MATHEMATICAL RESEARCH, INC.

HOUSTON, TEXAS 77004

NASA-CR-92216

METHODS OF THE

GALERKIN TYPE

THIS REPORT WAS PREPARED WITH PARTIAL SUPPORT OF NASA-MSC(HOUSTON)
CONTRACT NAS 9-6964

TABLE OF CONTENTS

Chapter I

INTRODUCTION

In his paper published in 1915, B. G. Galerkin [46] solved a number of problems on the equilibrium and stability of rods and plates by a method which has become associated with his name, the Galerkin method. Following this publication, a large amount of work has appeared in which the method has been extensively applied to the practical solution of diverse problems. The lack of proofs on convergence has never hindered its application. The first substantiation of the Galerkin method was given in 1940 by Repman [55] in application to Fredholm type integral equations, and in the same year an analogous result for a certain special ordinary fourth-order differential equation was obtained by Petrov [52]. In 1950 Mikhlin [51] obtained a fairly general sufficient condition for the convergence of the Galerkin method and demonstrated the application of this criterion to a number of problems which included the previously published results.

Through the years many modifications to the so-called classical Galerkin method have been suggested and used, in the vast majority of cases without proof of convergence. Justification is usually made on the basis of comparison of successive approximation or by comparison with experiment or other approximate solutions. It is only in recent years (since about 1957) that mathematicians (almost exclusively Russian) have begun to use functional analysis to obtain proofs for approximation methods and, in particular, the Galerkin and related methods.

The classical Galerkin method can briefly be described as follows:

Let A be a linear operator defined on a set D_A which is dense in some separable Hilbert space where a solution to the equation

$$Au - f = 0 \tag{1.1}$$

is required. A set of linearly independent elements $\{\varphi_n\}$ in D_A are selected and are called coordinate functions. For each integer n, an approximate solution of (1.1) is constructed in the form of a linear combination of the first n coordinate functions with constant coefficients, that is,

$$u_n(X) = \sum_{k=1}^{n} a_n \varphi_n(X) \tag{1.2}$$

The coefficients a_n are determined by replacing u by u_n in (1.1) and requiring the left-hand side of (1.1) to be orthogonal to the first n coordinate functions $\varphi_1, \ldots, \varphi_n$. This leads to the following system of n linear equations in the n unknowns a_k:

$$\sum_{k=1}^{n} (A\varphi_k, \varphi_j)\, a_k = (f, \varphi_j) \qquad j = 1, \ldots, n \tag{1.3}$$

Closely connected with the Galerkin method is the method of weighted residuals. In this method, a second set of linearly independent functions $\{\psi_n\}$ in D_A, called weight functions, is selected. The n^{th} approximate solution is again given by (1.2), but the coefficients in this case are obtained as the solution of the following system:

$$\sum_{k=1}^{n} (A\varphi_k, \psi_j) a_k = (f, \psi_j) \qquad j = 1, \ldots, n \tag{1.4}$$

In connection with these and similar methods there are several outstanding and very practical problems. It is desireable to have some definition of an optimal approximation to the exact solution using methods of this type. Once such a definition is given, the question of selecting a set of optimal coordinate functions and a set of optimal weight functions remains. In addition, the question of the convergence of the approximate solutions to the exact solution is ever present. Furthermore, from a practical point of view, it is important to develop computational techniques which make efficient use of these methods as well as to anticipate any computational problems involved.

The purpose here is to attack the above problem areas as well as to present a concise, lucid guide to efficient and economical application of these methods and their modifications. In addition a rather complete classification of Galerkin type methods is given with existing convergence theorems. Finally, the last chapter includes a bibliography which represents a rather extensive search. Most of the references not included are those which make a straight-forward application of a method.

Chapter II

GENERAL APPROXIMATION THEORY

It is the purpose of this chapter to state the general definitions used in the sequel and to present an outline of approximation theory [49] from a somewhat general point of view which will include some later methods as special cases.

A. General Definitions

A linear space over the field of real (complex) numbers is a set V of elements called points or vectors satisfying the following axioms.

(1) To every pair, f and g, of elements in V there corresponds an element $f + g$ called the sum of f and g. With respect to the operation $+$, V is an abelian group, i.e.

(a) $+$ is commutative, $f + g = g + f$,

(b) $+$ is associative, $f + (g + h) = (f + g) + h$,

(c) there exists in V a unique element 0 such that $f + 0 = f$ for every f in V, and

(d) for each f in V there corresponds a unique element $-f$ in V such that $f + (-f) = 0$.

(2) To every pair, α and f, where α is a real (complex) number and f is an element of V, there corresponds an element αf in V, called the product of α and f, such that

(a) scalar multiplication is associative, $(\alpha\beta)f = \alpha(\beta f)$,

(b) $1f = f$, $0f = 0$ for every f in V,

(c) scalar multiplication is distributive with respect to vector addition, $\alpha(f + g) = \alpha f + \alpha g$, and

(d) multiplication by vectors is distributive with respect to scalar addition, $(\alpha + \beta)f = \alpha f + \beta f$.

Elements $f_1, \ldots, f_n$ in V are <u>linearly independent</u> if the relation

$$\alpha_1 f_1 + \ldots + \alpha_n f_n = 0$$

holds only for $\alpha_1 = \ldots = \alpha_n = 0$; otherwise $f_1, \ldots, f_n$ are <u>linearly dependent</u>. $\alpha_1 f_1 + \ldots + \alpha_n f_n$ is called a <u>linear combination</u> of the elements $f_1, \ldots, f_n$. A linear space V is <u>n-dimensional</u> if V contains n linearly independent elements and if any set of $n + 1$ elements in linearly dependent. If for each positive integer n, V contains n linearly independent elements, then V is <u>infinite dimensional</u>. A non-empty subset M of V is a <u>subspace</u> if for every pair, f_1 and f_2, of elements of M, every linear combination, $\alpha_1 f_1 + \alpha_2 f_2$, is also contained in M. If $f_1, \ldots, f_n$ are linearly independent in V, the space of all linear combinations of $f_1, \ldots, f_n$ is called the <u>subspace spanned by</u> $f_1, \ldots, f_n$ and is denoted by $S_n(f)$ and $f_1, \ldots f_n$ is called a <u>basis</u> for $S_n(f)$.

A linear space V is an <u>inner product space</u> if for each pair of elements, f and g, in V there is a real (complex) number (f, g),

called the inner (or scalar) product of f and g which satisfies the following conditions:

(1) $(g, f) = \overline{(f, g)}$, the complex conjugate of (f, g),

(2) $(\alpha_1 f_1 + \alpha_2 f_2, g) = \alpha_1 (f_1, g) + \alpha_2 (f_2, g)$,

(3) $(f, f) \geq 0$, and

(4) $(f, f) = 0$ if and only if $f = 0$.

The <u>norm</u> of the element f in V, denoted by $\| f \|$, is defined by

$$\| f \| = \sqrt{(f, f)}$$

The following properties of the norm can be derived from the properties of the inner product.

(1) $\| \alpha f \| = | \alpha | \, \| f \|$

(2) $| (f, g) | \leq \| f \| \, \| g \|$, the Cauchy inequality

(3) $\| f + g \| \leq \| f \| + \| g \|$, the triangle inequality

Two elements f and g are said to be <u>orthogonal</u> if

$$(f, g) = 0.$$

An element f is <u>normalized</u> if

$$\| f \| = 1.$$

An inner product space V becomes a metric space if the distance between two elements, f and g, in V, denoted by $d(f, g)$ is defined as

$$d(f, g) = \|f - g\|.$$

From the properties of the norm, it can be shown that the function d satisfies the usual conditions for a metric, i.e.

(1) $d(f, g) = d(g, f)$

(2) $d(f, g) \leq d(f, h) + d(h, g)$

(3) $d(f, g) \geq 0$

(4) $d(f, g) = 0$ if and only if $f = g$.

If V is a metric space and $\{f_n\}_{n=1}^{\infty}$ is a sequence of elements of V, then f in V is said to be the limit of the sequence f_n, written $f_n \to f$ or $\lim_{n\to\infty} f_n = f$, if

$$\lim_{n\to\infty} d(f_n, f) = 0.$$

The sequence $\{f_n\}_{n=1}^{\infty}$ is said to converge to f. This type of convergence is called strong convergence or norm convergence. The sequence $\{f_n\}_{n=1}^{\infty}$ is said to be a Cauchy (or fundamental) sequence, if for each $\epsilon > 0$ there is an integer N_ϵ such that for all $m, n > N_\epsilon$

$$d(f_m, f_n) \leq \epsilon.$$

A metric space V is called complete if every Cauchy sequence converges to an element of V.

A Hilbert space H is an infinite dimensional inner product space which is a complete metric space with respect to the metric induced by the inner product. H is a separable Hilbert space if H contains a sequence of elements $\{f_n\}_{n=1}^{\infty}$ such that $\overline{S(f)}$, the closure of $S(f)$, that is, the subspace spanned by the set $\{f_n\}$ along with limits of all Cauchy sequences in $S(f)$, is the space H. Thus if H is a separable Hilbert space, each element f of H can be approximated with arbitrarily prescribed accuracy by a finite linear combination of the elements $\{f_n\}_{n=1}^{\infty}$.

The space $L_2[a, b]$ of functions which are square integrable (in the Lebesque sense) on $[a, b]$ is an example of a separable Hilbert space. In this case, denoting by μ the measure on $[a, b]$, the inner product of f and g in $L_2[a, b]$ is defined as

$$(f, g) = \int_a^b fg d\mu$$

and the norm of f thus becomes

$$\| f \| = \sqrt{(f, f)} = \left[\int_a^b f^2 d\mu\right]^{\frac{1}{2}}.$$

Norm convergence in L_2 is called mean convergence.

Let G be a subspace of a separable Hilbert space H. It can be shown that to each element h in H there corresponds a unique element

g in G such that

$$\|h - g\| = \inf_{g' \in G} \|h - g'\|.$$

Furthermore, $h - g$ is orthogonal to the subspace G, i.e., for each g' in G

$$(h - g, g') = 0,$$

and h can be represented in the form

$$h = g + f$$

where $g \in G$ and f is orthogonal to G. The element g is called the orthogonal projection of h on the subspace G.

A set $M = \{f_n\}_{n=1}^{\infty}$ which satisfies

(1) $(f_m, f_n) = \begin{cases} 0 \text{ for } m \neq n \\ 1 \text{ for } m = n, \text{ and} \end{cases}$

(2) $(f_n, f_n) = 1$ for $n = 1, 2, \ldots$

is called an orthonormal system or sequence. From a given set of linearly independent (finite or countable) set of elements, $\{g_n\}$, it is possible to construct an orthonormal set, $\{f_n\}$, so that each f_n is a linear combination of the g_m, $m = 1, \ldots, n$. One such method of orthonormalization is the Gram-Schmidt process. An orthonormal (linearly independent) system is said complete if it is not contained in any larger orthonormal (linearly independent) set.

Let D denote a subspace of the separable Hilbert space H. A function T which relates to each element f in D a particular element $Tf = g$ in H is called an operator in the space H with domain D. The set

$$\Delta_T = \{g \epsilon H : Tf = g, \ f \epsilon D\}$$

is called the range of T. The domain of T is denoted by D_T. The operator which maps each element onto itself is called the identity operator and is denoted by E.

If the operator T is one-to-one, that is, for each distinct pair, f_1 and f_2, in D_T, $Tf_1 \neq Tf_2$ in Δ_T, then T has an inverse denoted by T^{-1} which maps Δ_T onto D. Furthermore, $T^{-1}g = f$ if and only if $Tf = g$.

Let S and T be two operators such that $\Delta_T \cap D_S$ is not empty. Then the product of S and T is defined as

$$STf = S(Tf).$$

An operator T is linear if for each pair f and g in D_T and scalars α and β

$$T(\alpha f + \beta g) = \alpha Tf + \beta Tg.$$

The norm of an operator T, denoted by $\| T \|$, is defined as

$$\| T \| = \sup \| Tf \|$$

as f ranges over D_T and $\|f\| \leq 1$.

An operator T is bounded if $\|T\|$ is finite.

The operator T is said to be continuous at a point f_o in D_T if for each $\epsilon > o$ there exists a $\delta = \delta(\epsilon) > o$ such that for each f in D_T with

$$\|f - f_o\| < \delta \text{ we have } \|Tf - Tf_o\| < \epsilon.$$

If T is an arbitrary bounded linear operator on H, there exists a unique bounded linear operator T^* such that $\|T\| = \|T^*\|$ and for each f and g in H

$$(f, T_g) = (T^*f, g).$$

The operator T^* is called the adjoint of T. It is easy to see that $(T^*)^* = T$. If $T^* = T$, then T is said to be self-adjoint.

A self-adjoint operator T is defined to be positive if for each f in D_T

$$(Tf, f) \geq 0.$$

If $(Tf, f) = 0$ implies $f = 0$, then T is said to be positive definite. The energy product of two elements f and g in D_T is (Tf, g), and the energy norm, denoted by $\|f\|_T$ is

$$\|f\|_T = \sqrt{(Tf, f)}.$$

A sequence of elements $\{f_n\}_{n=1}^{\infty}$ in D_T is said to converge in energy to the element f in D_T if

$$\lim_{n\to\infty} \|f_n - f\|_T = 0.$$

It can be shown that energy convergence implies mean convergence [38].

A linear operator T is said to be symmetric if D_T is dense in H and for f and g in D_T

$$(Tf, g) = (f, Tg).$$

It is easily verified that a symmetric operator T such that $D_T = H$ is self-adjoint.

A set C is compact if every infinite sequence in C contains a convergent subsequence. A linear operator T such that $D_T = H$ is said to be completely continuous if it maps each bounded set into a set which is compact.

Let H be a separable Hilbert space and G be a subspace of H such that each element h in H has a unique representation in the form

$$h = g + f$$

where g is the orthogonal projection of h on G. The operator which maps h into its orthogonal projection g on G is called the orthogonal projection operator and is denoted by P_G or simply P if the subspace G has been previously specified. An orthogonal projection operator is linear, bounded, and $\| P \| = 1$. In addition, $P^2 = P$ and $P^* = P$.

A complex number λ is called an eigenvalue of the linear operator T if there exists a element $f \neq 0$ such that

$$Tf = \lambda f.$$

The vector f is called an eigenvector belonging to the eigenvalue λ.

A subspace G of H is called an invariant subspace of the operator T if every element of D_T in G is mapped by the operator T into an element also in G.

Given a complete orthonormal set $\{f_n\}$ in $\overline{S(f)}$ each function g has the unique representation

$$g = \sum_{i=1}^{\infty} (g, f_i) f_i$$

and this is called the Fourier series of g relative to the set $\{f_n\}$. The orthogonal projection of g on $S_n(f)$ can be written

$$\sum_{i=1}^{n} (g, f_i) f_i .$$

B. Theorems

Let $\overline{X}$ be a complete subspace of the normed space X and P denote a projection from X onto $\overline{X}$, i. e.

$$P(X) = \overline{X}; \; P^2 = P.$$

We first consider an equation of the form

$$Kx \equiv x - \lambda Hx = y \tag{2.1}$$

in the space X, called the exact equation, and an equation

$$\overline{K}\overline{x} \equiv \overline{x} - \lambda \overline{H}\overline{x} = Py \tag{2.2}$$

in $\overline{X}$, called the approximate equation. The operators H and $\overline{H}$ are linear.

In order to make the rather abstract setting more easily understood, we will take an example to illustrate the theory as we proceed.

Let $\{\varphi_i\}$ be an orthonormal sequence in L_2. Let $X = \overline{S(\varphi)}$ and $\overline{X} = S_n(\varphi)$. Then the operator P could be defined as the orthogonal projection from X onto $\overline{X}$.

The following conditions will play an important role in the sequel.

I. For every x in $\overline{X}$

$$\| PH\overline{x} - \overline{H}\overline{x} \| \leq \eta \| \overline{x} \|.$$

II. For every x in X there is an $\overline{x}$ in $\overline{X}$ such that

$$\| Hx - \overline{x} \| \leq \eta_1 \| x \|.$$

III. For each y in X there is a $\overline{y}$ in $\overline{X}$ such that

$$\| y - \overline{y} \| \leq \eta_2 \| y \|.$$

IV. If a solution exists for equation (2.2) then the solution is unique.

In our example, condition I could be satisfied if $\overline{H} = PH$. In such case $\eta = 0$ is an acceptable choice. Furthermore, conditions II and III would be satisfied provided n is sufficiently large since the sequence $\{\varphi_i\}$ is complete.

Condition IV will always hold for completely continuous operators.

We hope to be able to solve the approximate equation and have an approximation to the exact equation. Suppose that we have a sequence of approximate equations and solutions obtained from these. Then the space

$\overline{X}$, together with the operators $\overline{H}$ and $\overline{P}$ and the constants in conditions I, II, and III depend on an index n, which we shall omit for the sake of simplicity.

