MATHEMATICS OF COMPUTATION, VOLUME 28, NUMBER 127, JULY 1974, PAGES 719-730

On the Stability of Uniformly Asymptotically Diagonal Systems

R. S. Anderssen and B. J. Omodei

Abstract. In a number of papers ([1], [2]), Delves and Mead have derived some useful (though limited) rate of convergence results which can be applied to variational approximations for the solution of linear positive definite operator equations when the coordinate system is uniformly asymptotically diagonal. Independently, Mikhlin [5] has examined the stability of such variational approximations in the case of positive definite operators and concluded that the use of strongly minimal coordinate systems is a necessary and sufficient condition for their stability. Since, in general, the Delves and Mead results will only be applicable to actual variational approximations when their uniformly asymptotically diagonal system is at least strongly minimal, we examine the properties of uniformly asymptotically diagonal systems in terms of the minimal classification of Mikhlin.

We show that

(a) a normalized uniformly asymptotically diagonal system is either nonstrongly minimal or almost orthonormal;

(b) the largest eigenvalue of a normalized uniformly asymptotically diagonal system is bounded above, independently of the size of the system;

(c) the special property of normalized uniformly asymptotically diagonal systems mentioned in (b) is often insufficient to prevent their yielding unstable results when these systems are not strongly minimal.

1. Introduction. The aim of this paper is to relate the work of Delves and Mead on uniformly asymptotically diagonal systems to Mikhlin's work on the minimal classification and necessary and sufficient conditions for the stability of variational methods.

We restrict our attention to linear operator equations

$$\mathbf{A}u = f \qquad (f \in \mathfrak{H}), \tag{1.1}$$

where $\mathbf{A}$ is a positive definite symmetric operator defined on a domain $\mathfrak{D}(\mathbf{A})$, a linear manifold dense in a given separable Hilbert space $\mathfrak{H}$, and the range of $\mathbf{A}$ is in $\mathfrak{H}$. By positive definite we mean that $(\mathbf{A}u, u) \geqq \gamma^2 \|u\|^2$ for all $u \in \mathfrak{D}(\mathbf{A})$, where γ is a positive constant. For a suitably complete set of coordinate elements $\{\phi_i\}$, the approximate *Ritz-solution* (or variational solution) of (1.1) is defined to be

$$u_n = \sum_{i=1}^{n} a_i^{(n)} \phi_i \tag{1.2}$$

with the unknown $a_i^{(n)}$ defined by the *Ritz-system*

$$R_n a^{(n)} = f^{(n)} \tag{1.3}$$

where

Received August 2, 1973.

AMS (MOS) subject classifications (1970). Primary 65N30; Secondary 15A57, 41A25.

Key words and phrases. Ritz-solution, uniformly asymptotically diagonal systems, minimal classification, stability of variational methods.

$$a^{(n)} = [a_1^{(n)}, a_2^{(n)}, \dots, a_n^{(n)}]^T,$$
$$(R_n)_{ij} = (\mathbf{A}\phi_j, \phi_i) \qquad (i,j = 1,2,\dots,n), \quad \text{and}$$
$$f^{(n)} = [(f,\phi_1),(f,\phi_2),\dots,(f,\phi_n)]^T.$$

In the limit as $n \to \infty$, (1.3) and R_n become,respectively, the *infinite Ritz-system* and the *infinite Ritz-matrix R.*

Under the assumption that **A** is positive definite, the energy inner product and norm are defined by $[u,v] = (\mathbf{A}u, v)$ and $|||u||| = [u,u]^{1/2}$, u, $v \in \mathfrak{D}(\mathbf{A})$, respectively.

