A NUMERICAL INVESTIGATION OF THE RAYLEIGH-RITZ METHOD FOR THE SOLUTION OF VARIATIONAL PROBLEMS

J.L. LLOYD

Ph.D. Thesis July 1972

University of Newcastle-upon-Tyne

ACKNOWLEDGEMENTS

I wish to thank my supervisor, Dr. K. Wright, for his advice and encouragement throughout the preparation of this thesis.

Thanks are also due to Mrs. Pat Kirton, who typed the manuscript, and Miss Isobel Dixon, for duplicating copies of the thesis.

This research was supported by the Science Research Council and by the University of Newcastle-upon-Tyne.

Abstract

The results of a numerical investigation of the Rayleigh-Ritz method for the approximate solution of two-point boundary value problems in ordinary differential equations are presented. Theoretical results are developed which indicate that the observed behaviour is typical of the method in more general applications.

In particular, a number of choices of co-ordinate functions for certain second order equations are considered. A new algorithm for the efficient evaluation of an established sequence of functions related to the Legendre polynomials is described, and the sequence is compared in use with a similar sequence related to the Chebyshev polynomials. Algebraic properties of the Rayleigh-Ritz equations for these and other co-ordinate systems are discussed. The Chebyshev system is shown to lead to equations with convenient computational and theoretical properties, and the latter are used to characterize the asymptotic convergence of the approximations for linear equations. These results are subsequently extended to a certain type of non-linear equation.

An orthonormalization approach to the solution of the Rayleigh-Ritz equations which has been suggested in the literature is compared in practice with more usual methods, and it is shown that the properties of the resulting approximations are not improved. Since it is known that the method requires more work than established ones it cannot be recommended.

Quadrature approximations of elements of the Rayleigh-Ritz matrices are investigated, and known results for a restricted class of quadrature

approximation are extended towards the more general case.

In a final chapter extensions of the material of earlier chapters to partial differential equations are described, and new forms of the 'finite element' and 'extended Kantorovich' methods are proposed. A summary of the conclusions discerned from the investigation is given.

Contents

Index of Figures and Tables

Chapter One

Variational Calculus; Problems and Methods

1.1 A Historical Introduction.

In 1696, at a time when the study of the maxima and minima of functions of a finite number of variables had already played, as Courant(1) says, a decisive role in the development of differential and integral calculus, John Bernoulli suggested the following problem:

"Among all paths joining the points A and B, find that path along which a mass particle, subject only to the influence of gravity, will travel from A to B in the shortest possible time".

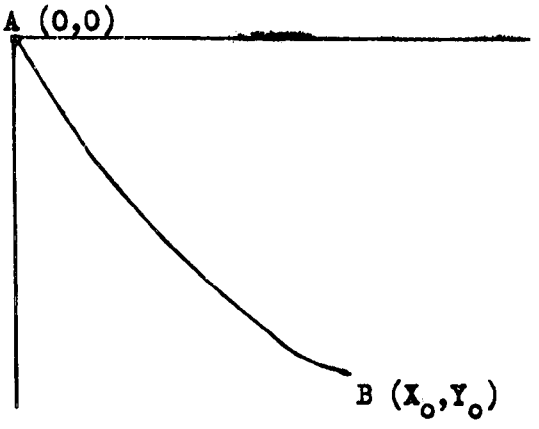


Fig.1

This problem, known as the brachistochrone, was novel in that it involved an infinite number of variables, the position of all points on the curve AB. The mathematical expression for T, the time taken, is straightforward and is given by

$$T = \int_0^{X_0} \sqrt{\frac{1 + y'^2}{2gy}}\, dx$$

The time T is therefore a function of the curve y, which must satisfy

$$y(0) = 0 \quad , \quad y(X_0) = Y_0$$

Other problems came to be expressed in this form. One important problem which can be so expressed is the determination of the path of light through a medium. Fermat's principle states that light travels between two points so that the time taken is a minimum with respect to times taken on other possible paths. This is clearly closely related to the brachistochrone, and can be expressed mathematically in the following manner.

Given two points $A(x_1,y_1)$ and $B(x_2,y_2)$, in a medium for which the velocity of light v at any point (x,y) is given by $v = v(x,y)$, determine a curve $y = y(x)$ joining A and B such that

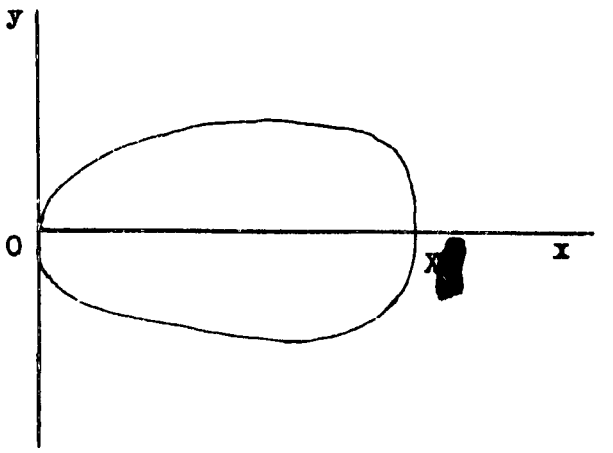
$$T = \int_{x_1}^{x_2} \frac{\sqrt{1 + y'^2}}{v(x,y)} \, dx$$

is a minimum.

Another class of variational problems which arose were the isoperimetric problems, of which the classical example is

'Determine the equation of the closed curve of fixed length which encloses the maximum area'.

It is straightforward to see that the curve must be convex, and we may assume that it is symmetric about some axis, which we shall assume to be the x axis



The problem can be expressed

$$\text{Minimize} \quad A = 2 \int_0^{X_0} y(x)\,dx$$

subject to

$$2 \int_0^{X_0} \sqrt{1 + y'^2}\,dx = L \text{ , a constant}$$

and

$$y(0) = y(X_0) = 0$$

Notice that in this problem the upper limit X_0 of each integration is not fixed, but varies with the curve y, subject to the restriction on the length of y.

Problems such as these were formulated and solved by such mathematicians as the Bernoulli's and Euler, using methods which were specific to the particular problem, and often very ingenious. A general approach was developed later by Lagrange, who reduced the general problem to the solution of a differential equation with auxiliary conditions, usually two point boundary conditions. The equivalence of the variational and differential formulations of particular problems was used by many mathematicians, including Gauss, to prove the existence of solutions of certain types of differential equation.

Euler expressed the laws of mechanics in variational terms, and formulations such as this, including those invoked by Hamilton in the theory of optics, became popular in the physical sciences. A simple example of such a formulation is the principle that the equilibrium positions of a mechanical system are stationary points of the expression for the potential energy of the system, and that a position of stable equilibrium is one which not only is a stationary point of the potential energy, but also makes that energy a minimum. As a

particular example, consider the case of a string of length l fixed at $x = 0$ and $x = l$, under the action of an external force $f(x)$ acting in a direction perpendicular to the x axis. Then if we let $y(x)$ denote the configuration of the string, and assume that $y(x)$ is "small" for all x, we obtain an expression for the total reduced potential energy of the system as

$$I(y) = \int_0^1 \left(y'(x)^2 + y(x).f(x) \right) dx \qquad (1.1)$$

The position of equilibrium of the string is then obtained by minimizing $I(y)$ over the set of functions $Y = \{ y(x) : y(0) = y(1) = 0,\ y \text{ is continuous in } (0,1) \}$.

Integrands of this type, in which we have a combination of homogeneous quadratic and linear expressions, are very important in the calculus of variations, because they occur so frequently in physical situations. More examples can be found in Courant (1,p.131-2).

The use of variational formulations of problems to obtain numerical approximations to their solution stems largely from the work of Rayleigh (1) and Ritz (1,2). Rayleigh's most important contribution was the use of particular trial functions for $y(x)$ in integrands related to (1.1) above to obtain estimates of the frequency of vibrations of mechanical systems, whilst Ritz provided the systematic approach to the substitution of trial functions which now forms the basis of the Rayleigh-Ritz method. Using this systematic approach Ritz was able to give the first satisfactory explanation of the nodal lines on a vibrating clamped plate. For an account of the contributions of Rayleigh and Ritz, and a review of Ritz's work by Rayleigh, see Gould(1).

An important development shortly following the work of Ritz was the Galerkin method (e.g. Kantorovich and Krylov (1)), derived in 1915.

Although this is not a variational method, and is wider in application than the Rayleigh-Ritz method, in certain circumstances the two methods are identical and it is important for this reason. Kantorovich proposed a variant of the Rayleigh-Ritz method, (Kantorovich and Krylov (1); for recent extensions see Kerr (1)) as applied to variational problems in which the function y is a function of more than one variable. The finite element method (e.g. Zienkiewicz and Cheung (1)), is closely related to the Rayleigh-Ritz method, and has a distinguished practical history. Practical use of the Rayleigh-Ritz method is particularly important in the works of engineers such as Timoshenko(1) and Von Karman (Von Karman and Biot(1)). During the mid 1930's an extensive study of existence theory for variational problems was undertaken by Bliss (1) and co-workers at Chicago.

The development of high-speed computation facilities has emphasised two of the problems which Courant(2) saw in 1949; the first, the selection of suitable trial functions for $y(x)$ has been examined by Mikhlin (Mikhlin (1),(4), Mikhlin and Smolitskiy(1)), and will concern us at some length, and the second, the problem of error estimates and bounds has been tackled recently by Ciarlet, Schultz and Varga(1). These workers, and others have also investigated the use of piecewise continuous functions as trial functions, and provided a closer connection between the Rayleigh-Ritz and the finite element method. This work will also be thoroughly discussed in a later chapter.

1.2 The simplest problem of variational calculus

The problems of variational calculus which will concern us can be expressed in general terms as follows:

Let Ω be a region of n dimensional space with boundary $\delta\Omega$.

Let $Y = \{y(\underline{x})\}$ be a set of functions of points $\underline{x}$, $\underline{x} \in \Omega$ such that

$$I(y) = \int_{\Omega} F(x,y,y' \cdots y^{(k)})\, d\Omega \qquad \ldots(1.2)$$

exists, and assume that there exists at least one $y_o(\underline{x}) \in Y$ such that

$$I(y_o) \leq L < \infty$$

and that there exists an $\Pi > -\infty$ such that

$$I(y) > \Pi \qquad \text{for all } y \in Y.$$

Determine a function $\bar{y}(\underline{x}) \in Y$ such that $I(\bar{y}) \leq I(y)$ for all $y \in Y$.

Additional conditions, in the form of boundary conditions, are usually imposed on the set Y. These take the form

$$g_i\,(x,y,y' \cdots y^{(k-1)}) = 0 \quad \text{on} \quad \delta\Omega_i$$

The set Y must be such that the integral (1.2) exists, and we give the following specification of the set Y, which is the most restrictive we shall require

$$Y = \{y(\underline{x}) : y \in C^{k-1}(\Omega),\ y \text{ satisfies } g_i(x,y,y' \cdots y^{k-1}) = 0 \text{ on } \delta\Omega_i\}$$

$y \in C^{k-1}(\Omega)$ implies that y has $k-1$ continuous derivatives in Ω and that $y^{(k)} = \frac{d^k y}{d\underline{x}^k}$ is square-summable in Ω.

This problem, the so-called 'simplest problem' of variational calculus, includes most of the problems we have so far encountered. An exception is the iso-perimetric problem, in which the condition imposed on the boundary is not a simple closed expression in $y(\underline{x})$. We consider only the simplest problem and its multi-dimensional extensions in the rest of this work.

As has been indicated, even amongst the simplest problems of

variational calculus there are a number of important divisions. We have already noted the distinction between integrals of combinations of homogeneous quadratic and linear forms, such as those occurring in the problem of the vibrating string, and the more general integral of problems such as the brachistochrone. The former are the most important, and most tractable problem for the methods of variational calculus. We shall defer discussion of all but the 'simple quadratic' integrals, and some closely related types, until Chapter Six.

Another distinction can be made with regard to the dimension of the space Ω, that is, the dimension of the vector $\underline{x}$. One dimensional problems, in which $y(x)$ is a function of a scalar argument, are an important subset of the set of simple variational problems, and we shall be primarily concerned with practical problems of this type. Nonetheless, many results are applicable to the one dimensional and multi-dimensional cases, and parts of the theory will be developed without reference to the dimension of the problem.

A third distinction, the distinction of the order of $F(\underline{x},y,y'\cdots y^k)$, which is k, the order of the highest derivative of $y(\underline{x})$ occurring in F, will be of no importance in the subsequent discussion.

1.3 Related Problems

Problems which can be expressed in the form of the simplest problem of the calculus of variations can often be expressed in many other forms. Perhaps the relationship with certain types of differential equations is most well known, and for our purposes it is so important that we devote paragraph 1.4 and subsequent paragraphs to its discussion. Here we wish to indicate briefly connections with two other important areas of mathematical study, the fields of dynamic

programming and optimal control.

In 1921 Hadamard (1) said, discussing the brachistochrone problem

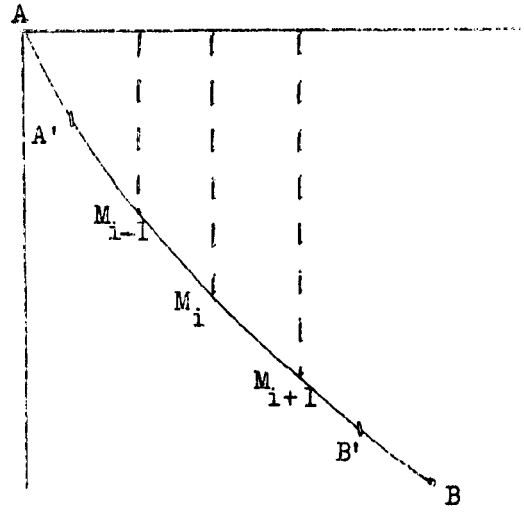


Fig.1a

"For the line considered to be the brachistochrone between A and B, it is necessary that the whole arc A'B' of this line (see Fig.1a) be the brachistochrone between A' and B'. It is this principle which, applied to the small section M_{i-1} M_{i+1} provides us with a solution. It is clear that this is general and that it will recurr in all similar problems".

This principle, expressed by Hadamard, has been taken up recently as the 'Principle of Optimality' of dynamic programming as applied to discrete optimization problems (Bellman (1), Bellman and Dreyfus (1)) and extended to continuous problems by Dreyfus (1). A full discussion of the connections between dynamic programming and variational calculus is given by Dreyfus (1).

The equivalence of certain problems in optimal control and variational problems is established by Hestenes(1), (see also Gumowski and Mira (1)). As an example of this equivalence we give the optimal control formulation of the brachistochrone. Recall that as a variational problem we had

Determine $\bar{y}(x) \in Y$,

$$Y = \left\{ y(x):\ y \in C^1(0,X_o),\ y(0) = 0,\ y(X_o) = y_o \right\}$$

such that

$$I(\bar{y}) \leq I(y) = \int_0^{X_o} \sqrt{\frac{1+y'^2}{2gy}}\ dy \qquad \forall_{y \in Y}$$

For an optimal control formulation we describe the system parametrically. Let t, the time, be the independent variable, and let x(t), y(t) denote the position of the particle at time t. Let u(t) be the gradient of the curve joining (0,0) and (X_o,Y_o) (see Fig.1). We refer to $\begin{pmatrix} x \\ y \end{pmatrix}_t$ as the state vector and u(t) as the control. x(t), y(t) satisfy the differential equations of motion

$$\dot{x} = \sqrt{2gy}\ \cos\ u$$

$$\dot{y} = \sqrt{2gy}\ \sin\ u$$

with the initial conditions $\begin{pmatrix} x \\ y \end{pmatrix}_{t=0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

Define terminal conditions of the form

$$f_1(x,y,t) = 0\ ,\quad f_2(x,y,t) = 0$$

In this case we have

$$f_1\ (x,y,t) = x - X_o$$

$$f_2\ (x,y,t) = y - Y_o$$

and let $t = T_o$ be the first time for which these conditions hold.

Define a criterion function $g(x(T_o),y(T_o),T_o)$. In this case $g(x(T_o),y(T_o),T_o) = T_o$.

Determine the control u(t) such that the criterion function g is minimized.

1.4 Differential Equations

The work of Euler and Lagrange on the solution of the problems of

variational calculus produced a very close inter-relationship between certain types of differential equation and these problems. Because of the wide understanding of the principles of differential equations, and methods of deriving their solution, this relationship has provided the most usual method of solution of variational problems. We now develop this connection.

Solutions of the problem of determining a minimizing function $\bar{y}(x)$ for the integral (1.2) may be considered to be of two types. If $\bar{y}(x)$ is such that

$$I(\bar{y}) \leq I(y) \qquad \text{for all } y \in Y \text{ such that}$$

$$|y(x) - \bar{y}(x)| < \varepsilon \quad , \quad x \in \Omega$$

and

$$|y^{j}(x) - \bar{y}^{j}(x)| < \varepsilon_j \quad , \quad x \in \Omega, \; j=1\ldots k \qquad \ldots(1.3)$$

then $\bar{y}(x)$ is said to afford a minimum of $I(y)$ in a 'weak neighbourhood' of itself, or briefly, to be a 'weak solution' of the problem. If the conditions (1.3) may be omitted, then $\bar{y}(x)$ is said to afford a minimum of $I(y)$ in a 'strong neighbourhood' or to be a 'strong solution' of the problem. It is apparent that any strong solution is also a weak solution.

The Euler-Lagrange method of solution of a variational problem exploits a necessary condition for a function $y(x)$ to be a weak solution. This condition may be stated as

$$\sum_{S=0}^{k} (-1)^{s} \frac{d^{s}}{dx^{s}} \left(\partial F / \partial y^{(s)} \right) = 0 \qquad \ldots(1.4)$$

where the partial derivatives $\partial F / \partial y^{(s)}$ are obtained by considering F as a function of $k+2$ independent variables $x, y, y' \ldots y^{k}$.

Derivations of this condition are to be found in many texts on

Variational Calculus, of which we may mention Bliss (1), Elsgolc (1), Fox (1) and Bolza (2). The boundary conditions imposed on the differential equations are the conditions

$$g_i\ (x,y,y' \cdots y^{k-1}) = 0 \quad \text{on} \quad \delta\Omega_i$$

imposed on the variational problem, together possibly with certain additional boundary conditions, called natural boundary conditions, of which we shall say more later in this section.

The differential equation (1.4) assumes a simple and important form when $F(x,y,y' \cdots y^k)$ is a simple quadratic integrand, that is, where

$$F(x,y,y' \cdots y^k) = \sum_{s=0}^{k} p_s(x) \left(\frac{d^s}{dx^s}\, y(x)\right)^2 + 2f(x).y(x) \qquad \ldots(1.5)$$

since then we have

$$\sum_{t=0}^{k} (-1)^t \frac{d^t}{dx^t}\left(\frac{\partial F}{\partial y(t)}\right) = \sum_{s=0}^{k} (-1)^s \frac{d^s}{dx^s}\left(p_s(x)\ \frac{d^s y}{dx^s}\ (x)\right) + f(x) = 0 \qquad \ldots(1.6)$$

which is a linear differential equation.

Similarly, if

$$F(x,y,y' \cdots y^k) = \sum_{s=0}^{k} p_s(x) \left(\frac{d^s}{dx^s}\ y(x)\right)^2 + 2\int_0^{y(x)} f(x,\eta)\,d\eta \qquad \ldots(1.7)$$

which is the case in certain physical systems which include a non-linear forcing term, the differential equation becomes

$$\sum_{s=0}^{k} (-1)^s \frac{d^s}{dx^s}\left(p_s(x)\ \frac{d^s y(x)}{dx^s}\right) + f(x,y) = 0 \qquad \ldots(1.8)$$

which equation is linear in derivatives of y, although non-linear in y. Problems of this type are known as mildly-non-linear differential equations, and will be discussed in Chapter Four.

The equations (1.6) and (1.8) can be written in the forms

$$L(y(x)) + f(x) = 0$$

and

$$L(y(x)) + f(x,y) = 0$$

respectively, where L is a linear differential operator of order 2k

$$L\,y(x) = \sum_{s=0}^{k} (-1)^s \frac{d^s}{dx^s}\left(p_s(x) \frac{d^s}{dx^s} y(x) \right)$$

The operator L is not only linear but also, when subject to certain boundary conditions, self-adjoint. (For definitions and a discussion of adjoint and self adjoint operators see e.g. Lanczos (2), p.179 et seq.).

The knowledge that the Euler equation of a simple quadratic variational problem corresponds to a self adjoint linear differential equation is perhaps more useful when considered in reverse; that to every self adjoint differential equation there corresponds a simple quadratic variational problem which, from the relations above, can be easily written down. For example Bolza (ref. Collatz (1, p.208)) has shown that, in theory at least, any second order (non-singular) differential equation may be written in self-adjoint form. In practice the transformation to self-adjoint form may require the solution of a partial differential equation and may not be so easily achieved. However, equations of the form

$$p(x)y'' + r(x)y' + q(x)y + f(x) = 0$$

may be expressed as the Euler equation of a simple quadratic variational

problem by first multiplying throughout by

$$\rho(x) = \exp\left(\int_0^x \frac{r(\eta)-p'(\eta)}{p(\eta)} \, d\eta \right)$$

when they can be reduced to the form

$$\frac{d}{dx}\left(p(x)\rho(x)\frac{d}{dx}\,y(x)\right) + q(x)\rho(x)y(x) + f(x)\rho(x) = 0$$

$$= \sum_{s=0}^{1} (-1)^s \frac{d^s}{dx^s}\left(p_s^*(x) \frac{d^s}{dx^s}\, y(x)\right) + f^*(x) = 0$$

provided that the integral $\int_0^x \frac{r(\eta)-p'(\eta)}{p(\eta)} \, d\eta$ exists.

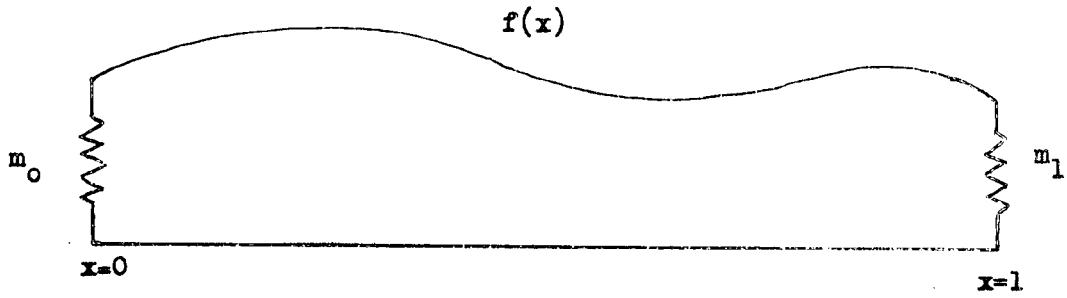
1.5 The relationship between variational and differential boundary conditions:

The determination of a particular solution of a differential equation of order $2k$ requires, in general, $2k$ boundary conditions. The equivalent variational problem, where one exists, will sometimes require $2k$ boundary conditions, but sometimes the quadratic and linear forms occurring in the variational formulation of the problem are such that any function $y(x)$ which minimizes $I(y)$ will automatically satisfy certain conditions, and the number of conditions which it is necessary to impose will be less than $2k$. Conditions which are automatically required of a solution by the form of the integral $I(y)$ are said to be 'natural conditions' of the variational problem.

As an example we take the following problem, discussed by Courant (1), p.139.

Consider a homogeneous string which is elastically tied at both ends by forces of intensity m_0 per unit displacement at $x=0$, and m_1 per unit

displacement at x=1, and subject to an external force f(x). Determine the equilibrium position of the system.



As a differential problem, we have the following equation for u(x), the equilibrium position of the system :-

$$\frac{d^2u}{dx^2} = f(x) \qquad 0 < x < 1$$

subject to the boundary conditions

$$m_o\, u(0) = u'(0) \qquad x = 0$$

$$m_o\, u(1) = -u'(1) \qquad x = 1$$

The total energy of the system is given by

$$I(u) = \int_0^1 \left\{ \frac{1}{2}\left(\frac{du}{dx}\right)^2 + f.u \right\} dx + m_o u^2(0) + m_1 u^2(1) \qquad \ldots(1.9)$$

We can consider the function u_o which minimizes the functional I(u) as one member of a class of functions of one parameter ε ,

$$v_\varepsilon (x) = u_o(x) + \varepsilon.h(x) \quad , \quad h(x) \neq 0$$

where h is required only to possess sufficient continuity properties such that $I(v_\varepsilon)$ is defined. Then the statement $I(u_o) \leq I(v_\varepsilon)$ for all functions v_ε implies the following condition

$$\left.\frac{dI(v_\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} = 0$$

We have

$$I(v_\varepsilon) = I(u_o + \varepsilon h)$$

$$= \int_o^1 \left\{ \frac{1}{2}(u_o' + \varepsilon h')^2 + f.(u_o + \varepsilon h) \right\} dx + m_o(u_o(0) + \varepsilon h(0))^2 + m_1(u_o(1) + \varepsilon h(1))^2$$

and

$$\frac{dI(V_\varepsilon)}{d\varepsilon} = \int_o^1 \{u_o' h' + f.h\} dx + m_o u_o(0).h(0) + m_1 u_o(1).h(1)$$

Integrating the first term by parts, and taking minus signs throughout we obtain

$$\left.\frac{dI(v_\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} = \int_o^1 h.(u_o'' - f)\, dx - (m_1 u_o(1) + u_o'(1)).h(1) - (m_o u_o(0) - u_o'(0)).h(0)$$

$$= 0$$

We recall that this condition is satisfied by any function u_o which minimizes $I(u)$, and that this must be true irrespective of the function $h.(x)$, subject only to $h(x) \neq 0$. Selecting in turn

a) $h(0) = h(1) = 0$

b) $h(0) = 0$, $h(1) \neq 0$

c) $h(0) \neq 0$, $h(1) = 0$

we obtain the Euler equation

$$u_o'' - f = 0 \qquad \ldots(1.10)$$

with the conditions

$$\left.\begin{aligned} u'(1) + m_1 u(1) &= 0 \\ u'(0) - m_o u(0) &= 0 \end{aligned}\right\} \qquad \ldots(1.11)$$

which must be satisfied by u_o.

We have shown that if u_o minimizes $I(u)$ then u_o must satisfy the differential conditions (1.10) and (1.11). We have not shown, however, that for m_o, m_1 finite, we need not impose these conditions on

the functions with which we attempt to minimize $I(u)$, whilst for m_o, m_1 infinite, corresponding to the fixed boundary conditions

$$u(0) = u(1) = 0$$

this condition must be imposed on trial functions. To do this we note that we implicitly require the potential energy of the system to be finite, and if $m_o \to \infty$ then $I(u) \to \infty$ for all u not satisfying $u(0) = 0$, and hence this condition, and similarly the condition $u(1) = 0$ when $m_1 \to \infty$, have to be imposed.

Given a self adjoint differential equation of order $2k$,

$$Ly = f$$

and $2k$ boundary conditions of the form

$$g_i(x,y,y' \cdots y^{t_i}) = 0 \qquad i = 1 \cdots 2k$$

there is a simple criterion for determining whether g_i is a natural or essential boundary condition for the corresponding variational problem.

If $t_i \leq k - 1$ then the boundary condition g_i is essential, otherwise it is natural.

1.6 The Ritz and Galerkin methods.

The fundamental method for the solution of variational problems other than that of reduction to the Euler equation, is the method of minimizing sequences. Let $y_n(x)$ be a sequence of functions complete in the space Y of admissible functions. Then y_n is a minimizing sequence for the functional $I(y)$ if and only if

$$I(y_{n+1}) \leq I(y_n) \qquad \text{for all } n$$

and $\lim_{n\to\infty} I(y_n)$ exists, say $\lim_{n\to\infty} I(y_n) = I_o$

Then $I_o = \inf I(y)$, $y \in Y$

and

$\lim_{n\to\infty} y_n = y_o$ is the solution of the variational problem.

The Ritz method is one method of constructing such a sequence; there are many others, examples of which are given in Mikhlin and Smolitskiy (1) and Gumowski and Mira (1). The basis of the Ritz method is to select a finite sequence of functions $\phi_1 \cdots \phi_n(x)$ which form a basis of a subspace Y_n of the space Y of admissible trial functions; that is, ϕ_i must satisfy the essential boundary conditions of the problem, and be sufficiently differentiable for $I(\phi_i)$ to be defined. Then a trial solution

$$y_n = \sum_{i=1}^{n} \alpha_i \phi_i(x)$$

is computed, where the coefficients α_i are determined from the necessary condition for a minimum of $I(y_n)$ in the subspace Y_n, i.e.

$$\frac{dI(y_n)}{d\alpha_i} = 0 \quad , \quad i = 1 \cdots n \qquad \ldots(1.12)$$

A sequence of such approximate solutions y_n constitutes, subject to certain conditions concerning the convergence of the sequence of subspaces, and similar problems (see e.g. Hilbert (1)), a minimizing sequence for the variational problem. More detailed discussions of the approximation and convergence theory will be given throughout this thesis, particularly in Chapter Two.

In the particular case of simple quadratic problems, where

$$I(y) = \int_o^1 \left\{ \sum_{s=0}^{k} p_s(x) \left(\frac{d^s}{dx^s} y(x) \right)^2 + 2f(x).y(x) \right\} dx$$

the equations (1.12) assume the particularly simple linear form

$$A \underline{\alpha} + \underline{b} = 0 \qquad \ldots(1.13)$$

where

$$A = (a_{ij}) \qquad i,j = 1 \cdots n$$

$$\underline{b} = (b_i) \qquad i = 1 \cdots n$$

$$a_{ij} = \int_0^1 \sum_{s=0}^{k} p_s(x) \frac{d^s}{dx^s} \phi_i(x) \frac{d^s}{dx^s} \phi_j(x) \, dx \qquad \ldots(1.14)$$

$$b_i = \int_0^1 f(x) . \phi_i(x) \, dx$$

whilst for mildly non-linear problems with integrands of the form (8) the equations are

$$A \underline{\alpha} + g(\underline{\alpha}) = 0 \qquad \ldots(1.15)$$

where $A = (a_{ij})$, $i,j = 1 \cdots n$ is defined by (1.14)

and

$$g_i(\underline{\alpha}) = \int_0^1 f(x, \sum_{j=1}^{n} \alpha_j \phi_j(x)) . \phi_i(x) \, dx \qquad i = 1 \cdots n$$

The Galerkin method is not a variational method, and can be successfully applied to many differential equations which will not admit a variational treatment. However, in the case of differential problems corresponding to variational problems with integrands of types (1.5) or (1.7), that is the problems (1.6) or (1.8) the Ritz and Galerkin methods produce identical systems of linear or mildly non-linear equations. The Galerkin method can be described as follows.

Let

$$G(x,y,y' \cdots y^t) = 0 \quad \text{in a region } \Omega \qquad \ldots(1.16)$$

be the given differential equation, with boundary conditions

$$g_i(x,y,y' \cdots y^{t-1}) = 0 \quad \text{on} \quad \delta\Omega_i \qquad \cdots(1.17)$$

$$i = 1 \cdots t$$

Let Y, the domain of existence of G, be the set of all functions satisfying (1.17) for which the differential operator $G(x,y,y' \cdots y^t)$ is defined. Select in Y a set of n linearly independent functions $\phi_1(x) \cdots \phi_n(x)$, defining a subspace Y_n of Y.

Define

$$y_n = \sum_{i=1}^{n} \gamma_i \phi_i(x)$$

to be an approximate solution of (1.16), (1.17) if the residual

$$G(x,y_n,y_n' \cdots y_n^t)$$

is orthogonal in the scalar product of some space containing Y to the functions $\phi_1, \phi_2 \cdots \phi_n$.

The Galerkin method is normally applied in this form. If the residual is made orthogonal to a set of n linearly independent functions $\psi_i(x)$, $i = 1 \cdots n$, $\psi_i(x) \neq \phi_i(x)$, the method is referred to as the 'method of projections' - see Collatz (1).

When the Galerkin method is applied to differential equations of the type of (1.6) the resulting equations are (assuming an integral norm)

$$\int_0^1 \left\{ \sum_{s=0}^{2k} (-1)^s \sum_{i=1}^{n} \gamma_i p_s(x) \frac{d^{2s}}{dx^{2s}} (\phi_i(x)) + f(x) \right\} \cdot \phi_j(x)dx \quad \cdots(1.18)$$

whilst for equations of the form (1.8) we have

$$\int_0^1 \left\{ \sum_{s=0}^{2k} (-1)^s \sum_{i=1}^{n} \gamma_i p_s(x) \frac{d^{2s}}{dx^{2s}} (\phi_i(x)) + f(x, \sum_{i=1}^{n} \gamma_i \phi_i(x)) \right\} \cdot \phi_j(x)dx \quad \cdots(1.19)$$

If we assume homogeneous boundary conditions

$$\frac{d^s}{dx^s} y \Big|_{y=0} = \frac{d^s y}{dx^s} \Big|_{y=1} = 0 \qquad \cdots(1.20)$$

$$s = 0,1..k-1$$

then the terms

$$\int_0^1 (-1)^s \frac{d^{2s}}{dx^{2s}} \phi_i(x) \cdot \phi_j(x)\, dx$$

occurring in (1.18) and (1.19) can be expressed as

$$\int_0^1 \frac{d^s}{dx^s} \phi_i(x) \cdot \frac{d^s}{dx^s} \phi_j(x)\, dx \qquad \ldots(1.21)$$

and the equations (1.18) and (1.19) reduce to the forms given in (1.13) and (1.15).

If the differential equation does not have boundary conditions of the form (1.20) then the integration by parts performed to derive (1.21) produces boundary terms similar to those occurring in the expression (1.9) of Courant's example (p.14).

The equivalence of variational problems with certain differential equations extends in this way to encompass methods which when applied to these particular problems are equivalent. It is possible, therefore, to make theoretical deductions concerning the Ritz method from a theoretical knowledge of the Galerkin method, whilst properties of the Ritz method may (though not necessarily) imply the same properties for the Galerkin method. It is the case, however, that theoretical study of the Ritz method has provided more specific theoretical results than those available for the more general Galerkin method, and so the former approach has not been necessary. Conversely, the generality of the Galerkin method is such that implications of variational theory have not, in general, been shown to hold for all cases of the Galerkin method.

Chapter Two

Variational Theory for Differential Equations

We are concerned for most of this chapter with variational problems involving integrands of the type (1.5) and the corresponding differential problems (1.6). Integrands such as (1.7) and their corresponding problems, are considered in section 2.8.

2.1 Energy

The physical concept of energy, which was useful in Chapter One for the discussion of Courant's example, can be expressed in abstract mathematical terms. In this section we develop some of the theory of this abstraction, and show how the application of energy principles to linear differential operators of certain types leads to a formal account of the Rayleigh-Ritz method. Much of the treatment of this section will follow the works of Mikhlin (1,2,3) and Mikhlin and Smolitskiy (1).

Let H be an arbitrary Hilbert space, with scalar product (u,v) and associated norm $\|u\| = (u,u)^{\frac{1}{2}}$. Let L be an arbitrary linear operator acting on H. Then the domain of existence of L, denoted $D(L)$, the set of all functions $u \in H$ such that Lu is defined, satisfies $D(L) \subseteq H$. $D(L)$ may be strictly within H, for example, let

$$H = \mathcal{L}_2\,[0,1] \quad \text{, the set of square integrable functions on } [0,1] \text{ ,}$$

and

$$L = {}^{d^2}/_{dx^2} \text{ .}$$

Then $D(L) = \mathcal{L}_2^2\,[0,1]$, the set of functions on $[0,1]$ having square-integrable second derivatives, and clearly

$$D(L) \subseteq \mathcal{L}_2\,[0,1] = H \text{ .}$$

An operator L is symmetric if and only if, for any functions $u, v \in D(L)$ we have the identity $(u.Lv) \equiv (Lu.v)$.

A symmetric operator L is positive if and only if, for any $u \in D(L)$, the relation $(Lu.u) \geq 0$ holds and $(Lu.u) = 0$ if $u \equiv 0$.

A symmetric operator L is positive definite if and only if $(Lu.u) \geq \gamma^2 \|u\|^2$, γ real const, $\gamma \neq 0$.

If an operator L is positive or positive definite then $(Lu.v)$ is a scalar product for certain functions u,v in some space. We denote $(Lu.v)$ by $(u.v)_L$ and refer to it as the energy product; the corresponding Hilbert space is denoted by H_L and referred to as the energy space of the operator L. Associated with the energy product is the energy norm $\|u\|_L$, defined by $\|u\|_L = (Lu.u)^{\frac{1}{2}} = (u.u)_L^{\frac{1}{2}}$

If L is a positive definite operator then $H_L \subseteq H$, if L is only positive then this need not be so. To investigate this we consider the case where L is a linear differential operator, since this is the particular case of interest to us. The distinction is made in terms of generalized derivatives of a function, which we now introduce. In a sense these generalized derivatives allow the extension of the method of integration by parts, and as a result of this the scalar product $(u.v)_L$ can be applied to functions u,v which do not satisfy $u,v \in D(L)$.

Let Ω be a finite region of m dimensional space, with boundary $\delta\Omega$ and let $\bar{\Omega} = \Omega + \delta\Omega$. By a bounded strip of width ε in Ω we mean the set of all points $x \in \Omega$ such that for a particular $\varepsilon > 0$ there exists a point $\bar{x} \in \delta\Omega$ satisfying

$$\|x - \bar{x}\|_\Omega < \varepsilon$$

where $\|x - \bar{x}\|_\Omega$ denotes some norm in the space Ω. Denote by ϕ_k the set of all functions which are k times differentiable in Ω

and zero in some bounded strip of width $\epsilon \geq 0$ in Ω.

If $\phi \in \Phi_k$, then for any function $u \in C^k(\Omega)$ we have the relation

$$\int_\Omega u \frac{\partial^k \phi}{\partial x_{i_1} \cdots x_{i_k}} d\Omega_k = (-1)^k \int_\Omega \phi \frac{\partial^k u}{\partial x_{i_1} \cdots x_{i_k}} d\Omega_k$$

which we can verify by integration by parts.

Now consider a function $v \in \mathcal{L}_2(\Omega)$. If there exists a function $w \in \mathcal{L}_2(\Omega)$ such that for any $\phi \in \Phi_k$ we have

$$\int_\Omega v \cdot \frac{\partial^k \phi}{\partial x_{i_1} \cdots x_{i_k}} d\Omega_k = \int_\Omega \phi \cdot w \, d\Omega_k$$

then we say w is the k^{th} generalized derivative of v.

(Note: $d\Omega_k = \partial x_{i_1} \cdots \partial x_{i_k}$)

The k^{th} generalized derivative, $z(x)$, of the function v is denoted

$$\frac{\partial^k v}{\partial x_{i_1} \cdots \partial x_{i_k}}$$

as are the conventional derivatives.

As an example of a positive definite operator and its corresponding energy space we can take the example given above, where the Hilbert space $H = \mathcal{L}_2[0,1]$ and the operator $L = d^2/dx^2$ acts on functions $y(x)$ satisfying the differentiability conditions imposed by L and $y(0) = y(1) = 0$.

Then

$$H = \mathcal{L}_2[0,1], \quad D(L) \subset C^1[0,1].$$

$$(u,v)_L = \int_0^1 v \cdot \frac{d^2u}{dx^2} dx.$$

Clearly $(u,v)_L$ is defined for functions u,v possessing square-summable first generalized derivatives, that is, the energy space H_L is given by

$$H_L = W_1^2[0,1]$$

where $W_k^p[\Omega]$ is the Sobolev space of functions having all k^{th} generalized derivatives which are p-summable in Ω, and the functions themselves are p-summable in Ω. We note that clearly

$$H_L = W_1^2[0,1] \subset H = \mathcal{L}_2[0,1]$$

and also that

$$D(L) \subset H_L .$$

The condition $D(L) \subseteq H_L$ is a necessary condition for an energy space and the corresponding domain of existence of an operator; we have for a positive definite operator L

$$D(L) \subseteq H_L \subseteq H .$$

($H_L \subseteq H$ was first proved by Friedrichs (1)).

To show that these relations need not hold if L is only positive, we consider the following examples of the Laplacian operator in three dimensions, taken from Mikhlin (3, pp.16 & 17).

Let $S(R)$ denote the sphere

$$x_1^2 + x_2^2 + x_3^2 = R^2$$

and let Ω be the exterior of $S(1)$. Again we take $H = \mathcal{L}_2[\Omega]$. Let $M \subset H$ be the set of functions u in H satisfying

a) u is twice continuously differentiable in $\Omega + S(1)$

b) $u = 0$ on $S(1)$

c) $\int_{S(R)} \bar{u} \frac{du}{dn} \to 0$ as $R \to \infty$

d) $u \in M \Rightarrow \Delta u \in \mathcal{L}_2(\Omega)$, where Δ denotes the three-dimensional Laplacian operator.

Then $-\Delta$ is positive (but not positive definite) on M, and so we have an inner product

$$(u,v)_{-\Delta} = \int_\Omega \sum_{i=1}^{3} \frac{\partial u}{\partial x_i} \frac{\partial \bar{v}}{\partial x_i} \, \partial\Omega$$

and the associated energy space $H_{-\Delta}$.

The problem

$$\Delta u = {}^{1}/_{R^4} \quad , \quad u = 0 \text{ on } S(1)$$

has the solution

$$\hat{u} = {}^{1}/_{2}({}^{1}/_{R^2} - {}^{1}/_{R}) ,$$

so that $\hat{u} \in H_{-\Delta}$, $\hat{u} \notin \mathcal{L}_2(\Omega)$

and therefore

$$H_{-\Delta} \not\subset H = \mathcal{L}_2(\Omega).$$

Furthermore, the problem

$$\Delta u = {}^{1}/_{4} \; R^{-5/2} \ln R \qquad u = 0 \text{ on } S(1)$$

has the solution

$$\hat{u} = R^{-1/2} \ln R , \text{ and}$$

$$\hat{u} \notin H_{-\Delta}, \quad u \notin H = \mathcal{L}_2(\Omega) .$$

<u>The Energy Problem</u>

We consider the linear equation $Ly = f$ where L is a positive operator defined on a Hilbert space H and f is defined on an arbitrary Hilbert space. Define

$$I(y) = (Ly.y) - 2(y.f) \qquad \ldots(2.1)$$

The following theorem is given by Mikhlin and Smolitskiy (1).

Let L be a positive operator. If the equation $Ly = f$ has a solution then this solution strictly minimizes $I(y)$.

Conversely, if there exists an element which minimizes $I(y)$, this element satisfies the equation $Ly = f$.

Since we have already noted that $I(y)$ is defined for certain functions for which Ly is not, it is important to clarify the wording of this theorem. For the case with which we are primarily concerned, where L is a second order differential operator and y a function of

one variable, this clarification may be obtained from the Dubois-Raymond lemma or from Hilbert's derivation of the Euler equation (Bolza (2,p.22)), both of which show that in this case, even if the function y_0 minimizing $I(y)$ has only a continuous first derivative then it must satisfy the Euler equation $Ly = f$. Additionally, the Hilbert derivation implies the existence of a second derivative of y_0 for all values of x for which (using the notation of § 1.2, § 1.4)

$$\frac{d^2}{dy'dy'} \; F(x,y_0(x),y_0'(x)) \neq 0$$

Mikhlin (3) points out that this result does not extend to multi-dimensional variational problems. Instead, for these general problems, we have to assume that $y_0(\underline{x}) \notin D(L)$, $y_0(\underline{x}) \subset H_L$ and it can be shown (Mikhlin and Smolitskiy (1), p.156) that $y_0(\underline{x})$ satisfies an equation related to the equation $Ly = f$, and such solutions are said to be generalized solutions of $Ly = f$.

If the operator L is positive, but not positive definite, then the equation $Ly = f$ has a solution iff the scalar product $(y.f)$ of $I(y)$ is bounded above for all $y \in H_L$, since then, by the Reisz theorem, there exists a function $y_0 \in H_L$ s.t. $(y.f) = (y.y_0)_L$ and y_0 is a solution of the minimization problem for $I(y)$, and a solution in some sense of the problem $Ly = f$.

The problem:-

Given a positive or positive definite operator L, find y_0 such that

$$I(y_0) \leq I(y) = (Ly.y) - 2(y.f) \qquad \forall_{y \in H_L} \;, \; y_0 \in H_L$$

is referred to as the energy problem, or sometimes as the energy method. The phrase 'energy problem' is preferred throughout this thesis.

Methods of Solution of the Energy Problem

The methods of minimizing sequences, mentioned in § 1.6, are the most important methods of solution of the energy problem. We are concerned with the Rayleigh-Ritz method, and we give now a more formal description than that of § 1.6.

The Rayleigh-Ritz method produces a sequence of approximate solutions of the energy problem. The n^{th} Rayleigh-Ritz approximation, denoted by $y_n(x)$, is of the form

$$y_n(x) = \sum_{i=1}^{n} a_n(i)\phi_i$$

where $\phi_i \in H_L$, $i = 1 \ldots n$ are linearly independent, and the constants $a_n(i)$ are determined from

$$\frac{\partial}{\partial a_n(j)} I\left(\sum_{i=1}^{n} a_n(i)\phi_i\right) = 0 \quad , \quad j = 1 \ldots n$$

the necessary condition for a minimum of the function of the n variables $a_n(i)$.

Substituting for I the expression (2.1) we have the equations

$$\sum_{i=1}^{n} a_n(i)(L\phi_i,\phi_j) - (f.\phi_j) = 0 \quad , \quad j = 1 \ldots n \ ,$$

written in matrix form as

$$A_n \underline{a}_n = \underline{f}_n \qquad \ldots(2.2)$$

where $A_n = \left\{ a_{ij} : a_{ij} = (L\phi_i.\phi_j) = (\phi_i.\phi_j)_L \ , \quad i,j = 1 \ldots n \right\}$

$$\underline{f}_n = \left\{ f_j \ . \ f_j = (f.\phi_j) \ , \quad j = 1 \ldots n \right\}$$

The functions ϕ_i are referred to as co-ordinate functions or basis functions and the matrix A_n is the Gram matrix of these

functions in the space H_L .

2.2 Theoretical criteria for the selection of co-ordinate systems.

Though the selection of co-ordinate systems is of considerable importance in the practical computational use of the Rayleigh-Ritz method, the criteria imposed by the analytic theory are of a simple nature. It is necessary that the functions ϕ_i satisfy the following

$\phi_i \in H_L$

ϕ_i , i=1..∞ form a complete system in H_L

ϕ_i are linearly independent for all n , since otherwise the Gramm matrix A_n is singular and equations (2.2) have no unique solution.

The particular energy problem with which we are concerned is that for which the equation

$$Ly = f$$

takes the form (from (1.5))

$$\sum_{s=0}^{k} (-1)^s \frac{d^s}{dx^s} \left(p_s(x) \frac{dy^s}{dx^s}\right) = f(x) \qquad \ldots(2.3)$$

subject to boundary conditions

$$\frac{d^r}{dx^r} y(0) = \frac{d^r}{dx^r} y(1) = 0 \quad , \quad r=0, 1..k-1 \qquad \ldots(2.4)$$

that is, problems for which

$$I(y) = \int_0^1 \left\{ \sum_{s=0}^{k} p_s(x) \left(\frac{d^s y}{dx^s}\right)^2 - 2 f(x) y \right\} dx \qquad \ldots(2.5)$$

subject to boundary conditions (2.4)

for which $H_L = \left\{ y(x): \; y \in C^{k-1}[0,1] \; , \; \frac{d^r}{dx^r} y(0) = \frac{d^r}{dx^r} y(1) = 0 \right\}$

Basis functions in use for problems of this type, and for the similar problem given by (1.7) are of two broad types, distinguished in terms of the concept of the 'support' of a function. A function $\phi(x)$ is said to be a global basis function on $[0,1]$ iff $\phi(x) = 0$ at only a finite number of points on $[0,1]$. A function is said to have compact support iff there exist two numbers α, β such that $0 \leq \alpha < \beta \leq 1$ and $\alpha = 0 \Rightarrow \beta \neq 1$, or $\beta = 1 \Rightarrow \alpha \neq 0$ and $\phi(x)$ is a global function on $[\alpha, \beta]$, $\phi(x) \equiv 0$ on $[0,\alpha)$, $(\beta, 1]$.

Global basis functions have been considered theoretically by Kantorovich and Krylov (1), Mikhlin (1,2,3), Mikhlin and Smolitskiy (1), and extensively used in practice. Functions with compact support have been the underlying basis for finite difference and finite element methods, and are considered theoretically in a variational setting in papers by Ciarlet, Schultz, Varga, and others. In § 2.3 and § 2.4 we consider examples of functions with complete and compact support respectively.

2.3 Basis functions with complete support

Co-ordinate systems with this property were used for many years to the exclusion of all others, and as Schultz says, (1,p.303), if it were not of importance computationally for A to possess certain features, basis functions of this type would always be used. The most frequently used systems for problem (2.5) are the following

$$\phi_i(x) = x^i . x^{k-1}(1-x)^{k-1} \qquad \ldots(2.6)$$

and

$$\phi_i(x) = \sin(i \pi x) \qquad \ldots(2.7)$$

$(k=1$ in $(2.5))$

or certain weightings and combinations of these. Important for our

purposes in this context are the functions

$$\phi_i(x) = x^{i+1}(1-x)\ (i+1)^{i+1} / i^i \qquad \text{(k=1 in (2.5))} \qquad \ldots(2.8)$$

and

$$\phi_i(x) = \frac{\sqrt{2}}{i\pi} \sin i\pi x \qquad \text{(k=1 in (2.5))} \qquad \ldots(2.9)$$

Ciarlet, Schultz and Varga (1) use the functions

$$\phi_i(x) = \int_0^x P_{i-1}(2t-1)dt \qquad \ldots(2.10)$$

where $P_{i-1}(x)$ is the Legendre polynomial of degree i-1 on $[-1,1]$, and we shall also examine the functions

$$\phi_i(x) = x(1-x)\ P_{i-1}(2x-1) \qquad \ldots(2.11)$$

and

$$\phi_i(x) = x(1-x)\ T_{i-1}(2x-1) \qquad \ldots(2.12)$$

where $T_{i-1}(x)$ is the Chebyshev polynomial of degree i-1 on $[-1,1]$. These three co-ordinate systems will be shown to possess computationally convenient properties.

2.4 Basis functions having compact support

The wide variety of piecewise polynomial and spline basis functions which have been introduced by Ciarlet, Schultz and Varga (1), and their collaborators may be considered as follows.

Let $\overline{\Pi}_N = \{x_0, x_1 \ldots x_{N+1}\}$ be a partition of the interval $[0,1]$ such that $x_0 = 0$, $x_{N+1} = 1$, $x_i < x_{i+1}$, $i = 0 \ldots N$. Then the approximate solution $y_n(x)$ of the equations (2.2), where the basis $\phi_i(x)$, $i = 1 \ldots n$, is a piecewise polynomial or spline basis defined on $\overline{\Pi}_N$, will be a function in the space $Q_o^m(\overline{\Pi}_N)$ of functions $q(x)$ having the properties $q(x) = v_i(x)$ on each interval $[x_i, x_{i+1}]$,

$i = 0 \ldots N$ where $v_i(x)$ are polynomials of degree $2m-1$ in x where

$$x \in (x_i , x_{i+1}) \qquad \ldots(2.13).$$

$v_i(x)$, $i = 0 \ldots N$ are such that $q(x) \subset C^t[0,1]$, $t \geq m-1$, (this implies continuity properties at the nodes x_i of the mesh Π_N)...(2.14)

$$v_0(0) = 0 , \qquad v_N(1) = 0 \qquad \ldots(2.15)$$

The particular type of piecewise polynomial or spline representation for y_n is determined by the continuity properties implied in (2.14). The following examples are given to indicate the differences. Other types may be found in Ciarlet, Schultz, Varga (1) and Schultz and Varga (1).

Hermite piecewise-continuous polynomials ...(2.16)

At each point x_i , $i = 0 \ldots N+1$ of the partition Π_N consider that there are m interpolation parameters d_i^s , $0 \leq s \leq m-1$, $0 \leq i \leq N+1$. Then in each interval $[x_i, x_{i+1}]$ there is a unique interpolating polynomial $v_i(x)$ of degree $2m-1$ such that

$$\frac{d^s}{dx^s} v_i(x_i) = d_i^s \quad , \quad \frac{d^s}{dx^s} v_i(x_{i+1}) = d_{i+1}^s$$

$$0 \leq s \leq k-1$$

Clearly a function $q(x)$ defined as above in terms of these $v_i(x)$ satisfies $q(x) \in C^{m-1}[0,1]$, and taking $d_0^s = d_{N+1}^s = 0$ we can satisfy the boundary conditions of the problem, i.e. (2.4). There remain $m(N+2) - 2k$ interpolation parameters to be determined by the Ritz procedure.

Spline piecewise continuous basis functions ...(2.17)

These are the functions which satisfy (2.13) .. (2.15), with $t = 2m-2$ in (2.13). As such they may be considered as Hermite piecewise continuous polynomials satisfying additional continuity

requirements so that $q(x) \in C^{2m-2}[0,1]$. To do this we consider as parameters the quantities d_o^s, d_{N+1}^s $s = 0,1\ m-1$ and d_i^o, $i = 1 \ldots N$. The values d_i^s, $i = 1 \ldots N$, $s = 1 \ldots m-1$ are then determined in terms of the parameters so that $q(x)$, defined by

$$q(x) = v_i(x) \quad , \quad x_i \leq x \leq x_{i+1}$$

satisfies $q(x) \in C^{2m-2}[0,1]$. There are then $n + 2(m-k)$ parameters, since the $2k$ values $d_o^s = 0$, $d_{N+1}^s = 0$, $s = 0,1 \ldots k-1$ are prescribed by the boundary conditions.