The obvious question is whether or not such a sequence of approximations converges to the exact solution.

Theorem 2.1. If the following conditions are satisfied:

(1) K has a linear inverse

(2) $\overline{K}$ satisfies IV for each n

(3) I, II, and III hold for each n

(4) $\lim_{n\to\infty} \eta = 0$; $\lim_{n\to\infty} \eta \|P\| = 0$; $\lim_{n\to\infty} \eta_2 \|P\| = 0$

then the approximate equations are soluable for sufficiently large n and

$$\lim_{n\to\infty} \|x - \overline{x}_n\| = 0.$$

These sufficient conditions for convergence can be phrased in another form when $\overline{H} = PH$.

Theorem 2.2. If $\overline{H} = PH$ and the following conditions are satisfied:

(1) X is complete

(2) $\lim_{n\to\infty} P_n x = x$

(3) H is completely continuous

(4) λ is not the reciprocal of an eigenvalue of H

then the approximate solutions converge to the exact solution.

Corollary: The characteristic values of H are limits of sequences of characteristic values of $\overline{H}_n$.

In our example, X is complete and the orthogonal projectors converge to the identity so (1) and (2) of Theorem 2.2 are satisfied.

We shall now consider equations in which the left-hand side does not contain the identity operation explicitly and moreover represents an operation from the original space not into itself but into another normed space. In spite of these differences as compared with the equations discussed above, these are reducible to the former type.

We suppose that X and Y are normed spaces, from each of which complete subspaces $\overline{X}$ and $\overline{Y}$, respectively, have been chosen. We suppose that F is a linear operator projecting Y onto $\overline{Y}$.

As before, we consider two equations, the exact

$$K_1 x \equiv Gx - \lambda Tx = y_1 \tag{2.3}$$

and the corresponding approximate equation

$$\overline{K}_1 \overline{x} \equiv G\overline{x} - \lambda \overline{T}\overline{x} = Fy_1. \tag{2.4}$$

Here, G and T (and K_1) are linear operators mapping X into Y while $\overline{T}$ (and $\overline{K}_1$) are linear operators from $\overline{X}$ into $\overline{Y}$. We further assume that G has a linear inverse and establishes a one-to-one correspondence between $\overline{X}$ and $\overline{Y}$. In this setting conditions I through III above become:

Ib. For every x in X

$$\|F\overline{T}\overline{x} - \overline{T}\overline{x}\| \leq \mu \|\overline{x}\|$$

IIb. For every x in X there is a $\overline{y}$ in $\overline{Y}$ such that

$$\|Tx - \overline{y}\| \leq \mu_1 \|x\|$$

IIIb. For each y_1 in Y there is an element $\overline{y}_1$ in $\overline{Y}$ such that

$$\|y_1 - \overline{y}_1\| \leq \mu_2 \|y_1\|$$

Writing, under the above assumptions, (2.1) and (2.2) as

$$Kx \equiv G^{-1}K_1x \equiv x - \lambda G^{-1}Tx = G^{-1}y_1 \tag{2.5}$$

$$\overline{K}\overline{x} \equiv G_1^{-1}\overline{K}_1\overline{x} \equiv \overline{x} - \lambda G^{-1}\overline{T}\overline{x} = G^{-1}Fy_1 \tag{2.6}$$

it can be shown that conditions I through III hold.

The preceding theorems can be restated in this setting.

Theorem 2.3. If the following conditions are satisfied:

(1) K_1 has a linear inverse,

(2) $\overline{K}_1$ satisfies IV for each n,

(3) Ib, IIb, IIIb are satisfied for each n,

(4) $\lim_{\mu\to\infty} \mu = 0$; $\lim_{\mu\to\infty} \mu_1 \|F\| = 0$; $\lim_{\mu\to\infty} \mu_2 \|F\| = 0$,

then the approximate equations are soluable for sufficiently large n, and the sequence of approximate solutions converges to the exact solution.

As before when the approximate equation is obtained by projection of the exact equation, the theorem can be formulated differently.

Theorem 2.4. If $\overline{T} = FT$ and the following conditions are satisfied:

(1) Y is complete

(2) $\lim_{n\to\infty} F_n x = x$

(3) $G^{-1}T$ is completely continuous

(4) K_1^{-1} exists

then the approximate equations are soluable for sufficiently large n, and the approximate solutions converge to the exact solution.

C. Summary

From the above we conclude that when presented an operator equation to solve, we must decide into which general category it falls. That is, decide whether it is of type 1 with the identity explicit or type 2. Then several other definitions must be made. There is a great deal of flexibility in these, but certain choices are sometimes more judicious than others. There is no general rule since the choice is influenced by the character of the operators as well as the boundary conditions and the particular method of solution employed.

The following is an outline of the situation.

Type 1. $x - \lambda Hx = y$

Define: (1) the space X and the subspace $\overline{X}_n$

(2) the projection P_n

(3) the operator $\overline{H}_n$

At this point the approximate equation is determined, and it may be possible to check conditions I through IV. It must then be determined if either Theorem 2.1 or Theorem 2.2 applies to prove that the approximate solutions converge.

Regarding type 2, it should be pointed out that the first step here is to write the given equation in the proper form. There may be more than one way to do this, and one version may be more amenable than another. Again, there is no general principle. However, it must be kept in mind that G must be one-to-one and have a linear inverse.

Type 2. $Gx - \lambda Tx = y$

Define: (1) the spaces X, $\overline{X}$, Y and $\overline{Y}$

(2) the projection F

(3) the operator $\overline{T}$

As above the approximate equation is now determined and, perhaps, condition Ib through IIIb, and IV can be checked. Also it remains to determine if the theorems apply to prove convergence.

The above procedures are exhibited in the examples.

Chapter III

CLASSIFICATION OF METHODS OF THE GALERKIN TYPE

Since Galerkin first used his technique to solve a problem in mechanics, a vast number of modifications have been suggested. The purpose of this section is to classify these suggested modifications and to outline their application. Certain of these fall into the general approximation theory considered in the last chapter. Some of them have been rather widely used but are without convergence proofs. These will be carefully noted along with the limitations of the other methods.

We first state the definitions applicable to this chapter, then classify and describe the methods and in the final section summarize the results and draw certain conclusions.

The methods all involve the solution of the equation

$$L(u) = 0 \tag{3.1}$$

where u is defined on a domain D bounded by C, L is an operator in a separable Hilbert space and on C

$$B_i(u) = g_i, \; i = 1, \ldots, p \tag{3.2}$$

are the appropriate number of boundary conditions. It should be noted here, and will be pointed out below, that the methods discussed are primarily designed for boundary valve problems and certain careful modifications must be made to even apply any of these methods to initial value or other types of problems. The methods to be described involve the selection

of two linearly independent sets of functions: the first called coordinate functions $\{v_i\}_{i=1}^{\infty}$ and the second called weight functions $\{w_i\}_{i=1}^{\infty}$.

An n^{th} approximation to the solution is obtained in the form

$$u_n = f(c_1, \ldots, c_n, v_1, \ldots v_n) \tag{3.3}$$

where f is a function of some selected form (e.g., a linear combination) and the functions (perhaps constants) c_i are determined by one of the procedures described below.

A. Definitions

Definition 3.1: The operator residual, denoted by $R_L(u_n)$, is the result of operating by L on u_n. Thus $R_L(u_n) = L(u_n)$.

Definition 3.2: The component boundary residuals, denoted by $R_{B_i}(u_n)$, are the result of substituting u_n for u in the boundary conditions. Thus $R_{B_i}(u_n) = B_i(u_n) - g_i$, $i = 1, \ldots, p$.

Definition 3.3: The boundary residual, denoted by $R_B(u_n)$, is the sum of the absolute values of the component boundary residuals. Thus $R_B(u_n) = \sum_{i=1}^{p} \left| R_{B_i}(u_n) \right|$.

Definition 3.4: A method of solution is called an interior method provided that the coordinate functions are chosen to satisfy the boundary conditions and the functions $\{c_i\}$ are determined by requiring that the operator residual be orthogonal to the first n weight functions throughout the region D. That is

$$(R_L(u_n),\ w_j) = 0 \quad j = 1,\ \ldots,\ n. \tag{3.4}$$

Definition 3.5: A method of solution is called a boundary method provided that the coordinate functions are chosen to satisfy the operator equation, and the functions $\{c_i\}$ are determined by requiring that the boundary residual be orthogonal to the first n weight functions on the boundary C. That is

$$(R_B(u_n),\ w_j) = 0 \quad j = 1,\ \ldots,\ n. \tag{3.5}$$

Definition 3.6: A method of solution is called a mixed method if it is neither interior nor boundary.

It should be noted that generally in the mixed method described above both the boundary and operator residuals must be considered. If we require that both the boundary residual and the operator residual be orthogonal to the first n weight functions, we will obtain $2n$ equations for the n unknowns $\{c_i\}$. It has been suggested that such a situation be remedied by:

(1) Discarding some of the equations; the choice is arbitrary as long as the equations are independent although consideration is given to those characteristics of the problem which are most important [29]. Exactly how such a decision should be made is not stated.

(2) Using $\frac{n}{2}$, instead of n, weight functions (n must, of course, be even) [13].

(3) Adding the residuals and making this sum orthogonal to the first n weight functions [36].

None of these suggestions are proved to be valid and are merely ways of obtaining n equations in n unknowns. The sole motivation for the rather arbitrary choice seems to be a matter of taste.

B. Specific Methods

There are four fundamental classifications of methods of the Galerkin type.

I. General: This is simply to solve the system (3.4) using the form (3.3) as it stands. Unfortunately, it can be quite difficult to achieve such a solution, and a certain amount of ingenuity may be required to select an appropriate form for f, to satisfy the boundary conditions, etc.

II. Shulesko Technique: This is a mixed method wherein $p + 1$ sets of coordinate functions are chosen, $\{v_i\}_{i=1}^{\infty}$ and $\{t_{ki}\}_{i=1}^{\infty}$, $k = 1, \ldots p$. The n^{th} approximation is sought in the form

$$u_n = \sum_{k=1}^{p} \sum_{i=1}^{n} a_{ki} t_{ki} + \sum_{i=1}^{n} b_i v_i$$

Furthermore two complete sets of weight functions $\{s_j\}$, $\{w_j\}$ are selected. The $\{a_{ik}\}$ are then determined by making the component boundary residuals orthogonal to the first n weight functions s_j. The b_i are determined by making the operator residual orthogonal to the first n weight functions w_j. [41].

III. Method of Weighted Residuals: A set of functions $\{v_i\}_{i=1}^{\infty}$ is chosen to satisfy the boundary conditions. The n^{th} approximation

takes the form $u_n = \sum_{i=1}^{n} c_i v_i$ and the $\{c_i\}$ are determined by making the operator residual orthogonal to the first n of any set of linearly independent weight functions $\{w_j\}_{j=1}^{\infty}$ [48].

IV. Classical Galerkin: A set of coordinate functions of the required number of variables, linearly independent and complete, is chosen to satisfy the boundary conditions. The n^{th} approximation has the form $u_n = \sum_{i=1}^{n} c_i v_i$ where the $\{c_i\}$ are constants determined by making the operator residual orthogonal to the set of coordinate functions. Thus the coordinate functions themselves become the weight functions [38].

Notice that II includes III in the case in which $\{t_{ki}\} \equiv \{v_i\}$ and satisfy the boundary conditions; it includes IV in the case in which $\{t_{ki}\} \equiv \{v_i\}$ and satisfy the boundary conditions and $\{w_j\} \equiv \{v_i\}$. In addition, III includes IV when $\{v_i\} \boxminus \{w_j\}$.

There are several slight modifications and variations on the above theme. These depend on the specific characteristics of the problem to be solved; for example, whether or not the operator is linear, the boundary conditions homogeneous, etc. Some of these are indicated below.

(1) Ames Method

When the problem is an initial value problem, where the range of t is the non-negative real numbers, the n^{th} approximation is taken in the

form $u_n = \sum_{i=1}^{n} c_i v_i$ where the coordinate functions satisfy the initial conditions. The operator residual is then made orthogonal to a set of weight functions (according to III or IV) on the interval $[0, T]$ where T is regarded as the time when steady state is achieved. Hopefully, T may be estimated from the physical nature of the problem [31]. See Chapter V for a further discussion of this.

(2) Bickley Method

When a linear problem involves the variables x and t, the n^{th} approximation may be taken in the form $u_n = \sum_{i=1}^{n} c_i(t)v_i(x)$. Thus the members of $\{c_i(t)\}$ are unknown functions of t and may be determined from the conditions of III or IV [18].

(3) Crandall Method

When L is a partial differential operator, it is possible to reduce the problem to an ordinary differential system if we are given in addition to (3.1) and (3.2) an appropriate number of initial conditions:

$$I_k(u) = h_k \qquad k = 1, \ldots, r \tag{3.6}$$

The n^{th} approximation is taken in the form $u_n = \sum_{1=1}^{n} c_i(t)v_i$ where $\{v_i\}_{i=1}^{\infty}$ is linearly independent, complete, and satisfies the boundary conditions but not the initial conditions or the differential equation. Thus, in addition to the operator residual, we define an initial residual to be:

$$R_I(u_n) = \sum_{k=1}^{r} \left| h_k - I_k\left(\sum_{i=1}^{n} c_i(0)v_i \right) \right|.$$

Making these residuals orthogonal to the first n of a complete set of weight functions $\{w_j\}_{j=1}^{\infty}$, we obtain the two systems of equations:

$$(a) \int_T R_I(u_n) w_j dT = 0$$

and

$$(b) \int_D R_L(u_n) w_j dD = 0$$

for $j = 1, \ldots, n$. From equation (b) we get a system of ordinary differential equations in the $\{c_i(t)\}$, and equation (a) will specify initial conditions for them, allowing solution either exactly or approximately [34].

(4) Generalized Moments Method

W. V. Petryshyn [53] has suggested a method for the solution of the eigenvalue problem

$$T_u - \lambda S u = 0 \tag{3.7}$$

where the operators T and S are linear, unbounded and nonsymmetric, and T is K-positive definite. An operator is said to be K-positive definite provided there is a closeable operator K with $D_K \geq D_T$ mapping D_T onto a dense subset of H and two positive constants α_1 and α_2 such that

$$(Tu, Ku) \geq \alpha_1 \|u\|^2 \qquad u \epsilon D_T$$

and

$$\|Ku\|^2 \leq \alpha_2 (Tu, Ku) \qquad u \epsilon D_T.$$

An operator K is closeable if $\{u_n\}$ converging to zero and $\{Ku_n\}$ converging to f imply f is zero. He also suggested the method for

$$Au + Bu = f \tag{3.8}$$

where A is K-positive definite, and B is a linear unbounded operator [54]. The method is the method of weighted residuals where the weight functions are chosen as the K-images of the coordinate functions. The author points out that if T (or A) is self-adjoint, the method reduces to the classical Galerkin method (i.e., $K = I$).

Certain of the above are special cases of the general theory of the preceding chapter. The classical Galerkin method is a special case in the following way. If the operator L has the form of type 1, the space X is taken as the completion of the subspace of L_2 spanned by coordinate functions and the subspace $\overline{X}$ as the subspace spanned by the first n coordinate functions. Further, P is taken as the orthogonal projection operator from X onto $\overline{X}$ and $\overline{H}$ is defined as PH. The approximate equation then gives, on taking the inner product with φ_j,

$$(\overline{x} - \lambda \overline{H}\,\overline{x},\ \varphi_j) = (Py,\ \varphi_j) \qquad j = 1,\ \ldots\ n.$$

Simplifying, using the self-adjointness of P and the fact that $P(\varphi_j) = \varphi_j$, $j \leq n$, gives

$$(\overline{x},\ \varphi_j) - \lambda\ (\overline{H}\,\overline{x},\ \varphi_j) = (y,\ \varphi_j) \qquad j = 1,\ \ldots\ n \tag{3.6}$$

which is exactly the Galerkin system of equations. Since

$$\overline{x} = \sum_{i=1}^{n} a_i \varphi_i$$

equation (3.6) gives n equations in the n unknowns a_i.

If the operator is of type 2, then X and $\overline{X}$ are taken as above. Further, Y and $\overline{Y}$ are taken as $G(X)$ and $G(\overline{X})$, respectively. The operator F is taken as the orthogonal projection from Y onto $\overline{Y}$. Then the approximate equation is

$$G\overline{x} - \lambda FT\overline{x} = Fy.$$

As above this gives rise to the system

$$(G\overline{x}, \varphi_j) - \lambda(FT\lambda, \varphi_j) = (Fy, \varphi_j)$$

This system determines the n^{th} approximation.