The energy space $\mathfrak{H}_\mathbf{A}$ is the completion of $\mathfrak{D}(\mathbf{A})$ with respect to the energy norm. It is shown in Mikhlin [3, Section 5] that the positive definite operator **A** can be extended to a selfadjoint operator the domain of which is in $\mathfrak{H}_\mathbf{A}$. Denoting the extension of **A** as **A**, it follows from Mikhlin [3, Sections 5, 8] that

(i) (1.1) defines a one-one mapping from $\mathfrak{D}(\mathbf{A})$,the domain of the extension, onto $\mathfrak{H}$, and

(ii) if u_0 denotes the exact solution of (1.1), then

$$\lim_{n\to\infty} |||u_n - u_0||| = 0.$$

For these reasons, we restrict our attention to such operators, thus ensuring the existence and uniqueness of the exact solution, and the convergence of the Ritz-solution to the exact solution in $\mathfrak{H}_\mathbf{A}$. Furthermore, from Mikhlin [4, p. 324], convergence in $\mathfrak{H}$ is also ensured.

We note in passing that approximate variational solutions can be constructed for operators which are positive, i.e., $(\mathbf{A}u, u) > 0$ for all $u \neq 0$, $u \in \mathfrak{D}(\mathbf{A})$, but not positive definite. However, care must be exercised to ensure that (1.1) has a solution and that u_n converges to u_0 in $\mathfrak{H}$ {see, for example, Mikhlin [3, Section 6]}.

Next,we note that though the above conditions guarantee the convergence of u_n to u_0, they do not ensure the stability of the inversion of (1.3) with respect to rounding error perturbations. This has been cogently demonstrated by Mikhlin [5, Section 8]. Mikhlin has shown that, at least for positive definite operators, the stability of the estimation of $a_i^{(n)}$ and u_n depends upon the choice of coordinate elements $\{\phi_i\}$.

Two independent classes of coordinate systems have been discussed in recent literature; viz., the uniformly asymptotically diagonal systems of Delves and Mead, and the minimal systems of Mikhlin. For their uniformly asymptotically diagonal systems, Delves and Mead derive rate of convergence results but ignore the question of stability, whereas Mikhlin shows that a necessary and sufficient condition for stability is the strong minimality of the $\{\phi_i\}$ (the relevant definitions and results are cited in Section 2). Since it appears therefore that the Delves and Mead results will probably only be useful for uniformly asymptotically diagonal systems which are stable, we examine the properties of normalized uniformly asymptotically diagonal systems in terms of the minimal classification of Mikhlin. After developing preliminaries in Section 2, we establish in Section 3 that

(a) a normalized uniformly asymptotically diagonal system is either nonstrongly minimal or almost orthonormal, and

(b) the largest eigenvalue of a normalized uniformly asymptotically diagonal system is bounded above, independently of the size of the system.

In addition, we consider two classes of coordinate systems which show that neither strongly minimal systems nor normalized uniformly asymptotically diagonal systems are subclasses of the other. By constructing a specific example, we show in Section 4 that

(c) normalized uniformly asymptotically diagonal systems which are nonstrongly minimal can exhibit instability.

2. Preliminaries. In this section, we state relevant definitions and results from the theory developed by Delves and Mead, and Mikhlin.

The Delves and Mead Results {*see* [1], [2]}.

Definition 2.1.An infinite Hermitian matrix R is said to be *uniformly asymptotically diagonal* (U.A.D.) of degree p (> 0) if, for all i and for all $n > i$, there exists a positive constant C such that

$$|R_{in}|/(|R_{nn}|\,|R_{ii}|)^{1/2} < Cn^{-p}. \tag{2.1}$$

Remark 2.1. If (2.1) is true for all n and i, $n > i$, then it can be shown that (2.1) will hold for all n and i, $n \neq i$.

Remark 2.2. Regarding terminology, we refer to a set of coordinate elements $\{\phi_i\}$ as a U.A.D. system with respect to an operator **A**, if and only if the corresponding infinite Ritz-matrix R is U.A.D.

Remark 2.3. Delves and Mead [1] have shown that the U.A.D. property is invariant under an arbitrary renormalization of the coordinate elements $\{\phi_i\}$. Therefore, we can always renormalize to ensure that

$$R_{ii} = 1 \qquad (i = 1,2,\dots).$$

For this reason, Delves and Mead ([1], [2]) only considered matrices R which had been normalized in this way, and their theorems are proved for such systems.