These piecewise continuous basis functions, and others considered by Ciarlet, Schultz and Varga (1) are special cases of the L-spline functions (not necessarily polynomial functions) defined by Schultz and Varga (1), and Perrin, Price and Varga (1). The complexity of a general treatment of such basis functions is such that we avoid a summary of it. In practice only particular cases, such as those given above, seem useful because of the difficulty of determining the functions $v_i(x)$ in the more general cases. In §2.5, where convergence results given for the piecewise continuous functions defined above can be strengthened for the more general class of L-spline functions this is indicated.

In passing we comment that in practice a further distinction exists between Rayleigh-Ritz approximations computed using basis functions such as (2.6)..(2.12) and those computed using (2.16), (2.17). That is that the sequence index n of the approximation $y_n = \sum_{i=1}^{n} a_n(i)\, \phi_i(x)$ is determined by the two parameters m and N of the respective subspaces, and thus the sequence $\{y_n\}$ should more properly be considered as $\{y_{m,N}\}$.

Convergence of the Rayleigh-Ritz Method

There are two approaches to the convergence problem for the Rayleigh Ritz method, which may be conceived as the problem of convergence of the coefficients of the approximation, and the problem of convergence of the function approximation defined by these coefficients and the sequence of basis functions. We look at each separately.

2.5 The convergence of Rayleigh Ritz coefficients

Not only is it important to discover under what circumstances the Rayleigh Ritz coefficients converge at all, but also it is convenient to have some estimate of the rate of convergence of the coefficients for different expansion sets, so as to estimate the number of terms likely to be required in the expansion of a particular Rayleigh-Ritz approximation, and indeed, to determine the computational feasibility of such an approximation. After some initial convergence results, taken from Mikhlin (1),(4), we examine the work of Delves and Mead (1),(2) on the estimation of convergence rates from simple properties of the Rayleigh-Ritz matrix. The theorems of Mikhlin rely on concepts which we find it more convenient to introduce fully in § 3.1; we briefly indicate them here.

A sequence of functions $\left\{\phi_i\right\}_{i=1}^{\infty}$ is minimal in a Hilbert space H if the space H_k formed by linear combinations of $\phi_1, \phi_2 \ldots \phi_{k-1}, \phi_{k+1} \ldots$ is a proper subspace of H.

A sequence of functions $\left\{\phi_i\right\}_{i=1}^{\infty}$ is strongly minimal in H if the Gramm matrix A_n of $\left\{\phi_i\right\}_{i=1}^{n}$ in H is positive definite for all n.

The following results describe the convergence of the sequence of Rayleigh-Ritz coefficient vector elements $\underline{a}_n(i)$ $\quad i=1 \ldots n,\ n=1 \ldots \infty$

Theorem:

If $\{\phi_i\}_{i=1}^{\infty}$ is minimal in H_L, then there exist constants α_k, $k = 1 \ldots \infty$ such that

$$\lim_{n\to\infty} a_n(k) = \alpha_k \quad , \quad k = 1,2 \ldots \infty \qquad \ldots(2.18)$$

Under fairly general circumstances (Mikhlin (4), p.14, p.23) the limit process is uniform in k.

Theorem

If $\{\phi_i\}_{i=1}^{\infty}$ is strongly minimal in H_L then

$$\lim_{n\to\infty} \| \underline{a}_n - \underline{\alpha} \|_{l_2} = 0$$

where $\underline{\alpha}$ is the vector $(\alpha_k : k = 1 \ldots \infty)$ defined above and

$$\| \underline{x} \|_{l_2} = \left[\sum_{i=1}^{\infty} |x(i)|^2 \right]^{\frac{1}{2}}$$

We can consider that the constants α_k defined in (2.18) are generalized Fourier coefficients of $y(x)$ in terms of the co-ordinate system $\{\phi_i\}_{i=1}^{\infty}$, so that we may take

$$y(x) = \sum_{k=1}^{\infty} \alpha_k \phi_k(x)$$

where y satisfies $Ly = f$.

The Rayleigh-Ritz approximation given by the solution of equations (2.2), i.e.

$$y_n(x) = \sum_{k=1}^{n} a_n(k) \phi_k(x)$$

clearly has an error

$$\varepsilon_n(x) = y(x) - y_n(x) = \sum_{i=1}^{n} (\alpha_i - a_n(i))\phi_i(x) + \sum_{i=n+1}^{\infty} \alpha_i \phi_i(x)$$

Since the variational solution $y_n(x)$ is invariant under a non-singular linear transformation, we may assume that the functions $\{\phi_i\}$ are orthonormal with respect to some scalar product, so that

$$\|\varepsilon_n(x)\|^2 = \sum_{i=1}^{n} (\alpha_i - a_n(i))^2 + \sum_{i=n+1}^{\infty} \alpha_i^2$$

and thus the convergence rate of the variational process depends on

the rates at which $\alpha_i \to 0$ as $i \to \infty$ (Fourier convergence),

$|\alpha_i - a_n(i)| \to 0$ as $n \to \infty$ (horizontal convergence)

and $|\alpha_i - a_i(i)| \to 0$ as $i \to \infty$ (diagonal convergence)

Rates of convergence are given by Delves and Mead (1),(2), and these lead to practical criteria by which co-ordinate systems may be judged, at least from the viewpoint of analytic computation. The terms 'horizontal' and 'diagonal' convergence were introduced by Delves and Mead and refer to a triangular array of the coefficients of successive Rayleigh-Ritz approximations.

$$\begin{array}{lllll} a_1(1) & a_2(1) & a_3(1) \ldots a_n(1) \ldots & & \alpha_1 \\ & a_2(2) & a_3(2) \ldots a_n(2) \ldots & & \alpha_2 \\ & & a_3(3) \ldots a_n(3) \ldots & & \alpha_3 \\ & & & \text{"Horizontal" Convergence} & \\ & & a_n(n) \ldots & \downarrow & \\ & \text{"Diagonal" Convergence} \searrow & & \text{Fourier Convergence} \downarrow & \\ & & & 0 & \end{array}$$

In certain circumstances not all convergence problems may be present in a problem, for example, if the basis functions $\{\phi_i\}_{i=1}^{\infty}$ are orthogonal with respect to the energy norm, then the matrix A_n of (2.2) is diagonal and the solutions of (2.2) are the Fourier coefficients of $y_n(x)$; the horizontal and diagonal convergence problems are not present. This is a somewhat artificial case, and from a general viewpoint all three are important. The convergence of Fourier approximation may be determined from the data of the problem, i.e. from the forms of L and f of the equation $Ly = f$, and this problem, for one dimensional problems, has been considered recently by Delves and Mead (3).

We concentrate now on information concerning the convergence problem derived from the forms of the matrix A_n and free terms b_n of equation (2.2).

The important definition is the following

A matrix A is <u>asymptotically diagonal</u> of degree p if for fixed j and all i

$$\frac{|A(i,j)|}{\{|A(i,i)| \cdot |A(j,j)|\}^{\frac{1}{2}}} \leq C_j \, i^{-p} \qquad \ldots(2.19)$$

where C_j and p are positive constants.

A matrix A is <u>uniformly asymptotically diagonal</u> of degree p if (2.19) holds and there exists a constant C ; $0 < C < \infty$ such that $C_j < C$ for all j .

With these definitions, the following results of Delves and Mead (2) are the most important for our purposes.

Theorem (Delves and Mead (2), Thm. 0)

If the operator $L : H \to H$ and the sequence $\{\phi_i\}$ is orthonormal in H , then the matrix A , $a_{ij} = (\phi_i, \phi_j)_L$ is asymptotically diagonal of degree $p \geq 1/2$ provided $|A(i,i)| \geq C > 0$ for all i .

[Note that A is the infinite matrix formed from $A(i,j) = a_{ij}$, $i, j = 1 \ldots \infty$].

We recall the definition of the energy functional of the problem $Ly = f$, that is

$$I(y) = (Ly,y) - 2(y\,f)$$

Then we have

Theorem (Delves and Mead (2), Thm.5)

Let L be a positive symmetric (Hermitian) operator, and define $y_o(x)$,

$$y_o(x) : \min_{y \in H_L} I(y) = I(y_o)$$

Also, let the matrix A : $A(i,j) = (\phi_i,\phi_j)_L$ be uniformly asymptotically diagonal of degree $p > 1/2$, let $A_{ii} = 1$, and let

$$|\alpha(i)| = |(\phi_i,y)| \leq K\, i^{-q} \quad , \quad q > 1/2 \quad , \quad K > 0$$

where A and $\underline{b}$ are the (infinite) matrix and vector of (2.2). Then if $p + 2q > 2$ the inequality

$$\|\varepsilon_n\|_L \leq \gamma_1\, n^{-2q+1} + \gamma_2\, n^{-(p+2q-2)}$$

holds, (γ_1, γ_2 constants > 0)

We note that weaker results applicable to general algebraic problems in which uniformly asymptotically diagonal matrices occur are also given by Delves and Mead (2).

In the above theorem we have a dependence on the rate at which the unknown Fourier coefficients decrease. The following theorem provides an estimate of this rate

Theorem (Delves and Mead (2), Thm.6)

Let A satisfy the conditions of the above theorem, and assume

$$b(i) = (\phi_i,f) \leq c\, i^{-r} \qquad r \geq 1 .$$

Then there exists a constant D such that

$$|\alpha(i)| \leq D\, i^{-s} \quad , \quad s = \min(p,r)$$

for all integers i .

This theorem describes the Fourier convergence problem of the basis functions. Delves and Mead (2) give the following result for the variational convergence problem.

Theorem (Delves and Mead (2), Thm.7)

With the assumptions of the above theorem

$$|\alpha(i) - a_n(i)| \leq D_1 n^{-(2p-1)} i^{-(r-1)} + D_2\, n^{-q'}$$

$$i = 1 \ldots n$$

where $q' = \min(p+r-1,\ 2p-1,\ 2p+r-2)$ and D_1 , D_2 are constants.

It is clear from the last two theorems that the Fourier convergence problem is generally dominant. In (1), Delves and Mead pursue an approach which leads not only to estimates of the Fourier convergence rate but which also may have considerable significance in the determination of an initial approximation from which equations (2.2) may be solved iteratively.

Theorem (Delves and Mead (1), p.212)

Let the matrix A be asymptotically diagonal of degree p , and, assuming $\alpha(i) \neq 0$ for any i , let

$$|\alpha(i)| \leq D_3\, i^{-q} \qquad q > 1 .$$

Then the coefficients $a'(i)$ defined by

$$\sum_{j=1}^{n} A_n(i,j)a'(j) = b(i) \qquad \ldots(2.20)$$

decrease asymptotically at the same rate as the Fourier coefficients $\alpha(i)$

provided $p > q+1$.

If p is sufficiently large, the coefficients

$$a_n'' \; : \; A_n(i,i)\; a''(i) \;=\; b(i) \qquad \ldots(2.21)$$

also have this property.

These convergence results are important in practical computational terms. We can contrast them with results on the convergence of the approximation $y_n(x)$ and its derivatives to those of the solution $y(x)$. We continue now to examine results of this type.

2.6 Convergence of the Rayleigh-Ritz approximation

The results of this section are primarily those of Ciarlet, Schultz and Varga, for basis functions which are trigonometric functions, or polynomials, both continuous and piecewise continuous. A discussion of the most general result of Rayleigh-Ritz convergence, i.e. Ciarlet, Schultz and Varga (1), Thm. 4, is deferred until we consider mildly non-linear differential problems in § 2.8.

We give first the appropriate theorems for global basis functions.

Polynomial functions of degree n satisfying the boundary conditions given in (2.4) are of the form

$$x^k(1-x)^k \left[a_o + a_1 x + \ldots + a_{n-2k}\, x^{n-2k} \right] ,$$

so that each polynomial is an element of a subspace P_o^n of H_L with dimension $n-2k+1$. For representations in this space of functions we quote the following theorem without proof.

Theorem (Ciarlet, Schultz, Varga (1), Thm.5, p.403)

If $y(x) \in C^t\,[0,1]$ and $y(x)$ satisfies $y^{(r)}(0) = y^{(r)}(1) = 0$ for $r = 0,1 \ldots k-1$, then there exists a sequence of polynomials of degree n , $\left\{ \tilde{p}_n(x) \right\}_{n=w}^{\infty}$ with $w = \max(t, 2k-1)$ such that $\tilde{p}_n(x) \in P_o^n$ and

$$\left\| \frac{d^r}{dx^r} (y - p_n) \right\|_\infty \leq \frac{K}{(n-k)^{t-k}} \omega \left(\frac{d^t y}{dx^t} , \frac{1}{n-k} \right)$$

$$0 \leq r \leq k$$

where $\omega(u, \delta)$ is the modulus of continuity of u defined by

$$\omega(u, \delta) = \max_{\substack{x, x_o \\ |x - x_o| < \delta}} |y(x) - y(x_o)|$$

As a consequence of this theorem we can establish the following.

Theorem (Ciarlet, Schultz, Varga (1), Thm.6, p.405)

Let $\psi(x)$ be the solution of the differential problem (2.3). Then $\psi(x) \in C^t [0,1]$, $t \geq 2k$. Let $\hat{p}_n(x)$ be the function in P_o^n which minimizes the corresponding energy functional (2.5), where $n \geq t$. Then there exists a constant M dependent on t and n such that

$$\left\| \hat{p}_n - \psi \right\|_\infty \leq \frac{M}{(n-k)^{t-k}} \left\| \frac{d^t \psi}{dx^t} \right\|_\infty \quad \ldots(2.22)$$

We remark that, since ψ is assumed to be a classical solution of (2.3) then $\psi \in D(L)$ and $\hat{p}_n \in D(L)$. Thus $\hat{p}_n$, the unique minimizing element of P_o^n for the energy functional $I(y)$ .. (2.5) satisfies

$$I(\hat{p}_n) \geq I(\psi) \geq I(\hat{y})$$

where $\hat{y}$ is the function which minimizes $I(y)$ over H_L.

It follows at once that the sequence of Rayleigh-Ritz approximations p_{k+v}, $v = 1,2 \ldots$ converges with order at least equal to k, i.e. if $\psi \in C^{2k} [0,1]$, $\psi \notin C^{2k+1} [0,1]$ then $t - k = k$.

These results are proved in Ciarlet, Schultz, Varga (1) for mildly non-linear problems of the form (1.8) which we shall consider in section 2.8, provided that certain assumptions are made concerning the coefficients $p_s(x)$, $s = 0 .. k$ and the right hand side $f(x,y)$. We shall detail

these requirements and their relevance to these theorems in section 2.8.

Weaker results than (2.21) and (2.22) concerning polynomial approximation and the Rayleigh-Ritz method are given by Mikhlin (4,p.129), (5). These results include, however, convergence results for approximation in more than one dimension as well as convergence in certain Sobolev norms, including derivatives, giving results comparable to results of a later section obtained by Ciarlet, Schultz, Varga (p.48).

Results similar to (2.21), (2.22) can be established when the approximation is a trigonometric polynomial if the solution $y(x)$ is known to be periodic or if periodic boundary conditions are prescribed. Such results are given by Ciarlet, Schultz and Varga (2) ; and Birkhoff and Fix (1).

Convergence theorems of the above types are developed by Ciarlet, Schultz, Varga (1), Shultz and Varga (1), Perrin, Price and Varga (1), for piecewise continuous basis functions such as (2.16) and (2.17). The strongest of these results are those developed by Schultz and Varga (1), using the generality of L-spline theory; however the following theorems are of more practical importance and indicate the types of result which can be derived for specific piecewise continuous basis functions in practical use.

Theorem (Ciarlet, Schultz, Varga (1), Thm.10, pp.409-10).

Let $\psi(x)$, the solution of (2.3), (2.4) be of class $C^t[0,1]$ with $t \geq 2m \geq 2k$, let $\overline{\Pi}$ be any partition of $[0,1]$, and let $q(x)$ be the unique function which minimizes $I(y)$ .. (2.5) over the space $H_o^m(\overline{\Pi})$ defined by (2.16). Then there exists a constant M, independent of $\overline{\Pi}$, such that

$$\left\| q(x) - \psi(x) \right\|_\infty \leq M \left\| \frac{d^{2m}}{dx^{2m}} \psi \right\|_\infty (\bar{h}(\overline{\Pi}))^{2m-k}$$

where $\bar{h}(\overline{\Pi}) = \max_i (x_i - x_{i-1})$, $x_i \in \overline{\Pi}$.

We note the comment in Ciarlet, Schultz, Varga (1) that the exponent of $\bar{h}(\overline{\Pi})$, $2m-k$, is in a sense the best possible.

For elements of the space $Sp_o^m(\overline{\Pi})$ defined by (2.17) we have

Theorem (Ciarlet, Schultz, Varga (1), Thm.16, p.416)

Let $\psi(x)$, the solution of (2.3),(2.4) be of class $C^t[0,1]$ with $t \geq 2m \geq 2k$, let $\{\overline{\Pi}_i\}_{i=1}^{\infty}$ be any sequence of partitions of $[0,1]$ with $\lim_i \bar{h}(\overline{\Pi}_i) = 0$, and let $r_i(x)$ be the unique function which minimizes (37) over $Sp_o^m(\overline{\Pi}_i)$. Then there exists a constant M , independent of the sequence $\overline{\Pi}_i$, such that

$$\left\| r_i(x) - \psi(x) \right\| \leq M \left\| \frac{d^{2m}}{dx^{2m}} \psi \right\|_{\infty} \cdot (\bar{h}(\overline{\Pi}_i))^{2m-1-k}$$

Unlike the result above it appears that the exponent $2m-1-k$ could be improved to $2m-k$. See Ciarlet, Schultz, Varga (1, p.417).

2.7 Convergence of the Residual

We have seen that the function $y(x)$ which minimizes $I(y) = (Ly,y) - 2(y,f)$ does not necessarily lie in the domain of existence $D(L)$, and it is therefore not possible for such functions that the approximations $y_n(x)$ satisfy

$$\lim_{n \to \infty} Ly_n(x) = f$$

under general assumptions. Under what conditions this convergence can be established is a problem which has been considered by Mikhlin, Vainikko and others; see Mikhlin (4, p.108), and Vainikko (1). The following theorem, from Mikhlin, characterizes the situation for the Rayleigh-Ritz method; it is a particular case of the theorem established by Vainikko for general projection methods (including least squares, Galerkin, and the method of moments).

Theorem (Mikhlin (4,p.109))

Let L and M be positive definite operators defined on a separable Hilbert space H , satisfying D(L) = D(M). Assume that the eigen values of M are discrete, and let the corresponding eigenfunctions be the sequence $\left\{ \phi_i(x) \right\}_{i=1}^{\infty}$. If this sequence of functions are used as co-ordinate functions for the Rayleigh-Ritz method, and

$$y_n(x) = \sum_{i=1}^{n} a_n(i)\, \phi_i(x)$$

is the resulting approximation, where $\underline{a}_n$ satisfies (2.2) then

$$\| Ly_n - f \| \to 0 \qquad \text{as} \qquad n \to \infty$$

The operators L and M are said to be similar: see § 3.1.

For the differential equation

$$L(y) = \frac{d}{dx} \left(p_1(x) \frac{dy}{dx} \right) + p_o(x)y = f \qquad \ldots(2.23)$$

where $p_1(x) > 0$, $p_o(x) \geq 0$, with the boundary conditions $y(0) = 0$ $y(1) = 0$, a similar operator is

$$M(y) = \frac{d^2y}{dx^2}$$

with the same boundary conditions. Thus if the basis

$$\left\{ \phi_i \; : \; \phi_i(x) = C_i \sin i \pi x \, , \quad i = 1 \ldots \infty \right\} \qquad \ldots(2.24)$$

is used for problem (2.23) then the residual will converge to zero. The role of the normalization factor C_i with respect to practical computation will be considered in Chapter Three.

An alternative study of the convergence of the residual for projection methods is that given by Kantorovich (1), which is somewhat less restrictive than the results of Mikhlin, given above.

Let $L : D(L) \to R(L)$, and let $D(L) = H_1$, $R(L) = H_2$. Let $H_1^{(n)}$ be a finite dimensional subspace of H_1 for each n . In the

Rayleigh Ritz method this will be the subspace spanned by $\{\phi_i\}_{i=1}^{n}$. Assume that for arbitrary $v \in H_1$

$$E_n(v) = \inf_{v_n \in H_1^{(n)}} \| v - v_n \| \to 0 \quad \text{as} \quad n \to \infty \qquad \ldots(2.25)$$

Let $\lambda_1^{(n)}$, $\lambda_n^{(n)}$ be the smallest and largest eigenvalues of the Gramm matrix of $\{\phi_i\}_{i=1}^{n}$. Then it can be shown that

$$\| Ay_n - f \| \to 0 \quad \text{as} \quad n \to \infty$$

if

$$\left[\frac{\lambda_n^{(n)}}{\lambda_1^{(n)}} \right]^{\frac{1}{2}} E_n(y) \to 0 \quad \text{as} \quad n \to \infty \qquad \ldots(2.26)$$

The practical limitation of this result is that the problem of determining a suitable subspace for which (2.25) holds; the result of Mikhlin above provides such a subspace in terms of the eigenfunctions of a similar operator. On the other hand, these eigenfunctions may not permit rapid convergence of the Rayleigh-Ritz solution. For example, the basis (2.24) will not provide rapidly converging approximations to the solution of (2.23) unless this solution is periodic, and a polynomial basis would be preferable in the general case. We shall examine the usefulness of these respective criteria in Chapters Three and Four, with reference to specific possible co-ordinate systems.

2.8 Mildly non-linear differential problems

We consider now differential equations of the form

$$\sum_{s=0}^{i} (-1)^{s+1} \frac{d^s}{dx^s} \left(p_s(x) \frac{d^s y}{dx^s} \right) = f(x,y) \qquad \ldots(2.27)$$

subject to

$$y(0) = y(1) = 0 \qquad \ldots(2.28)$$

for which the functional of the energy problem assumes the form

$$I(y) = \int_0^1 \left\{ \frac{1}{2} \sum_{s=0}^{k} p_s(x) \left(\frac{d^s y}{dx^s} \right)^2 + \int_0^{y(x)} f(x,\eta)\,d\eta \right\} dx \qquad \ldots(2.29)$$

The differential problem (2.27) (2.28) will be termed a mildly non-linear problem, since it is linear in derivatives of y but non-linear in y. The problems (2.3) and (2.5) are easily seen to be special cases of (2.27) and (2.29). In order that the problems (2.27) and (2.29) have unique solutions it is necessary that we make explicit a number of assumptions which have been implicit or unnecessary in our treatment of linear differential problems. In doing this we are again following the work of Ciarlet, Schultz and Varga (1), and their collaborators. We shall not comment further on two extensions of this work, in which Gladwell (1) considers differential equations of the form

$$\sum_{s=0}^{k} (-1)^s \frac{d^s}{dx^s} \left(p_s(x) \frac{d^s y}{dx^s} \right) = \sum_{i=0}^{k} f_i\left(x, y^{(i)}\right)^{(i)}$$

where

$f_i\left(x, y^{(i)}\right)^{(i)}$ is the ith derivative of a function of x and $y^{(i)}$, but is independent of any other derivative of y, or y itself, subject to linear homogeneous essential boundary conditions, and non-linear natural boundary conditions, and Ciarlet, Schultz and Varga (5) treat a wider class of differential equations by the Galerkin method, using monotone operator theory.

Under the following assumptions it can be shown that there is a unique solution of the energy problem:- Minimize (2.21) subject to (2.22). These are given by Ciarlet, Schultz and Varga (1, § 8). To introduce

them we must define

$$\Lambda = \inf_{\phi \in S} \frac{\int_0^1 \sum_{s=0}^{k} p_s(x) \left(\frac{d^s}{dx^s} \phi(x) \right)^2}{\int_0^1 (\phi(x))^2 \, dx}$$

Under the assumptions $p_j(x) \in C^j [0,1]$ and that there exist constants β and K such that

$$\sup_{x \in [0,1]} |\phi(x)| \leq K \left[\int_0^1 \left\{ \sum_{s=0}^{k} p_s(x) \left(\frac{d^s}{dx^s} \phi(x) \right)^2 + \beta (\phi(x))^2 \right\} dx \right]^{1/2} \quad \ldots(2.30)$$

then $\Lambda > -\infty$ (C,S,V (1, p.396, Lemma 1)).

We require the assumption on $f(x,y)$ that there exists a constant γ such that

$$\frac{f(x,u) - f(x,v)}{u-v} \geq \gamma > -\Lambda \quad \text{for all } x \in [0,1] \quad \ldots(2.31)$$

and $-\infty < u, v < \infty$, $u \neq v$.

This condition is implied by

$$\frac{\partial f(x,y)}{\partial y} \geq \gamma > -\Lambda \quad \ldots(2.32)$$

The condition (2.31) can be replaced in certain circumstances, (see C.S.V. (1), p.419, 8.3), but only by requiring certain other, and very specific properties of the problem, for example that the solution y of (2.27), (2.28) is positive in (0,1). Similarly, the coefficients $p_s(x)$, $s=0 \ldots k$, may be assumed only piecewise continuous; again restrictions are implied on the solution. With the conditions (2.30), (2.31) and provided $p_j(x) \in C^j [0,1]$, Ciarlet, Schultz and Varga show

that the functional $I(y)$ ... (2.29) has a unique minimum and that this is achieved for the function $y = \psi(x)$ where $\psi(x)$ is a classical solution of (2.27), (2.28).

The role of the assumptions (2.30) and (2.31) or (2.32) is not immediately clear. Their importance lies in the proof of

$$I(\psi) \leq I(y) \qquad \text{for all } y \neq \psi$$

Letting $y(x) = \psi(x) + \varepsilon(x)$ we obtain by integrating by parts, and using the boundary conditions

$$\varepsilon^{j}(0) = \varepsilon^{j}(1) = 0 , \qquad j = 0 \,..\, k-1 ,$$

$$I(y) = I(\psi) + 1/2 \int_0^1 \sum_{s=0}^{k} p_s(x) \frac{d^s}{dx^s} (\varepsilon(x))^2 \, dx + \int_0^1 \left\{ \int_{\psi}^{\psi+\varepsilon} f(x,\eta) - f(x,\bar{y}) \, d\eta \right\} dx$$

Then using (2.31) or (2.32) we have

$$\int_{\psi}^{\psi+\varepsilon} \left(f(x,\eta) - f(x,\psi) \right) d\eta \cdots \geq \frac{\gamma}{2} \varepsilon^2(x)$$

and using (2.30)

$$\int_0^1 \sum_{s=0}^{k} p_s(x) \left(\frac{d^s}{dx^s} (\varepsilon(x)) \right)^2 dx > \Lambda \int_0^1 \varepsilon^2(x)$$

so that

$$I(y) \geq I(\psi) + \frac{\Lambda + \gamma}{2} \int_0^1 \varepsilon^2(x) \, dx$$

By definition, $\gamma > -\Lambda$ and hence $I(\psi) < I(y)$ for all y for which $I(y)$ is defined and which satisfy the boundary conditions (2.28).

Under conditions (2.30), (2.31), (2.32) all the theorems concerning the convergence of the Rayleigh-Ritz approximation to the solution $\Psi(x)$ hold. In order to ensure convergence to derivatives of $\Psi(x)$ it is necessary to replace (2.30) with the stronger assumption

$$\left\| \frac{d^l}{dx^l} \phi \right\|_\infty \leq K \left[\int_0^1 \left\{ \sum_{s=0}^{k} p_s(x) \left(\frac{d^s \phi}{dx^s} \right)^2 + \beta (\phi(x))^2 \right\} dx \right]^{\frac{1}{2}}$$

or

$$\left\| \frac{d^{l+1} \phi}{dx^{l+1}} \right\|_2 \leq K \left[\int_0^1 \left\{ \sum_{s=0}^{k} p_s(x) \left(\frac{d^s \phi}{dx^s} \right)^2 + \beta (\phi(x))^2 \right\} dx \right]^{\frac{1}{2}}$$

(see Ciarlet, Schultz, Varga (1, p.395)).

We have now outlined the theory of convergence of Rayleigh-Ritz approximation for variational problems related to certain linear and mildly non-linear problems, when the approximating functions are chosen in a number of ways. Throughout, we have assumed that coefficients in the approximation are evaluated analytically, and it remains to be seen whether certain choices of basis function are more expedient when these coefficients are evaluated numerically. We look at this problem in Chapter Three.

Chapter Three

Numerical Considerations in the Application of the Rayleigh-Ritz method to Linear Differential Equations

Although the convergence results of Chapter Two, with the exception of those concerning asymptotically diagonal matrices, depend only on the choice of the sequence of subspaces in which to obtain an approximate solution, and not on the selection of co-ordinate functions forming a basis for this subspace, the correct choice of the co-ordinate functions is essential in order to perform the Rayleigh-Ritz method numerically. In this chapter we shall use examples to indicate the properties which it is desirable that a co-ordinate system should possess. Criteria of this kind are given in many of the works of Mikhlin, particularly (1), (4). Many of these criteria are relevant to other function expansion methods, such as least squares (Andersenn (1), Dshishkariani (1)), the method of moments, and Galerkin's method.

Other approaches to the problems arising when the Rayleigh-Ritz method is applied numerically have been given in the literature. Delves (1) describes a method in which the effects of rounding errors in the numerical processes are minimized, whilst Babuska, Prager and Vitasek (1) report the results of a numerical study in which errors of a known magnitude are introduced into the Rayleigh-Ritz equations and the effect on the solution analysed.

3.1 Computational Results for Simple Quadratic Problems.

The variational problems considered in this section are of the form

$$\min \quad I(y) = \int_0^1 \left\{ p_1(x)(y'(x))^2 + p_0(x)(y(x))^2 + 2f(x)y(x) \right\} dx \quad \ldots(3.1)$$

subject to $y(0) = y(1) = 0$.

For the sequence of co-ordinate functions $\{\phi_i\}_{i=1}^n$ this leads to the system of Rayleigh-Ritz equations (2.2)

$$A_n \underline{a}_n = \underline{b}_n \quad \ldots(3.2)$$

and the nth Rayleigh-Ritz approximate solution

$$y_n(x) = \sum_{i=1}^{n} a_n(i)\, \phi_i(x) \quad \ldots(3.3)$$

When the Rayleigh-Ritz method is applied numerically, the solution of equations (3.2) is affected by errors introduced at two stages. First, numerical representation of the elements of the matrix A_n and vector $\underline{b}_n$ introduces errors into the values of these elements, and second, numerical methods of solution of the equations, for example Gaussian elimination or triangular decomposition, introduce further errors. If $\mathcal{E}_n$ denotes the matrix of errors in the elements of A_n , and $\underline{\delta}_n$ the vector of errors in $\underline{f}_n$, then we can define $\underline{a}'_n$ and $\bar{y}(x)$ in the following manner.

Let $\underline{a}'_n$ denote the algebraic solution of the equations

$$(A_n + \mathcal{E}_n)\, \underline{a}'_n = \underline{b}_n + \underline{\delta}_n \quad \ldots(3.4)$$

and define

$$\bar{y}_n(x) = \sum_{i=1}^{n} \underline{a}'_n(i)\, \phi_i(x)$$

We also define $\underline{a}''_n$, $\bar{\bar{y}}_n(x)$ by the following. Let $\underline{a}''_n$ denote the vector solution obtained when equations (3.4) are solved numerically; then

$$(A_n + \mathcal{E}_n)(\underline{a}''_n + \delta \underline{a}_n) = \underline{b}_n + \underline{\delta}_n \quad \ldots(3.5)$$

where $$\delta \underline{a}_n = \underline{a}'_n - \underline{a}''_n$$

and $$\bar{\bar{y}}_n(x) = \sum_{i=1}^{n} \underline{a}''_n \ (i)\ \phi_i(x) \qquad \ldots(3.6)$$

We remark that a realistic numerical solution is of the form (3.6).

The numerical experiments which follow were performed on an I.B.M. 360/67 using either single word length or double word length precision. These represent either 7 or 15 significant figures. The scalar products of A_n and $\underline{b}_n$ were computed using a 20 point Gaussian quadrature formula, so that for arbitrary $\{\phi_i\}_{i=1}^{n}$, $p_o(x)$, $p_1(x)$, $f(x)$ the errors $\underline{\varepsilon}_n$ and $\underline{\delta}_n$ represent quadrature errors and rounding errors; this presents no complication. For certain examples, where $\{\phi_i\}_{i=1}^{n}$, $p_o(x)$, $p_1(x)$, $f(x)$ are polynomials, the quadrature formulae are exact for some subsystem $\{\phi_i\}_{i=1}^{m}$ of the co-ordinate system, and the errors in these cases are exclusively due to rounding error.

The equations were solved using the method of Gaussian Elimination with row interchanges, for $n = 2,3 \ldots 20$. This does not take advantage of the symmetry of the equations, nor does it readily admit the addition of the single row and column which are introduced in the progression from a Rayleigh-Ritz system of size n to a system of size $n+1$. Both these advantages are obtained if the equations (3.2) are solved using triangular decomposition, and particularly the Cholesky decomposition. More efficient, but not necessarily more accurate, computation could be achieved in this way.

The approximation $\bar{\bar{y}}_n(x)$ was computed in a straightforward manner from (3.6). No attempts to transform these series into others prior to evaluation of the solution are considered. As an estimate of the L_∞ error norm of the error in $\bar{\bar{y}}(x)$ we take the measure

$$\bar{e}_n = \max_{x=x_\alpha} \left| \bar{\bar{y}}(x) - y(x) \right| \qquad \ldots(3.7)$$

where $x_\alpha = \alpha h$, $\alpha = 0 \ldots k$, $h = 1/k$ for some k , usually $k = 20$.

As simple test examples we consider the following:-

Problem L1.

$$\min \quad I(y) = \int_0^1 \left\{ y'^2 - y^2 - 2xy \right\} dx$$

subject to $y(0) = y(1) = 0$, ...(3.8)

which has the solution

$$y(x) = \sin(x) / \sin(1) - x \qquad ...(3.9)$$

The corresponding differential equation is

$$y'' + y + x = 0$$

with the boundary conditions (3.8).

This example is also considered by Kantorovich and Krylov (1, p.269).

Problem L2.

$$\min \quad I(y) = \int_0^1 \left\{ y'^2 + 4y^2 + 8 \cosh(1)\, y \right\} dx$$

subject to $y(0) = y(1) = 0$ ...(3.10)

with the solution

$$y(x) = \cosh(2x-1) - \cosh(1)$$

The corresponding differential equation is

$$y'' = 4y + 4 \cosh(1)$$

subject to (3.10). This example is considered by Ciarlet, Schultz and Varga (1, p.426).

Problem L3.

$$\min \quad I(y) = \int_0^1 \left\{ (x+1)y'^2 - \frac{x(x+2)}{(x+1)}\, y^2 - 2(x+1)^2 y \right\} dx$$

subject to $y(0) = y(1) = 0$ ...(3.11)

with solution

$$y(x) = 3.6072\, J_1(x) + 0.75195\, Y_1(x) - 1 - x$$

This example is obtained from the Bessel differential equation

$$t^2z'' + tz' + (t^2-1)z = 0$$

with the boundary conditions

$$z(1) = 1 , \quad z(2) = 2$$

by the transformations

$$z = y+t , \quad t = x+1$$

and division by $x+1$, giving the differential equation

$$(x+1)y'' + \frac{x(x+2)}{x+1} y + (x+1)^2 = 0$$

with boundary conditions (3.11). This example, in which the coefficients $p_0(x)$, $p_1(x)$ are not so simple as in problems L1 and L2, is considered by Kantorovich and Krylov (1,p.270). Evaluation of $y(x)$ for this problem utilizes the values for the Bessel functions $J_1(x)$ and $Y_1(x)$ from the tables of Abramowitz and Stegun (1).

We consider first the solution of these problems in terms of the basis functions

$$\left\{ S_i^1 \right\}_{i=1}^{\infty} = \frac{\sqrt{2}}{i\pi} \sin (i\pi x) \qquad \ldots(3.12)$$

and

$$\left\{ P_i^1 \right\}_{i=1}^{\infty} = x^i (1-x) \qquad \ldots(3.13)$$

These functions are used extensively in examples of the application of the Rayleigh-Ritz method in the literature; see e.g. Mikhlin (4), Mikhlin and Smolitskiy (1), Kantorovich and Krylov (1). In Tables I and II we give $\bar{e}_n$ and the coefficients $a_n''(i)$ for problem L1 and the bases (3.12) and (3.13) respectively, for certain adjacent pairs of values of n. Similar tables for problems L2 and L3 may be found in Appendix A. In addition, in Table I, the exact values of the Rayleigh-Ritz coefficients are also given; this is straightforward since the

matrix A_n is diagonal for problem L1 and the basis (3.12). Also given is the estimate of the maximum error in the Rayleigh-Ritz approximation $y_n(x)$, given by

$$e_n = \max_{x=x_\alpha} \left| y_n(x) - y(x) \right|$$

where the points x_α are those used in the application of (3.7).

More precisely, we have

$$A_n(i,j) = 2 \int_0^1 \cos i\pi x \cos j\pi x - \frac{1}{ij\pi^2} \sin i\pi x \cdot \sin j\pi x \quad dx$$

$$= \begin{cases} 1 - \dfrac{1}{i^2\pi^2} & , \quad i = j \\ 0 & , \quad i \neq j \end{cases}$$

and

$$b_n(i) = \frac{\sqrt{2}}{i\pi} \int_0^1 x \sin i\pi x \quad dx = \frac{(-1)^{i-1} \cdot \sqrt{2}}{i^2\pi^2}$$

so that

$$a_n(i) = \frac{(-1)^{i-1} \cdot \sqrt{2}}{i^2\pi^2 - 1}$$

and

$$y_n(x) = 2 \sum_{i=1}^{n} (-1)^{i-1} \sin i\pi x \; / \; ((i^2\pi^2 - 1)\, i\pi)$$

A number of points of note emerge from Tables I and II. In Table I the error $\bar{e}_n$ decreases smoothly as n increases. Further, the values of the coefficients $a_n''(i)$ do not vary greatly with n, and we note also that the coefficients $a_n''(i)$ become smaller as i increases, for fixed n. This apparent convergence, and other features displayed in this Table are those generally required of a reliable numerical method.

The situation in Table II is not so encouraging. Here we have a

Behaviour of the Solution Vector and the Approximate Solution.

Problem L1 $\qquad \phi_i = \frac{\sqrt{2}}{i\pi} \sin i\pi x$

n	2	3	7	8	11	12	Exact
$\bar{e}_n$	3.859650'-3	1.904899'-3	4.048160'-4	2.847872'-4	1.020543'-4	7.764621'-5	
e_n	3.859645'-3	1.904899'-3	4.048682'-4	2.848654'-4	1.016520'-4	7.969140'-5	
e_n^*	3.859616'-3	1.904865'-3	4.048573'-4	2.848499'-4	1.020415'-4	7.763326'-5	
	+1.594447'-1	+1.594447'-1	+1.594447'-1	+1.594447'-1	+1.594447'-1	+1.594447'-1	+1.594448'-1
	-3.675349'-2	-3.675350'-2	-3.675352'-2	-3.675352'-2	-3.675352'-2	-3.675352'-2	-3.675340'-2
		+1.610239'-2	+1.610238'-2	+1.610239'-2	+1.610239'-2	+1.610239'-2	+1.610235'-2
			-9.012881'-3	-9.012881'-3	-9.012881'-3	-9.012881'-3	-9.012680'-3
$a_n''(i)$			+5.755066'-3	+5.755010'-3	+5.755014'-3	+5.755014'-3	+5.754910'-3
			-3.991961'-3	-3.991965'-3	-3.991968'-3	-3.991834'-3	-3.991506'-3
			+2.930339'-3	+2.930347'-3	+2.930221'-3	+2.930224'-3	+2.930339'-3
				-2.242829'-3	-2.242662'-3	-2.245205'-3	-2.242453'-3
					+1.774150'-3	+1.774147'-3	+1.771226'-3
					-1.437535'-3	-1.412066'-3	-1.434351'-3
					+1.162711'-3	+1.162711'-3	+1.185206'-3
						-1.110446'-3	-9.957681'-4

Table I.

Behaviour of the Solution Vector and the Approximate Solution.

Problem L1 $\phi_i = x^i(1-x)$

n	2	3	7	8	11	12
$\bar{e}_n$	3.026127'-4	2.489983'-5	1.219660'-5	9.596347'-6	2.670288'-5	6.247309'-6
e^*_n	3.025191'-4	2.507028'-5	5.354866'-11	2.764134'-11	2.652861'-11	2.681174'-11
a''_n (i)	+1.924085'-1	+1.877603'-1	+1.876336'-1	+1.885666'-1	+1.892988'-1	+1.881749'-1
	+1.707373'-1	+1.941637'-1	+2.075043'-1	+1.789945'-1	+1.682944'-1	+1.954954'-1
		-2.232645'-2	-1.615116'-1	+1.174669'-1	+9.820651'-2	-8.644186'-2
			+5.276064'-1	-7.274599'-1	-5.092235'-2	+3.418415'-1
			-9.399538'-1	+2.021488'+0	-8.568961'-1	-5.466980'-1
			+7.964375'-1	+2.992083'+0	+1.178898'+0	-9.663411'-1
			-2.603114'-1	+2.228110'+0	+2.858857'+0	+4.873112'+0
				-6.570082'-1	-7.833037'+0	-6.597326'+0
					+5.406525'+0	+2.171780'+0
					+1.644933'-1	+2.831546'+0
					-9.677417'-1	-2.700788'+0
						+6.530638'-1

Table II

polynomial basis and constant or polynomial coefficients, so that the errors in A_n and $\underline{b}_n$ are caused only by rounding error, and not by the generally more severe effects of quadrature approximation. However, the coefficients $a_n''(i)$ do not tend to a limit as either n or i increase, nor does the error reduce smoothly. This basis has generated a numerical approximation with few desirable properties. Only one feature of this Table provides any redemption for this method; all of the results obtained using the polynomial basis functions (3.13) for $n = 1,2,3...20$ are more accurate than any of the results obtained using the trigonometric basis (3.12).

It can be seen from the corresponding double precision results for this problem given as e_n^* in Tables I and II that the above remarks hold equally in that case.

There is an immediate explanation for the greater accuracy obtained with the basis (3.13) rather than that obtained using (3.12). The solution (3.9) of problem L1 is not periodic, and this is reflected in a slow rate of convergence when the function is approximated by periodic functions.

We must recognise the two important problems illustrated by the results of Tables I and II. The fundamental aim of our application of the Rayleigh-Ritz method is to produce an approximation to the solution of the variational problem; an approximating function which agrees closely with the true solution. It is necessary, however, to recognise a good approximate solution in some way, and this is generally achieved by considering successive vectors of coefficients; for example, if

$$\left| a_n''(i) - a_{n+1}''(i) \right| < \varepsilon_1 \qquad 1 \le i \le n \qquad \ldots(3.14)$$

and

$$\left| a_{n+1}''(n+1) \right| < \varepsilon_2 \qquad \ldots(3.15)$$

where ε_1 , ε_2 are small, then $\underline{a}''_{n+1}$ may be considered to be a solution. It is, of course, possible to evaluate successive approximating functions and compare these, though this represents a considerable increase in computation. The two problems therefore are

1. The determination of an approximating subspace in which the solution of a variational problem may be conveniently and accurately represented by a small number of basis functions.

2. The determination of a suitable basis for the subspace, for which conditions (3.14), (3.15) may be expected to imply the accuracy of the approximating function, and vice-versa.

The first of these problems is a convergence problem in approximation theory whose solution requires an a-priori understanding of the nature of the solution of the variational problem. Some results of this type are given in Chapter Two. A paper of Delves and Mead (3, to appear) will consider the a-priori determination of properties of the solution.

The second problem is referred to as a stability problem. The irregular behaviour of the coefficients in Table II is caused by the numerical errors involved in the solution of (3.5). The effects of such errors in the determination of the coefficients of a Rayleigh-Ritz approximation have been considered by Mikhlin (1),(4), Samokish (1), Andersenn (1), and others; we examine this work in § 3.3, after first introducing certain properties of arbitrary systems of functions.

3.2 A Minimal-Orthonormal Classification of Functions.

Let H be an arbitrary Hilbert space, and let $\{\phi_i : \phi_i \in H\}$ be a finite or infinite sequence of functions in H. The sequence $\{\phi_i\}$ is said to be minimal in H if and only if eliminating one element of the sequence reduces the dimension of the subspace spanned by the system, i.e

Let $\{\phi_i\} = \phi_1, \phi_2, \phi_3, \ldots$ be a sequence in H and let H_k be the subspace of H spanned by $\phi_1, \phi_2 \ldots \phi_{k-1}, \phi_{k+1} \ldots$. The sequence $\{\phi_i\}$ is minimal in H iff $\phi_k \notin H_k \ \forall k$. As a consequence of this definition we have immediately:-

The sequence of functions $\{\phi_i\}$ is non-minimal in H iff there exists an integer j such that, given $\varepsilon > 0$, there exists an integer N and scalars $\alpha_1 \ldots \alpha_{j-1}, \alpha_{j+1} \ldots \alpha_N$ such that

$$\left\| \phi_j - \sum_{\substack{k=1 \\ k \neq j}}^{N} \alpha_k \phi_k \right\|_H \leq \varepsilon \qquad \ldots(3.16)$$

where $\| \cdot \|_H$ denotes the norm in H.

Thus any finite sequence of linearly independent functions in H is minimal in H; any finite system of linearly dependent functions is non-minimal in H. In particular, in this case we may take $\varepsilon = 0$ in (3.16). Any infinite sequence of functions orthonormal in H is minimal in H. An often quoted result (see Mikhlin (4), Kaczmarz and Steinhous (1)) is the following:- The sequence $\{x^n\}_{n=0}^{\infty}$ is non-minimal in $\mathcal{L}_2[0,1]$.

The sequence $\{\phi_i\}$ is said to be strongly minimal in H iff the smallest eigenvalue $\lambda_1^{(n)}$ of the Gramm matrix A_n

$$A_n(i,j) = (\phi_i, \phi_j)_H \qquad i,j = 1 \ldots n$$

of the sequence $\{\phi_i\}_{i=1}^{n}$ is bounded below by a strictly positive

constant independent of n ; i.e. there exists an $w \neq 0$ s.t

$$\lambda_1^{(n)} \geq w^2 \qquad \forall \ n \ .$$

The sequence $\{\phi_i\}$ is said to be <u>almost orthonormal</u> in H iff it is strongly minimal in H and the largest eigenvalue $\lambda_n^{(n)}$ of the Gramm matrix A_n is bounded above by a positive constant independent of n , i.e. there exist real numbers $w \neq 0$, $W \neq 0$ satisfying $|w| \leq |W|$ such that

$$w^2 \leq \lambda_1^{(n)} \leq \lambda_n^{(n)} \leq W^2 \qquad \ldots(3.17)$$

The sequence $\{\phi_i\}$ is <u>orthonormal</u> in H iff $w = W = 1$ in (3.17).

The following theorem (Andersenn (1), p.134 Thm.3.3) summarizes the relationships between these classes of functions, and condenses several results of Mikhlin (1), (4).

"Orthonormal systems are almost orthonormal, almost orthonormal systems are strongly minimal, and strongly minimal systems are minimal. At no stage is the converse (necessarily) true".

We have also the following result (Mikhlin, (4), p.6). Let the sequence $\{\phi_i\}$ be minimal in H . Then there exist scalars α_1 , $\alpha_2 \ldots \alpha_n$, ... such that the sequence $\{\alpha_i \phi_i\}$ is strongly minimal in H .

In order to pursue Mikhlin's study of the selection of co-ordinate systems the following theorems are useful. We remark that, when a sequence of co-ordinate functions $\{\phi_i\}$ are chosen for the solution of the equation $Ly = f$, (where L is a positive definite operator) by the Rayleigh-Ritz method, the resulting matrix A_n (3.2) is exactly the Gramm matrix of the subsequence $\{\phi_i\}_{i=1}^{n}$ in the energy space H_L of the operator L .

Theorem (Mikhlin (4), p.7-8)

Let M and N be positive definite, self-adjoint operators in some Hilbert space H, satisfying $H_M \subseteq H_N$. If $\{\phi_i : \phi_i \in H_M, i=1,2\ldots\}$ is a minimal (strongly minimal) sequence in H_N then it forms a minimal (strongly minimal) sequence in H_M.

In particular, if $N = I$, then $H_N = H$, and since M is positive definite, we have $H_M \subseteq H_N = H$. The result above implies that any minimal (strongly minimal) sequence in H is minimal (strongly minimal) in H_M.

Definition.

Two positive definite self-adjoint operators M and N are <u>semi-similar</u> iff $H_M = H_N$. M and N are <u>similar</u> iff $D(M) = D(N)$. Similar operators are necessarily semi-similar.

Theorem (Mikhlin,(4), p.11)

Let $\{\phi_i\}$ be a complete, almost orthonormal sequence in H_N, and let M and N be semi-similar. Then $\{\phi_i\}$ form a complete, almost orthonormal sequence in H_M.

Theorem (Mikhlin & Smolitskiy (1))

Let M and N be similar, and assume $|(My, Ny)| \leq m \|My\|^2$ $\forall y \in H_M$, $m = \text{const} > 0$. Let $\{\phi_i\}$ be the orthonormalized eigenfunctions of N, assumed complete in H_N. Then $\{\phi_k\}$ are a complete, almost orthonormal sequence in H_M.

3.3 <u>Stability</u>

Loosely speaking, Rayleigh-Ritz computations are stable in some sense if the differences $\|y_n(x) - \bar{y}_n(x)\|$ and $\|\underline{a}_n - \underline{a}'_n\|$ or $\|y_n(x) - \bar{\bar{y}}_n(x)\|$ and $\|\underline{a}_n - \underline{a}''_n\|$ are small whenever $\|\epsilon_n\|$, $\|\delta_n\|$ in (3.4), (3.5) are small. More precisely, we indicate

the four cases separately, following the definitions of Mikhlin.

The vector solution of the Rayleigh-Ritz equations (3.2) is stable with respect to errors in the equations, of the form (3.4), iff there exist constants p,q,r, independent of n, such that for all $\|\mathcal{E}_n\| \leq r$ and all $\underline{\delta}_n$ the equations (3.4) are solvable and

$$\|\underline{a}'_n - \underline{a}_n\| \leq p\,\|\mathcal{E}_n\| + q\,\|\underline{\delta}_n\| \quad \ldots(3.18)$$

The Rayleigh-Ritz approximation $y_n(x)$ defined by (3.3) is stable with respect to errors in equations (3.2) of the form (3.4) iff there exist constants p_1, q_1, r_1, independent of n, such that for all $\|\mathcal{E}_n\| \leq r_1$, and all $\underline{\delta}_n$

$$\|\bar{y}_n(x) - y_n(x)\| \leq p_1\,\|\mathcal{E}_n\| + q_1\,\|\underline{\delta}_n\| \ldots(3.19)$$

The inevitable magnification of the initial errors $\mathcal{E}_n$, $\underline{\delta}_n$ by numerical methods of solution is not considered in these definitions. We therefore also introduce the following.

The vector solution of the Rayleigh-Ritz equations (3.2) is stable with respect to errors in the equations of the form (3.5) and errors introduced by the method of solution iff there exist constants P, Q, R such that for all $\|\mathcal{E}_n\| \leq R$ and all $\underline{\delta}_n$

$$\|\underline{a}_n - \underline{a}''_n\| \leq P\,\|\mathcal{E}_n\| + Q\,\|\underline{\delta}_n\| \quad \ldots(3.20)$$

The Rayleigh Ritz approximation $y_n(x)$ is stable with respect to errors in the equations (3.2) of the form (3.5), and errors introduced by the method of solution iff there exist constants P_1, Q_1, R_1 such that, for all $\|\mathcal{E}_n\| \leq R_1$ and all $\underline{\delta}_n$

$$\|y_n(x) - y''_n(x)\| \leq P_1\,\|\mathcal{E}_n\| + Q_1\,\|\underline{\delta}_n\| \ldots(3.21)$$

For brevity we introduce the following terminology. We speak of (3.18) (resp (3.19) as the <u>numerical stability</u> of the Rayleigh-Ritz

solution vector (resp. the Rayleigh-Ritz approximation) and of (3.20) (resp.(3.21)) as the complete numerical stability of the Rayleigh-Ritz solution vector (resp. the Rayleigh-Ritz approximation).

As has already been indicated, the requirement that the Rayleigh-Ritz method be numerically stable in any of the senses (3.18) ... (3.21) imposes further restrictions on the choice of basis functions. Mikhlin's criteria for the selection of suitable basis functions rest on the following theorems, which summarize results of Mikhlin (4), Mikhlin and Smolitskiy (1).

Theorem

In order that the Rayleigh-Ritz vector solution and the Rayleigh-Ritz approximation be numerically stable (in the sense of (3.18), (3.19)) for the solution of the equation $Ly = f$, it is necessary and sufficient that the co-ordinate system $\{\phi_i\}$ be strongly minimal in the energy space H_L.

Theorem

In order that the Rayleigh-Ritz vector solution and the Rayleigh-Ritz approximation be completely numerically stable (in the sense of (3.20), (3.21)) it is necessary and sufficient that the co-ordinate system $\{\phi_i\}$ be almost orthonormal in the energy space H_L.