The method of weighted residuals can also be considered as a special case of the general theory under certain conditions. In this method the same approximate equations as above are employed, but the inner products are taken with weight functions $\{w_i\}$ other than the coordinate functions. This gives

$$(\overline{x} - \lambda\overline{H}\,\overline{x}, w_j) = (Py, w_j) \quad j = 1, \ldots n$$

and

$$(G\overline{x}, w_j) - \lambda(FT\overline{x}, w_j) = (Fy, w_j) \quad j = 1, \ldots n$$

for types 1 and 2, respectively. The self-adjointness of P (or F) can be invoked to write

$$(Py, w_j) = (y, Pw_j)$$

However, unless w_j is in the span of the first n coordinate functions, $P\, w_j \neq w_j$. Therefore, it can be concluded that the method of weighted residuals is a special case of the general theory only when the span of the first n coordinate functions includes the first n weight functions. (See in this connection Theorem 4.5 of Chapter IV.)

This same problem arises when the coordinate functions are not orthogonal. In this case the projection used is not orthogonal and hence not self-adjoint. That is, the projection is defined as truncation of the expansion in terms of the coordinate functions at n terms. This means, in particular, that the right hand side of the Galerkin system should be $(P\, y,\ \varphi_j)$ and not $(y,\ \varphi_j)$ to be a special case of the general theory. It does, indeed, seem clear that a convergence proof using $(P\, y,\ \varphi_j)$ would imply convergence using the right hand side $(y,\ \varphi_j)$ since $P\, y$ converges to y. However, the simple way out of this apparent dilemma about convergence as well as accuracy of corresponding approximations is to utilize the following theorem.

Theorem 3.1: If the coordinate functions $\{\varphi_i\}$ for the linear problem

$$Ay = f \quad .$$

are linearly independent but not orthogonal in L_2 then the n^{th} approximation is the same as the n^{th} approximation using the sequence $\{\varphi_i^*\}$ obtained by orthonormalization of $\{\varphi_i\}$.

Proof: Let G be the nonsingular $n \times n$ matrix which takes $\overline{\varphi}$ to $\overline{\varphi}^*$, i.e.,

$$\bar{\varphi}^* = \begin{Bmatrix} \varphi_1^* \\ \vdots \\ \varphi_n^* \end{Bmatrix} = G\bar{\varphi} = \begin{Bmatrix} \varphi_1 \\ \vdots \\ \varphi_n \end{Bmatrix}.$$

Let $P = (a_{i_j})$ be the $n \times n$ matrix where $a_{i_j} = (A\varphi_i, \varphi_j)$ and $Q = (b_{i_j})$ be the $n \times n$ matrix where $b_{i_j} = (A\varphi_i^*, \varphi_j^*)$. In addition, let

$$y_n = \bar{c}\,\bar{\varphi} \quad \text{and} \quad y_n^* = \bar{d}\,\bar{\varphi}^* = \bar{d}G\bar{\varphi}$$

where $\bar{c} = (c_1, \ldots, C_n)$ and $\bar{d} = (d_1, \ldots, d_n)$. The Galerkin systems are then

$$P\bar{c}' = R = \begin{Bmatrix} (f, \varphi_1) \\ \vdots \\ (f, \varphi_n) \end{Bmatrix}, \quad Q\bar{d}' = S = \begin{Bmatrix} (f, \varphi_1^*) \\ \vdots \\ (f_n, \varphi_n^*) \end{Bmatrix}$$

where prime denotes transpose.

If it can be shown that $\bar{c} = \bar{d}G$, the proof will be complete. We have

$$L(\bar{\varphi}^*) = GL(\bar{\varphi})$$

so that

$$L(\bar{\varphi}^*)\,\bar{d}^{*\prime} = GL(\bar{\varphi})\bar{\varphi}^{*\prime} = GL(\bar{\varphi})\bar{\varphi}'G'.$$

Integrating the extremes of this gives

$$Q = GPG'$$

so

$$P = G^{-1}QG'^{-1}.$$

Thus

$$c' = P^{-1}R = (G'Q^{-1}G)\ (G^{-1}S) = G'Q^{-1}S = G'\bar{d}'$$

and taking transposes gives the result.

The above theorem does answer the dilemma because orthonormal coordinate functions $\{\varphi_i^*\}$ gives

$$(Pf,\ \varphi_i^*) = (f,\ P\varphi_i^*) = (f,\ \varphi_i^*)$$

since P is an orthogonal projection, is self-adjoint and is, of course, the identity on its range. This means that the "correct" right hand side is used in the Galerkin system when the coordinate functions are orthogonal. However, by the theorem, the approximate solution with nonorthogonal linearly independent functions is the same even though the right hand side in this system is $(f,\ \varphi_i)$.

C. Summary

The number of convergence proofs for the above methods is quite small. As noted above in certain cases a few techniques are covered by the earlier theorems from the general theory.

The motivation for accepting these methods probably stems from the following well known theorem.

Theorem 3.2. An orthogonal sequence of functions is complete if and only if the only function orthogonal to each member of the sequence is equivalent

to zero. This theorem implies that if

$$(L(u_n), \varphi_j) = 0$$

for all j then

$$L(u_n) = 0$$

i.e., u_n is a solution, whenever $\{\varphi_j\}$ is a complete orthogonal sequence. Of course, only n values of j are used so that u_n is an "approximate" solution. In practice and in fact in most of the examples in the literature, the coordinate functions are neither orthogonal nor complete. Hence the appeal of this argument wanes.

The following is a summary of essentially all of the theorems proved in regard to convergence of the above methods.

Theorem 3.3. If the sequence of coordinate functions is complete, the classical Galerkin method converges in the mean to the exact solution of a Fredholm integral equation [38].

Theorem 3.4. Suppose $Au = f$ has a unique solution and can be written as

$$A_o u + Ku = f$$

where A_o is symmetric and positive-bounded-below in a Hilbert space H. Then if $A_o^{-1} K$ is completely continuous, the classical Galerkin method with a complete sequence of coordinate functions converges in energy to the exact solution. A similar result holds for the eigenvalue problem [38]. As an application of Theorem 3.4 the following result can be proven.

Theorem 3.5. Consider the problem

$$(-1)^m u^{(2m)} - \lambda Ku = f$$

$$u(a) = u'(a) = \ldots = u^{(m-1)}(a) = 0$$

$$u(b) = u'(b) = \ldots = u^{(m-1)}(b) = 0$$

where K is a linear differential operator of order $2m-1$ such that all coefficients are sufficiently differentiable. Suppose there is a unique solution. Then the classical Galerkin method with a complete sequence of coordinate functions converges to the exact solution. A similar result holds for the eigenvalue problem [38].

Theorem 3.6. Suppose, on some domain D,

$$-\sum_{i,\,k=1}^{m} \frac{\partial}{\partial x_i}\left(A_{i_k} \frac{\partial u}{\partial x_k}\right) + \sum_{i=1}^{m} B_i \frac{\partial u}{\partial x_i} + Cu = f$$

where u vanishes on the boundary of D, the coefficients may be variable, and the equation is elliptic. Then the classical Galerkin method with a complete sequence of coordinate functions converges in energy to the exact solution [38].

Theorem 3.7. Suppose, on the interval $0 \leq x \leq \pi$ and for $t \geq 0$,

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial t} - gu = f$$

where

$$u(x, 0) = u(0, t) = u(\pi, t) = 0,$$

f, g and their first two derivatives are continuous and $f(0, 0) = f(\pi, 0) = 0$. Then Bickley's variation of the classical Galerkin method

with $\{\sin nx\}$ as coordinate functions converges to the exact solution [47].

Theorem 3.7 can be slightly generalized. For example a first derivative term in x could be present and the coefficient of the first derivative in t could be a positive function, and the same conclusion will hold.

The following is a generalization of Theorem 3.4.

Theorem 3.8. If the equation

$$Au + Bu = f$$

has a unique solution and $T = A^{-1}B$ can be extended to a completely continuous operator then for sufficiently large n, there is a unique n^{th} approximate solution from the generalized moments methods, and this sequence converges to the solution [54].

Recall that in using the generalized moments method A should be K-positive definite. The practical difficulty in making use of Theorem 3.8 is that there is no method to find K or even to decide if a given operator is K-positive definite.

Considering the above theorems, it is apparent that the convergence of the many suggested variations remains for the most part unproved. Papers dealing with a method per se are few compared with those that simply apply a method without regard to convergence. The acceptance of these techniques appears to stem from the fact that the solutions

"agree" in some sense with either experiment or approximate solutions obtained in other ways.

Chapter IV

RESULTS ON METHODS OF THE GALERKIN TYPE

Most of the theorems, definitions and computational techniques stated in this chapter could not be found in the literature. The theorems either generalize existing results or give new information. The techniques of solution of a problem are either a consequence of a theorem or the result of experience with examples.

We will consider throughout this chapter the problem

$$Ay = f \tag{4.1}$$

subject to homogeneous boundary conditions where A is a self-adjoint linear operator defined in a separable Hilbert space. If the boundary conditions are not homogeneous a slight modification to the problem will usually make them so. In particular we investigate the question of selecting "optimal" sets of coordinate functions and weight functions for the method of weighted residuals.

Throughout this chapter, it is assumed that (4.1) has a unique solution which will be denoted by y^*. In particular, then, the operator L is one-to-one and has the property that for any linearly independent set $\{\varphi_i\}_{i=1}^{n}$ in D_L, the set $\{L(\varphi_i)\}_{i=1}^{n}$ is linearly independent in the range of L.

A. Definitions

Definition 4.1. A linearly independent set of functions $\{\varphi_i\}_{i=1}^{n}$ is an L_2 (energy) optimal Galerkin coordinate set with respect to the

operator A provided that the Galerkin method using this set of functions as coordinate functions gives the best L_2 (energy) approximation in $S_n(\varphi)$ to y^*.

Definition 4.2. Suppose each of $\{\varphi_i\}_{i=1}^{\infty}$ and $\{w_i\}_{i=1}^{\infty}$ is a linearly independent set of functions. Then $(\{\varphi_i\}, \{w_i\})$ is called an L_2 (energy) optimal weighted residual pair with respect to A provided this pair used in the method of weighted residuals gives the best L_2 (energy) approximation in $S_n(\varphi)$ to y^*.

Definition 4.3. The operators A_1 and A_2 with the same domain are strongly similar provided there is a linear operator P such that $PA_1 = A_2P$.

Definition 4.4. The operators A_1 and A_2 with the same domain are similar provided there exist continuous linear operators P_1 and P_2 such that $A_1P_1 = A_2$ and $A_2P_2 = A_1$.

Definition 4.5. A pair of linearly independent sets of functions $(\{\varphi_i\}, \{w_i\})$ is an L_2 (energy) almost optimal weighted residual pair with respect to A provided there is a number C such that for all i

$$\| A(w_i) - \varphi_i \| < C$$

where the norm used in the L_2 (energy) norm. It is obvious that optimal implies almost optimal.

Definition 4.6. Suppose $\{\varphi_i\}$ is a linearly independent set of coordinate functions for the Galerkin method and y_n is the n^{th}

approximate solution to (4.1). Let B_φ denote the operator which maps the right hand side, f, to y_n and $R_\varphi = A^{-1} - B_\varphi$. Then $\{\varphi_i\}$ will be called an R-optimal Galerkin coordinate set if for all sets $\{\psi_i\}$

$$\| R_\varphi \| \leq \| R_\psi \| .$$

The set will be called almost R-optimal if there is a positive number C such that

$$\| R_\varphi \| \leq C \inf \| R_\psi \|$$

where the inf is taken over all $\{\psi_i\}$ in the space.

In connection with the above definition it should be pointed out that examination of the norm of the remainder operator could very well not be an indication of the smallness of $\| y^* - y_n \|$ since the operator norm is defined as

$$\| R_\varphi \| = \inf \| R_\varphi g \|$$

and there could be some function g such that

$$\| R_\varphi \| < \| R_\varphi g \| < \| R_\varphi f \| = \| y^* - y_n \| .$$

B. Fundamental Theorems

The only previously published attempt to find an optimal use of the methods considered herein is found in [45]. The approach used in this reference is to introduce the nonstandard energy norm. (See Definitions,) This allows certain interesting results to be established.

In [45] the problem (4.1) with homogeneous boundary conditions where A is a positive definite self-adjoint operator is investigated. They consider the domain of A to be L_2, H_1, and H_2 respectively, which are determined by the three inner products (u, v), (Au, v) and (Au, Av).

In this setting the following theorems are obtained by these authors.

Theorem 4.1. The best n^{th} approximation in energy to the solution of (4.1) is given by the classical Galerkin method regardless of the coordinate system. If the coordinate functions are complete, then the approximations converge in energy to the solution.

Theorem 4.2. The set $\{\varphi_i\}$ is an R-optimal coordinate set for (4.1) if the span of the first n elements is the same as the span of the first n eigenfunctions of the problem

$$Au = \lambda u$$

with homogeneous boundary conditions.

Theorem 4.3. An almost R-optimal set for a problem defined by an operator A_1 is an almost R-optimal set for a problem defined by an operator A_2 if A_1 and A_2 are similar.

These same theorems are also proved in the space obtained using the inner product (Au, Av).

It is clear that the optimal approximation to the solution y^* by a function in $S_n(\varphi)$ will be the orthogonal projection, y_n, of y^* onto $S_n(\varphi)$. Also

$$y^* = y_n + y_n' , \quad (y_n' , \varphi_i) = 0, \ i = 1, \ \ldots \ n.$$

in whatever space is being considered.

The following result from this investigation is in the spirit of the above theorems.

Theorem 4.4. In the norm induced by (Au, v) the pair $(\{\varphi_i\}, \{A\varphi_i\})$ is an optimal weighted residual pair for (4.1) for any linearly independent coordinate set $\{\varphi_i\}$.

Proof: We know

$$[y_o, \varphi_i] = [y_n, \varphi_i].$$

That is,

$$(Ay_o, A\varphi_i) = (Ay_n, A\varphi_i).$$

However $Ay_o = f$ and letting $A\varphi_i = w_i$ gives

$$(f, w_i) = (Ay_n, w_i)$$

which is the weighted residual system. Notice that the self-adjointness of A is not used. A somewhat stronger and much more useful result than Theorem 4.4 is the following for self-adjoint operators.

Theorem 4.5. The coordinate set $\{\varphi_i\}$ is an L_2 optimal Galerkin set for A in (4.1) if and only if $S_n(\varphi)$ is an invariant subspace of A.

Proof: Suppose $S_n(\varphi)$ is invariant under A. Then $A\varphi_j$ is in $S_n(\varphi)$, $j \leq n$ and

$$(y_o, A\varphi_j) = (y_n, A\varphi_j).$$

Since A is self adjoint this gives

$$(Ay_o, \varphi_j) = (Ay_n, \varphi_j)$$

or

$$(f, \varphi_j) = (Ay_n, \varphi_j)$$

which says y_n is the n^{th} Galerkin approximation.

Suppose $S_n(\varphi)$ is not invariant under A. Then for some $j \leq n$, $A\varphi_j$ is not in $S_n(\varphi)$. Then

$$(Ay_n, \varphi_j) = (y_n, A\varphi_j) \neq (y_o, A\varphi_j) = (Ay_o, \varphi_j) = (f, \varphi_j).$$

so that y_n is not the n^{th} Galerkin approximation.

Corollary: The eigenfunctions of

$$Au = \lambda u$$

subject to homogeneous boundary conditions form an L_2 optimal Galerkin coordinate set.

Proof: $S_n(\varphi)$ is invariant for each n and hence Theorem 4.5 gives the result.

Theorem 4.5 is useful since if a set of coordinate functions have been chosen, it can easily be determined if $S_n(\varphi)$ is an invariant subspace and hence if the Galerkin method yields the best L_2-approximation.

Once a set of coordinate functions is selected the question arises, "Is it possible to choose weight functions such that the pair of sequences will form an L_2-optimal weighted residual pair?" The next theorem answers this question in the affirmative under most circumstances. That is, for the problem (4.1) to have a solution, the function f, the right-hand side, must satisfy certain conditions. We will call such functions admissable. Then we have another theorem.

Theorem 4.6. Corresponding to any set $\{\varphi_i\}_{i=1}^{\infty}$ of linearly independent admissable functions which satisfy the homogeneous boundary conditions there exists a sequence $\{w_i\}_{i=1}^{\infty}$ of functions such that $(\{\varphi_i\}, \{w_i\})$ is an L_2-optimal weighted residual pair.

Proof: Let w_i be the solution to $Ay = \varphi_i$ subject to the homogeneous boundary conditions. Existence is assured since φ_i is admissable. We know that $(y_o, \varphi_j) = (y_n, \varphi_j)$ so that

$$(Ay_n, w_j) = (y_n, Aw_j) = (y_o, Aw_j) = (Ay_o, w_j) = (f, w_j).$$

Thus $(Ay_n, w_j) = (f, w_j)$ which is the system of equations for the method of weighted residuals. This proves $(\{\varphi_i\}, \{w_i\})$ is an L_2-optimal weighted residual pair.

