $\{\phi_j\}_1^\infty$ would be almost orthonormal in $\underline{\underline{H}}_{\underset{\sim}{A}}$. Further, it follows from Theorems 2.2 and 2.3 that a sufficient condition for $\hat{\underline{\underline{H}}}$ and $\underline{\underline{H}}_{\underset{\sim}{A}}$ to be imbedded in each other is that $\hat{\underline{\underline{H}}}$ correspond to the energy space $\underline{\underline{H}}_{B}$ of an operator $\underset{\sim}{B}$ wich is either similar or semi-similar to $\underset{\sim}{A}$. In fact, we have established

Proposition 4.2. *For the spectral norm, a sufficient condition for the relative error stability of the Ritz-Galerkin spectral process is that the (orthonormal) system* $\{\phi_j\}_1^\infty$ *in* $\underline{\underline{H}}$ *be almost orthonormal in the energy space* $\underline{\underline{H}}_{\underset{\sim}{A}}$.

A similar proposition holds for the Bubnov-Galerkin process.

If $\underset{\sim}{A}$ and $\underset{\sim}{L}$ are unbounded operators, then it is well-known that, for the approximations u_n generated by some of the standard variational methods such as the Ritz-Galerkin and Bubnov-Galerkin (but excluding least squares), there is no guarantee that the residuals $\underset{\sim}{A}\, u_n - f$ and $\underset{\sim}{L}\, u_n - f$ will converge. As for stability, this difficulty can be circumvented by imposing additional conditions on the choice of the $\{\phi_j\}_1^\infty$.

In fact, the basic result is given by (cf. Mikhlin [14], §22)

Theorem 4.1. *Let $\underset{\sim}{A}$ and $\underset{\sim}{B}$ be similar positive definite operators with domains contained in the separable Hilbert space $\hat{\underline{\underline{H}}}$ and $\underset{\sim}{B}$ have a discrete spectrum. If the coordinate system $\{\phi_j\}_1^\infty$ consists of the normalized eigenfunctions of $\underset{\sim}{B}$, then the residual $\underset{\sim}{A}\, u_n - f$ converges to zero when the approximations u_n are constructed using the Ritz-Galerkin process.*

Proof (Vainikko [20]). The key step is to convert $\underset{\sim}{A}\, u_n - f$ to a form which allows the properties of the $\{\phi_j\}_1^\infty$ to be exploited; namely, $\underset{\sim}{B}\, \phi_j = \mu_j \phi_j$, where the μ_j denote the eigenvalues of $\underset{\sim}{B}$ corresponding to the eigenfunctions $\{\phi_j\}_1^\infty$. We assume that u_f(the solution of $\underset{\sim}{A}u=f$) takes the form

(29) $$u_f = \sum_{j=1}^{\infty} c_j \phi_j \ ,$$

and, with respect to the metric of $\hat{\underline{\underline{H}}}$, define $\underset{\sim}{P}_n$ to be the following orthogonal projection

(30) $$\underset{\sim}{P}_n : \hat{\underline{\underline{H}}} \to \underline{\underline{H}}^{(n)} = \mathrm{span}(\phi_1, \phi_2, \ldots, \phi_n) \ .$$

We write $\underset{\sim}{P}^{(n)} = \underset{\sim}{I} - \underset{\sim}{P}_n$.

The proof first exploits a consequence of Theorem 2.2, the boundedness of $\underset{\sim}{A}\, \underset{\sim}{B}^{-1}$ and $\underset{\sim}{A}^{-1}\, \underset{\sim}{B}$:

(31) $$\| \underset{\sim}{A}\, u_n - f \| = \| \underset{\sim}{A}(u_n - u_f) \| \le \| \underset{\sim}{A}\, \underset{\sim}{B}^{-1} \| \ \| \underset{\sim}{B}(u_n - u_f) \| \ .$$

The importance of this step is that it brings $\underset{\sim}{B}$ into direct relationship with the Ritz-Galerkin approximation u_n constructed from the $\{\phi_j\}_1^\infty$. In addition, because $\underset{\sim}{P}^{(n)}\, u_n = 0$, it follows that