We remark here that the requirement that the co-ordinate system be almost orthonormal in the energy space H_L in order to achieve complete numerical stability is exactly equivalent to the condition that the Rayleigh-Ritz matrix have bounded condition number. The most appropriate condition number $K(A_n)$ of the symmetric matrix A_n is defined by

$$K(A_n) = \frac{\lambda_n^{(n)}}{\lambda_1^{(n)}} \qquad \ldots(3.22)$$

and the above remark is an easy consequence of the definition of almost orthonormal systems and the equivalence of the Rayleigh-Ritz matrix A_n and the Gramm matrix in H_L .

An alternative view of stability problems for a class of methods, the "abstract Galerkin processes", which includes the Rayleigh-Ritz method as a special case, has been presented by Samokish (1). We rely on the exposition of Mikhlin (4, pp.76-81) for details of this work. Although Samokish is regarding a stable process as one in which small errors in the data and method of solution produce small errors in the results, there is no formal definition of this view. Instead, two measures of stability are introduced. The first, for the Rayleigh-Ritz method, reduces to $K(A_n)$, defined by (3.22). By considering the equations

$$A_n \underline{\alpha}_n = \underline{\beta}_n \qquad \ldots(3.23)$$

where $\beta_n(i)$ are arbitrary numbers satisfying

$$\sum_{i=1}^{n} \left| \beta_n(i) \right|^2 = 1 \qquad \ldots(3.24)$$

and defining

$$z_n(x) = \sum_{i=1}^{n} \alpha_n(i)\, \phi_i(x)$$

the quantity

$$\mu_n = \frac{\max \left\| z_n(x) \right\|}{\min \left\| z_n(x) \right\|} \qquad \ldots(3.25)$$

is introduced, where the operators 'max' and 'min' act over all vectors satisfying (3.24). With these measures, Samokish utilizes the following assumptions. The Rayleigh-Ritz solution vector is completely numerically stable (3.20) if $K(A_n)$ is bounded independent of n . The Rayleigh-Ritz approximation is completely numerically stable (3.21)

if μ_n is bounded independent of n.

The definition of μ_n may perhaps seem arbitrary, but it may be considered in the following manner. The set of all equations of the form (3.23) for all $\underline{\beta}_n$ satisfying (3.24) includes, except for a scaling factor and the degenerate case $\underline{\delta}_n = -\underline{b}_n$, all possible right hand sides of the equation

$$A_n \underline{a}_n = \underline{b}_n + \underline{\delta}_n \qquad \ldots(3.26)$$

that is, for any $\underline{\delta}_n \neq -\underline{b}_n$ there exists a scalar k such that

$$k(\underline{b}_n + \underline{\delta}_n) \in B_n \ , \quad B_n = \left\{ \underline{\beta}_n : \sum_{j=1}^{n} \beta_n(j)^2 = 1 \right\}$$

Therefore the set

$$Z_n = \left\{ z_n(x) : z_n = \sum_{j=1}^{n} \alpha_n(j)\, \phi_j(x) \ , \ A_n \underline{\alpha}_n = \underline{\beta}_n \ , \ \underline{\beta}_n \in B_n \right\}$$

contains all possible numerically computed approximations to the solution $\bar{z}_n(x)$ of the equations

$$A_n \underline{a}'_n = \underline{b}_n + \underline{\delta}_n \ ,$$

each scaled by some factor $1/k$. μ_n is therefore a measure of the range of norms of functions in Z_n, and hence a measure of the range of norms of functions defined by equations (3.26).

Mikhlin shows that Samokish's definition of complete numerical stability in terms of the boundedness of $K(A_n)$ and μ_n is equivalent to the criterion of almost-orthonormality imposed by Mikhlin to achieve (3.20), (3.21). This is achieved by showing that for the Rayleigh-Ritz method μ_n is bounded independent of n iff $K(A_n)$ is bounded.

Samokish provides an additional concept of stability, which he

terms 'computationally-stable'. The Rayleigh-Ritz solution vector (resp. approximation) is computationally stable if $K(A_n)$ (resp. μ_n) "do not grow too rapidly". This informal concept makes the point that it is not necessary to have a stable method if good approximations can be obtained with a small number of co-ordinate functions. We might prefer to say that such co-ordinate systems are 'sufficiently stable'.

Criteria which enable it to be asserted that a co-ordinate system is 'sufficiently stable' for a particular problem and particular accuracy requirements have been given by Mikhlin (4). These relate to the (impractical) case in which the perturbed Rayleigh-Ritz equations (3.4) are solved algebraically rather than numerically. If the co-ordinate sequence $\{\phi_i\}$ is not strongly minimal in the energy space H_L then the corresponding Rayleigh-Ritz method is not numerically stable. Assuming that this is the case, and letting $\lambda_1^{(n)}$ denote the smallest eigenvalue of the Rayleigh-Ritz matrix, we have $\lambda_1^{(n)} \to 0$ as $n \to \infty$. Let ε_n in (3.4) satisfy

$$\|\varepsilon_n\| \leq \beta \lambda_1^{(n)} \qquad 0 < \beta < 1$$

Then the bounds

$$\|\underline{a}_n' - \underline{a}_n\| \leq \frac{c.(\lambda_1^{(n)})^{-\frac{1}{2}} \|\varepsilon_n\| + (\lambda_1^{(n)})^{-1} \|\underline{\delta}_n\|}{1 - \beta}$$

where $c \geq \|y(x)\|_L$ is an upper bound for the energy norm of $y(x)$, and

$$\|\bar{y}_n(x) - y_n(x)\| \leq \|\underline{a}_n' - \underline{a}_n\| \left(\|\delta_n\| + \|\varepsilon_n\| . \left[(\lambda_1^{(n)})^{-\frac{1}{2}} + \|\underline{a}_n' - \underline{a}_n\| \right] \right)$$

hold.

Thus, for example, if

$$\|\varepsilon_n\| = O\left(\left(\lambda_1^{(n)}\right)^{1+k}\right)$$

and

$$\|\underline{\delta}_n\| = O\left(\left(\lambda_1^{(n)}\right)^{\frac{1}{2}+k}\right)$$

then

$$\|\underline{a}_n - \underline{a}'_n\| = O\left(\left(\lambda_1^{(n)}\right)^{k-\frac{1}{2}}\right) \qquad \ldots(3.27)$$

and

$$\|y_n(x) - \bar{y}_n(x)\| = O\left(\left(\lambda_1^{(n)}\right)^{k}\right) \qquad \ldots(3.28)$$

In this case we can see that it is possible to evaluate the right-hand side of equations (3.5) less accurately than the matrix A_n without detriment to the order of accuracy of the resulting approximations. We also note that in the case $0 < k < \frac{1}{2}$ (3.27) diverges, whilst (3.28) converges. This situation, in which the coefficients of an approximate solution are unstable although the corresponding function approximation remains stable (if not convergent) is well known in several types of numerical computation, and we have examples of this for the Rayleigh-Ritz method in Table II and elsewhere.

3.4 Rescaled co-ordinate systems

It seems surprising that the criteria for the stability of the Rayleigh-Ritz method should indicate that a rescaling of a co-ordinate system $\{\phi_i\}$ of the form

$$\{\psi_i(x) : \psi_i = d(i)\,\phi_i(x)\}$$

may generate an unstable approximation where $\{\phi_i\}$ generates a stable one. Here the $d(i)$ are scalar functions of i, and we assume $d(i) \neq 0$, so that the co-ordinate system $\{\psi_i\}$ remains complete in the energy space H_L. Mikhlin illustrates this by a theoretical discussion (4, p.139) concerning the co-ordinate systems

$$\{\phi_i\} = \sqrt{2i+1} \int_0^x P_i^*(t)\, dt$$

and $$\{\psi_i\} = \int_0^x P_i^*(t)\, dt$$

for problems involving the operator d^2/dx^2 on $(-1,1)$, with boundary conditions $y(-1) = y(1) = 0$. For this operator the functions $\{\phi_i\}$ are orthonormal, and consequently generate completely numerically stable approximations for similar operators of the form

$$p_0(x) \frac{d^2}{dx^2} - p_1(x) \quad ; \quad y(-1) = y(1) = 0$$

where $$p_0(x) \ , \ p_1(x) > 0$$

The functions $\{\psi_i\}$, on the other hand, are not orthonormal, and in fact for the operator d^2/dx^2 generate the Rayleigh-Ritz matrix A_n.

$$A_n(i,j) = \begin{cases} 0 & i \neq j \\ 2i+1 & i=j \end{cases} \qquad i=1 \ldots n$$

Hence the functions $\{\psi_i\}$ are only strongly minimal, and a Rayleigh-Ritz approximation in terms of these is not completely numerically stable, in view of the theory.

To consider the practical effects of simple rescaling of the co-ordinate system we consider the basis functions

$$\{S_i^2\}_{i=1}^{\infty} = \sin i \pi x \qquad \ldots(3.29)$$

and

$$\{P_i^2\}_{i=1}^{\infty} = \frac{(i+1)^{i+1}}{i^i} \cdot x^i(1-x) \qquad \ldots(3.30)$$

used in the solution of problems L1, L2, L3. (3.29) and (3.30) may be obtained from the functions (3.12), (3.13) by rescaling so that the

maximum value of each basis function is unity. Results for these calculations are summarized in Tables III and IV, and Tables A V, A VI, A VII, A VIII of Appendix A.

A comparison of the error $\bar{e}_n$ given in Tables I and III (similarly Tables A I, A V and A III, A VII) indicates that the solutions $\bar{\bar{y}}_n(x)$ defined in terms of the co-ordinate systems $\{S_i^1\}$ and $\{S_i^2\}$ differ insignificantly. Furthermore, it can be seen from Table III (similarly A V, A VII) that the coefficients $a_n''(i)$ appear to be stable, and we note also that these coefficients are, as would be expected since $\frac{\sqrt{2}}{i\,\tilde{\pi}} < 1$, uniformly smaller than the corresponding coefficients in Table I (resp. A I, A III). In fact, if we denote by $a_n''(i, \{S_i^j\})$ the coefficients of the basis functions $\{S_i^j(x)\}$, $j = 1,2$, then it is easily verified that the coefficients of Tables I and III (similarly A I and A V and A III and A VII) satisfy to reasonable numerical accuracy the relation

$$a_n''(i, \{S_i^1\}) = \frac{i\,\tilde{\pi}}{\sqrt{2}}\, a_n''(i, \{S_i^2\}) \qquad \ldots(3.31)$$

For example, Table V lists for comparison the values of the left and right hand sides of relation (3.31) for the coefficients given in Tables I and III, for $n = 8$.

In spite of the remarkable agreement in the accuracy of the solution vectors and the approximate solutions obtained using the basis functions $\{S_i^1\}$ and $\{S_i^2\}$, it is easy to show that these co-ordinate systems have different theoretical stability properties. If we denote the Rayleigh-Ritz matrices of size n for the basis $\{S_i^1\}$ and $\{S_i^2\}$ by $A_n\{S_i^1\}$ and $A_n\{S_i^2\}$ we have for the problem L1

$$A_n\{S_i^1\}(i,j) = \begin{cases} 1 - \dfrac{1}{i^2\,\tilde{\pi}^2} & i = j \\ 0 & i \neq j \end{cases}$$

Behaviour of the Solution Vector and the Approximate Solution

Problem L1 $\quad \phi_i = \sin i\pi x$

$\bar{e}_n$	2	3	7	8	11	12
$\bar{e}_n$	3.859594'-3	1.904862'-3	4.047900'-4	2.847648'-4	1.020953'-4	7.768347'-5
e_n	3.859645'-3	1.904899'-3	4.048682'-4	2.848654'-4	1.016520'-4	7.969140'-5
$a_n''(1)$	+7.177549'-2	+7.177543'-2	+7.177543'-2	+7.177543'-2	+7.177543'-2	+7.177543'-2
	-8.272431'-3	-8.272435'-3	-8.272435'-3	-8.272435'-3	-8.272431'-3	-8.272431'-3
		+2.416203'-3	+2.416202'-3	+2.416202'-3	+2.416202'-3	+2.416202'-3
			-1.014305'-3	-1.014305'-3	-1.014305'-3	-1.014305'-3
			+5.181388'-4	+5.181345'-4	+5.181350'-4	+5.181350'-4
			-2.995009'-4	-2.995014'-4	-2.995021'-4	-2.994921'-4
			+1.884510'-4	+1.884515'-4	+1.884434'-4	+1.884437'-4
				-1.262037'-4	-1.261944'-4	-1.263374'-4
					+8.873929'-5	+8.873914'-5
					-6.47128'-5	-6.356625'-5
					+4.758160'-5	+4.758160'-5
						-4.165843'-5

Table III

Behaviour of the Solution Vector and the Approximate Solution

Problem L1

$$\phi_i = \frac{(i+1)^{i+1}}{i^i}\, x^i(1-x)$$

$\bar{e}_n$	2	3	7	8	11	12
$\bar{e}_n$	3.023743'-4	2.515688'-5	2.611428'-6	4.135070'-6	9.059904'-6	6.437299'-6
	+4.810279'-2	+4.694134'-2	+4.705405'-2	+4.707158'-2	+4.705481'-2	+4.706804'-2
	+2.529388'-2	+2.876275'-2	+2.854369'-2	+2.801466'-2	+2.802679'-2	+2.886649'-2
		-2.469549'-3	-4.480655'-3	+3.718649'-4	+2.624824'-3	-1.053394'-2
			+8.488773'-3	-1.188183'-2	-2.890636'-2	+4.175611'-2
			-1.302598'-2	+3.206521'-2	+8.176981'-2	-9.532493'-2
a_n^* (i)			+9.227033'-3	-4.528377'-2	-1.103773'-1	+1.025441'-1
			-2.582144'-3	+3.147361'-2	+7.656192'-2	-3.019357'-2
				-8.607059'-3	-8.123635'-2	-3.751891'-2
					+1.522049'-1	+5.681281'-2
					-1.387720'-1	-7.163411'-2
					+4.427732'-2	+6.220636'-2
						-2.082470'-2

Table IV

A Comparison of the Solution Vectors and the

Approximate Solutions of Problem L1

using the co-ordinate systems $\{S_i^1\}$ and $\{S_i^2\}$ for $n = 8$.

	$a_8''(i, \{S_i^2\})$	$\frac{i\pi}{\sqrt{2}}\, a_8''(i, \{S_i^2\})$	$a_8''(i, \{S_i^1\})$	Exact $a_8(i, \{S_i^1\})$
$\bar{e}_n$	2.847648'-4		2.847872'-4	2.848654'-4
$a_n''(i)$	+7.177543'-2	+1.594685'-1	+1.594447'-1	+1.594448'-1
	-8.272435'-3	-3.675890'-2	-3.675552'-2	-3.675340'-2
	+2.416202'-3	+1.610481'-2	+1.610239'-2	+1.610235'-2
	-1.014305'-3	-8.997697'-3	-9.012881'-3	-9.012680'-3
	+5.181345'-4	+5.755877'-3	+5.755014'-3	+5.754190'-3
	-2.995014'-4	-3.992539'-3	-3.991965'-3	-3.991506'-3
	+1.884515'-4	+2.930871'-3	+2.930347'-3	+2.930339'-3
	-1.262037'-4	-2.243145'-3	-2.242829'-3	-2.242453'-3

Table V

and

$$A_n\left[S_i^2\right](i,j) = \begin{cases} \frac{1}{2}(i^2\pi^2 - 1) & i=j \\ 0 & i\neq j \end{cases}$$

Since these are diagonal matrices, their eigenvalues are the diagonal elements, and their condition numbers are given by

$$K_n\left(A_n\left[S_i^1\right]\right) = \left(1 - \frac{1}{n^2\pi^2}\right) / \left(1 - \frac{1}{\pi^2}\right) \quad \ldots$$

i.e. $$K_n\left(A_n\left[S_i^1\right]\right) < 1 / (1 - 1/\pi^2) \qquad \ldots(3.32)$$

and

$$K_n\left(A_n\left[S_i^2\right]\right) = (n^2\pi^2 - 1) / (\pi^2 - 1)$$

Clearly, from (3.32) the basis $\left\{S_i^1\right\}$ is almost orthonormal in the energy space of the operator of problem L1, and thus generates a completely numerically stable sequence of Rayleigh-Ritz approximations, whilst the basis $\left\{S_i^2\right\}$ is only strongly minimal, and cannot be expected to do so. This is not, however, reflected in the Rayleigh-Ritz approximations obtained using the basis $\left\{S_i^2\right\}$. Figs. II and III illustrate the behaviour of the maximum error $\bar{e}_n$ of the Rayleigh-Ritz approximation and the condition number of the Rayleigh-Ritz matrix for $n = 2 \ldots 20$ for problem L1 and the basis functions $\left\{S_i^1\right\}$ and $\left\{S_i^2\right\}$ respectively. Similar graphs are given as Figs. A I and A II in Appendix A for the problem L2. In each case we plot $\log_{10} \bar{K}_n$ and $\log_{10} \bar{e}_n$ against n. The unpredictable behaviour of $\bar{e}_n$ for larger values of n ($n \geq 12,13$) in these graphs is attributed to the effects of quadrature error. Note also that $\bar{K}_n$, the condition number of the matrix A_n obtained by evaluating $\lambda_1^{(n)}$, $\lambda_n^{(n)}$ numerically, using a Sturm-sequence-bisection algorithm, and shown on the graphs for $n > 12$, exhibits irregular behaviour. For $n \leq 12$ we have $\bar{K}_n \doteq K_n$.

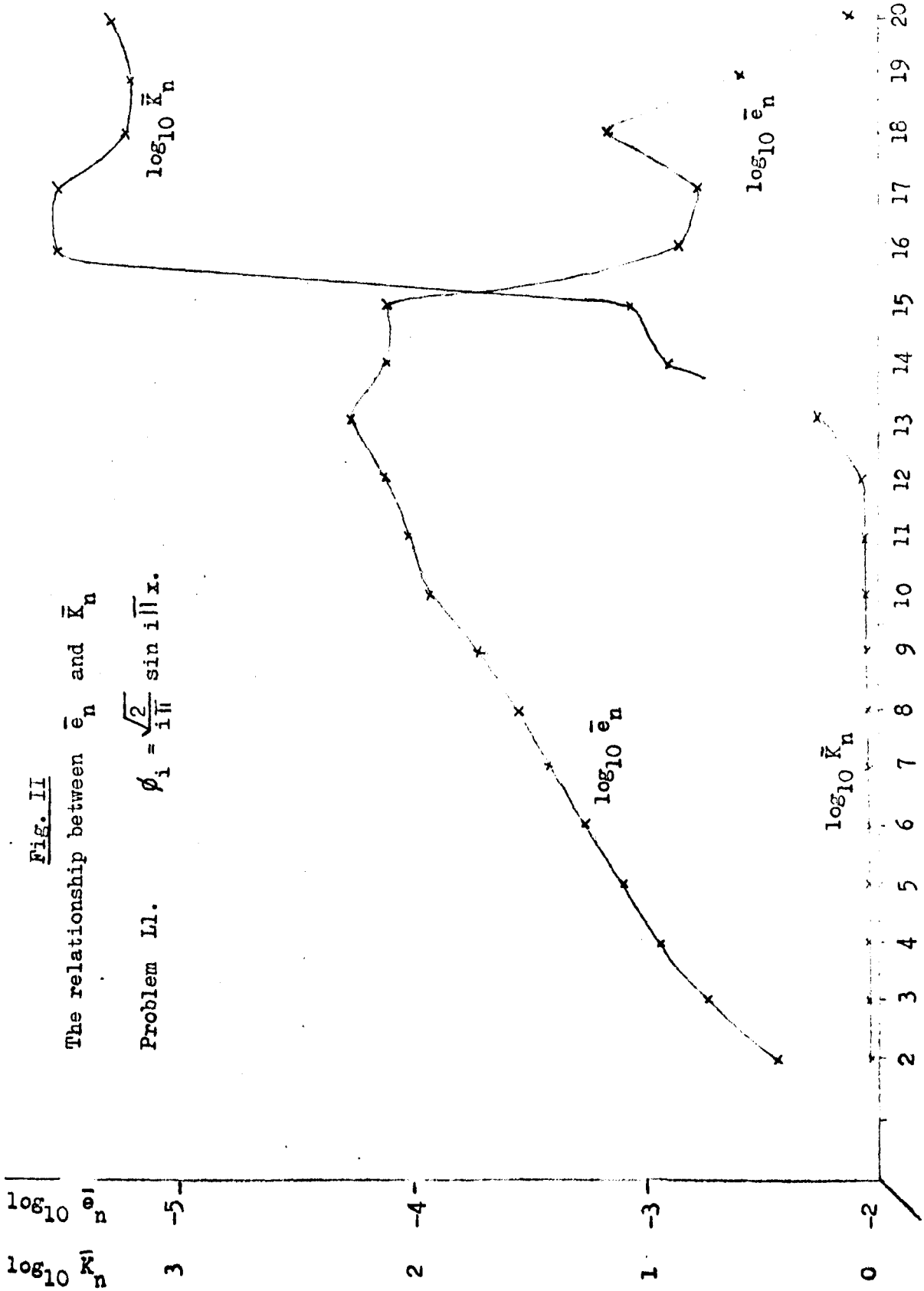


Fig. II

The relationship between $\bar{e}_n$ and $\bar{K}_n$

Problem 11. $\phi_i = \frac{\sqrt{2}}{i\,\overline{11}} \sin i\,\overline{11}\,x.$

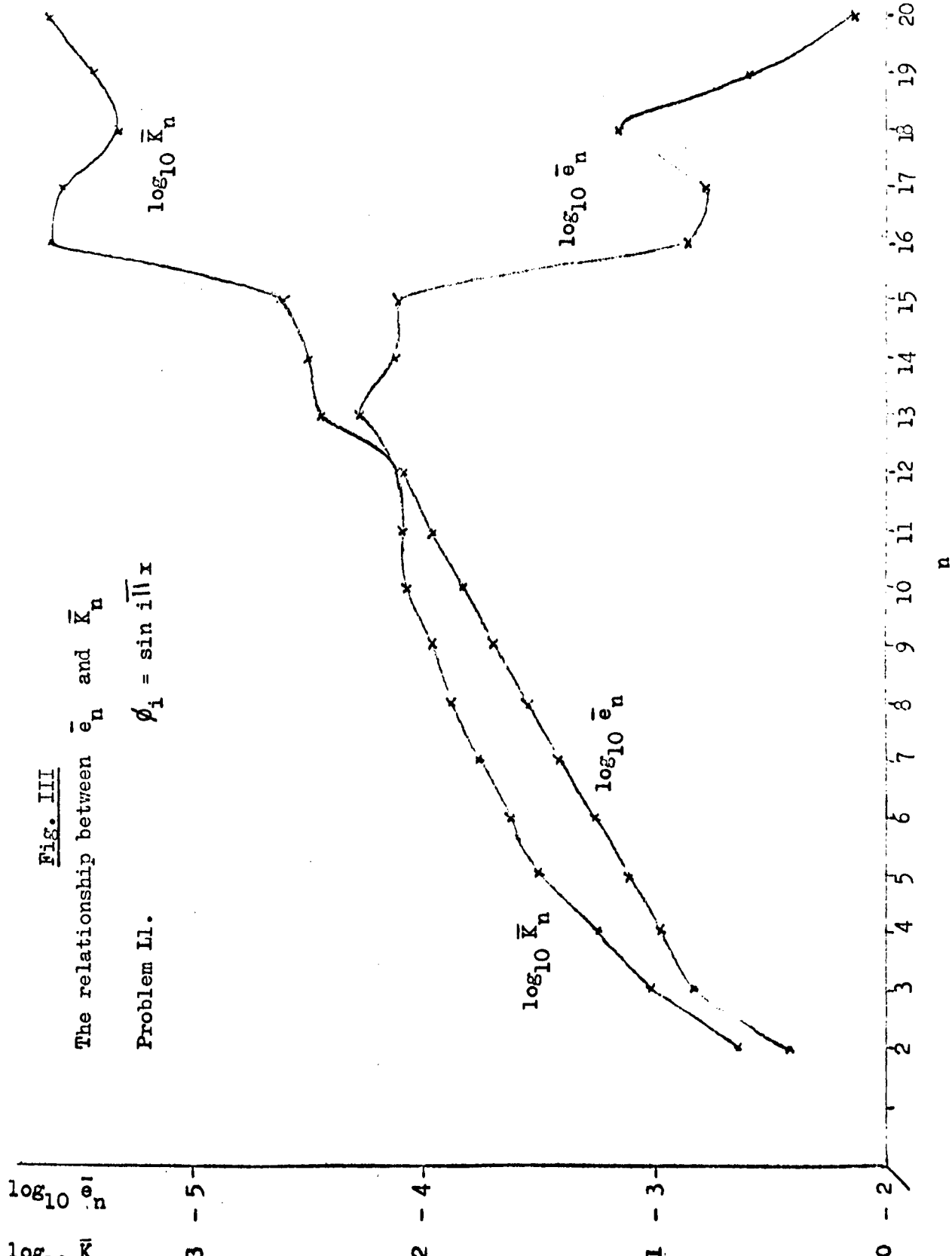


Fig. III

The relationship between $\bar{e}_n$ and $\bar{K}_n$

Problem L1. $\phi_i = \sin i\overline{11}\, x$

Similar comparisons to those made for the co-ordinate systems $\{S_i^1\}$ and $\{S_i^2\}$ on the basis of Tables I and III may be made for the co-ordinate systems $\{P_i^1\}$ and $\{P_i^2\}$ from Tables II and IV, but it is not possible to make the same deductions. First, although the values for $\bar{e}_n$ shown in Tables II and IV are similar in size, the difference between them is often of the same order of magnitude as the errors themselves. Secondly, the relationship which existed between the coefficients of Tables I and III does not hold for the coefficients of Tables II and IV; we do not have

$$a_n''(i, \{P_i^1\}) = \frac{(i+1)^{i+1}}{i^i} a_n''(i, \{P_i^2\}) \qquad \ldots(3.33)$$

For example, taking $n = 8$, $i = 5$ in (3.33) we obtain

$$a_n''(i, \{P_i^1\}) = +2.021488$$

$$\frac{(i+1)^{i+1}}{i^i} a_n''(i, \{P_i^2\}) = +0.478731$$

The unpredictable behaviour of the error $\bar{e}_n$ for the basis functions $\{P_i^1\}$ and P_i^2 is shown in Fig.IV and Fig.V respectively. This time this behaviour cannot be attributed to quadrature error, since the Gauss formulae used are exact in this case. However, the Rayleigh-Ritz matrix for these problems resembles the Hilbert matrix, whose instability is well known. For the basis $\{P_i^1\}$ and problem L1 the matrix A_n is given by

$$A_n(i,j) = \frac{ij}{i+j-1} - \frac{i+j+2}{i+j} + \frac{ij+i+j}{i+j+1} + \frac{2}{i+j+2} - \frac{1}{i+j+3}$$

The condition number plotted graphically in Figs. III and IV is again obtained by a Sturm-sequence-bisection algorithm, and for such matrices this procedure is not reliable, so that whilst the smooth behaviour exhibited for $n \leq 7$ may be indicative of the real situation, one is

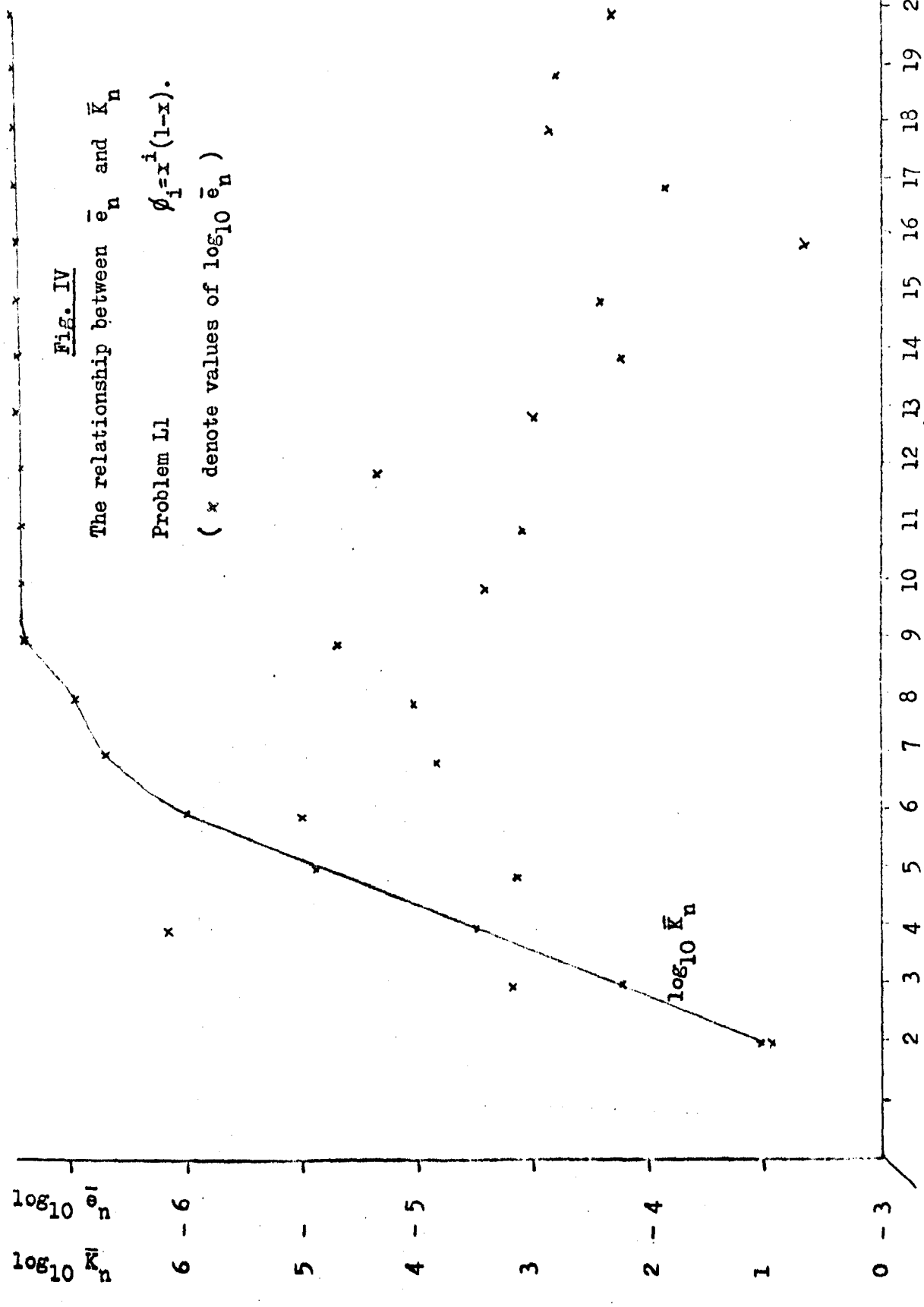


Fig. IV

The relationship between $\bar{e}_n$ and $\bar{K}_n$

Problem L1 $\phi_i = x^i(1-x)$.

(* denote values of $\log_{10} \bar{e}_n$)

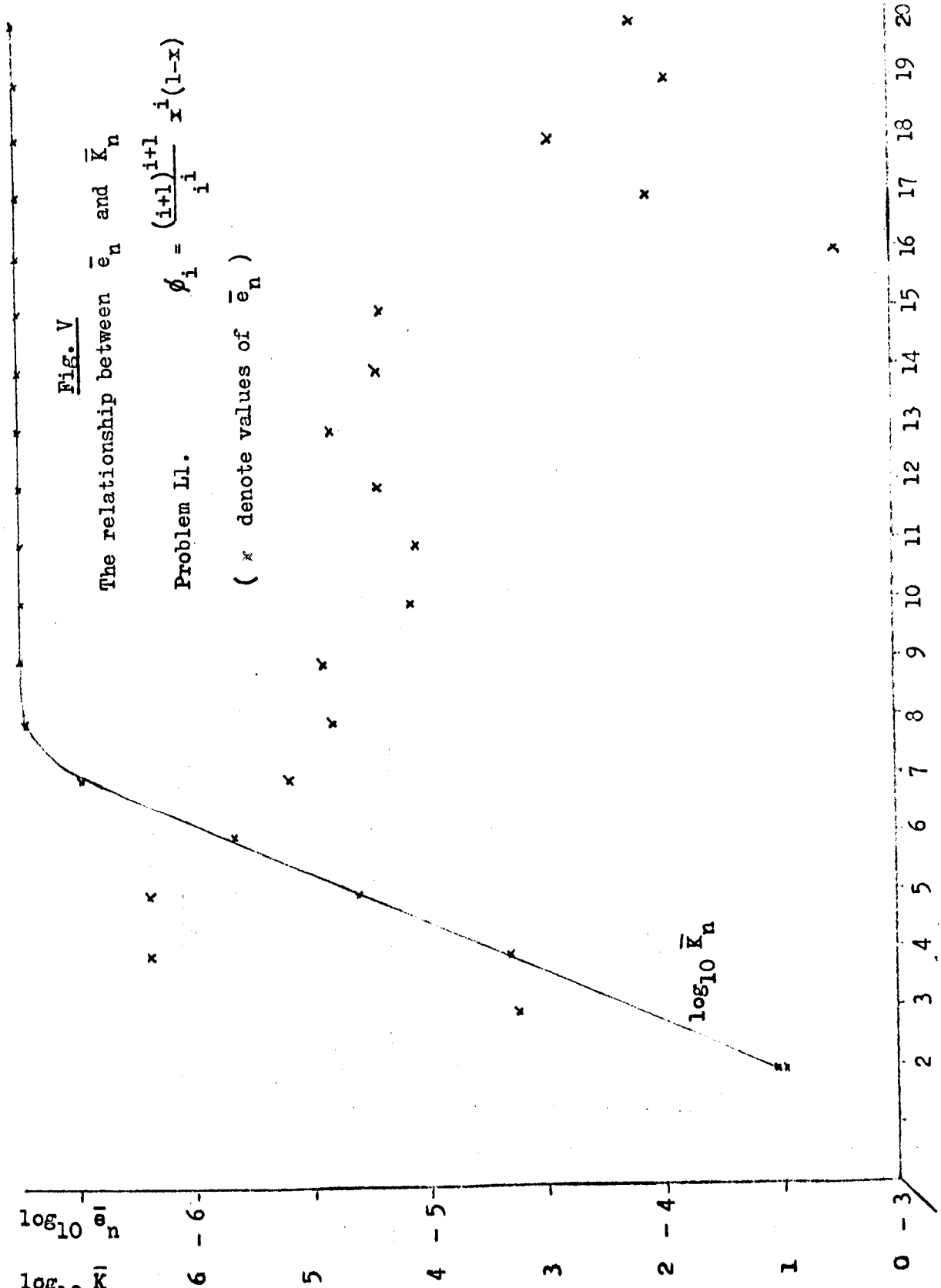


Fig. V

The relationship between $\bar{e}_n$ and $\bar{K}_n$

Problem L1. $\phi_i = \frac{(i+1)^{i+1}}{i^i} x^i(1-x)$

(* denote values of $\bar{e}_n$)

not justified in assuming that the behaviour for larger n is as shown. In fact, for $n = 10$ in Fig.III and $n = 9$ in Fig.IV, the numerical procedure determines a small negative eigenvalue; since the matrix A_n is positive definite this is clearly in error.

That the relationship (3.33) should not hold for the coefficients $a_n''(i, \{P_i^1\})$ and $a_n''(i, \{P_i^2\})$ need not be surprising; we might hope that, defining $C_n''(i, \{P_i^j\})$ by

$$\sum_{i=1}^{n} a_n''(i, \{P_i^j\})\ P_i^j(x) = \sum_{i=1}^{n+1} C_n''(i, \{P_i^j\})\ x^i \qquad \ldots(3.34)$$

[i.e. for $j = 1$

$$\sum_{i=1}^{n} a_n''(i, \{P_i^1\})\ x^i(1-x) = \sum_{i=1}^{n+1} C_n''(i, \{P_i^1\})\ x^i$$

so that

$$C_n''(1, \{P_i^1\}) = a_n''(1, \{P_i^1\})$$

$$C_n''(k, \{P_i^1\}) = a_n''(k, \{P_i^1\}) - a_n''(k-1, \{P_i^1\}) \qquad k = 2 \ldots n$$

$$C_n''(n+1, \{P_i^1\}) = a_n''(n, \{P_i^1\}) \quad]$$

we would have

$$C_n''(k, \{P_i^1\}) = C_n''(k, \{P_i^2\}) \qquad \ldots(3.35)$$

In Table VI we demonstrate the coefficients $C_8''(k, \{P_i^1\})$ and $C_8''(k, \{P_i^2\})$, indicating that this is not the case, though clearly, from Tables II and IV, the polynomials given by the right hand side of (3.34) for $j = 1,2$, have similar values over the range $[0,1]$.

A Comparison of the coefficients of the expanded polynomials obtained from the solutions of problem L1 using the co-ordinate systems $\{P_i^1\}$ and $\{P_i^2\}$ for n=8

i	$C_8''(i, P_i^1)$	$C_8''(i, P_i^2)$
1	+1.885666'-1	+1.882863'-1
2	-9.572148'-3	+8.126348'-4
3	-6.152755'-2	-1.855731'-1
4	-8.449268'-1	-1.485677'-1
5	+2.748948'+0	+6.237728'-1
6	+9.713144'-1	-1.278051'+0
7	-7.646923'-1	+1.440501'+0
8	-2.885118'+0	-8.399349'-1
9	+6.570082'-1	+1.987547'-1

Table VI

We have considered numerical examples of the effect of rescaling co-ordinate systems on the accuracy of numerically computed Rayleigh-Ritz approximations. A theoretical discussion of the properties of such a rescaling, i.e. of considering the relationship between the Rayleigh-Ritz approximations in terms of a system $\phi_i(x)$, and the system

$$\psi_i(x) = d(i)\,\phi_i(x) \quad ,$$

rests on the notions of 'optimally scaled matrices' discussed by Bauer (1), Stoer and Witzgall(1), and Van der Sluis (1,2). Denoting by $y_n(x)$ the Rayleigh-Ritz approximations

$$y_n(x) = \sum_{i=1}^{n} a_n(i)\phi_i(x) = \sum_{i=1}^{n} \alpha_n(i)\psi_i(x)$$

we require

$$A_n\, \underline{a}_n = \underline{b}_n$$

and

$$D_n\, A_n\, D_n\, \underline{\alpha}_n = D_n\, \underline{b}_n$$

where A_n is the Rayleigh-Ritz matrix $(\phi_i , \phi_j)_L$, $i,j = 1,\ldots,n$ and D_n is the diagonal matrix with elements $D_n(i,i) = d(i)$, $D_n(i,j) = 0$, $i \neq j$. Pre-multiplication by D_n corresponds to a row scaling of A_n and post-multiplication to a column scaling. The latter is known to be irrelevant, but the former can affect the order of pivoting in the numerical solution of the Rayleigh-Ritz equations, and thus the accuracy of the numerical solution. This is clearly more likely when the off-diagonal elements of the matrix are large by comparison with the diagonal elements than when they are small, and this is a property which distinguishes the polynomial basis from the trigonometric one, so that the inconsistency of solutions defined in terms of the polynomial basis used here may be attributed to this effect of rescaling.

We remark here that the property that rescaling has little effect when off-diagonal elements are small is closely analogous to the fact that if a matrix is uniformly asymptotically diagonal of degree p then pre or post multiplication by a diagonal matrix does not affect this property (§ 2.5, also Delves and Mead (1)).

A further discussion of the effects of pre and post multiplication of the Rayleigh-Ritz matrix by matrices of particular type will be considered in § 3.6.

3.5 The use of Chebyshev and Legendre Polynomials

The high accuracy displayed when low order polynomial approximations are compared with smooth, non-periodic solutions such as those of problems L1, L2, L3 suggest, along with the convergence properties of polynomial expansions (Mikhlin (4), Ciarlet, Schultz, Varga (1), also § 2.6) that if the condition difficulties encountered with polynomial basis functions, such as $\{P_i^1\}$ and $\{P_i^2\}$ can be avoided, then polynomial approximation will have much to recommend it. A number of alternatives, all of which display more satisfactory computational properties, are possible. We consider in this and subsequent sections the co-ordinate systems

$$\left\{ Tch_i(x) : \quad Tch_i(x) = x(1-x)\, T^*_{i-1}(x) \right\} \qquad i \geq 1 \qquad \ldots(3.36)$$

$$\left\{ L_i^1(x) : \quad L_i^1(x) = x(1-x)\, P^*_{i-1}(x) \right\} \qquad i \geq 1 \qquad \ldots(3.37)$$

and

$$\left\{ L_i^2(x) : \quad L_i^2(x) = \sqrt{2i+1} \int_0^x P_i^*(x) \right\} \qquad i \geq 1 \qquad \ldots(3.38)$$

where $T_k^*(x)$ and $P_k^*(x)$ denote respectively the shifted Chebyshev

and Legendre polynomials defined on the interval $[0,1]$.

The system (3.38) has been considered by Mikhlin (4) and Ciarlet, Schultz and Varga (1). No study appears to have been made of the systems (3.36) and (3.37), though McDonald (1) considers the use of a co-ordinate system involving shifted Chebyshev polynomials in the application of the Kantorovich method to certain second order partial differential equations. We shall consider the Kantorovich method in a subsequent chapter. Chebyshev polynomials have been widely used in other areas of approximation theory; for a general survey see, for example, Fox and Parker (1).

The most convenient of the above co-ordinate systems is (3.36), principally because of the simple closed form of the Chebyshev polynomial. We shall see subsequently that for certain problems the bases (3.36) and (3.37) generate matrices with an unusual and convenient splitting property; it is trivial to see that for problems of the form (3.1) in which $k=1$, $p_1(x)=1$, $p_o(x)=0$, the functions (3.38) form an orthonormal co-ordinate system, and are thus almost orthonormal for all operators similar to $-d^2/dx^2$.

We again solve as test examples the problems L1, L2, L3 used previously, and note here important features of the computations.

The basis (3.36) generates approximate solutions to problems L1 and L2 which converge rapidly to the known solutions as n increases, until the overriding persistence of numerical error limits the accuracy at $O(_{10}-7)$. The convergence process can be observed to continue past this point if the Rayleigh-Ritz equations are constructed and solved using double precision arithmetic. This behaviour is illustrated in Tables VII and VIII for problems L1 and L3 and in Table A VIII of Appendix A for problem L2. An

important property of the Chebyshev basis $\{Tch_i\}$ can be observed in the behaviour of the coefficients $a''_n(i)$ tabulated in Table VII, particularly by examining the coefficients for $n = 11, 12, 13$. We notice that

$$\left| a''_{11}(i) - a''_{12}(i) \right| \doteq \begin{cases} 10^{-9} & i = 1,3,5,7,9,11 \\ 10^{-5} & i = 2,4,6,8,10 \end{cases}$$

and

$$\left| a''_{12}(i) - a''_{13}(i) \right| \doteq \begin{cases} 10^{-6} & i = 1,3,5,7,9,11 \\ 10^{-11} & i = 2,4,6,8,10,12 \end{cases}$$

...(3.39)

We remind outselves of the progress of the Rayleigh-Ritz iteration. In progressing from an approximation in terms of n functions to one in terms of $n+1$ functions an additional unknown is introduced into the equations, so that the Rayleigh-Ritz system is transformed from

$$A_n \; \underline{a}_n = \underline{b}_n$$

to

$$A_{n+1} \; \underline{a}_{n+1} = \underline{b}_{n+1}$$

where the matrix A_{n+1} satisfies

$$A_{n+1}(i,j) = A_n(i,j) \qquad i,j = 1...n$$

The relationships (3.39) suggest that the transition from a system of size n to one of size $n+1$ affects only those elements of the solution vector whose indices have the same parity as $n+1$, and, for problems L1 and L2 this is indeed the case, though not for problem L3.

This can be shown by considering the elements of A_n, defined for variational problems of the form (3.1) by

Behaviour of the Solution Vector and the Approximate Solution

Problem L1 $\quad \phi_i = x(1-x)T^*_{i-1}(x)$

n	2	3	7	8	11	12	13
$\bar{e}_n$(short)	3.026703'-4	2.509355'-5	5.364418'-7	5.364185'-7	5.960464'-7	7.748603'-7	8.344650'-7
$\bar{e}_n$(long)	3.025191'-4	2.507028'-5	5.186285'-11	3.005634'-11	3.104306'-11	3.199334'-11	3.052655'-11
$a^n_n(i)$	+2.777770'-1	+2.760581'-1	+2.760625'-1	+2.760625'-1	+2.760647'-1	+2.760647'-1	+2.760667'-1
	+8.536660'-2	+8.536654'-2	+8.502727'-2	+8.503174'-2	+8.503883'-2	+8.504813'-2	+8.504813'-2
		-2.926426'-3	-2.917837'-3	-2.917837'-3	-2.913669'-3	-2.913667'-3	-2.909523'-3
			-2.654716'-4	-2.612352'-4	-2.544133'-4	-2.453247'-4	-2.453247'-4
			+6.507062'-6	+6.507155'-6	+1.044420'-5	+1.044477'-5	+1.444902'-5
			+1.953085'-6	+5.625322'-6	+1.189062'-5	+2.047428'-5	+2.047428'-5
			+2.073393'-7	+2.073381'-7	+3.725146'-6	+3.725239'-6	+7.482358'-6
				+2.625749'-6	+7.951918'-6	+1.571846'-5	+1.571849'-5
					+2.823937'-6	+2.824082'-6	+6.205603'-6
					+3.750246'-6	+1.027446'-5	+1.027446'-5
					+1.439878'-6	+1.439949'-6	+4.268334'-6
						+4.549001'-6	+4.549003'-6
							+1.964800'-6

Table VII

$$A_n(i,j) = \int_0^1 \left\{ p_1(x)\, \phi_i'(x)\, \phi_j'(x) + p_o(x)\, \phi_i(x)\phi_j(x) \right\} dx$$

In the case in which $p_1(x)$, $p_o(x)$ are constants, say $p_1(x)=k_1$, $p_o(x)=k_o$, and $\phi_i = x(1-x)T^*_{i-1}(x)$ we have

$$A_n(i,j) = k_1 \int_0^1 \left\{ ((1-2x)T^*_{i-1}(x) + x(1-x)T^{*'}_{i-1}(x)) \right.$$
$$\left. ((1-2x)T^*_{j-1}(x) + x(1-x)T^*_{j-1}(x)) \right\} dx$$
$$+ k_o \int_0^1 \left\{ x^2(1-x^2)T^*_{i-1}(x)T^*_{j-1}(x) \right\} dx$$

Transforming the integrals to the interval (-1,1), and using the relationships

$$\frac{d}{dx} T_i(x) = \begin{cases} i\,(2T_{i-1} + \ . \ + 2T_2 + T_o) & i \text{ odd} \\ i\,(2T_{i-1} + \ . \ . \ + 2T_3 + 2T_1) & i \text{ even} \end{cases}$$

(Abramowitz and Stegun (1)) and the property

$$\int_{-1}^1 T_i(x)\, T_j(x) = 0 , \qquad i + j \text{ odd}$$

we can show that

$$I_1 = \int_{-1}^1 (1-x^2)^2\, T_i'\, T_j'\, dx = 0 \qquad i + j \text{ odd}$$

$$I_2 = \int_{-1}^1 2x(1-x^2)(T_i\, T_j' + T_j T_i')\, dx = 0 \qquad i + j \text{ odd}$$

$$I_3 = \int_{-1}^1 4x^2\, T_i T_j\, dx = 0 \qquad i + j \text{ odd}$$

$$I_4 = \int_{-1}^1 (1-x^2)^2\, T_i T_j\, dx = 0 \qquad i + j \text{ odd}$$

and hence that

$$A_n(i,j) = k_1(I_1 + I_2 + I_3) + k_o I_4$$

$$= 0 \qquad \text{if } i + j \text{ is odd}$$

The matrix A_n therefore has the form

$$\begin{matrix} a_{11} & 0 & a_{33} & 0 & & & a_{n-2,n} \\ 0 & a_{22} & 0 & a_{24} & & & 0 \\ a_{31} & 0 & a_{33} & 0 & & & \\ 0 & a_{42} & 0 & & & & \\ & & & & a_{n,n-2} & 0 & a_{n,n} \end{matrix}$$

with non-zero entries on the diagonal and alternate sub and super diagonals, forming a chequerboard pattern. The equations can therefore be arranged into two independent systems of the form

$$A_n^{(1)} \underline{a}_n^{(1)} = \underline{b}_n^{(1)}$$

$$A_n^{(2)} \underline{a}_n^{(2)} = \underline{b}_n^{(2)} \qquad \ldots(3.40)$$

where

$$A_n^{(1)}(i,j) = A_n(2i-1,\ 2j-1), \qquad i,j = 1 \ldots \left[\tfrac{n+1}{2}\right]$$

$$A_n^{(2)}(i,j) = A_n(2i,2j) \qquad i,j = 1 \ldots \left[\tfrac{n}{2}\right]$$

$$a_n^{(1)}(i) = a_n(2i-1), \quad \underline{b}_n^{(1)}(i) = \underline{b}_n(2i-1), \quad i = 1 \ldots \left[\frac{n+1}{2}\right]$$

$$a_n^{(2)}(i) = a_n(2i), \quad \underline{b}_n^{(2)}(i) = \underline{b}_n(2i), \quad i = 1 \ldots \left[\tfrac{n}{2}\right] \qquad \ldots(3.41)$$

and $[k]$ is the integer part of k.

Thus for linear problems with constant coefficients, and using the basis $\{Tch_i\}$, the Rayleigh-Ritz equations may be considered as two disjoint systems determining the odd and even coefficients of the approximate solution. This property has not been exploited in the solution of problems discussed in this chapter, but will play a very important role in the study of mildly non-linear problems which we undertake in Chapter Four. We give here (Table VIII) numerical results for problem L3, which does not have constant coefficients and for which the Rayleigh-Ritz equations do not form disjoint sets. It will be seen that the property (3.39) which illustrated this disjunction for problem L1 does not occur. We remark that the error $\bar{e}_n$ shown in Table VIII is determined by comparison with the given solution, in which the decimal constants are given only to 5 significant figures, and from tables of the Bessel functions involved (Abramowitz and Stegun). It is felt that this accounts, at least in part, for the rather larger values of $\bar{e}_n$ as compared with those obtained for problem L1, shown in Table VII.

The basis functions (3.37) display very similar properties to those of (3.36), and indeed, the first two co-ordinate functions of each system are identical. The splitting property characterized by (3.40) and (3.41) again holds for problems with constant coefficients, and the convergence and stability properties of the two co-ordinate systems (3.36) and (3.37) are very similar. Evidence of this can be found in Table IX, where the basis (3.37) is used to solve problem L1, and in Appendix A (Tables A IX and A X).
Although the modified Legendre basis $\{L_i^1\}$ compares equally with the modified Chebyshev basis $\{Tch_i\}$ from the viewpoints of stability and convergence, it is felt that the convenience of the simple closed forms of the Chebyshev functions give them considerable computational

Behaviour of the Solution Vector and the Approximate Solution

Problem L3

$$\phi_i = x(1-x)T^*_{i-1}(x)$$

n	2	3	7	8	11	12
$\bar{e}_n$	3.979801'-4	2.601742'-4	1.171827'-4	1.168251'-4	1.167059'-4	1.172423'-4
$a''_n(i)$	+8.117174'-1	+8.134221'-1	+8.138625'-1	+8.138618'-1	+8.138641'-1	+8.138625'-1
	-4.351347'-3	-4.989116'-3	-7.292385'-3	-7.284786(-3	-7.270894'-3	-7.250778'-3
		+2.730763'-3	+3.486576'-3	+3.485223'-3	+3.490349'-3	+3.486789'-3
			-1.739069'-3	-1.731917'-3	-1.718539'-3	-1.698955'-3
			+2.955526'-4	+2.943696'-4	+2.992474'-4	+2.958608'-4
			-4.387863'-5	-3.771558'-5	-2.542813'-5	-6.949143'-6
			+8.260959'-6	+7.430742'-6	+1.186019'-5	+8.781357'-6
				+4.287780'-6	+1.473822'-5	+3.142384'-5
					+3.712246'-6	+1.126689'-6
					+7.374252'-6	+2.131584'-5
					+2.506777'-6	+7.409455'-7
						+9.488190'-6

Table VIII

Behaviour of the Solution Vector and the Approximate Solution

Problem L1 $\phi_i = x(1-x)P^*_{i-1}(x)$

n	2	3	7	8	11	12
$\bar{e}_n$	3.026723'-4	2.466142'-5	2.980232'-7			
$a''_n(i)$	+2.777770'-1	+2.770377'-1	+2.770349'-1			
	+8.536648'-2	+8.536642'-2	+8.518546'-2			
		-3.901936'-3	-3.895344'-3			
			-4.280556'-4			
			+1.194163'-5			
			+1.378951'-6			
			+4.619441'-7			

Table IX

advantages. The evaluation of the scalar products of the Rayleigh-Ritz matrix by a quadrature technique, which we shall examine in Chapter Five, requires many evaluations of the basis functions and their derivatives. The Legendre recurrence relation may be applied to evaluate the functions, but only indirectly to evaluate the derivatives. This presents a barrier from the viewpoint of computer time. The basis functions $L_i^2(x)$ defined in (3.38) can be evaluated more efficiently, however. Although these functions are considered theoretically by Ciarlet, Schultz and Varga (1) and Mikhlin (4), no account of a practical evaluation algorithm appears to have been given, and we therefore illustrate some of the computational properties of this co-ordinate system.