Notice that if $S_n(\varphi)$ is invariant then any basis for $S_n(\varphi)$ and its image under A could be used as an optimal weighted residual pair. Also, in accordance with Theorem 4.5, any basis for $S_n(\varphi)$ is an L_2 optimal Galerkin coordinate set.

Theorem 4.7. If A_1 is strongly similar to A_2 and $\{\varphi_i\}$ is an optimal Galerkin coordinate set with respect to A_1 then $\{P(\varphi_i)\}$ is an optimal Galerkin coordinate set with respect to A_2.

Proof: By Theorem 4.5, $S_n(\varphi)$ is invariant under A_1. Since A_2 is strongly similar to A_1 we have

$$A_2P(S_n(\varphi)) = PA_1(S_n(\varphi)) = P(S_n(\varphi))$$

or $P(S_n(\varphi))$ is invariant under A_2 and again by Theorem 4.5 any basis for $P(S_n(\varphi))$ is an optimal Galerkin coordinate set with respect to A_2. This completes the proof.

Theorem 4.8. Suppose A_1 and A_2 are similar and that $(\{\varphi_i\}, \{w_i\})$ is an optimal weighted residual pair with respect to A_1. Then $(\{\varphi_i\}, \{P(w_i)\})$ is is an optimal weighted residual pair with respect to A_2.

Proof: We have $A_1w_i = \varphi_i$ and by similarity $A_2P_2(w_i) = A_1(w_i) = \varphi_i$. Hence the result follows.

Theorem 4.9. Suppose A_1 and A_2 are similar. Then an almost optimal weighted residual pair with respect to A_1 is an almost optimal weighted residual pair with respect to A_2.

Proof: Suppose $(\{\varphi_i\}, \{w_i\})$ are almost optimal for A_1. Since A_1 and A_2 are similar

$$\|A_2(w_i) - \varphi_i\| = \|P_1A_1(w_i) - \varphi_i\| \leq \|P_1A_1(w_i)\| + \|\varphi_i\|$$

Since P_1 is continuous and linear it is bounded so that there is a K such that

$$\|P_1A_1(w_i)\| \leq K \|A_1(w_i)\|.$$

Furthermore there is a number C such that

$$C > \|A_1(w_i) - \varphi_i\| \geq \big|\|A(w_i)\| - \|\varphi_i\|\big|$$

We can take the $\{\varphi_i\}$ such that $\|\varphi_i\| = 1$. Hence

$$\|A(w_i)\| < C + 1$$

so that

$$\|P_1A_1(w_i)\| \leq K(C + 1)$$

and the first inequality above gives

$$\|A_2(w_i) - \varphi_i\| \leq K(C + 1) + 1 \equiv M.$$

This completes the proof.

An additional and very significant result comes from using optimal coordinate and weight functions.

Theorem 4.10. If the optimal coordinate functions are used in the Galerkin method or an optimal pair is used in the method of weighted residuals then

the approximate solutions converge to the exact solution.

Proof: By optimality the approximate solutions are orthogonal projections of the exact solution on a sequence of monotonically increasing subspaces. Hence the approximate solutions converge to the exact solution.

From this theorem it is seen that if optimal coordinate and weight functions exist the method will converge.

Corollary to Theorem 4.6. If the conditions of Theorem 4.6 are satisfied the method of weighted residuals converges using the optimal pair.

Finally, it can be remarked that when optimal functions are used the sequence of norms $\| y_n - y^* \|$ is nonincreasing. This follows since each y_n is the orthogonal projection $S_n(\varphi)$ and $S_{n+1}(\varphi) \supset S_n(\varphi)$.

C. Improving Approximations

A description will now be given for the selection of an optimal weighted residual pair.

The usual procedure in the method of weighted residuals consists in selecting coordinate functions first and then weight functions. The choice of coordinate functions may be suggested by the problem but is usually a matter of taste. For example, if the solution is known to be periodic then of course periodic coordinate functions would be chosen. Frequently, however, the only constraint is that they should be linearly independent and satisfy the homogeneous boundary conditions. Once the

coordinate functions are selected the weight function choice is either based on the ease of resultant arithmetic (i.e., the computation of the inner products) or is completely arbitary. They should, of course, be linearly independent.

The following is suggested for these selections and is of course based on the preceding theorems. Select a sequence w_i of linearly independent functions that satisfy the homogeneous boundary conditions and such that there is enough "arbitrariness" (e.g., sufficiently many arbitrary constants) in each w_i so that Aw_i can be "forced" (e.g., the constants chosen) to satisfy the homogeneous boundary conditions.

It follows from Theorem 4.6 that the pair $\left(\{\varphi_i\}, \{w_i\}\right)$, where $\varphi_i = Aw_i$, form an optimal weighted residual pair. This method of choice is exhibited in the examples.

Another observation in connection with the above is quite helpful and can be used to obtain additional approximations.

Let

$$\overline{y}_n = \sum^{n} a_i \, (Aw_i)$$

denote the n^{th} approximate solution. The system of equations to solve using $\{Aw_i\}$ as coordinate functions and $\{w_i\}$ as weight functions is

$$(Ay_n, w_j) = (f, w_j) \quad j = 1, \ldots n$$

or

$$\sum^{n} a_i \, (A^2 w_i, w_j) = (f, w_j)$$

or

$$\sum^{n} a_i \, (Aw_i, Aw_j) = \sum a_i \, (\varphi_i, \varphi_j) = (f, w_j).$$

Interchanging the roles of weight and coordinate functions, now using

$$\bar{\bar{y}}_n = \sum^{n} b_i \, w_i$$

for the approximate solution, the system to solve is

$$(Ay_n, Aw_j) = (f, Aw_j) \qquad j = 1, \ldots n$$

or

$$\sum^{n} b_i \, (Aw_i, Aw_j) = \sum b_i \, (\varphi_i, \varphi_j) = (f, Aw_j).$$

Hence we conclude that the coefficient matrix to be inverted in each system is the same so that very little extra effort is needed to determine both $\bar{y}_n$ and $\bar{\bar{y}}_n$, i.e., merely the additional calculation of the inner products on the right.

Once these two approximations are known they can be used to give an additional approximation of y^* as a linear combination of $\bar{y}_n$ and $\bar{\bar{y}}_n$ which involves only a 2×2 matrix.

D. Computational Considerations

In [45] a coordinate set is defined to be numerically optimal provided it is complete and the eigenvalues of the coefficient matrix for the n^{th} approximation in the Galerkin method are bounded above and below independent of n. It is then shown that a numerically optimal set is invariant under similarity.

These ideas could easily be extended to the method of weighted residuals but have not been for two reasons. First, as shown in the examples, unless some prior information is available to give insight into the character of the problem, a random selection of coordinate functions may be just as good if not better than a set obtained using the idea of similarity.

Second, it is not reasonable to calculate eigenvalues of the coefficient matrix. The motivation for this idea is simply not to have the entries in the matrix get too large or too small and therefore create difficulties in machine calculations. This can usually be avoided by multiplication of each coordinate function by an appropriate constant. For example, if the coordinate functions were $\sin n\pi x$ and the inner products in the coefficient matrix were of order n, and hence increasing greatly as the size of the matrix increased, by using $\left\{\frac{\sin n\pi x}{\sqrt{n}}\right\}$ as coordinate functions, this difficulty would be overcome with no essential change in the problem. Therefore, the general rule is to properly "normalize" the coordinate and the weight functions to prevent the inner product entries in the coefficient matrix from becoming arbitrarily large or small.

It should be pointed out that the n^{th} approximation may not be as good as the $(n - 1)$st approximation. That is, the sequence consisting of the norms of the difference between the n^{th} approximation and the exact solution is not monotone decreasing. This can be observed in the examples. In practice, however, where the exact solution is not known there is in general no way to decide which particular approximation is better. All that is assured, if the method employed has been shown to

converge at all, in that for sufficiently large n the n^{th} approximate solution can be made as close as desired in norm to the exact solution.

The following computational scheme is suggested by the preceding section and by observation of the examples.

Suppose several approximations have been obtained using $\{\varphi_i\}$ as coordinate functions. A sequential limit for the first few coefficients may be guessed giving, say, an n^{th} approximation y_n with some certainty. Then other approximations can be obtained using $\{y_n, \varphi_{n+1}, \varphi_{n+2}, \cdots\}$ as coordinate functions. This will necessitate inverting a much smaller matrix and hence a considerable saving in machine time.

Chapter V

A RELATED METHOD

The purpose of this chapter is to present a method for the solution of the type of problems under consideration, to prove that it is an optimal method and to prove that the method converges. The method is no more difficult to apply than those previously considered and has several reasons to be prefered.

Suppose throughout this chapter L is a linear operator defined in a separable Hilbert space and y^* is the unique solution to

$$L(u) = f \tag{5.1}$$

subject to homogeneous boundary conditions. Let $\{\varphi_i\}$ and $\{\psi_i\}$ be sets of linearly independent functions which satisfy the boundary conditions. Furthermore, let

$$u_n = \sum_{i=1}^{n} a_i \varphi_i, \qquad v_n = \sum_{i=1}^{n} b_i \psi_i .$$

Definition 5.1. The function u_n is called an nth optimal approximation to the solution of 5.1 with respect to the set $\{\varphi_i\}$ provided

$$\| L(u_n) - f \| \leq \| L\left(\sum_{i=1}^{n} c_i \varphi_i \right) - f \|$$

for any choice of the coefficients c_i.

Definition 5.2. The set $\{\varphi_i\}$ is a better coordinate set for 5.1 than $\{\psi_i\}$ provided that if u_n and v_n are the respective n^{th} optimal approximations then

$$\| L(u_n) - f \| \leq \| L(v_n) - f \|$$

These definitions are exactly the criteria commonly used, that is, we ask what function most nearly satisfies the operator equation.

The first result is a necessary and sufficient condition for a function to be an optimal approximation.

Lemma 5.1. The function u_n is the n^{th} optimal approximation with respect to $\{\varphi_i\}$ if and only if $L(u_n)$ is the orthogonal projection of f on $S_n(L(\varphi))$.

Proof: We have

$$\| L(u_n) - f \| = \| \sum^{n} a_i L(\varphi_i) - f \| \leq \| \sum c_i L(\varphi_i) - f \|$$

for any choice of the c_i's. This is exactly the characterization of the orthogonal projection.

Lemma 5.2. If $\{\bar{\varphi}_i\}$ is an orthonormal basis for $S_n(L(\varphi))$ and u_n is the n^{th} optimal approximation with respect to $\{\varphi_i\}$ then

$$L(u_n) = \sum^{n} (f, \varphi_i)\, \bar{\varphi}_i. \tag{5.2}$$

Proof: By Lemma 5.1, $L(u_n)$ is the orthogonal projection of f on $S_n(L(\varphi))$ and truncation of the Fourier series gives the orthogonal projection.

Notice that equation 5.2 can be used to determine the approximation u_n. That is, having selected the $\{\varphi_i\}$ one can calculate $\{L(\varphi_i)\}$ and find (say by the Gram-Schmidt process) an orthonormal basis $\{\bar{\varphi}_i\}$ for $S_n(L(\varphi))$.

Then equation 5.2 can be solved directly for the coefficients a_i in u_n using the linear independence of $\{\bar{\varphi}_i\}$ or a set of linear independent functions $\{w_i\}$ can be used to give the system

$$\sum^{n} a_i(L(\varphi_i),\ w_j) = \sum^{n} (f,\ \bar{\varphi}_i)(\bar{\varphi}_i,\ w_j) \qquad j = 1,\ \ldots\ n$$

for the determination of the a_i's.

Theorem 5.1. The set $\{\varphi_i\}$ is a better coordinate set for 5.1 than $\{\psi_i\}$ provided the orthogonal projection of f on $S_n(L(\varphi))$ is a better approximation to f than the orthogonal projection of f on $S_n(L(\psi))$, that is

$$\Big\|\sum_{i=1}^{n} (f,\ \bar{\varphi}_i)\ \bar{\varphi}_i - f\Big\| < \Big\|\sum_{1}^{n} (f,\ \bar{\psi}_i)\ \bar{\psi}_i - f\Big\| \ .$$

Proof: By Lemma 5.2 the norm inequality becomes

$$\|L(u_n) - f\| < \|L(v_n) - f\|$$

which gives the result.

The question of convergence of this method is answered by the following result.

Corollary to Lemma 1: The norms $\|L(u_n) - f\|$ converge monotonically to zero provided f is in $\overline{S(L(\varphi))}$.

Proof: Since $L(u_n)$ is the orthogonal projection of f on the subspaces $S_n(L(\varphi))$ and these increase in dimension it follows that the norms converge to zero. The monotonicty comes from the fact that, since

$$S_n(L(\varphi)) \subset S_{n+1}(L(\varphi))$$

the orthogonal projection of f on $S_{n+1}(L(\varphi))$ could be no worse an approximation to f than the orthogonal projection of f on $S_n(L(\varphi))$.

To illustrate the application of the method presented in this chapter, we consider the equation

$$Ly \equiv y'' + y = -x$$

which $y(0) = y(1) = 0$. As coordinate functions which satisfy the boundary conditions, we select

$$\varphi_n(x) = x^n(1 - x) \quad n = 1, 2, \ldots.$$

To obtain an approximation in the form

$$y_2(x) = a_1\varphi_1(x) + a_2\varphi_2(x)$$

we determine an orthonormal basis $\{\bar{\varphi}_1, \bar{\varphi}_2\}$ for $S_2(L\varphi)$. An acceptable choice is

$$\bar{\varphi}_1(x) = \sqrt{\frac{30}{101}}\,(-2 + x - x^2)$$

$$\bar{\varphi}_2(x) = \sqrt{\frac{210}{2437}}\,(-6 + 13x - 3x^2 + 2x^3)$$

We now require that

$$Ly_2 = (-x, \bar{\varphi}_1)\,\bar{\varphi}_1 + (-x, \bar{\varphi}_2)\,\bar{\varphi}_2$$

which is an equation involving the two unknowns a_1 and a_2. By expressing the left-hand side as a linear combination of $\bar{\varphi}_1$ and $\bar{\varphi}_2$, it

is possible to solve directly for a_1 and a_2. However, if the inner product of each side with any two linearly independent functions is taken, two equations in a_1 and a_2 result, i.e.,

$$(Ly_2,\ w_m) = \left(\sum_{p=1}^{2} (-x,\ \bar{\varphi}_p)\ \bar{\varphi}_p,\ w_m \right) \quad m = 1,\ 2.$$

We remark that if, in particular, w_m are chosen to be $\bar{\varphi}_m$, the resulting system of equations are the same as would be obtained from the method of weighted residuals. For our example, to simplify computation we select

$$w_1 = 1,\ w_2 = x.$$

Then the system of equations, after simplification becomes

$$2a_1 + a_2 = \frac{55}{101}$$

$$a_2 = \frac{413}{2437}\ .$$

It may be the case that the n^{th} approximation obtained by this method is not the orthogonal projection of the solution on the span of the first n coordinate functions. However, we have the following result.

Theorem 5.2. If the operator L has a bounded linear inverse L^{-1} and if y_n is the n^{th} optimal approximation to the solution of

$$Ly = f$$

then

$$\lim_{n\to\infty} \| y_n - y \| = 0.$$

Proof: We have

$$\| y_n - y \| = \| L^{-1} (Ly_n - f) \| \quad \| L^{-1} \| \, \| Ly_n - f \|$$

and since $\| Ly_n - f \|$ can be made arbitrarily small the result follows.

Chapter VI

APPLICATIONS AND EXAMPLES

The purpose of this section is to apply several of the above methods to specific problems in order to illustrate the details of application, to demonstrate the approximation accuracy, and to point out certain difficulties which may arise.

We apply the Galerkin method and the method of weighted residuals to approximate the solutions to the following problems.

I. $-y'' - y = x$ with boundary conditions $y(0) = y(1) = 0$

II. $$y(x) - \int_0^1 K(x, \xi)\, y(\xi)\, d\xi = \frac{1}{6} x\,(1 - x^2)$$

where

$$K(x, \xi) = \begin{cases} x(1 - \xi) \text{ for } x \leq \xi \\ \xi(1 - x) \text{ for } \xi \leq x \end{cases}$$

III. $((1 + x)y')' = x$ with boundary conditions $y(0) = y(1) = 0$

In each of the problems considered, the operator is linear, self-adjoint, and positive definite. In addition, the operator in problem II is bounded, and hence continuous.

