THE DISCRETE LEAST SQUARES METHOD FOR 2m-th ORDER ELLIPTIC BOUNDARY-VALUE PROBLEMS

by

PETER HENRY SAMMON

B.Sc., University of British Columbia, 1973

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department

of

MATHEMATICS

and the

INSTITUTE OF APPLIED MATHEMATICS AND STATISTICS

We accept this thesis as conforming to

the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

December, 1974

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Mathematics

The University of British Columbia
Vancouver 8, Canada

Date December 6, 1974

Abstract

Many positive results are known for the Least Squares method of numerically computing an approximant solution for a 2m-th order elliptic boundary value problem on a real interval. However due to integrals that appear in the linear system that must be solved to find the approximant solution, the Least Squares method is computationally unattractive. Thus one is led to consider other, more practical, methods.

In this thesis, such a method is examined. It is first shown that minimizing a quantity involving discrete quadrature sums to find an approximation to the solution leads to a linear system in which quadrature sums, not integrals, appear. The approximation generated by this method, known as the Discrete Least Squares method, is thus easily computed.

Error estimates are then proven for the Discrete method. A certain polynomial interpolate is first constructed and then using Sobolev space theory and known estimates from the related Least Squares method, the desired estimates are obtained. These estimates are like the usual Least Squares estimates although more continuity of the coefficients of the problem is required in the final stages.

Comparisons are then drawn between the Discrete method and the Collocation method. It is noted that both methods offer easy computability and shown why the Discrete method can generate a smoother approximation in most situations.

Results of numerical computations are then provided which support the theory and a discussion of what problems the theory applies to is also given.

Table of Contents

Table of Contents (Contd.)

List of Tables

Acknowledgement

I wish to express my deep appreciation to my advisor, Dr. James M. Varah, for his continuing encouragement and interest in my studies and also for his assistance with the problem discussed in this thesis.

I would also like to thank Dr. A.T. Bui for his help and suggestions, Mrs. Y.S. Choo for her excellent typing and my family and friends, especially my parents Kathleen and Andrew Sammon and Christine Hellwig, for their interest and patience throughout the writing of this work.

The assistance of the University of British Columbia and the National Research Council is also gratefully acknowledged.

Chapter 0

Some Motivation

Many results are known for the Least Squares Method of numerically computing an approximate solution for a 2m-th order elliptic boundary value problem. However for reasons to be noted later, this method is computationally unattractive, if not impossible, in actual situations. In this thesis we will examine a more practical method, commonly known as the Discrete Least Squares Method. (To avoid misunderstandings, we will always use the term Continuous Least Squares for the usual approach).

To show how the Discrete method differs from the usual Continuous method and how the Discrete method compares with the well known Collocation Method we will consider these three methods as applied to the following differential equation problem. Consider

$$(0.1.1) \qquad Lu(x) = \sum_{r=0}^{m} (-1)^r D^r (a_r'(x) D^r u(x)) = f(x) \; ; \quad x \in [a, b] \; ,$$

Boundary Conditions : $D^r u(a) = D^r u(b) = 0$ for $0 \leq r \leq m-1$,

where D^r denotes differentiation. (The primes on the coefficients $\{a_j'\}$ are used for later notational reasons). We assume all the coefficient functions are suitably smooth and also that $a_j'(x) > 0$ for $x \in [a, b]$, $1 \leq j \leq m$. Thus the operator L is uniformly strongly elliptic on $[a, b]$.

Let S_o be a finite dimensional subspace of $L^2(a, b)$ that consists of suitably smooth functions which satisfy the boundary conditions

stated in (0.1.1). We will specify a particular choice later. Given this subspace each of the three methods finds an approximation in S_o, denoted by w, to the solution u of equation (0.1.1). If $\{e_i\}_{i=1}^q$ is a basis for S_o, then equivalently, each of the three methods finds a vector $\underline{d} = (d_1, \ldots, d_q)$ so that the corresponding function $w = \sum_{i=1}^{q} d_i e_i$ in S_o is "close" to u in some norm. We choose to use the usual L^2-norm on (a, b) (denoted by $||\cdot||_0$). To be practical, the method must be able to define $\underline{d}$ in some easily computed way.

The Continuous method determines w as the unique function in S_o that minimizes $||Lv - f||_0$ over all functions $v \in S_o$. This process would appear to give a reasonable approximation to u. In fact it will be shown that $\underline{d}$ is easily found as the solution to the linear system

$$(0.1.2) \qquad \sum_{r=1}^{q} \left\{ \int_a^b Le_r \cdot Le_s \right\} d_r = \left\{ \int_a^b f \cdot Le_s \right\} \quad \text{for } 1 \leq s \leq q .$$

But as mentioned before, actual computations are difficult or impossible because of the integrals in this system.

To avoid integrals we need some form of integral approximation. Given a partition $a = x_0 < x_1 < x_2 < \ldots < x_N < x_{N+1} = b$ of $[a, b]$ and a function $g \in C[a, b]$ we can approximate $\int_a^b g(x)\,dx$ by $\sum_{i=0}^{N} \sum_{j=1}^{K} w_{ij} g(z_{ij})$ where K is a positive integer, the $\{w_{ij}\}$ are certain positive numbers called <u>weights</u> and the $\{z_{ij}\}_{j=1}^K$ are certain points in (x_i, x_{i+1}), for $0 \leq i \leq N$, called the <u>Gauss</u> points. We will discuss this approximation in

more detail in Section 2.6.

Given these integral approximations we can now describe the Discrete method. The Discrete method determines w as the unique function in S_o that minimizes

$$\sum_{i=0}^{N} \sum_{j=1}^{K} w_{ij} \left\{ (Lv - f)(z_{ij}) \right\}^2$$

over all functions $v \in S_o$. Thus the Discrete method is equivalent to the Continuous method with approximate integration replacing integration. We will also show later that $\underline{d}$ is easily found as the solution to

$$(0.1.3) \qquad \sum_{r=1}^{q} \left\{ \sum_{i=0}^{N} \sum_{j=1}^{K} w_{ij} (Le_r \cdot Le_s)(z_{ij}) \right\} d_r = \left\{ \sum_{i=0}^{N} \sum_{j=1}^{K} w_{ij} (f \cdot Le_s)(z_{ij}) \right\}$$

$$\text{for} \quad 1 \leq s \leq q ,$$

where we note that (0.1.3) is the same as (0.1.2) with the appropriate integral approximations. Thus calculating $\underline{d}$ for the Discrete method from (0.1.3) becomes feasible as each term of the system only involves a finite sum of function values.

The motivation for the Collocation method is different from the norm minimizing concepts used above in the Continuous method. For the Collocation method in its most general form we pick q distinct points $\{y_i\}_{i=1}^{q}$ in $[a, b]$ and require that w satisfy

$$(0.1.4) \qquad Lw(y_i) = \sum_{r=1}^{q} Le_r(y_i)\, d_r = f(y_i) \qquad \text{for} \quad 1 \leq i \leq q ,$$

which gives an easily computed system for $\underline{d}$. However if good approximation

qualities are to be obtained, the choice of the points must be made carefully.

To further discuss the methods we must specialize our choice for the approximation space S_o. We choose to use a space of piecewise polynomials of odd degree which is known to have good approximation qualities. More precisely given two integers n, z where $n - 1 \leq z \leq 2n - 2$ and a partition of $[a, b]$ as before, we let S_o be the collection of all functions $e \in C^z[a, b]$ that are polynomials of degree $2n - 1$ on (x_i, x_{i+1}) for $0 \leq i \leq N$ and that satisfy the boundary conditions. In this case $q = (2n - 1 - z)N + (2n - 2m)$. We will choose $z \geq 2m - 1$ also so that $Le \in L^2(a, b)$ for $e \in S_o$.

If we order the points $\{z_{ij}\}$ and $\{w_{ij}\}$ of the Discrete method lexicographically and call them $\{s_j\}$ and $\{t_j\}$ respectively then (0.1.4) becomes

$$S^T D S \underline{d} = S^T D \underline{f} \tag{0.1.5}$$

where $S = [Le_r(s_j)]_{j=1}^{K(N+1)}{}_{r=1}^{q}$ is a $K(N+1) \times q$ matrix, $D = \mathrm{diag}(t_1, \ldots, t_{K(N+1)})$ and $\underline{f} = [f(s_j)]_{j=1}^{K(N+1)}$. Now (0.1.5) is precisely the system, known as the <u>system of normal equations</u>, that must be solved to find the unique vector $\underline{d}$ that minimizes $||D^{\frac{1}{2}}(S\underline{v} - \underline{f})||$ over all $\underline{v} = (v_1, \ldots, v_q)$ in q-dimensional Euclidean space. Here $||\cdot||$ denotes the $K(N+1)$-dimensional Euclidean norm. We note that $K = 2n - 2m$ will be our final choice for K so $n > m$ is required. Thus both w and $\underline{d}$ are

found as solutions to least squares minimization problems under the Discrete least squares method.

If, on the other hand, we have $q = K(N + 1)$ (which is equivalent to $z = 2m - 1$) and pick the points $\{y_i\}$ for the Collocation method as the Gauss points, that is $\{y_i\}_{i=1}^{q} = \{s_i\}_{i=1}^{K(N+1)}$, we get good approximation properties (Russell and Varah [10]). In this case the system (0.1.4) for $\underline{d}$ becomes $S\underline{d} = \underline{f}$.

Thus the system (0.1.5) which determines the Discrete method and also finds the vector $\underline{d}$ that minimizes $||D^{\frac{1}{2}}(S\underline{v} - \underline{f})||$ over all $\underline{v} = (v_1, \ldots, v_q)$ may be viewed as an extension of the Collocation method, which describes the particular case when S is square and $S\underline{d} = \underline{f}$ is directly solvable. Essentially under the Discrete method we are collocating at more points than unknowns, scaling the equations by $D^{\frac{1}{2}}$ and looking for $\underline{d}$ as a least squares solution in $K(N+1)$-dimensional Euclidean space.

We have seen that both the Discrete and Collocation methods are practical for computations. However the Discrete method has the advantage that $2n - 2 \geq z \geq 2m - 1$ is allowed. Since $z = 2m - 1$ is required for Collocation, we can find an approximation w by the Discrete method that has <u>more continuity</u> than that available from the more restrictive Collocation method. This in fact allows a more convenient basis to be chosen when $z = 2n - 2$ that allows the formation and solution of $S^T DS\underline{d} = S^T D\underline{f}$ for the Discrete method to be about as much work as solving $S\underline{d} = \underline{f}$ for the Collocation method, where a more restricted continuity of basis is required.

Thus because of the computational competitiveness of the Discrete method when $z = 2n - 2$ and the higher continuity of the approximation obtained it seems worthwhile to show that the Discrete method has approximation properties similar to those known for the Collocation method. This will be the objective of this thesis.

We begin by stating known results from Sobolev space and piecewise polynomial space theory in Chapter 1, as well as our assumptions on the differential equation problem. In Chapter 2 we define formally the Continuous and Discrete methods. We prove known results for the Continuous method and also examine a certain type of function approximation that will be necessary for further results. We then examine integral approximations and prove the convergence results for the Discrete method. Finally in Chapter 3 we examine our hypotheses from a practical point of view, compare the Collocation method and the Discrete method again and give some numerical results.

Chapter 1

The Preliminaries

1.1 Sobolev Spaces

We wish to find an approximation to the solution u of the linear self adjoint differential equation

$$Lu = \sum_{r,s=0}^{m} (-1)^r D^r (a'_{rs} D^s u) = f ,$$

Boundary Conditions : $D^r u(a) = D^r u(b) = 0$ for $0 \leq r \leq m - 1$, on the interval (a, b), where $a'_{rs} = a'_{sr}$ for $0 \leq r, s \leq m$. Note that we are allowing a slightly more general operator than the operator of equation (0.1.1).

We begin the formal description of our methods by examining some Sobolev space theory (see Agmon [2]).

Definition (1.1.1) : (1) *Let* $C^r(c, d)$ *denote the space of* r *times continuously differentiable functions on* (c, d) .

(2) *Let* $C^\infty(c, d) = \bigcap_{j=0}^{\infty} C^j(c, d)$.

(3) *Let* $\operatorname{supp} g(x) = \left\{ \text{closure of } \{x : g(x) \neq 0\} \right\}$

(4) *Let* $C_0^\infty(c, d) = \{g \in C^\infty(c, d) : \operatorname{supp} g \subset (c, d)\}$.

(5) *Let* $C^{*r}(c, d) = \{g \in C^r(c, d) : \sum_{s=0}^{r} \int_c^d (D^s g)^2 < \infty\}$.

(6) *Let* $||\cdot||_{r,(c,d)} = \left\{ \sum_{s=0}^{r} \int_c^d (D^s(\cdot))^2 \right\}^{1/2}$.

(7) *Let the usual* L^p *spaces be denoted by* $L^p(c, d)$ *for* $p \geq 1$.

Similar definitions hold for $C^r[c, d]$ and $C^\infty[c, d]$. Clearly $||\cdot||_{r,(c,d)}$ is a norm on $C^{*r}(c, d)$ associated with the inner product

$$(\cdot, \cdot)_{r,(c,d)} = \int_c^d \sum_{s=0}^{r} D^s(\cdot)D^s(\cdot) .$$

<u>Definition</u> (1.1.2) : *Let* $H^r(c, d)$ = (*completion of* $C^{*r}(c, d)$ *under* $||\cdot||_{r,(c,d)}$).

Thus $H^r(c, d)$ is a Hilbert space.

We say $g \in L^2(c, d)$ has <u>weak</u> L^2-<u>derivatives up to order</u> r if there is a function $g_s \in L^2(c, d)$ for $0 \leq s \leq r$ so that

$$\int_c^d e\, g_s = (-1)^s \int_c^d g\, D^s e \quad \text{for } e \in C_0^\infty(c, d) .$$

<u>Definition</u> (1.1.3) : *Let* $W^r(c, d) = \{ g \in L^2(c, d) : g$ *has weak* L^2-*derivatives up to order* $r \}$.

We will ignore interval designations if the interval involved is (a, b). Agmon [2; p. 11] shows that $W^r(c, d) = H^r(c, d)$ for $r \geq 0$. We note that $W^0(c, d) = H^0(c, d) = L^2(c, d)$.

We will now state a form of the Sobolev imbedding theorem found in Agmon [2; p. 32] .

<u>Theorem</u> (1.1.4) : *If* $g \in W^r(c, d)$ *for* $r \geq 1$ *then there is a* $g' \in C^{r-1}[c, d]$ *so that* $g = g'$ *a.e.*

We will identify g with g' in the usual way. We also have a stronger result .

<u>Theorem</u> (1.1.5) : *If* $g \in W^r(c, d)$ *for* $r \geq 1$ *then* $D^{r-1}g$ *is absolutely continuous on* $[c, d]$.

<u>Proof</u> : See Berezanskii [4; p. 25] . ///

Thus $g \in W^r(c, d)$ implies that $D^r g \in L^2(c, d)$. Hence :

<u>Theorem</u> (1.1.6) : *We have* $g \in W^r(c, d)$ *for* $r \geq 1$ *if and only if we have* $D^r g \in L^2(c, d)$ *and* $D^{r-1}g$ *is absolutely continuous on* $[c, d]$ *for* $r \geq 1$.