The basis functions are

$$L_i^{(2)}(x) = \int_0^x P_i^*(t)\,dt \quad , \qquad i = 1 \ldots \infty$$

where P_i^* are the shifted Legendre polynomials on the interval $[0,1]$. Evaluation of the derivatives $\frac{d}{dx}(L_i^{(2)}(x))$ is easily accomplished using

$$\frac{d}{dx}(L_i^{(2)}(x)) = P_i^*(t) = \begin{cases} 1 & i = 0 \\ 2x-1 & i = 1 \\ \frac{1}{i+1}((2i-1)\,.\,(2x-1)\,.\,P_i^*(x) \\ \qquad - i\,P_{i-1}^*(x)), & i > 1 \end{cases}$$

where the general case is the usual recurrence relation for the shifted Legendre polynomials (e.g. Sansone (1), p.177) and the case $i = 0$ is given only for use with this recurrence, that is, the function $L_0^{(2)}(x) = \int_0^x P_0^*(t)dt$ is not one of the basis functions.

Evaluation of the basis functions themselves proceeds from the relationship for Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$ on $(-1,1)$,

$$2n \int_0^x (1-t)^{\alpha} (1+t)^{\beta} \; P_n^{(\alpha,\beta)}(t)\, dt$$

$$= P_{n-1}^{(\alpha+1,\beta+1)}(0) - (1-x)^{\alpha+1}(1+x)^{\beta+1} P_{n-1}^{(\alpha+1,\beta+1)}(x)$$

(see Abramowitz and Stegun (1), p.775, 22.13.1)

The Legendre polynomials on (-1,1), $P_n(x)$ are a special case of the Jacobi polynomials given by $\alpha = \beta = 0$, so that

$$2n \int_0^x P_n(t)dt = P_{n-1}^{(1,1)}(0) - (1-x^2)\, P_{n-1}^{(1,1)}(x)$$

Considering for the moment the basis $\{\phi_i\}$ on (-1,1) given by

$$\phi_i(x) = \sqrt{2i+1} \int_{-1}^x P_i(t)\, dt$$

we have

$$\phi_i(x) = \sqrt{2i+1} \left(\int_{-1}^0 P_i(t)dt + \int_0^x P_i(t)dt \right)$$

$$= \sqrt{\frac{2i+1}{2i}} \left(P_{i-1}^{(1,1)}(0) - (1-x^2)\, P_{i-1}^{(1,1)}(x) - P_{i-1}^{(1,1)}(0) \right)$$

$$= \sqrt{2i+1}(1-x^2)\, P_{i-1}^{(1,1)}(x) \,/\, 2i$$

Transformation of this result to the interval $0 \leq t \leq 1$ finally gives

$$L_i^{(2)}(x) = \sqrt{2i+1}.t(t-1)\; P_{i-1}^{(1,1)}(2t-1)/i$$

and the Jacobi polynomials $P_{i-1}^{(1,1)}(x)$ are given by the recurrence (Abramowitz and Stegun, (1), p.782, 22.7.1)

$$P_i^{(1,1)}(x) = (a_{3i}\, x\, P_{i-1}^{(1,1)}(x) - a_{4i}\, P_{i-2}^{(1,1)}(x)\,)/a_{1i}$$

where

$$a_{1i} = 4i(i+1)(i+2)$$

$$a_{3i} = 2i(2i+1)(2i+2)$$

$$a_{4i} = 4i^2(i+2)$$

and the initial values $P_o^{(1,1)}(x) = 1$, $P_1^{(1,1)}(x) = 2x$

The basis functions $L_i^{(2)}(x)$ can therefore be evaluated by the application of the recurrence relation, and in particular by a variant of the algorithm described by Clenshaw (1) for the evaluation of Chebyshev polynomials.

We comment now on the results of the application of the Rayleigh-Ritz method using this co-ordinate system to the test problems of this chapter, and in particular consider the stability properties of the related basis $\int_0^x P_i^*(x)$ to consider the points made by Mikhlin (4, p.138-9). Tables X and XI summarize the application of the two co-ordinate systems to the problem L1. Similar results for problem L2 are given in Tables A XI and A XII of Appendix A. It is felt that the results shown are sufficient to indicate that the basis $L_i^{(2)}(x)$ and the basis $(2i+1)^{-\frac{1}{2}} L_i^{(2)}(x)$ have similar computational properties, and that stability considerations are not restrictive when convergence is rapid. Further, a comparison with the results expressed in Table VII indicates that the errors obtained with the basis functions $x(1-x)\, T_{i-1}^*(x)$ and with $(2i+1)^{-\frac{1}{2}} L_i^{(2)}(x)$ (i.e. $x(x-1)\, P_{i-1}^{(1,1)*}(x)$) are of the same order of magnitude, though the latter are uniformly smaller. It would seem that these co-ordinate systems may be of some importance in the approximation of the solution of a given variational problem whose solution is known not to have special properties (e.g. which is not periodic or which does not have a discontinuous derivative of low order).

Behaviour of the Solution Vector and the Approximate Solution

Problem L1

$$\phi_i(x) = \sqrt{2i+1} \int_0^x P_i^*(x)dx$$

n	2	3	7	8
$\bar{e}_n(x)$	3.025531'-4	2.467632'-5	2.384185'-7	2.980232'-7
	-1.603748'-1	-1.603962'-1	-1.603962'-1	-1.603962'-1
	-3.817707'-2	-3.817707'-2	-3.817818'-2	-3.817818'-2
		+1.769360'-3	+1.769379'-3	+1.769379'-3
			+2.041966'-4	-2.041965'-4
$a_n^n(i)$			-6.040485'-6	-6.040485'-6
			-7.574862'-7	-7.569331'-7
			-4.880311'-7	-4.880346'-7
				-4.953981'-7

Table X

Behaviour of the Solution Vector and the Approximate Solution

Problem L1

$$\phi_i(x) = \int_0^x P_i^*(x)dx$$

n	2	3	7	8
$e_n(x)$	3.026723'-4	2.461671'-5	3.576278'-7	4.172325'-7
$a_n^n(i)$	-2.777770'-1	-2.778139'-1	-2.778139'-1	-2.778139'-1
	-8.536648'-2	-8.536642'-2	-8.536893'-2	-8.536893'-2
		+4.681374'-3	+4.681426'-3	+4.681426'-3
			+6.126439'-4	+6.126439'-4
			-2.019894'-5	-2.019894'-5
			-2.668125'-6	-2.665932'-6
			-1.918064'-6	-1.918078'-6
				-2.245284'-6

Table XI

3.6 Algebraic Transformations of Co-ordinate Systems

An abstract algebraic formulation of the effects of considering different scalings and combinations of a particular set of basis functions is possible, and we outline this here. Unfortunately, as we commented in § 3.3, even the simple case of diagonal scaling leads to inconclusive theoretical discussions of optimal scaling of matrices. In this section we examine choices of combination matrix which correspond to improving the minimal-orthonormal properties of the basis. One of these is obtained by the Gramm-Schmidt process, but we shall show that, for at least some of the purposes for which we are employing the Rayleigh-Ritz method, this method has disadvantages.

In the numerical considerations which have occupied much of this chapter, we have supposed that in applying the Rayleigh-Ritz method to the solution of the problem (3.1) with homogeneous boundary conditions, we select a sequence of co-ordinate functions $\{\phi_i\}_{i=1}^{\infty}$ which satisfy the boundary conditions and the additional assumptions of Chapter Two, and determine an approximation of the form (3.3) from the Rayleigh-Ritz equations. If we return to the viewpoint taken in considering convergence properties of the Rayleigh-Ritz method in Chapter Two, we see that the selection of a sequence of basis functions $\{\phi_i\}_{i=1}^{\infty}$ can be seen as the selection of a sequence of finite dimensional subspaces $\{S_n\}_{n=1}^{\infty}$ of the energy space of the problem, and then the selection of a basis for this space. The theoretical convergence properties of the Rayleigh-Ritz method are independent of the choice of basis, but not of the choice of sequence of subspace. On the other hand, the numerical properties of the method are determined by the choice of basis functions within a particular subspace.

Consider the variational problem

$$\min_{y} I(y) = (Ly,y) - 2(f,y)$$

subject to $y(0) = y(1) = 0$,

and let $\{S_n\}_{n=1}^{\infty}$ be a sequence of subspaces of the energy space of L . We assume $S_n \subset S_{n+1}$, so that we can consider a sequence of functions $\{\phi_i\}_{i=1}^{\infty}$ with the property that the subsequences $\{\phi_i\}_{i=1}^{n}$ form a basis for the space S_n for all n . The basis functions $\{\phi_i\}_{i=1}^{n}$ will be known as the fundamental basis functions for the space S_n . The Rayleigh-Ritz approximation in S_n , obtained in terms of the basis $\{\phi_i\}$ will be denoted $y_n(x, \{\phi\})$ and is defined by

$$y_n(x, \{\phi\}) = \sum_{i=1}^{n} \underline{a}_n(i)\, \phi_i(x)$$

where $\underline{a}_n$ satisfies

$$A_n \underline{a}_n = \underline{b}_n \qquad \ldots(3.41)$$

and $A_n(i,j) = (\phi_i, L\phi_j)$ and $\underline{b}_n(i) = (f,\phi_i)$

Let $\{\psi_i\}_{i=1}^{\infty}$ be another sequence of basis functions with the property that $\{\psi_i\}_{i=1}^{n}$ forms a basis for S_n for all n . Clearly, we can write

$$\psi_i = \sum_{j=1}^{n} c_{ij}\, \phi_j \qquad , i = 1\ldots n$$

and extending the usual matrix operator to sequences of functions we denote

$$\{\psi_i\} = C\,\{\phi_i\}$$

where $C = (c_{ij})$ is a matrix with the property that, if C_n denotes the $n \times n$ submatrix of C ; $C_n(i,j) = c_{ij}$, $i,j=1\ldots n$ then C_n is non-singular for all n . In the above C is clearly lower triangular. Although if C is not of this form

the sequence

$$\{\psi_i\}_{i=1}^{n} = C_n \{\phi_i\}_{i=1}^{n}$$

does form a basis of S_n, this relationship does not hold for all $k < n$. In terms of the basis $\{\psi_i\}$ we obtain the Rayleigh-Ritz approximate solution

$$y_n(x, \{\psi\}) = \sum_{i=1}^{n} \alpha_n(i)\, \psi_i(x)$$

from the equations

$$A_n^* \underline{\alpha}_n = \underline{b}_n^* \qquad \ldots(3.43)$$

where

$$A_n^*(i,j) = (\psi_i \cdot L\, \psi_j) \text{ and } \underline{b}_n^*(i) = (f \cdot \psi_i)$$

The equations (3.42) and (3.43) are related by expressing (3.43) as

$$C_n A_n C_n^T \underline{\alpha}_n = C_n \underline{b}_n \qquad \ldots(3.44)$$

since we have $A_n^* = C_n A_n C_n^T$ and $\underline{b}_n^* = C_n \underline{b}_n$. From (3.44) we deduce $\underline{\alpha}_n = (C_n^T)^{-1} \underline{a}_n$, so that

$$y_n(x, \{\psi_i\}) = \sum_{i=1}^{n} \sum_{j=1}^{n} C_n^{-1}(i,j)\, a_n(i)\, \psi_i(x)$$

and expressing $\{\psi_i\}_{i=1}^{n}$ as $C_n \{\phi_i\}_{i=1}^{n}$ we have

$$y_n(x, \{\psi\}) = y_n(x, \{\phi\}) \qquad \ldots(3.45)$$

/We remark here that a formalism of this kind has been used by Reid (1) in connection with the finite element method/.

The equality (3.45) is an idealized relationship attainable only in analytic computation. In practical application of the Rayleigh-Ritz method errors are introduced by the numerical processes, and one

would like a choice of the matrix C (i.e. of the functions ψ_i) which would minimize the effect of the numerical errors. In the literature of linear algebra there have been two approaches to this problem. These correspond to the cases $C = c_{ij}$, $c_{ij} = \begin{cases} d(i) & i=j \\ 0 & i \neq j \end{cases}$ i.e. C is diagonal, and $C_n A_n C_n^T = I_n$ where I_n is the identity matrix of size n. Both approaches are pleasing, the former for the simplicity of C, and the latter for the simplicity of solution of the resulting equations, where the coefficients $\underline{\alpha}_n$ are given by $\underline{\alpha}_n = C_n^T \underline{b}_n$. The suggestion that such orthonormalization might be an important tool in connection with the Rayleigh-Ritz method was made by Davis and Rabinowitz (1), but little attention seems to have been paid to this remark subsequently. There is, however, no need to exclude more general transformations; for example, the transformation matrix C of the transformation from

$$\{\phi_i\} = x^i (1-x)$$

to

$$\{\psi_i\} = x (1-x) T^*_{i-1} (x)$$

is not diagonal, and the resulting matrix $C AC^T$ is not the identity matrix.

There are, however, a number of remarks which can be made concerning the two special cases, which relate particularly to the Mikhlin stability analysis in this chapter. The first comment rests on a result of Dovbysh (1) (see Mikhlin (4), p.6).

Let the sequence $\{\phi_i\}_{i=1}^{\infty}$ be minimal in the appropriate energy space H_L. Then there exist scalars $d(i)$, $i = 1,2...$ such that the sequence $\{\psi_i\}_{i=1}^{\infty}$, $\psi_i = d(i)\,\phi_i$ is strongly minimal in the energy space H_L.

Clearly, this is a special case of the general transformation in which C is the diagonal matrix $D : d_{ii} = d(i)$ $i=1,2\ldots$. The determination of the scalars $d(i)$ rests on the construction of a sequence of functions $\{\theta_j\} \in H_L$ which are bi-orthogonal to $\{\phi_i\}$, that is, the relation

$$(\phi_i , \theta_j) = \delta_{ij}$$

(see Mikhlin (4, p.3)). If the scalars $d(i)$ are now chosen so that

$$\sum_{i=1}^{\infty} (d(i))^{-2} \| \theta_i \|_L = \gamma^2$$

where γ is an arbitrary non-zero constant, then the functions $\psi_i = d(i)\,\phi_i$ will be strongly minimal, and the eigenvalues $\lambda_i^{(n)}$ of the Gramm matrix of size n of the sequence $\{\psi_i\}$ in H_L will satisfy

$$\lambda_i^{(n)} \geq \gamma^{-2} \qquad \text{for all } n .$$

The determination of the scalars $d(i)$ by this method is not straightforward. For example, one might assume that a variation of the Gramm-Schmidt procedure for orthogonalizing functions could be used, but this is not so, as we indicate below.

Since the bi-orthogonal sequence of functions $\theta_{j(x)}$ satisfies $\theta_{j(x)} \in H_L$, we may write

$$\theta_{j(x)} = \sum_{k=1}^{\infty} \alpha(j,k)\, \phi_k (x) \qquad j = 1 \ldots \infty$$

for scalars $\alpha(j,k)$ to be determined from the bi-orthogonality condition

$$(\phi_i, \theta_j) = \sum_{k=1}^{\infty} \alpha(j,k)\, (\phi_i \ \phi_k)_L = \delta_{ij}$$

$$i = 1 \ldots \infty$$

Even for the single function $\Theta_1(x)$ this is an infinite system for the determination of the scalars $\alpha(1,k)$, and the determination of d(i) is therefore impractical. In the case of a finite subsequence of functions $\{\phi_i\}_{i=1}^{n}$ and a positive definite operator L the smallest eigenvalue of the Rayleigh-Ritz matrix is strictly positive, and the construction is unnecessary.

Thus we suggest that the result of Dovbysh is here of only theoretical value, and we remind ourselves that the condition that a co-ordinate system is strongly minimal is insufficient to give complete numerical stability.

We might, however, use the Gramm-Schmidt algorithm to orthonormalize the basis functions $\{\phi_i\}$ with respect to the scalar product of the energy space H_L, or with respect to the scalar product in an energy space H_M, where M and L are similar operators, in the latter case obtaining a system of functions which are almost orthonormal in H_L. Considering the case of orthonormalizing in H_L, we seek to define functions

$$\psi_i = \sum_{j=1}^{i} c_{ij}\,\phi_j$$

satisfying $(\psi_i, \psi_j)_L = \delta_{ij}$

This orthonormalization may be performed numerically, and in particular the Modified Gramm-Schmidt algorithm of Rice (1) is to be preferred. In general it will be necessary to approximate the scalar product in H_L by some quadrature rule. The resulting Rayleigh-Ritz matrix in terms of the basis $\{\psi_i\}$ is then the identity matrix, and the

coefficients $\underline{\alpha}_n$ of (3.43) are given trivially by $\underline{\alpha}_n(i) = (f, \psi_i)$. The determination of these coefficients is completely numerically stable, since the basis $\{\psi_i\}$ is orthonormal in H_L. The situation is not, however, as convenient as it might seem, if we wish to evaluate the solution function. [The comments which follow no doubt have less relevance if orthonormalization of the basis functions were utilized in the application of the Rayleigh-Ritz method to the eigenvalue problem, provided that only the eigenvalue, and not the eigenfunction, is required].

The approximate solution $y_n(x)$ is given by

$$y_n(x, \{\psi_i\}) = \sum_{i=1}^{n} \underline{\alpha}_n(i) \psi_i(x) ,$$

but where the basis $\{\psi_i\}$ has been constructed from $\{\phi_i\}$ by numerical orthonormalization, will be evaluated from

$$y_n(x, \{\psi_i\}) = \sum_{i=1}^{n} \alpha_n(i) \sum_{j=1}^{i} c_{ij} \phi_j(x) \qquad \ldots(3.46)$$

In the case that the basis $\{\phi_i\}$ are 'nearly' linearly dependent, which is the case when the basis is only minimal or strongly minimal, at least for large n the orthonormalization coefficients c_{ij} may be large in modulus, and of alternating sign, and the evaluation of (3.46) can be severely affected by cancellation error. We illustrate these points by considering the solution of the test problems of this chapter using a sequence of basis functions $\{\psi_i\}_{i=1}^{n}$ orthonormal in H_L derived from the fundamental basis functions $\{\phi_i\}_{i=1}^{n}$ by application of the Modified Gramm-Schmidt process described by Rice(1).

In practice this algorithm seems best applied in the following manner. Let $\underline{\gamma}_1$, $\underline{\gamma}_2$ be n component vectors and define

$$(\underline{\gamma}_1 , \underline{\gamma}_2) = (\sum_{i=1}^{n} \underline{\gamma}_1(i)\phi_i(x) . \sum_{j=1}^{n} \underline{\gamma}_2(j)\phi_j(x))_L$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \underline{\gamma}_1 (i) \underline{\gamma}_2 (j) (\phi_i , \phi_j)_L \qquad ...(3.47)$$

and

$$\| \underline{\gamma}_1 \|^2 = (\underline{\gamma}_1 , \underline{\gamma}_1) \qquad ...(3.48)$$

The function $\phi_i(x)$ is therefore represented by the unit vector $\underline{e}_n^i = (0,0,... 0,1,0 .. 0)^T$ where the non-zero element occupies the i^{th} position. The sequence $\{\Psi_i\}_{i=1}^n$ can therefore be constructed by applying the Modified Gramm-Schmidt process to the n unit vectors $\underline{e}_n^i$, i = 1 ... n , i.e. to the columns of the identity matrix I_n using (3.47) and (3.48) to define the scalar product and norm required by the process. The scalar products $(\phi_i , \phi_j)_L$ are evaluated initially, and these, together with the expressions (3.47) and (3.48) may be evaluated in extended precision arithmetic if necessary.

The results of applying this orthonormalization process to the solution of problem L1 using the basis $\phi_i(x) = x^i(1-x)$ as the fundamental basis are summarized in the style used previously, in Table XII. We give only the cases n = 2,3,7,8. For n = 11 , the orthonormalization procedure generates a function $\Psi_{10}(x)$ with 'negative' norm, so that the Gramm-Schmidt procedure fails. This may be attributed to rounding errors in the Gramm-Schmidt process being magnified because of the near-dependence of the fundamental basis, since the problem does not occur using double precision arithmetic. For the results given, the coefficients $a_n''(i)$ decrease rapidly for small i as would be expected for an orthonormal basis. However, it will be noticed that this process does not continue throughout the

Behaviour of the Solution Vector and the Approximate Solution

Problem L1 — Orthonormal Basis $\{\psi_i\} = C\{\phi_i\} : \phi_i(x) = x^i(1-x)$

n	2	3	7	8
$\bar{e}_n(x)$	3.026723'-4	2.485513'-5	3.481283'-5	6.061789'-5
	+1.521444'-1	+1.521444'-1	+1.521444'-1	+1.521444'-1
	+3.772085'-2	+3.772085'-2	+3.772085'-2	+3.772085'-2
		-1.761256'-3	-1.761256'-3	-1.761256'-3
			-2.012251'-4	-2.012251'-4
$a_n''(i)$			-1.049042'-5	-1.049042'-5
			+9.059902'-5	+9.059902'-5
			-3.967285'-4	-3.967285'-4
				+4.720687'-4

Table XII

Orthogonal Combination Matrix C^T: n = 7

Problem L1 $\qquad\qquad$ $\phi_i(x) = x^i(1-x)$

+1.825743'+0	-2.263173'+0	+2.639687'+0	-2.997811'+0	+3.311882'+0	-3.575060'+0	+3.331100'+0
	+4.526356'+0	-1.330399'+1	+2.706434'+1	-4.643788'+1	+7.205810'+1	-9.485606'+1
		+1.330396'+1	-6.320428'+1	+1.859026'+2	-4.350891'+2	+8.115766'+2
			+4.213496'+1	-2.789687'+2	+1.093061'+3	-3.007251'+3
				+1.395103'+2	-1.206916'+3	+5.420003'+3
					+4.841914'+2	-4.678116'+3
						+1.548143'+3

Table XIII

vectors $\underline{a}_7''$, $\underline{a}_8''$, and later coefficients begin to increase. This is again a consequence of the near-dependence of the basis $\{\phi_i(x)\}$. Additionally, the elements of the matrix C by which the orthonormal basis $\{\psi_i\} = C\{\phi_i\}$ is generated are large in modulus and have an alternating sign distribution. This can be seen clearly in Table XIII where C is given for the case $n = 7$. Thus the computation of

$$y_n(x, \{\psi_i\}) = \sum_{i=1}^{n} \underline{\alpha}_n(i) \sum_{j=1}^{i} c_{ij}\, \phi_j(x)$$

is likely to be affected by cancellation error, and it will be noticed that the results given in Table XII do not improve on those given in Table II for the simple basis $\phi_i(x) = x^i(1-x)$. Similar conclusions may be drawn from Tables A XIII and A XIV which relate to the application of this method to the problem L2.

We consider also the orthonormalization of one of the co-ordinate systems which has previously proved useful in the solution of these problems. Tables XIV and XV indicate the results obtained for problem L1 using orthonormalizing techniques with the fundamental basis $\phi_i(x) = x(1-x)\, T^*_{i-1}(x)$. The situation here is greatly improved; again the coefficients of the solution vector $\underline{\alpha}_n$ decrease rapidly, but furthermore the matrix C has elements which are not large in magnitude, so that cancellation is not a problem, as can be verified from the accuracy of the results. Again, the error of the approximation produced by this orthonormalizing technique is little different (though in fact marginally greater) than the error obtained by direct application of the fundamental basis. It is clear that the modified Gramm-Schmidt algorithm should be regarded as a numerical method for solving the problem

Behaviour of the Solution Vector and the Approximate Solution.

Problem L1 — Orthonormal Basis from $\phi_i = x(1-x)T^*_{i-1}(x)$

n	2	3	7	8	11	12
$\bar{e}_n(x)$	3.026723'-4	2.521276'-5	7.152557'-7	5.364418'-7	1.311302'-6	1.370906'-6
$a''_n(i)$	+1.521444'-1	+1.521444'-1	+1.521444'-1	+1.521444'-1	+1.521444'-1	+1.521444'-1
	+3.771979'-2	+3.771979'-2	+3.771979'-2	+3.771979'-2	+3.771979'-2	+3.771979'-2
		-1.759826'-3	-1.759826'-3	-1.759826'-3	-1.759826'-3	-1.759826'-3
			-2.036504'-4	-2.036504'-4	-2.036504'-4	-2.036504'-4
			+5.636362'-6	+5.636362'-6	-5.636362'-6	-5.636362'-6
			+2.062646'-6	+2.062646'-6	+2.062646'-6	+2.062646'-6
			+3.362074'-7	+3.362074'-7	+3.362074'-7	+3.362074'-7
				+3.615626'-6	+3.615265'-6	+3.615265'-6
					+1.259671'-6	+1.259671'-6
					+6.350773'-6	+6.350773'-6
					+2.793020'-6	+2.793020'-6
						+9.132504'-6

Table XIV

Orthogonal Combination Matrix C^T: n = 7

Problem L1

$$\phi_i = x(1-x)T^*_{i-1}(x)$$

+1.825743'+0	+2.894645'-6	+9.766665'-1	+5.916479'-6	+7.758984'-1	+1.990745'-5	+6.614903'-1
	+2.263177'+0	+4.933896'-6	+1.680796'+0	+1.216180'-5	+1.407036'+0	+2.100088'-5
		+1.662970'+0	+4.659541'-6	+1.448842'+0	+8.752800'-6	+1.275785'+0
			+1.316802'+0	+4.400969'-6	+1.266725'+0	+1.121716'-5
				+1.090612'+0	+1.261133'-5	+1.122963'+0
					+9.309718'-1	+3.566823'-6
						+8.122079'-1

Table XV

in terms of the fundamental basis, and not as a method of generating a new basis. In particular we have demonstrated that the stability properties of the numerical process are determined by those of the fundamental basis $\phi_i(x)$ and not by those of the numerically orthonormalized basis $\psi_i(x)$.

3.7 Linear Differential Equations with Singular Boundary Points.

In this section we demonstrate the application of the Rayleigh-Ritz method to the solution of two self-adjoint differential equations with singular points at the boundary of the region, corresponding to the equations considered by Mayers (1).

The original equations are

$$xy'' - y' - x = 0 \qquad \ldots(3.49)$$

and

$$x^2y'' - 2y + {}^3/2\, x^2 = 0 \qquad \ldots(3.50)$$

with boundary conditions

$$y(0) = y(1) = 0 \qquad \ldots(3.51)$$

These can be written in self adjoint form, so that (3.49) and (3.50) become

$$\frac{d}{dx}\left(\frac{1}{x}\frac{dy}{dx}\right) - \frac{1}{x} = 0 \qquad \ldots(3.52)$$

and

$$\frac{d}{dx}\left(\frac{dy}{dx}\right) + \frac{2}{x^2}y + \frac{3}{2} = 0 \qquad \ldots(3.53)$$

respectively, again subject to the boundary conditions (3.51). To the self adjoint equations (3.52) and (3.53) there correspond the following variational problems, which we shall refer to as S1 and S2 .

S1 :
$$\min_y \int_0^1 \frac{1}{x}\,(y'^2 + 2y)\,dx$$

and

$$S2: \quad \min_{y} \int_0^1 \left\{ y'^2 + \frac{2}{x^2} y^2 + 3y \right\} dx$$

again with boundary conditions (3.51). It is known that the equations (3.49) and (3.50), with boundary conditions (3.51) have the exact solution

$$y(x) = \frac{1}{2} x^2 \log x$$

(Mayers (1)), and clearly the third derivative of the solution $y(x)$ is infinite at $x = 0$. This weak boundary singularity does not seriously affect the Rayleigh-Ritz procedure used. The scalar products which are encountered, i.e.

$$\int_0^1 \left(\frac{1}{x} \frac{d\phi_i(x)}{dx} \cdot \frac{d\phi_j(x)}{dx} \right) dx \qquad \ldots(3.54)$$

and

$$\int_0^1 \left(\frac{d\phi_i(x)}{dx} \cdot \frac{d\phi_j(x)}{dx} + \frac{2}{x^2} \phi_i(x)\phi_j(x) \right) dx$$

are finite in value for any co-ordinate system $\{\phi_i\}$ satisfying the boundary conditions and, in the case of (3.54) satisfying

$$\left. \frac{d\phi_i}{dx} \right|_{x=0} = 0 \text{ for all } i \qquad \ldots(3.54')$$

Tables XVI, XVII, XVIII indicate the results obtained by applying the bases

$$\phi_i(x) = \frac{\sqrt{2}}{i\pi} \sin i\pi x$$

$$\phi_i(x) = x^i(1-x) \qquad \ldots(3.55)$$

and

$$\phi_i(x) = x(1-x)\, T^*_{i-1}(x)$$

respectively* to the solution of problem S1, whilst similar results

* Although these bases do not satisfy (3.54'), finite approximations to the infinite integrals (3.54) are obtained using Gaussian quadrature, and the resulting Rayleigh-Ritz approximations for problem S1 approximately satisfy $y'_n(0) = 0$.

Behaviour of the Solution Vector and the Approximate Solution

Problem S1 $\phi_i(x) = \frac{\sqrt{2}}{i\pi} \sin(i\pi x)$

n	2	3	7	8	11	12
$\bar{e}_n(x)$	1.783527'-2	1.242201'-2	2.824314'-3	2.261354'-3	1.035027'-3	8.579081'-4
	-1.724290'-1	-1.788035'-1	-1.906504'-1	-1.914668'-1	-1.924572'-1	-1.927132'-1
	+1.235743'-1	+9.793889'-2	+8.944236'-2	+8.900272'-2	+8.818358'-2	+8.803891'-2
		+3.742832'-2	-3.027280'-4	-1.194456'-3	-2.032842'-3	-2.264896'-3
			+3.053147'-2	+2.992591'-2	+2.883799'-2	+2.867068'-2
			+5.942240'-3	+4.162535'-3	+2.998370'-3	+2.703378'-3
			+1.830011'-2	+1.709909'-2	+1.512589'-2	+1.490595'-2
$a''_n(i)$			+1.782085'-2	+5.724548'-3	+3.393063'-3	+2.933648'-3
				+1.986742'-2	+1.003199'-2	+9.719673'-3
					+3.986399'-3	+2.951983'-3
					+8.388321'-3	+7.362261'-3
					+1.038918'-2	+3.457058'-3
						+1.223174'-2

Table XVI

Behaviour of the Solution Vector and the Approximate Solution

Problem S1 $\phi_i(x) = x^i(1-x)$

n	2	3	7	8	11	12
$\bar{e}_n(x)$	9.994446'-3	3.192781'-3	1.259446'-4	2.020090'-3	1.167431'-4	-1.144930'-4
$a_n^n(i)$	-3.972539'-2	-2.361520'-2	-9.285293'-3	-1.734776'-2	-9.188226'-3	-9.186685'-3
	-5.872157'-1	-9.235911'-1	-1.566350'+0	-7.818103'-1	-1.571813'+0	-1.572115'+0
		+5.069252'-1	+4.849440'+0	-8.815558'+0	+4.863986'+0	+4.871896'+0
			-1.224855'+1	+7.799633'+1	-1.183051'+1	-1.189457'+1
			+1.808265'+1	-2.657529'+2	+1.514397'+1	+1.535365'+1
			-1.382144'+1	+4.431003'+2	-6.333017'+0	-6.577276'+0
			+4.219692'+0	-3.591184'+2	-5.350572'+0	-5.387358'+0
				+1.130474'+2	+1.111291'+1	+1.084900'+1
					-1.781547'+1	-1.574112'+1
					+1.809849'+1	+1.458384'+1
					-6.814606'+0	-4.324956'+0
						-6.576579'-1

Table XVII

Behaviour of the Solution Vector and the Approximate Solution

Problem S1 $\qquad \phi_i(x) = x(1-x)\, T^*_{i-1}(x)$

n	2	3	7	8	11	12
$\bar{e}_n(x)$	9.994625'-3	3.192842'-3	1.595020'-4	9.191035'-5	-1.022592'-5	8.139758'-6
	-3.333336'-1	-2.953146'-1	-3.063508'-1	-3.075482'-1	-3.070433'-1	-3.070994'-1
	-2.936086'-1	-2.083348'-1	-2.254268'-1	-2.278202'-1	-2.269003'-1	-2.270212'-1
		+6.336468'-2	+4.611796'-2	+4.380849'-2	+4.479882'-2	+4.468762'-2
			-1.366389'-2	-1.592051'-2	-1.506235'-2	-1.518016'-2
			+7.757280'-3	+5.724895'-3	+6.653457'-3	+6.547540'-3
			-1.999099'-3	-3.954812'-3	-3.233555'-3	-3.344821'-3
$a^n_n(i)$			+2.500827'-3	+1.033043'-3	+1.848142'-3	+1.751686'-3
				-1.396933'-3	-9.359025'-4	-1.036573'-3
					+6.218126'-4	+5.403607'-4
					-2.386630'-4	-3.232212'-4
					+1.930710'-4	+1.360570'-4
						-5.893238'-5

Table XVIII

for problem S2 are presented as Tables A XV, A XVI, A XVII of Appendix A.

Similar conclusions may be drawn here as have been indicated earlier. First, the trigonometric polynomial approximation converges slowly, since the solution is not periodic. The convergence of the coefficients of the solution vector is no longer regular. The simple polynomial basis (3.55) again demonstrates that, whilst it may produce reasonably accurate solutions, no reliance may be placed on its convergence, and the situation in which coefficients of the solution vector are large in magnitude and alternate in sign again appears. Finally, the use of the modified Chebyshev basis again provides approximations which have high accuracy and solution vectors whose coefficients are stable and which decrease rapidly to zero.

3.8 Summary

A full treatment of the numerical application of the Rayleigh-Ritz method for linear differential equations requires a discussion of the error of the approximation. Since the error of the approximations calculated in this chapter depends on both the numerical errors in the solution of the Rayleigh-Ritz equations, and on the quadrature errors in their construction, these approximations are not covered by the error bounds of Ciarlet, Schultz and Varga (1) or Gladwell (1), which apply in the case of analytic computation. We defer a further study of the errors inherent in our approximation until Chapter Five.

However, it is well known that the error in variational calculations is often estimated by considering the magnitude and rate of convergence of the coefficients of the approximate solution.

We have shown that, whilst accurate solutions may sometimes be obtained when the coefficients of an approximate solution in terms of a particular basis do not converge to zero, we can place some reliance on the accuracy of a solution whose coefficients do tend to zero in this way. As might be expected, we can also comment that, unless a problem has a solution which is known to be periodic, a co-ordinate system which consists of trigonometric polynomials has few merits, and that for solutions which are 'smooth' expansions in terms of a polynomial basis may be very satisfactory. In particular we recommend the basis

$$\phi_i(x) = x(1-x)\, T^*_{i-1}(x)$$

as having desirable computational and convergence properties. We shall see in the next chapter that this also proves to be the case for a much wider class of problem than those so far considered.

Chapter Four

Numerical Considerations in the Application of the Rayleigh-Ritz method to Mildly Non-Linear Differential Equations

Whilst the linear differential equations and the corresponding simple quadratic variational problems which we considered in Chapter Three are sufficient to illustrate a number of important features of the Rayleigh-Ritz method as a numerical process, most recent interest in the method has been concerned with the solution of mildly non-linear differential equations of the form (2.27) which we restate here for convenience. Solve the differential equation

$$\sum_{s=0}^{k} (-1)^s \frac{d^s}{dx^s} \left(p_s(x) \frac{d^s y}{dx^s} \right) = f(x,y) \qquad 0 \leq x \leq 1 \qquad \ldots(4.1)$$

subject to the boundary conditions

$$y^{(r)}(0) = y^{(r)}(1) = 0 \qquad r=0 \,..\, k-1 \qquad \ldots(4.2)$$

and certain conditions on $p_s(x)$ and $f(x,y)$. These last conditions are, for example, those of Ciarlet, Schultz and Varga (1). [These have recently been extended by Gladwell (1), who considers the case in which the right hand side of (4.1) is replaced by $\sum_{i=0}^{k-1} f_i(x,y^{(i)})^{(i)}$ where $y^{(i)} = \frac{d^i}{dx^i} y(x)$ and each of the functions $f_i(x,y^{(i)})$ satisfies rather less restrictive conditions]. Under these conditions the solution of (4.1) minimizes the variational integral

$$I(y) = \int_0^1 \left\{ \sum_{s=0}^{k} p_s(x) \left(\frac{d^s y}{dx^s} \right)^2 - 2 \int_0^y f(x,\eta) d\eta \right\} dx$$

subject to (4.2) over the space of functions

$$\left\{ y \in Y : y^{(r)}(0)=y^{(r)}(1)=0 \ , \ r=0..k-1, \ \ y \in C^{k-1}[0,1] \ , \right.$$

$$\left. \frac{d^k y}{dx^k} \in \mathcal{L}^2[0,1] \right\}$$

The Rayleigh-Ritz approximation $y_n(x)$ of size n is defined by

$$y_n(x) = \sum_{i=1}^{n} a_n(i)\phi_i(x) \qquad \ldots(4.3)$$

where $\phi_i(x)$ are suitably chosen basis functions and

$$A_n \underline{a}_n = \underline{g}_n(\underline{a}_n) \qquad \ldots(4.4)$$

where $$A_n(i,j) = (\phi_i , \phi_j)$$

$$= \int_0^1 \sum_{s=0}^{k} p_s(x) \frac{d^s\phi_i}{dx^s} \frac{d^s\phi_j}{dx^s} \, dx$$

and $$\underline{g}_n(a_n) = g_n(a_n)(i) = \int_0^1 f(x,y_n(x))\, \phi_i(x) \, dx \qquad i = 1 \ldots n$$

The matrix A_n is an $(n \times n)$ real symmetric matrix, and is assumed to be positive definite. That A_n is positive definite is implied, for example, by the conditions of Ciarlet, Schultz and Varga (1). In general, as in Chapter Three, the elements of A_n and $\underline{g}_n(\underline{a}_n)$ are replaced by a quadrature approximation, though if the coefficients $p_s(x)$ and the chosen basis $\{\phi_i(x)\}$ are sufficiently simple, the matrix elements $A_n(i,j)$ may be evaluated analytically. We leave a full study of the effect of these quadrature errors until Chapter Five, but remark here that except where explicitly stated, the numerical results quoted in this chapter have been obtained by applying numerical quadrature to the evaluation of the elements of A_n and $\underline{g}_n$.

4.1 Details of the Application of the Rayleigh-Ritz Method.

In this section we describe the method of solution of problems of the form (4.1) using Rayleigh-Ritz techniques. We begin with a general approach and subsequently describe particular cases in which convenient modifications to this approach may be made.

Applied to mildly non-linear differential equations of the form (4.1) the Rayleigh-Ritz method may be seen to consist of an outer and an inner iteration. The outer iteration consists of the determination of a sequence of approximate solutions $y_n(x)$ for a sequence of values of n, where $y_n(x)$ is given in terms of a basis $\{\phi_i\}_{i=1}^{n}$ and is defined by (4.3) where the coefficient vector $\underline{a}_n$ is given by (4.4). This iteration may be terminated when some condition of the form

$$h\,(y_n(x),\, y_{n-1}(x)\,) \;<\; \delta$$

where $h\,(\,a(x),\,b(x)\,)$ is generally some norm or seminorm of $a(x) - b(x)$. For example, we might have

$$h\,(\,a(x),\,b(x)\,) = \max_{0 \le x \le 1} \left|a(x) - b(x)\right| \qquad \ldots(4.5)$$

or

$$h\,(a(x),\,b(x)\,) = \sqrt{\sum_{t=1}^{r} \left(a(x_t) - b(x_t)\right)^2} \qquad \ldots(4.6)$$

where $0 \le x_1 < x_2 \ldots < x_r \le 1$. (4.5) defines a norm on the function $a(x) - b(x)$, whilst (4.6) is a seminorm of this function. More frequently a relation of the form

$$\left\| \underline{a}_{n+1} - \underline{a}_n \right\| \;<\; \varepsilon \qquad \ldots(4.7)$$

is used, for simplicity. We have seen in Chapter Three that relationships of the form (4.5) or (4.6) may hold when (4.7) does not, and in these cases use of (4.7) may require additional computation; in

general, as we have seen, one prefers to choose a basis for which (4.7) is true if and only if (4.5) or (4.6) hold.

The inner iteration performs the solution of the non-linear equations (4.4). That is, for fixed n and a given $\underline{a}_n^{(o)}$ we determine a sequence of vectors $\underline{a}_n^{(r)}$, $r = 1,2.....$ such that

$$\lim_{r \to \infty} \underline{a}_n^{(r)} = \underline{a}_n$$

This iterative solution of the non-linear equations is terminated by a condition of the form

$$\left\| \underline{a}_n^{(r)} - a_n^{(r-1)} \right\| < \delta_n \qquad \ldots(4.8)$$

when we take $\underline{a}_n = \underline{a}_n^{(r)}$

A great variety of iterative methods for the solution of non-linear equations have been proposed, (see, for example, Ortega and Rheinboldt (1)). A very simple approach has been taken in the examples quoted in this chapter, though in certain special cases this has been modified. We take an initial estimate $\underline{a}_n^{(o)}$ of the solution vector and from the sequence of iterates $\underline{a}_n^{(r)}$ defined by

$$A_n \underline{a}_n^{(r)} = \underline{g}_n (\underline{a}_n^{(r-1)}) \qquad r=1,2.. \qquad \ldots(4.9)$$

until the condition (4.8) is satisfied. This simple iteration converges if

$$\rho \left(A_n^{-1} \cdot \left. \frac{\partial \underline{g}_n}{\partial \underline{a}_n} \right|_{\underline{a}_n = \underline{a}_n^r} \right) < 1 \qquad \ldots(4.10)$$

where $\rho(A)$ is the spectral radius of a matrix A. Though the condition (4.10) is stringent, good initial approximation vectors $\underline{a}_n^o$ are generally available from previous outer iterations. In the examples given the initial estimate

$$\underline{a}_n^{(o)T} = (\underline{a}_{n-1}^T, 0) \qquad \ldots(4.11)$$

has been used, though a more sophisticated approach based on extrapolation from the elements of the vector $\underline{a}_{n-1}$ might be used. In practice, if the coefficients $\underline{a}_n(i)$ are tending rapidly to zero there is little to be gained by a more elaborate choice than (4.11). It remains only to choose an initial estimate $\underline{a}_n^{(o)}$ for the first outer iteration. In this situation the matrix A is small (in our case (2x2)) and if necessary the relation (4.10) may be directly checked. In fact the method has proved very insensitive to the choice of initial vectors.

A more rapidly convergent iterative solution of the equations (4.4), and more sophisticated techniques of estimation of $\underline{a}_n^o$ may lead to significantly fewer evaluations of vectors $\underline{g}_n(\underline{a}_n^{r-1})$ in (4.9), with considerable gains in time where this vector is evaluated by a quadrature rule.

The relation (4.8) indicates that the termination condition for the inner iteration may depend on n, the number of basis functions used in the expansion of $y_n(x)$. It is felt that since the principal purpose of the first outer iterations is to provide good initial estimates for subsequent iterations the iterative solution of small Rayleigh-Ritz systems need not be as accurately performed as that for larger ones. Accordingly, in the results which are given δ_n has one of the following forms

$$\delta_n = n \cdot \varepsilon \, 10^{-n}$$

or

$$\delta_n = \varepsilon . 10^{-n}$$

where ε is a small positive constant and typically $\varepsilon = 5 \times 10^{-5}$ or $\varepsilon = 1 \times 10^{-5}$.

The inner iteration of the application of the Rayleigh-Ritz method to mildly non-linear problems is modified in certain special cases. These relate to the case where in (4.1) we have $k=1$, $p_1(x)=1$, $p_0(x)=0$; i.e. to differential equations of the form

$$\frac{d^2y}{dx^2} = f(x,y) \qquad \ldots(4.12)$$

and to particular choices of the basis functions $\phi_i(x)$ with which $y_n(x)$ is determined.

The simplest case is that in which the basis is given by

$$\phi_i(x) = \sqrt{2i+1} \int_0^x P_i^*(x)dx \qquad i=1,2.....$$

where $P_i^*(x)$ are the shifted Legendre polynomials. In this case the matrix A_n has elements given by

$$A_n(i,j) = \sqrt{2i+1}\ \sqrt{2j+1} \int_0^1 P_i^*(x)\, P_j^*(x)dx = \delta_{ij}$$

so that $A_n = I_n$, where I_n is the identity matrix of order n , so that equations (4.4) become

$$I_n \underline{a}_n = \underline{a}_n = g_n(\underline{a}_n)$$

and the iterative scheme (4.9) is reduced to

$$\underline{a}_n^{(r)} = g_n(\underline{a}_n^{(r-1)})$$

i.e. $$a_n^{(r)}(i) = g_n(\underline{a}_n^{(r-1)})(i)$$

with again a termination condition of the form (4.7).

This very simple Rayleigh-Ritz method has the disadvantage that evaluation of the functions $\underline{\phi}_i(x)$, best achieved from the recurrence given in § 3.4 (p.92) , is very time consuming. This will be clear from the subsequent discussion of test examples.

The second special case applies to problems of the form (4.12) when the chosen basis functions are either

$$\underline{\phi}_i(x) = x(1-x)\, T^*_{i-1}(x)$$

or

$$\underline{\phi}_i(x) = x(1-x)\, P^*_{i-1}(x)$$

where $T^*_i(x)$ are the shifted Tchebyshev polynomials on $[0,1]$ and $P^*_{i-1}(x)$ are the shifted Legendre polynomials. In either of these cases the matrix A_n assumes the special form indicated in Chapter Three in which

$$A_n(i,j) = 0 \qquad i+j \text{ odd.}$$

Hence A_n can be re-ordered so that A_n is reducible. Let

$$A^{(1)}_n(i,j) = A_n(2i-1, 2j-1) \qquad i,j = 1 \ldots \left[\frac{n+1}{2}\right]$$

and

$$A^{(2)}_n(i,j) = A_n(2i, 2j) \qquad i,j = 1 \ldots \left[\frac{n}{2}\right]$$

and define vectors $g^{(1)}_n$, $g^{(2)}_n$, u_n , v_n

$$\underline{g}^{(1)}_n(i) = g_n(2i-1) \qquad i = 1 \ldots \left[\frac{n+1}{2}\right]$$

$$\underline{g}^{(2)}_n(i) = g_n(2i) \qquad i = 1 \ldots \left[\frac{n}{2}\right]$$

and

$$\underline{u}_n(i) = \underline{a}_n(2i-1) \qquad i = 1 \ldots \left[\frac{n+1}{2}\right]$$

$$\underline{v}_n(i) = \underline{a}_n(2i) \qquad i = 1 \ldots \left[\frac{n}{2}\right]$$

where $[x]$ denotes 'integer part of x' . Then equations (4.4) can be rewritten

$$\left.\begin{aligned} A_n^{(1)} \underline{u}_n &= g_n^{(1)}(u_n, v_n) \\ A_n^{(2)} \underline{v}_n &= g_n^{(2)}(u_n, v_n) \end{aligned}\right\} \quad \ldots(4.13)$$

Starting from given initial estimates $\underline{u}_n^0$, $\underline{v}_n^0$ the equations (4.13) may be solved iteratively by the relations

$$\left.\begin{aligned} A_n^{(1)} \underline{u}_n^r &= g_n^{(1)}(\underline{u}_n^{r-1}, \underline{v}_n^{r-1}) \\ A_n^{(2)} \underline{v}_n^r &= g_n^{(2)}(\underline{u}_n^r, \underline{v}_n^{r-1}) \end{aligned}\right. \quad \ldots(4.14)$$

where each of the linear systems of equations for $\underline{u}_n^r$, $\underline{v}_n^r$ is solved by some numerical method, for example by Gaussian Elimination. This reduction to two connected systems of the form (4.14) provides considerable economy in the solution of the Rayleigh-Ritz equations for each value of n .

4.2 Some numerical results for mildly non-linear problems

In this section we report certain numerical experiments demonstrating the application of various choices of polynomial co-ordinate systems to the variational solution of mildly non-linear differential equations satisfying the restrictions of Ciarlet, Schultz and Varga (1). In section 4.4 examples which do not satisfy these conditions but which satisfy the weaker conditions of Gladwell (1) will be considered.

The test examples reported here are the following.

Problem N1.

The solution of the differential equation

$$y'' = \tfrac{1}{2}(y + x + 1)^3 \qquad y(0) = y(1) = 0$$

minimizes the variational integral

$$I(y) = \int_0^1 \left\{ y'^2 + \int_0^y (\eta + x + 1)^3 \, d\eta \right\} dx$$

subject to boundary conditions $y(0) = y(1) = 0$.

The exact solution is given by

$$y(x) = 2/(2-x) - x - 1 .$$

This example is considered by Ciarlet, Schultz and Varga (1, p.425).

Problem N2.

The solution of the differential equation

$$y'' = e^y \quad ; \quad y(0) = y(1) = 0$$

minimizes the variational integral

$$I(y) = \int_0^1 \left\{ y'^2 + 2\int_0^y \exp(\eta)\, d\eta \right\} dx$$

subject to $y(0) = y(1) = 0$. This solution is given by

$$y(x) = -\ln 2 + 2 \ln (c \sec (c(x-0.5)/2))$$

where $c \doteq 1.3360557$.

This example is considered also by Ciarlet, Schultz and Varga,(1, p.424).

Problem N3.

As a third test example we consider the differential equation

$$z'' = 6xz^2 \qquad z(0) = z(1) = 1 \qquad \ldots(4.15)$$

examined by Collatz (1, p.201), who shows that the operator T defined by

$$T(z) = z'' - 6xz^2$$

is monotone in the rectangle $0 \leq x \leq 1$, $0 \leq z(x) \leq 1$; i.e. if z

satisfies $T(z) = 0$ and $w_1(x)$, $w_2(x)$ satisfy

$$T(w_1(x)) < 0 , \quad T(w_2(x)) > 0$$

then the inequalities

$$w_1(x) < z(x) < w_2(x) \qquad \ldots(4.16)$$

hold. In particular the relation (4.16) is known to hold when

$$w_1(x) = 1 - x + x^3 , \quad w_2(x) = 1 - 0.43(x-x^4) .$$

The differential equation (4.15) is transformed by the substitution $y(x) = z(x) - 1$ to the form

$$y'' = x(y+1)^2 \qquad y(0) = y(1) = 0 \qquad \ldots(4.17)$$

Since $w_1(x) \leq z(x) \leq w_2(x)$, $y(x)$ satisfies

$$-x + x^3 \leq y(x) \leq -0.43(x-x^4)$$

The solution of (4.17) minimizes the variational integral

$$I(y) = \int_0^1 \left\{ y'^2 + 2 \int_0^y x(\eta+1)^2 \, d\eta \right\} dx$$

with boundary conditions $y(0) = y(1) = 0$.

We consider the approximate solution of these problems in terms of some of the polynomial co-ordinate systems we have used previously. The systems used are

$$\phi_i(x) = x^i(1-x) \qquad \ldots(4.18)$$

$$\phi_i(x) = x(1-x) \, T^*_{i-1}(x) \qquad \ldots(4.19)$$

$$\phi_i(x) = \sqrt{2i+1} \int_0^x P^*_i(t)dt \qquad \ldots(4.20)$$

We do not consider basis functions which may be obtained from these by simple diagonal scalings, nor do we consider the co-ordinate system

$\phi_i(x) = x(1-x)\, P^*_{i-1}(x)$ in view of its cumbersome computational properties. Approximations $y_n(x)$ in terms of each of the systems (4.18), (4.19), (4.20) for values of n in the range $2 \leq n \leq 10$ are considered. Numerical results for problems N1 and N2 using the basis (4.20) have been reported by Ciarlet, Schultz and Varga (1) for $n = 2,4,6$. All numerical experiments have been performed on an IBM 360/67 computer using either single precision (8 sig.figs.) or double precision (15 sig.figs.) arithmetic. A Gauss-Legendre quadrature formula using 20 points was used to evaluate the elements of the matrices A_n and the successive right hand side vectors of (4.9) or (4.14) respectively, except that, since for each of the test problems the matrix A_n generated from the basis (4.20) is the identity matrix I_n, this was substituted directly.

The results for problem N1 using each of the co-ordinate systems (4.18), (4.19),(4.20) are given in Tables XIX, XX, XXI respectively, whilst similar results are given for the problems N2, N3 in Appendix B. It is immediately apparent that there is no significant difference between the accuracies of the approximate solution produced in each co-ordinate system. The errors in the approximations obtained by Ciarlet, Schultz and Varga for this problem are indicated as e_{csv} in Table XXI, and we must attribute the minor difference to the use of different computing facilities.

The situation with respect to the determination of the coefficients of the approximate solution vector and the progress of the iterative procedure does, however, depend considerably on the choice of co-ordinate system from (4.18), (4.19), (4.20). It is not useful to compare the number of iterations performed for very small values of n, since this depends on the accuracy of the initial estimate $\underline{a}_n^0$, but once a reasonable approximation has been obtained a comparison can be made.