We consider two complete linearly independent sets of functions which satisfy the boundary conditions in each of the problems,

$$\varphi_n(x) = x^n(1 - x), \quad n = 1, 2, \ldots$$

$$\varphi_n(x) = \sin n\pi x, \quad n = 1, 2, \ldots$$

Thus, applying the general approximation theory of Chapter II, we take the space X to be the completion of the subspace of L_2 spanned by $\varphi_n(x)$, $n = 1, 2, \ldots$.

To determine whether or not the Galerkin method will give the best L_2 approximation using a fixed set of coordinate functions, we apply Theorem 4.5. That is, we determine whether or not the given operator applied to a linear combination of the first n coordinate functions gives an element which can also be written as such a linear combination. If the Galerkin method does not give the best approximation, applying Theorem 4.6, we determine a second linearly independ set of functions, $w_n(x)$, $n = 1, 2, \ldots$, to be used as weight functions in the method of weighted residuals which will give the best approximation.

In each of the following examples, we find eight approximations, the third through the tenth. The n^{th} approximation to the solution of $Ay = f$ is given by

$$y_n(x) = \sum_{j=1}^{n} a_{nj}\varphi_j(x)$$

We then evaluate the n^{th} approximation at each of the points $x_i = 0.1(i - 1)$, $i = 1, \ldots, 10$ and compare these with the exact solution at each point. Finally, since the exact solution y is known, we are able to evaluate the norm of the difference of the exact and approximate solutions.

The three norms considered are defined as follows.

The L_2-norm is

$$\|y - y_n\|_L = (y - y_n, y - y_n)^{\frac{1}{2}}$$

The A-norm is

$$\|y - y_n\| = (Ay - y_n, y - y_n)^{\frac{1}{2}}$$

$$= (f - y_n, y - y_n)^{\frac{1}{2}},$$

i.e., the energy norm.

The sup-norm is

$$\|y - y_n\| = \sup_{x\epsilon[0,1]} |y(x) - y_n(x)|$$

In example 1, we use the Galerkin method to solve Problem I. We restate the problem equivalently as

$$y'' + y = -x \text{ with boundary conditions } y(0) = y(1) = 0$$

As coordinate functions which satisfy the boundary conditions, we select

$$\varphi_n(x) = x^n(1 - x), \quad n = 1, 2, \ldots.$$

Thus

$$A\varphi_n = \varphi_n'' + \varphi_n = n(n - 1)x^{n-2} - n(n + 1)x^{n-1} + x^n - x^{n+1}$$

Applying the Galerkin method, we seek a j^{th} approximation in the form

$$y_j(x) = \sum_{n=1}^{j} a_n\varphi_n(x)$$

where the j unknowns a_n, $n = 1, \ldots, j$ are determined from the following system of j equations

$$\sum_{n=1}^{j} a_n(A\varphi_n, \varphi_m) = (-x, \varphi_m), \quad m = 1, \ldots, j$$

where

$$(A\varphi_n, \varphi_m) = \frac{m(m-1) + n(n-1)}{(m+n-1)(m+n)(m+n+1)} - \frac{m + n + 4}{(m+n+2)(m+n+3)}$$

$$(-x, \varphi_m) = - \frac{1}{(m+2)(m+3)}$$

The exact solution to Problem 1 is

$$y(x) = \frac{\sin x}{\sin 1} - x.$$

A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	A(9)	A(10)
1.8777-01	1.9414-01	-2.3411-02							
1.8838-01	1.8863-01	-1.0525-02	-8.5912-03						
1.8860-01	1.8837-01	-9.4872-03	-1.0124-02	7.7438-04					
1.8839-01	1.8844-01	-9.8572-03	-9.0942-03	-6.1035-05	4.2725-04				
1.8860-01	1.8826-01	-8.7891-03	-5.3711-03	-9.7656-04	-1.9531-03	2.1973-03			
1.8840-01	1.8845-01	-1.0742-02	-6.5918-03	-1.0254-02	1.3889-02	-9.2773-03	2.0752-03		
1.8839-01	1.8844-01	-1.0666-02	-7.0190-03	-7.0801-03	1.3889-02	-1.4862-02	9.8909-03	-2.3727-03	
1.8840-01	1.8818-01	-7.6141-03	-1.8250-02	1.7857-02	-1.8219-02	1.1111-02	-7.8125-03	7.5758-03	-3.0441-03

X	Y(X)	Y3(X)	Y4(X)	Y5(X)	Y6(X)	Y7(X)	Y8(X)	Y9(X)	Y10(X)
.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
.1	.018642	.018625	.018641	.018642	.018642	.018642	.018641	.018641	.018642
.2	.036098	.036105	.036098	.036098	.036098	.036104	.036055	.036094	.036098
.3	.051195	.051220	.051195	.051195	.051197	.051225	.051184	.051185	.051195
.4	.062783	.062803	.062783	.062784	.062787	.062860	.062755	.062766	.062784
.5	.069747	.069746	.069746	.069749	.069756	.069893	.069694	.069722	.069748
.6	.071018	.070998	.071018	.071021	.071033	.071245	.070933	.070987	.071022
.7	.065585	.065561	.065585	.065588	.065606	.065883	.065465	.065554	.065592
.8	.052502	.052496	.052503	.052505	.052527	.052827	.052362	.052477	.052514
.9	.030902	.030918	.030902	.030904	.030923	.031155	.030783	.030889	.030916
1.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
L2 NORM		1.658-05	3.696-07	1.742-06	1.343-05	1.840-04	7.714-05	1.941-05	6.391-06
L NORM		2.040-04	5.630-06	7.745-06	7.833-05	9.624-04	4.433-04	7.284-05	5.350-05
SLP NORM		2.565-05	5.895-07	2.875-06	2.491-05	3.250-04	1.409-04	3.197-05	1.399-05

Example 1.

In Example 2, we use the method of weighted residuals to solve Problem I.

As coordinate functions which satisfy the boundary conditions, again we select

$$\varphi_n(x) = x^n(1 - x), \quad n = 1, 2, \ldots.$$

As weight functions which will give the best approximations, we solve

$$Aw_n = \varphi_n, \quad n = 3, \ldots, 10$$

which gives

$$w_n = A_n \sin x - C_{n_1} \cos x + \sum_{k=1}^{n+2} C_{n_k} x^{k-1}$$

where

$$A_n = \frac{1}{\sin 1}\left(C_{n_1} \cos 1 - \sum_{k=1}^{n+2} C_{n_k}\right)$$

$$C_{n(n+2)} = -1, \; C_{n(N+1)} = 1 \text{ and}$$

$$C_{n_k} = -k(k+1)\, C_{n(k+2)}, \quad k = 1, \ldots, n.$$

Applying the method of weighted residuals, we seek a j^{th} approximation in the form

$$y_j(x) = \sum_{n=1}^{j} a_n \varphi_n(x)$$

where the j unknowns a_n, $n = 1, \ldots, j$ are determined from the following system of j equations

$$\sum_{n=1}^{j} a_n (A\varphi_n, w_m) = (-x, w_m) \qquad m = 1, \ldots, j$$

where

$$(A\varphi_n, w_m) = (\varphi_n, Aw_m) = (\varphi_n, \varphi_m) = \frac{1}{(m+n+1)(m+n+2)(m+n+3)}$$

$$(-x, w_m) = -\sum_{k=1}^{m+2} \frac{C_{m_k}}{k+1} + C_{m_1}\left(\frac{1}{\sin 1} - 1\right) - \sum_{k=1}^{m+2} C_{m_k}\left(\frac{\cos 1}{\sin 1} - 1\right)$$

A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	A(9)	A(10)
1.8766-01	1.9638-01	-2.3436-02							
1.8837-01	1.8864-01	-1.0513-02	-8.6060-03						
1.8839-01	1.8829-01	-9.1553-03	-1.0986-02	1.2207-03					
1.8839-01	1.8843-01	-9.9030-03	-8.6060-03	-1.2372-03	8.3923-04				
1.8834-01	1.8923-01	-1.3603-02	-1.8232-03	-5.2032-03	-8.7837-04	1.9608-03			
1.8832-01	1.8951-01	-1.4282-02	-1.8232-03	-4.3945-03	-8.7837-04	-4.8828-04	2.4414-03		
1.8843-01	1.8791-01	-7.8735-03	-1.0376-02	-1.2372-03	-8.7837-04	4.8828-03	-4.8933-04	-2.1667-03	
1.8903-01	1.8128-01	8.6212-03	-1.8232-03	-5.5695-02	-8.7837-04	6.2408-02	-4.8933-04	-3.7946-04	-2.6154-02

X	Y(X)	Y3(X)	Y4(X)	Y5(X)	Y6(X)	Y7(X)	Y8(X)	Y9(X)	Y10(X)
.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
.1	.018642	.018617	.018641	.018641	.018642	.018641	.018642	.018642	.018651
.2	.036098	.036095	.036098	.036096	.036098	.036099	.036101	.036098	.036085
.3	.051195	.051211	.051195	.051191	.051196	.051196	.051199	.051199	.051184
.4	.062783	.062798	.062783	.062777	.062786	.062785	.062786	.062796	.062789
.5	.069767	.069767	.069767	.069758	.069752	.069751	.069748	.069770	.069752
.6	.071018	.071003	.071020	.071005	.071024	.071025	.071017	.071052	.071004
.7	.065585	.065570	.065587	.065569	.065591	.065592	.065583	.065629	.065569
.8	.052502	.052506	.052504	.052487	.052506	.052509	.052505	.052548	.052516
.9	.030902	.030925	.030903	.030891	.030903	.030907	.030915	.030934	.030907
1.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
L2 NORM		1.491-05	1.009-06	9.614-06	3.396-06	4.242-06	5.062-06	2.624-05	1.263-05
A NORM		2.092-04	7.227-06	4.342-05	1.300-05	2.264-05	7.757-05	1.270-04	2.987-04
SUP NORM		2.522-05	2.077-06	1.635-05	5.841-06	7.080-06	1.408-05	4.650-05	2.742-05

Example 2.

In Example 3, we use the method of weighted residuals to solve Problem 1.

As coordinate functions which satisfy the boundary conditions, again we select

$$\varphi_n(x) = x^n(1-x), \quad n = 1, 2, \ldots.$$

As weight functions, we use

$$w_n(x) = \sin n\pi x, \quad n = 1, 2, \ldots$$

Applying the method of weighted residuals, we seek a j^{th} approximation in the form

$$y_j(x) = \sum_{n=1}^{j} a_n \varphi_n(x)$$

where the j unknowns a_n, $n = 1, \ldots, j$ are determined from the following system of j equations

$$\sum_{n=1}^{j} a_n (A\varphi_n, w_m) = (-x, w_m) \qquad m = 1, \ldots, j$$

where

$$(A\varphi_n, w_m) = \left(1 - (m\pi)^2\right) \int_0^1 x^n(1-x) \sin m\pi x dx$$

$$(-x, w_m) = \frac{1}{m\pi} (-1)^m.$$

A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	A(9)	A(10)
8.1269-03	3.4200-01	7.3638-01							
8.1278-03	-7.1406-01	7.3638-01	2.0717+00						
1.2995-01	-7.1406-01	-5.7481+00	2.0717+00	6.8706+00					
1.2995-01	8.9257-01	-5.7481+00	-1.8100+01	6.8705+00	2.5441+01				
9.6243-02	8.9256-01	9.6303+00	-1.8101+01	-8.8022+01	2.5441+01	1.0187+02			
9.6211-02	-6.2109-01	9.6307+00	7.6913+01	-8.8021+01	-4.3676+02	1.0187+02	4.3231+02		
1.0038-01	-6.2045-01	-1.5172+01	7.6863+01	5.4016+02	-4.3644+02	-2.2111+03	4.3198+02	1.9175+03	
1.0854-01	3.5598-01	-1.5204+01	-2.0230+02	5.4098+02	3.5750+03	-2.2140+03	-1.1627+04	1.9199+03	8.8319+03

X	Y(X)	Y3(X)	Y4(X)	Y5(X)	Y6(X)	Y7(X)	Y8(X)	Y9(X)	Y10(X)
.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
.1	.018642	.006472	-.004846	.002144	.014809	.022975	.017407	.000988	-.011132
.2	.036098	.016957	-.014185	-.021635	.005256	.062238	.112636	.092456	-.050531
.3	.051195	.037170	-.017615	-.065102	-.065203	.034677	.262038	.536926	.554517
.4	.062783	.063059	-.006500	-.107254	-.200344	-.177579	.178615	1.106598	2.666458
.5	.069747	.090805	.023540	-.118933	-.349733	-.606775	-.593375	.510885	4.401882
.6	.071018	.114821	.070168	-.076374	-.413963	-1.077009	-2.090766	-2.861868	-.643158
.7	.065585	.127756	.121761	.032287	-.286620	-1.184796	-3.400172	-8.206575	-16.846351
.8	.052502	.120481	.155023	.165567	.052536	-.529025	-2.666043	-9.574152	-30.274595
.9	.030902	.082115	.132502	.222241	.380905	.622287	.779675	-.274248	-7.876163
1.0	.000000	.000000	.000000	.000000	.000000	.000000	.000001	.000005	.000015
L2 NORM		3.777-02	6.243-02	1.246-01	2.753-01	6.509-01	1.631+00	4.306+00	1.195+01
A NORM		2.192-01	4.653-01	1.014+00	2.428+00	6.394+00	1.816+01	5.455+01	1.718+02
SUP NORM		6.828-02	1.109-01	1.914-01	4.859-01	1.287+00	3.545+00	1.017+01	3.033+01

Example 3.

In Example 4, we use the Galerkin method to solve Problem 1.

As coordinate functions which satisfy the boundary conditions, we select

$$\varphi_n(x) = \sin n\pi x, \quad n = 1, 2, \ldots.$$

Thus

$$A\varphi_n = \varphi_n'' + \varphi_n = \left(1 - (n\pi)^2\right) \sin n\pi x.$$

Applying the Galerkin method, we seek a j^{th} approximation in the form

$$y_j(x) = \sum_{n=1}^{j} a_n \varphi_n(x)$$

where the j unknowns a_n, $n = 1, \ldots, j$ are determined from the following system of j equations

$$\sum_{n=1}^{j} a_n \, (A\varphi_n, \varphi_m) = (-x, \varphi_m), \quad m = 1, \ldots, j$$

where

$$(A\varphi_n, \varphi_m) = \begin{cases} \frac{1}{2}\left(1 - (m\pi)^2\right) & \text{if } m = n \\ 0 & \text{if } m \neq n \end{cases}$$

$$(-x, \varphi_m) = \frac{1}{m\pi} (-1)^m.$$

Thus

$$a_n = \frac{2(-1)^n}{n\pi[1 - (n\pi)^2]}$$

and

$$y_j(x) = \frac{2}{\pi} \sum_{n=1}^{j} \frac{(-1)^n}{\pi\left[1 - (n\pi)^2\right]} \sin n\pi x.$$

A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	A(9)	A(10)
7.1775-02	-8.2724-03	2.4162-03							
7.1775-02	-8.2724-03	2.4162-03	-1.0143-03						
7.1775-02	-8.2724-03	2.4162-03	-1.0143-03	5.1812-04					
7.1775-02	-8.2724-03	2.4162-03	-1.0143-03	5.1812-04	-2.9947-04				
7.1775-02	-8.2724-03	2.4162-03	-1.0143-03	5.1812-04	-2.9947-04	1.8845-04			
7.1775-02	-8.2724-03	2.4162-03	-1.0143-03	5.1812-04	-2.9947-04	1.8845-04	-1.2618-04		
7.1775-02	-8.2724-03	2.4162-03	-1.0143-03	5.1812-04	-2.9947-04	1.8845-04	-1.2618-04	8.8592-05	
7.1775-02	-8.2724-03	2.4162-03	-1.0143-03	5.1812-04	-2.9947-04	1.8845-04	-1.2618-04	8.8592-05	-6.4568-05

X	Y(X)	Y3(X)	Y4(X)	Y5(X)	Y6(X)	Y7(X)	Y8(X)	Y9(X)	Y10(X)
.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
.1	.018642	.019272	.018308	.018026	.018541	.018693	.018619	.018647	.018647
.2	.036098	.036619	.036023	.036023	.036199	.036020	.036140	.036087	.036087
.3	.051195	.050947	.051543	.051025	.051201	.051259	.051139	.051211	.051211
.4	.062783	.061980	.062945	.062945	.062660	.062770	.062845	.062760	.062760
.5	.069767	.069359	.069359	.069877	.069877	.069689	.069689	.069778	.069778
.6	.071018	.071705	.070740	.070740	.071025	.071136	.071061	.070977	.070977
.7	.065585	.066682	.066086	.065567	.065391	.065450	.065570	.065641	.065641
.8	.052502	.052354	.052950	.052950	.052774	.052595	.052475	.052423	.052423
.9	.030902	.028997	.029962	.030480	.030765	.030917	.030991	.031019	.031019
1.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
L2 NORM		8.535-04	4.626-04	2.825-04	1.869-04	1.311-04	9.609-05	7.287-05	5.679-05
A NORM		1.242-02	8.576-03	6.368-03	4.967-03	4.013-03	3.329-03	2.819-03	2.427-03
SUP NORM		2.000-03	1.214-03	8.151-04	5.825-04	4.375-04	3.403-04	2.739-04	2.216-04

Example 4.