We will also need :

<u>Definition</u> (1.1.7) : *Let* $H_o^r(c, d)$ = (*completion of* $C_o^\infty(c, d)$ *under* $||\cdot||_{r,(c,d)}$) .

Note that $H_o^r(c, d)$ is a closed subspace of $H^r(c, d)$. We also have the (non-optimal) result :

<u>Theorem</u> (1.1.8) : (1) *If* $g \in H_o^r(c, d)$ *for* $r \geq 1$ *then* $g \in H^r(c, d)$ *and* $D^s g(c) = D^s g(d) = 0$ *for* $0 \leq s \leq r-1$.

(2) *If* $g \in C^r[c, d]$ *and* $D^s g(c) = D^s g(d) = 0$ *for* $0 \leq s \leq r-1$ *then* $g \in H_o^r(c, d)$.

Proof : See Berezanskii [4; p. 21] and Agmon [2; p. 131] . ///

The spaces $W^r(c, d)$ are called Sobolev spaces and they will provide the framework for much of the remainder of our analysis. We note that due to Theorem (1.1.6) we have Taylor expansions in Sobolev spaces.

Theorem (1.1.9) : *If* $g \in W^r(c, d)$, $x, y \in [c, d]$ *and* $1 \leq s \leq r$ *then*

$$g(x) = \sum_{j=0}^{s-1} \frac{D^j g(y)}{j!} (x - y)^j + \frac{1}{(s-1)!} \int_y^x D^s g(t)(x - t)^{s-1} \, dt.$$

Proof : See Hewitt and Stromberg [8; p. 287]. ///

We now state a result that comes from the theory of interpolation between Banach spaces. We let $L(B_1, B_2)$ denote the space of linear operators and let $B(B_1, B_2)$ denote the space of bounded linear operators from a normed linear space B_1 to a normed linear space B_2 . (For a proof of the following and some discussion see Adams [1; pp 334-339]). We have :

Theorem (1.1.10) : *If* $T \in B(W^0(c, d), Y) \cap B(W^r(c, d), Y)$ *for some Banach space* Y *and there are constants* M , M_1 *so that*

$$||Tg||_Y \leq M_0 ||g||_{0,(c,d)} \quad \textit{for} \quad g \in W^0(c, d)$$

$$||Tg||_Y \leq M_1 ||g||_{r,(c,d)} \quad \textit{for} \quad g \in W^r(c, d) ,$$

then given $0 < \theta < 1$ *so that* θr *is an integer we have that there is a constant* C *so that* :

(1) $T \in B(W^{\theta r}(c, d), Y)$

(2) $||Tg||_Y \leq CM_0^{1-\theta}M_1^{\theta}||g||_{\theta r,(c,d)}$ *for* $g \in W^{\theta r}(c, d)$.

1.2 Piecewise Polynomial Spaces

We will now take a closer look at the space of functions from which our approximations will come. This section will follow the definitions and results from Schultz's paper [11].

Definition (1.2.1) : (1) *For each* $N \geq 0$ *let* $P_N(a, b)$ *denote the set of all partitions* Δ *of the form*

$$\Delta : a = x_0 < x_1 < x_2 < \ldots < x_N < x_{N+1} = b .$$

Let $P(a, b) = \bigcup_{N=0}^{\infty} P_N(a, b)$.

(2) *For* $\Delta \in P(a, b)$ *and* $n, z \geq 1$ *where* $n-1 \leq z \leq 2n-2$ *we let*

$$S(2n-1, \Delta, z) = \{e(x) \in C^z[a, b] : e \text{ is a polynomial of degree } 2n-1 \text{ on } (x_i, x_{i+1}) \text{ for } 0 \leq i \leq N\} .$$

We now fix our attention on some $S(2n-1, \Delta, z)$ which we denote by S, which satisfies

$$n > m \qquad \text{and} \qquad z \geq 2m - 1 .$$

(Note that if $z = 2n - 2$ then $z > 2m - 2$ so $z \geq 2m - 1$). The reasons for these restrictions will appear later.

Given a function g we want to consider an interpolate, denoted

by Ig, in S ; that is Ig is a function in S whose derivative values at certain prescribed points equal the corresponding values of the derivatives of g . Thus we define :

<u>Definition</u> (1.2.2) : *Let* $I : C^{n-1}[a, b] \longrightarrow S$ *be given by*

For $g \in C^{n-1}[a, b]$, Ig *is the* <u>*unique*</u> *function in* S *that satisfies* : $D^r Ig(a) = D^r g(a)$ *for* $0 \leq r \leq n-1$, $D^r Ig(b) = D^r g(b)$ *for* $0 \leq r \leq n-1$ *and* $D^r Ig(x_i) = D^r g(x_i)$ *for* $1 \leq i \leq N$, $0 \leq r \leq 2n-2-z$.

We need one other approximation space.

<u>Definition</u> (1.2.3) : *Let* S_o *be the subspace of* S *given by*

$$S_o = \{e \in S : D^r e(a) = D^r e(b) = 0 \quad \text{for} \quad 0 \leq r \leq m-1 \} .$$

Since $S \subset C^m[a, b]$, this definition makes sense. Note that if $g \in C^{n-1}[a, b]$ satisfies the conditions $D^r g(a) = D^r g(b) = 0$ for $0 \leq r \leq m-1$, then $Ig \in S_o$.

<u>Definition</u> (1.2.4) : *Let* $h = \max \{ x_{i+1} - x_i : 0 \leq i \leq N \}$, $h' = \min \{ x_{i+1} - x_i : 0 \leq i \leq N \}$ *and for* $g \in W^r(x_i, x_{i+1})$ *for* $0 \leq i \leq N$ *let*

$$||g||_{r,\Delta} = \left(\sum_{i=0}^{N} \sum_{s=0}^{r} \int_{x_i}^{x_{i+1}} (D^s g)^2 \right)^{1/2} .$$

We assume $h' \leq h \leq 1$ although this is not crucial. Note that $||\cdot||_{r,\Delta}$ is like the usual Sobolev norm. In fact for $g_1, g_2 \in W^r(x_i, x_{i+1})$

for $0 \leq i \leq N$ we have that

$$(||g_1 + g_2||_{r,\Delta})^2$$

$$= (||g_1||_{r,\Delta})^2 + (||g_2||_{r,\Delta})^2 + 2 \sum_{i=0}^{N} \sum_{s=0}^{r} \int_{x_i}^{x_{i+1}} D^s g_1 \, D^s g_2$$

$$\leq (||g_1||_{r,\Delta} + ||g_2||_{r,\Delta})^2 .$$

We let α be a bound on the <u>mesh ratio</u> ; that is $1 \leq (h/h') \leq \alpha$. In what follows we will use the symbol C to denote a generic constant that depends on $m, n, z, a, b,$ the mesh ratio bound α and the coefficient functions of L . The coefficient C is hardly ever the same in two different places.

From Schultz [11; p. 510, p. 512] we have :

<u>Theorem</u> (1.2.5) : *If* $g \in W^{2n}$ *then*

$$||D^r(g - Ig)||_0 \leq Ch^{2n-r}||D^{2n}g||_0 \quad \textit{for} \quad 0 \leq r \leq n .$$

<u>Theorem</u> (1.2.6) : *If* $g \in W^p$ *for* $n < p \leq 2n$, $4n-2p-1 \leq z \leq 2n-2$ *then*

$$\left(\sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} (D^r(g - Ig))^2 \right)^{1/2} \leq Ch^{(p-r)}||D^p g||_0 \quad \textit{for} \quad n < r \leq p .$$

From these two results we get the immediate corollary :

<u>Theorem</u> (1.2.7) : *If* $g \in W^{2n}$ *then for* $0 \leq r \leq 2n$ *we have*

$$||g - Ig||_{r,\Delta} \leq Ch^{2n-r}||g||_{2n}$$

where we may replace $||\cdot||_{r,\Delta}$ *by* $||\cdot||_r$ *if* $0 \leq r \leq 2m$.

Proof : We have the following with obvious simplifications if $r \leq n$:

$$||g - Ig||^2_{r,\Delta}$$

$$= \sum_{s=0}^{n} ||D^s(g - Ig)||_0^2 + \sum_{s=n+1}^{r} \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} (D^s(g - Ig))^2$$

$$\leq \sum_{s=0}^{n} C(s)h^{2(2n-s)}||D^{2n}g||_0^2 + \sum_{s=n+1}^{r} C(s)h^{2(2n-s)}||D^{2n}g||_0^2$$

$$\leq C||g||^2_{2n} \sum_{s=0}^{r} h^{2(2n-s)} \leq Ch^{2(2n-r)}||g||^2_{2n}$$

where we have used Theorems (1.2.5) and (1.2.6) and $h \leq 1$. ///

Clearly S_o is a finite dimensional subspace of L^2 . Thus we can choose a basis $\{e_i\}$ for $1 \leq i \leq q$ of S_o . The particular choice of basis is not important for the analysis as it does not enter into the determination of any estimates. However the choice is important for practical computations, as noted in Chapter 0 .

Note $S_o \subset H_o^m \cap W^{2m}$ by Theorems (1.1.6) and (1.1.8).

1.3 Restrictions on L

We now have enough background to state all our assumptions on the operator L . We will assume without further mention, throughout this section and throughout Chapter 2, that :

(1) $a'_{rs} \in C^{\max(4n-4m,2m)+r}[a, b]$ for $0 \leq r, s \leq m$.

(2) given $g \in W^p$ for $0 \leq p \leq \max(4n-4m, 2m)$ there exists an unique $v \in H_o^m \cap W^{p+2m}$ so that $Lv = g$ in W^p and

$$||v||_{p+2m} \leq C||g||_p = C||Lv||_p .$$

We will discuss in Section 3.1 what kind of operators have these properties. We note that given $g \in W^{2m}$ we have by Leibniz's rule :

$$Lg = \sum_{r,s=0}^{m} (-1)^r \sum_{p=0}^{r} \binom{r}{p} D^{r-p} a'_{rs} D^{p+s} g = \sum_{j=0}^{2m} a_j D^j g$$

where we define

$$a_j = \sum_{r=0}^{m} \left(\sum_{\substack{0 \leq p,s \leq m \\ p+s=j}} (-1)^r \binom{r}{p} D^{r-p} a'_{rs} \cdot \begin{Bmatrix} 0 & \text{if} & p < r \\ 1 & \text{if} & p \geq r \end{Bmatrix} \right)$$

for $1 \leq j \leq 2m$. Note $a_j \in C^{\max(4n-4m,2m)}[a, b]$ for $1 \leq j \leq 2m$ by our standing assumption (1).

We now show the general result :

<u>Theorem</u> (1.3.1) : *If* $g \in W^{p+2m}(c, d)$, $a_j \in C^p[a, b]$ *for* $1 \leq j \leq 2m$ *and* $(c, d) \subset (a, b)$ *then* $Lg \in W^p(c, d)$ *and*

$$||Lg||_{p,(c,d)} \leq C||g||_{p+2m,(c,d)} .$$

<u>Proof</u> : Using Leibniz's rule and the continuity of the functions $\{a_j\}$ we know $Lg \in W^p(c, d)$ by Theorem (1.1.6) and

$$||Lg||^2_{p,(c,d)} = \sum_{r=0}^{p} \int_c^d \left(D^r \left(\sum_{s=0}^{2m} a_s D^s g \right) \right)^2$$

$$= \sum_{r=0}^{p} \int_c^d \left(\sum_{s=0}^{2m} \sum_{t=0}^{r} \binom{r}{t} D^{r-t} a_s \, D^{s+t} g \right)^2$$

$$\leq C||g||^2_{p+2m,(c,d)} \quad . \qquad ///$$

Now since we can use Theorem (1.3.1) for $0 \leq p \leq \max(4n-4m, 2m)$ (by assumption (1)) we see that L is a continuous linear operator from $W^{p+2m} \cap H_o^m$ to W^p and by (2) we see that there exists a continuous linear operator $L^{-1} : W^p \longrightarrow W^{p+2m} \cap H_o^m$, for $0 \leq p \leq \max(4n-4m, 2m)$.

We now formally state these conclusions :

<u>Definition</u> (1.3.2) : *For* $r \geq 0$ *let* $V^r = W^r \cap H_o^m$.

<u>Theorem</u> (1.3.3) : *Given* $g \in V^{p+2m}$ *for* $0 \leq p \leq \max(4n-4m, 2m)$ *we know* $Lg \in W^p$ *and*

$$||g||_{p+2m} \leq C||Lg||_p \leq C||g||_{p+2m} \quad .$$

<u>Theorem</u> (1.3.4) : *Given* $g \in W^p$ *for* $0 \leq p \leq \max(4n-4m, 2m)$ *we know there exists an unique* $v \in V^{p+2m}$ *so that* $Lv = g$ *and*

$$||v||_{p+2m} \leq C||Lv||_p \quad .$$

Chapter 2

The Least Squares Methods : Continuous and Discrete

2.1 Introduction

We have established the setting for our analysis and we now are going to formally define and analyze the Continuous and Discrete least squares methods discussed in Chapter 0. We note that since we already have an estimate for the type of approximations possible in S_o, via Theorem (1.2.7), it would be satisfying to see our methods provide the same order of accuracy. We will see this is possible in some cases.

2.2 The Continuous Method

We will assume that $f \in W^{2n-2m}$ throughout this chapter so that the unique solution u is in V^{2n} by Theorem (1.3.4). We also have $||u||_{2n} \leq C||f||_{2n-2m}$.

We note that Sections 2.2 and 2.3 are elaborations of work done by Russell and Varah [10].

We now define the basic operator .

Definition (2.2.1) : *Let* $A : V^{2m} \longrightarrow S_o$ *be given by*

For $g \in V^{2m}$ *let* $Ag = \left\{ e \in S_o : ||Le - Lg||_0^2 \text{ is minimal} \right\}$.

We will soon show that A is in fact a well defined linear operator and hence Au is a well defined function in S_o. Thus Au is the Continuous least squares approximation discussed in Chapter 0 .

We will need :

<u>Definition</u> (2.2.2) : *Let* $b(\cdot, \cdot)$ *denote the bilinear map*

$$b(\cdot, \cdot) = \int_a^b L(\cdot)L(\cdot) \quad \text{on} \quad W^{2m} \times W^{2m} .$$

That $b(\cdot, \cdot)$ is well defined follows from Theorem (1.3.1). We now have :

<u>Theorem</u> (2.2.3) : A *is in fact a well defined linear operator from* V^{2m} *to* S_o ; *that is* $A \in L(V^{2m}, S_o)$.

<u>Proof</u> : We first show that A is well defined as an operator following Schultz [12; p. 69]. Let $g \in V^{2m}$.

Let $e \in S_o$. Then $e = \sum_{i=1}^{q} c_i e_i$ for some coefficients $\{c_i\}$. Now define

$$F(\underline{c}) = F(c_1, \ldots, c_q) = ||Lg - \sum_{i=1}^{q} c_i Le_i||_0^2 .$$

We will show that there is an <u>unique</u> $\underline{c}'$ minimizing F .