Definition 2.2. A U.A.D. system of degree p is said to be "*nice*", if

(i) $p > 1$,

(ii) the system is normalized such that $R_{ii} = 1$ $(i = 1,2,\dots)$, and

(iii) $0 < C \leqq C(p)$ where C is the constant in Definition 2.1, and

$$C(p) = \frac{(p-1)}{(8p^2-8p-1)}[((8p-7)^2 + (8p^2-8p-1))^{1/2} - (8p-7)].$$

Remark 2.4. In Delves and Mead ([1], [2]), "nice" systems are used extensively. In fact, only one of the rate of convergence theorems in [1] apply to non-nice systems.

Dropping the restriction on C in (iii) leads to the following definition.

Definition 2.3. A *normalized uniformly asymptotically diagonal* (N.U.A.D.) system of degree p is a U.A.D. system satisfying

(i) $p > 1$, and

(ii) $R_{ii} = 1$ $(i = 1,2,\dots)$.

Remark 2.5. We could have defined a less restrictive U.A.D. system than the N.U.A.D. system, but this would not change the character of the results derived below, but only lead to a rather cumbersome presentation.

The Mikhlin Results {*see* [5]}.

Definition 2.4. A system of elements of a Hilbert space $\mathfrak{H}$ is called a *minimal system* in this space, if the deletion of any one of the elements from the system restricts the closed linear subspace generated by the new set to a proper closed subspace of the closed linear space generated by the original set.

Consider the countable system of elements $\{\phi_i\}$ lying in the Hilbert space $\mathfrak{H}$, and the corresponding Gram matrix of the first n elements of $\{\phi_i\}$, viz.

$$\begin{bmatrix} (\phi_1,\phi_1)_{\mathfrak{H}} & (\phi_1,\phi_2)_{\mathfrak{H}} & \cdots & (\phi_1,\phi_n)_{\mathfrak{H}} \\ (\phi_2,\phi_1)_{\mathfrak{H}} & (\phi_2,\phi_2)_{\mathfrak{H}} & \cdots & (\phi_2,\phi_n)_{\mathfrak{H}} \\ \cdots & \cdots & \cdots & \cdots \\ (\phi_n,\phi_1)_{\mathfrak{H}} & \cdots & \cdots & (\phi_n,\phi_n)_{\mathfrak{H}} \end{bmatrix}.$$

Since this matrix is Hermitian and nonnegative semidefinite, its eigenvalues are nonnegative and can be written in increasing order as

$$0 \leqq \lambda_1^{(n)} \leqq \lambda_2^{(n)} \leqq \cdots \leqq \lambda_n^{(n)}.$$

Definition 2.5. The system $\{\phi_i\}$ is called *strongly minimal* in $\mathfrak{H}$, if there exists a positive constant λ_0 independent of n such that

$$\inf \lambda_1^{(n)} = \lim_{n\to\infty} \lambda_1^{(n)} \geqq \lambda_0 > 0.$$

Definition 2.6. The system $\{\phi_i\}$ is called *almost orthonormal* in $\mathfrak{H}$, if there exist positive constants λ_0, Λ_0 such that for all n and m, $m \leqq n$, the following inequality holds:

$$0 < \lambda_0 \leqq \lambda_m^{(n)} \leqq \Lambda_0.$$

Remark 2.6.

(i) Every strongly minimal system is minimal.

(ii) A minimal system can be renormalized to yield a strongly minimal system {see Dovbyš [6]}.

We shall now introduce Mikhlin's definition of numerical stability. We consider the *exact Ritz-process*

$$R_n a^{(n)} = f^{(n)}. \tag{2.2}$$

Let $\gamma_{km} = \overline{\gamma}_{mk}$ denote the (small) errors arising in the evaluation of the inner products $(A\phi_k, \phi_m)$, and Γ_n the matrix with elements γ_{km} $(k, m = 1, 2, \ldots, n)$. Let $\delta^{(n)}$ be the corresponding error in $f^{(n)}$. Instead of the exact Ritz-process (2.2), we solve the following *nonexact Ritz-process*:

$$(R_n + \Gamma_n) b^{(n)} = f^{(n)} + \delta^{(n)} \tag{2.3}$$

where $b^{(n)}$ is the column-vector of the *nonexact Ritz-coefficients*.