In Example 5, we use the method of weighted residuals to solve Problem I.

As coordinate functions which satisfy the boundary conditions, we select

$$\varphi_n(x) = \sin n\pi x, \quad n = 1, 2, \ldots,$$

and as weight functions, we use

$$w_n(x) = x^n(1-x), \quad n = 1, 2, \ldots.$$

Applying the method of weighted residuals, we seek a j^{th} approximation in the form

$$y_j(x) = \sum_{n=1}^{j} a_n(A\varphi_n, w_m) = (-x, w_m) \quad m = 1, \ldots, j$$

where

$$(A\varphi_n, w_m) = \left(1 - (n\pi)^2\right) \int_0^1 x^m(1-x) \sin n\pi x dx$$

$$(-x, w_m) = \frac{1}{m\pi} (-1)^m.$$

A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	A(9)	A(10)
1.0101-01	-5.9832-02	6.1117-03							
1.0020-01	-5.3567-02	-7.8481-03	9.3065-03						
9.8943-02	-3.9279-02	-5.9995-02	8.5027-02	-3.7860-02					
9.7867-02	-2.2819-02	-1.4630-01	2.8717-01	-2.5476-01	8.6761-02				
9.7067-02	-7.1217-03	-2.5607-01	6.5232-01	-8.7988-01	6.1781-01	-1.7702-01			
9.6499-02	6.6439-03	-3.7836-01	1.1902+00	-2.1760+00	2.3614+00	-1.4018+00	3.4994-01		
9.6131-02	1.8272-02	-5.0521-01	1.8942+00	-4.4060+00	6.9553+00	-6.0354+00	3.1233+00	-6.9333-01	
9.5819-02	2.7967-02	-6.3184-01	2.7525+00	-7.8174+00	1.4940+01	-1.8934+01	1.5206+01	-6.9916+00	1.3997+00

X	Y(X)	Y3(X)	Y4(X)	Y5(X)	Y6(X)	Y7(X)	Y8(X)	Y9(X)	Y10(X)
.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
.1	.018642	.000989	.001980	.001955	-.000668	.003521	.000160	.000554	.003848
.2	.036098	.000280	.005959	.013719	.014476	-.004618	.015174	.030698	-.026912
.3	.051195	.026702	.022225	.012033	.047238	.071244	-.043680	-.077060	.281328
.4	.062783	.057304	.059574	.025412	-.024938	.101780	.402301	.108238	-1.347255
.5	.069747	.094897	.108050	.121078	-.010593	-.369731	-.300082	1.537386	4.852569
.6	.071018	.127641	.140248	.233317	.383080	.175780	-1.421901	-5.107751	-6.687048
.7	.065585	.140510	.135055	.186701	.530220	1.577915	3.453134	5.880348	-7.896488
.8	.052502	.122088	.096908	-.011525	-.177709	-.031632	2.045007	11.416292	43.171435
.9	.030902	.071326	.047250	-.113600	-.685092	-2.404034	-7.174450	-19.764418	-51.558666
1.0	.000000	.000000	-.000000	-.000000	-.000000	-.000000	-.000001	-.000002	-.000007
L2 NORM		6.200-02	3.919-02	8.506-02	2.990-01	9.173-01	2.643+00	7.514+00	2.160+01
L NORM		2.360-01	2.369-01	9.771-01	4.104+00	1.438+01	4.692+01	1.495+02	4.770+02
SUP NORM		7.606-02	7.350-02	1.717-01	7.214-01	2.435+00	7.414+00	2.198+01	6.537+01

Example 5.

To conclude the analysis of Problem I, we compare the results of Examples 1 through 5.

A comparison of the first three examples using as coordinate functions

$$\varphi_n(x) = x^n(1-x), \quad n = 1, 2, \ldots$$

shows that the Galerkin method gives good results. In the second example where an optimal weighted residual pair is used, it can be seen that in most cases, the approximations are not much, if any, better than those obtained by the Galerkin method. In those cases where the Galerkin method apparently gives a better approximation, the discrepancy is attributed to the inaccuracy of numerical evaluation of the inner product expressions in the optimal weighted residual pair case and to the error produced in the numerical evaluation of the integrals defining the respective norms. This is substantiated by Theorem 4.10 since the sequence of norms $\{\|y_n - y_o\|\}$ is non-increasing which is not reflected in this example. A comparison of examples 2 and 3 shows more clearly that the optimal weighted residual pair gives the best approximation using this set of coordinate functions.

In general, the problem of solving $Aw_n = \varphi_n$ to obtain an optimal weighted residual pair is as difficult as solving the original problem $Ay = f$. In Problem I, in particular, after obtaining the optimal weight functions, it was found that these functions were very close to being orthogonal to the right-hand side of the original equation. By

examination of the inner-product expression (f, w_n), it can be seen that as n increases, we have (f, w_n) represented as the sum of three increasingly large factors to produce a term which tends to zero. Thus to obtain the accuracy required, it was necessary to evaluate (f, w_n) using double precision arithmetic.

In summary, we conclude that although the Galerkin method did not give the orthogonal projection of the solution on the span of the first n coordinate functions $\varphi_n(x) = x^n(1-x)$, the increased difficulty of finding such a projection did not warrant the increased effort.

In Examples 4 and 5 coordinate functions

$$\varphi_n(x) = \sin n\pi x, \quad n = 1, 2, \ldots$$

were used. In this case

$$A\varphi_n = [1 - (n\pi^2)] \sin n\pi x$$

which shows that the Galerkin method gives the orthogonal projection since, as in Theorem 4.5, the span of the first n coordinate functions is invariant under the operator A. Thus, as we would expect, the n^{th} approximation using the Galerkin method in Example 4 is much better than the corresponding approximation obtained by the method of weighted residuals in Example 5.

As a final comparison we note that in Examples 2 and 4 orthogonal projections on the spans of the respective first n coordinate functions are obtained in each case. However, each approximation in Example 2 is

better than the corresponding approximation in Example 4. In fact, each approximation in Example 1, which is not the orthogonal projection, is also better than the corresponding approximation in Example 4. This indicates that the coordinate functions $\varphi_n(x) = x^n(1-x)$, $n = 1, 2, \ldots$ are a better set for Problem I than $\varphi_n(x) = \sin n\pi x$, $n = 1, 2, \ldots$. Although no criteria for the selection of a best set of coordinate functions is given, we remark that any additional information about the problem such as physical characteristics, periodicity of the solution, etc., are helpful in determining an optimal set of coordinate functions. In the case of Problem I, we speculate, for example, that coordinate functions $\varphi_n(x) = x^n(1-x)$ give a better approximation than $\varphi_n(x) = \sin n\pi x$ since the right-hand side of the equation, $-x$, is certainly not periodic.

In Example 6, we use the Galerkin method to solve Problem II.

As coordinate functions we select

$$\varphi_n(x) = x^n(1-x), \quad n = 1, 2, \ldots.$$

Thus

$$A\varphi_n = \varphi_n(x) - \int_0^1 K(x, \xi)\, \varphi_n(\xi)\, d\xi$$

$$= -\frac{2x}{(n+1)(n+2)(n+3)} + x^n(1-x) + \frac{x^{n+2}}{(n+1)(n+2)} - \frac{x^{n+3}}{(n+2)(n+3)}.$$

Applying the Galerkin method, we seek a j^{th} approximation in the form

$$y_j(x) = \sum_{n=1}^{j} a_n \varphi_n(x)$$

where the j unknowns a_n, $n = 1, \ldots, j$ are determined from the following system of j equations

$$\sum_{n=1}^{j} a_n(A\varphi_n, \varphi_m) = \frac{1}{6}(x - x^3, \varphi_m), \quad m = 1, \ldots, j$$

where

$$(A\varphi_n, \varphi_m) = \frac{2}{(m+n+1)(m+n+2)(m+n+3)} + \frac{2}{(n+1)(n+2)(n+3)} \times$$

$$\left[\frac{2n + m + 6}{(m+n+3)(m+n+4)(m+n+5)} - \frac{1}{(m+2)(m+3)}\right]$$

$$\frac{1}{6}(x - x^3, \varphi_m) = \frac{2m + 7}{3(m+2)(m+3)(m+4)(m+5)}$$

By means of Green's function, the boundary value problem, Problem I, can be transformed to the integral equation of problem II. Thus the solution to Problem II is also

$$y(x) = \frac{\sin x}{\sin 1} - x.$$

We observe here that, to within the accuracy of numerical methods, the n^{th} approximation obtained by solving Problem II is the same as the corresponding approximation obtained by solving Problem I (Example 1). This fact is better illustrated by considering the coordinate functions

$$\varphi_n(x) = \sin n\pi x, \quad n = 1, 2, \ldots.$$

Applying the Galerkin method, we seek a j^{th} approximation in the form

$$y_j(x) = \sum_{n=1}^{j} a_n\varphi_n(x)$$

where the j unknowns a_n, $n = 1, \ldots, j$ are determined from the following system of j equations

$$\sum_{n=1}^{j} a_n (A\varphi_n, \varphi_m) = \frac{1}{6}(x - x^3, \varphi_m) \quad m = 1, \ldots, j$$

where

$$(A\varphi_n, \varphi_m) = \begin{cases} \frac{1}{2}\, 1 - (m\pi)^2 & \text{if } m = n \\ 0 & \text{if } m \neq n \end{cases}$$

$$\frac{1}{6}(x - x^3, \varphi_m) = \frac{1}{m\pi}(-1)^m$$

Thus

$$a_n = \frac{2(-1)^n}{n\pi[1 - (n\pi)^2]}$$

which is the same as the result obtained in Example 4.

We conclude that all observations made concerning the operator and approximate method of determining a solution to Problem I are equally valid in determining a solution to Problem II.

A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	A(9)	A(10)
1.8766-01	1.9438-01	-2.3437-02							
1.8837-01	1.8864-01	-1.0529-02	-8.6060-03						
1.8839-01	1.8852-01	-9.3994-03	-1.0010-02	6.1835-04					
1.8840-01	1.8838-01	-1.0040-02	-8.6365-03	-1.1243-03	9.9182-04				
1.8834-01	1.8922-01	-1.3702-02	-1.6534-03	-5.4092-03	-7.9966-04	1.9604-03			
1.8837-01	1.8887-01	-1.2817-02	-1.6534-03	-8.7891-03	-7.9966-04	6.0359-03	-4.1504-03		
1.8840-01	1.8856-01	-1.0498-02	-9.1553-03	-1.1243-03	-7.9966-04	3.1128-03	-4.4678-04	-8.8501-04	
1.8837-01	1.8891-01	-1.2848-02	-1.6534-03	-7.5989-03	-7.9966-04	4.6387-03	-4.4678-04	-3.4688-04	-7.4768-04

X	Y(X)	Y3(X)	Y4(X)	Y5(X)	Y6(X)	Y7(X)	Y8(X)	Y9(X)	Y10(X)
.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
.1	.018642	.018617	.018641	.018641	.018641	.018641	.018641	.018643	.018642
.2	.036098	.036095	.036098	.036096	.036096	.036098	.036097	.036099	.036098
.3	.051195	.051211	.051195	.051193	.051191	.051194	.051191	.051191	.051194
.4	.062783	.062798	.062783	.062782	.062777	.062781	.062772	.062768	.062781
.5	.069747	.069747	.069747	.069747	.069760	.069745	.069726	.069715	.069743
.6	.071018	.071003	.071018	.071020	.071010	.071018	.070974	.070966	.071011
.7	.065585	.065570	.065586	.065588	.065578	.065584	.065510	.065516	.065572
.8	.052502	.052505	.052503	.052505	.052499	.052501	.052397	.052431	.052483
.9	.030902	.030925	.030902	.030903	.030902	.030902	.030796	.030850	.030882
1.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
L2 NORM		1.491-05	3.510-07	1.807-06	4.703-06	1.021-06	5.583-05	4.081-05	1.021-05
A NORM		2.091-04	5.998-06	1.105-05	2.175-05	1.192-05	4.015-04	2.071-04	7.746-05
SUP NORM		2.520-05	6.463-07	3.241-06	7.946-06	1.875-06	1.130-04	7.263-05	2.082-05

Example 6.

In Example 7, we use the Galerkin method to solve Problem III.

As coordinate functions which satisfy the boundary conditions, we select

$$\varphi_n(x) = x^n(1 - x), \quad n = 1, 2, \ldots.$$

Applying the Galerkin method, we seek a j^{th} approximation in the form

$$y_j(x) = \sum_{n=1}^{j} a_n \varphi_n(x)$$

where the j unknowns a_n, $n=1, \ldots, j$ are determined from the following system of j equations

$$\sum_{n=1}^{j} a_n (A\varphi_n, \varphi_m) = (x, \varphi_m), \quad m = 1, \ldots, j$$

where

$$(A\varphi_n, \varphi_m) = \int_0^1 \left((1 + x)\varphi_n'' + \varphi_n'\right)\varphi_m dx = -\int_0^1 (1 + x)\varphi_n'\varphi_m' dx$$

$$= -\frac{mn}{m+n-1} + \frac{(mn + m + n)}{m+n} - \frac{(1 - mn)}{m+n+1} - \frac{(m + 1)(n + 1)}{m+n+2}$$

$$(x, \varphi_m) = \frac{1}{(m+2)(m+3)}$$

The exact solution to Problem III is

$$y(x) = \frac{1}{4}\left[x(x - 2) + \frac{\ln (1 + x)}{\ln 2}\right].$$

A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	A(9)	A(10)
-1.4005-01	-5.8989-02	1.9663-02							
-1.3954-01	-6.6559-02	3.6458-02	-1.0804-02						
-1.3936-01	-6.8882-02	4.5353-02	-2.3697-02	6.2066-03					
-1.3931-01	-6.9923-02	5.1208-02	-3.7720-02	2.1240-02	-5.8899-03				
-1.3969-01	-6.0608-02	-2.2461-02	2.1875-01	-3.9258-01	3.3390-01	-1.1816-01			
-1.3926-01	-7.1037-02	5.9143-02	-5.9814-02	4.1016-02	1.3672-02	-4.3823-02	1.7776-02		
-1.3939-01	-6.7902-02	3.2837-02	3.4180-02	-1.2012-01	1.0498-01	5.3691-02	-1.0107-01	4.5288-02	
-1.3928-01	-7.0625-02	5.7007-02	-5.8960-02	4.5654-02	-1.2207-03	-2.2217-02	-2.2827-02	4.9713-02	-2.0752-02

X	Y(X)	Y3(X)	Y4(X)	Y5(X)	Y6(X)	Y7(X)	Y8(X)	Y9(X)	Y10(X)
.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
.1	-.013124	-.013154	-.013126	-.013124	-.013124	-.013122	-.013124	-.013124	-.013125
.2	-.024241	-.024234	-.024237	-.024241	-.024242	-.024239	-.024241	-.024246	-.024244
.3	-.032872	-.032839	-.032869	-.032872	-.032873	-.032853	-.032870	-.032884	-.032882
.4	-.038643	-.038616	-.038645	-.038644	-.038644	-.038554	-.038639	-.038665	-.038671
.5	-.041259	-.041257	-.041264	-.041259	-.041260	-.041053	-.041255	-.041292	-.041317
.6	-.040482	-.040503	-.040484	-.040480	-.040483	-.040155	-.040482	-.040518	-.040583
.7	-.036116	-.036142	-.036114	-.036115	-.036118	-.035725	-.036134	-.036130	-.036271
.8	-.028001	-.028009	-.027998	-.028000	-.028003	-.027662	-.028056	-.027961	-.028199
.9	-.016000	-.015985	-.016001	-.015999	-.016002	-.015851	-.016082	-.015907	-.016180
1.0	-.000000	-.000000	-.000000	-.000000	-.000000	-.000000	-.000000	-.000000	-.000000
L2 NORM		2.138-05	2.949-06	8.225-07	1.131-06	2.115-04	3.228-05	3.799-05	1.052-04
A NORM		3.151-04	5.434-05	1.049-05	1.233-05	1.377-03	4.953-04	6.659-04	9.407-04
SUP NORM		3.452-05	5.015-06	1.600-06	1.878-06	3.921-04	8.218-05	9.473-05	2.023-04

Example 7.

In Example 8, we use the Galerkin method to solve Problem III.