Now

$$F(\underline{c}) = \sum_{i,j=1}^{q} c_i c_j b(e_i, e_j) - 2 \sum_{i=1}^{q} c_i b(e_i, g) + ||Lg||_0^2 .$$

Thus F is a quadratic form in $\underline{c}$ so F has a minimum $\underline{c}'$ if and only if the following two conditions are satisfied :

(1) $(\partial F/\partial c_i)(\underline{c}') = 0$ for $1 \leq i \leq q$

(2) The matrix $H = [(\partial^2 F/\partial c_i \partial c_j)(\underline{c}')]_{i,j=1}^{q}$ is positive definite.

Condition (1) is satisfied precisely when $\underline{c}'$ satisfies

$$\sum_{j=1}^{q} b(e_i,e_j)c_j' = b(e_i,g) \quad \text{for } 1 \leq i \leq q . \tag{2.2.4}$$

This is a linear system that may be solved for $\underline{c}'$. We will show that an unique solution exists.

Let $\underline{d} = (d_1, \ldots, d_q)$ and $B = [b(e_i,e_j)]_{i,j=1}^{q}$. Then

$$\underline{d}^T B \underline{d} = b(\sum_{i=1}^{q} d_i e_i, \sum_{i=1}^{q} d_i e_i) \geq C||\sum_{i=1}^{q} d_i e_i||_{2m}^2$$

by Theorem (1.3.3) since $S_o \subset V^{2m}$. Thus B is positive definite. Hence an unique solution $\underline{c}'$ to (2.2.4) exists.

A calculation shows $H = 2B$ so we know H is positive definite and that condition (2) is satisfied when we calculate $\underline{c}'$ from (2.2.4).

Thus we can calculate an unique solution $\underline{c}'$ from (2.2.4) that minimizes $F(\underline{c})$. Hence setting

$$Ag = \sum_{i=1}^{q} c_i' e_i ,$$

we see that A is well defined as an operator.

Now we will show that A is linear. Let $g_1, g_2 \in V^{2m}$ and consider $A(g_1 + dg_2)$ for real d.

We can associate the form F as before and note that conditions (1) and (2) apply as before. Now condition (1) is satisfied for $\underline{c}'$ precisely when

$$\sum_{i=1}^{q} b(e_i,e_j)c_j' = b(e_i,g_1) + db(e_i,g_2) \quad \text{for } 1 \leq i \leq q.$$

Thus in an obvious notation : $\underline{c}'(g_1 + dg_2) = \underline{c}'(g_1) + d\underline{c}'(g_2)$. Condition (2) is satisfied since $H = 2B$ as before. Thus

$$A(g_1 + dg_2) = A(g_1) + dA(g_2) . \qquad ///$$

Hence A is a well defined linear operator which may be calculated in practice by solving (2.2.4) for $\underline{c}'$. We note that by (2.2.4) we have $b(g - Ag,e) = 0$ for $e \in S_o$.

By Theorem (1.3.3), for $g \in V^{2m}$ we have

$$C||g||_{2m}^2 \leq ||Lg||_0^2 = b(g,g) \leq C||g||_{2m}^2 .$$

Thus we conclude that $b(\cdot, \cdot)$ is an inner product on the subspace V^{2m} of W^{2m} and the norm associated with this inner product is equivalent to the $||\cdot||_{2m}$ norm on V^{2m} . Thus from (2.2.4) we see that Ag is the unique projection of $g \in V^{2m}$ onto the subspace S_o of V^{2m} in the inner product $b(\cdot, \cdot)$.

We also have :

<u>Theorem</u> (2.2.5) : *Considering* S_o *and* V^{2m} *as subspaces of* W^{2m} *we have that* $A \in \mathcal{B}(V^{2m}, S_o)$.

<u>Proof</u> : Let $g \in V^{2m}$. Now $||L(Ag - g)||_0^2 \leq ||L(0 - g)||_0^2$ by the minimization property of the operator A . Thus using Theorem (1.3.3) we have

$$\begin{aligned} ||Ag||_{2m} &\leq ||Ag - g||_{2m} + ||g||_{2m} \\ &\leq C||L(Ag - g)||_0 + ||g||_{2m} \\ &\leq C||g||_{2m} . \end{aligned}$$

Hence $A \in \mathcal{B}(V^{2m}, S_o)$. ///

<u>Theorem</u> (2.2.6) : *Given* $g \in V^{2n}$ *we have*

$$||g - Ag||_{2m} \leq Ch^{2n-2m}||g||_{2n} .$$

<u>Remark</u> (2.2.7) : We know that $f \in W^{2n-2m}$. Thus we have

$$||u - Au||_{2m} \leq Ch^{2n-2m}||f||_{2n-2m}$$

since $||u||_{2n} \leq C||f||_{2n-2m}$ by Theorem (1.3.4).

<u>Proof</u> : Since $g \in V^{2n}$ we know $\mathcal{I}g \in S_o$ exists. Hence using the minimization properties of A we have

$$||L(g - Ag)||_0^2 \leq ||L(g - \mathcal{I}g)||_0^2 .$$

Thus by Theorems (1.3.3) and (1.2.7) we have

$$||g - Ag||_{2m} \leq C||L(g - Ag)||_0 \leq C||L(g - Ig)||_0$$

$$\leq C||g - Ig||_{2m} \leq Ch^{2n-2m}||g||_{2n} . \qquad ///$$

Thus we see that Au and Iu both provide the same order of accuracy as approximations to u when measured in the $||\cdot||_{2m}$ norm.

Note that the positive definite system (2.2.4) which must be solved to find Au is just equation (0.1.2) as discussed before. Thus we only require knowledge of L and f to calculate the function Au in practice.

2.3 Further Error Bounds for the Continuous Method

We now wish to obtain estimates on the quantity $||u - Au||_0$. The main objective is the following theorem whose proof uses Banach space interpolation theory and a trick, due to Nitsche.

Theorem (2.3.1) : *Suppose that* $f \in W^{2n-2m}$. *Then*

$$||u - Au||_0 \leq Ch^{\min(2n,4n-4m)}||f||_{2n-2m} .$$

We remark that we do not obtain convergence at the interpolate rate in the case where $4n - 4m < 2n$ since by Theorems (1.2.7) and (1.3.4) we see

$$||u - Iu||_0 \leq Ch^{2n}||u||_{2n} \leq Ch^{2n}||f||_{2n-2m} .$$

However for large enough n we get convergence at the interpolate rate.

We will need the following two Lemmas before we can prove the Theorem. The proof of the first uses techniques from Bramble and Schatz [5; p. 4].

<u>Lemma</u> (2.3.2) : *Say* $v \in W^{2m}$ *where* $2n-2m > 2m$. *Then there is an* $e \in S_o$ *so that* $||v - Le||_0 \leq Ch^{2m}||v||_{2m}$.

<u>Proof</u> : Let $LS_o = \{Le : e \in S_o\}$. Now LS_o is a finite dimensional subspace of W^0 by Theorem (1.3.4) since $S_o \subset V^{2m}$. Hence LS_o is closed in W^0 .

Let $Y = W^0/LS_o$ be the usual quotient space which is a Banach space under the norm

$$[g] \in Y \longrightarrow ||[g]||_Y = \inf\{||g - e||_0 : e \in LS_o\}$$

where $[g]$ denotes the equivalence class of $g \in W^0$.

Say $g \in W^0$. Then

$$\inf\{||g - e||_0 : e \in LS_o\} \leq ||g - 0||_0 = ||g||_0 . \tag{2.3.3}$$

Say $g \in W^{2n-2m}$. Then we will show that

$$\inf\{||g - e||_0 : e \in LS_o\} \leq Ch^{2n-2m}||g||_{2n-2m} . \tag{2.3.4}$$

Since $g \in W^{2n-2m}$ there is an unique $g' \in V^{2n}$ so that $Lg' = g$ and $||g'||_{2n} \leq C||g||_{2n-2m}$, by Theorem (1.3.4). Now $Ig' \in S_o$ is defined so by Theorems (1.2.7) and (1.3.3) we have

$$\inf\{||g - e||_0 : e \in LS_o\} \leq ||g - L(Ig')||_0$$

$$= ||L(g' - Ig')||_0 \leq C||g' - Ig'||_{2m}$$

$$\leq Ch^{2n-2m}||g'||_{2n} \leq Ch^{2n-2m}||g||_{2n-2m} .$$

Thus we have demonstrated equation (2.3.4).

Now we define $T : W^0 \longrightarrow Y$ by $Tg = [g]$ for $g \in W^0$, which is the usual map associating the equivalence class of g in Y with the function g.

Let $g \in W^0$. Then (2.3.3) implies

$$||Tg||_Y = ||[g]||_Y \leq ||g||_0 .$$

Let $g \in W^{2n-2m}$. Then (2.3.4) implies

$$||Tg||_Y = ||[g]||_Y \leq Ch^{2n-2m}||g||_{2n-2m} .$$

Thus $T \in \mathcal{B}(W^0, Y) \cap \mathcal{B}(W^{2n-2m}, Y)$. Now setting $r = 2n-2m$, $\theta = 2m/(2n-2m)$, we can apply Theorem (1.1.10) to show $T \in \mathcal{B}(W^{2m}, Y)$ and

$$||Tg||_Y = \inf\{||g - e||_0 : e \in LS_o\}$$

$$\leq C\, 1^{1-\theta}\, (Ch^{2n-2m})^{\theta}\, ||g||_{2m}$$

$$= Ch^{2m}||g||_{2m}$$

for $g \in W^{2m}$.

Thus $\inf\{||v - Le||_0 : e \in S_o\} \leq Ch^{2m}||v||_{2m}$. Hence choosing a new $C = 2C$, we have the result. ///

We also have :

<u>Lemma</u> (2.3.5) : *Say* $v \in W^{2m}$ *where* $2n-2m \leq 2m$. *Then letting* v' *be the unique function in* V^{4m} *that satisfies* $Lv' = v$ *we have*

$$||v - L(Iv')||_0 \leq Ch^{2n-2m}||v||_{2m} .$$

<u>Proof</u> : The existence and uniqueness of $v' \in V^{4m}$, as stated in the Lemma, is guaranteed by Theorem (1.3.4). We also have

$$||v'||_{4m} \leq C||Lv'||_{2m} = C||v||_{2m} .$$

Now $v' \in V^{4m} \subset V^{2n}$ since $2n \leq 4m$. Thus $Iv' \in S_o$ exists and hence by Theorems (1.2.7) and (1.3.3) we have

$$\begin{aligned} ||v - L(Iv')||_0 &= ||L(v' - Iv')||_0 \\ &\leq C||v' - Iv'||_{2m} \leq Ch^{2n-2m}||v'||_{2n} \\ &\leq Ch^{2n-2m}||v'||_{4m} \leq Ch^{2n-2m}||Lv'||_{2m} \\ &= Ch^{2n-2m}||v||_{2m} . \end{aligned}$$

///

Now we are ready for the proof of the main result.

<u>Proof of Theorem</u> (2.3.1) : Suppose $||u - Au||_0 \neq 0$; otherwise the result is trivial. Let v be the unique function in V^{2m} that satisfies $Lv = (u - Au)/||u - Au||_0$ $(\in W^0)$ whose existence is guaranteed by Theorem (1.3.4). Note that we also have $||v||_{2m} \leq C||Lv||_0 = C$.

Now since $v \in V^{2m}$ and $u - Au \in V^{2m}$ we know that $u - Au$ and v

satisfy the endpoint conditions given by Theorem (1.1.8) for $r = m$.

Thus integrating by parts and using the self adjointness of L we have

$$||u - Au||_0 = \int_a^b (u - Au)^2/||u - Au||_0$$

$$= \int_a^b (u - Au)\cdot Lv$$

$$= \sum_{r,s=0}^{m} (-1)^r \int_a^b (u - Au)D^r(a'_{rs}D^s v)$$

$$= \sum_{r,s=0}^{m} \int_a^b a'_{rs}\, D^r(u - Au)\, D^s v$$

$$= \sum_{r,s=0}^{m} (-1)^s \int_a^b D^s(a'_{rs}D^r(u - Au))v$$

$$= \int_a^b L(u - Au)v \ .$$

(The above representation of $||u - Au||_0$ is Nitsche's trick).

By equation (2.2.4) we know that given $e \in S_o$, we have $b(u-Au, e) = 0$. Thus for any $e \in S_o$,

$$||u - Au||_0 = \int_a^b L(u - Au)\, v - b(u - Au, e)$$

$$= \int_a^b L(u - Au)\, (v - Le)$$

$$\leq \ ||L(u - Au)||_0 ||v - Le||_0$$

$$\leq \ C||u - Au||_{2m} ||v - Le||_0$$

by Theorem (1.3.3). Recall that by Theorem (2.2.6) we know

$$||u - Au||_{2m} \leq Ch^{2n-2m} ||f||_{2n-2m} .$$

Now, if $2n-2m > 2m$ then by Lemma (2.3.2) we can choose $e \in S_o$ so that $||v - Le||_0 \leq Ch^{2m} ||v||_{2m}$.

Or if $2n-2m \leq 2m$ then by Lemma (2.3.5) we can choose $e \in S_o$ so that $||v - Le||_0 \leq Ch^{2n-2m} ||v||_{2m}$.

Hence we can choose $e \in S_o$ so that

$$||v - Le||_0 \leq Ch^{\min(2m,2n-2m)} ||v||_{2m} \leq Ch^{\min(2m,2n-2m)} .$$

With this choice we thus have

$$\begin{aligned} ||u - Au||_0 &\leq C||u - Au||_{2m} ||v - Le||_0 \\ &\leq Ch^{\min(2m,2n-2m)+2n-2m} ||f||_{2n-2m} \\ &= Ch^{\min(2n,4n-4m)} ||f||_{2n-2m} . \end{aligned}$$

///

2.4 A Discussion

We have examined the Continuous least squares method and its associated error bounds. Now we intend to examine the Discrete least squares method.

We will soon define a new operator much like the operator A defined before. We will show that the calculation of values for our new operator will involve the solution of a system like (2.2.4) with approximate integration. Thus the new system will be amenable to automatic computation as discussed in Chapter 0 .

However much background is needed before we can describe the necessary integral approximations and prove error results. We thus find it necessary to leave the mainstream of our development and look at some general techniques. We will return to our problem in Section 2.7.

2.5 An Interpolate

We wish to look now at a specific way of finding polynomial interpolates that will prove very useful later (see Section 2.8). The results and methods of proof of our main theorem come from Ciarlet and Raviart [6].

Theorem (2.5.1) : *Say* $0 < Q \leq M-1$. *Let* $c \leq c_1 < c_2 < \ldots < c_{M-Q} \leq d$ *be* $M-Q$ *points in* $[c, d]$. *Let* $c_j' = (2c_j-d-c)/(d-c)$ *for* $1 \leq j \leq M-Q$ *noting that* $y = (2x-d-c)/(d-c)$ *is the unique affine transformation of* $[c, d]$ *onto* $[-1, 1]$.