Behaviour of the Solution Vector and the Approximate Solution

Problem N1 $\phi_i(x) = x^i(1-x)$

n	ESTIMATE	2	3	4	5	6	10
$\bar{e}_n$		3.669807'-3	6.494669'-4	1.079375'-4	1.864391'-5	3.155990'-6	2.706643'-9
	1.0	-4.474689'-1	-5.128811'-1	-4.970389'-1	-5.006487'-1	-4.998626'-1	-4.999997'-1
	0	-4.677538'-1	-1.464643'-1	-2.881863'-1	-2.378296'-1	-2.535195'-1	-2.500140'-1
	0		-3.202393'-1	+9.348510'-3	-1.916721'-1	-9.765705'-2	-1.247145'-1
	0			-2.195192'-1	+8.168708'-2	-1.531579'-1	-6.523508'-2
$a_n(i)$	0				-1.505418'-1	-1.076669'-1	-1.673853'-2
	0					-1.032613'-1	-6.163793'-2
	0						+8.194446'-2
	0						-1.103734'-1
	0						+6.966051'-2
	0						-2.289152'-2
	ITERATIONS	13	10	10	8	7	16 20* 20* 20*

Long Precision $\delta_n = n.5._{10}-(5+n)$

Table XIX

Behaviour of the Solution Vector and the Approximate Solution

Problem N1 $\phi_i(x)=x(1-x)T^*_{i-1}(x)$

n	ESTIMATE	2	3	4	5	6	10
$\bar{e}_n$		3.731896'-3	6.494664'-4	1.079385'-4	1.775936'-5	3.155982'-6	2.706616'-9
	1.0	-6.813458'-1	-7.062307'-1	-7.062262'-1	-7.070771'-1	-7.070774'-1	-7.071067'-1
	0	-2.338768'-1	-2.333518'-1	-2.423185'-1	-2.423221'-1	-2.426300'-1	-2.426406'-1
	0		-4.002991'-2	-3.999128'-2	-4.157371'-2	-4.157398'-2	-4.163050'-2
	0			-6.859975'-3	-6.856142'-3	-7.132706'-3	-7.142664'-3
$a_n(i)$	0				-1.176108'-3	-1.175675'-3	-1.225432'-3
	0					-2.016823'-4	-2.102509'-4
	0						-3.602321'-5
	0						-6.180619'-6
	0						-1.017831'-6
	0						-1.746687'-7
	ITERATIONS	13	10	9	7	5	5 6 4 5

Long Precision $\delta_n=n.5_{10}-(5+n)$

Table XX

Behaviour of the Solution Vector and the Approximate Solution

Problem N1

$$\phi_i(x) = \sqrt{2i+1} \int_0^x P_i^*(t)\,dt$$

n	ESTIMATE	2	3	4	5	6	9+
$\bar{e}_n$		3.731930'-3	6.494546'-4	1.079387'-4	1.864398'-5	3.156062'-6	1.516039'-8
e_{csv}		3.76 '-3		1.10 '-4		3.29 '-6	
	1.0	+3.933752'-1	+3.938597'-1	+3.938864'-1	+3.938878'-1	+3.938878'-1	+3.938878'-1
	0	+1.045929'-1	+1.043581'-1	+1.044237'-1	+1.044275'-1	+1.044277'-1	+1.044277'-1
	0		+2.420781'-2	+2.418445'-2	+2.419309'-2	+2.419360'-2	+2.419362'-2
	0			+5.226647&-3	+5.223727'-3	+5.224903'-3	+5.224976'-3
$a_n(i)$	0				+1.080715'-3	+1.080318'-3	+1.080494'-3
	0					+2.169661'-4	+2.169352'-4
	0						+4.263098'-5
	0						+8.241889'-6
	0						+1.573495'-6
	ITERATIONS	12	9	8	8	5	6 5 5

Long Precision

$$n = n.5_{10}-(5+n)$$

Table XXI

We note in particular the poor convergence of the process when the basis (4.18) is used; an iteration limit of 20 iterations was imposed on the solution of the equations (4.9) or (4.14) at each stage, where this was reached an * has been placed in the tables for emphasis. That this situation can arise for the basis (4.18) is indicative of the poor convergence of the coefficients of the approximate solution vector when this basis is used, which can also be observed in the table. The convergence of the solution vector coefficients and of the iterative procedure for each of the co-ordinate systems (4.19) and (4.20) is good. We would emphasize again that solution in terms of the basis (4.20) is expensive of computer time because of the difficulty of evaluating the basis functions; as an indication we point out that the solutions quoted in Tables XIX, XX and XXI (including solutions for $n = 7,8,9$) were obtained in 124 secs, 173 secs and 300 secs respectively where the mark + in Table XXI indicates that solution of the system of size $n = 10$ in terms of the basis (4.20) had not been completed in 300 secs. Similar timing comparisons can be made for problems N2 and N3.

We compare briefly now the accuracy and numerical behaviour of solutions obtained using different precision of arithmetic; specifically we look at the solution of problem N1 in terms of the basis functions (4.18) and (4.19) using single precision (7 sig.figs.) and double precision (15 sig.figs.) arithmetic. Table XXII indicates that for the basis (4.19) the accuracy of the approximate solution is not severely affected by a reduction in the precision of the computations until the limit of this reduced accuracy is approached (the solution is of maximum magnitude $\doteq 1.1' - 1$) and it appears that the iterative solution of the equations is not adversely affected, though this behaviour is not continued; the iterative solution of the equations for $n = 9,10$ is not achieved using 20 iterations in single precision

Double Precision

n	2	3	4	5	6	7	8
e_n	3.7'-3	6.5'-4	1.1'-4	1.8'-5	3.2'-6	5.5'-7	9.1'-8
ITERS.	13	10	9	7	5	5	6

Single Precision

n	2	3	4	5	6	7	8
e_n	3.7'-3	6.5'-4	1.1'-4	1.8'-5	4.0'-6	1.1'-6	1.2'-6
ITERS.	13	10	7	6	4	5	3

The Effect of Single and Double Precision Arithmetic on the accuracy and number of iterations.

Problem N1 $\phi_i(x)=x(1-x)T^*_{i-1}(x)$

Table XXII

Double Precision

n	2	3	4	5	6	7	8
e_n	3.7'-3	6.5'-4	1.1'-4	1.9'-5	3.2'-6	5.5'-7	9.1'-8
ITERS.	13	10	10	8	7	16	20*

Single Precision

n	2	3	4	5	6	7	8
e_n	3.7'-3	6.5'-4	1.1'-4	1.9'-5	7.3'-6	2.0'-5	1.1'-5
ITERS.	15	20*	8	20*	20*	20*	20*

The Effect of Single and Double Precision Arithmetic on the accuracy and number of iterations.

Problem N1 $\phi_i(x) = x^i(1-x)$

Table XXIII

arithmetic. The effect of a reduction in the precision of the computation is much more marked when we consider solution in terms of the basis (4.18), as can be seen from the results summarized in Table XXIII. Although low order solutions are produced in single precision arithmetic which agree with those obtained using double precision, the convergence of the iteration is severely affected, and the iteration limit of 20 iterations is exceeded on several occasions. Furthermore, the iterative processes seem unlikely ever to converge; in Table XXIV we give the last six iterates of the process in the case $n = 6$.

We notice that the first and last rows of the Table, i.e. $\underline{a}_6^{15}$ and $\underline{a}_6^{(20)}$ are identical, and hence that the iteration will continue to oscillate, so that convergence will not occur.

It appears then that, although the solution of mildly non-linear problems in terms of the basis functions $\phi_i = x^i(1-x)$ may produce solutions equally accurate with those obtained using either the basis functions (4.19) and (4.20), and may be more economical of computer time, the numerical processes involved are unstable, and may lead to the non-convergence of iterative methods of solution. We would therefore recommend that either the basis functions (4.19) or (4.20) be used for mildly non-linear problems having solutions which may be well represented by polynomials, and express a preference on the grounds of computational convenience for the basis (4.19). Furthermore, we show in the next section that this basis generates matrices A_n with a useful theoretical property governing the convergence of iterative processes for the solution of equations. In § 4.4 we demonstrate that the numerical comparisons which we have made in this section remain valid for a wider class of problems.

Iterates $\underline{a}_6^r$, r = 15-20. Problem N1, $\phi_i(x) = x^i(1-x)$

r	$\underline{a}_6^r(1)$	$\underline{a}_6^r(2)$	$\underline{a}_6^r(3)$	$a_6^r(4)$	$a_6^r(5)$	$a_6^r(6)$
15	-5.001976'-1	-2.475424'-1	-1.306860'-1	-7.557272'-2	+2.643902'-2	-7.206231'-2
16	-5.002686'-1	-2.462569'-1	-1.380121'-1	-5.790004'-2	+7.537025'-3	-6.467634'-2
17	-5.005817'-1	-2.404756'-1	-1.705644'-1	+1.950022'-2	-7.423257'-2	-3.303568'-2
18	-5.002124'-1	-2.469784'-1	-1.348512'-1	-5.409252'-2	+1.312359'-2	-6.656747'-2
19	-5.004919'-1	-2.419709'-1	-1.627390'-1	+1.881488'-3	-5.639078'-2	-3.970960'-2
20	-5.001976'-1	-2.475424'-1	-1.306860'-1	-7.557272'-2	+2.643902'-2	-7.206231'-2

Table XXIV

4.3 Iterative Convergence : U.A.D. matrices.

In this section we outline a property of Rayleigh-Ritz approximation in terms of the basis

$$\phi_i(x) = x(1-x)\, T^*_{i-1}(x)$$

which helps to justify the use of this basis for the solution of certain types of variational problem. The initial result is the following.

Theorem 1

Let B be the matrix with elements

$$B(i,j) = \int_{-1}^{1} \phi_i'(x).\phi_j'(x)dx \qquad i,j = 1,2....$$

If $\phi_i(x) = (1-x^2)\, T_{i-1}(x)$, then the matrix B is asymptotically diagonal of degree three (3).

Proof.

The elements of the matrix B are given by

$$B(i+1,j+1) = \begin{cases} 0 & i+j \text{ odd} \\ r(i,j) & i+j \text{ even} \end{cases}$$

where

$$r(i,j) = -\frac{3}{\overline{i+j}^2-9} - \frac{1}{\overline{i+j}^2-1} - \frac{1}{\overline{i-j}^2-1} - \frac{3}{\overline{i-j}^2-9}$$

$$+ \frac{4\{(i^2-j^2)^2(2i^2+2j^2-20) + 9(\overline{i+j}^2 + \overline{i-j}^2)}{(\overline{i+j}^2-9)(\overline{i+j}^2-1)(\overline{i-j}^2-1)(\overline{i-j}^2-9)}$$

$$+ \frac{24i^2j^2(2i^2 + 2j^2 - 10)}{(\overline{i+j}^2-9)(\overline{i+j}^2-1)(\overline{i-j}^2-1)(\overline{i-j}^2-9)}$$

Hence

$$B(i+1,i+1) = O(i^2) + O(i^{-2}) + c$$

and for sufficiently large i, and fixed j,

$$B(i+1,i+1) \doteq k_i i^2$$

$$B(i+1,j+1) \doteq d_i\, i^{-2}$$

so that

$$\frac{|B(i,j)|}{\{B(i,i).\ B(j,j)\}^{\frac{1}{2}}} \doteq \frac{|d_i|\, i^{-2}}{\{(k_i i^2 + c_i)(k_j j^2 + c_j)\}^{\frac{1}{2}}}$$

$$\doteq \frac{|d_{ij}|\ .\ i^{-3}}{|k_i|.|B(j,j)|^{\frac{1}{2}}}$$

Hence the matrix B is asymptotically diagonal of degree 3. In order that B be uniformly asymptotically diagonal we require that there exist constants w, W such that

$$B(i+1,j+1) < W \qquad \forall_{i,j},\ i \neq j \qquad \ldots(4.21)$$

and

$$B(j+1,j+1) \geq w > 0 \qquad \forall_j \qquad \ldots(4.22)$$

To prove (4.22) we have

$$B(j+1,j+1) = r(j,j) = -\frac{3}{2j^2-9} - \frac{1}{2j^2-1} + 4/3$$

$$+ \frac{4j^2}{(2j^2-9)(2j^2-1)} + \frac{16j^4(j^2-5)}{3(2j^2-9)(2j^2-1)}$$

For $j > 2$ we have

$$B(j+1,j+1) \geq \frac{4}{3} - \frac{3}{2j^2-9} - \frac{1}{2j^2-1} \geq 16/17$$

and for $j = 0,1,2$,

$$B(1,1) = 8/3$$

$$B(2,2) = 48/21$$

$$B(3,3) = 116/21$$

so that (4.22) holds with $w = 16/17$. Clearly (4.21) holds in view of the asymptotic behaviour of $A(i,j)$, $i \neq j$, so that the theorem is proved.

Corollary 1

The Rayleigh-Ritz matrix A for the solution of variational problems of the form

$$\min_{y} I(y) = \int_0^1 \left\{ y'^2 + 2 \int_0^y f(x,\eta)\, d\eta \right\} dx$$

in terms of the basis $\phi_i(x) = x(1-x)T^*_{i-1}(x)$ is uniformly asymptotically diagonal of degree three.

Proof

The matrix A has elements

$$A(i,j) = \int_0^1 \phi'_i(x)\, \phi'_j(x)\, dx$$

which satisfy

$$A(i,j) = \tfrac{1}{8} B(i,j)$$

and from Delves & Mead (2), p.703, the property that a matrix is uniformly asymptotically diagonal is invariant under a diagonal transformation.

Corollary 2

The Rayleigh-Ritz matrices $A^{(1)}$, $A^{(2)}$ defined by

$$A^{(1)}(i,j) = A(2i-1,2j-1) \qquad i,j=1..\left[\frac{n+1}{2}\right]$$

$$A^{(2)}(i,j) = A(2i,2j) \qquad i,j=1..\left[n/2\right]$$

are asymptotically diagonal of degree three.

Proof

$$\frac{|A^{(1)}(i,j)|}{(|A^{(1)}(i,i)|.|A^{(1)}(j,j)|)^{\frac{1}{2}}} = \frac{A(2i-1,2j-1)}{A(2i-1,2i-1)\ A(2j-1,2j-1)}^{\frac{1}{2}}$$

Put $k = 2i-1$, $m = 2j-1$.

For fixed m

$$\frac{|A(k,m)|}{(|A(k,k)|.|A(m,m)|)^{\frac{1}{2}}} \leq C\,k^{-3}$$

from Theorem 1 and Corollary 1, and hence

$$\frac{|A(k,m)|}{(|A(k,k)|.|A(m,m)|)^{\frac{1}{2}}} \leq C\,k^{-3} = C(2i-1)^{-3}$$

so that

$$\frac{|A^{(1)}(i,j)|}{(|A^{(1)}(i,i)||A^{(1)}(j,j)|)^{\frac{1}{2}}} \leq C'\,i^{-3}$$

and similarly for the matrix $A^{(2)}$.

Thus the results of Delves and Mead, given in Chapter Two, are applicable when the basis $\phi_i(x) = x(1-x)T^*_{i-1}(x)$ is used to solve linear differential equations of the particularly simple type

$$y'' = f(x) \qquad y(0) = y(1) = 0$$

More importantly, we consider an extension of the results of Delves and Mead to the iterative solution of the Rayleigh-Ritz equations (4.4) or (4.13), i.e.

$$A_n\,\underline{a}_n = \underline{g}_n\,(\underline{a}_n) \qquad \ldots(4.23)$$

or

$$A_n^{(1)}\,\underline{u}_n = g_n^{(1)}\,(\underline{u}_n\ ,\ \underline{v}_n)$$

$$A_n^{(2)}\,\underline{v}_n = g_n^{(2)}\,(\underline{u}_n\ ,\ \underline{v}_n)$$

where

$$\underline{a}_n(2i-1) = \underline{u}_n(i) \qquad i=1 \,..\, \left[\frac{n+1}{2}\right]$$

$$\underline{a}_n(2i) = \underline{v}_n(i) \quad . \quad . \quad . \qquad i=1 \,..\, \left[n/2\right]$$

which determine an approximate solution of the form

$$y_n(x) = \sum_{i=1}^{n} a_n(i)\, \phi_i(x)$$

of the differential equation

$$y'' = f(x,y) \;:\; y(0) = y(1) = 0 \qquad \ldots(4.24)$$

where $g_n^{(1)}$, $g_n^{(2)}$ are defined in § 4.1. Such results clearly apply to the solution of differential equations of the form

$$y'' = h(x)y + g(x) \;:\; y(0) = y(1) = 0$$

provided that the linear term in y is included in the right hand side vector g_n of (4.22).

Let the solution $y(x)$ of (4.24) have a generalized Fourier expansion in terms of the co-ordinate system $\{\phi_i\}$ given by

$$y(x) = \sum_{i=1}^{\infty} \underline{\alpha}(i)\, \phi_i(x)$$

Let $R_s = \left\{ \underline{v} : \exists\ C \text{ s.t } \underline{v}(i) < Ci^{-s} \right\}$

We can now prove the following.

Theorem 2

Let A be a U.A.D. matrix of degree $p > 1/2$, satisfying $|A_{ii}| = 1$ and $A_{ij} < C(p)i^{-p}$ where $C(p)$ is defined by Delves and Mead (2, Thm.1). Let $\underline{a}^{(o)}$ be a chosen initial vector such that $\underline{g}(\underline{a}^{o}) \in R_s$ and define the sequence of vectors $\underline{a}^{(r)}$ $r = 1,2 \ldots$ by

$$A\,\underline{a}^{(r)} = \underline{g}\,(a^{(r-1)}) = \underline{g}^{(r-1)} \qquad \ldots(4.25)$$

If $\underline{g}^{(r-1)} \in R_{s_r}$, $s_r > 0$

then $a^{(r)} \in R_t$ for all r

where $t = \min (s_r , p)$.

Proof

For each r we solve the linear system of equations

$$A \underline{a}^{(r)} = \underline{g}^{(r)}$$

where A is a U.A.D. matrix of degree p and $\underline{g}$ is a vector in R_{s_r} . Hence by Delves and Mead (2, Thm.6) $\underline{a}^{(r)} \in R_{t_r}$ where $t_r = \min (p,s_r)$. Hence the sequence of vectors $\underline{a}^{(r)}$ are all contained in R_t where $t = \min_r (t_r)$. Q.E.D.

Corollary

If the iterative solution of (4.25) is convergent then $\underline{a} \in R_t$.

We note several consequences of these results. The condition $A(i,i) = 1$ does not apply in the case of the Rayleigh-Ritz matrix obtained from (4.24) by substitution of an approximation in terms of the co-ordinate system $\phi_i(x) = x(1-x)T_{i-1}(x)$. However, we have shown that $|A(i,i)| = C_i i^2$ for some constant C_i, and it is known (Delves and Mead, (2,p.703)) that the U.A.D. properties of a matrix are invariant under a diagonal transformation. Thus we can consider a diagonal transformation by the matrix

$$D : D(i,i) = (C_i\, i^2)^{-1}$$

such that $A(i,i) = 1$.

A major weakness of Theorem 2 is the assumption

$$\underline{g}\,(\underline{a}^{(r-1)}) \in R_{s_r} \quad \forall\, r$$

which appears difficult to verify. However, we consider the system

$$DA\underline{a}^{(r)} = D\underline{g}(\underline{a}^{(r-1)}) \qquad \ldots(4.26)$$

where A is the Rayleigh-Ritz matrix for (4.24) with the basis $\phi_i(x) = x(1-x)T^*_{i-1}(x)$.

Writing $DA = A'$, $D\underline{g} = \underline{g}'$ we have the system

$$A'\underline{a}^{(r)} = \underline{g}'^{(r-1)} \qquad \ldots(4.27)$$

where A' is a U.A.D. matrix of degree 3 with $A'(i,i) = 1$, and where $\underline{g}'^{(r-1)} \in R_2$ provided there exists a constant M s.t $\underline{g}^{(r-1)}(i) < M \ \forall \ i$. Hence if A' satisfies the remaining condition of Theorem 2, i.e. $|A'(i,j)| < C(3)i^{-p}$ then $\underline{a}^{(r)}$ defined by (4.27)satisfy $\underline{a}^{(r)} \in R_2$. The elements $\underline{g}^{(r-1)}(i)$ are defined by

$$\underline{g}^{(r-1)}(i) = \int_0^1 f(x, \sum_{j=1}^{\infty} \underline{a}^{(r-1)}(j)\,\phi_j(x)) \cdot \phi_i(x)dx$$

so that

$$\underline{g}^{(r-1)}(i) \leq \left\{ \left[\int_0^1 f(x, \sum_{j=1}^{\infty} a^{(r-1)}(j)\phi_j(x)) dx \right]^2 \cdot \left[\int_0^1 \phi_i(x)dx \right]^2 \right\}^{\frac{1}{2}}$$

by the Cauchy Inequality. We can now state

Theorem 3.

Let $\{\phi_i(x)\}$ be the co-ordinate system

$$\phi_i(x) = x(1-x)T^*_{i-1}(x) \qquad \ldots(4.28)$$

and define W to be the set of functions

$$W = \{w(x): \ f(x,w) < M_1 \ , \ w(x) \in H_L\}$$

and suppose

$$y^{(r)}(x) = \sum_{i=1}^{\infty} \underline{a}^{(r)}(i)\phi_i(x) \ , \ y(x) = \sum_{i=1}^{\infty} \underline{\alpha}(i)\phi_i(x)$$

satisfy $y^{(r)} \in W$, $r = 0,1\ldots$, $y \in W$

where $\underline{a}^{(r)}$ are the solutions of equations (4.26). Then $\underline{a}^{(r)} \in R_2$ for all r .

Proof.

We have established $a^{(r)} \in R_2$ provided $\left|g^{(r-1)}(i)\right| < M$.

Since, for $\phi_i(x)$ given by (4.28)

$$\left[\int_0^1 \phi_i(x)dx\right]^2 \leq \text{const} = C_1$$

then for all $w(x) \in W$ $\exists\, M = C_1 M_1$

such that

$$g^{(r-1)}(i) < M .$$

Since $y^{(r-1)}(x) \in W$, we have $a^{(r)} \in R_2$

Corollary.

If the iterative solution of equations (4.27) is convergent, then $\alpha \in R_2$.

The results given above indicate the behaviour of the coefficients of the generalized Fourier expansion of the solution $y(x)$ of problem (4.23) in terms of the cc-ordinate system (4.28), and also the behaviour of the successive iterates in an iterative solution of the (infinite) system of equations (4.27). Nothing has been said concerning the situation for finite matrices, or concerning the particular iterative process used. It would seem probable that results similar to Delves and Mead (1, Thm., p.212) can be given for particular iterative methods of solution of a finite system of equations. This result asserts that the elements of the first iterate $\underline{a}_n^{(1)}$ of the Gauss-Seidel or Jacobi methods

of solution of a finite system of equations have the same asymptotic behaviour as the exact solution of the corresponding infinite system of linear equations provided that the matrix A is sufficiently asymptotically diagonal (p large enough).

It is felt that the results of this section account in part for the high reliability of the variational solutions in terms of the basis (4.28) quoted in this chapter. It can also be shown that the Rayleigh-Ritz matrix in terms of the basis

$$\phi_i(x) = x(1-x)\, P^*_{i-1}(x) \qquad \ldots(4.29)$$

is uniformly asymptotically diagonal of degree 3 and thus that Theorem 3 holds in the case of (4.29) as well as in the case of (4.28). Finally in this section we suggest that the results given here lend weight to the optimistic comment of Delves and Mead (2, p.25) that A.D and U.A.D matrices may be of considerable importance in variational calculations.

4.4 Numerical Results for q-bounded problems.

The conditions given by Ciarlet, Schultz and Varga (1), which apply to the numerical examples considered in § 4.2, have recently been extended by Gladwell (1), who shows that a wider class of problems can be solved using the Rayleigh-Ritz method. We consider now numerical results obtained by the Rayleigh-Ritz method for the solution of two examples given by Gladwell which satisfy his conditions but not those of Ciarlet, Schultz and Varga. Additionally we demonstrate the success of the Rayleigh-Ritz method for certain particular problems which are not covered by either the requirements of Gladwell or Ciarlet, Schultz and Varga.

The examples considered are

Problem Q1.

The solution of the differential equation

$$y'' = \tfrac{1}{2}(y + x + 1)^3 - \frac{1}{x(1-x)}\left(y - \frac{2}{2-x} + x + 1\right)$$

subject to $y(0) = y(1) = 0$ is given by

$$y(x) = {}^{2}/_{(2-x)} - x - 1 .$$

This solution minimizes the variational integral

$$I(y) = \int_0^1 \left\{ y'^2 - 2\int_0^y \left[\frac{1}{x(1-x)}\left(\eta - \frac{2}{2-x} + x + 1\right) - \tfrac{1}{2}(\eta + x + 1)^3 \right] d\eta \right\} dx$$

subject to the boundary conditions $y(0) = y(1) = 0$.

Problem Q2.

The second problem we consider is similar to that given by Gladwell (1, p.62). Specifically, we replace the derivative boundary condition given there, $y'(0) = 0$ by the condition $y(0) = 0$. Then, defining

$$\lambda = \frac{e}{2(1-e)} \quad , \quad y_0(x) = \cos\left(\lambda(1-e^{-x})\right)$$

the differential equation

$$(e^x y')' = e^x - \lambda^2 e^{-x}(K(y-x+1) + (1-K)y_0(x) + (y_0(x) - y + x - 1)^{2i+1} \qquad \ldots(4.30)$$

(K = const , $i = 1, 2, 3\ldots$)

with the boundary conditions $y(0) = y(1) = 0$ has the solution $y(x) = y_0(x)+x-1$. The equation (4.30) has the form of (4.1) but the right hand side of (4.30) does not satisfy the conditions of Ciarlet, Schultz and Varga (1), but in the case $K < e^2/4$ it satisfies

the extended conditions of Gladwell. We consider the solution of this problem for a number of values of K, including some cases $K > e^2/4$, for which a theoretical justification of the method has not been given. We also consider the case in which the exponent $2i+1$ occurring in (4.30) is replaced by $2i$, for which the problem again does not satisfy Gladwell's conditions.

The solution of (4.30) minimizes the variational integral

$$I(y) = \int_0^1 \left\{ e^x y'^2 - 2 \int_0^y \left[\lambda^2 e^{-x} (K(\eta - x + 1) + (1-K) y_0(x) + (y_0(x) - (\eta - x + 1))^{2i+1} \right] d\eta \right\} dx \qquad \ldots(4.31)$$

subject to the boundary conditions $y(0) = y(1) = 0$.

This example is interesting from an additional viewpoint, in that it is the only example of a mildly non-linear equation which we shall consider for which $p_1(x) \neq 1$, so that the elements of the matrices A_n are not so readily evaluated.

We consider only solution of the problems Q1 and Q2 in terms of the basis $\phi_i(x) = x(1-x)\, T^*_{i-1}(x)$. Solutions of problem Q1 are given in Table XXV, and of problem Q2 for the values $i = 1$, $K = 0.5$ in Table XXVI. Solutions of problem Q2 for other values of K and i are given as Tables BVII, BVIII, BIX in Appendix B. No difficulties are encountered and both the solution vectors and the approximate solutions display stable and convergent behaviour. For problem Q2 it can be seen that the convergence of the iterative solution of the equations depends on K and i. Specifically, as K is increased towards $e^2/4$, or as i decreases towards zero, more iterations are required for the determination of each approximate solution vector $\underline{a}_n$. This reflects the behaviour of the integrand

Behaviour of the Solution Vector and the Approximate Solution

Problem Q1 $\quad \phi_i(x) = x(1-x)T^*_{i-1}(x)$

n	ESTIMATE	4	5	6	7	8	10
$\bar{e}_n(x)$		1.082485'-4	1.775997'-5	3.156666'-6	5.522174'-7	9.081012'-8	2.730828'-9
	-0.5	-7.062235'-1	-7.070770'-1	-7.070774'-1	-7.071058'-1	-7.071058'-1	-7.071067'-1
	-0.25	-2.423171'-1	-2.423219'-1	-2.426300'-1	-2.326301'-1	-2.426403'-1	-2.426406'-1
	0	-3.998988'-2	-4.157364'-2	-4.157396'-2	-4.162868'-2	-4.162869'-2	-4.163050'-2
	0	-6.860071'-3	-6.856057'-3	-7.132702'-3	-7.132725'-3	-7.142345'-3	-7.142664'-3
	0		-1.176115'-3	-1.175668'-3	-1.223749'-3	-1.223751'-3	-1.225432'-3
$a_n(i)$	0			-2.016829'-4	-2.016291'-4	-2.099582'-4	-2.102509'-4
	0				-3.458997'-5	-3.458312'-5	-3.602322'-5
	0					-5.932915'-6	-6.180621'-6
	0						-1.017834'-7
							-1.746698'-7
	ITERATIONS	9	5	4	4	4	4 4

Long Precision $\quad \delta_n = n5_{10}-(3+n)$

Table XXV

Behaviour of the Solution Vector and the Approximate Solution

Problem Q2 : K = 0.5, i = 1 $\quad \phi_i(x) = x(1-x)T^*_{i-1}(x)$

n	ESTIMATE	2	3	4	5	6	10
$e_n(x)$		1.261579'-2	1.208426'-3	7.948973'-5	4.423306'-5	9.148378'-6	2.617497'-9
$a_n(i)$	1.0	+2.772160'-1	+3.453961'-1	+3.474917'-1	+3.471648'-1	+3.470793'-1	+3.470030'-1
	0	-4.806177'-1	-5.190254'-1	-5.338607'-1	-5.336877'-1	-5.330513'-1	-5.329933'-1
	0		+1.081963'-1	+1.117820'-1	+1.111715'-1	+1.110113'-1	+1.108642'-1
	0			-1.114156'-2	-1.100610'-2	-1.043511'-2	-1.038202'-2
	0				-4.434962'-4	-5.634499'-4	-6.924685'-4
	0					+4.063258'-4	+4.476025'-4
	0						-9.081614'-5
	0						+1.133954'-5
	0						-7.539886'-7
	0						-4.670775'-8
	ITERATIONS	8	6	7	6	6	5 5 4 4

Long Precision $\quad \delta_n = n.5.10^{-(5+n)}$

Table XXVI

of the problem in terms of the criteria given by Gladwell. For example, if $K > e^2/4$ then the integrand of (4.31) does not satisfy these criteria, and we have no theoretical justification for the use of the Rayleigh-Ritz method. However, we find that in this case the method proves satisfactory. Table XXVII gives Rayleigh-Ritz approximations to the solution of problem Q2 for $K = 2$, $i = 1$ obtained in terms of the basis $\phi_i(x) = x(1-x)T^*_{i-1}(x)$.

If the exponent $2i+1$ occurring in the right hand side of (4.30) is replaced by $2i$, $i = 1,2,3\ldots$ then the variational formulation of this problem:-

Problem Q2'

$$\min_y \int_0^1 \left\{ e^x y'^2 - 2 \int_0^y \left[\lambda^2 e^{-x} (K(\eta - x + 1) + (1-K)y_0(x) + (y_0(x) - (\eta - x + 1))^{2i} \right] d\eta \right\} dx$$

subject to $y(0) = y(1) = 0$

also fails to satisfy the criteria of Gladwell (1) for any value of K, and again there is no theoretical justification for the use of the Rayleigh-Ritz method. In the particular case $K = 0.5$, $i = 1$, however, we demonstrate the success of the method in Table XXVIII in which the basis $\phi_i(x) = x(1-x)T^*_{i-1}(x)$ is used to obtain a sequence of Rayleigh-Ritz approximations.

4.5 Conclusions

Numerical results for certain examples of mildly non-linear differential equations have been presented which show that for suitable choices of polynomial co-ordinate system accurate solutions can be obtained in terms of a small number of basis functions

Behaviour of the Solution Vector and the Approximate Solution

Problem Q2 : K = 2, i = 1 $\quad \phi_i(x) = x(1-x)T^*_{i-1}(x)$

n	ESTIMATE	2	3	4	5	6	10
$e_n(x)$		1.268371'-2	1.380878'-3	8.381422'-5	4.545780'-5	-9.248011'-6	-2.665064'-9
	1.0	+2.793716'-1	+3.458856'-1	+3.474726'-1	+3.471604'-1	+3.470786'-1	+3.470030'-1
	0	-4.769798'-1	-5.194710'-1	-5.333396'-1	-5.336774'-1	-5.330506'-1	-5.329933'-1
	0		+1.080319'-1	+1.117693'-1	+1.111672'-1	+1.110101'-1	+1.108642'-1
	0			-1.114026'-2	-1.100174'-2	-1.043468'-2	-1.038202'-2
$a_n(i)$	0				-4.422355'-4	-5.640330'-4	-6.924684'-4
	0					+4.061073'-4	+4.476025'-4
	0						-9.081603'-5
	0						+1.133950'-5
	0						-7.539300'-7
	0						-4.669007'-8
	ITERATIONS	20	14	17	14	14	12 12 10 6

Long Precision $\quad S_n = n.5_{10}-(5+n)$

Table XXVII

Behaviour of the Solution Vector and the Approximate Solution

Problem Q2' : k = 0.5, i = 1 $\quad \phi_i(x) = x(1-x)T^*_{i-1}(x)$

n	ESTIMATE	2	3	4	5	6	10
$e_n(x)$		1.261316'-2	1.208479'-3	7.948959'-5	4.423302'-5	9.148381'-6	2.617497'-9
	1.0	+2.772362'-1	+3.453963'-1	+3.474917'-1	+3.471648'-1	+3.470793'-1	+3.470030'-1
	0	-4.806289'-1	-5.190255'-1	-5.338607'-1	-5.336877'-1	-5.330513'-1	-5.329933'-1
	0		+1.081963'-1	+1.117820'-1	+1.111715'-1	+1.110113'-1	+1.108642'-1
	0			-1.114156'-2	-1.100610'-2	-1.043511'-2	-1.038202'-2
	0				-4.434962'-4	-5.634499'-4	-6.924685'-4
$a_n(i)$	0					+4.063258'-4	+4.476025'-4
	0						-9.081614'-5
	0						+1.133954'-5
	0						-7.539886'-7
	0						-4.670775'-8
	ITERATIONS	8	6	7	6	6	5 5 4 4

Long Precision $\quad \delta_n = n.5_{10}-(5+n)$

Table XXVIII

and that the coefficients of such approximations display good convergence and stability properties. A theoretical justification of the convergence rate of the generalized Fourier expansion of the solution has been given for a particular co-ordinate system. Additionally, in this special case it has been shown that the Rayleigh-Ritz matrix assumes a special form which permits decomposition into two matrices of approximately half size; the consequent gains in the elimination and back-substitution steps of the solution of equations (4.14) are a factor of 4 and 2 respectively at each iteration.

We contrast this situation with the justification of the admittedly powerful techniques of piecewise polynomial co-ordinate systems given by Schultz (1, p.303).

"Indeed, were it not for the fact that sparseness (of the Rayleigh-Ritz matrix) is so important and orthonormalization so difficult, we would always use polynomial-type subspaces."

We must point out that the above remark was made in the wider context of elliptic boundary value problems. Nevertheless, the impression that orthonormality and/or sparseness are desirable properties for the application of the Rayleigh-Ritz method to the one-dimensional problems we have considered is emphasized by the use in the papers of Ciarlet, Schultz and Varga (1) of the basis

$$\phi_i(x) = \sqrt{2i+1} \int_0^x P_i^*(t)\, dt \qquad \ldots(4.32)$$

which is an orthonormal basis for problems of the form $y'' = f(x,y)$. We feel that it has been demonstrated that this emphasis is in some ways unnecessary, and that other choices of co-ordinate system may be equally satisfactory. In particular it has been shown that the

co-ordinate system

$$\phi_i(x) = x(1-x)\, T^*_{i-1}(x)$$

has several useful and important theoretical and computational properties.

We would also like to point out the application of the Rayleigh-Ritz method to examples of the class of problems considered by Gladwell (1), and the examples which are not included in that theory. Clearly, further extension of the assumptions on the form of the non-linear term $f(x,y)$ of the mildly non-linear differential equations is possible. The situation in respect of the application of the Rayleigh-Ritz method to mildly non-linear differential equations is a familiar one in numerical analysis; there exist theoretical criteria which guarantee that the method may be applied with success to certain problems, and there exist certain other problems to which the method can be applied but for which applications there is no theoretical justification.

Chapter Five

Errors of quadrature and approximation

In this chapter we outline and develop a number of results necessary for a full understanding of the numerical experiments and results of the previous chapters. These concern the behaviour and effect of two kinds of error implicit in the numerical solution of linear and mildly non-linear differential equations by the application of the Rayleigh-Ritz method.

First we consider the errors introduced into the Rayleigh-Ritz process by the approximation of certain integrals which arise, e.g.

$$A_n(i,j) = (\phi_i, \phi_j)_L = \sum_{s=0}^{k} \int_0^1 p_s(x)\, \phi_i^{(s)}(x)\, \phi_j^{(s)}(x)\, dx \qquad \ldots(5.1)$$

$$i,j = 1 \ldots n$$

and

$$b_n(i) = (f, \phi_i) = \int_0^1 f(x)\, \phi_i(x)\, dx \qquad \ldots(5.2)$$

$$i = 1 \ldots n$$

or

$$b_n(i, \underline{a}^{(r-1)}) = \int_0^1 f(x, \sum_{j=1}^{n} \underline{a}_n^{(r-1)}(j)\, \phi_j(x)).\phi_i(x)dx \qquad \ldots(5.3)$$

$$i = 1 \ldots n$$

by finite quadrature sums. That is, we take

$$A_n(i,j) = \sum_{s=0}^{k} \sum_{t=1}^{m_1} w_t^{(1)}\, p_s(x_t^{(1)})\, \phi_i^{(s)}(x_t^{(1)})\, \phi_j^{(s)}(x_t^{(1)}) \qquad \ldots(5.4)$$

and

$$b_n(i) = \sum_{t=1}^{m_2} w_t^{(2)}\, f(x_t^{(2)})\, \phi_i\, (x_t^{(2)}) \qquad \ldots(5.5)$$

$$i = 1 \ldots n$$

or

$$b_n^{(r)}(i) = \sum_{t=1}^{m_3} w_t^{(3)} \; f(x_t^{(3)}, \sum_{j=1}^{n} a_n^{(r-1)}(i)\, \phi_j(x_t^{(3)}))\, \phi_i(x_t^{(3)}) \qquad \ldots(5.6)$$

$$i = 1 \ldots n$$

where $(\underline{w}_t^{(1)}, \underline{x}_t^{(1)})$, $(\underline{w}_t^{(2)}, \underline{x}_t^{(2)})$, $(w_t^{(3)}, x_t^{(3)})$ are the weights and abscissae of certain quadrature formulae. Expressions (5.4), (5.5), (5.6) indicate that there is no need to use the same quadrature rule on each of the many integrands occurring in the Rayleigh-Ritz method, and much economy might be gained in this way. The quadrature rule used might be made to depend on i,j,n in each of (5.1), (5.3), (5.3), and a separate quadrature rule could be applied to each term in the summation over s in (5.1), and different rules for each value of r in the iterative evaluation of (5.3). In fact, in the numerical experiments reported previously the same quadrature rule has been used to approximate each of the expressions (5.1), (5.2), (5.3), and the theoretical analysis which we give will rely on this assumption.

The second type of errors which we must consider are those most important for an assessment of the Rayleigh-Ritz method. That is, we seek estimates and/or bounds for the accuracy of approximate solutions; e.g. relationships of the form

$$\left| y_n^m(x) - y(x) \right| \leq \varepsilon \qquad 0 \leq x \leq 1 \qquad \ldots(5.7)$$

or

$$\left| y_n^m(x) - y(x) \right| = O(n^{-i}) + O(m^{-j}) \qquad \ldots(5.8)$$

where $y(x)$ is the solution of the given differential equation and $y_n^m(x)$ a Rayleigh-Ritz approximation dependent on the number of basis functions used, n , and the number of points in the quadrature

rule, m . We shall see that such results are not readily available, and in general we have only relations of the form

$$\left|y_n(x) - y(x)\right| \leq \varepsilon \qquad 0 \leq x \leq 1 \qquad \ldots(5.9)$$

or

$$\left|y_n(x) - y(x)\right| = O(n^{-i})$$

5.1 Consistent Quadrature Schemes.

The application of quadrature schemes to the evaluation of the right hand side vectors (5.2) and (5.3) has been considered by Herbold (1), Herbold, Schultz and Varga (1), and criteria determined which allow the selection of quadrature rules which preserve the rate of convergence of the approximate solutions as n increases, for the case where the basis functions are piecewise continuous on a given sequence of meshes; such consistent schemes sometimes generate approximations which coincide with finite difference approximations. We examine the requirements for consistent quadrature schemes for piecewise approximation so that we can consider their relevance for global approximations; and comment briefly on a relationship with finite difference methods.

In this discussion of quadrature approximations we use a notation similar to that used by Lyness (2). We denote by $Q^m_{m_o}[\sigma(x),a,b]$ the quadrature approximation

$$\int_a^b \sigma(x)\,dx \doteq Q^m_{m_o}[\sigma(x),a,b] = \sum_{t=1}^{m_o} w_t\,\sigma(x_t)$$

and write $Q^m_{m_o}[\sigma(x)]$ when a = 0, b = 1 . We define the algebraic (trigonometric) degree of $Q^m_{m_o}[\sigma,a,b]$.

$Q^m_{m_o}[\sigma(x),a,b]$ is said to be a quadrature rule of exact

algebraic (trigonometric) degree m iff

$$\int_a^b \sigma(x)\,dx - Q_{m_o}^{m}[\sigma(x),a,b] = 0 \qquad \ldots(5.10)$$

whenever $\sigma(x)$ is an algebraic (trigonometric) polynomial of degree $\leq m$, but (5.10) does not hold if $\sigma(x)$ is an arbitrary polynomial of higher degree.

We use the notation $Q_{m_o}[\sigma(x),a,b]$ to denote a quadrature approximation of unspecified degree m $(m \geq 0)$. No assumption is made concerning the weights w_t or abscissae x_t of the quadrature rule, though we shall in general take $a \leq x_t \leq b$, and on occasion require $w_t > 0$. With the notation here defined we can write

$$\underline{b}_n^{m_o}(i) = Q_{m_o}[f(x).\phi_i(x)] \doteq \underline{b}_n(i)$$

and

$$b_n^{m_o}(i, \underline{a}_n) = Q_{m_o}\left[f(x, \sum_{j=1}^{n} \underline{a}_n(j)\,\phi_j(x).)\phi_i(x)\right] \doteq \underline{b}_n(i,\underline{a}_n)$$

We can then define the Rayleigh-Ritz solution $y_n(x)$ and a quadrature approximated Rayleigh-Ritz solution $y_n^{m_o}(x)$ by

$$y_n(x) = \sum_{i=1}^{n} \underline{a}_n(i)\,\phi_i(x)$$

and

$$y_n^{m_o}(x) = \sum_{i=1}^{n} \underline{a}_n^{m_o}(i)\,\phi_i(x)$$

where in the linear case $\underline{a}_n$ and $\underline{a}_n^{m_o}$ satisfy the equations

$$A_n\,\underline{a}_n = \underline{b}_n$$

and

$$A_n\,\underline{a}_n^{m_o} = \underline{b}_n^{m_o}$$

respectively, and in the non-linear case the equations

$$A_n \; \underline{a}_n = \underline{b}_n (\underline{a}_n)$$

and

$$A_n \; \underline{a}_n^{m_o} = \underline{b}_n^{m_o} (\underline{a}_n^{m_o})$$

The notion of a consistent quadrature scheme for the application of the Rayleigh-Ritz method with piecewise continuous basis functions is the following. Let $y(x)$ be the solution of the given differential equation, and assume that in the given approximation subspace we have results of the form

$$\left\| y(x) - y_n(x) \right\|_\gamma \leq K_1 \, (\bar{\bar{\Pi}})^{-\nu} \qquad \ldots(5.11)$$

where $\bar{\bar{\Pi}} = \max_{1 \leq i \leq N} (x_i - x_{i-1})$ is the usual parameter of a partition $\bar{\Pi}$ and $\| \cdot \|_\gamma$ is a norm related to the given problem, (cf. Ch.Two and Ciarlet, Schultz, Varga (1)). Then the quadrature scheme Q_{m_o} is consistent in the given approximation subspace if we have results of the form

$$\left\| y_n(x) - y_n^{m_o}(x) \right\|_\gamma \leq K_2 \, (\bar{\bar{\Pi}})^{-k}$$

where $k \geq \nu$, since then the convergence rate of the true and calculated quadrature-approximated solutions is governed by

$$\left\| y(x) - y_n^{m_o}(x) \right\|_\gamma \leq K \, (\bar{\bar{\Pi}})^{-\nu}$$

and the convergence rate of (5.11) is preserved.

In the case of piecewise approximation a particular type of quadrature rule, the composite quadrature rule, is important. Given a quadrature rule $Q_{m_1} [\sigma, \alpha, \beta]$ and a partition $\bar{\Pi} = (a = x_o < x_1 < \ldots < x_N = b)$ the corresponding composite rule is

$$Q_{m_o} [\sigma(x), a, b, \bar{\Pi}] = \sum_{i=0}^{N-1} Q'_{m_1} [\sigma(x), x_i, x_{i+1}] \qquad \ldots(5.12)$$

where $Q'_{m_1}[\sigma(x), x_i, x_{i+1}]$ $i = 0 \ldots N-1$ are obtained from $Q_{m_1}[\sigma, \alpha, \beta]$ by a linear transformation. Such rules include the familiar composite trapezium and Simpson's rules, and certain other rules, including some of those considered by Herbold, Schultz and Varga (1) which are based on interpolation. We say that:-

A composite quadrature rule $Q^m_{m_o}[\sigma(x), a, b, \overline{\Pi}]$ is of exact piecewise polynomial degree m with respect to a partition $\overline{\Pi}$ iff

$$\left| Q^m_{m_o}[\sigma(x), a, b, \overline{\Pi}] - \int_a^b \sigma(x)dx \right| = 0 \qquad \ldots(5.13)$$

for all functions $\sigma(x)$ which are piecewise polynomial functions of degree $k \leq m$ on each interval (x_i, x_{i+1}), $i = 0..N-1$, with possible discontinuities only at points x_i $i=0..N$, but (5.13) does not hold for arbitrary piecewise polynomials of degree $k > m$.

We summarize results of Herbold, Schultz and Varga (1, Thms. 4,5, Corollary p.113) on consistent quadrature schemes for mildly non-linear differential equations.

Theorem

Let C be any collection of quasi-uniform partitions $\overline{\Pi}: 0 = x_1 < x_2 \ldots < x_{N+1} = 1$, and for each $\overline{\Pi}$, let $S_n(\overline{\Pi})$ be a finite dimensional subspace of H_L consisting of polynomial L-splines such that $v(x) \in S_n(\overline{\Pi})$ implies $v(x)$ is a polynomial of degree at most n_o on each subinterval defined by $\overline{\Pi}$, and suppose there exists a set of linearly independent functions $w_j(x)$, $j = 1..n$ which form a basis for $S_n(\overline{\Pi})$. Let $f(x,v(x))$ satisfy $\frac{\partial^s f}{\partial x^s}(x,v(x))$ is continuous in each subinterval of $\overline{\Pi}$ for $0 \leq s \leq m_2$, and let $y(x)$, the solution of the given equation satisfy $y(x) \in C^{n_o+1}[0,1]$.

Let $Q'_{m_1}[\sigma(x), x_i, x_{i+1}, \overline{\Pi}]$ be a quadrature rule of polynomial

degree at least one satisfying $w_i \geq 0$ and (for each i)

$$\left| Q_{m_1}^{1} \left[G(x), x_i, x_{i+1}, \overline{\Pi} \right] - \int_{x_i}^{x_{i+1}} G \quad (x)dx \right| < K_o h_i^{m_2+1} \left\| D^{m_o} G \right\|_\infty$$

and let $Q_{m_o}^{2n_o} \left[G, 0, 1, \overline{\Pi} \right]$ be a quadrature rule of the form (5.12) of piecewise polynomial degree at least $2n_o$.

Then there exists a positive constant K such that

$$\left| y_n^{2n_o}(x) - y_n(x) \right| \leq K \left(\overline{\Pi} \right)^{-r} \qquad \ldots(5.14)$$

where $r = \min \left(m_2 - n_o + k - 1 , n_o - k + 1 \right)$

and therefore the quadrature scheme $Q_{m_o}^{2n_o} \left[G(x), 0, 1, \overline{\Pi} \right]$ is consistent with the appropriate L-spline bound given in § 2.6 if $m_2 \geq 2 + 2n_o - 2k$.

A number of the conditions of this Theorem may be relaxed when the given differential equation is linear. The conditions on the polynomial and piecewise polynomial degree of the quadrature formula, which ensure a unique solution of the non-linear Rayleigh-Ritz equations, may be dispensed with as may the restriction that the partitions $\overline{\Pi} \in C$ are quasi-uniform. The restriction to subspaces consisting of polynomial L-splines, which is necessary to ensure the boundedness of derivatives of the quantity

$$\left\| \frac{d^{m_o}}{dx^{m_o}} \left\{ f(\cdot, w)(w - y_n(x)) \right\} \right\|_{L_\infty \left[x_j, x_{j+1} \right]}$$

for all $0 \leq j \leq N$ and arbitrary $w \in S_n(\overline{\Pi})$ can also be omitted, so that the result holds for arbitrary piecewise polynomial subspaces and arbitrary partitions $\overline{\Pi}$ in the linear case.

Herbold, Schultz and Varga (1) show that if the simple piecewise linear basis

$$w_i(x) = \begin{cases} (x-x_i)/h_{i-1} & x_{i-1} \le x \le x_i \\ (x_{i+1} - x)/h_i & x_i \le x \le x_{i+1} \\ 0 & \text{otherwise} \end{cases} \qquad \ldots(5.15)$$

is used to solve the equation

$$y'' = f(x)$$

$$y(0) = y(1) = 0$$

then a suitable consistent quadrature is the trapezium rule approximation to the integrals

$$\int_{x_{i-1}}^{x_i} f(x)w_i(x)dx \quad , \quad \int_{x_i}^{x_{i+1}} f(x)w_i(x)dx$$

If $h_i = [x_{i+1} - x_i] = h$, a constant, this scheme yields the usual finite difference approximation

$$y_{i-1} - 2y_i + y_{i+1} = h^2 f(x_i)$$

to the differential equation $y'' = f(x)$, and the approximation extends to

$$y_{i-1} - 2y_i + y_{i+1} = h^2 f(x_i, y_i) \qquad \ldots(5.16)$$

in the non-linear case.

For the equation

$$y'' + ky = f(x) \quad , \quad k = \text{const}$$

and the basis (5.15) the resulting Rayleigh-Ritz equations have the form

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h} + \frac{hk}{6}\left[y_{i-1} + 4y_i + y_{i+1}\right] = \int_{x_{i-1}}^{x_{i+1}} f(x)w_i(x)dx \qquad \ldots(5.17)$$

so that, if the right hand side of (5.17) is approximated by Simpson's Rule, we are led to

$$y_{i-1} - 2y_i + y_{i+1} = \frac{h^2}{6}\left[f(x_{i-1},y_{i-1}) + 4f(x_i,y_i) + f(x_{i+1},y_{i+1})\right] \quad \ldots(5.18)$$

instead of (5.16). It can be shown (Appendix C) that the schemes (5.16) and (5.18) have local truncation errors of the form αh^2 and $-\alpha h^2$, and hence there is no advantage (and the disadvantage of additional computation) in using (5.18). That the schemes must have truncation errors of equal magnitude and opposite sign can easily be seen, however, directly from the forms of (5.16) and (5.18). Let $\underline{y}^{(1)}$ be a solution of (5.16) and $\underline{y}^{(2)}$ of (5.18). Then the vector $\bar{\underline{y}} = \frac{1}{2}(\underline{y}^{(1)} + \underline{y}^{(2)})$ satisfies the equations

$$\bar{y}_{i-1} - 2\bar{y}_i + \bar{y}_{i+1} = \frac{h^2}{2}\left[f(x_i,\bar{y}_i) + \frac{1}{6}\left(f(x_{i-1},\bar{y}_{i-1}) + 4f(x_i,\bar{y}_i) + f(x_{i+1},\bar{y}_{i+1})\right)\right]$$

$$= \frac{h^2}{12}\left[f(x_{i-1},\bar{y}_{i-1}) + 10f(x_i,\bar{y}_i) + f(x_{i+1},\bar{y}_{i+1})\right] \quad \ldots(5.19)$$

and the scheme (5.19) is the well known Mehrstellenverfahren* which has local truncation error $O(h^4)$. A full truncation error analysis of the scheme (5.18) together with numerical examples are given in Appendix C.

The criteria of consistent quadrature schemes are not directly relevant to global approximation in view of the different nature of the basis functions and the different parameterization of the approximation subspace. An appropriate definition of a consistent quadrature scheme for global approximation might be the following.

The quadrature schemes $Q_{m_0(n)}^{m(n)}$ n=1,2.. form a sequence of quadrature schemes consistent with the bound

$$\left\| y_n - y \right\|_{\infty} \leq Kn^{-\nu} \qquad (\S\ 2.6)$$

*See Collatz (1, p.168).

if there exists a constant K_1 such that

$$\left\| y_n - y_n^{m_o} \right\|_\infty \leq K_1 n^{-\nu} \qquad \ldots(5.20)$$

Attempts to derive characterization theorems for consistent quadrature schemes of this type have foundered. Much of the proof of Herbold, Schultz and Varga (1), Thm.5, remains valid, but certain auxiliary results from the theory of L-spline approximation are not applicable to polynomial approximation, and we are unable to proceed further along these lines. One useful result which emerges from this study, however, is the following application of Herbold, Schultz, Varga (1), Thm.4, (which we state below in an appropriate form) to the use of global polynomial approximations.

Theorem

Given any finite dimensional subspace S_n of H_L, let $\phi_i(x)$, $i = 1..n$ be a basis of this space, and let $Q_{m_o}^m \left[\sigma(x),0,1\right]$ be a quadrature rule of polynomial degree at least zero satisfying

$$Q_{m_o}^m \left[\phi_i \cdot \phi_j, 0, 1\right] - \int_0^1 \phi_i \cdot \phi_j \, dx = 0 \qquad \ldots (5.20') \qquad 1 \leq i, j \leq n$$

and $w_t > 0 \quad 1 \leq t \leq m_o$.

