ON SOLVING QUASILINEAR ELLIPTIC EQUATIONS BY VARIATIONAL METHODS, USING PIECEWISE POLYNOMIAL TRIAL FUNCTIONS.

ABSTRACT

The local properties of piecewise polynomial functions defined on bounded regions of Euclidean spaces are studied in considerable generality. Aspects of the resulting theory are used to discretise a quasilinear elliptic equation, and various methods for solving the resulting nonlinear algebraic equations are investigated.

Preface

This thesis is concerned with the construction of approximate solutions to a certain nonlinear elliptic equation, which arises frequently in practice and which includes as special cases many commonly occuring linear elliptic equations. The construction of the solutions has two clearly discernable aspects, namely, reduction to an algebraic problem and solution of the algebraic problem. The first problem is solved here by the use of the Ritz method in conjunction with piecewise polynomial trial functions. The second problem is approached through the use of iterative methods.

Part 1 of the thesis contains some introductory material about elliptic equations and variational methods. It also contains the seeds of an approach to a third problem, not mentioned above, that of assessing the quality of an approximate solution. But generally speaking this latter problem has not been considered in any depth, and still awaits a satisfactory treatment.

In order to place the construction and analysis of piecewise polynomial trial functions on a rigourous footing it was necessary to develop a multidimensional generalisation of equal interval Lagrange interpolation. This theory, which is in a reasonable state of completion is perfectly general in that it holds for spaces of any finite dimension and for polynomials of arbitrary degree. Its application to solving linear elliptic equations by finite element methods has yet to be fully exploited.

The theory is developed in Part 2.

Part 3 contains a derivation of the nonlinear algebraic problem corresponding to the differential equation. This Part also contains some miscellaneous algebraic details concerning the algebraic system and also existence and uniqueness proofs for the algebraic problems.

Part 4 contains some general results to do with iterative methods for solving the equations. A number of iterative methods in use are studied theoretically against the background of a simple general theory set out at the beginning. The results in Part 4 are perhaps not as incisive as one would like: but this seems unavoidable as it is caused by the very generality of the problem itself. More precise results seem possible only when specfic problems are dealt with.

Finally, an appendix contains the briefest account of some computations carried out by the author. As is pointed out, the computations were not intended to yield serious experimental results, but simply to obtain some practice with the various algorithms.

The author is properly aware of the limitations of the thesis. However, it is felt that a start has been made on the solution of some important problems.

Acknowledgments

I would like to thank Dr. M.J.M.Bernal, who was my supervisor for the degree, for much advice and encouragment and especially for his critical reading of the manuscript which resulted in the detection of several errors.

I have also to acknowledge a very great debt that I owe to Dr. R.W.Thatcher. If I have learned anything about how to do mathematics it is in no small measure due to this person.

CONTENTS

Part 1: Ellipticity and variational methods

1. Introduction
2. Ellipticity
3. Maximum principles
4. The equation $div[\gamma(|\nabla u|^2)\nabla u] = f(x)$
5. Uniqueness results
6. Direct methods in the calculus of variations
7. Conditions for a sequence to be minimizing
8. On the variational formulation of boundary value problems
9. Statement of the variational problem
10. Semicontinuity of $J(u)$
11. Resume of Part 1

Part 2: Piecewise polynomial trial functions p. 40

1. Scope and purpose
2. Simplexes and polynomials
3. Interpolation
4. Representation results
5. Examples
6. Triangulations and continuous trial functions
7. Error of interpolation
8. Error of derivative interpolation
9. Additional remarks on the interpolation theory
10. Denseness of trial functions
11. Resume and notes