Definition 2.7. The Ritz-process is *stable*, if there exist constants $p, q,$ and r independent of n such that, for $\|\Gamma_n\| \leqq r$ and arbitrary $\delta^{(n)}$, the nonexact Ritz-process is soluble and the following inequality holds:

$$\|b^{(n)} - a^{(n)}\| \leqq p\|\Gamma_n\| + q\|\delta^{(n)}\|.$$

In the opposite case, we say that the Ritz-process is *unstable*.

THEOREM 2.1. {See Mikhlin [5, Section 9].} *In order that the Ritz-process of* (1.1) *be stable, it is necessary and sufficient that its generating coordinate system be strongly minimal in the corresponding energy space; i.e., the eigenvalues of R_n are uniformly bounded away from zero.*

The solution of the exact Ritz-process (2.2) yields the *exact approximate Ritz-solution*

$$u_n = \sum_{k=1}^{n} a_k^{(n)} \phi_k,$$

and the solution of the nonexact Ritz-process (2.3) yields the *nonexact approximate Ritz-solution*

$$v_n = \sum_{k=1}^{n} b_k^{(n)} \phi_k.$$

The definition of stability for the solution of the exact Ritz-process (2.2) is analogous to Definition 2.7, and a theorem almost identical to Theorem 2.1 is valid {see Mikhlin [5, Section 10]}.

3. The Minimal Classification of N.U.A.D. Systems. In this section, we show that N.U.A.D. systems are either nonstrongly minimal or almost orthonormal. We start by deriving conditions under which N.U.A.D. systems can be almost orthonormal.

THEOREM 3.1. *The largest eigenvalue of an N.U.A.D. system is bounded above independently of n.*

Proof. We use Brauer's theorem {see [7]} to obtain the upper bound

$$\Lambda_0 = 1 + Cp/(p-1).$$

THEOREM 3.2. *The smallest eigenvalue of an N.U.A.D. system is bounded away from* 0 *independently of n, provided*

$$C < (p-1)/p, \tag{3.1}$$

where C and p are defined in Definition 2.1.

PROOF. We use Gerschgorin's theorem {see [8]} to obtain the lower bound

$$\lambda_0 = 1 - Cp/(p-1) > 0.$$

COROLLARY 3.1. *An N.U.A.D. system is almost orthonormal provided* $C < (p-1)/p$.

Proof. This is an immediate consequence of Theorems 3.1 and 3.2.

COROLLARY 3.2. *The "nice" systems of Delves and Mead are almost orthonormal.*

Proof. A "nice" system is an N.U.A.D. system satisfying

$$C < C(p) = \frac{(p-1)}{(8p^2-8p-1)}[\{(8p-7)^2 + (8p^2-8p-1)\}^{1/2} - (8p-7)].$$

It can be shown that $C(p) < (p-1)/p$ for $p > 1$, and hence, using Corollary 3.1, Corollary 3.2 is proved.

COROLLARY 3.3. *N.U.A.D. systems are either nonstrongly minimal or almost orthonormal.*

Proof. Since the eigenvalues of an N.U.A.D. system are bounded above independently of n, an N.U.A.D. system which is strongly minimal must also be almost orthonormal, proving the corollary.

Next, we show by constructing two classes of systems that

(i) there exist almost orthonormal systems which are normalized but not N.U.A.D., and

(ii) there exist N.U.A.D.systems which are not strongly minimal and hence not almost orthonormal.