Suppose $g \in W^M(c, d)$. *Then* :

(1) *There exist* M *unique polynomials* $\{p_j'(y)\}$ *of degree* $M-1$ *on* $[-1, 1]$ *that satisfy* :

(i) *For* $1 \leq j \leq Q$: $D^Q p_j'(c_i') = 0$ *for* $1 \leq i \leq M-Q$

$$D^r p_j'(c_1') = \delta_{r(j-1)} \quad \textit{for} \quad 0 \le r \le Q-1$$

(ii) *For* $Q+1 \le j \le M$: $D^Q p_j'(c_i') = \delta_{i(j-Q)}$ *for* $1 \le i \le M-Q$

$$D^r p_j'(c_1') = 0 \quad \textit{for} \quad 0 \le r \le Q-1 .$$

(2) *There exists an unique polynomial* $Eg(x)$ *of degree* $M-1$ *on* $[c, d]$ *that satisfies* $D^Q Eg(c_j) = D^Q g(c_j)$ *for* $1 \le j \le M-Q$ *and* $D^r Eg(c_1) = D^r g(c_1)$ *for* $0 \le r \le Q-1$.

(3) *For* $0 \le r \le M$ *and* $(d-c) \le 2$ *we have*

$$||g - Eg||_{r,(c,d)} \le CG(d-c)^{M-r} \, ||g||_{M,(c,d)}$$

where $G = \max\{|D^r p_j'(y)| : -1 \le y \le 1, \ 0 \le r \le M-1, \ 1 \le j \le M-Q\}$.

(*This theorem constructs an interpolate* Eg *of* g *on* $[c, d]$ *of a generalized* <u>*Hermite*</u> *type*).

<u>Proof</u> : We first prove the following lemma.

<u>Lemma</u> : *Say* $s \le d_1 < d_2 < \ldots < d_{M-Q} \le t$ *are* $M-Q$ *distinct points in* $[s, t]$. *Then there exist* M *unique polynomials* $\{g_j\}$ *of degree* $M-1$ *on* $[s, t]$ *that satisfy* :

(i) *For* $1 \le j \le Q$: $D^Q g_j(d_i) = 0$ *for* $1 \le i \le M-Q$

$$D^r g_j(d_1) = \delta_{r(j-1)} \quad \textit{for} \quad 0 \le r \le Q-1$$

(ii) *For* $Q+1 \le j \le M$: $D^Q g_j(d_i) = \delta_{r(j-Q)}$ *for* $1 \le i \le M-Q$

$$D^r g_j(d_1) = 0 \quad \textit{for} \quad 0 \le r \le Q-1 .$$

<u>Proof of Lemma</u> : We first consider the cases when $1 \leq j \leq Q$.

We let $g_j(x) = \sum_{r=1}^{M} d_{rj}x^{r-1}$ where we will uniquely specify the coefficients. Since $D^Q g_j$ is a polynomial of degree $M-Q-1$ that is required to vanish at $M-Q$ points we must let $d_{rj} = 0$ for $Q+1 \leq r \leq M$. Thus $g_j(x)$ will be determined when we fix d_{rj} for $1 \leq r \leq Q$.

Now $D^{Q-1}g_j(d_1) = (Q-1)!d_{Qj} = \delta_{jQ}$ is desired. Using this equation we can uniquely determine d_{Qj} .

Now $D^{Q-2}g_j(d_1) = (Q-1)!d_{Qj}d_1 + (Q-2)!d_{(Q-1)j}$ and $D^{Q-2}g_j(d_1) = \delta_{j(Q-1)}$ is desired. Using this equation we can uniquely define $d_{(Q-1)j}$.

Continuing in this fashion we can define the rest of our coefficients uniquely for $1 \leq j \leq Q$. Thus $g_j(x)$ is uniquely determined for $1 \leq j \leq Q$ so that the necessary properties hold.

We now consider the cases where $Q+1 \leq j \leq M$.

We again let $g_j(x) = \sum_{r=1}^{M} d_{rj}x^{r-1}$ where we will again uniquely specify the coefficients. Now $D^Q g_j(x)$ is a polynomial of degree $M-Q-1$ that is required to satisfy $(D^Q g_j)(d_r) = \delta_{r(j-Q)}$ for $1 \leq r \leq M-Q$.

Let $(D^Q g_j)(x) = \sum_{r=1}^{M-Q} d'_{rj}x^{r-1}$. Now by the notation of the last paragraph we have $d_{(r-Q)j}r(r-1)\ldots(r-Q+1) = d'_{rj}$ for $1 \leq r \leq M-Q$. Thus uniquely specifying d'_{rj} for $1 \leq r \leq M-Q$ we have uniquely specified d_{rj}

for $Q+1 \leq r \leq M$. Let the matrix D be defined by $D = [(d_r)^{i-1}]_{r,i=1}^{M-Q}$. Now $\det D$ is a Vandermonde determinant and is nonzero as $d_r \neq d_i$ for $r \neq i$. Thus we can uniquely specify $\{d'_{rj}\}$ as the solution of the system

$$\sum_{i=1}^{M-Q} (D)_{ri}\, d'_{ij} = \delta_{r(j-Q)} \qquad \text{for } 1 \leq r \leq M-Q .$$

Hence d_{rj} for $Q+1 \leq r \leq M$ has been uniquely specified.

Now $D^{Q-1} g_j(d_1) = 0$ is required. We thus uniquely define d_{Qj} by letting

$$(Q-1)!\,d_{Qj} = -D^{Q-1} \left(\sum_{r=Q+1}^{M} d_{rj} x^{r-1} \right) \Bigg|_{x=d_1} .$$

Now $D^{Q-2} g_j(d_1) = 0$ is required. We thus uniquely define $d_{(Q-1)j}$ by letting

$$d_1 (Q-1)!\,d_{Qj} + (Q-2)!\,d_{(Q-1)j} = -D^{Q-2} \left(\sum_{r=Q+1}^{M} d_{rj} x^{r-1} \right) \Bigg|_{x=d_1} .$$

Continuing in this fashing we can define the rest of the coefficients, and thus $g_j(x)$ for $Q+1 \leq j \leq M$, so that the necessary properties hold.

Thus the Lemma is proved. ///

Now we apply this Lemma with $s = -1$, $d = 1$ and $d_j = c'_j$ to construct the M unique polynomials $\{p'_j(y)\}$ as described in conclusion (1) of the theorem.

We also apply this Lemma with $s = c$, $t = d$ and $d_j = c_j$ to construct M unique polynomials of degree $M-1$ on $[c, d]$, which we denote by $\{p_j(x)\}$, that satisfy the conclusions of the Lemma.

Now $g \in C^Q[c, d]$ so we can let Eg be defined by

$$Eg(x) = \sum_{j=1}^{Q} D^{j-1}g(c_1)p_j(x) + \sum_{j=1}^{M-Q} D^Q g(c_j)p_{j+Q}(x) .$$

Thus by our definition of the $\{p_j\}$ we see that Eg is the <u>unique</u> polynomial on $[c, d]$ of degree $M-1$ that satisfies conclusion (2) of the theorem.

Now consider $g'(y) = g(((d-c)y + d + c)/2)$. Note $g' \in W^M(-1, 1)$. Thus we can define $E'g'(y)$ as the unique polynomial of degree $M-1$ on $[-1, 1]$ that satisfies $D^r E'g'(c_1') = D^r g'(c_1')$ for $0 \le r \le Q-1$ and $D^Q E'g'(c_j') = D^Q g'(c_j')$ for $1 \le j \le M-Q$ as was done above on $[c, d]$. Thus

$$E'g'(y) = \sum_{j=1}^{Q} D^{j-1}g'(c_1')p_j'(y) + \sum_{j=1}^{M-Q} D^Q g'(c_j')p_{j+Q}'(y)$$

where the $\{p_j'\}$ were defined previously. It is easily seen that $Eg(((d-c)y + d + c)/2)$ satisfies the same conditions as $E'g'(y)$ by a simple calculation using the known conditions satisfied by Eg on $[c, d]$. Thus by the uniqueness of our constructions we have

$$E'g'(y) = Eg(((d-c)y + d + c)/2) .$$

Now we will show :

<u>Lemma</u> : *For* $0 \leq r < M$ *and* $-1 \leq y \leq 1$ *we have*

$$D^r E' g'(y) = D^r g'(y) + \sum_{j=1}^{M-Q} R_y^Q(c_j') D^r p_{j+Q}'(y) + \sum_{j=1}^{Q} R_y^{(j-1)}(c_1') D^r p_j'(y)$$

where

$$R_y^r(x) = \frac{1}{(M-r-1)!} \int_y^x (D^M g'(t))(x-t)^{M-r-1}\, dt \quad \textit{for} \quad x \in [-1, 1].$$

<u>Proof of Lemma</u> : Since $g' \in W^M(-1, 1)$ we know we can Taylor expand g' about any $x \in [-1, 1]$ by Theorem (1.1.9). Hence for $-1 \leq x, y \leq 1$ and $0 \leq r < M$ we can define

$$P_x^r g'(y) = \sum_{j=r}^{M-1} \frac{1}{(j-r)!} D^j g'(x)\ (y-x)^{j-r}$$

and $R_x^r g'(y)$ as noted above. Note that

$$(D^j g')(c_r') = P_x^j g'(c_r') + R_x^j g'(c_r')$$

for $0 \leq j < M$ and $1 \leq r \leq M-Q$.

Now let

$$P_x g'(y) = \sum_{r=0}^{M-1} \frac{1}{(r)!} D^r g'(x)\ (y-x)^r$$

for $-1 \leq x, y \leq 1$. A simple calculation shows that $D^r P_x g'(y) = P_x^r g'(y)$ for $0 \leq r < M$.

Now E' is a linear operator defined for functions in $W^M(-1, 1)$ so that $E' P_x g'(y)$ is defined. Because of the uniqueness of our constructions we have $E' P_x g'(y) = P_x g'(y)$ for $-1 \leq x, y \leq 1$.

Hence for $0 \leq r < M$ and $-1 \leq x, y \leq 1$ we have

$$D^r E'g'(y) = \sum_{j=1}^{Q} R_x^{(j-1)} g'(c_1') D^r p_j'(y) + \sum_{j=1}^{M-Q} R_x^Q g'(c_j') D^r p_{j+Q}'(y)$$

$$+ D_y^r \left(\sum_{j=1}^{Q} D^{j-1} P_x g'(c_1') p_j'(y) + \sum_{j=1}^{M-Q} D^Q P_x g'(c_j') p_{j+Q}'(y) \right)$$

$$= D_y^r (E' P_x g'(y)) + \sum_{j=1}^{Q} R_x^{(j-1)} g'(c_1') D^r p_j'(y) + \sum_{j=1}^{M-Q} R_x^Q g'(c_j') D^r p_{j+Q}'(y)$$

$$= D_y^r (P_x g'(y)) + \sum_{j=1}^{Q} R_x^{(j-1)} g'(c_1') D^r p_j'(y) + \sum_{j=1}^{M-Q} R_x^Q g'(c_j') D^r p_{j+Q}'(y) .$$

Now choose $x = y$ in the above. Thus for $0 \leq r < M$ we have

$$D_y^r (P_x g'(y)) \big|_{x=y} = D^r g'(y) .$$

Thus the Lemma is proved. ///

Now let G be defined as in conclusion (3) and let $0 \leq r < M$.

Thus

$$||g - Eg||_{r,(c,d)}^2 = \sum_{s=0}^{r} \int_c^d (D^s(g - Eg)(x))^2 \, dx$$

$$= \sum_{s=0}^{r} \int_{-1}^{1} \left(\left(\frac{2}{d-c}\right)^s D_y^s (g' - E'g')(y) \right)^2 \left(\frac{d-c}{2}\right) dy$$

$$= \sum_{s=0}^{r} \int_{-1}^{1} \left(\left(\frac{2}{d-c}\right)^s \left(\sum_{j=1}^{M-Q} R_y^Q g'(c_j') D^s p_{j+Q}'(y) + \right.\right.$$

$$\left.\left. \sum_{j=1}^{Q} R_y^{(j-1)} g'(c_1') D^s p_j'(y) \right)^2 \left(\frac{d-c}{2}\right) dy$$

$$\leq C^2G^2 \left(\frac{2}{d-c}\right)^{1-2r} \int_{-1}^{1} (D^M g'(y))^2 \, dy$$

$$= C^2G^2 \, (d-c)^{2(M-r)} \int_c^d (D^M g(x))^2 \, dx$$

$$\leq C^2G^2 \, (d-c)^{2(M-r)} ||g||^2_{M,(c,d)}$$

where we have used the second Lemma and changes in the integration variable. Noting that the case $r = M$ is done since Eg has degree $M-1$ on $[c, d]$, the theorem is proved. ///

2.6 Quadrature Sums

We now describe a well known method of approximating integrals that will be used in Section 2.7.

Say $g \in C[c, d]$ and we want to find a number close to the number $\int_c^d g(x)\, dx$. Let $K > 0$. We choose to find points $\{z_j\}$ for $1 \leq j \leq K$ (so that $c < z_1 < \ldots < z_K < d$) and numbers $\{w_j\}$ for $1 \leq j \leq K$ (so that $w_j > 0$) so that the quadrature sum

$$\sum_{j=1}^{K} w_j g(z_j) \quad \text{is close to} \quad \int_c^d g(x)\, dx .$$

The points $\{z_j\}$ are called nodes or quadrature points and the numbers $\{w_j\}$ are called quadrature weights.

Given K nodes and K weights, we say that the corresponding quadrature sum has a degree of precision r if r is the largest integer

so that

$$\int_c^d g = \sum_{j=1}^{K} w_j g(z_j)$$

for all polynomials g of degree r.

We will choose our nodes and weights so the corresponding quadrature sum has maximum degree of precision. The unique nodes and weights characterized by maximum degree of precision are called the Gaussian nodes and weights and the corresponding quadrature is called Gaussian quadrature.

To define Gaussian quadrature, we need some notation. Let $g_0, g_1, \ldots$ be the unique sequence of orthonormal polynomials obtained by applying the Gram-Schmidt process to the sequence of polynomials $1, x, x^2, \ldots,$ using the inner product $\int_c^d (\cdot)(\cdot)$. Thus $(g_i, g_j)_{0,(c,d)} = \delta_{ij}$ for $i, j \geq 0$ and g_r has degree r. We note that it is well known that g_r has r distinct roots in (c, d). The polynomials $\{g_r\}$ are called the Legendre polynomials on $[c, d]$.

Now we can state :

Theorem (2.6.1) : (1) *A quadrature sum over* K *distinct nodes can have a degree of precision of at most* $2K - 1$.

(2) *Case* (1) *is attained if and only if the* K *nodes* $\{z_j\}$ *are the* K *roots of* g_K *in* (c, d).

(3) *The weights* $\{w_j\}$ *used when case* (1) *is attained are unique, strictly positive and satisfy*

$$w_j = \int_c^d \prod_{j \neq i=1}^{K} \left(\frac{x - z_i}{z_j - z_i} \right) dx \qquad \text{for } 1 \leq j \leq K .$$

<u>Proof</u> : See Isaacson and Keller [9; p. 328]. ///

Hence the nodes for Gaussian quadrature are the K roots of g_K and the weights are as given in Theorem (2.6.1). We will often call the nodes the <u>Gauss points</u> .

Let $\{z_j'\}$ for $1 \leq j \leq K$ be the K Gauss points on $[-1, 1]$ with associated weights $\{w_j'\}$ for $1 \leq j \leq K$. Let $g_o', g_1', g_2', \ldots$ be the Legendre polynomials on $[-1, 1]$.