(32) $$\underset{\sim}{B}(u_f - u_n) = \underset{\sim}{B}(\underset{\sim}{P}_n + \underset{\sim}{P}^{(n)})(u_f - u_n) = \underset{\sim}{B}\, \underset{\sim}{P}_n(u_f - u_n) + \underset{\sim}{B}\, \underset{\sim}{P}^{(n)}\, u_f \ .$$

Appropriate estimates for the terms on the right hand side of (32) are derived from the following consequences of the definitions of the $\{\phi_j\}_1^{\infty}$ and $\underset{\sim}{P}^{(n)}$: $\underset{\sim}{B}^{\alpha}$, $0 < \alpha < 1$, and $\underset{\sim}{P}_n$ commute; and

$$(33) \qquad \|\underset{\sim}{B}^{\alpha}\, \underset{\sim}{P}_n\| = \mu_n^{\alpha} , \qquad \|\underset{\sim}{B}^{-\alpha}\, \underset{\sim}{P}^{(n)}\| = \mu_{n+1}^{-\alpha} .$$

In fact, using (29), it follows that

$$(34) \qquad \|\underset{\sim}{B}\, \underset{\sim}{P}^{(n)}\, u_f\| = \|\underset{\sim}{P}^{(n)}\, \underset{\sim}{B} \sum_{j=1}^{\infty} c_j\phi_j\| = \| \sum_{j=n+1}^{\infty} c_j\mu_j\phi_j\| \to 0$$

as $n \to \infty$. In addition, using (33) and the best approximation properties of u_n in $\underline{\underline{H}}_{\underset{\sim}{A}}$, it can be shown that

$$(35) \qquad \|\underset{\sim}{B}\, \underset{\sim}{P}_n(u_f - u_n)\| \le \|\underset{\sim}{A}^{\frac{1}{2}}\, \underset{\sim}{B}^{-\frac{1}{2}}\|\, \|\underset{\sim}{B}^{\frac{1}{2}}\, \underset{\sim}{A}^{-\frac{1}{2}}\|\, \|\underset{\sim}{B}\, \underset{\sim}{P}^{(n)}\, u_f\| .$$

The convergence of the residual $\underset{\sim}{A}\, u_n - f$ now follows from Theorem 2.2, (31), (32), (34) and (35). #

This proof depends crucially on the ϕ_j being eigenfunctions of a positive definite operator $\underset{\sim}{B}$ which is similar to $\underset{\sim}{A}$ and has a discrete spectrum. However, it does not rule out the possibility that some subclass of the almost orthonormal systems in $\underline{\underline{H}}_{\underset{\sim}{B}}$ might also guarantee convergence of the residual $\underset{\sim}{A}\, u_n - f$. Nevertheless, it clearly illustrates a further limitation on the numerical performance of spectral methods when the orthonormal system is chosen arbitrarily from $\underline{\underline{H}}$.

§5. CONCLUDING REMARKS

As explained in the Introduction, the aim of this paper was to show how theory developed by Mikhlin [14] for studying the numerical performance of variational methods could be adapted for an examination of the numerical performance of spectral methods. For this reason, attention has been limited to stationary problems. In particular, it has been shown that the construction of spectral methods, using arbitrary orthonormal systems in $\underline{\underline{H}}$, is sufficient to guarantee absolute error stability, but not relative error stability nor convergence of the residual of an unbounded operator. In addition, the properties which orthonormal systems in $\underline{\underline{H}}$ must satisfy to guarantee relative error stability and convergence of the residual are discussed.

The basic characterization developed here extends naturally to the study of the numerical performance of spectral methods for time dependent problems , and eigenvalue problems. However, a discussion is beyond the scope of this paper. Appropriate results for eigenvalue problems can be found in Mikhlin [14] and Chatelin [3]. In addition, deeper results than those derived here would be obtained if a more specific exploitation of the theory of variational methods was applied to the study of spectral methods. Source references for such material are Kreiss and Oliger [10],Gottlieb and Orszag [8], Voigt _et al_ [19] and Hussaini _et al_ [9].