Part 3: Derivation and properties of the algebraic system

1. Introduction
2. General form of the equations

3. Existence and uniqueness of solutions

4. A special case of the equations

5. Resume and notes

Part 4: Solution of the algebraic equations

1. Introduction and objectives

2. Iterative methods

3. Derivative free algorithms

4. Choosing

5. Derivative linked algorithms

6. Convergence of SOR

7. Resume and notes

Appendix

On computation

References

1.1 Introduction

The present section contains an account of certain theoretical aspects of quasilinear elliptic equations inasmuch as these are pertinent to the problem of computing approximate solutions to them. The theory of the existence of solutions to the general quasilinear equation receives no attention here: however, for completeness it will be mentioned that this theory is now in a well advanced state, to which it has been brought in the past 70 years - and mostly in the past 35 years, by, amongst others, S.N. Bernstein, J. Leray, J. Schauder, O.A. Ladyzhenskaya and N.N. Uraltseva [1] , and J. Serrin. A comprehensive paper [2] , of the last author contains an up to date account of many of the results. (These remarks in no way pretend to a balanced account of the development of the existence theory: for example, a great deal of work was done by other people on specific equations in the class, (e.g. [3]) and on related, but more general problems (e.g. [4])).

In the next paragraph we shall collect together some basic definitions about elliptic operators and related topics. It is well known that the eigenvalue spectrum of the matrix constructed from the coefficients of a 2nd order partial differential equation plays a decisive role in determining the allowable characteristics of solutions to the equation (an example of such a characteristic would be whether the solution could have the character of a wave motion). The precise type of eigenvalue spectrum we shall allow is mentioned below: one of its very important consequences is the maximum principle (Theorem 3.1.), from which will follow a uniqueness theorem of use to us.

1.2 Ellipticity

The most general quasilinear elliptic equation that will be considered is

$$\sum_{i,j=1}^{n} a_{ij}(x, u, u_x)\, u_{x_i x_j} + a(x, u, u_x) = 0 \qquad (2,1)$$

where the symbols have the following meanings: x is a point of R_n, $u \equiv u(x)$, $u_x = \nabla u \equiv (u_{x_1}, \dots, u_{x_n})^T \equiv \left(\frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_n}\right)^T$. It will always be assumed that $a_{ij}(x) \equiv a_{ji}(x)$. Let us consider initially the case in which the coefficients a_{ij} are independent of u, u_x. In this case, if a was of the form

$$a = \sum_{i=1}^{n} a_i u_{x_i} + b(x)\cdot u - f(x)$$

the equation would be linear. If a has the more general form, but a_{ij} depend only on x the equation is called semilinear. The expression

$$\sum_{i,j=1}^{n} a_{ij}\, u_{x_i x_j}$$

is called the principal part of the equation.

Let Ω be a given bounded region of R_n where functions $a_{ij}(x)$ $a_i(x)$, $b(x)$, $f(x)$ are defined and measurable. The linear differential operator L defined in

$$Lu = \sum_{i,j=1}^{n} a_{ij}(x)\, u_{x_i x_j} \qquad (2,2)$$

is called <u>elliptic</u> if $\forall \quad x \in \Omega$ we have

$$\sum_{i,j=1}^{n} a_{ij}(x)\, \xi_i \xi_j > 0 \qquad (2,3)$$

$\forall \quad \xi = (\xi_1, \xi_2, \dots, \xi_n) \in R_n, \quad \xi \neq 0$. This is

equivalent to requiring that at each point $x \in \Omega$ the (symmetric) matrix $[a_{ij}(x)]$ is positive definite. But as $x \to x_0 \in \partial\Omega$ ($\partial\Omega$ = boundary of Ω) it may be that $[a_{ij}(x)]$ degenerates into a non-negative definite matrix. The case in which this does not happen is important, and is the subject of the next definition:

L is called uniformly elliptic if $\exists$ positive constants m, M such that

$$m \| \xi \|_2^2 \leq \sum_{i,j=1}^{n} a_{ij}(x) \xi_i \xi_j \leq M \| \xi \|_2^2 \qquad (2,4)$$

($\| \ \|_2$ is the Euclidean norm in R_n), $\forall\ x \in \Omega$. The right hand inequality will be satisfied if, e.g. the coefficients a_{ij} are continuous in $\bar{\Omega} = \Omega + \partial\Omega$.