Now $g_r(((d-c)y + c + d)/2)$ is a polynomial of degree r on $[-1, 1]$. Also for $r, s \geq 0$ we have

$$\int_{-1}^{1} g_r(((d-c)y+c+d)/2) g_s(((d-c)y+c+d)/2) \, dy = \left(\frac{2}{d-c}\right) \delta_{rs} .$$

Thus by the uniqueness ensured by the Gram-Schmidt process we have that

$$g_r'(y) = \left(\frac{d-c}{2}\right)^{1/2} g_r(((d-c)y + d + c)/2) .$$

Thus the Gauss points $\{z_j\}$ on $[c, d]$ and the Gauss points $\{z_j'\}$ on $[-1, 1]$ are related by

$$z_j = ((d-c)z_j' + d + c)/2 \qquad \text{for } 1 \leq j \leq K .$$

Now by Theorem (2.6.1) we have

$$w_j = \int_c^d \prod_{j \neq i=1}^{K} \left(\frac{x - z_i}{z_j - z_i} \right) dx$$

$$= \left(\frac{d-c}{2}\right) \int_{-1}^{1} \prod_{j \neq i=1}^{K} \left(\frac{y - z_i'}{z_j' - z_i'} \right) dy = \left(\frac{d-c}{2}\right) w_j'$$

for $1 \leq j \leq K$.

Hence the weights $\{w_j\}$ on $[c, d]$ and the weights $\{w_j'\}$ on $[-1, 1]$ are related by

$$w_j = ((d-c)/2)w_j' \qquad \text{for } 1 \leq j \leq K .$$

These relations between the parameters of the Gaussian quadrature rule for different intervals will prove to be very useful later. In fact we soon will be discussing the Gauss points on (x_i, x_{i+1}) for $0 \leq i \leq N$ and will use the fact that the correct affine transformation of these Gauss points gives the Gauss points on $[-1, 1]$ (which are independent of i and h). How these transformations will aid obtaining error bounds has already been hinted at in Theorem (2.5.1) when we defined the peculiar constant G that relied on affine transformations of interpolation points.

2.7 Preliminary Error Bounds Needed for the Discrete Method

We now return to our particular differential equation problem by defining the specific integral approximations that will be used for the Discrete method.

It is known that the error involved in approximating the integral

of some function over an interval is proportional to some power of the length of the interval (Isaacson and Keller [9]). With this in mind we are now able to define our integral approximations as follows.

<u>Definition</u> (2.7.1) : *Let* $K > 0$ *and recall* $\Delta : a = x_o < \ldots < x_{N+1} = b$. *We denote the* K *Gauss points in* $[x_i, x_{i+1}]$ *by* $\{z_{ij}\}$ *for* $1 \leq j \leq K$ *and the associated weights by* $\{w_{ij}\}$ *for* $1 \leq j \leq K$. *We also denote the* K *Gauss points in* $[-1, 1]$ *by* $\{z_j'\}$ *for* $1 \leq j \leq K$ *and the associated weights by* $\{w_j'\}$ *for* $1 \leq j \leq K$.

Then given $g(x)$ so that $g \in C[x_i, x_{i+1}]$ for $0 \leq i \leq N$ we can approximate $\int_a^b g$ by the <u>composite quadrature sum</u> given by $\sum_{i=0}^{N} \sum_{j=1}^{K} w_{ij} g(z_{ij})$.

We note that from Section 2.6 we have

$$z_{ij} = ((x_{i+1}-x_i)z_j' + x_{i+1} + x_i)/2 \; ; \qquad w_{ij} = \frac{(x_{i+1}-x_i)}{2} w_j'$$

for $1 \leq j \leq K$. Of course the definitions of $\{z_j'\}$ and $\{w_j'\}$ are independent of h and i .

We now prove the basic integral approximation result :

<u>Theorem</u> (2.7.2) : *Say* $g \in W^{2K}(x_i, x_{i+1})$ *for* $0 \leq i \leq N$. *Then*

$$\left| \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} g(x)dx - \sum_{i=0}^{N} \sum_{j=1}^{K} w_{ij} g(z_{ij}) \right|$$

$$\leq Ch^{2K} \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} |D^{2K} g(t)| \, dt \; .$$

<u>Proof</u> : We note $D^{2K}g \in L^2(x_i, x_{i+1}) \subset L^1(x_i, x_{i+1})$ for $0 \le i \le N$. Now by Theorem (1.1.9) we can Taylor expand $g(x)$ about x_i for $0 \le i \le N$. Thus, letting

$$P_i g(x) = \sum_{r=0}^{2K-1} \frac{1}{r!} D^r g(x_i)(x - x_i)^r$$

and

$$R_i g(x) = \frac{1}{(2K-1)!} \int_{x_i}^{x} (D^{2K}g(t))(x - t)^{2K-1}\, dt ,$$

we have for $x_i \le x \le x_{i+1}$ that

$$g(x) = P_i g(x) + R_i g(x) .$$

We now recall that $w_{ij} = ((x_{i+1}-x_i)/2)w_j'$ for $0 \le i \le N$, $1 \le j \le K$ and that Gaussian quadrature is accurate for polynomials of degree $2K-1$ by Theorem (2.6.1). Noting that $P_i g(x)$ is a polynomial of degree $2K-1$ on $[x_i, x_{i+1}]$ for $0 \le i \le N$ we thus have

$$\left| \sum_{i=0}^{N} \left(\int_{x_i}^{x_{i+1}} g(x)\, dx - \sum_{j=1}^{K} w_{ij} g(z_{ij}) \right) \right|$$

$$= \left| \sum_{i=0}^{N} \left\{ \left(\int_{x_i}^{x_{i+1}} P_i g(x)\, dx - \sum_{j=1}^{K} w_{ij} P_i g(z_{ij}) \right) \right.\right.$$

$$\left.\left. + \left(\int_{x_i}^{x_{i+1}} R_i g(x)\, dx - \sum_{j=1}^{K} w_{ij} R_i g(z_{ij}) \right) \right\} \right|$$

$$= \left| \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} \frac{1}{(2K-1)!} \int_{x_i}^{x} (D^{2K}g(t))(x-t)^{2K-1}\, dt\, dx \right.$$

$$\left. + \sum_{i=0}^{N} \frac{1}{(2K-1)!} \sum_{j=1}^{K} \frac{(x_{i+1}-x_i)}{2} w_j' \int_{x_i}^{z_{ij}} (D^{2K}g(t))(z_{ij}-t)^{2K-1} dt \right|$$

$$\leq Ch^{2K} \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} |D^{2K}g(t)|\, dt \,. \qquad ///$$

We now need more notation .

<u>Definition</u> (2.7.3) : (1) *If* $g \in C^r(x_i, x_{i+1})$ *for* $0 \leq i \leq N$ *then let*

$$||g||'_{r,\Delta} = \left(\sum_{i=0}^{N} \sum_{s=0}^{r} \sum_{j=1}^{K} w_{ij} (D^s g(z_{ij}))^2 \right)^{1/2}$$

(2) *If* $g_1, g_2 \in C^{2m}(x_i, x_{i+1})$ *for* $0 \leq i \leq N$ *then let*

$$b'(g_1, g_2) = \sum_{i=0}^{N} \sum_{j=1}^{K} w_{ij} Lg_1(z_{ij}) Lg_2(z_{ij})$$

where we recall $\Delta : a = x_o < \ldots < x_{N+1} = b$.

We note that $||\cdot||'_{r,\Delta}$ is an approximation to $||\cdot||_{r,\Delta}$ where we have replaced integration on (x_i, x_{i+1}) with a quadrature sum. Similarly $b'(\cdot, \cdot)$ is an approximation to the bilinear map $b(\cdot, \cdot)$.

Now $||\cdot||'_{r,\Delta}$ satisfies the triangle inequality and hence is clearly a seminorm; say $g_1, g_2 \in C^r(x_i, x_{i+1})$ for $0 \leq i \leq N$. Then

$$(||g_1 + g_2||'_{r,\Delta})^2$$

$$= \sum_{s=0}^{r} \sum_{i=0}^{N} \sum_{j=1}^{K} w_{ij}\left[(D^s g_1(z_{ij}))^2 + (D^s g_2(z_{ij}))^2\right]$$

$$+ 2 \sum_{s=0}^{r} \sum_{i=0}^{N} \sum_{j=1}^{K} w_{ij} D^s g_1(z_{ij}) D^s g_2(z_{ij})$$

$$\leq (||g_1||'_{r,\Delta})^2 + (||g_2||'_{r,\Delta})^2$$

$$+ 2 \sum_{s=0}^{r} \sum_{i=0}^{N} \sum_{j=1}^{K} (w_{ij}^{1/2}|D^s g_1(z_{ij})|)(w_{ij}^{1/2}|D^s g_2(z_{ij})|)$$

$$\leq (||g_1||'_{r,\Delta} + ||g_2||'_{r,\Delta})^2$$

where we have used the positivity of the weights.

We can now state and prove the following with the additional assumptions on the coefficients $\{a_j\}$ of L noted.

<u>Theorem</u> (2.7.4) : *Suppose* $g_1, g_2 \in W^{2K+2m}(x_i, x_{i+1})$ *for* $0 \leq i \leq N$, $g_1, g_2 \in W^{2m}$ *and* $a_j \in C^{2K}[a, b]$ *for* $1 \leq j \leq 2m$. *Then*

$$|(b-b')(g_1, g_2)| \leq Ch^{2K}||g_1||_{2K+2m,\Delta}||g_2||_{2K+2m,\Delta} .$$

<u>Proof</u> : Note that $2K+2m > 2m$ so that $b'(g_1, g_2)$ is defined. We will now apply Theorem (2.7.2) noting that $Lg_1, Lg_2 \in W^{2K}(x_i, x_{i+1})$ for $0 \leq i \leq N$ by Theorem (1.3.1). Thus

$$|(b-b')(g_1, g_2)| \leq Ch^{2K} \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} |D^{2K}Lg_1Lg_2|$$

$$\leq Ch^{2K} \left(\sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} \left| \sum_{r=0}^{2K} (D^{2K-r}Lg_1)(D^rLg_2) \right| \right)$$

$$\leq Ch^{2K} ||g_1||_{2K+2m,\Delta} \, ||g_2||_{2K+2m,\Delta}$$

where we have used the continuity of the coefficients. ///

We now recall an inequality due to Schmidt and described in Bellman [3]. This inequality will eventually lead to some corollaries of the previous theorem.

<u>Proposition</u> (2.7.5) : *If* $g(x)$ *is a polynomial of degree* s *on* $[-1, 1]$ *then*

$$\int_{-1}^{1} (Dg)^2 \leq ((s+1)^4/2) \int_{-1}^{1} (g)^2 .$$

<u>Lemma</u> (2.7.6) : *If* $g(x)$ *is a polynomial of degree* s *on* $[c, d]$ *then*

$$(d-c)||Dg||_{0,(c,d)} \leq (s+1)^2 2^{1/2} ||g||_{0,(c,d)} .$$

<u>Proof</u> : Note that $g(((d-c)y + d + c)/2)$ is a polynomial of degree s on $[-1, 1]$. Thus by Proposition (2.7.5) we have

$$\int_{c}^{d} (Dg)^2 = \frac{d-c}{2} \int_{-1}^{1} \big((Dg)(((d-c)y + c + d)/2)\big)^2 \, dy$$

$$\leq \frac{(s+1)^4}{(d-c)} \int_{-1}^{1} \big(g(((d-c)y + c + d)/2)\big)^2 \, dy$$

$$= \frac{2(s+1)^4}{(d-c)^2} \int_c^d (g(x))^2 \, dx \, . \qquad ///$$

We now show the very important result :

<u>Theorem</u> (2.7.7) : *Say* e *is a polynomial of degree* p *on* (x_i, x_{i+1}) *for* $0 \le i \le N$ *and* $0 \le r \le s \le p$. *Then*

$$||e||_{s,\Delta} \le Ch^{r-s} ||e||_{r,\Delta} \, .$$

<u>Proof</u> : Using Lemma (2.7.6) repeatedly we have for $r < s$:

$$||e||^2_{s,\Delta} = \sum_{j=0}^{r} \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} (D^j e)^2 + \sum_{j=r+1}^{s} \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} (D^j e)^2$$

$$\le ||e||^2_{r,\Delta} + \sum_{i=0}^{N} \sum_{j=1}^{s-r} C(j)(h')^{-2j} \int_{x_i}^{x_{i+1}} (D^r e)^2$$

$$\le ||e||^2_{r,\Delta} + Ch^{2(r-s)} \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} (D^r e)^2$$

$$\le Ch^{2(r-s)} ||e||^2_{r,\Delta}$$

where we have used $h' \le h \le 1$. Noting that the theorem is trivial for $r = s$, we are done. ///

We now have the following corollaries to Theorem (2.7.4) using Theorem (2.7.7).

<u>Corollary</u> (2.7.8) : *Say* $a_j \in C^{2K}[a, b]$ *for* $1 \leq j \leq 2m$. *Then* :

(1) *If* $g \in W^{2m}$, $g \in W^{2K+2m}(x_i, x_{i+1})$ *for* $0 \leq i \leq N$ *and* $e \in S_o$ *then letting* $r = \max(0, 2K+2m-2n+1)$ *we have*

$$|(b-b')(g, e)| \leq Ch^r ||g||_{2K+2m,\Delta} \, ||e||_{2m,\Delta}$$

(2) *If* $e_1, e_2 \in S_o$ *then letting* $r = \max(-2K, 2K-2(2n-2m)+2)$ *we have*

$$|(b-b')(e_1, e_2)| \leq Ch^r ||e_1||_{2m,\Delta} \, ||e_2||_{2m,\Delta} \quad .$$

<u>Proof</u> : We will just demonstrate the first case as the second case is similar. Applying Theorems (2.7.4) and (2.7.7) we have

$$\begin{aligned} |(b-b')(g, e)| &\leq Ch^{2K} ||g||_{2K+2m,\Delta} \, ||e||_{2K+2m,\Delta} \\ &= Ch^{2K} ||g||_{2K+2m,\Delta} \, ||e||_{\min(2K+2m,2n-1),\Delta} \\ &\leq Ch^{2K+2m-\min(2K+2m,2n-1)} \, ||g||_{2K+2m,\Delta} \, ||e||_{2m,\Delta} \\ &= Ch^r ||g||_{2K+2m,\Delta} \, ||e||_{2m,\Delta} \end{aligned}$$

where we have used the fact that $e \in S_o$ has degree $2n-1$ on (x_i, x_{i+1}) for $0 \leq i \leq N$. ///

The second part of Corollary (2.7.8) will be decisive in our choice for K, the number of quadrature points. It seems natural to expect K to only depend on n and m but <u>not</u> h . However the particular choice

for K cannot be made until we make one more definition and examine a theorem. This is done in the next section where we define the Discrete method.

2.8 The Discrete Method

To specify our method we will define a mapping much like the A mapping of Definition (2.2.1). The mapping will depend on K, the number of quadrature points in each interval.

Definition (2.8.1) : *Let* $A' : V^{2n} \longrightarrow S_o$ *be given by : For* $g \in V^{2n}$ *let*

$$A'(g) = \left\{ e \in S_o : (||L(e-g)||'_{0,\Delta})^2 \text{ is minimal} \right\} .$$

We note that for $e \in S_o$ and $g \in V^{2n}$ the quantity $||L(e-g)||'_{0,\Delta}$ is defined as $L(e-g) \in C(x_i, x_{i+1})$ for $0 \leq i \leq N$ by the continuity of the coefficients of L .