Definition 3.1. A coordinate system will be said to be of *Class* A, if its infinite Ritz-matrix R has the following form:

$$(3.2)\qquad R = \begin{bmatrix} 1 & a_1 & a_2 & a_3 & . & . & . \\ \bar{a}_1 & 1 & 0 & 0 & . & . & . \\ \bar{a}_2 & 0 & 1 & 0 & . & . & . \\ \bar{a}_3 & 0 & 0 & 1 & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \end{bmatrix},$$

and satisfies the following conditions:

(a) $|a_i| \geqq c/i$ for some constant $c > 0$, and

(b) $\sum_{i=1}^{\infty} |a_i|^2 < 1$.

Such a Ritz-matrix is

$$(3.3)\qquad R = \begin{bmatrix} 1 & \alpha(\frac{1}{2})^q & \alpha(\frac{1}{3})^q & \alpha(\frac{1}{4})^q & . & . & . \\ \alpha(\frac{1}{2})^q & 1 & 0 & 0 & . & . & . \\ \alpha(\frac{1}{3})^q & 0 & 1 & 0 & . & . & . \\ \alpha(\frac{1}{4})^q & 0 & 0 & 1 & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \end{bmatrix},$$

where $\frac{1}{2} < q \leqq 1$ and $\sum_{i=2}^{\infty} i^{-2q} < 1/\alpha^2$.

Clearly conditions (a) and (b) of Definition 3.1 are satisfied.

A system which generates such a Ritz-matrix is as follows: Let the operator **A** be the identity operator in $L_2(0,\pi)$, and as coordinate functions, we take

$$\phi_1(x) = \sqrt{\frac{3}{\pi^3}}\,(\pi - x), \qquad \phi_i(x) = \sqrt{\frac{2}{\pi}}\,\sin ix \qquad (i = 2,3,\ldots).$$

THEOREM 3.3. *Systems of Class* A *are almost orthonormal and normalized but not N.U.A.D.*

Proof. Initially, we determine the eigenvalues for the Ritz-matrix R_n. It is easily shown that $\det(R_n - \lambda^{(n)} I_n) = 0$ yields

$$(3.4)\qquad \lambda_i^{(n)} = 1, \qquad 1 \pm \sqrt{\sum_{i=1}^{n-1} |a_i|^2},$$

and thus, for all n,

$$\lambda_n^{(n)} \leqq 1 + \sqrt{\sum_{i=1}^{\infty} |a_i|^2} = \Lambda_0 \quad \text{and} \quad \lambda_1^{(n)} \geqq 1 - \sqrt{\sum_{i=1}^{\infty} |a_i|^2} = \lambda_0.$$

From condition (b) of Definition 3.1, we obtain

$$0 < \lambda_0 \leqq \lambda_s^{(n)} \leqq \Lambda_0$$

(for all n and s, $n \geqq s$), i.e., the system is almost orthonormal.

From condition (a) of Definition 3.1, the system cannot be N.U.A.D.

Definition 3.2. A system will be said to be of *Class* B, if its infinite Ritz-matrix R has the same form as (3.2) and satisfies the following conditions:

(a) $|a_i| \leqq Ci^{-p}$ where C is a positive constant and $p > 1$, and

(b) $\sum_{i=1}^{\infty} |a_i|^2 = 1$ with infinitely many $a_i \neq 0$.

Such a Ritz-matrix is given by (3.3) with

$$\frac{1}{\alpha^2} = \sum_{i=2}^{\infty} i^{-2q} \quad \text{and} \quad q > 1.$$

A system which generates such a Ritz-matrix is as follows: Let

$$\mathbf{A}u = -d^2u/dx^2$$

with $\mathfrak{D}(\mathbf{A})$ the set of twice continuously differentiable functions which satisfy $u(0) = u(\pi) = 0$, and $\mathfrak{H} = L_2(0, \pi)$. As coordinate functions, we take

$$\phi_1(x) = \tfrac{1}{2}\sqrt{\frac{5}{\pi^5}}\, x(x - \pi)(x - 2\pi),$$

$$\phi_{i+1}(x) = \sqrt{\frac{2}{\pi}}\, \frac{1}{i} \sin ix \qquad (i = 1, 2, 3, \ldots).$$

THEOREM 3.4. *Systems of Class* B *are N.U.A.D. but not strongly minimal.*

Proof. Since $|a_i| \leqq Ci^{-p}$ where $p > 1$, it is easily verified that the system is N.U.A.D.