In the general case (2,1), the definition of ellipticity must depend on the precise form of the function u. Apart from this, the definition follows the linear case: let Ω be a given region of R_n where functions $a_{ij}(x, u, u_x)$ are defined and measurable, u being a certain function defined, along with its derivatives in Ω. The operator $L_{(u)}$ defined in

$$L_{(u)} v = \sum_{i,j=1}^{n} a_{ij}(x, u, u_x) v_{x_i x_j}$$

is called elliptic (with respect to u), if, $\forall\ x \in \Omega$ we have

$$\sum_{i,j=1}^{n} a_{ij}(x, u, u_x) \xi_i \xi_j > 0$$

$\forall\ \xi \in R_n, \ \xi \neq 0$.

With regard to the uniform ellipticity of nonlinear operators,

there seems to be no universally accepted definition and different people use different, and not equivalent definitions. For our purposes, the following obvious extension of (2,4) will be sufficient:

$L_{(u)}$ is called uniformly elliptic (with respect to u) if, in addition to being elliptic, $\exists$ positive constants m, M such that

$$m \|\xi\|_2^2 \leq \sum_{i,j=1}^{n} a_{ij}(x, u, u_x)\, \xi_i \xi_j \leq M \|\xi\|_2^2 \qquad (2,4)'$$

$\forall \quad x \in \Omega$.

We may also talk about an <u>equation</u> being elliptic. This simply means that the operator constructed from the principal part of the equation is elliptic. Of most interest concerning non-linear equations is the question of their ellipticity (or non-ellipticity) with respect to a given solution: i.e. if u satisfies the equation

$$\sum_{i,j=1}^{n} a_{ij}(x, u, u_x)\, u_{x_i x_j} + a(x, u, u_x) = 0$$

can it be said that the equation is elliptic? Since, in general, the solutions are not known explicitly, this can be a difficult question to answer. However for many equations (including a large class of practical problems) an unequivocal answer (affirmative) can be given in advance of actually knowing a solution. Examples of this are given in section 1.4. On the other hand, it has to be pointed out that there are other classes of equations (e.g. the Monge-Ampere equation) which are elliptic for some solutions and which fail to be elliptic for other solutions. But there may be physical grounds for expecting an elliptic solution. Another example of this kind of difficulty occurs with the flow of compressible gases, where although (roughly speaking) the flow in a region is governed by a single equation, the equation may be elliptic only in certain parts of the region. Physically, this occurs when the

flow undergoes a change from being subsonic (corresponding to ellipticity) through what is known as the transonic region, to a supersonic region where the governing equation becomes hyperbolic. For more details of these problems, reference [5] could be consulted.

Concerning the region Ω in which a solution of (2,1) is sought, we shall never assume less than the following: $\partial\Omega$ can be covered with a finite number of spheres in such a way that the part of $\partial\Omega$ in the i^{th} sphere can be described by an equation $x_n = g_i(x_1, \cdots, x_{n-1})$ where g_i has Hölder continuous 2nd derivatives, with exponent $\alpha < 1$. This is simply an assumption which gives grounds for applying known existence theorems.

A general boundary value problem is the following: find a function $u \in C^{(2)}(\Omega) \cap C^{(0)}(\bar{\Omega})$ satisfying (2,1) in Ω and the condition

$$\alpha \frac{\partial u}{\partial n} + \beta u = \phi \qquad (2,5)$$

whenever $x \in \partial\Omega$. In (2,5) functions α, β are subjected to the single condition $|\alpha| + |\beta| > 0$. $\frac{\partial u}{\partial n}$ denotes differentiation of u in the direction of the outward normal to $\partial\Omega$. Actually, we shall be concerned mostly with the cases $\alpha = 1$, $\beta = 0$ on one part of the boundary (which may be vacuous), whilst $\alpha = 0$, $\beta = 1$ on the remainder (which may be vacuous). A special form of (2,1) to which our attention will be restricted, is introduced in section 1.4 We shall first obtain some results which hold for (2,2) generally.