We will now show that A' is a well defined linear operator for certain choices of K .

Theorem (2.8.2) : *For* $K \geq 2n-2m$, $a_j \in C^{2K}[a, b]$ *for* $1 \leq j \leq 2m$, *and sufficiently small* h, A' *is in fact a well defined linear operar from* V^{2n} *to* S_o ; *that is* $A' \in L(V^{2n}, S_o)$.

Proof : The proof follows the same line as that of Theorem (2.2.3). We first show that A' is well defined as an operator. Let $g \in V^{2n}$.

Let $e \in S_o$. Then $e = \sum_{i=1}^{q} c_i e_i$ for some coefficients $\{c_i\}$. Again define a functional

$$F'(\underline{c}) = \left(||Lg - \sum_{i=1}^{q} c_i Le_i||'_{0,\Delta} \right)^2 .$$

We again look for an <u>unique</u> $\underline{c}'$ minimizing F' .

We know that

$$F'(\underline{c}) = \sum_{i,j=1}^{q} c_i c_j b'(e_i,e_j) - 2 \sum_{i=1}^{q} c_i b'(e_i,g) + \left(||Lg||'_{0,\Delta}\right)^2 .$$

Thus F' is a quadratic form in $\underline{c}$ so F' has a minimum at $\underline{c}'$ if and only if the following two conditions are satisfied :

(1) $(\partial F'/\partial c_i)(\underline{c}') = 0$ for $1 \leq i \leq q$

(2) The matrix $H' = [(\partial^2 F'/\partial c_i \partial c_j)(\underline{c}')]_{i,j=1}^{q}$ is positive definite.

Now condition (1) is satisfied precisely when $\underline{c}'$ satisfies

$$\sum_{j=1}^{q} b'(e_r, e_j) c'_j = b'(e_r, g) \quad \text{for } 1 \leq r \leq q . \tag{2.8.3}$$

This is a linear system that may be solved for $\underline{c}'$. We will show that an unique solution exists.

Let $\underline{d} = (d_1, \ldots, d_q)$ and $B' = [b'(e_i, e_j)]_{i,j=1}^{q}$. Then

$$\underline{d}^T B' \underline{d} = b'(\sum_{i=1}^{q} d_i e_i, \sum_{i=1}^{q} d_i e_i)$$

$$= (b'-b)(\sum_{i=1}^{q} d_i e_i, \sum_{i=1}^{q} d_i e_i) + b(\sum_{i=1}^{q} d_i e_i, \sum_{i=1}^{q} d_i e_i).$$

Now

$$b(\sum_{i=1}^{q} d_i e_i, \sum_{i=1}^{q} d_i e_i) \geq C||\sum_{i=1}^{q} d_i e_i||_{2m}^2$$

by Theorem (1.3.3) since $S_o \subset V^{2m}$. Also we have that

$$|(b'-b)(\sum_{i=1}^{q} d_i e_i, \sum_{i=1}^{q} d_i e_i)| \leq Ch^2||\sum_{i=1}^{q} d_i e_i||_{2m}^2$$

by Corollary (2.7.8), part (2), since $K \geq 2n-2m$ and $S_o \subset W^{2m}$. Hence letting K' denote the constant obtained from Theorem (1.3.3), we have

$$\underline{d}^T B' \underline{d} \geq (K' - Ch^2)||\sum_{i=1}^{q} d_i e_i||_{2m}^2 .$$

Thus for sufficiently small h we have

$$\underline{d}^T B' \underline{d} \geq (K'/2)||\sum_{i=1}^{q} d_i e_i||_{2m}^2 .$$

Thus B' is a positive definite matrix for sufficiently small h and hence an unique solution $\underline{c}'$ to (2.8.3) exists.

Now a calculation shows $H' = 2B'$ so we know H' is positive definite also, for sufficiently small h. Thus condition (2) is satisfied when we calculate $\underline{c}'$ from (2.8.3).

Thus we can calculate a $\underline{c}'$ uniquely from (2.8.3) which minimizes $F'(\underline{c})$. Setting $A'g = \sum_{i=1}^{q} c_i' e_i$ we see that A' is well defined as an operator. The linearity of A' follows as in the proof of Theorem (2.2.3). ///

Thus we have shown A' is a well defined operator which may be calculated in practice by solving the linear system given by (2.8.3) for $\underline{c}'$. Of course the important point here is that (2.8.3) is like (2.2.4) except that the integrals have been replaced by composite quadrature sums. Since the quadrature sums are amenable to automatic computation we have a practical method, as noted in Chapter 0 .

It also may be seen why $K \geq 2n-2m$ is necessary. If K was less than $2n-2m$ then the power of h obtained from Corollary (2.7.8) would be zero or negative and we could not show that A' was well defined. In the sequel we will always assume $K = 2n-2m$, which is the practical choice. Note that $a_j \in C^{2K}[a, b]$ for $1 \leq j \leq 2m$ then, by Section (1.3).

Owing to the difficulties with the non-norm like behaviour of $b'(\cdot, \cdot)$ on V^{2m} we cannot obtain a characterization of our operator A' as a projection. However we can prove a result much like Theorem (2.2.6) although the proof is more difficult. In the following theorem the statement "for h sufficiently small" means that h must be small enough to obtain the necessary estimate used in the proof of Theorem (2.8.2). We recall that the "smallness" of h only depends on constants.

<u>Theorem</u> (2.8.4) : *Suppose* $K = 2n-2m$ *and* h *is sufficiently small. Then given* $g \in V^{2n}$ *we have*

$$||g - A'g||_{2m} \leq Ch^{2n-2m}||g||_{2n} .$$

<u>Remark</u> (2.8.5) : Suppose $K = 2n-2m$, h is sufficiently small and $f \in W^{2n-2m}$. Then we have

$$||u - A'u||_{2m} \leq Ch^{2n-2m}||f||_{2n-2m}$$

since $||u||_{2n} \leq C||f||_{2n-2m}$ by Theorem (1.3.4).

<u>Proof</u> : We start by showing

$$||(A - A')g||_{2m} \leq C||Ag - g||'_{2m,\Delta} .$$

Now as in the proof of Theorem (2.8.2) we have

$$\underline{d}^T B'\underline{d} = b'(\sum_{i=1}^{q} d_i e_i, \sum_{i=1}^{q} d_i e_i) \geq (K'/2)||\sum_{i=1}^{q} d_i e_i||^2_{2m}$$

where $\underline{d} = (d_1, \ldots, d_q)$, since h is sufficiently small. Let $\underline{d}$ represent the coefficients of $Ag - A'g \in S_o$. Thus

$$b'(Ag-A'g, Ag-A'g) \geq C||Ag - A'g||^2_{2m} .$$

From (2.8.3) we know $b'(e, g-A'g) = 0$ for $e \in S_o$. Hence

$$b'(Ag-A'g, Ag-A'g) = b'(Ag-A'g, Ag-g) .$$

Thus using Holder's inequality we have

$$C||Ag - A'g||_{2m}^2 \leq b'(Ag-A'g,\ Ag-g)$$

$$\leq \sum_{i=1}^{N} \sum_{j=1}^{K} (w_{ij}^{1/2}|L(Ag-A'g)(z_{ij})|)(w_{ij}^{1/2}|L(Ag-g)(z_{ij})|)$$

$$\leq b'(Ag-A'g,\ Ag-A'g)^{1/2} b'(Ag-g,\ Ag-g)^{1/2} \quad .$$

Now using Theorem (1.3.3) and Corollary (2.7.8) we have

$$b'(Ag-A'g,\ Ag-A'g)$$

$$= b(Ag-A'g,\ Ag-A'g) + (b'-b)(Ag-A'g,\ Ag-A'g)$$

$$\leq C||Ag - A'g||_{2m}^2 + Ch^2||Ag - A'g||_{2m}^2$$

$$\leq C||Ag - A'g||_{2m}^2 \quad .$$

By a simple computation using the continuity of the coefficients of L (as in Theorem (1.3.1)) we have

$$b'(Ag-g,\ Ag-g) \leq C(||Ag - g||'_{2m,\Delta})^2 \ .$$

Thus using the results above we have

$$C||Ag - A'g||_{2m}^2 \leq b'(Ag-A'g,\ Ag-A'g)^{1/2} b'(Ag-g,\ Ag-g)^{1/2}$$

$$\leq C||Ag - A'g||_{2m}\ ||Ag - g||'_{2m,\Delta} \ .$$

So $||Ag - A'g||_{2m} \leq C||Ag - g||'_{2m,\Delta}$ and we have our result.

We now apply Theorem (2.5.7) with $c = x_i$, $d = x_{i+1}$, $M = 2n$, $Q = 2m$, $c_j = z_{ij}$ for $1 \leq j \leq 2n-2m$ to obtain a polynomial $E_i g$ of

degree $2n-1$ on $[x_i, x_{i+1}]$ that satisfies :

$$(1) \quad D^{2m}E_i g(z_{ij}) = D^{2m}g(z_{ij}) \qquad \text{for } 1 \le j \le 2n-2m$$

$$(2) \quad D^r E_i g(z_{i1}) = D^r g(z_{i1}) \qquad \text{for } 0 \le r \le 2m-1 .$$

Also, for $0 \le s \le 2n$, we have

$$(2.8.6) \qquad \left(\sum_{r=0}^{s} \int_{x_i}^{x_{i+1}} (D^r(g-E_i g))^2\right)^{1/2} \le CGh^{2n-s}\left(\sum_{r=0}^{2n} \int_{x_i}^{x_{i+1}} (D^r g)^2\right)^{1/2} .$$

Recalling the definition of G and the remarks following Definition (2.7.1) we see that G depends only on $\{z_j'\}$ for $1 \le j \le 2n-2m$. Thus G is independent of h and i and hence we can denote G by C . We have also the following for $1 \le j \le K$ and $0 \le i \le N$:

$$\sum_{r=0}^{2m} w_{ij}(D^r(g-E_i g)(z_{ij}))^2 = \sum_{r=0}^{2m-1} w_{ij}(D^r(g-E_i g)(z_{ij}))^2 .$$

Now $g - E_i g \in C^{2m+1}[x_i, x_{i+1}]$ for $0 \le i \le N$. Thus we can use a one term Taylor expansion of $(D^r(g-E_i g))^2(x)$ about x_i for $x_i \le x \le x_{i+1}$ and $0 \le r \le 2m$. Recalling that our quadrature sum is exact for polynomials of degree zero, by (2.8.6) we have

$$\left|\sum_{i=0}^{N} \sum_{r=0}^{2m-1} \left(\sum_{j=1}^{K} w_{ij}\left(D^r(g-E_i g)(z_{ij})\right)^2 - \int_{x_i}^{x_{i+1}} (D^r(g-E_i g))^2\right)\right|$$

$$= \left|\sum_{i=0}^{N} \sum_{r=0}^{2m-1} \left\{\sum_{j=1}^{K} w_{ij}\left(\left(D^r(g-E_i g)(x_i)\right)^2 + \int_{x_i}^{z_{ij}} D\left(D^r(g-E_i g)(t)\right)^2 dt\right)\right.\right.$$

$$- \int_{x_i}^{x_{i+1}} \left(\left(D^r(g-E_i g)(x_i)\right)^2 + \int_{x_i}^{x} D\left(D^r(g-E_i g)(t)\right)^2 dt \right) dx \Bigg\} \Bigg|$$

$$= \left| \sum_{i=0}^{N} \sum_{r=0}^{2m-1} \left(\sum_{j=1}^{K} w_{ij} \int_{x_i}^{z_{ij}} D\left(D^r(g-E_i g)(t)\right)^2 dt \right.\right.$$

$$\left.\left. - \int_{x_i}^{x_{i+1}} \int_{x_i}^{x} D\left(D^r(g-E_i g)(t)\right)^2 dt\, dx \right) \right|$$

$$\leq \sum_{i=0}^{N} \sum_{r=0}^{2m-1} \left(\sum_{j=1}^{K} \frac{h}{2} w_j' + h \right) \int_{x_i}^{x_{i+1}} \left| D\left(D^r(g-E_i g)(t)\right)^2 \right| dt$$

$$\leq Ch \sum_{i=0}^{N} \sum_{r=0}^{2m} \int_{x_i}^{x_{i+1}} \left(D^r(g-E_i g)\right)^2$$

$$\leq Ch^{2(2n-2m)+1} \, ||g||_{2n}^2$$

where $\{w_j'\}$ are the Gauss weights on $[-1, 1]$.

Thus from the above we have

$$\sum_{i=0}^{N} \sum_{r=0}^{2m} \sum_{j=1}^{K} w_{ij} \left(D^r(g-E_i g)(z_{ij})\right)^2 \tag{2.8.7}$$

$$\leq Ch^{2(2n-2m)+1} ||g||_{2n}^2 + \sum_{i=0}^{N} \sum_{r=0}^{2m-1} \int_{x_i}^{x_{i+1}} \left(D^r(g-E_i g)\right)^2$$

$$\leq Ch^{2(2n-2m)+1} ||g||_{2n}^2 \; .$$

Now Ag and $E_i g$ are both polynomials of degree $2n-1$ on $[x_i, x_{i+1}]$ for $0 \leq i \leq N$. Thus by Theorems (2.7.2) and (2.7.7) we have

$$\left|\sum_{i=0}^{N}\sum_{r=0}^{2m}\left(\sum_{j=1}^{K} w_{ij}\left(D^r(Ag-E_ig)(z_{ij})\right)^2 - \int_{x_i}^{x_{i+1}}\left(D^r(Ag-E_ig)\right)^2\right)\right|$$

$$\leq Ch^{2K}\sum_{i=0}^{N}\sum_{r=0}^{2m}\int_{x_i}^{x_{i+1}}\left|D^{2K}\left(D^r(Ag-E_ig)\right)^2\right|$$

$$\leq Ch^{2K}\sum_{i=0}^{N}\sum_{r=0}^{2K+2m}\int_{x_i}^{x_{i+1}}\left(D^r(Ag-E_ig)\right)^2$$

$$= Ch^{2K}\sum_{i=0}^{N}\sum_{r=0}^{2n-1}\int_{x_i}^{x_{i+1}}\left(D^r(Ag-E_ig)\right)^2$$

$$\leq Ch^{2K-2K+2}\sum_{i=0}^{N}\sum_{r=0}^{2m}\int_{x_i}^{x_{i+1}}\left(D^r(Ag-E_ig)\right)^2 .$$

Thus from the above we have

$$\text{(2.8.8)}\qquad \sum_{i=0}^{N}\sum_{r=0}^{2m}\sum_{j=1}^{K} w_{ij}\left(D^r(Ag-E_ig)(z_{ij})\right)^2$$

$$\leq (1 + Ch^2)\left(\sum_{i=0}^{N}\sum_{r=0}^{2m}\int_{x_i}^{x_{i+1}}\left(D^r(Ag-E_ig)\right)^2\right) .$$

Now define $Eg(x)$ as follows :

$$Eg(x) = \left\{ E_ig(x) \text{ if } x_i < x < x_{i+1} \text{ for some } i \text{ or zero otherwise} \right\} .$$

Thus by (2.8.6) we know $||g - Eg||_{2m,\Delta} \leq Ch^{2n-2m}||g||_{2n}$, by (2.8.7) we know $||g - Eg||'_{2m,\Delta} \leq Ch^{2n-2m}||g||_{2n}$, by (2.8.8) we know $||Ag - Eg||'_{2m,\Delta} \leq C||Ag - Eg||_{2m,\Delta}$ and from the beginning of the proof we

know $\quad ||Ag - A'g||_{2m} \leq C||Ag - g||'_{2m,\Delta}$.