From (3.4) in Theorem 3.3,

$$\lambda_1^{(n)} = 1 - \sqrt{\sum_{i=1}^{n-1} |a_i|^2}.$$

Since $\sum_{i=1}^{\infty} |a_i|^2 = 1$, the system is nonstrongly minimal and the theorem is proved.

Thus, we have shown that N.U.A.D. systems are sometimes nonstrongly minimal, and, when strongly minimal, they are almost orthonormal.

4. The Numerical Stability of N.U.A.D Systems. In this section, we show that the use of N.U.A.D. systems to solve positive definite operator equations can lead to unstable solutions. In Section 3, we showed that N.U.A.D. systems are sometimes strongly minimal and sometimes not. Hence, on the basis of Theorem 2.1 of Section 2, the corresponding Ritz-process for N. U. A. D. systems can be stable or unstable. In particular, the use of an N.U.A.D. system of Class B can yield unstable results.

In order to confirm that instability does arise through the use of such systems, we construct an N.U.A.D. system which is nonstrongly minimal in the energy space, and use it to solve a positive definite operator equation.

Let the operator equation be

$$\mathbf{A}u = -d^2u/dx^2 = f, \qquad f \in L_2(0,\pi), \tag{4.1}$$

with $\mathfrak{D}(\mathbf{A})$ the set of twice continuously differentiable functions which satisfy $u(0) = u(\pi) = 0$. It is easily shown that this defines a positive definite symmetric operator.

As coordinate functions, we take the Class B system

$$\begin{aligned} \phi_1(x) &= \frac{x^2(x-\pi)}{\alpha}, \quad \text{where } \alpha = \sqrt{\frac{2\pi^5}{15}}, \\ \phi_{i+1}(x) &= \sqrt{\frac{2}{\pi}}\,\frac{1}{i}\sin ix \qquad (i = 1,2,3,\ldots). \end{aligned} \tag{4.2}$$

Since $R_{ij} = [\phi_j, \phi_i] = (\mathbf{A}\phi_j, \phi_i)$, it follows that

$$\begin{aligned} R_{ii} &= 1 \qquad (i = 1,2,\ldots), \\ R_{i+1,j+1} &= 0 \qquad (i \neq j;\, i,j = 1,2,\ldots), \\ R_{1,i+1} = R_{i+1,1} &= \begin{cases} -\dfrac{2\sqrt{2\pi}}{i^2\alpha} & (\text{for } i \text{ odd}), \\ \dfrac{6\sqrt{2\pi}}{i^2\alpha} & (\text{for } i \text{ even}). \end{cases} \end{aligned} \tag{4.3}$$

Since $|R_{i,j+1}| \leqq (6\sqrt{2\pi}/\alpha)j^{-2}$ (for all i and j, $i \leqq j$), the system is N.U.A.D. with $p = 2$.

Because $\cos ix$ $(i = 1,2,\ldots)$ form an orthogonal basis for the subspace of integrable functions $f \in L_2(0,\pi)$ which satisfy $\int_0^\pi f(x)\,dx = 0$, it follows that, given any positive ε, there exists an integer n and constants $a_1, a_2, \ldots, a_n$ such that

$$\left|\left|\left| \phi_1 - \sqrt{\frac{\pi}{2}} \sum_{i=1}^{n} a_i \phi_{i+1} \right|\right|\right| = \left\| \frac{1}{\alpha}(3x^2 - 2\pi x) - \sum_{i=1}^{n} a_i \cos ix \right\| < \varepsilon.$$

On the basis of Mikhlin [5, Theorem 1.1], this proves the nonminimality and hence nonstrong minimality of the coordinate system (4.2) in the energy space.