Then there exists a unique approximate solution $y_n^{m_o}(x)$ of the problem

$$Ly = \sum_{s=0}^{k} (-1)^s \frac{d^s}{dx^s} \left(p_s(x) \frac{d^s y}{dx^s}\right) = f(x,y)$$

subject to $y^\ell(0) = y^\ell(1) = 0, \quad \ell = 0..k-1$ defined by

$$y_n^{m_o}(x) = \sum_{i=1}^{n} \underline{a}_n^{m_o}(i) \, \phi_i(x)$$

where

$$A_n \, \underline{a}_n^{m_o} = \underline{b}_n^{m_o} \left(\underline{a}_n^{m_o}\right)$$

and

$$A_n = (\phi_i, \phi_j)_L \quad , \quad b_n^{m_o}\ (a_n^{m_o}) \text{ is defined by (5.10).}$$

We deduce the following.

Corollary.

Let S_n be a finite dimensional subspace of H_L consisting of functions $v(x)$ which are polynomials of degree at most n, and let $\{\phi_i\}_{i=2k}^{n}$ be a basis of this space. Then the above result holds if $Q_{m_o}^{m}\ [\alpha(x), 0,1]$ is an $n+1$ point Gauss-Legendre quadrature rule.

In order to investigate numerically the effect of quadrature approximations to

$$\int_0^1 f(x, y_n(x)).\ \phi_i(x)dx$$

we consider the differential equation

$$y'' = (y + x + 1)^3 \qquad \ldots(5.21)$$
$$y(0) = y(1) = 0$$

previously considered in Chapter Four, and estimate the constant ν of (5.20) for Gauss-Legendre quadrature approximations using different numbers of points. Equation (5.21) is particularly suited to this computation since for a given m point Gauss-Legendre quadrature rule the elements $A_n(i,j)$ $i,j=1..m$ are evaluated exactly (except for rounding errors) but the elements $b_n^{m_o}\ (a_n^{m_o})\ (1)$ only for $i=1,\ldots\frac{2m-1}{3}$. We use here the basis $\phi_i(x) = x(1-x)\ T_{i-1}^{*}\ (x)$.

The known solution of (5.21) is given by $y(x) = 2/(2-x)-x-1$, which satisfies $y(x) \in C^t\ [0,1]$ for all $t > 0$. In this case it is well known (e.g. Ciarlet, Schultz, Varga (1), p.406) that the classical Rayleigh-Ritz approximation in terms of a global polynomial basis

converges exponentially to the true solution, i.e. there exists a constant μ , $0 \leq \mu < 1$ such that

$$\lim_{n \to \infty} \left(\left\| y_n(x) - y(x) \right\|_\infty \right)^{1/n} \leq \mu \qquad \text{for all } n$$

In Figs. VI, VII we indicate graphically the behaviour of

$$e_n^{m_o} \doteq \left\| y_n^{m_o}(x) - y(x) \right\|_\infty$$

for $m_o = 8$, 20 and $n = 2 \ldots$, $\min(m_o-1,10)$ by plotting $-\ln e_n^{m_o}$ against n . Notice that $m_o = 20$ generates, except for rounding error, the classical Rayleigh-Ritz approximation, since all quadratures are in this case exact, and it is clear from Fig.VI that exponential convergence is obtained. In the case $m_o = 8$ the classical Rayleigh-Ritz approximation is not generated for n=6, n=7, but it is clear from Fig.VII that the same exponential convergence prevails.

We consider also the example

$$y'' - \exp(y) = 0 \quad ; \quad y(0) = y(1) = 0$$

with the known solution

$$y(x) = -\ln 2 + 2 \ln \, c \sec \left(c(x-\tfrac{1}{2})/2\right)$$

The classical Rayleigh-Ritz approximation in terms of a polynomial co-ordinate system is in this case not generated by a Gaussian quadrature for any value of n . Yet it is clear (Fig.VIII) that each of the subsequences

$$y_{2i-1}^{m_o}(x) \, , \quad y_{2i}^{m_o}(x) \, , \quad i = 1,2,3$$

converge exponentially, whilst it is believed that the breakdown of this in the case $y_7^{m_o}(x)$ is attributable to the approaching limits of machine accuracy.

We remark that in no case has the determination of a 'quadrature approximated' Rayleigh-Ritz solution $y_n^{m_o}(x)$ $n \geq m_o$ of a mildly

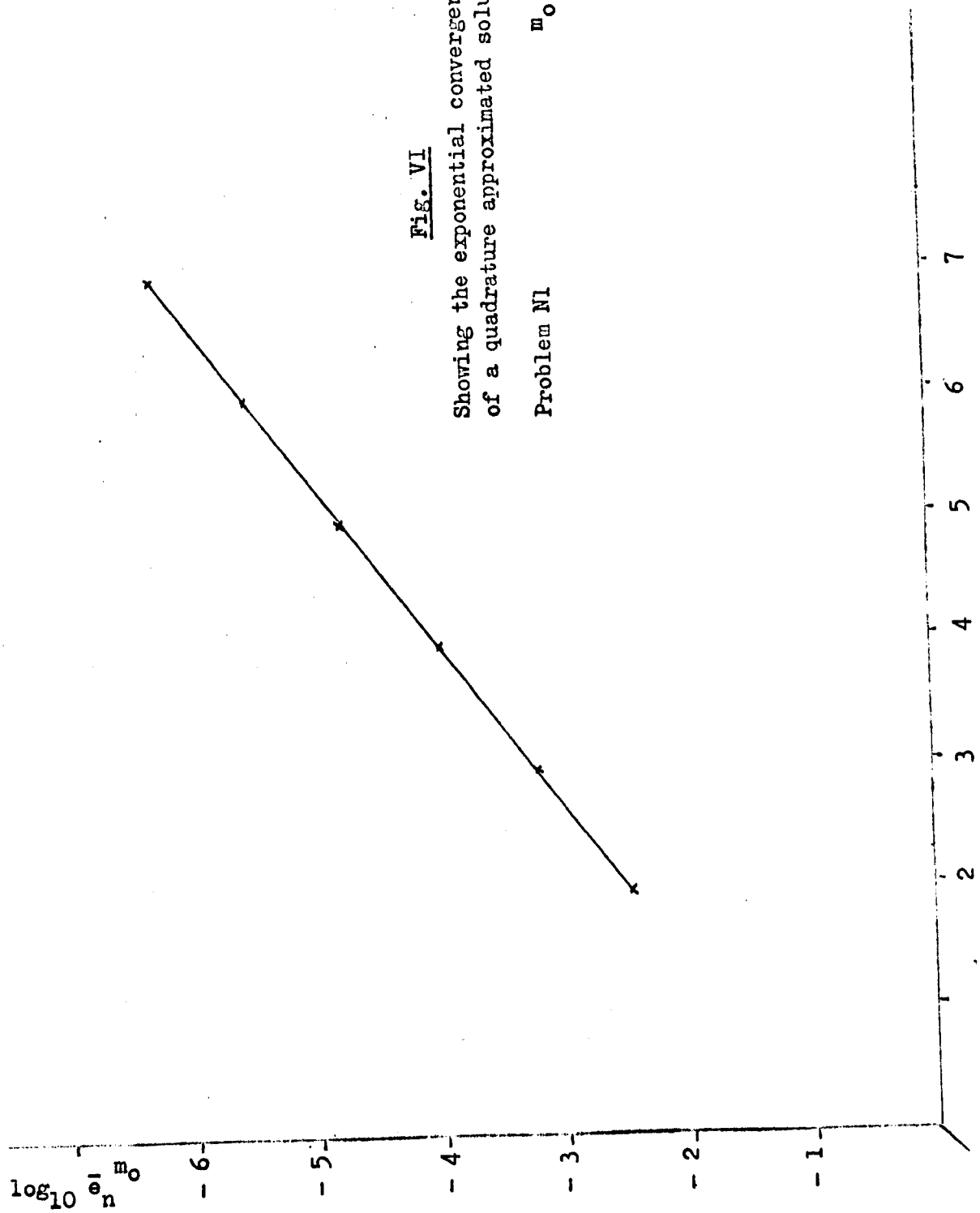


Fig. VI

Showing the exponential convergence of a quadrature approximated solution

Problem N1 $\quad m_o = 8$

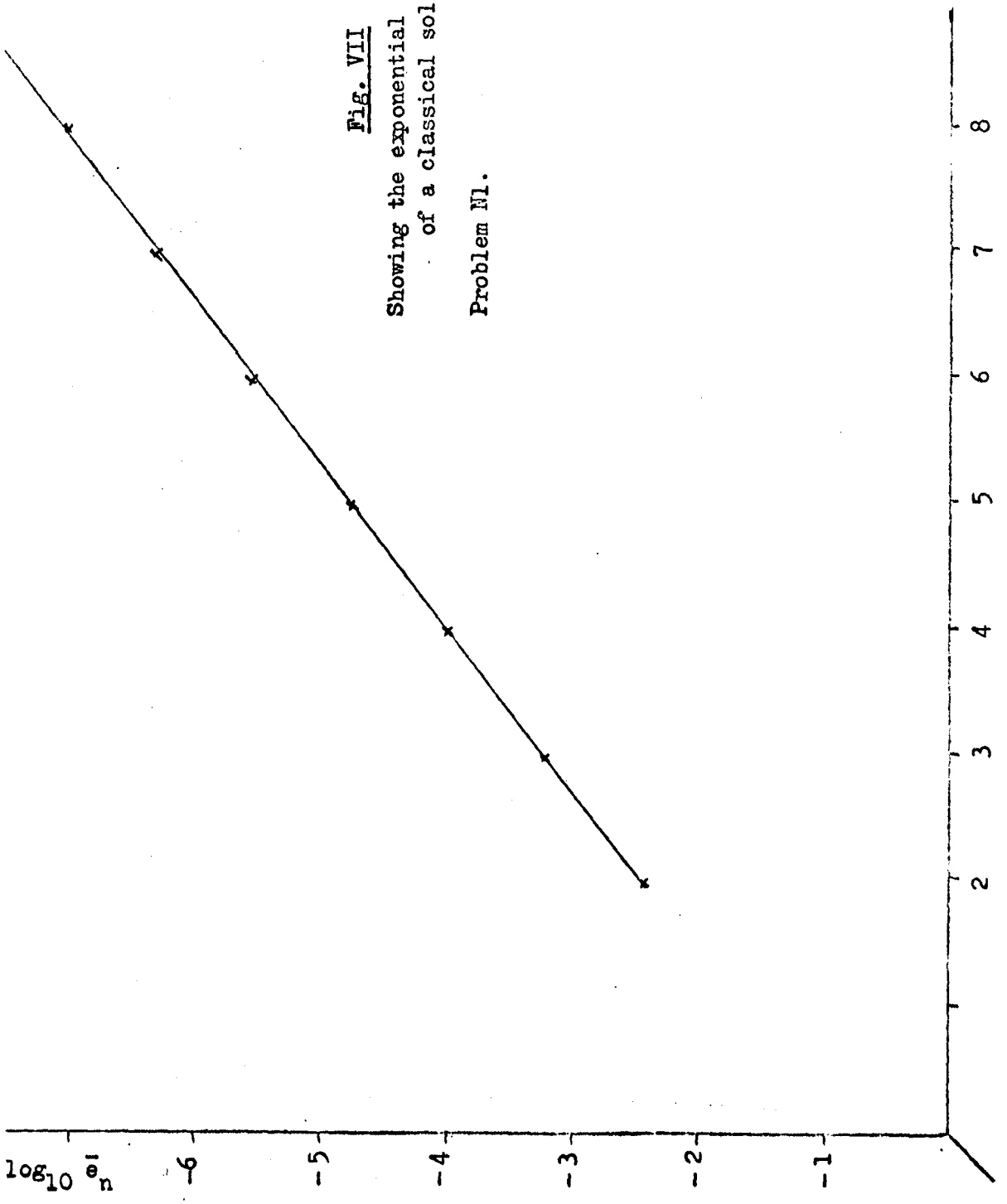


Fig. VII

Showing the exponential convergence of a classical solution

Problem M1.

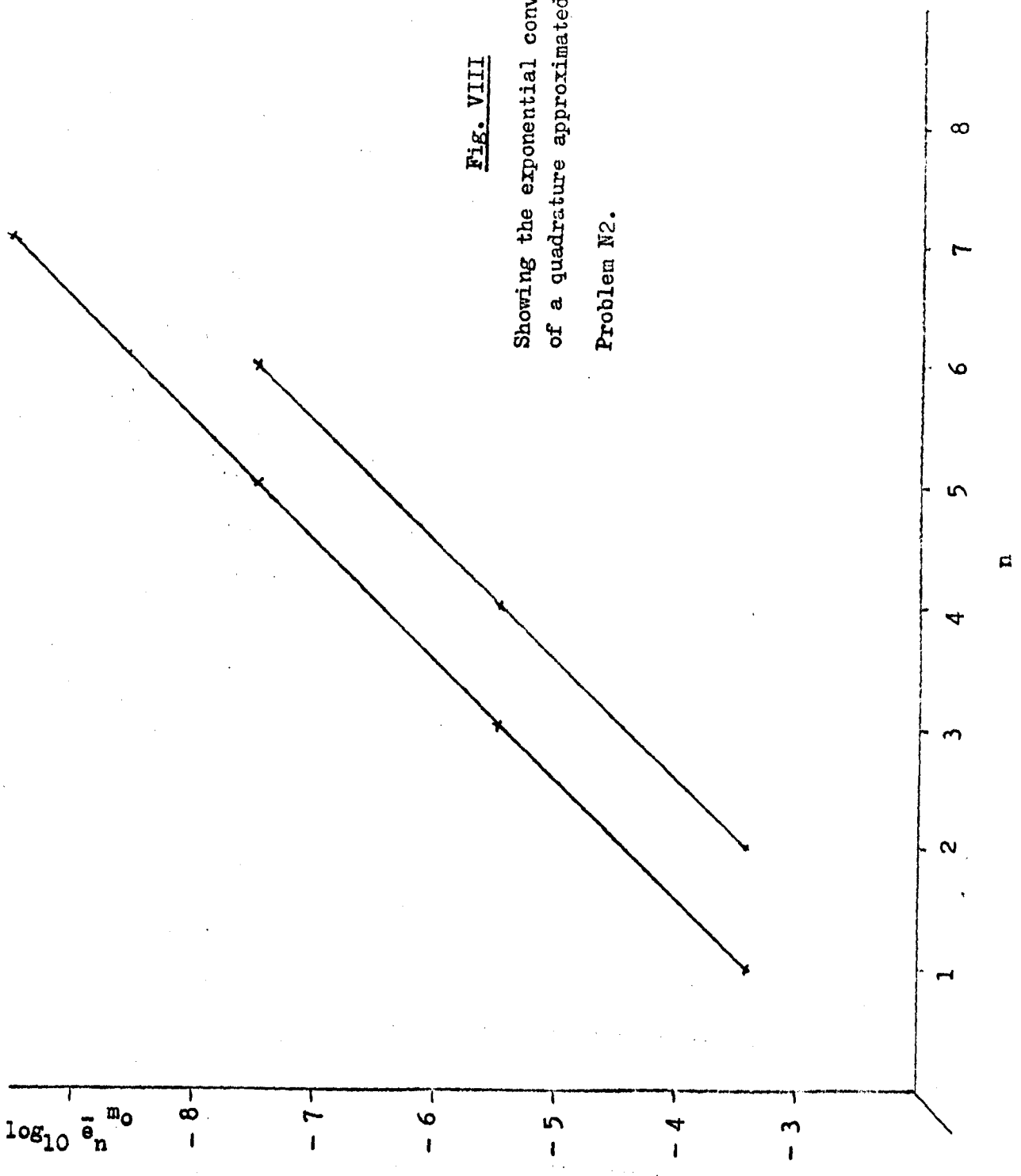


Fig. VIII

Showing the exponential convergence of a quadrature approximated solution

Problem N2. $m_o = 8$

non-linear differential equation proved successful, and accordingly it may be that this condition for the uniqueness of a solution given in the above theorem is not only sufficient, as proved, but also necessary.

5.2 General Quadrature Approximation

We have already noted that we must consider not only quadrature approximations to the elements of the right hand side vectors arising in the Rayleigh-Ritz method, but also to the elements of the matrix A, that is, to expressions $(\phi_i, \phi_j)_L$ where

$$(\phi_i, \phi_j)_L = \int_0^1 \sum_{s=0}^{k} p_s(x) \frac{d^s\phi_i}{dx^s} \frac{d^s\phi_j}{dx^s} dx$$

We recall the notation

$$I(y) = \int_0^1 \left[\sum_{s=0}^{k} p_s(x) \left(\frac{d^s y}{dx^s}\right)^2 + 2 \int_0^y f(x,\eta) d\eta \right] dx$$

and introduce

$$I_{m_o}^{m}(y) = Q_{m_o}^{m}\left[\sum_{s=0}^{k} p_s(x) \left(\frac{d^s y}{dx^s}\right)^2 + 2 \int_0^y f(x,\eta) d\eta, 0, 1 \right]$$

$$= \sum_{j=0}^{m_o} w_j \left[\sum_{s=0}^{k} p_s(x_j) \left(\frac{d^s y}{dx^s}\Big|_{x=x_j}\right)^2 + 2 \int_0^y f(x,\eta) d\eta \Big|_{x=x_j} \right]$$

These definitions will be of use, in an extended form, in Chapter Six. If the expression $y_n^{m_o}(x) = \sum a_n^{m_o}(i)\, \phi_i(x)$ is substituted above the necessary conditions for a minimum of the resulting expression, i.e.

$$\frac{d\, I_{m_o}^{m}(y_n^{m_o})}{d\, a_n^{m_o}(i)} = 0$$

generate the equations

$$A_n^{m_o} \; \underline{a}_n^{m_o} = \underline{b}_n^{m_o} (\underline{a}_n^{m_o})$$

where $A_n^{m_o}(i,j)$, $b_n^{m_o}(\underline{a}_n^{m_o})(i)$ are given by expressions of the form (5.4) and (5.6) respectively. We wish to prove that for a given subspace S_n of H_L spanned by the basis $\{\phi_i\}_{i=1}^n$ there exists a unique function $y_n^{m_o}(x)$ such that

$$I_{m_o}^m (y_n^{m_o}(x)) \leq I_{m_o}^m (y) \qquad \forall y \in S_n$$

We prove first a number of auxiliary results. A number of the assumptions of Ciarlet, Schultz and Varga (1) are given in a discretized form. We assume there exist constants β, $K > 0$ independent of n such that

$$\| u \|_\infty^{m_o} = \sup_{\substack{x=x_t \\ t=0..m_o}} |u(x)| \leq K \, Q_{m_o}^m \left[\sum_{s=0}^{k} p_s(x) \left(\frac{d^2 u(x)}{dx^s} \right)^2 + \beta \, (u(x))^2 \right]$$

for all $u(x) \in S_n$

and introduce Λ'_n defined by

$$\Lambda'_n = \inf_{\substack{u \in S_n \\ u \not\equiv 0}} \frac{Q_{m_o}^m \left[\sum_{s=0}^{k} p_s(x) \left(\frac{d^s u}{dx^s} \right)^2 \right]}{Q_{m_o}^m \left[u(x)^2 \right]} \qquad \ldots(5.22)$$

and $\Lambda' = \inf_n \Lambda'_n$.

We now establish Lemma 1 and Lemma 2, the proofs of which are similar.

Lemma 1

Let $\gamma > -\Lambda'$. Then

$$g[u] = \left(Q^{m}_{m_o} \left[\sum_{s=0}^{k} p_s(x) \left(\frac{d^s u}{dx^s} \right)^2 + \gamma (u(x))^2 \right] \right)^{\frac{1}{2}} \qquad \ldots(5.23)$$

is a seminorm on S_n provided $w_t \geq 0 \quad t = 0,1..m_o$.

Lemma 2

Let $S_n = P_n$ where P_n is a subspace of H_L consisting of polynomials of degree at most n . Then the expression (5.23) is a norm on P_n provided $w_t > 0$, $t = 0,1..m_o$, and $m_o > n$.

Proof of Lemma 2

We must verify

i) $g[u] \geq 0 \qquad \forall u \in P_n$

ii) $g[cu] = c\, g[u] \qquad c = \text{constant}$

iii) $g[u] = 0 \quad \text{iff} \quad u \equiv 0$

iv) $g[u+v] \leq g[u] + g[v]$

i) follows immediately from the restrictions $w_t > 0$, $\gamma > -\Lambda'$, and (5.22), ii) is a consequence of the definition of $Q^{m}_{m_o}$, iii) follows since $u \in P_n$, $u \not\equiv 0$ is such that $\left(\frac{d^s u}{dx^s} \right)^2$ has at most $n-s$ distinct zeros $(s = 0,1..k)$ and hence $w_t > 0 \quad m_o > n$ are sufficient to ensure $g[u] > 0$, $u \not\equiv 0$ whilst $g[u] = 0$, $u \equiv 0$ is immediately obvious. To prove iv), squaring both sides gives

$$g[u+v]^2 \leq g[u]^2 + g[v]^2 + 2\, g[u].g[v]$$

which reduces to

$$Q^{m}_{m_o} \left[\sum_{s=0}^{k} p_s(x) \frac{d^s u}{dx^s} . \frac{d^s v}{dx^s} + \gamma . u(x) . v(x) \right]$$

$$\leq Q^{m}_{m_o} \left[\sum_{s=0}^{k} p_s(x) \left(\frac{d^s u}{dx^s} \right)^2 + \gamma u^2 \right]^{1/2} . Q^{m}_{m_o} \left[\sum_{s=0}^{k} p_s(x) . \left(\frac{d^s v}{dx^s} \right)^2 + \gamma v^2 \right]^{1/2} \qquad \ldots(5.24)$$

Defining the vectors

$$W_r = w_t\, p_o(x_t) + \gamma, \qquad r = 1 \ldots m+1$$

$$W_r = w_t\, p_s(x_t), \qquad r = m_o + 2 \ldots R, \quad s \geq 1$$

$$U_r = \left.\frac{d^s u}{dx^s}\right|_{x=x_t} \qquad r = 1,2 \ldots R$$

$$V_r = \left.\frac{d^s v}{dx^s}\right|_{x=x_t} \qquad r = 1,2 \ldots R$$

where $r = ts + 1$, $R = (m_o + 1).(k + 1)$

(5.24) can be expressed as

$$\left(\sum_{r=1}^{R} W_r U_r V_r\right)^2 \leq \left(\sum_{r=1}^{R} W_r U_r^2\right).\left(\sum_{r=1}^{R} W_r V_r^2\right)$$

which has the form of a finite Cauchy-Buniakowski inequality (Liusternik and Sobolev (1)). Hence we may write

$$\|u\|_{\gamma}^{m_o} = g[u].$$

The proof of Lemma 1 is identical except that iii) is not required, and for general spaces S_n is not necessarily true.

Corresponding to Ciarlet, Schultz and Varga (1, Lemma 1, p.396) we have

Lemma 3

Let $S_n = P_n$. Then

$$\Lambda' \geq \frac{1}{k^2} - \beta$$

which follows from the definition of Λ' and since

$$\|u\|_{\infty}^{m_o} \leq \|u\|_{L_2}^{m_o} \qquad \text{where}$$

$$\|u\|_{L_2}^{m_o} = \left(\varrho_{m_o}^{m}\,[u^2(x)]\right)^{\frac{1}{2}}$$

Theorem.

Let $$\frac{\partial f(x,v)}{\partial u} \geq \gamma > -\Lambda \qquad \ldots(5.25)$$

and let $S_n = P_n$, let $m_o > n$ and $w_t > 0$, $t = 0, .. m_o$. Then there exists a unique function $y_n^{m_o}(x)$ which minimizes $I^m_{m_o}(y)$ over P_n .

Proof.

$$I^m_{m_o}(u_n) = \|u_n\|^{m_o}_o + Q^m_{m_o}\left[2\int_0^{u_n} f(x,\eta)\,d\eta\right]$$

where $$u_n = \sum_{i=1}^{n=2k+1} a_n(i)\,\phi_i(x) \qquad \in P_n$$

Using (5.25), clearly (e.g. Ciarlet, Schultz, Varga (1), p.396)

$$I^m_{m_o}(u_n) \geq \|u_n\|^{m_o}_o + Q^m_{m_o}\left[\gamma u_n^2(x)\right]$$

$$= \|u_n\|^{m_o}_\gamma$$

Hence, as $I^m_{m_o}$ is a continuous function on the finite dimensional subspace P_n , bounded below by 0 and satisfying

$$\lim_{\|u\|\to\infty} I^m_{m_o}(u) = +\infty$$

for any norm $\|u\|$ on P_n , since all norms on this finite subspace are equivalent, a standard compactness argument shows that there exists at least one $\bar{u}_n$ such that

$$I^m_{m_o}(\bar{u}_n) \leq I^m_{m_o}(u_n) \quad \text{for all } u \in P_n$$

To prove that $\bar{u}_n$ is unique, we continue to follow the arguments of Ciarlet, Schultz, Varga((1), p.398-9). $I^m_{m_o}(u_n)$ is twice differentiable with respect to the elements $a_n(i)$. We have

$$\frac{\partial I^{m}_{m_o}(u_n)}{\partial a_n(i)} = Q^{m}_{m_o}\left[\sum_{s=0}^{k} p_s(x) \sum_{j=1}^{n-2k+1} a_n(j) \frac{d^s \phi_j}{dx^s} \frac{d^s \phi_i}{dx^s}\right]$$

$$+ Q^{m}_{m_o}\left[f(x, \sum_{j=1}^{n} a_n(j)\, \phi_j(x)).\phi_i(x)\right]$$

and $1 \leq i \leq n-2k+1$

$$\frac{\partial^2 I^{m}_{m_o}(u_n)}{\partial a_n(i)\partial a_n(j)} = Q^{m}_{m_o}\left[\sum_{s=0}^{k} p_s(x) \frac{d^s}{dx^s}\phi_i(x) \frac{d^s \phi_j}{dx^s}\right]$$

$$+ Q^{m}_{m_o}\left[\frac{\partial f}{\partial u}(x, \sum_{l=1}^{n-2k+1} a_n(l)\phi_l).\phi_i.\phi_j\right] \quad \ldots(5.26)$$

We define $B^{m_o}(u_n)$ to be the n-2k+1 x n-2k+1 matrix with elements

$$b_{ij} = \frac{\partial^2 I^{m}_{m_o}(u_n)}{\partial a_n(i)\partial a_n(j)} \qquad i,j = 1 .. n-2k+1$$

and show that $B^{m_o}(u_n)$ is uniformly positive definite; i.e. that for any functions

$$u_n = \sum_{i=1}^{n-2k+1} a_n(i)\, \phi_i(x), \quad y_n = \sum_{i=1}^{n-2k+1} \alpha_n(i)\phi_i(x)$$

$\in P_n$ there exists a positive constant c such that

$$\underline{\alpha}_n^T B^{m_o}(u_n)\, \underline{\alpha}_n \geq c\, \underline{\alpha}_n^T . \underline{\alpha}_n \qquad \ldots(5.27)$$

But substituting (5.26) and from the definition of $\| u \|^{m_o}_{\gamma}$, clearly

$$\underline{\alpha}_n^T B^{m_o}(u_n)\, \underline{\alpha}_n \geq \| y_n \|_{\gamma}$$

and since all norms on the finite space P_n are consistent we have verified that $B^{m_o}(\underline{u}_n)$ is uniformly positive definite.

With this result established the arguments of Ciarlet, Schultz, Varga (1, p.398-9) follow immediately, so that the uniqueness of $y_n^{m_o}(x)$ is proved.

We notice that the restriction (5.20) evident in the Theorem of Herbold, Schultz and Varga (1) has here not been required, but that the above Theorem severely restricts the choice of approximation subspace. It would appear that generalization of the above result to the case of piecewise polynomial approximation can be readily accomplished; in this respect we state the following theorem for which the above proof also holds.

Theorem.

Let $\overline{\Pi} : 0 = x_o < x_1 \ldots < x_N = 1$ be a partition of $[0,1]$, and let $S_n(\overline{\Pi})$ be a subspace of H_L consisting of functions $v(x)$ such that $v(x) = w_i(x)$ $x_i \leq x \leq x_{i+1}$, $i = 0 \ldots N-1$, where $w_i(x)$ is a polynomial of degree n in the interval $[x_i, x_{i+1}]$. Let $Q'_{m_1}[\sigma(x), 0, 1]$ be a quadrature rule satisfying

$$Q'_{m_1}[\sigma(x), 0, 1] = \sum_{j=0}^{m_1} w_t \, \sigma(\xi_j),$$

$$w_t > 0, \qquad \xi_j \in (0,1) \qquad j = 0 \ldots m_1$$

and let $Q_{m_o}[\sigma(x), 0, 1, \overline{\Pi}]$ be the corresponding composite quadrature rule of the form (5.12). Then there exists a unique function $y_n^{m_o}(x) \in S_n(\overline{\Pi})$ which minimizes the expression

$$I_{m_o}(y) = Q_{m_o}\left[\sum_{s=0}^{k} p_s(x)\left(\frac{d^s}{dx^s}y\right)^2 + 2\int_0^y f(x,\eta)\,d\eta,\ 0,1,\overline{\Pi}\right]$$

subject to the conditions

$$y^{l}(0) = y^{l}(1) = 0 , \qquad l = 0, 1 \ldots k-1 .$$

We remark that the above theorems do not show that

$$\lim_{m_o \to \infty} \left(y_n^{m_o}(x) \right) = y_n(x)$$

where $y_n(x)$ minimizes $I(y)$ in the subspaces P_{n-k} or $S_n(\overline{\Pi})$ and so the convergence of $y_n^{m_o}(x)$ to $y(x)$ where $y(x)$ is the solution of the given differential equation is not established by these results.

As an example using non-Gaussian quadrature we consider the quadrature approximated Rayleigh-Ritz solutions obtained in terms of a global polynomial basis, of the differential equation

$$y'' = \exp(y) \; ; \; y(0) = y(1) = 0$$

obtained by minimizing

$$I_{m_o}(y) = Q_{m_o}\left[y'^2 + 2\int_0^y \exp(\eta)\, d\eta \right]$$

over the subspace P_{n-1}, using as a basis the functions $\phi_i(x) = x(1-x)\, T^*_{i-1}(x)$. The expression

$$Q_{m_o}[\sigma(x)] = \frac{1}{m_o} \sum_{i=0}^{m_o}{}'' \sigma(i/m_o)$$

denotes the trapezium quadrature rule and $\sum''$ indicates that first and last terms in the summation are taken with weight 1/2. In Table XXIX we list the values of the approximation so produced at the points $x = i.h$ $i = 1..9$ $h = 0.1$ for $n = 2-7$, and also the values of the true solution. This approximation is obtained with $m_o = 10$ and we should not be surprised at the low accuracy of these solutions

Behaviour of the Approximate Solution determined using

Trapezium Rule guadrature, $m_o = 10$

Problem N2 $\qquad \phi_i(x) = x(1-x)\, T^*_{i-1}(x)$

n	TRUE	2	3	4	5	6	7
0.1	-4.143561'-2	-3.994132'-2	-3.726778'-2	-3.726778'-2	-4.018395'-2	-4.018395'-2	-5.864745'-2
0.2	-7.326837'-2	-7.100679'-2	-6.962891'-2	-6.962891'-2	-7.781684'-2	-7.781684'-2	-8.957194'-2
0.3	-9.579983'-2	-9.319441'-2	-9.455207'-2	-9.455207'-2	-1.004141'-1	-1.004141'-1	-1.028812'-1
0.4	-1.092377'-1	-1.065101'-1	-1.102292'-1	-1.102292'-1	-1.100589'-1	-1.100589'-1	-1.214435'-1
0.5	-1.147036'-1	-1.109481'-1	-1.155754'-1	-1.155754'-1	-1.124016'-1	-1.124016'-1	-1.317592'-1
0.6	-1.092377'-1	-1.065101'-1	-1.102292'-1	-1.102292'-1	-1.100589'-1	-1.100589'-1	-1.214435'-1
0.7	-9.579983'-2	-9.319641'-2	-9.455207'-2	-9.455207'-2	-1.004141'-1	-1.004141'-1	-1.028812'-1
0.8	-7.326837'-2	-7.100679'-2	-6.962891'-2	-6.962891'-2	-7.781684'-2	-7.781684'-2	-8.957194'-2
0.9	-4.143561'-2	-3.994132'-2	-3.726778'-2	-3.726778'-2	-4.018395'-2	-4.018395'-2	-5.864745'-2

Table XXIX

Behaviour of Solution Vector Determined by

Trapezium Rule Quadrature, $m_o = 10$

Problem N2 $\qquad \phi_i = x(1-x)\, T^*_{i-1}(x)$

n	TRUE	2	3	4	5	6	7
	-4.591899'-1	-4.437924'-1	-4.437924'-1	-4.246335'-1	-4.261621'-1	-4.261621'-1	-5.079843'-1
	0	+6.276577'-11	+6.264950'-11	+8.486195'-11	+8.492623'-11	+8.238167'-11	+8.159274'-11
	-4.408692'-3		+3.766821'-2	+3.766821'-2	+7.119080'-2	+7.119080'-2	+5.388678'-2
$a_n(i)$	0			+2.616281'-11	+2.594028'-11	+3.043847'-11	+3.005369'-11
	-3.364178'-5				+4.774627'-2	+4.774627'-2	+1.159240'-1
	0					+6.862119'-12	+6.734933'-12
	-2.922086'-7						+8.108988'-2

Table XXX

The Behaviour of Rayleigh-Ritz Approximations using

Trapezium Rule quadrature, $m_o = 20$

Problem N2 $\quad \phi_i(x)=x(1-x)\ T^*_{i-1}(x)$

n	TRUE	2	3	4	5	6	7
e_n^{-20}		7.427937'-4	1.335644'-3	1.335633'-3	1.700439'-3	1.700439'-3	2.378551'-3
e_n^{-10}		2.755535'-3	4.294349'-3	4.294349'-3	5.374064'-3	5.374064'-3	2.052271'-2
$a_n(i)$	-4.591899'-1	-4.532848'-1	-4.484259'-1	-4.484260'-1	-4.397072'-1	-4.397072'-1	-4.393216'-1
	0	+6.723690'-11	+6.721788'-11	+1.139027'-10	+1.140181'-10	+1.300976'-10	+1.300933'-10
	-4.408692'-1		+8.328005'-3	+8.328017'-3	+2.863406'-2	+2.863407'-2	+4.161364'-2
	0			+4.004373'-11	+3.980822'-11	+6.145184'-11	+6.137825'-11
	-3.364178'-5				+1.800285'-2	+1.800285'-2	+3.885991'-2
	0					+2.059899'-11	+2.051226'-11
	-2.922086'-7						+2.099935'-2

Table XXXI

even for moderate values of n since the integrand contains polynomials of high degree. The coefficients of the approximate expansion, displayed in Table XXX, become increasingly innaccurate as n increases, reflecting the larger errors in the matrix A_n and the vectors g_n $(\underline{a}_n)$. In Table XXXI we display the coefficients of the approximation obtained with $m_o = 20$, and give a comparison of the maximum errors in $y_n^{m_o}(x)$, denoted e_n^{-10} and e_n^{-20}, in the two cases. It is clear that there is some improvement when m_o is increased, though $m_o = 20$ remains too small.

This example and our earlier use of Gaussian quadrature illustrate that where we expect the integrands of $I(y_n)$ to be smooth functions, as is the case with a polynomial basis and a differential equation with smooth coefficients, Gaussian quadrature is, not surprisingly, to be preferred.

5.3 The overall error of Rayleigh-Ritz approximations.

We have seen that the error in a numerical solution of the differential equation

$$Ly = \sum_{s=0}^{k} (-1)^2 \frac{d^s}{dx^s}\left(p_s \frac{d^s y}{dx^s}\right) = f(x,y)$$

$$y^r(0) = y^r(1) = 0 \qquad r = 0,1 \ldots k-1$$

by the Rayleigh-Ritz method is due to a cumulation of errors from three sources. These are the error introduced by truncation, that is by the determination of the n^{th} Rayleigh-Ritz approximation

$$y_n(x) = \sum_{i=1}^{n} \underline{a}_n(i)\, \phi_i(x)$$

instead of

$$y(x) \approx \sum_{i=1}^{\infty} \underline{a}(i)\, \phi_i(x) \quad ,$$

by the effects of quadrature errors in the evaluation of the elements of the Rayleigh-Ritz equations, and by the effects of rounding errors and the magnification of both rounding and quadrature errors by the method of numerical solution of the equations. Error bounds including the effects of all three of these sources of error do not appear in the literature. Indeed, the truncation error problem seems to be the only one for which there is an adequate treatment. Results concerning the magnitude of $\| y_n - y \|_\infty$ are typified by the a-priori and a-posteriori bounds of Ciarlet, Schultz and Varga (1) (see also Gladwell (1), Schultz (2)). We recall the assumptions of § 2.6, i.e. there exist constants K , β such that either

$$\left\| w(x) \right\|_\infty \leq K \int_0^1 \left\{ \sum_{s=0}^{k} p_s(x) \left(\frac{d^s w}{dx^s} \right)^2 + \beta\, w(x)^2 \right\} dx \qquad \ldots(5.28)$$

or

$$\left\| D^l w(x) \right\|_\infty \leq K \int_0^1 \left\{ \sum_{s=0}^{k} p_s(x) \left(\frac{d^s w}{dx^s} \right)^2 + \beta\, w(x)^2 \right\} dx \qquad \ldots(5.29)$$

or

$$\left\| D^l w(x) \right\|_2 \leq K \int_0^1 \left\{ \sum_{s=0}^{k} p_s(x) \left(\frac{d^s w}{dx^s} \right)^2 + \beta\, w(x)^2 \right\} dx \qquad \ldots(5.30)$$

where $0 \leq l \leq k$,

and

$$\bigwedge = \inf_{y} \frac{\int_0^1 \sum_{s=0}^{K} p_s(x) \left(\frac{d^s y}{dx^s} \right)^2 dx}{\int_0^1 y(x)^2\, dx}$$

and

$$\frac{\partial f(x,y)}{\partial y} \geq \gamma > -\bigwedge$$

and

$$\Gamma_o = \max \frac{\partial f(x,y)}{\partial y} \qquad 0 \leq x \leq 1 \ , \quad |y| \leq \frac{2kM}{\Lambda + \gamma}$$

$$M = \max \left| f(x,0) \right|$$

Then defining

$$\left\| w \right\|_\gamma = \int_0^1 \left\{ \sum_{s=0}^{k} p_s(x) \left(\frac{d^s w}{dx^s} \right)^2 + \gamma\, w(x)^2 \right\} dx$$

the a-priori bound (Ciarlet, Schultz, Varga (1), Thm.3, p.403)

$$\left\| y_n - y \right\|_\infty \leq K \left\| y_n - y \right\|_\gamma \leq C \inf_{w \in S_n} \left\| w - y \right\|_\infty$$

where

$$C = K \left(1 + \max \left(\frac{\Gamma_o - \gamma}{\Lambda + \gamma} , 0 \right) \right)^{\frac{1}{2}}$$

is valid, where we assume only (5.28). If either (5.29) or (5.30) hold then this a-priori bound can be extended to derivatives of order up to and including 1 , taking the form

$$\left\| \frac{d^t}{dx^t} (y_n - y) \right\| \leq \frac{C}{2^{1-t}} \inf_{w \in S_n} \left\| w - y \right\|_\infty \qquad 0 \leq t \leq 1$$

If we define the residual $\Upsilon(x)$ of a function $w(x)$, where $w(x) \in S$ by

$$\Upsilon(x) = \sum_{s=0}^{k} (-1)^s \frac{d^s}{dx^s} \left(p_s(x) \frac{d^s w}{dx^s} \right) - f(x,w)$$

then the a-posteriori bound

$$\left\| w - y \right\| \leq K \left\| w - y \right\|_\gamma \leq \frac{K}{\sqrt{\Lambda}} \left\| \Upsilon(x) \right\|_2 \qquad \ldots(5.31)$$

is valid, (Ciarlet, Scultz, Varga (1), Thm.17, p.421).

The bounds of Gladwell (1), p.45, p.55-7, represent extensions of bounds essentially of the above form to the more general problems

considered by Gladwell. In the case of non-linear problems of the type we have considered they reduce to the form of (5.31).

Though Gladwell (1, p.13) states that the determination of the constant K of the assumptions (5.28), (5.29) or (5.30) is on occasion possible, these error bounds ought more properly to be regarded as indicative of the pointwise convergence of the Rayleigh-Ritz solutions rather than as practical error bounds for general problems.

Bounds on the effect of the other terms are in general equally impractical. For example, in the case of consistent quadrature schemes for piecewise Rayleigh-Ritz approximation it is not practical to compute a value for the constant K of (5.14), (Herbold, Schultz, Varga, (1) Corollary, p.113) and the determination of constants P_1 , Q_1 for the definition of stability of Mikhlin (3.21) is not in practice achieved without numerical computation of the eigenvalues of the Rayleigh-Ritz matrix by some numerical process itself subject to error.

Accordingly we believe that the situation is that rigorous error bounds on practical Rayleigh-Ritz approximations are not produced by these means. In practice the situation is very similar to that for many other methods for evaluating series approximations to the solution of differential equations; for example, the Chebyshev series direct expansion method (Clenshaw (1), Lanczos (3)) and collocation (Clenshaw and Norton (1), Wright (1)). That is, we obtain an estimate of the reliability of the approximation

$$y_n(x) = \sum_{i=1}^{n} a_n(i)\, \phi_i(x)$$

by examining the behaviour of the coefficients $a_n(i)$ for a sequence of values of n. In practice we must assume that the computed

values of $\underline{a}_n(i)$ are the exact Rayleigh-Ritz coefficients so that careful selection of the basis, so as to minimize the affects of rounding error, and the quadrature, must be made as discussed above. From the computed coefficients $\underline{a}_n(i)$ and the assumption

$$y(x) = \sum_{i=1}^{\infty} \underline{a}(i)\phi_i(x)$$

we can write, in the usual fashion

$$\varepsilon_n(x) = y(x) - y_n(x)$$

$$= \sum_{i=1}^{n} (\underline{a}(i) - \underline{a}_n(i))\,\phi_i(x) - \sum_{j=n+1}^{\infty} a(j)\phi_j(x)$$

so that

$$\left|\varepsilon_n(x)\right| \leq \sum_{i=1}^{n} \left|\underline{a}(i) - \underline{a}_n(i)\right| \cdot \left|\phi_i(x)\right| + \sum_{j=n+1}^{\infty} \left|\underline{a}(j)\right| \cdot \left|\phi_j(x)\right|$$

Delves and Mead (1) suggest that the second term in this error quickly dominates, so that

$$\left\|\varepsilon_n(x)\right\|_{\infty} \leq (1+\sigma) \sum_{j=n+1}^{\infty} \left|a(j)\right| \cdot \left\|\phi_j(x)\right\|_{\infty}$$

where $\sigma \ll 1$. Accordingly, if the differences

$$\left|\underline{a}_n(i) - \underline{a}_{n-1}(i)\right| \cdot \left\|\phi_i(x)\right\|_{\infty} \qquad i=1,\ldots,n-1$$

are small by comparison with $\left|\underline{a}_n(n)\right| \cdot \left\|\phi_n(x)\right\|_{\infty}$ we can regard

$$E_n = \left|\underline{a}_n(n)\right| \cdot \left\|\phi_n(x)\right\|_{\infty}$$

as an estimate of $\left\|\varepsilon_n\right\|_{\infty}$. As with other series expansions it is more appropriate to consider

$$E_n^* = \max_{k=0,1,2} \left|\underline{a}_n(n-k)\right| \cdot \left\|\phi_{n-k}(x)\right\|_{\infty}$$

as an estimate of $\|\varepsilon_n\|_\infty$.

If the Rayleigh-Ritz matrix A is uniformly asymptotically diagonal of degree $p \geq 2$ it might be thought that more elaborate procedures might be used to estimate $\underline{a}(j)$, $j > n$, and hence to estimate

$$\sum_{j=n+1}^{\infty} |a(j)| \cdot \|\phi_j(x)\|_\infty$$

either by use of the prediction coefficients defined by Delves and Mead (1), using $(\underline{a}_n , Q)$ as an initial vector, or by an approach based on the following remarks. We shall show, however, that the estimates based on this second approach can be very pessimistic.

Typically, provided that the right hand side vector $\underline{b}$ or the sequence $\underline{b}_n$ $(\underline{a}_n^{(r)})$ satisfy certain conditions, (Delves and Mead (1), Thm.3, p.14, also Ch.4, Thm.2, p.139) we have

$$|\underline{a}(j)| < C\, j^{-q} \quad , \quad C > 0 , \qquad 1 \leq q \leq p$$

so that

$$\sum_{j=n+1}^{\infty} |\underline{a}(j)| \cdot \|\phi_j\|_\infty \leq C \sum_{j=n+1}^{\infty} \frac{\|\phi_j\|_\infty}{j^q}$$

If $\|\phi_j\|_\infty = K$, a constant, then

$$\|\varepsilon_n\| \leq C K \sum_{j=n+1}^{\infty} \frac{1}{j^q} \leq C K \int_{n+1}^{\infty} \frac{1}{j^q}$$

Hence, provided $q \geq 2$, we have

$$\|\varepsilon_n\| \leq \frac{C K}{(n+1)^{q-1}}$$

In the case of the application of the basis functions

$$\phi_i(x) = x(1-x)T^*_{i-1}(x) \qquad i = 1,2 \ldots$$

to the problem

$$Ly = \frac{d^2}{dx^2} y = f(x,y)$$

$$y(0) = y(1)$$

we have established (Ch.4, Thm.2, p.184) that provided $f(x,y) < M$, a constant, for all y 'near' the solution then $p = 3$, $q = 2$.

Noting that $K = 1/4$, we have

$$\|\varepsilon_n\|_\infty \leq \frac{C}{4(n+1)} \qquad \ldots(5.32)$$

If the coefficient vectors defined by the Rayleigh-Ritz method, $\underline{a}_n$, are such that the elements $\underline{a}_n(i)$ are tending smoothly to zero, we may estimate C by

$$C_n = \left| n^2 \ \underline{a}_n (n) \right| \qquad \ldots(5.33)$$

or more generally by

$$C_n^t = \max_{0 \leq 1 \leq t} \left| (n-1)^2 \ \underline{a}_n(n-1) \right| \qquad \ldots(5.34)$$

where t is some small integer. Substituting from (5.33) and (5.34) into (5.32) we obtain the estimates of the error bound

$$E_n^o = C_n / 4(n+1)$$

and

$$E_n^t = C_n^t / 4(n+1)$$

In Tables XXXII, XXXIII we illustrate the magnitude of these estimates for the problems

$$L(y) = y'' + y = -x$$

$$y(0) = y(1) = 0$$

and

$$L(y) = y'' = \tfrac{1}{2}(y+x+1)^3$$

for different values of n, and $t = 2$. The measured estimate

of $\|\varepsilon_n\|_\infty$ defined by

$$\|\varepsilon_n\|_\infty \doteq e_n = \max_{x_i} \left| y(x_i) - y_n(x_i) \right|$$

where $x_i = i/20$, $i = 0 \ldots 20$ is given for comparison. We observe immediately that the estimates E_n^o and E_n^2 and particularly the latter, are very wide bounds, and might well be thought impractical. The estimate E_n , based on the final coefficient $\underline{a}_n(n)$, is more satisfactory, and indeed seems very good in the case of the non-linear problem (Table XXXIII). We remark that the linear problem is somewhat exceptional in that its solution is approximated very well by low order polynomials.

The estimates E_n^c and E_n^2 do not seem to be without practical use, however. A comparison of Tables XXXII and XXXIII reveals that the magnitudes of these estimates become larger when the error e_n becomes larger, that is when numerical and quadrature errors have serious effect. This has been observed for a number of test examples. Accordingly the estimates E_n^o and E_n^t may be of value in providing a criterion for the automatic termination of the evaluation of the sequence of Rayleigh-Ritz approximations $y_n(x)$, $n = 1,2 \ldots$.

These methods of estimating the error of the Rayleigh-Ritz approximation are weak due to the asymptotic nature of U.A.D properties. Similar remarks apply to other techniques based on this theory, such as the prediction of the coefficients $a(n+1)$, $a(n+2)$ ... from $\underline{a}_n$ by application of the Gauss-Seidel or Jacobi iteration to the infinite Rayleigh-Ritz matrix, as described by Delves and Mead((1), p.212).

The determination of readily computable error estimates and bounds

Error Estimates

$y'' = -y-x$ $\qquad \phi_i(x)=x(1-x)T^*_{i-1}(x)$

n	e_n	E_n^0	E_n^2	E_n
7	5.36'-7 (5.18'-11)	3.1'-7	4.8'-6	5.0'-8
8	5.36'-7 (3.00)'-11	5.6'-7	6.0'-6	6.5'-7
12	5.96'-7 (3.19'-11)	1.3'-5	1.3'-5	1.1'-6

Table XXXII (see Table VII)

Error Estimates

$y'' = \frac{1}{2}(y+x+1)^3$ $\qquad \phi_i(x)=x(1-x)T^*_{i-1}(x)$

n	e_n	E_n^0	E_n^2	E_n
4	1.08'-4	5.4'-3	4.8'-2	1.7'-4
6	3.16'-6	3.0'-4	4.1'-3	5.0'-5
10	2.71'-9	3.8'-7	2.0'-5	4.3'-8

Table XXXIII (see Table XX)

for the Rayleigh-Ritz method is a problem which is not easily resolved. Accordingly we are of the opinion that the only practical criteria which can be applied to the determination of a suitable value of n, and the corresponding approximate solution $y_n(x)$ are those based on a careful choice of the basis functions ϕ_i and the quadrature scheme used guided by the remarks of Ch's. 3, 4 and 5, and a study of the convergence of the computed vectors $\underline{a}_n$ and functions $y_n(x)$. Of the estimates of $\| \varepsilon_n \|_\infty$ the value

$$E_n = \left| \underline{a}_n (n) \right| \cdot \| \phi_n \|_\infty$$

seems most satisfactory; it should nevertheless be used with caution.

Chapter Six

Extensions and Conclusions

In this last chapter we comment on extensions of the principles and methods of earlier chapters to more general variational problems than those previously considered here. In § 6.1 we consider an approach to the use of the finite element method for general second order elliptic partial differential equations, and in § 6.2 present a natural formulation of an approximate extended Kantorovich method (Kantorovich and Krylov (1), p.304, Kerr (1), Andersenn (2)), described with reference to Laplace's equation, but readily extended to other equations.

We consider briefly the possible application of a Rayleigh-Ritz method to the "simplest problem" of variational calculus in its general form, including the class of problems considered by Allen, (1,2). Finally, in § 6.4 we summarize the conclusions which may be reached from our numerical investigation of the Rayleigh-Ritz method emphasizing again many of the results and comments of this and earlier chapters.

6.1 An approximate finite element method for elliptic partial differential equations in two dimensions

We consider in this section the application of a finite element method involving the use of piecewise planar functions defined over triangular elements to second order elliptic partial differential equations in two dimensions, and suggest an approach which may be useful when the differential operator involved has coefficients dependent on the space variables.

The application of the finite element method in a number of forms and to various problems is described in a vast and expanding literature, to which we can make only token reference. An important text for many mechanical and mathematical aspects of the method is that of Kolar, Kratochvil, Zlamal and Zenicek (1), whilst a mathematical formulation appropriate to this section is given by Birkhoff, Schultz, Varga (1). Many aspects of the method are described in conference proceedings edited by Bramble and Hubbard (1) and Schoenberg (1). An extensive bibliography on the finite element method and its applications has recently been compiled by Whiteman (1).

For convenience we consider only positive definite self adjoint partial differential equations of the form

$$Lu = \frac{\partial}{\partial x}\left(p_1(x,y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(p_2(x,y)\frac{\partial u}{\partial y}\right) = f(x,y)$$

$$, \; x \in \Omega \qquad \ldots(6.1)$$

$$u = 0 \text{ on } \partial\Omega \qquad \ldots(6.2)$$

where Ω is a bounded region of R_2 with boundary $\partial\Omega$. Extensions of the proposed scheme to the case $Lu = f(x,y,u)$ and to the case in which L is not a self-adjoint operator, may be made using a Galerkin formulation.

Corresponding to the equation (6.1) we have the variational problem

$$I(u_o) \leq I(u) = \int_\Omega \left\{ p_1(x,y)\left(\frac{\partial u}{\partial x}\right)^2 + p_2(x,y)\left(\frac{\partial u}{\partial y}\right)^2 + 2f(x,y)u \right\} d\Omega$$

with the boundary conditions (6.2) above, where the functions $u(x,y)$ satisfy

$$u(x,y) \in \mathcal{L}_2^1(\Omega) .$$

It is known (e.g. Mikhlin (1,4)) that the function $u_o(x,y)$ is a weak or generalized solution of equation (6.1).