Hence by Theorem (2.2.6) and the above we have

$$||g - A'g||_{2m} \leq ||g - Ag||_{2m} + ||Ag - A'g||_{2m}$$

$$\leq Ch^{2n-2m}||g||_{2n} + C||Ag - g||'_{2m,\Delta}$$

$$\leq Ch^{2n-2m}||g||_{2n} + C||Ag - Eg||'_{2m,\Delta} + C||Eg - g||'_{2m,\Delta}$$

$$\leq Ch^{2n-2m}||g||_{2n} + C||Ag - Eg||_{2m,\Delta}$$

$$\leq Ch^{2n-2m}||g||_{2n} + C||Ag - g||_{2m} + C||g - Eg||_{2m,\Delta}$$

$$\leq Ch^{2n-2m}||g||_{2n} \; . \qquad ///$$

From Theorem (2.8.4) we can prove :

<u>Corollary</u> (2.8.5) : *Considering* V^{2n} *as a subspace of* W^{2n} *and* S_o *as a subspace of* W^{2m} *for* $K = 2n-2m$ *and* h *sufficiently small, we have that* $A' \in \mathcal{B}(V^{2n}, S_o)$.

<u>Proof</u> : Let $g \in V^{2n}$. Then by Theorem (2.8.4) we have

$$||A'g||_{2m} \leq ||g||_{2m} + ||g - A'g||_{2m}$$

$$\leq ||g||_{2n} + Ch^{2n-2m}||g||_{2n} \leq C||g||_{2n} \; .$$

Hence $A' \in \mathcal{B}(V^{2n}, S_o)$. ///

Now from our proven results we see that $A'u$, Au and Iu all provide the same order of accuracy as approximations to u when measured in the $||\cdot||_{2m}$ norm .

We note that the system (2.8.3) which must be solved to find $A'u$ is just equation (0.1.4) (or (0.1.5)) as discussed before in Chapter 0 . We also recall that for h sufficiently small, the smallness of h only depending on the problem's parameters, the system (2.8.3) is a positive definite one. Thus we only require knowledge of L and f to calculate $A'u$ in practice. Since (2.8.3) involves no integration but rather finite linear combinations of function values, we call the calculation of $A'u$ the <u>Discrete</u> least squares method, as noted before.

2.9 Further Error Bounds for the Discrete Method

We now want to get estimates on the quantity $||u - A'u||_0$ similar to estimates obtained for the Continuous method in Theorem (2.3.1). However several difficulties are involved. The trick used for Theorem (2.3.1) relies on the fact that $b(u-Au, e) = 0$ for $e \in S_o$. However by (2.8.3) we only know that $b'(u-A'u, e) = 0$ for $e \in S_o$ where $b'(\cdot, \cdot)$ involves only <u>approximate</u> integration. <u>However we can obtain our desired estimates if we assume more continuity of</u> u .

Thus we will now assume $f \in W^{2(2n-2m)}$ for the remainder of this section so that $u \in V^{4n-2m}$ and $||u||_{4n-2m} \leq C||f||_{4n-4m}$ by Theorem (1.3.4).

Now we state the main result. We note that the phrase "for sufficiently small h" means that h must be small enough for Theorem (2.8.4) to be used.

<u>Theorem</u> (2.9.1) : *Given that* $K = 2n-2m$ *and* $f \in W^{4n-4m}$ *then for sufficiently small* h *we have that*

$$||u - A'u||_0 \leq Ch^{\min(4n-4m,2n)}||f||_{4n-4m} .$$

<u>Proof</u> : We proceed as in the proof of Theorem (2.3.1). Suppose $||u - A'u||_0 \neq 0$. Let v be the unique function in V^{2m} that satisfies $Lv = (u - A'u)/||u - A'u||_0$ $(\in W^0)$ whose existence is guaranteed by Theorem (1.3.4). Note that we also have $||v||_{2m} \leq C||Lv||_0 = C$.

Since $v \in V^{2m}$ and $u - A'u \in V^{2m}$ we have by integration by parts as before :

$$||u - A'u||_0 = \int_a^b L(u - A'u)v .$$

By (2.8.3) we know that $b'(u-A'u, e) = 0$ for $e \in S_o$. Hence by Theorem (1.3.3) we have

$$(2.9.2) \qquad ||u - A'u||_0 = \int_a^b L(u - A'u)(v-Le) + (b-b')(u-A'u, e)$$

$$\leq C||u - A'u||_{2m}||v - Le||_0 + |(b-b')(u-A'u, e)|$$

for any $e \in S_o$.

Now we can use Lemmas (2.3.2) and (2.3.5) again to pick a

particular e . We consider two cases.

Say $2m \geq 2n-2m$. By Lemma (2.3.5) we can find a $v' \in V^{4m}$ so that $Lv' = v$, $||v'||_{4m} \leq C||v||_{2m} \leq C$ and

$$||v - L(Iv')||_0 \leq Ch^{2n-2m}||v||_{2m} \leq Ch^{2n-2m} .$$

Now $u - A'u$, $Iv' \in W^{2m}$ and $u - A'u$, $Iv' \in W^{2K+2m}(x_i, x_{i+1})$ for $0 \leq i \leq N$. Thus by Theorems (2.7.4) and (2.7.7) we have

$$\begin{aligned} &|(b-b')(u-A'u, Iv')| \\ &\quad \leq Ch^{2K}||u - A'u||_{2K+2m,\Delta} \, ||Iv'||_{2K+2m,\Delta} \\ &\quad = Ch^{2K}||u - A'u||_{2K+2m,\Delta} \, ||Iv'||_{2n-1,\Delta} \end{aligned}$$

since Iv' has degree $2n-1$ on each subinterval. Noting $4m \geq 2n$, we also have $||v'||_{2n-1,\Delta} \leq ||v'||_{4m,\Delta} \leq C$. Hence by Theorem (1.2.7) we see that

$$||v' - Iv'||_{2n-1,\Delta} \leq Ch||v'||_{2n,\Delta} \leq Ch .$$

Thus we have from the above :

$$\begin{aligned} ||Iv'||_{2n-1,\Delta} &\leq ||v' - Iv'||_{2n-1,\Delta} + ||v'||_{2n-1,\Delta} \\ &\leq Ch + C \leq C . \end{aligned}$$

Thus using Theorem (2.8.4) and choosing $e = Iv'$ in (2.9.2) we have that

(2.9.3) $$||u - A'u||_0$$

$$\leq C||u - A'u||_{2m}||v - L(Iv')||_0 + |(b-b')(u-A'u, Iv')|$$

$$\leq Ch^{2n-2m}||f||_{2n-2m}Ch^{2n-2m} + Ch^{4n-4m}||u - A'u||_{2K+2m,\Delta}$$

$$\leq Ch^{4n-4m}(||f||_{4n-4m} + ||u - A'u||_{4n-2m,\Delta}) \quad .$$

Now we consider the case when $2m < 2n-2m$. By Lemma (2.3.2) we can find an $e \in S_o$ so that

$$||v - Le||_0 \leq Ch^{2m}||v||_{2m} \leq Ch^{2m} .$$

By Theorem (2.7.4) again, we have

$$|(b-b')(u-A'u, e)| \leq Ch^{2K}||u - A'u||_{4n-2m,\Delta}||e||_{4n-2m,\Delta}$$

$$= Ch^{4n-4m}||u - A'u||_{4n-2m,\Delta}||e||_{2n-1,\Delta}$$

$$\leq Ch^{4n-4m-2n+1+4m}||u - A'u||_{4n-2m,\Delta}||e||_{4m,\Delta}$$

$$= Ch^{2n+1}||u - A'u||_{4n-2m,\Delta}||e||_{4m,\Delta}$$

where we have used Theorem (2.7.7) noting $4m < 2n$.

Let v' again be the unique function in V^{4m} that satisfies $Lv' = v$ and $||v'||_{4m} \leq C||v||_{2m} \leq C$. Now apply Theorem (2.5.1) to this situation with $M = 4m$, $Q = 2m$, $c_j = z_{ij}$ for $1 \leq j \leq 2m$ (we note $2m < K$) $c = x_i$ and $d = x_{i+1}$ to obtain a polynomial $E_i v'$ of degree $4m-1$ on

(x_i, x_{i+1}) that satisfies

$$\sum_{s=0}^{r} \int_{x_i}^{x_{i+1}} \left(D^s(v' - E_i v')\right)^2 \leq CGh^{2(4m-r)} \sum_{s=0}^{r} \int_{x_i}^{x_{i+1}} (D^s v')^2$$

for $0 \leq r \leq 4m$. As noted in the proof of Theorem (2.8.4), G depends only on the Gauss points in $[-1, 1]$ and hence we can denote G by C . We remark that the interpolation properties of $E_i v'$ will not be essential . Only the degree of $E_i v'$ and the above integral inequality will be used. Continuing, we define

$$Ev'(x) = \left\{ E_i v'(x) \text{ if } x_i < x < x_{i+1} \text{ for some } i \text{ or zero otherwise} \right\} .$$

Thus for $0 \leq r \leq 4m$ we have

$$||v' - Ev'||_{r,\Delta} \leq Ch^{4m-r}||v'||_{4m} \leq Ch^{4m-r} . \tag{2.9.4}$$

Hence using the above we have

$$\begin{aligned}
||e||_{4m,\Delta} &\leq ||e - v'||_{4m,\Delta} + ||v'||_{4m} \\
&\leq ||e - Ev'||_{4m,\Delta} + ||Ev' - v'||_{4m,\Delta} + C \\
&\leq ||e - Ev'||_{4m,\Delta} + C && \text{(by (2.9.4))} \\
&\leq Ch^{-2m}||e - Ev'||_{2m,\Delta} + C && \text{(by Theorem (2.7.7))} \\
&\leq Ch^{-2m}(||e - v'||_{2m} + ||v' - Ev'||_{2m,\Delta}) + C \\
&\leq Ch^{-2m}(||L(e - v')||_0 + Ch^{2m}) + C
\end{aligned}$$

(by Theorem (1.3.3) and (2.9.4))

$$= Ch^{-2m}(||Le - v||_0 + Ch^{2m}) + C \leq C .$$

Thus using (2.9.2) with the particular $e \in S_o$ guaranteed by Lemma (2.3.2), by Theorem (2.8.4) we have that

(2.9.5)
$$||u - A'u||_0$$
$$\leq C||u - A'u||_{2m}||v - Le||_0 + |(b-b')(u-A'u, e)|$$
$$\leq Ch^{2n-2m}||f||_{2n-2m}Ch^{2m} + Ch^{2n+1}||u - A'u||_{4n-2m,\Delta}\, C$$
$$\leq Ch^{2n}(||f||_{4n-4m} + ||u - A'u||_{4n-2m,\Delta}) .$$

Now since $u \in W^{4n-2m}$ we know $Iu \in S_o$ exists. Thus using Theorems (1.2.7), (2.7.7) and (2.8.4) we have that

(2.9.6)
$$||u - A'u||_{4n-2m,\Delta} \leq ||u||_{4n-2m} + ||A'u||_{2n-1,\Delta}$$
$$\leq ||A'u - Iu||_{2n-1,\Delta} + ||Iu - u||_{2n-1,\Delta} + ||u||_{2n-1} + ||u||_{4n-2m}$$
$$\leq ||A'u - Iu||_{2n-1,\Delta} + C||u||_{4n-2m}$$
$$\leq Ch^{2m-2n+1}||A'u - Iu||_{2m} + C||f||_{4n-4m}$$
$$\leq Ch^{2m-2n+1}(||A'u - u||_{2m} + ||u - Iu||_{2m}) + C||f||_{4n-4m}$$
$$\leq Ch^{2m-2n+1}(Ch^{2n-2m}||f||_{2n-2m} + Ch^{2n-2m}||u||_{2n}) + C||f||_{4n-4m}$$
$$\leq C||f||_{4n-4m} .$$

Thus by equations (2.9.3), (2.9.5) and (2.9.6) we have the desired result. ///

2.10 Some Extensions

We now summarize and slightly extend the results of Theorems (2.3.1) and (2.9.1).

Theorem (2.10.1) : (1) *Suppose that* $f \in W^{2n-2m}$. *Then*

$$||D^r(u - Au)||_{0,\Delta} \leq Ch^{\min(2n,4n-4m)-r}||f||_{2n-2m} \quad \textit{for} \quad 0 \leq r < 2m$$

$$||D^r(u - Au)||_{0,\Delta} \leq Ch^{2n-r}||f||_{2n-2m} \quad \textit{for} \quad 2m \leq r \leq 2n .$$

(2) *Suppose that* $f \in W^{2n-2m}$, $K = 2n-2m$ *and* h *is sufficiently small. Then*

$$||D^r(u - A'u)||_{0,\Delta} \leq Ch^{2n-r}||f||_{2n-2m} \quad \textit{for} \quad 2m \leq r \leq 2n .$$

(3) *Suppose that* $f \in W^{4n-4m}$, $K = 2n-2m$ *and* h *is sufficiently small. Then*

$$||D^r(u - A'u)||_{0,\Delta} \leq Ch^{\min(2n,4n-4m)-r}||f||_{4n-4m} \quad \textit{for} \quad 0 \leq r < 2m.$$

(4) *We note that in* (1), (2), (3) *we may replace* $||\cdot||_{0,\Delta}$ *by* $||\cdot||_0$ *if* $0 \leq r \leq z+1$.

Here "h sufficiently small" means that h must be small enough for Theorem (2.8.4) to be used. We note that these results reflect the

interpolate results of Theorem (1.2.7); that is, one "pays" for each derivative by a loss in a power of h. Also, for large enough n, $\min(2n,4n-4m) = 2n$ so that Theorem (1.2.7) is even more closely reproduced. Thus we get results for Au and $A'u$ that are very much like the interpolate results.