Letting $f = 1$, we obtain $R_n a^{(n)} = f^{(n)}$ with

$$\begin{aligned} f_1 &= -\pi^4/12\alpha, \\ f_{i+1} &= \begin{cases} 0 & (\text{for } i \text{ even}), \\ 2\sqrt{\dfrac{2}{\pi}}\, i^{-2} & (\text{for } i \text{ odd}). \end{cases} \end{aligned} \tag{4.4}$$

TABLE 4.1

The exact Ritz-coefficients $a_i^{(n)}$

n	i=1	5	10	15	20	25	30	35	40
5	-3.517259E-1	5.175848E-2							
10	-1.529826E-1	2.251226E-2	1.821856E-2						
15	-2.441850E-1	3.593322E-2	1.733488E-2	2.933324E-3					
20	-1.769356E-1	2.603708E-2	1.798648E-2	2.125476E-3	4.035747E-3				
25	-2.268062E-1	3.337582E-2	1.750327E-2	2.724557E-3	3.927326E-3	9.271062E-4			
30	-1.854837E-1	2.729499E-2	1.790365E-2	2.228163E-3	4.017163E-3	7.581942E-4	1.724371E-3		
35	-2.197839E-1	3.234247E-2	1.757131E-2	2.640201E-3	3,942593E-3	8.984018E-4	1.692361E-3	4.476466E-4	
40	-1.898548E-1	2.793822E-2	1.786130E-2	2.280671E-3	4,007660E-3	7.760618E-4	1.720292E-3	3.866882E-4	9.511935E-4

TABLE 4.2

The nonexact Ritz-coefficients $b_i^{(n)}$.

n	i=1	5	10	15	20	25	30	35	40
5	-3.551313E-1	5.225960E-2							
10	-1.741036E-1	2.562034E-2	1.801306E-2						
15	-3.357901E-1	4.941343E-2	1.644644E-2	4.033750E-3					
20	-3.694753E-1	5.437040E-2	1.612006E-2	4,438400E-3	3.616744E-3				
25	-6.526245E-1	9.603744E-2	1.337655E-2	7.839791E-3	3.001164E-3	2.667707E-3			
30	-8.648082E-1	1.272615E-1	1.132064E-2	1,038869E-2	2.539867E-3	3.535041E-3	1.089951E-3		
35	-1.390970	2.046891E-1	6.222527E-3	1.670932E-2	1.395969E-3	5.685809E-3	5.989318E-4	2.833067E-3	
40	-1.836431	2.702413E-1	1.906335E-3	2.206051E-2	4.275156E-4	7.506702E-3	1.832225E-4	3.740364E-3	1.014084E-4

TABLE 4.3

Energy norm of the error for the Ritz-solution.

n	Exact Ritz-solution	Non-exact Ritz-solution
5	0.070	0.070
10	0.020	0.022
15	0.012	0.012
20	0.007	0.008
25	0.005	0.006
30	0.004	0.007
35	0.003	0.007
40	0.002	0.008

Because of the simple form of R_n, we can solve the system of equations directly to yield explicit formulae for the Ritz-coefficients:

$$(4.5)\quad a_1^{(n)} = \sqrt{\frac{\pi^3}{7.5}}\left\{-\frac{\pi^4}{96} + \sum_{i=1,\, i\text{ odd}}^{n} \frac{1}{i^4}\right\}\Big/\left\{\frac{\pi^4}{60} - \left(\sum_{i=1,\, i\text{ odd}}^{n} \frac{1}{i^4} + 9 \sum_{i=2,\, i\text{ even}}^{n} \frac{1}{i^4}\right)\right\},$$

for i odd and $i \geqq 3$,

$$(4.6)\qquad a_i^{(n)} = -\frac{6\sqrt{15}}{\pi^2(i-1)^2}\, a_1^{(n)},$$

and, for i even and $i \geqq 2$,

$$(4.7)\qquad a_i^{(n)} = \frac{2\sqrt{15}}{\pi^2(i-1)^2}\left(\sqrt{\frac{\pi^3}{7.5}} + a_1^{(n)}\right).$$

Using these equations, we can examine the source of potential instability in the above example. The numerical instability that arises in the Ritz-coefficients $a_i^{(n)}$ is a direct consequence of the form of (4.5). In (4.5) both the numerator and the denominator approach zero as $n \to \infty$. Thus, as n increases, cancellation error will eventually dominate the computation of both numerator and denominator. The full effect of this cancellation error is carried over to (4.6), and to a lesser extent to (4.7), since $\sqrt{\pi^3/7.5} \gg |a_1^{(n)}|$.