1.3 Maximum principles

It follows from (2,1) that every solution of the quasilinear

equation satisfies a certain linear equation, namely that linear (elliptic) equation whose coefficients are obtained by the insertion of u in $a_{ij}(x, u, u_x)$ and in $a(x, u, u_x)$. Important results in the linear case are deduced from the 'maximum principle' and it follows from the above observation that this principle also has applications to quasilinear equations. We shall therefore prove the following result.

Theorem 3.1

Let Ω be given and let the operator L defined by (3,2) satisfy the ellipticity condition

$$\sum_{i,j=1}^{n} a_{ij}(x)\,\xi_i\,\xi_j \geq \gamma(x)\,\|\xi\|_2^2 \qquad \gamma(x) > 0, \quad x \in \Omega.$$

Assume that the coefficients a_{ij}, a, b satisfy the conditions that

$$\sum_{i=1}^{n} \frac{a_{ii}(x)}{\gamma(x)}, \quad \sum_{i=1}^{n} \frac{a_i(x)}{\gamma(x)}, \quad -\frac{b(x)}{\gamma(x)} \qquad (3,1)$$

are locally bounded in Ω and that the inequality

$$Lu \equiv \sum_{i,j=1}^{n} a_{ij}(x)\,u_{x_i x_j} + \sum_{i=1}^{n} a_i(x)\,u_{x_i} + b(x)u \geq 0 \quad (\leq 0) \qquad (3,2)$$

holds in Ω. If $b \leq 0$ and u attains a non-negative maximum M (non-positive minimum M') in Ω then $u \equiv M$, $(u \equiv M')$ in Ω.

Proof

Consider the case in which strict inequality holds in (3,2) and $b \leq 0$. By the ellipticity condition, we can always find a linear transformation of the x_i to y_i which is such that the principal part of the operator L transforms into the Laplace operator at a

given point $\bar{x} \in \Omega$. Suppose $\bar{x}$ is a non-negative relative maximum of u. Then

$$\frac{\partial u}{\partial x_i}(\bar{x}) = 0 \qquad \frac{\partial^2 u}{\partial x_i^2}(\bar{x}) \leq 0$$

and also

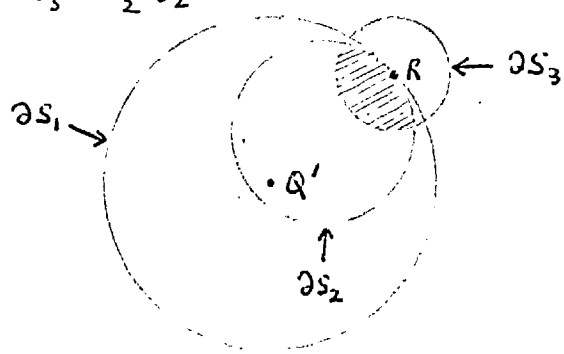
$$\frac{\partial u}{\partial y_i}(\bar{y}) = 0 \qquad \frac{\partial^2 u}{\partial y_i^2}(\bar{y}) \leq 0$$

From this we infer that u cannot have a non-negative relative maximum when strict inequalty holds in (3,2). We shall make use of this below.

We now prove the theorem by assuming that at point $P_1 \in \Omega$, $u(P_1) = M$ whilst for some other point P_2, $u(P_2) < M$. From this, a contradiction will be obtained. Let α be any arc lying in Ω joining P_1 and P_2, and let $d > 0$ be the greatest lower bound of the distances of each point of $\partial\Omega$ from α. Let Q be that point of α such that $u(Q) \equiv M$ and the distance $P_2 Q$ (along α) is as great as possible. Then inside that part of α terminated by P_2 and Q, $u < M$.

Let Q' be any point of this segment such that $|QQ'| < \frac{1}{2}d$ On Q' we construct the largest sphere possible with the property that inside it, $u < M$: this sphere is S_1 and there is a point on its boundary where $u \equiv M$.