As coordinate functions which satisfy the boundary conditions, we select

$$\varphi_n(x) = \sin n\pi x, \quad n = 1, 2, \ldots.$$

Applying the Galerkin method, we seek a j^{th} approximation in the form

$$y_j(x) = \sum_{n=1}^{j} a_n \varphi_n(x)$$

where the j unknowns a_n, $n = 1, \ldots, j$ are determined from the following system of j equations

$$\sum_{n=1}^{j} a_n (A\varphi_n, \varphi_m) = (x, \varphi_m) \quad m = 1, \ldots, j$$

where

$$(A\varphi_n, \varphi_m) = \begin{cases} -mn\left[(-1)^{m+n} - 1\right]\left[\dfrac{m^2 + n^2}{(m^2-n^2)^2}\right] & \text{if } m \neq n \\ -\dfrac{3}{4}(m\pi)^2 & \text{if } m = n \end{cases}$$

$$(x, \varphi_m) = -\frac{1}{m\pi}(-1)^m$$

A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	A(9)	A(10)
-6.2434-02	1.8922-03	-1.4154-03							
-6.2406-02	1.9069-03	-1.3561-03	3.1529-04						
-6.2413-02	1.8905-03	-1.3674-03	2.6247-04	-3.0106-04					
-6.2407-02	1.8925-03	-1.3633-03	2.6571-04	-2.8554-04	9.7623-05				
-6.2410-02	1.8891-03	-1.3649-03	2.6186-04	-2.8890-04	7.9811-05	-1.0984-04			
-6.2408-02	1.8896-03	-1.3640-03	2.6238-04	-2.8757-04	8.1030-05	-1.0317-04	4.2106-05		
-6.2409-02	1.8885-03	-1.3645-03	2.6147-04	-2.8813-04	7.9525-05	-1.0460-04	3.4042-05	-5.1712-05	
-6.2408-02	1.8887-03	-1.3641-03	2.6163-04	-2.8780-04	7.9738-05	-1.0400-04	3.4631-05	-4.8348-05	2.1853-05

X	Y(X)	Y3(X)	Y4(X)	Y5(X)	Y6(X)	Y7(X)	Y8(X)	Y9(X)	Y10(X)
.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
.1	-.013124	-.013146	-.012780	-.013154	-.013035	-.013152	-.013117	-.013143	-.013140
.2	-.024241	-.024489	-.024215	-.024278	-.024324	-.024218	-.024262	-.024224	-.024227
.3	-.032872	-.032968	-.033096	-.032790	-.032858	-.032881	-.032839	-.032889	-.032885
.4	-.038643	-.038413	-.038710	-.038672	-.038578	-.038659	-.038678	-.038626	-.038629
.5	-.041259	-.041019	-.041048	-.041348	-.041330	-.041224	-.041228	-.041279	-.041276
.6	-.040482	-.040637	-.040352	-.040395	-.040483	-.040533	-.040505	-.040460	-.040462
.7	-.036116	-.036567	-.036353	-.036077	-.036030	-.036073	-.036110	-.036145	-.036142
.8	-.028001	-.028088	-.028213	-.028182	-.028122	-.028025	-.027990	-.027966	-.027967
.9	-.016000	-.015370	-.015621	-.015875	-.015951	-.016023	-.016041	-.016052	-.016050
1.0	-.000000	-.000000	-.000000	-.000000	-.000000	-.000000	-.000000	-.000000	-.000000
L2 NORM		2.972-04	2.330-04	1.062-04	9.065-05	5.119-05	4.553-05	2.908-05	2.650-05
A NORM		6.539-03	5.401-03	3.465-03	3.057-03	2.227-03	2.027-03	1.585-03	1.470-03
SUP NORM		6.949-04	4.957-04	2.989-04	2.364-04	1.638-04	1.379-04	1.047-04	8.957-05

Example 8.

In Example 9 we use the method of weighted residuals to solve Problem III.

As coordinate functions which satisfy the boundary conditions, we select

$$\varphi_n(x) = x^n(1 - x), \quad n = 1, 2, \ldots.$$

It can be shown that the operator in Problem III, $A_1 y = \left((1 + x)y'\right)'$ is similar (Definition 4.5) to the operator $A_2 y = y''$. For weight functions, we solve

$$A_2 w_n = \varphi_n \qquad n = 1, 2, \ldots$$

to obtain

$$w_n(x) = \frac{x}{(n+1)(n+2)(n+3)} \left\{ x^{n+1} \left[(n + 3) - (n + 1)x\right] - 2 \right\}.$$

Thus, applying Theorem 4.9, since $\left\{\varphi_n, w_n\right\}_{n=1}^{\infty}$ is an optimal weighted residual pair with respect to the operator A_2, $\left\{\varphi_n, w_n\right\}_{n=1}^{\infty}$ is an almost optimal weighted residual pair with respect to the operator A_1.

Applying the method of weighted residuals, we seek a j^{th} approximation in the form

$$y_j(x) = \sum_{n=1}^{j} a_n \varphi_n(x)$$

where the j unknowns, a_n, $n = 1, \ldots, j$ are determined from the following system of j equations

$$\sum_{n=1}^{j} a_n(A\varphi_n, w_m) = (-x, w_m) \qquad m = 1, \ldots, j$$

where

$$(A\varphi_n, w_m) = \frac{1}{(m+1)(m+2)(m+3)} \left\{ (m+3) \left[\frac{n(m+2)}{m+n+1} - \frac{(mn+n+m+2)}{m+n+2} - \frac{(mn+2n+1)}{m+n+3} + \frac{(m+1)(n+1)}{m+n+4} \right] + \frac{2}{(n+1)(n+2)} \right\}$$

$$(x, w_m) = -\frac{(2m^2+9m+7)}{3(m+1)(m+2)(m+3)(m+4)(m+5)} .$$

A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	A(9)	A(10)
-2.9752-01	3.6327-01	-1.0749-01							
-4.2776-01	1.3719+00	-2.4021+00	1.4484+00						
-4.7957-01	1.9754+00	-4.6138+00	4.6326+00	-1.5682+00					
-4.6389-01	1.7662+00	-3.6857+00	2.8334+00	1.1243-03	-5.0082-01				
-4.2732-01	1.3180+00	-1.9110+00	1.6536-03	1.4600+00	7.9957-04	-4.8421-01			
-5.2112-01	2.3933+00	-5.0235+00	1.6535-03	9.5816+00	7.9964-04	-1.7440+01	1.1147+01		
-5.4671-01	2.8543+00	-7.8836+00	7.9890+00	1.1243-03	7.9964-04	-5.6416+00	4.4677-04	3.3626+00	
-5.4044-01	2.7537+00	-7.4064+00	7.2525+00	1.1243-03	7.9964-04	-4.0054+00	4.4678-04	3.4688-04	2.0938+00

X	Y(X)	Y3(X)	Y4(X)	Y5(X)	Y6(X)	Y7(X)	Y8(X)	Y9(X)	Y10(X)
.0	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
.1	-.013124	-.023676	-.028183	-.029133	-.028917	-.028303	-.029797	-.029892	-.029870
.2	-.024241	-.037179	-.038059	-.037520	-.037690	-.038054	-.036645	-.036422	-.036510
.3	-.032872	-.043137	-.040585	-.039864	-.039992	-.040400	-.039452	-.039508	-.039471
.4	-.038643	-.043731	-.040949	-.041112	-.041007	-.040887	-.042086	-.042229	-.042155
.5	-.041259	-.040690	-.040316	-.041072	-.040898	-.040536	-.041700	-.041694	-.041757
.6	-.040482	-.035294	-.037559	-.037905	-.037865	-.037779	-.036818	-.036756	-.036850
.7	-.036116	-.028372	-.031002	-.030481	-.030564	-.030827	-.029568	-.029594	-.029541
.8	-.028001	-.020304	-.020155	-.019614	-.019627	-.019796	-.021108	-.021186	-.021185
.9	-.016000	-.011020	-.007455	-.008162	-.008019	-.007805	-.008000	-.008003	-.008112
1.0	-.000000	-.000000	-.000000	-.000000	-.000000	-.000000	.000000	.000000	.000000
L2 NORM		7.636-03	8.144-03	8.191-03	8.185-03	8.158-03	8.272-03	8.284-03	8.278-03
A NORM		5.808-02	7.800-02	7.914-02	7.849-02	7.667-02	9.170-02	9.320-02	9.363-02
SUP NORM		1.297-02	1.575-02	1.629-02	1.617-02	1.586-02	1.673-02	1.677-02	1.675-02

Example 9.

A comparison of the results in Examples 7 and 9 shows that in all of the norms considered, the Galerkin method gives better approximations than the method of weighted residuals using an almost optimal weighted residual pair. Thus we are lead to conclude that in selection of the best approximation method, unless a priori information is available, the Galerkin method will give relatively good results.

To illustrate the use of the Galerkin method for partial differential equations, we consider the problem

IV. $\nabla^2 u + x \frac{\partial u}{\partial x} = 2x^2 + 2y^2 + 2x^2y^2 - 2x^2y - xy^2 + xy - 2x - 2y$

with boundary conditions

$u(x, 0) = u(x, 1) = u(0, y) = u(1, y) = 0.$

As coordinate functions which satisfy the boundary conditions, we select

$$\varphi_n(x) = \sin p\pi x \sin q\pi y$$

where the correspondence of p and q to n is given by

n	1	2	3	4	5	6	7	8	9	10
p	1	2	1	3	2	1	4	3	2	1
q	1	1	2	1	2	3	1	2	3	4

$$A\varphi_n = \nabla^2\varphi_n + x \frac{\partial u}{\partial x} = \left[-\pi^2(p^2 3 q^2) \sin p\pi x + p+x \cos p\pi x \right] \sin q\pi y.$$

Applying the Galerkin method, we seek a j^{th} approximation in the form

$$u_j(x) = \sum_{n=1}^{j} a_n \varphi_n(x)$$

where the j unknowns a_n, $n = 1, \ldots, j$ are determined from the following system of j equations

$$\sum_{n=1}^{j} a_n (A\varphi_n, \varphi_m) = (f, \varphi_m) \qquad m = 1, \ldots, j$$

where f is the right-hand side of equation IV and if

$\varphi_m(x) = \sin r\pi x \sin s\pi y$,

$$(A\varphi_n, \varphi_m) = \begin{cases} \dfrac{pr}{2(p^2-r^2)} (-1)^{p+r} & \text{if } q = s,\ p \neq r \\ -\dfrac{\pi^2}{4} (p^2 + q^2) - \dfrac{1}{8} & \text{if } q = s,\ p = r \\ 0 & \text{if } q \neq s \end{cases}$$

$$(f, \varphi_m) = \frac{1}{(s\pi)^3} \left[2A(r) - (s\pi)^2 B(r) \right] \left[(-1)^s - 1 \right]$$

where

$$A(r) = -\frac{1}{r\pi} (-1)^r + \left[\frac{4}{(r\pi)^3} - \frac{2}{(r\pi)} \right] \left[(-1)^r - 1 \right]$$

$$B(r) = \frac{4}{(r\pi)^3} \left[(-1)^r - 1 \right].$$

The solution to Problem IV is

$$u(x, y) = xy(1 - x)(1 - y).$$

We remark that this problem was considered by Mihklin [38]. He applied the Galerkin method as we have done, however, his results are incorrect due to sign errors in his calculations of $(A\varphi_n, \varphi_m)$ and (f, φ_m).

A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	A(9)	A(10)
6.6238-02	-3.4195-03	0.0000							
6.6118-02	-3.2660-03	0.0000	3.5338-03						
6.6118-02	-3.2660-03	0.0000	3.5338-03	0.0000					
6.6118-02	-3.2660-03	0.0000	3.5338-03	0.0000	2.4648-03				
6.6105-02	-3.2318-03	0.0000	3.5144-03	0.0000	2.4648-03	-5.5073-04			
6.6105-02	-3.2318-03	0.0000	3.5144-03	0.0000	2.4648-03	-5.5073-04	0.0000		
6.6105-02	-3.2318-03	0.0000	3.5144-03	0.0000	2.4641-03	-5.5073-04	0.0000	-4.9123-05	
6.6105-02	-3.2318-03	0.0000	3.5144-03	0.0000	2.4641-03	-5.5073-04	0.0000	-4.9123-05	0.0000

Example 10.

As a final example we consider the application of the Galerkin method to the initial value problem

V. $y'' + xy = 1$ with initial conditions $y(0) = y'(0) = 0$.

In attempting to apply the Galerkin method, we are immediately faced with two problems.

The first problem arises in the selection of coordinate functions which satisfy the initial condition. It is not difficult to satisfy the initial conditions, for

$$\varphi_n(x) = x^{n+1}, \quad n = 1, 2, \ldots$$

is one such set. However, the set

$$\varphi_n(x) = x^{n+1}(R - x), \quad n = 1, 2, \ldots$$

for any real number R, also satisfies the initial conditions. Thus, by an appropriate choice of R, we are able to force our approximations to vanish at any desired point.

Assuming an appropriate set of coordinate functions has been selected, the second problem arises in the calculation of inner products. Since

$$(f, g) = \int_a^b fg\,dx$$

we must have some finite interval [a, b] over which the integration is to be performed. We solve this problem by assuming that an approximate

solution is desired only for certain values of the independent variable. Thus we may assign to b, any value such that [a, b] contains all value in which we are interested. In [31], Ames suggests that b be chosen as some point at which steady-state is achieved (see discussion in Chapter III). We point out, however, that this is not always possible. In Problem V, in particular, noting that the solution is

$$y(x) = x^2 \sum_{n=0}^{\infty} (-1)^n \frac{n!}{(3n+2)!} (3x^3)^n$$

steady-state is never achieved; in fact, the solution becomes increasingly unstable as x increases.

We conclude that for Problem V the Galerkin method and the method of weighted residuals are not applicable unless a second zero of the function y is known, in which case, we approximate the solution as shown in the first three boundary value problems.

Chapter VII

THE NONLINEAR CASE

The methods under consideration can, in principle, be applied to nonlinear problems as noted in Chapter III. However, there are at least two difficulties involved. First, and less serious although by no means minor, is the fact that the system of equations obtained for the coefficients is nonlinear. Hence it may be an impossible task to solve them. Second, there are no known convergence proofs for this class of problems.

The above drawbacks have not prevented the use of a method. The first has been overcome by the use of very few terms and the second has been ignored. Authors invariably claim "good" results.

The following is a summary of the use of Galerkin's method in a nonlinear situation.

Cesari [33], in a very elegant paper, uses a Galerkin approach to give an existence proof for a solution to certain nonlinear equations. As an example he analyzes the problem

$$x'' + x + \alpha x^3 = \beta t, \quad 0 \leq t \leq 1$$

with boundary conditions

$$x(0) = 0, \quad x'(1) + hx(1) = 0$$

where α, β and h are numbers.

Following in this direction Urabe [42] considers the nonlinear periodic equation

$$\frac{dx}{dt} = X(x, t)$$

where x and $X(x, t)$ are vectors of the same dimension and $X(x, t)$ is periodic of period 2π in t. In particular he proves the following theorem.

Theorem: The existence of an isolated periodic solution, $\hat{x}$, of (7.1) implies the existence of Galerkin approximations x_m for sufficiently large m. The existence of Galerkin approximations x_m for sufficiently large m implies the existence of an exact solution provided a certain boundedness condition holds.

In a later publication Urabe [44] compares his earlier technique with a method of averaging when $X(x, t) = \lambda \bar{Y}(x, t)$ where λ is a small parameter. A somewhat more detailed numerical analysis is given in [43].

Locker [50] generalizes the work of Cesari to give an existence analysis of

$$Lx = Nx$$

where L is an unbounded linear operator and N is a nonlinear operator. When L is self-adjoint his results reduce to Cesari's. He also gives a simple example.

There is one other recent reference which mentions Galerkin's method in connection with nonlinear problems. Bellman [32] considers a method for replacing a nonlinear problem by a linear one. As an alternative to the main point of his paper, he briefly indicates that a one term Galerkin approximation could be used to implement his method.

In summary, there have been two uses of Galerkin's method on nonlinear problems. First, as indicated above, a few existence results have been obtained for certain problems by employing a Galerkin approach. Second, and at the other extreme, the method has been applied without regard to convergence.

To illustrate some of the difficulties which arise is attempting to apply the Galerkin method or method of weighted residuals to nonlinear equations, we consider the following differential equation.

$$(y')^2 + \pi^2 y^2 = \pi^2 \text{ with boundary conditions } y(0) = y(1) = 0$$

We note first that $y(x) = \sin \pi x$ is a solution to the equation, but that it is not unique, for $y(x) = -\sin \pi x$ is also a solution.