The finite element method we wish to consider generates approximations to $u_o(x,y)$ in the following manner. A triangulation T is super-imposed on the region Ω such that each triangle of T has at least one vertex in Ω not on $\delta\Omega$. We do not consider triangles having any vertex external to Ω. Algorithms for constructing such triangulations of a given region Ω have been described by George (1) and Reid (2). We depict such a triangulation below (Fig. IX)

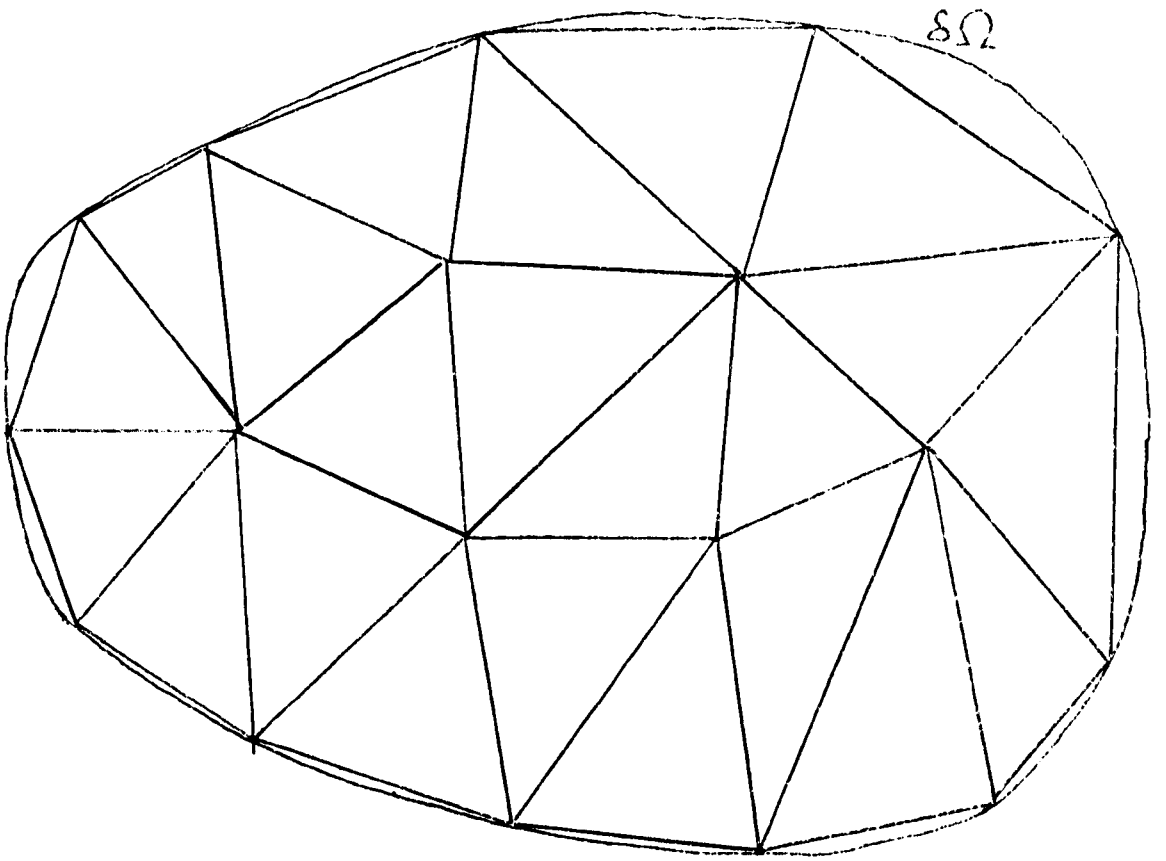


Fig. IX

For simplicity we now assume that the region Ω is approximated by the polygonal region Ω' whose boundary $\delta\Omega'$ is described by the

straight lines connecting 'adjacent' points of the triangulation T lying on $\delta\Omega'$, although Reid (2) considers the case in which the original region Ω is retained during the finite element procedure, and the simplification is not essential to our method. We now suppose that there are k vertices $P_1 .. P_k$ of the triangulation T internal to Ω' and $n - k$ vertices $P_{k+1} .. P_n$ lying on $\delta\Omega'$. In the simplest case, which is the one we pursue in detail, we associate with each internal vertex P_i, $i = 1 .. k$, a function $\phi_i(x,y)$ having the properties that

1) $\phi_i(x,y)$ is a planar function in each triangle of the triangulation T

2) $\phi_i(x_j,y_j) = \delta_{ij}$, $j = 1, .. , k$

where $P_j = (x_j,y_j)$

3) $\phi_i(x_j,y_j) = 0$ $\quad j = k+1, .. , n$.

Such a function is clearly non-zero only on triangles of T having $P_i = (x_i,y_i)$ as a vertex. Then an approximation $\bar{u}_T(x,y)$ to $u_0(x,y)$ of the form

$$\bar{u}_T(x,y) = \sum_{i=1}^{K} \bar{a}_i \phi_i(x,y)$$

is determined by the condition

$$I'(\bar{u}_T) \leq I'(u_T) = \int_{\Omega'} \left\{ p_1(x,y)\left(\frac{\partial u_T}{\partial x}\right)^2 + p_2(x,y)\left(\frac{\partial u_T}{\partial y}\right)^2 - 2f(x,y)u_T \right\} d\Omega'$$

where

$$u_T(x,y) = \sum_{i=1}^{k} a_i \phi_i(x,y) .$$

Substituting for $u_T(x,y)$ and differentiating with respect to

a_i , $i = 1 \,..\, k$, we obtain the equations

$$A\,\underline{\bar{a}} = \underline{b}$$

where A is the $k \times k$ matrix

$$A(i,j) = \int_{\Omega'} \left\{ p_1(x,y) \frac{\partial \phi_i}{\partial x} \cdot \frac{\partial \phi_j}{\partial x} + p_2(x,y) \frac{\partial \phi_i}{\partial y} \frac{\partial \phi_j}{\partial y} \right\} d\Omega'$$

$$i,j = 1 \,..\, k$$

and $\underline{b}$ the vector

$$b(i) = \int_{\Omega'} f(x,y)\, \phi_i(x,y)\, d\Omega'$$

$$i = 1 \,..\, k$$

We remark briefly that, in view of the form of the functions $\phi_i(x,y)$ the coefficients $\underline{\bar{a}}(i)$ satisfy

$$\underline{\bar{a}}(i) = \bar{u}_T(x_i, y_i)$$

so that the resulting coefficients are the values of the approximate solution at the vertices of T .

The expression for $A(i,j)$ admits immediate simplification if the vertices P_i , P_j are not immediate neighbours in T , that is, if no triangle has both P_i and P_j amongst its vertices, for then by definition of $\phi_i(x,y)$, $\phi_j(x,y)$, we have $A(i,j) = 0$. In fact, even if P_i , P_j are immediate neighbours in T , only two triangles have them both as vertices, so that, depicting these two triangles in Fig. X we can write

$$A(i,j) = \int_{T_1+T_2} \left\{ p_1(x,y) \frac{\partial \phi_i}{\partial x} \cdot \frac{\partial \phi_j}{\partial x} + p_2(x,y) \frac{\partial \phi_i}{\partial y} \frac{\partial \phi_j}{\partial y} \right\} d\Omega' \quad \ldots(6.3)$$

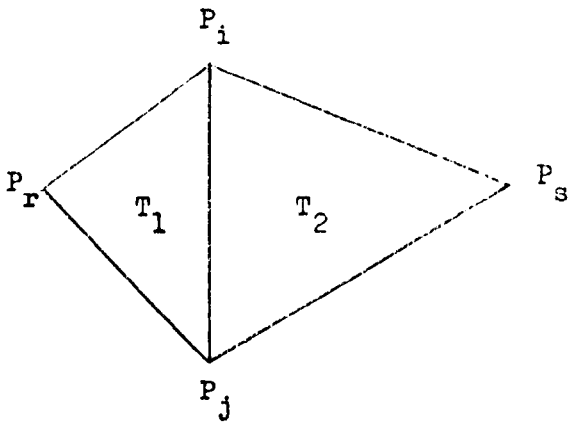


Fig. X

Similarly the region of integration Ω' in the definition of $\underline{b}(i)$ may be reduced to the region covered by all triangles of T having P_i as a vertex.

The evaluation of $A(i,j)$ and $\underline{b}(i)$ is not, however, trivial unless the coefficient functions $p_1(x,y)$, $p_2(x,y)$ are simple, and in practice the method is often only discussed with reference to the Laplace and Poisson equations where $p_1(x,y) = p_2(x,y) = 1$. Quadrature approximations are often necessary for more general problems, although little study seems to have been made of them. We can, however, cite the thesis of Herbold (1), who considers quadrature approximations to the elements of $\underline{b}$ for equations of this type. We now suggest a practical scheme for the case where $p_1(x,y)$, $p_2(x,y)$ are general functions of x and y which we believe has some merit. In particular we show how certain simplifications which can be applied in the case of Laplace's equation are also applicable to the new method, and briefly indicate how the new method may be applied using basis functions which are, for example, cubic rather than planar in each element.

Only one term of (6.3) is considered, since the expression may be formed by summation of similar terms. We take

$$C(i,j) = \int_{T_1} p_1(x,y) \frac{\partial \phi_i}{\partial x} \frac{\partial \phi_j}{\partial x} \, d\Omega' \qquad \ldots(6.4)$$

where T_1 is the triangle (Fig.X) with vertices (x_i,y_i), (x_j,y_j), (x_r,y_r). Reid (1) considers that in this triangle we may represent the planar function $\phi_i(x,y)$ by a linear combination of the Lagrange functions. Extending this, we define these Lagrange functions on T_1 by

$$l_u(x_v,y_v) = \delta_{uv} \quad , \quad u,v \in i,j,r$$

and the additional condition that l_u is linear in x and y, and express $p_1(x,y)$ approximately as

$$p_1(x,y) \doteq \sum_{u=i,j,r} p_1(x_u,y_u) \, l_u(x,y) \qquad x,y \in T_1$$

(i.e. $p_1(x,y)$ is approximated by its planar interpolant on T_1). Then the expression (6.4) may be approximately written as

$$C(i,j) \doteq \sum_{u=i,j,r} p_1(x_u,y_u) \int_{T_1} l_u(x,y) \frac{\partial \phi_i}{\partial x} \cdot \frac{\partial \phi_j}{\partial x} \, d\Omega'$$

Each of the integral terms is now a triple product of the Lagrange functions and their derivatives, for we note that

$$\phi_i(x,y) = l_i(x,y) \quad , \quad (x,y) \in T_1$$

so that

$$\int_{T_1} l_u(x,y) \frac{\partial \phi_i}{\partial x} \cdot \frac{\partial \phi_j}{\partial x} \cdot d\Omega' = \int_{T_1} l_u(x,y) \frac{\partial l_i}{\partial x} \cdot \frac{\partial l_j}{\partial y} \, d\Omega'$$

and the integration on the right is comparatively simple in comparison with our earlier expressions for $A(i,j)$.

George (1), has proposed a further scheme for the simplification of the evaluation of $A(i,j)$ for Laplace's equation which is readily applied to our more general problem. In the (ξ, η) plane we define a canonical triangle T° with vertices (0,0), (1,0), (0,1), and on this triangle the basis functions $\phi_i^\circ(\xi, \eta)$ $i = 1,2,3$ with the properties

1) $l_i^\circ(\xi, \eta)$ is a planar function

2) $l_i^\circ(\xi_j, \eta_j) = \delta_{ij}$, $i,j = 1 \ldots 3$, where the triangle T° has vertices (ξ_j, η_j). Let the linear transformation

$$(\xi, \eta) = J(x,y)$$

be such that $(x,y) \in T_1$ iff $(\xi, \eta) \in T^\circ$ and $(x,y) \in \Omega'$ iff $(\xi, \eta) \in \Omega^\circ$

Then

$$\int_{T_1} l_u(x,y) \frac{\partial l_i}{\partial x} \cdot \frac{\partial l_j}{\partial x} \, d\Omega'$$

$$= \int_{T^\circ} l_u^\circ(\xi, \eta) \frac{\partial l_i^\circ}{\partial \xi} \cdot \frac{\partial l_j^\circ}{\partial \xi} \cdot \det(J) \, d\Omega^\circ$$

$$+ \ldots\ldots$$

where the remaining terms are similar quadratic forms introduced by the transformation. Thus integrals of the form

$$\int_{T^\circ} l_u^\circ(\xi, \eta) \frac{\partial l_i^\circ}{\partial \xi} \cdot \frac{\partial l_j^\circ}{\partial \eta} \, d\Omega^\circ$$

may be evaluated once only, and then premultiplied by the appropriate Jacobian determinant and the value of the coefficient function at P_u , $u = i,j,r$, before being added in to $A(i,j)$.

We remark that the scheme is in reality a simple quadrature approximation of the integrals occurring in the finite element method derived so as to be exact when the coefficient functions and the basis functions are piecewise planar functions on T. In particular it generates the usual planar finite element approximation to the Laplacian operator since in this case the coefficient functions are exactly represented by their interpolant in the planar basis on T. The scheme extends to other finite element approximations, for example to bi-linear functions defined over rectangular elements, or to cubic functions on triangular elements. In this last case the usual parameterization of the cubic basis function is in terms of its values and those of its x and y derivatives at the vertices of the triangles together with the function value at the centre of the circumscribing circle. This would lead to approximations of the coefficient functions involving not only $p_1(x_i,y_i)$, but also

$$\frac{\partial}{\partial x} p_1(x,y)\Big|_{(x_i,y_i)} \quad \text{and} \quad \frac{\partial}{\partial y} p_1(x,y)\Big|_{(x_i,y_i)}$$

and a simple quadrature formula involving only values of the coefficient functions and their derivatives.

Finally in this section we comment that finite element procedures have often been regarded as variants of a finite difference procedure, and this view can clearly be applied to the approximate finite element method which we have described. Nonetheless, the 'finite difference' equations produced by our finite element method are of non-standard form, even on a rectangular mesh. For example, considering the use of bi-linear

functions on a square mesh, our method generates the same non-standard finite difference approximation of the Laplacian as the finite element approximation of Birkhoff, Schultz and Varga (1, p.253).

6.2 The Kantorovich and Extended Kantorovich Methods.

In this section we comment on some aspects of the 'method of reduction to ordinary differential equations' proposed by Kantorovich (see Kantorovich and Krylov (1), p.304-337, and for an example of its practical application McDonald (1)), and on an extension of this method suggested by Kerr (1), and studied in a particular form by Andersenn (2). Following Kantorovich, we outline the method and its extension, in relation to their application to Laplace's equation in two dimensions,

$$L\,u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = f(\underline{x}) \qquad \ldots(6.4')$$

defined on the rectangle Ω : $-a \leq x_1 \leq a$, $-b \leq x_2 \leq b$, and subject to the boundary conditions

$$u(-a) = u(a) = 0 \qquad \ldots(6.5)$$

$$u(-b) = u(b) = 0 \qquad \ldots(6.6)$$

Extension of the discussion to other differential equations involving a positive definite operator L , and to higher dimensions is readily achieved.

Corresponding to the equation (6.4') with boundary conditions (6.5), (6.6) we have the variational principle :-

Let $\bar{u}(x)$ be a function such that

$$I(\bar{u}) \leq I(u) = \int_\Omega \left\{ \left(\frac{\partial u}{\partial x_1}\right)^2 + \left(\frac{\partial u}{\partial x_2}\right)^2 + 2f(\underline{x})u \right\} d\Omega$$

where we consider only functions u, $\bar{u}$ satisfying the boundary conditions (6.5), (6.6) and such that

$$u, \bar{u} \in C^0[\Omega]$$

$$u, \bar{u} \in \mathcal{L}_2^1[\Omega]$$

The function $\bar{u}(\underline{x})$ is then a weak or generalized solution of (6.4) with the boundary conditions (6.5), (6.6), from the variational theory of Chapter Two.

The Kantorovich method determines a sequence of approximations to $\bar{u}(x)$, of the form

$$u_N(x) = \sum_{t=1}^{N} g_t^1(x_1)\, \theta_t^1(x_2) \qquad \ldots(6.7)$$

$$, N = 1,2\ldots$$

where the functions $\theta_t^1(x_2)$ are prescribed linearly independent functions satisfying (6.6). The functions $g_t^1(x_1)$ are determined so as to satisfy the remaining boundary conditions (6.5) and so that $I(u_N)$ is a minimum of $I(u)$ for all functions u of the form (6.7). We pursue the determination of $g_t^1(x_1)$ subsequently.

The extended Kantorovich method determines a sequence of approximations to $\bar{u}(x)$ of the form

$$u_N^{n,i}(\underline{x}) = \sum_{t=1}^{N} g_t^n(x_i)\, \theta_t^n(x_j) \qquad \ldots(6.8)$$

$$i = 1,2\ ;\ i \neq j$$

$$n = 1,2\ldots;\ N = 1,2\ldots$$

The function $\theta_t^n(x_2)$ is chosen as in the Kantorovich method for $n = 1$, and subsequently by the relations

$$\theta_t^n(x_1) = g_t^n(x_1)$$

$$\theta_t^n(x_2) = g_t^{n-1}(x_2)$$

The unknown functions $g_t^n(x_i)$ are determined at each iteration as in the Kantorovich method, by the requirements that they satisfy the boundary conditions (6.5) if $i = 1$ or (6.6) if $i = 2$, and to ensure that $I(u_N^{n,i}) \leq I(u)$ for all functions u of the form (6.8). We now take $i = 1$, for convenience. Substituting $u_N^{n,i}$ for u in $I(u)$, we have

$$I(u_N^{n,1}) = \int_\Omega \left\{ \left(\sum_{t=1}^{N} \Theta_t^n(x_2) \cdot \frac{dg_t^n(x_1)}{dx_1} \right)^2 + \left(\sum_{t=1}^{N} \frac{d\Theta_t^n(x_2)}{dx_2} \cdot g_t^n(x_1) \right)^2 + 2f(\underline{x}) \cdot \sum_{t=1}^{N} \Theta_t^n(x_2)\, g_t^n(x_2) \right\} d\Omega$$

Integrating with respect to x_2, and using the boundary conditions (6.5) to eliminate terms (see Kantorovich and Krylov (1), p.306-7) we obtain

$$I(u_N^{n,1}) = I(\underline{g}^n(x_1)) = \int_{-a}^{a} \left\{ \sum_{t=1}^{N} P_t^n \left(\frac{dg_t^n(x_1)}{dx_1} \right)^2 + \sum_{t=1}^{N} Q_t^n \left(g_t^n(x_1) \right)^2 - 2 \sum_{t=1}^{N} F_t^n g_t^n(x_1) \right\} dx_1 \qquad \ldots(6.9)$$

In the case of Laplace's equation, and the torsion equation considered by Kerr (1), Andersenn (2), the expressions P_t^n, Q_t^n, F_t^n, $t = 1 \ldots N$, are constants. For more general positive definite differential equations, these expressions will be functions of x_1.

The functional (6.9) may now be minimized over its parameters $g_t^n(x_1)$ in order to determine $u_N^{n,1}(\underline{x})$. This minimization is achieved by Kantorovich and Krylov (1), Kerr (1) and Andersenn (2) by solving analytically the corresponding Euler equations

$$\frac{\partial I\,(g_t^n\,(x_1))}{\partial g_t^n\,(x_1)} = 0 \quad , \qquad t = 1 \ldots N \qquad \ldots(6.10)$$

with the boundary conditions

$$g_t^n\,(-a) = g_t^n\,(a) = 0 \quad , \qquad t = 1 \ldots N \; .$$

These equations are linear, and for Laplace's equation and the torsion equation have constant coefficients, so that exact solutions are readily obtained, particularly for small values of N .

Where these equations do not have this simple form, it will in general be necessary to solve the differential equation (6.10) or the corresponding variational problem (6.9) approximately. We can therefore conceive a number of 'approximate extended Kantorovich' methods. From our viewpoint the approximate solution of the variational problem (6.9) is an interesting possibility. We assume that for each t , $t = 1 \ldots N$, k independent basis functions $\phi_{t,1}(x) \ldots \phi_{t,k}(x_1)$ are chosen, each satisfying $\phi_{t,r}(-a) = \phi_{t,r}(a) = 0$, and g_t^n is represented by

$$g_t^n\,(x_1) = \sum_{r=1}^{k} \alpha_{k,t}^{n}(r)\;\; \phi_{t,r}\,(x_1)$$

Substituting in $I(g^n)$, and minimizing with respect to the scalar parameters $\alpha_{t,r}^n$, $t = 1 \ldots N$, $r = 1 \ldots k$, leads to a linear system of Nk algebraic equations, the Rayleigh-Ritz equations for the minimization of (6.9).

We remark that this algorithm remains untried. It is apparent,

however, that the careful choice of co-ordinate system $\{\phi_{ti}\}$ will be governed by the same considerations which apply when the Rayleigh-Ritz method is used in the solution of ordinary differential equations, together, perhaps, with some additional criteria. Andersenn (2) has already pointed out that a careful selection of the function $\Theta_1^1(x_2)$ is necessary, even in the simple case $N = 1$, to ensure that the iterative determination of the sequence $u_1^n(\underline{x})$ converges. Further research along these lines is indicated.

6.3 The 'Simplest Problem' of Variational Calculus

Some investigations of a series expansion method for the determination of approximate solutions of the 'simplest problem' of variational calculus have been made, which are considered only briefly here. Amongst such problems are those considered by Allen (1), which may be stated (see Chapter One) as

Determine $\bar{y} = \bar{y}(x)$ s.t

$$I(\bar{y}) \leq I(y) = \int_0^1 F(x,y,y')\,dx \qquad \ldots(6.11)$$

subject to

$$y(0) = y(1) = 0 \qquad \ldots(6.12)$$

For convenience we do not include the similar boundary conditions $y(0) = \alpha$, $y(1) = \beta$, which are readily reduced to those above. The functions $y(x)$ are restricted to those for which the integral $I(y)$ exists and has finite value. To be more specific a knowledge of the form of $F(x,y,y')$ is required, and we wish to avoid further assumption concerning the nature of the integrand. In particular we do not restrict ourselves to those cases in which F is quadratic, or 'nearly so', in the dependent variables y, y', since these cases

correspond to those for which the Rayleigh-Ritz method has been derived in earlier chapters.

Allen (1), (2), and others have derived methods for this problem based on finite difference approximations of $y'(x)$ and simple quadrature approximations of the integral. Allen's methods closely resemble shooting methods for the solution of boundary value problems in ordinary differential equations. An analysis of theoretical properties of methods such as these for this 'simplest problem' is given by Stepleman (1).

We investigate an approach closely analogous to the Rayleigh-Ritz procedure. If $\phi_i(x)$, $i = 1, \ldots, \infty$ are a chosen set of basis functions satisfying the boundary conditions (6.12), and forming a complete sequence in the appropriate function space, and $\phi_i(x)$, $i = 1, \ldots, n$ are linearly independent for all n, then an approximation

$$y_n(x) = \sum_{i=1}^{n} a_n(i)\, \phi_i(x)$$

might be determined by the minimization of

$$I(y_n) = I(\underline{a}_n) = \int_0^1 F\left(x, \sum_{i=1}^{n} a_n(i)\phi_i(x), \sum_{i=1}^{n} a_n(i)\phi_i'(x)\right) dx \qquad \ldots(6.13)$$

The expression (6.13) is a function of the n variables $a_n(i)$, and minimization with respect to these might be achieved in a number of ways. Pursuing the analogy with the Rayleigh-Ritz method, the equations

$$\frac{\partial I(\underline{a}_n)}{\partial a_n(i)} = 0 \qquad i = 1, \ldots, n$$

are obtained. In the case that $F(x,y,y')$ has the simple form

considered in earlier chapters these are the usual Rayleigh-Ritz equations. More generally, these are non-linear equations and may have any non-negative number of solutions (as may the original problem). Though this method introduces no additional theoretical complexity to the problem, in practice differentiation of the function $I(a_n)$ may be inhibiting, and a more direct approach required. One of the many algorithms (see Powell (1)) for the minimization of a function of n variables might be employed. The purpose of this section is to illustrate that whilst this method is practicable, difficulties similar to those encountered using the Rayleigh-Ritz method in its usual form arise. In particular the selection of the co-ordinate system $\phi_i(x)$ again affects the performance of the method, as we show in a simple example below. The further point, that the minimization techniques locate local, rather than global minima, must also be born in mind.

Problem M1 :

Find $\bar{y}(x)$ s.t

$$I(\bar{y}) \leq I(y) = \int_0^1 \left\{ x^2 y'^2 + 12y^2 + 20xy \right\} dx$$

subject to

$$y(0) = y(1) = 0 .$$

This example has the unique solution

$$\bar{y} = x^3 - x$$

as may be verified from the Euler equation, and for which $I(\bar{y}) = -4/3$. We summarize in Tables XXXIV, XXXV the performance of the Davidon minimization algorithm (Fletcher-Powell (1)) using the basis functions

$$\phi_i(x) = x^i(1-x)$$

and

$$\phi_i(x) = x(1-x)\, T^*_{i-1}(x)$$

Convergence of Davidon Minimization, n=2

Problem M1 $\phi_i(x) = x^i(1-x)$

Function Evaluations	a_n (1)	a_n (2)	$I(y_n)$
0	-1.200000	-0.800000	-1.328000
17	-1.133464	-0.771484	-1.331689
26	-1.137132	-0.781446	-1.331733
37	-1.077520	-0.930959	-1.332415
49	-1.000427	-1.038437	-1.333028
68	-0.995719	-1.007085	-1.333332
80	-0.997066	-1.003156	-1.333332
87	-0.999401	-0.999447	-1.333333
96	-1.000484	-0.999259	-1.333333
108	-0.999828	-1.000008	-1.333333
Exact Solution	-1.000000	-1.000000	-1.333333'

Table XXXIV

Convergence of Davidon Minimization, n=2

Problem M1 $\phi_i(x) = x(1-x)T^*_{i-1}(x)$

Function Evaluations	a_n (1)	a_n (2)	$I(y_n)$
0	-0.600000	-1.400000	-0.901333
6	-1.440000	-1.280000	-1.256533
13	-1.369822	-0.665295	-1.323521
24	-1.501329	-0.499896	-1.333332
Exact Solution	-1.500000	-0.500000	-1.333333'

Table XXXV

with $n = 2$, for which the exact solution may be expressed exactly in terms of either basis. We note the comparatively rapid convergence of the minimization procedure when the trial function is expressed in terms of Chebyshev polynomials. This has been observed in this and other examples, and for larger values of n. In the case of problem M1, we can make the following remarks about the convergence of the Davidon algorithm. For any trial function $y_n(x)$ which is a linear combination of n independent basis functions, the function $I(y_n) = I(\underline{a}_n)$ is a quadratic function of the n parameters $\underline{a}_n(i)$, $i = 1...n$. For such functions the Davidon algorithm converges in n iterations provided that the one-dimensional minimization problems occurring in each iteration are solved exactly. In practice this will not be the case. From Tables XXXIV, XXXV, where each row except the first, corresponds to a Davidon iteration, we can see that convergence requires more than two iterations. We can attribute the differing convergence rates to the different accuracies with which the one-dimensional problems are solved, and the observed superiority of the Chebyshev expansion in the minimization algorithm may be a function of parameters of the one-dimensional minimization procedures.

The example considered above is of the simple quadratic type for which the Rayleigh-Ritz method in its usual form is appropriate. We experiment further with the following example considered by Allen (2, Ex. i) for which minimization techniques seem more appropriate.

Problem M2[†]:- Determine $\bar{z}(x)$ such that

[†]This example is given also by Allen (1, p.208, Ex.iii), where the expression $(1.05 \exp(1-4z'))$ is printed $(1.05 - \exp(1-4z'))$.

The latter form appears to be in error since in this case we believe $I(z)$ may assume arbitrarily large negative values.

$$I(\bar{z}) \leq I(z) = \int_0^1 f(x,z,z')dx$$

$$= \int_0^1 \frac{1.05 \exp(1-4z')(1-z)\cdot c(x)}{(1+z^2)} \left(2 - \frac{c(x)}{2(1-z)}\right) dx$$

subject to $z(0) = 0.5$, $z(1) = 0$, where

$$c(x) = (1 + x \exp(-6(x-0.4)^2))^{-1}$$

This problem arises in the study of the utilization of fuel by a ship, and its practical significance assures a unique solution (Allen (1)). To apply our minimization techniques we prefer to utilize the transformation

$$z(x) = y(x) + \tfrac{1}{2}(1-x)$$

and write

$$I(z) = I(y) = \int_0^1 f(x,y,y')\, dx$$

$$= \int_0^1 \frac{1.05 \exp(3-4y')(\frac{1}{2}(x+1)-y)\cdot c(x)}{(1 + (y - \frac{1}{2}(x-1))^2)} \left(2 - \frac{c(x)}{(x+1-2y)}\right) dx$$

with the homogeneous boundary conditions $y(0) = y(1) = 0$. The unknown function $y(x)$ is approximated by

$$y_n(x) = \sum_{i=1}^{n} \underline{a}_n(i)\, \phi_i(x)$$

where $\phi_i(x)$ are admissible functions, and the integrand approximated by a quadrature rule $Q_{m_o}[f, 0, 1]$, so that the function of $\underline{a}_n$,

$$I_{m_o}(y) = I_{m_o}(\underline{a}_n) = Q_{m_o}[f, 0, 1]$$

is minimized.

We have considered the basis functions

$$\phi_i(x) = x^i(1-x)$$

$$\phi_i(x) = x(1-x)T^*_{i-1}(x)$$

and

$$\phi_i(x) = \sin i\pi x$$

and have used eight and twelve point Gaussian quadrature schemes. The resulting functions, $I_{m_o}(\underline{a}_n)$, for various values of n, have been minimized using the Rosenbrock minimization algorithm (Rosenbrock (1); see also Palmer (1)) since in this case we prefer not to differentiate the integrand $f(x,y,y')$ with respect to y and y'. In Table XXXVI we present typical results of this procedure for various combinations of n, m_o and the basis $\phi_i(x)$. The minimum value $\bar{I}_{m_o}$ and the value of the corresponding function $z(x)$ at $x = 0.5$ are given. Table XXXVII records the progress of the procedure in the case $n = 4$, $m_o = 8$, $\phi_i(x) = x(1-x)T^*_{i-1}(x)$. The results are given at the end of each iteration of the Rosenbrock algorithm. N denotes the number of evaluations of the function $I_{m_o}(a_n)$ which have been performed. The asterisk * indicates that the step size parameter used in solving the sequence of one-dimensional minimization problems which occur in each iteration of the Rosenbrock algorithm was reduced after this point to enable increased accuracy to be obtained.

We contrast these results with those given by Allen (2), where the value $z(0.5) = 0.2233225 \pm 10^{-8}$ is given. We have applied one of the second order algorithms (approximation 5) due to Allen (1) to this problem and obtained the value $z(0.5) = 0.2281458$ with $h = 0.1$. Corresponding to the approximation given by this method we have obtained an approximation to $I(z)$ by application of Simpson's rule

A Comparison of Basis Functions in the Numerical Solution of Allen's Example by Minimization.

$\phi_i(x)$	n	m_o	$\bar{I}$	z(0.5)
$x^i(1-x)$	4	8	1.679943'+1	0.233556
	4	8	1.679617'+1	0.233797
$x(1-x)T^*_{i-1}(x)$	4	12	1.679617'+1	0.233797
	6	12	1.679614'+1	0.233876
sin i $\tilde{\Pi}$ x	4	12	1.679810'+1	0.234035

Table XXXVI

The Progress of Rosenbrock's Minimization Algorithm
in the Solution of Allen's Example, using a modified Chebyshev Basis

N	0	10	19	33	55 *	73	88
I(y)	+1.687511'+1	+1.681240'+1	+1.679749'+1	+1.679661'+1	+1.679624'+1	+1.679619'+1	+1.679617
$\underline{a}_n(i)$	-1.070000'-1	-8.700000'-2	-6.714442'-2	-6.986317'-2	-7.428655'-2	-7.327667'-2	-7.346390'-2
	0	-2.500000'-3	-6.196734'-3	-5.760375'-3	-5.179174'-3	-5.750956'-3	-6.206727'-3
	0	-2.500000'-3	-5.694190'-3	-6.993183'-3	-9.250436'-3	-8.710496'-3	-8.655659'-3
	0	+5.000000'-3	+5.798209'-3	+6.158782'-3	+6.484334'-3	+6.074610'-3	+5.919732'-3
z(0.5)	2.232500'-1	2.288749'-1	2.346374'-1	2.342824'-1	2.337409'-1	2.338584'-1	2.337979'-1

Table XXXVII

to the integration, taking $O(h^2)$ approximations to $z'(x_i)$ in terms of $z(x_i)$. The resulting value $(h = 0.1)$ is $\bar{I}(z) = 1.752381'+1$.

The disparity between our results for this problem and those of Allen, whilst not of major proportions, is not readily explained. We might suspect that in using an algebraic or trigonometric representation for $y(x)$, we are assuming that $y(x)$ has more continuous derivatives than is in fact the case, but if this were so then we should not be able to obtain a smaller value for $I(z)$ than by Allen's methods. Additionally, the assumption that Allen's methods are second order in h depends on the requirement that $y(x)$ has at least a Lipschitz bounded fourth derivative, and since Allen (1) has shown numerically that $O(h^2)$ convergence does occur, low order polynomial approximations can be justified. We therefore consider that we have shown that the combination of the techniques of the Rayleigh-Ritz method with those of numerical minimization algorithms prove a valuable alternative to 'discrete' methods such as those of Allen, for 'simplest' problems of the calculus of variations having smooth solutions.

6.4 Conclusions

It has been shown that a user of the Rayleigh-Ritz method for the numerical solution of two point boundary value problems in ordinary differential equations is presented with a number of choices which govern the success or failure of the method as he applies it. The stages in the application at which these choices must be made can be summarized as

i) The selection of a sequence of approximating subspaces,

ii) The selection of basis functions for each subspace of the chosen sequence,

iii) The evaluation of the elements of the Rayleigh-Ritz equations,

iv) The determination of the solution vector of the Rayleigh-Ritz equations,

v) The evaluation of the corresponding approximate solution,

vi) The estimation of the error of the approximation,

though we emphasize again that these choices are not independent. It has been the purpose of this investigation to evaluate criteria which have been proposed in the literature of the method by which such a choice may be governed, and to identify areas in which previously accepted criteria may be inadequate or too restrictive, and to suggest, on the basis of numerical and theoretical results, alternative and, we hope, more suitable criteria which may be applied.

The investigation has concentrated in parts on ordinary differential equations of second order, though much of the theory has been formulated for equations of arbitrary order, and some of it for partial as well as ordinary differential equations. We consider that second order ordinary differential equations are sufficient to demonstrate the effects of different choices at each of the six stages outlined above, at least for ordinary differential equations, and remark that this view seems widely held, for few examples of the application of the method to equations of higher order appear in the literature.

The selection of a sequence of subspaces in which to approximate the solution of the given equation should ideally be determined from a knowledge of the properties of this solution, for these, and particularly the degree of continuity of the solution, govern the rate of convergence of the Rayleigh-Ritz approximations, as is

indicated by the results of Ciarlet, Schultz, Varga (1), and others, described in Chapter Two. We have chosen to emphasize in this work the situation in which the solution can be assumed to be reasonably smooth, and high accuracy is required, and are thus led to the choice of subspaces of polynomial functions. We acknowledge that if the solution is known to have points of irregular behaviour, and particularly if the solution $y(x)$ only satisfies $y(x) \in C^t[0,1]$ for some t, $k \leq t \leq 2k$, then provided the points at which this irregular behaviour occurs can be isolated, and included in the mesh points of a piecewise approximation, this approximation will in general be more satisfactory than any polynomial approximation. Equally, if a solution is known to be periodic, it should be approximated in terms of periodic functions, whether they be periodic spline functions (Ciarlet, Schultz, Varga (2)) or trigonometric functions depending on the continuity properties of the solution.

The choice of a sequence of approximating subspaces being made, we have shown that an appropriate choice of co-ordinate system for each subspace is still required, and is governed by more stringent requirements than those of admissibility and independence. However, we have demonstrated that the criteria proposed by Mikhlin to ensure complete stability are unnecessarily restrictive for many practical problems, and have indicated a preference for the informal notion of 'sufficient stability' introduced by Samokish, which requires a choice of co-ordinate system in which rapid convergence may be expected. Numerical investigations of co-ordinate systems satisfying these criteria for second order ordinary differential equations have led us to derive an efficient evaluation algorithm for the co-ordinate system of integrated Legendre polynomials, which satisfies Mikhlin's criteria for these problems and which has been used by Mikhlin (4)

and Ciarlet, Schultz and Varga (1), and to a theoretical investigation of a co-ordinate system which has been shown by experiment to satisfy the less restrictive conditions of Samokish (1). For this system, based on Chebyshev polynomials, we have shown that for a certain class of problems the Rayleigh-Ritz matrix assumes a particularly convenient form, and have established that it is uniformly asymptotically diagonal of degree three, allowing the theorems of Delves & Mead (2) to be utilized to derive asymptotic convergence results for the resulting Rayleigh-Ritz approximations to the solution of linear equations. The results of Delves and Mead have been extended to apply in the case when this co-ordinate system is used to solve mildly non-linear equations, allowing asymptotic convergence results to be deduced for these also. This modified Chebyshev co-ordinate system has been used to solve successfully a number of test problems taken from the literature, including problems with a weak boundary singularity (Mayers (1)) and problems satisfying the extended existence results of Gladwell (1). In this connection the successful solution of problems to which no known existence results apply should be mentioned, indicating that further extensions of the existence theory for the Rayleigh-Ritz method are possible. We conclude that these modified Chebyshev polynomials represent a useful alternative co-ordinate system to the integrals of the Legendre polynomials, for second order ordinary differential equations having 'smooth' solutions.

A scheme for the construction of a co-ordinate system based on the orthonormalization of a prescribed co-ordinate system with respect to the energy norm of the co-ordinate functions, suggested by Davis and Rabinowitz (1), has been investigated. If the

orthonormalization could be performed exactly the resulting co-ordinate system would satisfy the criteria of Mikhlin, and thus be completely stable. In practice, the orthonormalization of the functions must be performed numerically, and it has been demonstrated by example that the stability properties of the resulting co-ordinate system are identical with those of the original prescribed system, and theoretical reasons for this have been investigated. It has been shown that the orthonormalization of the co-ordinate system corresponds to a method of solution of the Rayleigh-Ritz equations which is known to be less economical of computer time than standard methods. Since the accuracy of the resulting approximation is not increased the suggestion that orthonormalization of this form could be a powerful tool in variational calculation has to be rejected.

In the literature of the Rayleigh-Ritz method the effects of the quadrature approximations which are frequently made in the evaluation of elements in the Rayleigh-Ritz equations has previously been neglected, with the exception of the contributions of Herbold (1) and Herbold, Schultz and Varga (1). We have been able to modify some of these results, which are applicable when the co-ordinate system used consists of piecewise polynomials and only the free term vector is approximated by quadrature, to the case of a polynomial co-ordinate system. Additionally, extensions to the case in which quadrature approximations to the elements of both the matrix and the free term vector have been derived. It has unfortunately not been possible to extend all the results in this manner, and in particular we have been unable to show that the convergence rate which applies when the sequence of Rayleigh-Ritz approximations is computed exactly is preserved when quadrature errors are present.

However, this property has been shown to hold for certain test examples using a polynomial co-ordinate system and particular choices of quadrature rule.

We have indicated our belief that the error bounds given in the literature for the Rayleigh-Ritz method are of theoretical rather than practical significance, and in general are not easily computed. Accordingly, we express the opinion that simple and previously popular error estimates based on the magnitude of the coefficients in the Rayleigh-Ritz expansion have to be accepted. Estimates developed from the theory of Delves and Mead (2) are shown to be extremely pessimistic for a number of test examples. However, the behaviour of a sequence of these estimates provides a guide to the growth of rounding error in the Rayleigh-Ritz process, and this may have application in the selection of a particular member of the sequence of Rayleigh-Ritz approximations as a final solution.

In the last chapter extensions of aspects of the previous work to partial differential equations are considered. An approximate finite element method for second order elliptic equations, derived by applying an extension of the work of Herbold (1) for partial differential equations analogous to that used in Chapter Five to extend his work for ordinary differential equations, has been proposed. This method has a simple form, and in particular is more easily applied to general operators than standard finite element techniques. Further investigation of this approach is clearly necessary, both from a practical and a theoretical aspect, but in particular it is hoped that uniqueness, convergence and consistency results similar to those of Herbold (1) may also be derived in this case.

A unified description of the Kantorovich and 'extended Kantorovich' methods as applied to second order linear partial differential equations in two dimensions involving a positive definite differential operator has been presented, and extensions to other equations involving a positive definite differential operator indicated. An approximate method which arises in a natural manner from these methods has been outlined. It has been emphasized that the successful application of this method rests on the criteria which govern the application of the Rayleigh-Ritz method to ordinary differential equations, and the relevance of earlier discussions has thus been extended to a wider field.

An attempt to extend the Rayleigh-Ritz method to general formulations of variational problems of the so-called 'simplest' type, based on algorithms for the minimization of a function of n variables, has also been considered. Further research into this approach is required before its difficulties can be isolated, but in particular the high overhead of quadrature approximation is again a serious practical disadvantage.

In conclusion we comment on the usefulness of the Rayleigh-Ritz algorithm for the solution of the problems with which we have been primarily concerned, two-point boundary value problems in ordinary differential equations, in comparison with other algorithms. Broadly speaking, numerical algorithms for these problems approximate the solution in one of two ways, either in terms of pivotal values (values at a sequence of points) or as a series expansion of the approximate solution in terms of prescribed functions. These classifications are not independent, and we have seen that for the Rayleigh-Ritz method a particular choice of prescribed piecewise

polynomial co-ordinate system for a method regarded as being of the second type generates an approximation considered to be a classical example of methods of the first type. Nonetheless, we preserve elements of this classification in our brief comparison of these methods.

A choice between type of method must be made on the basis of required properties of the approximate solution and on the nature of the problem, and its exact solution. If the value of the approximate solution may be required at many arbitrary points, then a series expansion is frequently to be preferred. Further, we have seen that if the exact solution of a problem is sufficiently smooth, polynomial expansions are appropriate, and whilst this remark has been made here in the context of the Rayleigh-Ritz method it remains largely true for other series expansion methods. Pivotal methods, whether of the 'shooting' or 'boundary' type, are well suited to problems with solutions which are not known to be smooth, and these methods, particularly 'shooting' methods are often very readily applied to very general ordinary differential equations.

Of the methods for the determination of series expansion approximations to the solution of ordinary differential equations we comment only upon the more popular ones. These are the direct power series expansion method, the method of collocation and the least squares method in both its continuous and discrete forms, together with the Rayleigh-Ritz or Galerkin methods.

The direct power series expansion method, usually applied in terms of Chebyshev series (Lanczos (3)) is regarded as a method appropriate to the solution of linear equations with particularly simple coefficients, and is out of place when considered for a

wider class of problems.

The least squares method may be applied in two forms, corresponding to the choice of an integral least squares norm, or a discrete semi-norm. In the former case the resulting least squares equations for the linear differential equation $Ly = f$ are equivalent to those obtained by applying the Rayleigh-Ritz method to the differential equation $L^*Ly = L^*f$ where L^* denotes the conjugate operator to L (Mikhlin and Smolitskiy (1), p.226). In view of this similar difficulties regarding the selection of co-ordinate systems can be encountered in the least squares method and in the Rayleigh-Ritz method. However, one of the criteria applied to this selection of the co-ordinate system in the Rayleigh-Ritz case, that $\| Ly_n - f \| \to 0$ as $n \to \infty$, where $\| \cdot \|$ denotes a norm other than the energy norm, does not apply in the least squares case, where $\| Ly_n - f \|_2 \to 0$ as $n \to \infty$ from the definition of the least squares approximation. On the other hand, convergence of the least squares approximation $y_n(x)$ to the true solution $y(x)$ requires more restrictive differentiability assumptions on the true solution $y(x)$ and on elements of the co-ordinate system than is the case for Rayleigh-Ritz calculations.

The difficulties of numerical quadrature which are apparent in the evaluation of matrix and vector elements for both the Rayleigh-Ritz and continuous least squares methods are avoided by the discrete least squares method and by the collocation method (Wright (1), Clenshaw and Norton (1)), both of which are powerful algorithms for the solution of problems of the type we have considered. In particular there are powerful convergence results for the collocation method (Karpilovskaja (1)), for linear equations, comparable with the

results convergence of the Rayleigh-Ritz method using polynomial basis functions, given in Chapter Two. These include convergence results for the derivatives of the approximate solution up to and including the order of the differential equation. In view of these results, and the well known observation that the collocation method is a particular case of the discrete least squares method, we acknowledge the efficiency of the collocation method for linear equations. For non-linear equations such results have not been found in the literature.

We have shown therefore the power of the Rayleigh-Ritz method for the solution of linear and mildly non-linear boundary value problems in ordinary differential equations, and extended known results for this method in a number of directions. Though extensions of some of these results into the field of partial differential equations have been outlined, it is clear that further research in these areas remains to be undertaken.

References

(Items marked ** have not been consulted in the original)

Abramowitz and Stegun (1)
'Handbook of Mathematical Functions'.
Dover Publ. 1965

Allen (1)
'An investigation into direct numerical methods for solving some calculus of variations problems. Part 1. Second order methods'.
Computer Journal 9 p.205 1966

Allen (2)
'An investigation into direct numerical methods for solving some calculus of variations problems'.
Part 2. Fourth order methods'.
Unpublished paper.

Andersenn (1)
'The numerical solution of parabolic partial differential equations by variational methods'.
Numerische Mathematik 13 p.129 1969

Andersenn (2)
'A stability analysis of the extended Kantorovich method applied to the torsion problem'.
Numerische Mathematik 17 p.239 1971

Babuska, Prager and Vitasek (1)
'Numerical Processes in Differential Equations'.
Wiley (Prague) 1966

Bauer (1)
'Optimally scaled matrices'
Numerische Mathematik 5 p.73 1963

Bellman (1)
'Dynamic Programming'
Princeton University Press 1957

Bellman and Dreyfus (1)
'Applied Dynamic Programming'
Princeton University Press 1962

Birkhoff and Fix (1)**
'Rayleigh-Ritz approximation with trigonometric polynomials' (to appear)

Birkhoff, Schultz and Varga (1)
'Piecewise Hermite interpolation in one and two dimensions with applications to partial differential equations' Numerische Mathematik 11 p.232 1968

Bliss (1)
'Lectures on the Calculus of Variations'
Chicago University Press 1946

Bolza (1)**
'Vorlesungen über Variationsrechnung'
Leipzig 1909

Bolza (2)
'Lectures on the Calculus of Variations'
Chicago University Press 1904

Bramble and Hubbard (ed) (1)
'SYNSPADE : Proceedings of a Symposium on the Numerical Solution of Partial Differential Equations'
University of Maryland 1970

Ciarlet, Schultz and Varga (1)
'Numerical methods of high order accuracy for the solution of boundary value problems. I. One dimensional problem.
Numerische Mathematik 9 p.394 1967

Ciarlet, Schultz and Varga (2)
'Numerical methods of high order accuracy for the solution of boundary value problems. IV. Periodic Boundary Conditions'.
Numerische Mathematik 12 p.266 1968

Clenshaw (1)
'Curve fitting with a digital computer'
Computer Journal 2 p.170 1959

Clenshaw and Norton (1)
'The solution of non-linear ordinary differential equations in Chebyshev series'.
Computer Journal 6 p.88 1962

Collatz (1)
'The Numerical Treatment of Differential Equations'
Springer Verlag 1960

Courant (1)
'Advanced Methods in Applied Mathematics'
New York University Press 1941

Courant (2)
'Variational methods for the solution of problems of equilibrium and vibrations'
Bulletin of the American Mathematical Society
49 p.1 1943

Davis and Rabinowitz (1)
'Advances in orthonormalizing computation'.
Advances in Computers 2 p.56 1961

Delves (1)
'Round-off error in variational calculations'.
Journal of Computational Physics 3 p.17 1968

Delves (2)
'Lectures on variational methods'
Universities of Liverpool and Manchester
Summer School 1971

Delves and Mead (1)
'On the convergence rates of variational methods'
Proceedings of the Conference on the Numerical Solution of Differential Equations. Dundee 1969. Springer Verlag 1969

Delves and Mead (2)
'On the convergence rates of variational methods. I. Asymptotically diagonal systems'.
Mathematics of Computation 25 p.699 1971

Delves and Mead (3)
'On the convergence rate of generalized Fourier expansions' To appear.