<u>Proof of Theorem</u> (2.10.1) :

We start with (1). Noting that $Iu \in S_o$ is defined we have for $0 \leq r \leq 2m$ that

$$(2.10.2) \qquad \begin{aligned} ||D^r(u - Au)||_{0,\Delta} &\leq ||u - Au||_{r,\Delta} \\ &\leq ||u - Iu||_{r,\Delta} + ||Iu - Au||_{r,\Delta} \\ &\leq Ch^{2n-r}||u||_{2n} + Ch^{-r}(||Iu - Au||_0) \\ &\leq Ch^{2n-r}||f||_{2n-2m} + Ch^{-r}(||u - Iu||_0 + ||u - Au||_0) \\ &\leq Ch^{2n-r}||f||_{2n-2m} \\ &\qquad + Ch^{-r}\left(Ch^{2n}\ ||u||_{2n} + Ch^{\min(2n,4n-4m)}||f||_{2n-2m}\right) \\ &\leq Ch^{\min(2n,4n-4m)-r}||f||_{2n-2m} \end{aligned}$$

where we have used Theorems (1.2.7), (2.7.7) and (2.3.1) . For $2m \leq r \leq 2n-1$ we have

$$(2.10.3) \qquad ||D^r(u - Au)||_{0,\Delta} \leq ||u - Au||_{r,\Delta}$$

$$\leq ||u - Iu||_{r,\Delta} + ||Iu - Au||_{r,\Delta}$$

$$\leq Ch^{2n-r}||u||_{2n} + Ch^{2m-r}(||Iu - Au||_{2m,\Delta})$$

$$\leq Ch^{2n-r}||f||_{2n-2m} + Ch^{2m-r}(||Iu - u||_{2m,\Delta} + ||u - Au||_{2m,\Delta})$$

$$\leq Ch^{2n-r}||f||_{2n-2m} + Ch^{2m-r}(Ch^{2n-2m}||u||_{2n} + Ch^{2n-2m}||f||_{2n-2m})$$

$$\leq Ch^{2n-r}||f||_{2n-2m}$$

where we have used Theorems (1.2.7), (2.7.7) and (2.2.6). Now since the case $r = 2n$ is trivial as Au has only degree $2n-1$ we have shown (1).

Case (2) is proved by an argument similar to (2.10.3) only we use Theorem (2.8.4) instead of Theorem (2.2.6).

Case (3) is proved by an argument similar to (2.10.2) only we use Theorem (2.9.1) instead of Theorem (2.3.1).

Since case (4) is clear due to the continuity of functions in S_o, we are done. ///

Chapter 3

Practical Considerations

3.1 Properties of L

We listed the properties required of the operator L in Section 1.3. We now wish to indicate which operators have these properties.

We know that $a'_{rs} \in C^{\max(4n-4m,2m)+r}[a, b]$ for $1 \leq r, s \leq m$ already. Now we suppose that :

(1) $J(v, v) \geq C||v||_m^2$ for $v \in C_o^m(a, b)$.

(2) $a'_{mm} \geq \delta > 0$ for $x \in [a, b]$

where $J(v_1, v_2) = \sum_{r,s=0}^{m} \int_a^b a'_{rs} D^s v_2 D^r v_1$ is a bilinear form on $W^m \times W^m$ and

$$C_o^m(a, b) = \left\{ v \in C^m(a, b) : \text{supp } v \quad (a, b) \right\}.$$

Note that for $v \in C_o^{2m}(a, b)$ we have

$$C||v||_m^2 \leq J(v, v) = \sum_{r,s=0}^{m} \int_a^b a'_{rs} D^r v D^s v$$

$$= \sum_{r,s=0}^{m} \int_a^b (-1)^r D^r (a'_{rs} D^s v) v = \int_a^b Lv \cdot v$$

$$\leq ||Lv||_0 ||v||_0 \leq ||Lv||_0 ||v||_m$$

where we have used integration by parts. Thus we have

$$C||v||_m \leq ||Lv||_0 \tag{3.1.1}$$

for $v \in C_o^{2m}(a, b)$. Now since (3.1.1) holds for $v \in C_o^{\infty}(a, b) \subset C_o^{2m}(a, b)$ we have that (3.1.1) holds for $v \in H_o^m$. Note that by Theorem (1.3.1) we also know

$$||Lv||_0 \leq C||v||_{2m}$$

for $v \in W^{2m}$.

Now by Agmon [2; p. 102, p. 129] we have that given conditions (1) and (2) above and $g \in W^p$ for $0 \leq p \leq \max(4n-4m, 2m)$ there is an unique $v \in H_o^m \cap W^{2m+p}$ so that $Lv = g$ in W^p and

$$\begin{aligned} ||v||_{p+2m} &\leq C(||g||_p + ||v||_0) \\ &\leq C(||g||_p + ||v||_m) \\ &\leq C||g||_p = C||Lv||_p . \end{aligned}$$

Thus under (1) and (2) above we have the necessary conditions stated for L in Section 1.3.

Note that if L is of the form stated in (0.1.1) then (1) above is clearly satisfied if a_r' is sufficiently smooth and $a_r' \geq \delta > 0$ for $1 \leq r \leq m$ on $[a, b]$. Thus if L is of this form then the totality of restrictions :

(3.1.2) (1) $a'_r \in C^{\max(4n-4m,2m)+r}[a, b]$ for $0 \leq r \leq m$

(2) $a'_r \geq \delta > 0$ for $0 \leq r \leq m$ on $[a, b]$

are sufficient for Section 1.3.

3.2 A Comparison

We wish to briefly compare Theorem (2.9.1) with the following result on the usual Collocation method from, for example, Russell and Varah [10].

Theorem (3.2.1) : *Say* L *is of the form given by equation* (0.1.1), L *satisfies* (2) *of* (3.1.2), $z = 2m-1$ *and the coefficients of* L *and* f *are sufficiently smooth. Then we have the following result for the Collocation approximation* w :

$$||u - w||_0 \leq Ch^{\min(4n-4m,2n)}$$

where h *is sufficiently small* .

Now since we know from the preceding section that L satisfies the necessary properties needed for Section 1.3 if L is of the form given in (0.1.1) and satisfies (3.1.2), we have by Theorem (2.9.1) :

$$||u - A'u||_0 \leq Ch^{\min(4n-4m,2n)}$$

for h sufficiently small and f sufficiently smooth.

Thus we can see that the Collocation method and the Discrete method both generate approximations in S_o that have the same convergence properties. Thus both methods are feasible for practical computations, with the limitations outlined in Chapter 0 .

3.3 Discrete Least Squares Computations

By the use of numerical computations we now intend to examine Theorem (2.9.1) for the Discrete method. However before we state the examples we will make some preliminary remarks. (The results of this section were obtained by the author on the IBM 360/67 and the later IBM 370/168 computers at U.B.C.).

All calculations in the following examples utilized double precision arithmetic to minimize roundoff errors. The choices made for the various parameters in the examples are :

(1) $h = h' = 1/(N+1)$ for integer N . (Thus the meshes are uniform and $\alpha = 1$ is a mesh bound).

(2) $[a, b] = [0, 1]$.

(3) $n = 2$, $z = 2$. (Thus the space S is the space of cubic splines on $[0, 1]$).

(4) $m = 1$. (Thus $2n-2m = 2 > 0$).

The basis $\{e_j\}$ used for S_o is the B-spline basis on a uniform mesh

described in Schultz [12; p. 73] or de Boor [7]. For reasons noted before $K = 2n-2m = 2$ Gauss points in (x_i, x_{i+1}) (for $0 \leq i \leq N$) were used where $w_{ij} = (x_{i+1} - x_i)/2 = h/2$ (for $0 \leq i \leq N$, $1 \leq j \leq K$), are the associated weights. Note that $z = 2n-2 = 2 \neq 1 = 2m-1$. Thus the Discrete method gives an approximation with one more degree of continuity than that available from the Collocation method.

In each of the examples an estimate of

$$||u - A'u||_\infty = \max\{|u - A'u|(x) : 0 \leq x \leq 1\}$$

was made as follows. An initial set of calculations was made of $|u - A'u|(x)$ at $\geq N$ non-mesh points in $[0, 1]$. The maximum of these numbers was taken and a further calculation of $|u - A'u|(x)$ was made at twenty additional points in a neighborhood of where the largest error was observed. The maximum over all these values was then called $||u - A'u||'_\infty$.

Since the number of points checked in the initial calculations differed little from N in many cases, $||u - A'u||'_\infty$ is only an estimate of $||u - A'u||_\infty$ and no doubt some anomalies in α calculations (following) are due to this fact.

Now the following asymptotic result is expected :

$$||u - A'u||_\infty = Ch^\beta .$$

Thus given two values e_1 and e_2 of $||u - A'u||_\infty$ calculated at h_1 and h_2, we can see that $\beta = \log(e_1/e_2)/\log(h_1/h_2)$. With this in mind we

define α to be as in the above formula for β where we use $||\cdot||'_\infty$ values instead of $||\cdot||_\infty$ values. Hence we have a way to estimate β .

Now we can state the results of our calculations. Consider the following operators and solution functions :

(3.3.1) $$Lu = -D\big((1/\pi^2)Du\big) + e^x u = f \; ; \qquad u = \sin(\pi x)$$

(3.3.2) $$Lu = -D\big((e^{10x} - s)Du\big) + 100su = f \; ; \qquad u = (x^2-x)e^{10x}$$

where $s = .999999999$

(3.3.3) $$Lu = -D(1\cdot Du) + 4u = f \; ; \qquad u = \cosh(2x-1) - \cosh 1 \; .$$

In each of the above equations we can see that L and the associated f satisfy the assumptions of Section 1.3 (by the conditions (3.1.2)) and in fact they also satisfy the assumptions of Theorem (2.9.1). Thus we expect

$$||u - A'u||_0 \leq Ch^4 \; .$$

Referring to Table I we can see the results of the α calculations imply

$$||u - A'u||_0 \leq ||u - A'u||_\infty \approx ||u - A'u||'_\infty \lesssim Ch^4$$

in all cases. We also note the higher rate of convergence observed in the case of (3.3.2) in spite of the almost singular coefficient in L and the growth of u . This heightened convergence rate will probably approach 4.0 as the error $||u - A'u||'_\infty$ becomes fractionally small.

We now consider

$$(3.3.4) \qquad Lu = -D(e^x Du) + e^x u = f \ ;$$

$$u = \begin{cases} p(x) - \frac{29}{3}(x - \frac{1}{2})^4 & \text{for} \quad 0 \le x \le \frac{1}{2} \\ p(x) - \frac{79}{3}(x - \frac{1}{2})^4 & \text{for} \quad \frac{1}{2} < x \le 1 \end{cases}$$

where $p(x) = \frac{1}{6}x^3 + \frac{1}{4}x^2 + \frac{5}{8}x + \frac{29}{48}$. Now L and f satisfy the assumptions of Section 1.3 (again by the conditions (3.1.2)). Also we note that $f \in C^1[0, 1]$ and $D^2 f$ has a jump discontinuity at $x = \frac{1}{2}$. Thus by Theorem (1.1.6) we have $f \in W^{2n-2m} = W^2$ but $f \notin W^{2(2n-2m)} = W^4$. Hence the hypotheses of Theorem (2.9.1) do not apply so we <u>cannot</u> guarantee $||u - A'u||_0 \le Ch^4$. However by Theorem (2.3.1) we do have $||u - Au||_0 \le Ch^4$.

Referring to Table II we see that

$$||u - A'u||_0 \le ||u - A'u||_\infty \approx ||u - A'u||'_\infty \lesssim Ch^4 .$$

Thus the hypotheses of Theorem (2.9.1) would appear to be too strong. We suspect $f \in W^{2n-2m}$ should be sufficient for Theorem (2.9.1).

The Discrete least squares method was also tried on the following equations which do not satisfy the conditions discussed in Section 3.1 :

$$(3.3.5) \qquad Lu = (1/\pi)^2 D^2 u + (1/\pi)Du + u = f \ ; \qquad u = \sin(\pi x)$$

$$(3.3.6) \qquad Lu = D^2 u + Du = f \ ; \qquad u = (x^2 - x)e^x .$$

In both cases $\alpha = 4$ convergence was obtained using the Gauss points, as seen in Table II. Thus the method appears to be stronger than our initial analysis would indicate.

Other quadrature approximations were tried using (3.3.5) and (3.3.6) but no increase in the convergence rate was discovered.

Thus by these calculations we can see that the Discrete least squares approximation method can be used for practical problems.

Table I

Results for Equations (3.3.1) - (3.3.3)

	(3.3.1)		(3.3.2)		(3.3.3)	
N	$\|\|u - A'u\|\|'_\infty$	α	$\|\|u - A'u\|\|'_\infty$	α	$\|\|u - A'u\|\|'_\infty$	α
5	.13E - 3		.78E + 4		.46E - 4	
		4.0		3.0		3.9
10	.11E - 4		.13E + 4		.42E - 5	
		4.0		4.9		4.0
20	.82E - 6		.54E + 2		.32E - 6	
		4.0		5.5		4.2
30	.17E - 6		.64E + 1		.63E - 7	
		4.0		5.7		4.2
40	.56E - 7		.13E + 1		.20E - 7	
		4.0		5.8		3.8
50	.23E - 7		.36E + 0		.87E - 8	
		4.0		5.9		6.3
60	.11E - 7		.13E + 0		.28E - 8	
		4.0		6.0		1.3
70	.62E - 8		.51E - 1		.23E - 8	
						3.7
80	—		—		.14E - 8	
						3.9
90	—		—		.90E - 9	

Table II

Results for Equations (3.3.4) - (3.3.6)

	(3.3.4)		(3.3.5)		(3.3.6)	
N	$\|\|u - A'u\|\|'_\infty$	α	$\|\|u - A'u\|\|'_\infty$	α	$\|\|u - A'u\|\|'_\infty$	α
5	.19E - 2		.13E - 3		.83E - 4	
		4.9		3.7		4.0
10	.11E - 3		.14E - 4		.76E - 5	
		3.9		3.8		4.0
20	.83E - 5		.12E - 5		.57E - 6	
		4.0		4.1		4.1
30	.17E - 5		.25E - 6		.12E - 6	
		4.0		3.8		5.4
40	.55E - 6		.88E - 7		.26E - 7	
		3.8				
50	.24E - 6		—		—	
		4.7				
60	.11E - 6		—		—	
		3.2				
70	.63E - 7		—		—	
		3.9				
80	.38E - 7		—		—	
		4.0				
90	.24E - 7		—		—	

Bibliography

1. Adams, R.A., Sobolev Spaces, Academic Press, New York, (to appear - 1975).

2. Agmon, Shmuel, Lectures on Elliptic Boundary Value Problems, D. Van Nostrand Co., New York, 1965.

3. Bellman, Richard, A Note on an Inequality of E. Schmidt, Bull. Amer. Math. Soc., V50, 1944, pp 734-736.

4. Berezanskii, Ju. M., Expansions in Eigenfunctions of Self Adjoint Operators, American Math. Society, Providence, Rhode Island, 1968.

5. J.H. Bramble and A.H. Schatz, Least Squares Methods for 2m-th Order Elliptic Boundary-Value Problems, Math. Comp., V25, 1970, pp 1-33.

6. P.G. Ciarlet and P.A. Raviart, General Lagrange and Hermite Interpolation in R^n with Applications to Finite Element Methods, Arch. Rational Mech. Anal., V46, 1972, pp 177-199

7. De Boor, C., On Calculating with B-splines, J. Approx. Theory, V6, 1972, pp 50-62.

8. E. Hewitt and K. Stromberg, Real and Abstract Analysis, Springer-Verlag, New York, 1969.

9. E. Isaacson and H.B. Keller, Analysis of Numerical Methods, John Wiley and Sons, Inc., New York, 1966.

10. R.D. Russell and J.M. Varah, A Comparison of Global Methods for Linear Two-Point Boundary Value Problems, Dept. of Computer Sc., U.B.C., Technical Report, 1974.

11. Schultz, M.H., Error Bounds for Polynomial Spline Interpolation, Math. Comp., V24, 1970, pp 507-515.

12. Schultz, M.H., Spline Analysis, Prentice-Hall, Inc., New Jersey, 1973.