We can calculate the exact Ritz-coefficients for values of n ranging from 5,10, . . . ,40 with 7 significant digits of accuracy by taking the nature of (4.5), (4.6), and (4.7) into account. The results are given in Table 4.1. We calculate the non-exact Ritz-coefficients $b_i^{(n)}$ after rounding the values of f_i to 4 significant digits. The results are tabulated in Table 4.2. Comparing Tables 4.1 and 4.2, it can be seen that the coefficients for a system of size 40 differ by a factor of about 10 for all 9 coefficients shown. This was in fact true for all the coefficients $a_i^{(40)}$ $(i = 1, 2, \ldots, 40)$. Thus, the numerical instability of the Ritz-coefficients is clearly established for the above N.U.A.D. system.

Since the exact solution of (4.1) is known explicitly, viz., $u_0 = -\frac{1}{2}x(x - \pi)$, we can calculate the energy norm for the error in the exact approximate Ritz-solution u_n, viz. $|||u_0 - u_n|||$. These values accurate to 3 decimal places are given in the first column of Table 4.3 for $n = 5, 10, \ldots, 40$. Similarly, we can calculate the energy norm for the error in the nonexact approximate Ritz-solution v_n. The results are given in the second column of Table 4.3.

A comparison of the first and second columns of Table 4.3 illustrates clearly the effect of the numerical instability in the approximate Ritz-solution. In particular, for a system of size 40, the energy norm of the error in the nonexact solution is about 4 times as large as that in the exact solution.

Similar results were observed, when we calculated the L_2-norm of the error of the exact and nonexact approximate Ritz-solutions. However, the actual magnitudes of the norms of the errors were much smaller.

Computer Centre
Australian National University
Canberra, Australia

1. L. M. Delves & K. O. Mead, "On the convergence rates of variational methods. I. Asymptotically diagonal systems," *Math. Comp.*, v. 25, 1971, pp. 699 – 716. MR **46** #10227.

2. K. O. Mead & L. M. Delves, "On the prediction of expansion coefficients in a variational calculation," *J. Inst. Math. Appl.*, v. 10, 1972, pp. 166–175.

3. S. G. Mihlin, *The Problem of the Minimum of a Quadratic Functional*, GITTL, Moscow, 1952; English transl., Holden-Day Ser. in Math. Phys., Holden-Day, San Francisco, Calif., 1965. MR **16**,41; **30** #1427.

4. S. G. Mihlin, *Variational Methods in Mathematical Physics*, GITTL, Moscow, 1957; English transl., Macmillan, New York, 1964. MR **22** #1981; **30** #2712.

5. S. G. Mihlin, *The Numerical Performance of Variational Methods*, "Nauka", Moscow, 1966; English transl., Wolters-Noordhoff, Groningen, 1971. MR **34** #3747; **43** #4236.

6. L. N. Dovbyš, "A remark about minimal systems," *Trudy Mat. Inst. Steklov*, v. 96, 1968, pp. 188-189 = *Proc. Steklov Inst. Math.*, v. 96, 1968, pp. 235–238. MR **42** #8258.

7. A. Brauer, "Limits for the characteristic roots of a matrix," *Duke Math. J.*, v. 13, 1946, pp. 387-395. MR **8**, 192.

8. S. Gerschgorin, "Über die Abgrenzung der Eigenwerte einer Matrix," *Izv. Akad. Nauk SSSR, Ser. Fiz.-Mat.*, v. 6, 1931, pp. 749-754.