Dovbysh (1)**
'A note on minimal systems'
Trudy Matem in-ta im. V.A. Steklov 96 p.188 1968

Dreyfus (1)
'Dynamic Programming and the Calculus of Variations'
Academic Press 1965

Dzhishkariani (1)
'The least squares and Bubnov-Galerkin methods'.
U.S.S.R. Journal of Computational Mathematics and Mathematical Physics 8 5, p.235 1968
(English translation 1971)

Elsgolc (Els'gol'tz) (1)
'The Calculus of Variations'.
International Series of Monographs in Pure and Applied Science 1961

Fletcher and Powell (1)
'A rapidly convergent descent method for minimization'
Computer Journal 6 p.163 1963

Fox, C. (1)
'The Calculus of Variations'
Oxford University Press 1950

Fox, L. and Parker (1)
'Chebyshev Polynomials in Numerical Analysis'
Oxford University Press 1968

Freidrichs (1)**
'Spektraltheorie halbbeschrankler operatoren und anwendung auf die spectralzerlegung von differentialoperatoren'.
Math.Ann 109 p.456 1934

George (1)
'Computer implementation of the finite element method'
Stanford University report STAN-CS-71-208 1971

Gladwell (1)
'Rayleigh-Ritz methods for non-linear boundary value problems'.
Journal of the Institute of Mathematics and its Applications (to appear)

Gould (1)
'Variational Methods for Eigenvalue Problems'
University of Toronto Press 1957

Gumowski and Mira (1)
'Optimization in Control Theory and Practice"
Cambridge University Press 1968

Hadamard (1)
'Lecons sur le calculus de variations'
Hermann 1921

Herbold (1)
'Consistent Quadrature Schemes for the Numerical Solution of Boundary Value Problems by Variational Techniques'
Doctoral Thesis. Case Western Reserve University 1968

Herbold, Schultz and Varga (1)
'The effect of quadrature errors in the numerical solution of boundary value problems by variational techniques'.
Aequationes Mathematiae 3 p.247 1969

Hestenes (1)
'The Calculus of Variations and Optimal Control Theory'.
Wiley 1966

Hilbert (1)**
'Lectures' 1899
See:- Whittemore, 'Annals of Mathematics'.
Vol.II p.132 1901

Kaczmarz and Steinhous (1)
'Theorie der Orthogonalreihen'
Chelsea Publ. 1951

Kantorovich (1)
'Functional Analysis and Applied Mathematics'
(Russian, 1948)
English translation: National Bureau of Standards

Kantorovich and Krylov (1)
'Approximate Methods of Higher Analysis'.
Nordhoff 1958

Karpilovskaja (1)
'Convergence of the collocation method'.
Sov.Math. 4 p.1070 1963

Kerr (1)
'An extension of the Kantorovich method'.
Quarterly of Applied Mathematics 26 p.219 1968

Kolar, Kratochvil, Zlamal and Zenicek (1)
'Technical, Physical and Mathematical Principles of the Finite Element Method'.
Academia Publ. Prague 1971

Lanczos (1)
'Variational Principles of Mechanics'.
Toronto University Press 1949

Lanczos (2)
'Linear Differential Operators'
Van Nostrand Press 1961

Lanczos (3)
'Applied Analysis'.
Pitman 1957

Liusternik and Sobolev (1)
'Elements of Functional Analysis'
Ungar 1961

Mayers (1)
'The deferred approach to the limit in ordinary differential equations'.
Computer Journal 7 p.54 1964

McDonald (1)
'Solution of the incompressible boundary layer equations by the Galerkin-Kantorovich technique'.
Journal of the Institute of Mathematics and its Applications 6 p.115 1970

Mikhlin (1)
'The Numerical Realization of Variational Methods'.
(Russian). Nauka Moscow 1966

Mikhlin (2)
'Variational Methods in Mathematical Physics'.
(Russian) Gostelihizdat Moscow 1957
English translation:- McMillan 1964

Mikhlin (3)
'The Problem of the Minimum of a Quadratic Functional'.
Holden Day 1964

Mikhlin (4)
'The Numerical Performance of Variational Methods'.
Nordhoff 1971

Mikhlin (5)
'Some properties of polynomial approximation in the sense of Ritz'
(Russian) Dokl.Acad.Nauk 180 p.2 1968
English translation:- Sov.Math.Dokl. 3 p.614 1968

Mikhlin and Smolitskiy (1)
'Approximate Methods for the Solution of Differential and Integral Equations'.
Elsevier 1967

Noble (1)
'Applied Linear Algebra'.
Prentice Hall 1969

Ortega and Rheinboldt (1)
'Iterative Solution of Nonlinear Equations in Several Variables'.
Academic Press 1970

Palmer (1)
'An improved procedure for orthogonalizing the search vectors in Rosenbrock's and Swann's direct search optimization methods'.
Computer Journal 12 p.69 1969

Perrin, Price and Varga (1)
'On higher order numerical methods for nonlinear two point boundary value problems'.
Numerische Mathematik 13 p.180 1969

Powell (1)
'A survey of numerical methods for unconstrained optimization'.
S.I.A.M. Review 12 p.79 1970

Rayleigh (1)
'Theory of Sound'
McMillan 1896

Reid (1)
'Lectures on the finite element method'.
Universities of Liverpool and Manchester Summer School 1971

Reid (2)
'On the construction and convergence of a finite element solution of Laplace's equation'.
Technical Report T.P. 436 U.K.A.E.A. Harwell 1971

Rice (1)
'Experiments on Gramm-Schmidt orthogonalization'.
Mathematics of Computation 20 p.325 1966

Ritz (1)**
'Uber eine neue methode zur loesung gewisser variationsprobleme der mathematischen physik.
Journal Reine Agnew Math. 135 1908

Ritz (2)**
'Theory der transwersalchwingungen einer quadratishen platte mit freien raendern'.
Annalen der Physic 38 1909

Rosenbrock (1)
'An automatic method for finding the greatest or least value of a function'.
Computer Journal 3 p.175 1960

Samokish (1)**
'On the stability of the abstract method of Galerkin'.
(Russian) Vestn. Leningrad Gos. University ser matem mekh i astr 1 p.160 1964

Sansone (1)
'Orthogonal Functions'
Interscience 1959

Schoenberg (ed) (1)
'Approximation, with special emphasis on spline functions'
Academic Press 1969

Schultz (1)
'Multivariate spline functions and elliptic problems'.
in Schoenberg (ed) (1) 1969

Schultz (2)
'Error bounds for the Rayleigh-Ritz-Galerkin method'.
Journal of Mathematical Analysis and Applications 27 p.524 1969

Schultz and Varga (1)
'L-splines'.
Numerische Mathematik 10 p.345 1967

Stepleman (1)
'Finite dimensional analogues of variational problems in the plane'.
S.I.A.M. Journal of Numerical Analysis p.11 1971

Stroud (1)
'Approximate Calculation of Multiple Integrals'.
Prentice Hall 1971

Timoshenko (1)
'Vibration Problems in Engineering'
Van Nostrand 1928

Vainniko (1)
'On similar operators'.
(Russian) Dokl. Akad.Nauk. 179 1968
English translation:- Sov.Math.Dokl. 9 p.477 1968

Van der Sluis (1)
'Condition number and equilibration of matrices'.
Numerische Mathematik 14 p.14 1969

Van der Sluis (2)
'Conditioning, equilibration and pivoting in linear algebraic systems'.
Numerische Mathematik 15 p.74 1970

Von Karman and Biot (1)
'Mathematical Methods in Engineering'
McGraw Hill 1940

Whiteman (1)
'A bibliography for finite element methods'.
Dept. of Mathematics, Brunel University 1972

Wilkinson (1)
'Rounding Error in Algebraic Processes'.
H.M.S.O. 1963

Wright (1)
'Chebyshev collocation methods for ordinary differential equations'.
Computer Journal 6 p.358 1964

Zienkiewicz and Cheung (1)
'The Finite Element Method in Structural and Continuum Mechanics'.
McGraw Hill 1967

Appendix A

Additional Numerical Experiments concerning the Stability of the Rayleigh-Ritz method for Linear Differential Equations

Behaviour of the Solution Vector and the Approximate Solution

Problem L2 $\quad \phi_i = \frac{\sqrt{2}}{i\pi} \sin i\pi x$

n	2	3	7	8	11	12	Exact
$\bar{e}_n$	3.067046'-2	8.543193'-3	2.059637'-3	2.059757'-3	5.676150'-4	5.675553'-4	
e_n	3.067016'-2	8.543073'-3	2.059101'-3	2.059101'-3	5.638599'-4	5.638599'-4	
	-1.258719'+0	-1.258719'+0	-1.258719'+0	-1.258719'+0	-1.258719'+0	-1.258719'+0	-1.258719'+0
	+2.551434'-7	+3.906468'-7	+4.369090'-7	+4.369123'-7	+4.871040'-7	+4.871049'-7	0
		-1.880712'-1	-1.880712'-1	-1.880712'-1	-1.880712'-1	-1.880712'-1	-1.880707'-1
			+1.325641'-6	+1.325642'-6	+1.364883'-6	+1.364884'-6	0
			-6.962626'-2	-6.962626'-2	-6.962633'-2	-6.962633'-2	-6.962561'-2
$a_n^n(i)$			+3.415361'-6	+3.415366'-6	+3.489104'-6	+3.489023'-6	0
			-3.580279'-2	-3.580279'-2	-3.580123'-2	-3.580123'-2	-3.580304'-2
				+2.202146'-6	+2.296201'-6	+2.297730'-6	0
					-2.176207'-2	-2.176207'-2	-2.172907'-2
					+1.693847'-6	+1.678460'-6	0
					-1.429344'-2	-1.429344'-2	-1.456988'-2
						+6.771289'-7	0

Table A.I

Behaviour of the Solution Vector and the Approximate Solution

Problem L2 $\quad \phi_i = x^i(1-x)$

n	2	3	7	8	11	12
$\bar{e}_n$	8.018374'-3	6.729363'-5	8.434054'-5	4.559753'-5	8.267164'-5	8.088347'-5
$a_n''(i)$	-2.204382'+0	-2.348120'+0	-2.345106'+0	-2.347199'+0	-2.343614'+0	-2.343832'+0
	-2.938764'-5	+7.042962'-1	+6.031696'-1	+6.257339'-1	+5.469805'-1	+5.496667'-1
		-7.043280'-1	+2.168700'-1	+3.652343'-1	+7.863547'-1	+8.218269'-1
			-3.492091'+0	-5.729425'+0	-5.296747'+0	-6.221701'+0
			+6.312297'+0	+1.518153'+1	+4.583859'+0	+1.162742'+1
			-5.402607'+0	-2.127252'+1	+1.663523'+1	-9.521029'+0
			+1.760405'+0	+1.513413'+1	-4.657511'+1	+5.466616'+0
				-4.310890'+0	+4.025767'+1	-1.076542'+1
					-1.072923'+0	+6.026103'+0
					-1.647729'+1	+1.732836'+1
					+6.604581'+0	-2.377923'+1
						+8.461565'+0

Table A.II

Behaviour of the Solution Vector and the Approximate Solution

Problem L3 $\qquad \phi_i = \frac{\sqrt{2}}{i\pi} \sin i\pi x$

n	2	3	7	8	11	12
$\bar{e}_n$	9.993731'-3	2.412616'-3	2.801417'-4	2.908110'-4	2.669095'-4	2.600550'-4
$a_n^n(i)$	+4.638341'-1	+4.649453'-1	+4.650508'-1	+4.650483'-1	+4.650617'-1	+4.650608'-1
	-6.225671'-3	+1.306083'-3	+1.774430'-3	+1.772947'-3	+1.813214'-3	+1.812832'-3
		+5.296903'-2	+5.324144'-2	+5.323769'-2	+5.326077'-2	+5.325992'-2
			+1.448819'-3	+1.446035'-3	+1.508326'-3	+1.507817'-3
			+1.933184'-2	+1.937292'-2	+1.941706'-2	+1.941581'-2
			+6.446225'-4	+5.347889'-4	+7.778537'-4	+7.770825'-4
			+9.860436'-3	+9.797393'-3	+9.942493'-3	+9.940273'-3
				+4.550586'-4	+4.405113'-4	+4.380696'-4
					+6.022375'-3	+6.016566'-3
					+2.348949'-4	+2.360584'-4
					+3.922856'-3	+3.883826'-3
						-3.211936'-4

Table A.III

Behaviour of the Solution Vector and the Approximate Solution

Problem L3 $\phi_i = x^i(1-x)$

n	2	3	7	8	11	12
$\bar{e}_n$	3.964677'-4	2.598694'-4	1.155449'-4	1.156282'-4	1.156184'-4	1.156183'-4
$a_n''(i)$	+8.160603'-1	+8.211416'-1	+8.203907'-1	+8.128460'-1	+1.100451'+0	+8.252407'-1
	-8.705347'-3	-3.181756'-2	+6.843547'-2	-2.975450'-1	-8.021972'+0	-6.322640'-2
		+2.183716'-2	-1.007884'+0	-3.270527'+0	+7.217375'+1	+1.315409'-1
			+3.896978'+0	+1.425488'+1	-2.906025'+2	-5.585981'-1
			-6.881056'+0	-3.187196'+1	+5.819848'+2	+1.670747'+0
			+5.772976'+0	+3.857339'+1	-5.773002'+2	-1.107191'+0
			-1.864398'+0	-2.400769'+1	+2.589934'+2	-2.665423'+0
				+6.019235'+0	-5.425202'+1	+3.101179'+0
					-3.514027'+1	+1.242385'+0
					+1.238805'+2	-5.173576'-1
					-7.219279'+1	-3.154259'+0
						+1.905414'+0

Table A.IV

Behaviour of the Solution Vector and the Approximate Solution

Problem L2 $\qquad \phi_i = \sin i\widetilde{\Pi} x$

n	2	3	7	8	11	12	Exact
$\bar{e}_n$	3.067016'-2	8.543014'-3	2.059400'-3	2.059518'-3	5.679726'-4	5.679130'-4	
$a_n^n(i)$	-5.666235'-1	-5.666235'-1	-5.666235'-1	-5.666235'-1	-5.666235'-1	-5.666235'-1	-5.666234'-1
	+1.026310'-7	+1.362118'-7	+1.474327'-1	+1.474336'-1	+1.587446'-7	-1.587449'-7	0
		-2.822056'-2	-2.822054'-2	-2.822054'-2	-2.822054'-2	-2.822054'-2	-2.822057'-2
			+1.487271'-7	+1.487273'-7	+1.518779'-7	+1.518781'-7	0
			-6.268568'-3	-6.268568'-3	-6.268579'-3	-6.268579'-3	-6.268515'-3
			+2.509243'-7	+2.509246'-7	+2.566941'-7	+2.566873'-7	0
			-2.302469'-3	-2.302469'-3	-2.302368'-3	-2.302368'-3	-2.302434'-3
				+1.273960'-7	+1.328447'-7	+1.329424'-7	0
					-1.088488'-3	-1.088488'-3	-1.086836'-3
					+6.709620'-8	+6.630965'-8	0
					-5.849362'-4	-5.849362'-4	-5.962506'-4
						+2.884520'-8	0

Table A.V

Behaviour of the Solution Vector and the Approximate Solution

Problem L2

$$\phi_i = \frac{(i+1)^{i+1}}{i^i} x^i(1-x)$$

n	2	3	7	8	11	12
$\bar{e}_n$	8.018791'-3	6.681679'-5	4.929300'-5	8.100269'-5	1.480103'-3	5.662437'-5
$a_n''(i)$	-5.510975'-1	-5.870326'-1	-5.867050'-1	-5.874324'-1	-6.182686'-1	-5.861266'-1
	-2.675539'-6	+1.043433'-1	+9.633236'-2	+1.126902'-1	+7.352089'-1	+8.435368'-2
		-7.428562'-2	-1.846647'-2	-1.516352'-1	-5.025074'+0	+7.599061'-2
			-1.692720'-1	+3.573606'-1	+1.988800'+1	-5.423505'-1
			+2.526790'-1	-8.747731'-1	-4.557829'+1	+1.080667'+0
			-1.827185'-1	+1.153470'+0	+6.253292'+1	-1.288023'+0
			+5.115836'-2	-7.735217'-1	-5.329187'+1	+1.014116'+0
				+2.068609'-1	+3.189691'+1	-6.308184'-1
					-1.799464'+1	+3.334127'-1
					+9.296372'+0	-1.106451'-1
					-2.398888'+0	+1.454233'-2
						-2.130320'-3

Table A.VI

Behaviour of the Solution Vector and the Approximate Solution

Problem L3 $\quad \phi_i = \sin i\pi x$

n	2	3	7	8	11	12
$\bar{e}_n$	9.994327'-3	2.413152'-3	2.786516'-4	2.890825'-4	2.664327'-4	2.595782'-4
	+2.087971'-1	+2.092974'-1	+2.093446'-1	+2.093434'-1	+2.093492'-1	+2.093488'-1
	-1.401453'-3	+2.937775'-4	+3.991916'-4	+3.988593'-4	+4.079239'-4	+4.078368'-4
		+7.948096'-3	+7.988974'-3	+7.988419'-3	+7.991891'-3	+7.991756'-3
			+1.630371'-4	+1.627237'-4	+1.697341'-4	+1.696766'-4
			+1.744966'-3	+1.744163'-3	+1.748138'-3	+1.748024'-3
$a_n''(i)$			+4.835539'-5	+4.761746'-5	+5.835093'-5	+5.829300'-5
			+6.341049'-4	+6.300499'-4	+6.393810'-4	+6.392381'-4
				-2.561006'-5	+2.478292'-5	+2.464550'-5
					+3.012220'-4	+3.009314'-4
					+1.057282'-5	+1.062521'-5
					+1.605317'-4	+1.589340'-4
						-1.205196'-5

Table A.VII

Behaviour of the Solution Vector and the Approximate Solution

Problem L3

$$\phi_i = \frac{(i+1)^{i+1}}{i^i} x^i(1-x)$$

n	2	3	7	8	11	12
$\bar{e}_n$	3.973841'-4	2.596974'-4	1.235604'-4	1.617074'-4	1.828669'-4	2.561211'-4
	+2.040166'-1	+2.052860'-1	+2.065733'-1	+2.067685'-1	+2.085866'-1	+2.032431'-1
	-1.289450'-3	-4.717405'-3	-1.082940'-2	-1.335302'-2	-4.088308'-2	+4.349186'-2
		+2.305782'-3	+7.458947'-3	+1.861542'-2	+1.772055'-1	-3.218552'-1
			+1.720487'-2	-1.081941'-3	-4.340974'-1	+9.542376'-1
			-4.092942'-2	-4.328237'-2	+4.946390'-1	-1.270512'+0
$a_n^n(i)$			+3.336322'-2	+7.499170'-2	-6.647415'-3	+3.320995'-1
			-9.997639'-3	-5.531494'-2	-5.625248'-1	+1.112635'+0
				+1.550595'-2	+4.936668'-1	-1.533685'+0
					-8.754467'-2	+1.096158'+0
					-5.737963'-2	-5.845233'-1
					+1.782958'-2	+1.752035'-1
						-3.627368'-3

Table A.VIII

Behaviour of the Solution Vector and the Approximate Solution

Problem L2

$\phi_i = x(1-x)\, T^*_{i-1}(x)$

n	2	3	7	8	11	12
$\bar{e}_n$	7.512092'-3	6.848570'-5	1.609325'-6	2.443790'-6	2.980231'-6	2.741811'-6
$a''_n(i)$	-2.204399'+0	-2.260107'+0	-2.260634'+0	-2.260634'+0	-2.260646'+0	-2.260646'+0
	-3.046706'-6	-3.288793'-6	-2.032014'-5	-4.714167'-5	-8.884820'-5	-1.451001'-4
		-8.805703'-2	-8.903860'-2	-8.903860'-2	-8.906525'-2	-8.906531'-2
			-1.481899'-5	-4.010117'-5	-8.025800'-5	-1.350449'-4
			-7.269900'-4	-7.269906'-4	-7.521975'-4	-7.522000'-4
			-8.882419'-6	-3.077652'-5	-6.764875'-5	-1.193898'-4
			-3.819528'-6	-3.819528'-6	-2.635270'-5	-2.635325'-5
				-1.557500'-6	-4.690210'-5	-9.371082'-5
					-1.810425'-5	-1.810511'-5
					-2.199219'-5	-6.129839'-5
					-9.370720'-6	-9.371147'-6
						-2.735282'-5

Table A.IX

Behaviour of the Solution Vector and the Approximate Solution

Problem L2 $\phi_i = x(1-x)P^*_{i-1}(x)$

n	2	3	7	8
$\bar{e}_n(x)$	8.018791'-3	6.818769'-5	1.370906'-6	1.490116'-6
$a''_n(i)$	-2.204399'+0	-2.230754'+0	-2.230906'+0	-2.230906'+0
	-3.247331'-6	-3.460043'-6	-6.737194'-6	-7.879636'-6
		-1.174072'-1	-1.181630'-1	-1.181630'-1
			-7.404697'-6	-1.007034'-5
			-1.328536'-3	-1.328536'-3
			-3.97102'-6	-8.159448'-6
			-9.592214'-6	-9.592258'-6
				-5.673747'-6

Table A.X

Behaviour of the Solution Vector and the Approximate Solution

Problem L2

$$\phi_i(x) = \sqrt{2i+1}\int_0^x P_i(t)\,dt$$

n	2	3	7	8
$\bar{e}_n(x)$	8.019148'-3	6.800886'-5	1.072883'-6	+1.251697'-6
	+1.272711'+0	+1.274372'+0	1.274372'+0	+1.274372'+0
	+1.484564'-6	+1.575626'-6	+1.621495'-6	+1.621509'-6
		+5.325417'-2	+5.326225'-2	+5.326225'-2
			+2.304838'-6	+2.304959'-6
$a_n^n(i)$			+6.660175'-4	+6.660175'-4
			+1.968203'-6	+1.981231'-6
			+7.112133'-6	+7.112155'-6
				+2.934978'-6

Table A.XI

Behaviour of the Solution Vector and the Approximate Solution

$$\phi_i(x) = \int_0^x P_i(t)\,dt$$

Problem L2

n	2	3	7	8
$\bar{e}_n(x)$	8.018791'-3	6.830691'-5	1.311302'-6	1.490116'-6
	+2.204399'+0	+2.207274'+0	+2.207274'+0	+2.207274'+0
	+3.334250'-6	+3.537141'-6	+3.638447'-6	+3.638483'-6
		+1.408944'-1	+1.409158'-1	+1.409158'-1
			+6.817412'-6	+6.817807'-6
$a_n^n(i)$			+2.209962'-3	+2.209962'-3
			+7.041765'-6	+7.092304'-6
			+2.857999'-5	+2.858006'-5
				+1.301981'-5

Table A.XII

Behaviour of the Solution Vector and the Approximate Solution

Problem L2 — Orthonormal Basis from $\phi_i = x^i(1-x)$

n	2	3	7	8	11	12
$\bar{e}_n(x)$	8.018493'-3	6.014105'-5	3.486871'-5	6.479023'-5	6.163716'-4	5.082487'-4
	-1.505888'+0	-1.505888'+0	-1.505888'+0	-1.505888'+0	-1.505888'+0	-1.505888'+0
	-8.583068'-6	-8.583068'-6	-8.583068'-6	-8.583068'-6	-8.583-68'-6	-8.583068'-6
		-5.437374'-2	-5.437374'-2	-5.437374'-2	-5.437374'-2	-5.437374'-2
			-4.005432'-5	-4.005432'-5	-4.005432'-5	-4.005432'-5
			-6.103515'-4	-6.103515'-4	-6.103515'-4	-6.103515'-4
$\underline{a}''_n(i)$			-1.525879'-4	-1.525879'-4	-1.525879'-4	-1.525879'-4
			-3.051757'-4	-3.051757'-4	-3.051757'-4	-3.051757'-4
				+9.307856'-4	+9.307856'-4	+9.307856'-4
					-2.578735'-3	-2.578735'-3
					-4.048346'-3	-4.048346'-3
					+1.039505'-3	+1.039505'-3
						+3.342627'-3

Table A.XIII

Orthogonal Combination Matrix C^T : n = 7

Problem L2 $\phi_i = x^i(1-x)$

+1.463850'+0	-2.136630'+0	+2.643342'+0	-3.004469'+0	+3.314711'+0	-3.510157'+0	+2.592614'+0
	+4.273274'+0	-1.295247'+1	+2.674147'+1	-4.615870'+1	+7.034625'+1	-8.289505'+1
		+1.295250'+1	-6.219874'+1	+1.842778'+2	-4.224670'+2	+7.631538'+2
			+4.146691'+1	-2.763041'+2	+1.056733'+3	-2.989816'+3
				+1.381989'+2	-1.162841'+3	+5.642671'+3
					+4.652741'+2	-5.067519'+3
						+1.736699'+3

Table A.XIV

Behaviour of the Solution Vector and the Approximate Solution

Problem S2 $\quad \phi_i(x) = \sqrt{\frac{2}{i\pi}} \sin i\pi x$

n	2	3	7	8	11	12
$\bar{e}_n(x)$	6.527401'-3	5.923725'-3	1.535866'-3	1.139778'-3	7.211637'-4	4.667527'-4
	-1.883149'-1	-1.895375'-1	-1.931626'-1	-1.933572'-1	-1.935239'-1	-1.935703'-1
	+9.572040'-2	+9.513211'-2	+8.831941'-2	+8.795130'-2	+8.763456'-2	+8.754694'-2
		+8.788853'-3	-1.533694'-3	-2.085970'-3	-2.559677'-3	-2.691058'-3
			+2.987747'-2	+2.914348'-2	+2.851251'-2	+2.833767'-2
			+4.360727'-3	+3.441324'-3	+2.652891'-3	+2.434248'-3
			+1.697441'-2	+1.587457'-2	+1.492863'-2	+1.466572'-2
$a_n^n(i)$			+5.427565'-3	+4.137191'-3	+3.033392'-3	+2.727323'-3
				+1.101873'-2	+9.755838'-3	+9.420480'-3
					+3.026842'-3	+2.632835'-3
					+7.336101'-3	+6.757706'-3
					+2.972543'-3	+2.497274'-3
						+6.175361'-3

Table A.XV

Behaviour of the Solution Vector and the Approximate Solution

Problem S2 $\quad \phi_i(x) = x^i(1-x)$

n	2	3	7	8	11	12
$\bar{e}_n(x)$	5.803226'-3	1.720130'-3	3.876723'-4	3.894865'-4	3.223009'-4	2.186298'-4
$a_n^n(i)$	-9.375005'-2	-5.000058'-2	-5.526549'-4	-4.014935'-4	-2.169546'-2	-8.558641'-3
	-4.687501'-1	-7.749965'-1	-1.843196'+0	-1.848973'+0	-1.129760'+0	-1.573131'+0
		+3.499961'-1	+7.165195'+0	+7.230227'+0	+8.122797'+1	+4.730925'+0
			-2.048207'+1	-2.081726'+1	+2.663888'+0	-1.153619'+1
			+3.233145'+1	+3.323843'+1	-3.215400'+0	+1.992294'+1
			-2.568568'+1	-2.701338'+1	-1.720487'+1	-3.217942'+1
			+8.026410'+0	+9.018657'+0	+3.943324'+1	+4.319787'+1
				-2.961196'-1	-1.035058'+1	-2.166500'+1
					-4.195680'+1	-1.888598'+1
					+4.288789'+1	+1.369080'+1
					-1.241240'+1	+1.431690'+1
						-1.052300'+1

Table A.XVI

Behaviour of the Solution Vector and the Approximate Solution

Problem S2 $\quad \phi_i(x) = x(1-x)T^*_{i-1}(x)$

n	2	3	7	8	11	12
$\bar{e}_n(x)$	5.803446'-3	1.720237'-3	9.139300'-5	6.333386'-5	1.615658'-5	1.176447'-5
	-3.281248'-1	-3.062503'-1	-3.070344'-1	-3.072949'-1	-3.069368'-1	-3.069925'-1
	-2.343750'-1	-2.125003'-1	-2.261004'-1	-2.271205'-1	-2.270883'-1	-2.273091'-1
		+4.374912'-2	+4.475112'-2	+4.427361'-2	+4.500172'-2	+4.489445'-2
			-1.419412'-2	-1.521050'-2	-1.523403'-2	-1.545480'-2
			+6.426192'-3	+6.071455'-3	+6.827033'-3	+6.732291'-3
			-2.274219'-3	-3.249667'-3	-3.371217'-3	-3.590253'-3
$a^n_n(i)$			+1.384596'-3	+1.207243'-3	+1.975250'-3	+1.900259'-3
				-7.980900'-4	-1.028812'-3	-1.241278'-3
					+6.903937'-4	+6.403994'-4
					-2.819008'-4	-4.763621'-4
					+2.054354'-4	+1.824376'-4
						-1.494704'-4

Table A.XVII

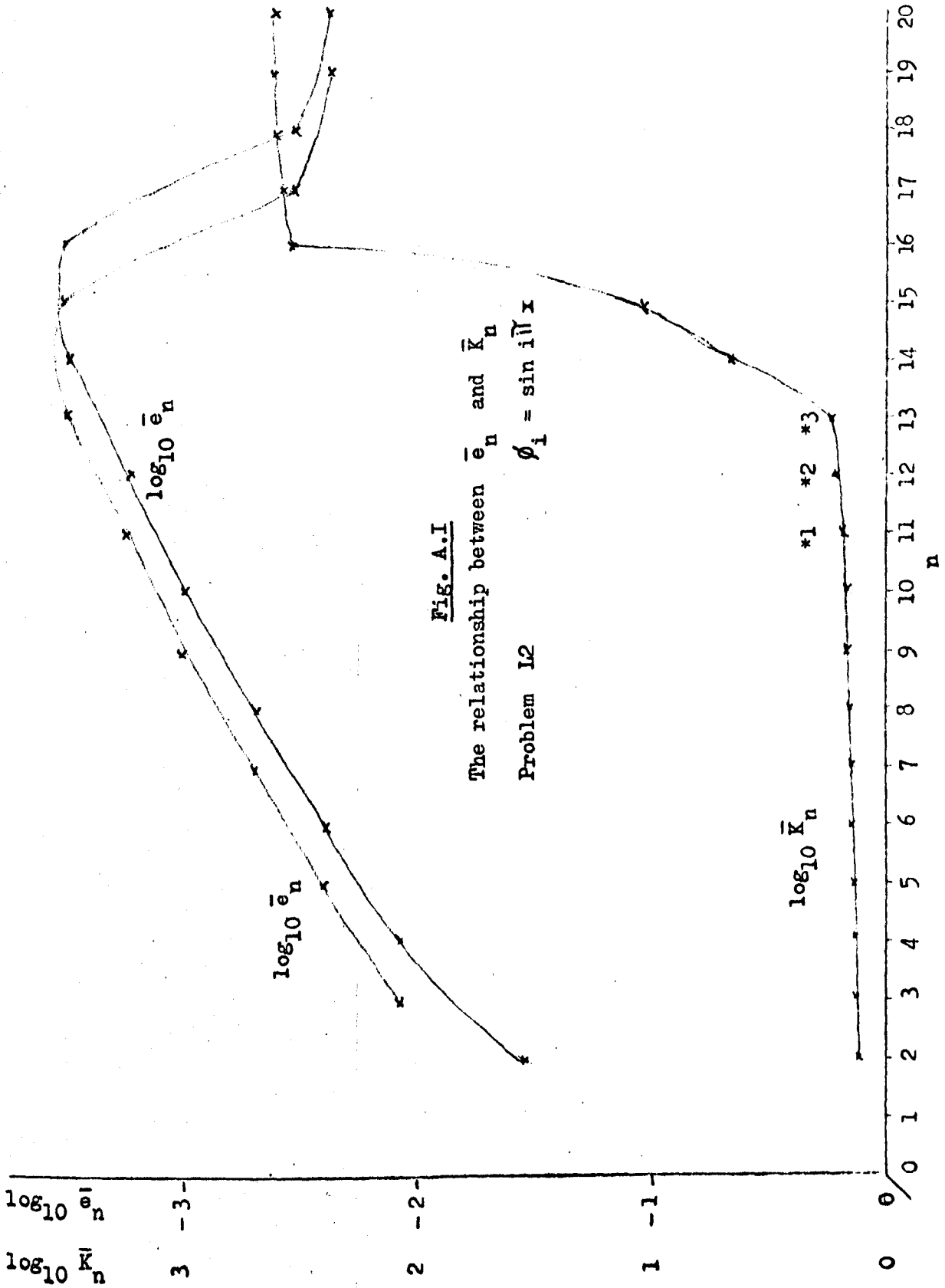


Fig. A.I

The relationship between $\bar{e}_n$ and $\bar{K}_n$

Problem L2 $\phi_i = \sin i\frac{\pi}{2} x$

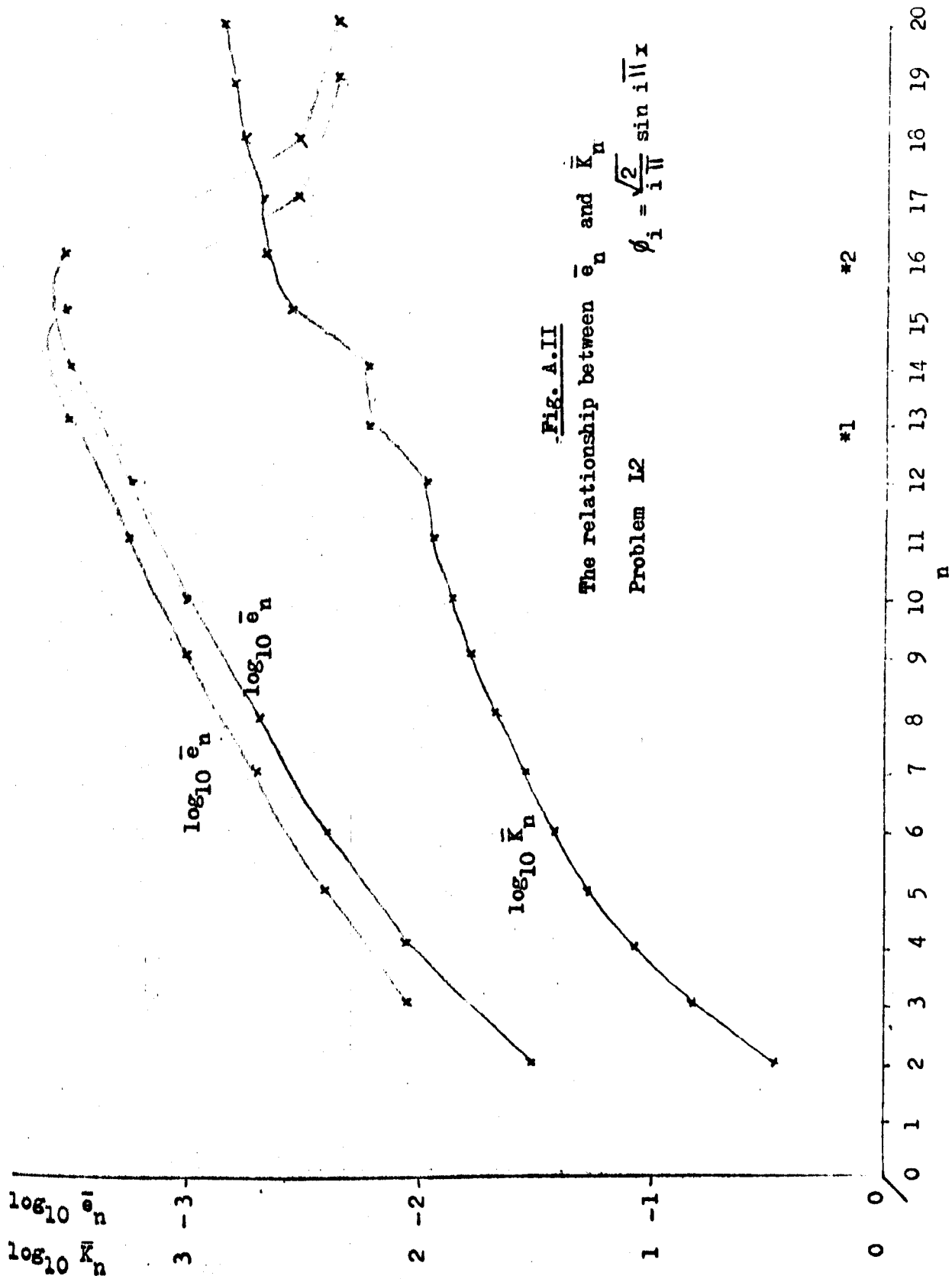


Fig. A.II

The relationship between $\bar{e}_n$ and $\bar{K}_n$

Problem L2 $\quad \phi_i = \frac{\sqrt{2}}{i\,\Pi} \sin i\Pi x$

Notes on Figures A.I and A.II

The numbered asterisks shown in these figures have the following significance.

Fig.A.I

*1 An incorrect large eigenvalue is determined by the numerical procedure for $n = 11$.

*2 An incorrect small eigenvalue is determined by the numerical procedure for $n = 12$.

*3 Incorrect large and small eigenvalues are determined alternately as n increases.

Fig.A.II

*1 An incorrect large eigenvalue is determined by the numerical procedure for $n = 13$.

*2 An incorrect small eigenvalue is determined by the numerical procedure for $n = 16$.

Appendix B

Additional Numerical Experiments concerning the Stability of the Rayleigh-Ritz Method for Mildly non-linear Differential Equations

Behaviour of the Solution Vector and the Approximate Solution

Problem N2 $\phi_i(x) = x^i(1-x)$

n	ESTIMATE	2	3	4	5	6	10
e_n		4.234559'-4	3.120355'-6	3.119467'-6	3.422730'-8	3.422903'-8	8.509098'-9
	1.0	-4.565084'-1	-4.635296'-1	-4.635296'-1	-4.636311'-1	-4.636311'-1	-4.636325'-1
	0	0	+3.491020'-2	+3.491020'-2	+3.632946'-2	+3.632946'-2	+3.636746'-2
	0		-3.491020'-2	-3.491020'-2	-4.058386'-2	-4.058386'-2	-4.090519'-2
	0			+5.418444'-14	+8.508805'-3	+8.508805'-3	+9.720019'-3
$a_n(i)$	0				-4.254402'-3	-4.254402'-3	-6.563773'-3
	0					-1.451856'-12	+2.303060'-3
	0						-1.198267'-3
	0						+3.685464'-4
	0						-9.164261'-5
	0						-2.086676'-7
	ITERATIONS	8	5	1	4	1	5 20* 20* 20*

Long Precision $\delta_n = n.5._{10}-(5+n)$

Table B I

Behaviour of the Solution Vector and the Approximate Solution

Problem N2 $\quad \phi_i(x) = x(1-x)T^*_{i-1}(x)$

n	ESTIMATE	2	3	4	5	6	10
e_n		4.234559'-4	3.120360'-6	3.119472'-6	3.424599'-8	3.422530'-8	8.509098'-9
	1.0	-4.565084'-1	-4.591658'-1	-4.591658'-1	-4.591897'-1	-4.591897'-1	-4.591899'-1
	0	+6.886373'-11	+6.887238'-11	+1.313644'-10	+1.313730'-10	+1.909028'-10	+3.042043'-10
	0		-4.363775'-3	-4.363775'-3	-4.408232'-3	-4.408232'-3	-4.408692'-3
	0			+4.860495'-11	+4.831744'-11	+1.018942'-10	+2.098215'-10
$a_n(i)$	0				-3.323755'-5	-3.323755'-5	-3.364178'-5
	0					+3.928965'-11	+1.355382'-10
	0						-2.922086'-7
	0						+7.552395'-11
	0						-2.846474'-9
	0						+2.965556'-11
	ITERATIONS	8	5	1	3	1	3 2 3 2

Long Precision $\quad \delta_n = n.5_{10}-5+n$

Table B II

Behaviour of the Solution Vector and the Approximate Solution

Problem N2

$$\phi_i(x) = \int_0^x P_i^*(t)dt$$

n	ESTIMATE	2	3	4	5	6	10
$e_n(x)$		4.234435'-4	3.120406'-6	3.119517'-6	3.418465'-8	3.416395'-8	8.509093'-9
e_{csv}		4.23' - 4		3.12' - 6		5.03' - 8	
	1.0	!2.635652'-1	+2.635878'-1	+2.635878'-1	+2.635879'-1	+2.635879'-1	+2.635879'-1
	0	+1.165856'-17	+1.155014'-17	+1.247171'-17	+1.334246'-17	+1.414545'-17	+1.241750'-17
	0		+2.638963'-3	+2.638963'-3	+2.639048'-3	+2.639048'-3	+2.639048'-3
	0			+4.682398'-18	+3.001884'-18	+3.100140'-18	+2.080312'-18
$a_n(i)$	0				+3.054156'-5	+3.054155'-5	+3.054202'-5
	0					-4.624799'-18	-4.397795'-18
	0						+3.550269'-7
	0						-2.771491'-18
	0						+4.125467'-9
	0						+3.039154'-18
	ITERATIONS	7	5	1	3	1	3 1 3 1

Long Precision

$$\delta_n = n.5_{10}-(5+n)$$

Table B III

Behaviour of the Solution Vector and the Approximate Solution

Problem N3 $\phi_i(x) = x^i(1-x)$

n	ESTIMATE	2	3	4	5	6	10
$y_n(0.25)$		-1.486763'-1	-1.498608'-1	-1.511273'-1	-1.509089'-1	-1.509232'-1	-1.509307'-1
$y_n(0.5)$		-2.399975'-1	-2.361573'-1	-2.362442'-1	-2.364812'-1	-2.364806'-1	-2.364674'-1
$y_n(0.75)$		-2.113200'-1	-2.122418'-1	-2.110550'-1	-2.108470'-1	-2.108340'-1	-2.108453'-1
	1.0	-6.258907'-1	-6.958431'-1	-6.509554'-1	-6.587685'-1	-6.561671'-1	-6.567726'-1
	0	-6.681989'-1	-3.297470'-1	-7.294529'-1	-6.206072'-1	-6.725118'-1	-6.568735'-1
	0		-3.356501'-1	+5.909725'-1	+1.572041'-1	+4.679204'-1	+3.452166'-1
	0			-6.163057'-1	+3.290605'-2	-7.426467'-1	-3.323586'-1
	0				-3.242768'-1	+5.279804'-1	-8.611922'-2
$a_n(i)$	0					-3.407206'-1	-8.728849'-2
	0						+4.098094'-1
	0						-6.713837'-1
	0						+4.659514'-1
	0						-1.472348'-1
	ITERATIONS	20*	17	18	15	14	20* 20* 20* 20*

Long Precision $\delta_n = n.5_{10}-(5+n)$

Table B IV

Behaviour of the Solution Vector and the Approximate Solution

Problem N3

$\phi_i(x) = x(1-x)T^*_{i-1}(x)$

n	ESTIMATE	2	3	4	5	6	10
$y_n(0.25)$		-1.486764'-1	-1.498608'-1	-1.511273'-1	-1.509089'-1	-1.509232'-1	-1.509307'-1
$y_n(0.5)$		-2.399978'-1	-2.361573'-1	-2.362442'-1	-2.364812'-1	-2.364806'-1	-2.364674'-1
$y_n(0.75)$		-2.113202'-1	-2.122418'-1	-2.110550'-1	-2.108470'-1	-2.108340'-1	-2.108453'-1
	1.0	-9.599912'-1	-9.865855'-1	-9.866627'-1	-9.885069'-1	-9.885095'-1	-9.886377'-1
	0	-3.340999'-1	-3.326986'-1	-3.581335'-1	-3.581479'-1	-3.591685'-1	-3.592139'-1
	0		-4.195627'-1	-4.168575'-2	-4.511515'-2	-4.511686'-2	-4.536418'-2
	0			-1.925955'-2	-1.923898'-2	-2.015507'-2	-2.019746'-2
	0				-2.533412'-3	-2.529851'-3	-2.747575'-3
$a_n(i)$	0					-6.654699'-4	-7.015560'-4
	0						-1.575402'-4
	0						-2.630752'-5
	0						-5.999963'-6
	0						-1.123338'-6
	ITERATIONS	20*	15	16	14	11	12 10 10 8

Long Precision

$\delta_n = n.5_{10}-(5+n)$

Table B V

Behaviour of the Solution Vector and the Approximate Solution

Problem N3

$$\phi_i(x) = \int_0^x P_i^*(t)dt$$

n	ESTIMATE	2	3	4	5	6	8 +
$y_n(0.25)$		-1.486766'-1	-1.498608'-1	-1.511273'-1	-1.509089'-1	-1.509232'-1	-1.509305'-1
$y_n(0.5)$		-2.399980'-1	-2.361573'-1	-2.362442'-1	-2.364812'-1	-2.364806'-1	-2.364668'-1
$y_n(0.75)$		-2.113204'-1	-2.122418'-1	-2.110550'-1	-2.108470'-1	-2.108340'-1	-2.108453'-1
	1.0	+5.542517'-1	+5.550713'-1	+5.552096'-1	+5.552117'-1	+5.552124'-1	+5.552125'-1
	0	+1.494140'-1	+1.487873'-1	+1.490881'-1	+1.491064'-1	+1.491069'-1	+1.491071'-1
	0		+2.537277'-2	+2.520917'-2	+2.524032'-2	+2.524423'-2	+2.524436'-2
	0			+1.467394'-2	+1.465827'-2	+1.466485'-2	+1.466557'-2
$a_n(i)$	0				+2.327931'-2	+2.324659'-3	+2.325976'-3
	0					+7.159005'-4	+7.154627'-4
	0						+1.837053'-4
	0						+3.434615'-5
	ITERATIONS	19	15	17	14	12	8 11

Long Precision

$\delta_n = n.5_{10}-(5+n)$

Table B VI

Behaviour of the Solution Vector and the Approximate Solution

Problem Q2 : k = 0.5, i = 3 $\quad \phi_i(x) = x(1-x)T^*_{i-1}(x)$

n	ESTIMATE	2	3	4	5	6	10
$e_n(x)$		1.261580'-2	1.208426'-3	7.948973'-5	4.423306'-5	9.148378'-6	2.617497'-9
	1.0	+2.772160'-1	+3.453961'-1	+3.474917'-1	+3.471648'-1	+3.470793'-1	+3.470030'-1
	0	-4.806176'-1	-5.190254'-1	-5.338607'-1	-5.336877'-1	-5.330513'-1	-5.329933'-1
	0		+1.081963'-1	+1.117820'-1	+1.111715'-1	+1.110113'-1	+1.108642'-1
	0			-1.114156'-2	-1.100610'-2	-1.043511'-2	-1.038202'-2
	0				-4.434962'-4	-5.634499'-4	-6.924685'-4
$a_n(i)$	0					+4.063258'-4	+4.476025'-4
	0						-9.081614'-5
	0						+1.133954'-5
	0						-7.539886'-7
	0						-4.670775'-8
	0						
	ITERATIONS	8	6	7	6	6	5 5 4 4

Long Precision $\quad \delta_n = n.5._{10}-(5+n)$

Table B VII

Behaviour of the Solution Vector and the Approximate Solution

Problem Q2 : k = 0.99, i = 1 $\quad \phi_i(x) = x(1-x)T^*_{i-1}(x)$

n	ESTIMATE	2	3	4	5	6	10
$e_n(x)$		1.265498'-2	1.246804'-3	8.058865'-5	4.459027'-5	9.178843'-6	2.628862'-9
	1.0	+2.777448'-1	+3.454895'-1	+3.474868'-1	+3.471635'-1	+3.470791'-1	+3.470030'-1
	0	-4.794123'-1	-5.191336'-1	-5.338548'-1	-5.336845'-1	-5.330511'-1	-5.329333'-1
	0		+1.081440'-1	+1.117780'-1	+1.111701'-1	+1.110109'-1	+1.108642'-1
	0			-1.114110'-2	-1.100470'-2	-1.043498'-2	-1.038202'-2
	0				-4.430951'-4	-5.63638 '-4	-6.924684'-4
	0					+4.062553'-4	+4.476025'-4
$a_n(i)$	0						-9.081610'-5
	0						+1.133953'-5
	0						-7.539695'-7
	0						-4.670200'-8
	0						
	ITERATIONS	11	7	9	8	8	7 7 6 4

Long Precision $\quad \delta_n = n.5_{10}-(5+n)$

Table B VIII

Behaviour of the Solution Vector and the Approximate Solution

Problem Q2 : k = 0.01, i = 1 $\quad \phi_i(x) = x(1-x)T^*_{i-1}(x)$

n	ESTIMATE	2	3	4	5	6	10
$e_n(x)$		1.256998'-2	1.178835'-3	7.855611'-5	4.390251'-5	9.119162'-6	2.608187'-9
	1.0	+2.767768'-1	+3.453349'-1	+3.474959'-1	+3.471659'-1	+3.470795'-1	+3.470030'-1
	0	-4.818031'-1	-5.189360'-1	-5.338660'-1	-5.336907'-1	-5.330516'-1	-5.329933'-1
	0		+1.082475'-1	+1.117859'-1	+1.111727'-1	+1.110117'-1	+1.108642'-1
	0			-1.114202'-2	-1.100749'-2	-1.043525'-2	-1.038202'-2
$a_n(i)$	0				-4.438890'-4	-5.632636'-4	-6.924685'-4
	0					+4.063955'-4	+4.476026'-4
	0						-9.0816178'-5
	0						+1.133955'-5
	0						-7.540076'-7
	0						-4.671388'-8
	ITERATIONS	3	3	3	3	3	3 3 3 3

Long Precision $\quad \delta_n = n.5_{10}-(5+n)$

Table B IX

Appendix C

On the Order of the Error of a New Finite Difference Formula.

In this Appendix we consider finite difference approximations of the solution of the ordinary differential equation

$$y'' = f(x,y) \quad , \quad x \in [0,1] \qquad \ldots(C\ 1.1)$$

with the boundary conditions

$$y(0) = y(1) = 0$$

and in particular present a formal analysis of the order of the error of the new finite difference approximation (5.18) proposed in § 5.1, together with numerical experiments in its use.

We assume that the equation (C 1.1) is such that a variational formulation exists (§ 2.6), and denoting its solution by $y^o(x)$ we have

$$I(y^o) \leq I(y) = \int_0^1 \left\{ y'^2 + 2 \int_0^y f(x,\eta)\,d\eta \right\} dx$$

Applying the Rayleigh-Ritz method to the determination of an approximate solution of the variational problem of the form

$$\bar{y}(x) = \sum_{i=1}^{n} \bar{y}_i \; w_i(x)$$

where $w_i(x)$ are the basis for the subspace of piecewise Hermite polynomials of degree one on a partition

$$\Pi : \quad 0 = x_o < x_1 \ldots < x_{n+1} = 1$$

$$x_i = i \times h \quad , \quad h = 1/n+1$$

leads to the equations (Herbold (1))

$$\frac{\bar{y}_{i+1} - 2\bar{y}_i = \bar{y}_{i-1}}{h} = \int_{x_{i-1}}^{x_{i+1}} f(x,\bar{y}(x))w_i(x)dx \qquad \ldots(C\ 1.2)$$

$$i = 1, \ldots, n \ .$$

Herbold shows that if the integral of (C 1.2) is expressed as

$$\int_{x_{i-1}}^{x_{i+1}} f(x,\bar{y}(x))\, w_i(x)dx = \int_{x_{i-1}}^{x_i} f(x,\bar{y}(x))w_i(x)dx + \int_{x_i}^{x_{i+1}} f(x,\bar{y}(x))w_i(x)dx$$

and each term approximated by the trapezium rule, then the familiar second order finite difference formula

$$\frac{\tilde{y}_{i+1} - 2\tilde{y}_i + \tilde{y}_{i-1}}{h^2} = f(x_i,\tilde{y}_i) \qquad \ldots(C\ 1.3)$$

is deduced, and similarly, that if the integrand $f(x,\bar{y}(x)).w_i(x)$ is approximated by $\bar{f}(x,\bar{y}(x)).w_i(x)$, where $\bar{f}(x,y)$ is the quadratic polynomial satisfying

$$\bar{f}(x_j,\tilde{y}_j) = f(x_j,\bar{y}_j) \quad , \quad j = i-1,\ i,\ i+1$$

and the integration performed, the fourth order **Mehrstellenverfarhen** scheme (Collatz (1)) results; i.e.

$$\frac{y^+_{i+1} - 2y^+_i + y^+_{i-1}}{h^2} = \frac{1}{12}\left(f(x_{i-1},y^+_{i-1}) + 10f(x_i,y^+_i) + f(x_{i+1},y^+_{i+1})\right) \qquad \ldots(C\ 1.4)$$

The new scheme may be derived in at least two ways. Trivially, we may choose to approximate the integral of (C 1.2) by Simpson's Rule (which is consistent in the sense of Herbold (1)) to get

$$\frac{\hat{y}_{i+1} - 2\hat{y}_i + \hat{y}_{i-1}}{h^2} = \frac{1}{6}\left(f(x_{i-1},\hat{y}_{i-1}) + 4f(x_i,\hat{y}_i) + f(x_{i+1},\hat{y}_{i+1})\right) \qquad \ldots(C\ 1.5)$$

Less arbitrarily, we consider the case in which

$$f(x,y) = ky + g(x,y)$$

where k is a constant, including linear equations of this type automatically. Integrating the right hand side of (C 1.2) exactly yields

$$\int_{x_{i-1}}^{x_{i+1}} f(x,y)\cdot w_i(x)dx = \int_{x_{i-1}}^{x_{i+1}} (ky + g(x,y))\cdot w_i(x)\ dx$$

$$= \frac{hk}{6}(y_{i-1} + 4y_i + y_{i+1}) + \int_{x_{i-1}}^{x_{i+1}} g(x,y)dx$$

The motivation for Simpson's rule approximation of the remaining integral term is clearly apparent, leading again to (C 1.5).

As indicated in § 5.1, it is clear that this new finite difference formula has an error of order h^2 with leading coefficient opposite in sign from that of (C 1.3) from the relation between (C 1.3), (C 1.4), (C 1.5) and the known orders of accuracy of the established methods. We present here a verification of this in terms of a formal error analysis for linear equations of the form

$$y'' + k(x)y + f(x) = 0$$

and write (C 1.3) and (C 1.5) in the forms

$$\frac{\tilde{y}_{i-1} - 2\tilde{y}_i + \tilde{y}_{i+1}}{h^2} + k_i\tilde{y}_i + f_i = 0$$

and

$$\frac{\hat{y}_{i-1} - 2\hat{y}_i + \hat{y}_{i+1}}{h^2} + \frac{1}{6}(k_{i-1}\hat{y}_{i-1} + 4k_i\hat{y}_i + k_{i+1}\hat{y}_{i+1}) = 0$$

To consider the errors of these approximations, we write

$$\tilde{\varepsilon}_i = y^0(x_i) - \tilde{y}_i$$

$$\hat{\varepsilon}_i = y^0(x_i) - \hat{y}_i$$

Then

$$k_i \tilde{\xi}_i = -\left(y^{o''}\Big|_{x_i} - \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} \right)$$

and, as is well known, this results in

$$k_i \tilde{\xi}_i = -2 \sum_{j=1}^{\infty} \frac{h^{2j}\, y^{o(2j+2)}}{(2j+2)!}\Bigg|_{x=x_i}$$

$$= -2S(y^o) \qquad \ldots(C\ 1.6)$$

which we take to be the definition of $S(y^o)$.

Similarly, for $\hat{\xi}_i$, we can write

$$-2\, S(y^o) + k_i \hat{\xi}_i - \frac{1}{6} \left(k_{i-1} \hat{y}_{i-1} - 2k_i \hat{y}_i + k_{i+1} \hat{y}_{i+1}\right)$$

$$+ \left(f_{i-1} - 2f_i + f_{i+1}\right) = 0$$

Assuming $y^{o(iv)}$ to be continuous, we can write

$$y^{o(iv)}(x) = -\frac{d^2}{dx^2} (ky + f)$$

$$= \frac{\left(k_{i-1} \hat{y}_{i-1} - 2k_i \hat{y}_i + k_{i+1} \hat{y}_{i+1}\right)}{h^2}$$

$$+ \frac{\left(f_{i-1} - 2f_i + f_{i+1}\right)}{h^2} + S(ky^o + f)$$

so that

$$-2S(y^o) + k_i \hat{\xi}_i - \frac{h^2}{6} \left(y^{o(iv)} - 2S(ky^o + f)\right) = 0$$

Expanding $-2S(y)$ from its definition, we have

$$k_i \hat{\xi}_i = \frac{h^2}{12}\, y^{o(iv)} + O(h^4)$$

and from (C 1.6)

$$k_i \xi_i = -\frac{h^2}{12} y^{o(iv)} + O(h^4)$$

Thus the conclusion that the finite difference algorithms (C 1.3), (C 1.5) have local truncation errors of order h^2 with coefficients equal in magnitude and opposite in sign is verified.

We report briefly numerical results from the application of the methods considered here to the linear ordinary differential equation

$$y'' - \frac{2y}{(x+2)^2} + \frac{1}{x+2} = 0$$

with the boundary conditions

$$y(0) = y(1) = 0$$

which has the solution

$$y(x) = -(19(x+2) - 5(x+2)^2 - \frac{36}{x+2}) / 38$$

This example is adapted from Collatz (1, p.178). The behaviour of the error as a function of h is indicated, illustrating the $O(h^2)$ error behaviour of the new method. In Table C I results for the finite different approximations (C 1.3), (C 1.4), (C 1.5) for a number of values of h are given. In Table C II the results of an application of the Rayleigh-Ritz method in the form of Chapter Three with the modified Chebyshev basis functions

$$\phi_i(x) = x(1-x) T^*_{i-1}(x)$$

are given for comparison.

The new method cannot be recommended as an alternative to the established central difference formula (C 1.3), since it requires three times as many function evaluations. We suggest, however, that the joint application of methods (C 1.3), (C 1.5) can be considered as a useful alternative to the direct application of (C 1.4). No

A Comparison of the Maximum Error of Three Finite Difference Formulae as a Function of h .

	$h = 0.1$	$h = 0.05$	$h = 0.025$
Method (C 1.3)	+2.661'-5	+6.669'-6	+1.669'-6
Method (C 1.5)	+1.362'-6	+3.542'-7	+8.967'-8
Method (C 1.4)	-2.677'-5	-6.679'-6	-1.669'-6

Table C I

The Behaviour of the Error of a Polynomial Rayleigh-Ritz Expansion.

n	2	6	10
e_{max}	1.274'-4	1.247'-8	1.869'-11

Table C II

additional function evaluations are required, and in the general case $y'' = f(x,y)$, (C 1.3) and (C 1.5) generate non-linear algebraic equations of the form

$$A\underline{y} = \underline{g}(y)$$

with the same matrix A , so that the same iteration matrix A may be used to solve these equations iteratively. As a result of

applying the two $O(h^2)$ methods jointly one readily obtains the solutions $\tilde{y}$, $\hat{y}$ and $y^+ = (\tilde{y} + \hat{y})/2$, for which the relations $y^o \in [\tilde{y} , \hat{y}]$ and $y^o \doteq y^+$ hold for sufficiently small h (neglecting rounding error effects). Thus readily computable expressions of the nature of asymptotic error bounds are provided by the joint application of these formulae.