In applying the method of weighted residuals, as coordinate functions which satisfy the boundary conditions, we select

$$\varphi_n(x) = x^n(1 - x), \quad n = 1, 2, \ldots$$

and as weight functions, we use

$$w_n(x) = x^{n-1}, \quad n = 1, 2, \ldots.$$

We seek a j^{th} approximation in the form

$$y_j(x) = \sum_{n=1}^{j} a_n \varphi_n(x)$$

where the j unknowns a_n, $n = 1, \ldots, j$ are determined from the following system of j equations

$$(A \sum_{n=1}^{j} a_n \varphi_n),\ w_m) = (\pi^2,\ w_m), \quad m = 1,\ \ldots,\ j. \tag{7.2}$$

Thus we obtain a system of non-linear equations which may be difficult or impossible to solve.

To illustrate, consider the case where $j = 2$. The system of equations (7.2) becomes

$$\left(\frac{\pi^2 + 10}{30}\right) a_1^2 - \left(\frac{\pi^2 - 10}{30}\right) a_1 a_2 + \left(\frac{71\pi^2 + 14}{105}\right) a_2^2 = \pi^2$$

$$\left(\frac{\pi^2 + 10}{60}\right) a_1^2 - \left(\frac{29\pi^2 - 49}{210}\right) a_1 a_2 + \left(\frac{485\pi^2 + 84}{840}\right) a_2^2 = \frac{\pi^2}{2}$$

Solving this system, assuming $a_2 \neq 0$, we obtain

$$a_1^2 = \frac{\pi^2}{\left[y(\pi^2 + 10) - 7(\pi^2 - 10)\ k + 2\ (71\pi^2 + 14)\ k^2\right]}$$

$$a_2 = ka_1, \text{ where } k = \frac{2(51\pi^2 - 28)}{201\pi^2 + 28}.$$

Thus, in seeking a two-term approximation, we find two possible solutions. Continuing in this manner, it can be seen that as the number of terms in the approximation increases, the number of possible solutions to the system of equations (7.2) also increases. It is apparent that without a priori information regarding the solution of a non-linear problem, a purely arbitrary choice among the possible solutions may not give a desirable approximation.

Chapter VIII

APPLICATIONS AND ANALYSES OF METHODS OF THE GALERKIN TYPE AVAILABLE IN THE LITERATURE

This chapter consists of a rather complete representative cross section of references to the methods under consideration. Following each reference is a remark to indicate how the article uses a method. Some insight into typical applications, frequence of use and care in application can be gained from these references.

The list does not include many of the very brief references to one or more of the various methods. A reference is included only in case there is a fairly complete discussion of a method or a nontrivial application. Inclusion of a reference is not an endorsement of the application. Convergence proofs or justification for use almost never occur.

The references are grouped by (A) Ordinary Differential Equations, (B) Partial Differential Equations and (C) Studies of Methods per se.

A. Ordinary Differential Equations

1. Birikh, "On the spectrum of small perturbations of plane-parallel Couette flow", (PMM) Journal of Applied Mathematics and Mechanics, Vol. 29, 1965, 946-949.

Solves a fourth order ordinary differential equation subject to homogeneous boundary conditions using the coordinate functions

$$V_n = K_n \left[\frac{\cosh ax}{\cosh a} - \frac{\cos \sqrt{\lambda_n - a^2}\, x}{\cos \sqrt{\lambda_n - a^2}} \right]$$

for even n and the same funtions with sinh and sin replacing cosh and cos, respectively, for odd n. The same problem is solved in [9] using different coordinate functions.

2. Bruslinskaja, "Limit cycles for equations of motion of a rigid body and Galerkin equations for hydrodynamics", Soviet Mathematics, Vol. 5, Aug. 1964, 1051-1054.

Solves a first order system of three equations with little comment.

3. Chandraskhar, "Hydrodynamic and hydromagnetic stability", Clarendon Press, Oxford, England.

Solves the system

$$(D^2 - a^2)^2 u = (1 + ax) v$$

$$(D^2 - a^2) v = - \lambda u$$

subject to $u = Du = v = 0$ for $x = 0, 1$ using as coordinate functions $v_n = \sin n\pi x$ and u_n the solutions of

$$(D^2 - a^2) u_n = (1 + ax) \sin n\pi x$$

$$u_n = Du_n = 0 \text{ for } x = 0, 1$$

and using $\sin n\pi x$ as weight functions.

4. DiPrima, "Applications of Galerkins method to a problem in hydrodynamic stability," Quarterly of Applied Mathematics, Vol. 13, 1, 1955, 55-62.

Solves a problem similar to that in [3] using e^{nx} times a polynomial as coordinate functions.

5. DiPrima, Walowit, Tsao, "Stability of flow between arbitrarily spaced concentric cylindrical surfaces including the effect of a radial temperature gradient," Journal of Applied Mechanics, Vol. 31, 1964, 585-593.

Solves a 2 × 2 homogeneous system of ordinary differential equations with homogeneous boundary conditions using $(x^2 - \frac{1}{4})^2 x^{n-1}$ as coordinate functions for both unknowns. Shows no work.

6. DiPrima, "Some variational principles for problems in hydrodynamic and hydromagnetic stability", Quarterly of Applied Mathematics, Vol. 18, 1961, 375-385.

Mentions the application of the Galerkin method to a problem similar to [3]. He uses cosines as one set of coordinate functions and those of [10] for the other.

7. Duncan, "Torsional oscillation of a cantilever when the stiffness is of composite origin," TR Aero Research Commission, Great Britian Air Ministry, Part 1, 1937, 1809, 471-483.

Solves a second order ordinary differential equation with homogeneous boundary conditions using polynomials as coordinate functions.

8. Frazer, Jones and Shan, "Approximation to functions and the solution of differential equations", TR Aereo Research Commission, Great Britian Air Ministry, Vol. 1, 1937, 1799, 517-549.

Solves ordinary differential equations of order less than four using certain polynomials as coordinate functions.

9. Gallagher and Mercer, "On the behavior of small disturbances in plane couette flow", Journal of Fluid Mechanics, Vol. 13, 1962, 91-100.

Solves a fourth order equation with homogeneous boundary conditions using coordinate functions similar to those in [10].

10. Harris and Reid, "On orthogonal functions which satisfy four boundary conditions – Tables", Astrophysical Journal Supplement, Vol. 3, 1958, 429-452.

An orthogonal sequence of functions is presented which satisfy

$$y^{(3)} = ay,\ y = y^{1} = 0 \text{ at } x = \pm 0.5$$

$$c_m(x) = \frac{\cosh \lambda_m x}{\cosh \frac{1}{2}\lambda_m} - \frac{\cos \lambda_m x}{\cos \frac{1}{2}\lambda_m}$$

$$S_m(x) = \frac{\sinh \mu_m x}{\sinh \frac{1}{2}\mu_m} - \frac{\sin \mu_m x}{\sin \frac{1}{2}\mu_m}$$

where λ_m and μ_m are the positive roots of

$$\tanh \tfrac{1}{2}\lambda + \tan \tfrac{1}{2}\lambda = 0$$

$$\coth \tfrac{1}{2}\mu - \cot \tfrac{1}{2}\mu = 0.$$

11. Krueger and DiPrima, "The stability of a viscous fluid between rotating cylinders with axial flow", Journal of Fluid Mechanics, Vol. 19, 1964, 528-538.

Solves a complex eigenvalue problem for a 2 × 2 system using Galerkins method with the coordinate functions of [10].

12. Kurzweg, "Magnetohydrodynamic stability of curved viscous flows", TRII-29, Princeton University, Department of Physics, 1961.

Applies the Galerkin method to three simultaneous fourth order linear ordinary differential equations with homogeneous boundary conditions using polynomial coordinate functions.

13. Penzes and Burgin, "Free vibrations of thin isotropic oblate-spheroidal shells", General Dynamics/Convair, San Diego, Calif.

Solves a second order homogeneous ordinary linear differential equation with variable coefficients using associated Legerdre functions with the Galerkin method where the constraints are that the solution be single valued and bounded.

14. Rao, "The fundamental flexural vibrations of a cantilever beam of rectangular cross section with uniform taper", Aeronautical Quarterly, May 1965, Vol. 16, 139-144.

Applies Galerkin method to a fourth order ordinary equation with homogeneous boundary conditions.

15. Sastry, "Influence of wall conductance on the stability of Hartmann flow in a curved channel", Physical Society of Japan Journal, Vol. 21, 6, June 1966.

Applies Galerkin method to an eigenvalue problem involving a fourth order differential equation.

16. Sennet and Wait, "Analog computer analysis of static beams", AICE, Engineering Mechanics Division Journal, Vol. 91, April (I) 1965, 47-62.

Compares a Galerkin solution to an analog computer solution of a fourth order equation and finds the Galerkin solution within 3% of the other.

17. Smith, "On the growth of Taylor-Garther vortices along highly concave walls", Quarterly of Applied Mathematics, Vol. 13, 1955, 233-262.

Applies Galerkin method to two simultaneous ordinary differential equations, one of third order and one of fourth, with homogeneous boundary conditions using e^{-an} times polynomials as coordinate functions.

B. Partial Differential Equations

18. Bickley, "Experiments in approximating solutions of a partial differential equation", Philosophical Magazine, (7) 32, 1941, 50-56.

Solves the one dimensional wave equation with certain boundary and initial conditions using his modification of the Galerkin method, collocation and least squares obtaining one and two term approximations.

19. Biezeno, Grammil, Blackie, "Theory of Elasticity", Engineering Dynamics, Vol. 1, 170-176.

Applies Galerkin method to $\nabla^4 u = f$ with homogeneous side conditions.

20. Bolotin, "Nonconservative Problems of the Theory of Elastic Stability", 1963, pp. 58-62, 108-109, 207-212, 247-265.

Makes several applications of the Galerkin method.

21. Cheng and Pan, "Stability analysis of gas lubricated, self-acting, plain, cylindrical journal bearings of finite length using Galerkins method", Journal of Basic Engineering, Vol. 87, 1965, 185-191.

Applies Galerkin method to a second order equation with homogeneous boundary conditions using cosines as coordinate functions.

22. Crandall, Engineering Analysis, McGraw-Hill, 1956, pp. 149, 233, 372.

Gives example to illustrate the method described herein on page 25.

23. Duncan, "Applications of the Galerkin method to the torsion and flexure of cylinders and prisims", Philosophical Magazine, Series 7, Vol. 25, 1938, pp. 634-649.

Applies Galerkin method to several examples of Poissons equation with polynomials as coordinate functions.

24. Finlayson and Scriven, "The method of weighted residuals and its relation to certain variational principles for the analysis of transport processes", Chemical Engineering Sciences, Vol. 20, 1965, pp. 305-404.

Discusses the method of weighted residuals and works a few examples.

25. Kaplan, "Some new methods of flux synthesis", Nuclear Science and Engineering, Vol. 13, 1962, pp. 22-31.

Applies Galerkin and related methods to diffusion equations subject to continuous flux and current and homogeneous boundary conditions.

26. Karnagnti, "The critical Reynold's number for the flow past a sphere", Journal of the Physical Society of Japan, Vol. 10, 1955, pp. 694-699.

Uses method of weighted residuals to solve nonlinear fourth order partial differential equation with nonhomogeneous boundary conditions using sines as coordinate functions and Legendre functions as weight functions.

27. Nemat-Nasser, Prasad, Henmann, "Destabalizing effect of velocity dependent forces in nonconservative continuous systems", AIAA Journal, Vol. 4, 7, July 1966, pp. 1276-1280.

Compares an exact solution of a fourth order partial differential equation to a two term Galerkin approximation.

28. Snyder and Stewart, "Velocity and pressure profiles for newtonian creeping flow in regular packed beds of spheres", AICLE Journal, Vol. 12, 1966, pp. 167-173.

Applies Galerkin method to a homogeneous system of linear partial differential equations with nonhomogeneous boundary conditions using sines and cosines as coordinate functions.

29. Snyder, Spriggs and Stewart, "Solutions of the equations of change by Galerkins method", AICLE Journal, 10, 1964, pp. 535-539.

Has discussion of methods in general with no proof of his recommendations. Uses Galerkin method on nonlinear partial differential equation example with known solution to conclude results are good.

30. Toos, "The Galerkin's vector for the dynamic problems of an elastic isotropic and nonhomogeneous body", Revue Romaine de Mathematique Pures et Appliques, Vol. 10, 6, pp. 855-861.

Applies the Galerkin method to a fourth order linear nonhomogeneous partial differential equation.

C. Method Studies

31. Ames, "Nonlinear partial differential equations in engineering", Mathematics in Science and Engineering, Vol. 18, Academic Press, 1965, pp. 243-270.

Discusses several variations of the method of weighted residuals. Covered herein on page 24.

32. Bellman and Richardson, "Linearization based on differential approximation and Galerkins method", Quarterly of Applied Mathematics, Vol. 24, April 1966.

Uses a one term approximation to help in gaining information to approximate a nonlinear equation in a very brief example.

33. Cesari, "Functional analysis and Galerkins method", Michigan Mathematical Journal, Vol. 11, 1964, pp. 385-414.

See page 90 of this report.

34. Crandall, Engineering Analysis, McGraw-Hill, 1956.

Discusses Galerkin method and applies to very simple examples.

35. Duncan, "The principles of the Galerkin method", TR Aeronautical Research Commission, Great Britian Air Ministry, Vol. 2, 1848, 1938, pp. 589-612.

Applies the Galerkin method to several problems quite similar to [7].

36. Finlayson and Scriven, "The method of weighted residuals - a review", Applied Mechanics Reviews, Vol. 19, 9, Sept. 1966, pp. 735-747.

Outlines the relation between Galerkins method and variational methods.

37. Kantorivich and Krylov, Approximate Methods of Higher Analysis, Wiley and Sons, 1964, pp. 258-304.

Gives good discussion of several approximate methods and solves several examples.

38. Mikhlin, Variational Methods in Mathematical Physics, Pergamon Press, 1964.

Discusses Galerkin methods and proves several convergence theorems stated herein.

39. Milne, Numerical Solution of Differential Equations, Wiley and Sons, 1953, pp. 114-116.

Describes the Galerkin method for nonhomogeneous boundary conditions.

40. Reiss, "Variational and related methods", bound as supplement to Problems of Mathematical Physics by Lebeder, Prentice Hall, 1965, pp. 404-407.

Explains several variations of the Galerkin method with suggestions but offers no proof. See Chapter III.

41. Shulesko, "A new method of solving boundary value problems of mathematical physics", Australian Journal of Applied Science, Vol. 10, 1959, pp. 1-7.

For a discussion of his suggestion see page 23.

42. Urabe, "Galerkins procedure for nonlinear periodic systems", Archive for Rational Mechanics and Analysis, Vol. 20, 1965, pp. 120-152.

See page 90 of this report.

43. Urabe and Reiter, "Numerical computation of nonlinear forced oscillations by Galerkins procedure", Journal of Mathematical Analysis and Applications, Vol. 14, 1966, pp. 107-140.

See page 91 of this report.

44. Urabe, "Periodic solutions of differential systems, Galerkins procedure and the method of averaging", Journal of Differential Equations, Vol. 2, 1966, pp. 265-280.

See page 91 of this report.

45. Babuska, Numerical Processes in Differential Equations, (New York: Interscience Publishers, 1966).

46. Galerkin, "Rods and plates. Series occuring in various questions concerning the elastic equilibrium of rods and plates," Engineers Bulletin (Vestnik inzhenerov) 19, 1915, 897-908.

47. Green, "An Expansion Method for Parabolic Partial Differential Equations," Journal of Research of the National Bureau of Standards, LI (September 1953), 127-132.

48. Hildebrand, Methods of Applied Mathematics. (Englewood Cliffs: Prentice Hall, 1952), p. 286.

49. Kantorovich, Functional Analysis in Normed Spaces, (New York: The MacMillan Co., 1964).

50. Locker, "An existence analysis for nonlinear equations in Hilbert space", Transactions of the American Mathematical Society, Vol. 128, 3, Sept. 1967, pp. 403-413.

51. Milhlin, "Some sufficient conditions for the convergence of Galerkin's method," Uch. zap. Len. gos. un-ta, 135, ser. matem. nauk, 21, 1950. (restated in [38])

52. Petrov, "Application of Galerkin's method to the problem of stability of the flow of a viscous liquid", PMM (Applied Mathematics and Mechanics), Vol. 4, 1940, 3-12.

53. Petryshyn, "On the eigenvalue problem $Tu - Su = 0$ with unbounded and nonsymmetric operators T and S," TID-4500, NYU, 1963.

54. Petryshyn, "Direct and iterative methods for the solution of linear operator equations in Hilbert space," Trans. Amer. Math. Soc., Vol. 105, 1962, 136-175.

55. Repman, "A problem in the mathematical bases of Galerkin's method for solving problems on the stability of elastic systems," PMM (Applied Mathematics and Mechanics) 4, 1940.