The material of Mikhlin [14] has been motivated by the need to have, for specific problems, reliable choices for the coordinate systems. For spectral methods, one needs the converse: for specific orthonormal systems, a catalogue is required which lists the numerical properties they guarantee for various classes of ordinary and partial differential equations as well as integral equations. Such information is contained in references like Gottlieb and Orszag [8], Orszag [16], and Delves and Freeman [4].

The sufficient conditions for relative error stability of §4 were derived using the spectral norm for matrices. Because the spectral condition number of a matrix is always bounded above by the maximum norm condition number, it follows that the Ritz-Galerkin process could yield an approximation which exhibits relative error stability in the ℓ_2-norm but not relative error stability in the maximum norm. This is not surprising since it is well known how to construct n-component vectors which, as a function of n, are arbitrarily large in maximum norm and bounded in ℓ_2. Clearly, to prove the Proposition 4.1 for the maximum norm would automatically guarantee its validity for the spectral norm. However, in terms of the properties of strongly minimal and almost orthonormal systems, the natural setting for Proposition 4.1 is the spectral norm.

ACKNOWLEDGEMENT

The author wishes to acknowledge helpful comments received from Alan McIntosh on an earlier draft of this paper.

REFERENCES

[1] R.S. Anderssen and B.J. Omodei, "On the stability of uniformly asymptotically diagonal systems", *Math. Comp.* 28 (1974), 719-730.

[2] K.E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley and Sons, New York, 1978.

[3] F. Chatelin, *Spectral Approximation of Linear Operators*, Academic Press, New York, 1983.

[4] L.M. Delves and T.L. Freeman, *Analysis of Global Expansion Methods: Weakly Asymptotically Diagonal Systems*, Academic Press, London, 1981.

[5] C.A.J. Fletcher, *Computational Galerkin Methods*, Springer-Verlag, Berlin, 1984.

[6] G. Forsythe and C.B. Moler, *Computer Solution of Linear Algebraic Equations*, Prentice-Hall, Englewood Cliffs, N.J., 1967.

[7] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[8] D. Gottlieb and S.A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, Penn., 1977.

[9] M.Y. Hussaini, C.L. Streett and T.A. Zang, *Spectral Methods for Partial Differential Equations*, Transactions of the First Army Conference on Applied Mathematics and Computing, ARO Report 84-1-,1984.

[10] H.-O. Kreiss and J. Oliger,Comparison of accurate methods for the integration of hyperbolic equations, *Tellus* 24 (1972), 199-215.

[11] S. Lewin, Über einige mit der Konvergenz Mittel verbundenen Eigenschaften von Funktionalfolgen, *Zeit. fur. Math.* 32 (1930).

[12] P. Linz, *Theoretical Numerical Analysis: An Introduction to Advanced Techniques*, John Wiley and Sons, New York, 1979.

[13] S.G. Mikhlin, *The Problem of the Minimum of a Quadratic Functional*, Holden-Day, San Francisco, 1965.

[14] S.G. Mikhlin, *The Numerical Performance of Variational Methods*, Wotters-Noordhoff Publishing, Groningen, The Netherlands, 1971.

[15] B.J. Omodei, "On the numerical stability of the Rayleigh-Ritz method", *SIAM J. Numer. Anal.* 14 (1977), 1151-1171.

[16] S.A. Orszag, "Spectral methods for problems in complex geometries", *J. Comp. Phys.* 37 (1980), 70-92.

[17] R. Peyret and T.D. Taylor, *Computational Methods for Fluid Flow*, Springer Series in Computational Physics, Springer-Verlag, New York, 1983.

[18] A.T. Taldykin, Systems of elements of a Hilbert space and series constructed with them, *Math. Sb.* 29 (1951), 79-120.

[19] R.G. Voigt, D. Golllieb and M.Y. Hussaini (Editors), *Proceedings of the Symposium on Spectral Methods for Partial Differential Equations*, SIAM, Philadelphia, 1984.

[20] G.M. Vainikko, "On similar operators", *Dokl. Akad. Nauk SSSR* 179 (1968), 1029-1031; English translation: *Soviet Math. Dokl.* 9 (1968), 477-480.

[21] J.H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, N.J., 1963.

Division of Mathematics and Statistics, CSIRO,
GPO Box 1965, Canberra, ACT 2601