Let S_2 be another sphere lying wholly inside S_1 and tangent to it at R and of radius r_2, and let S_3 be a third sphere with centre R and radius $r_3 = \frac{1}{2} r_2$



Now let $\partial S_3 = \partial S_3' + \partial S_3''$ where $\partial S_3'$ is $\partial S_3 \cap \bar{S}_2$ and $\partial S_3''$ is the remainder of ∂S_3. Since $\partial S_3'$ is closed, on it, $u < M - c$, where c = constant and > 0, and on $\partial S_3''$, $u < M$ by hypothesis.

Let x_0 be the centre of S_2 and let $\bar{d}_i = x_i - x_{0i}$, $\bar{d}^2 = \sum_{i=1}^{n} \bar{d}_i^2$ where x is any point of Ω.

We define the auxiliary function

$$\zeta(x) = e^{-\theta \bar{d}^2} - e^{-\theta r_2^2}$$

which vanishes on ∂S_2 and is respectively positive and negative inside and outside of it. Then in S_3

$$L\zeta = \theta e^{-\theta \bar{d}^2}\left[4\theta \sum_{i,j=1}^{n} a_{ij}\bar{d}_i\bar{d}_j - 2\left(\sum_{i=1}^{n} a_{ii} + a_i\bar{d}_i\right)\right] + b\zeta$$

$$\geqslant \theta e^{-\theta \bar{d}^2}\left[4\theta\gamma(x)\bar{d}^2 - 2\left(\sum_{i=1}^{n} a_{ii} + a_i\bar{d}_i\right)\right] + b\zeta$$

$$= \theta e^{-\theta \bar{d}^2}\gamma(x)\left[4\theta\bar{d}^2 - 2\left(\sum_{i=1}^{n}\frac{a_{ii}(x)}{\gamma(x)} + \frac{a_i\bar{d}_i}{\gamma(x)}\right) + \frac{b}{\gamma(x)}\left(e^{-\theta\bar{d}^2} - e^{-\theta r_1^2}\right)\right]$$

$$\geqslant \theta e^{-\theta \bar{d}^2}\gamma(x)\left[4\theta\bar{d}^2 - 2\left(\sum_{i=1}^{n}\frac{a_{ii}(x)}{\gamma(x)} + \frac{a_i\bar{d}_i}{\gamma(x)}\right) - \frac{|b|}{\gamma(x)}\left(e^{-\theta\left(\frac{\bar{d}}{2}\right)^2} - e^{-\theta r_1^2}\right)\right]$$

from which it is evident, using (3,1) that if θ is sufficiently large, say $\theta = \bar{\theta}$ then

$$L\zeta > 0 \quad , \qquad x \in S_3$$

Let

$$\omega = u + \varepsilon\zeta \quad , \qquad \varepsilon > 0$$

so that $L\omega > 0$ in S_3; by our previous result, ω cannot achieve a

non-negative maximum in S_3. Choose $0 < \varepsilon < \frac{c}{1-e^{-\theta l_2^2}}$. Now, on $\partial S_2'$ we have

$$\omega = u + \frac{c}{1-e^{-\theta l_2^2}} \cdot z \leq (M-c) + \frac{c}{1-e^{-\theta l_2^2}} z < M$$

Moreover, on $\partial S_2''$, $\omega < M$ since $z < 0$ there and $u \leq M$. But, at R, $\omega = M$. Since R is the centre of S_3 we infer that ω achieves a non-negative maximum inside S_3. This contradicts the previous finding and the theorem is proved.

It should be mentioned that theorems of the type proved above are well known. The present proof is based on a proof in [6] This book also contains some interesting historical remarks. From the theorem the following conclusions may be drawn.

Corollary 3.1

With the hyptheses of Theorem 3.1, let $u \in C^{(2)}(\Omega) \cap C^{(0)}(\bar{\Omega})$. Then

a) If $Lu = f$, $f \geq 0$, $(f \leq 0)$ then u has no non-negative maximum (non-positive minimum) in Ω unless it is constant.

b) If $Lu = f$, $b \equiv 0$, $f \geq 0$, $(f \leq 0)$ then u has no maximum (minimum) in Ω unless it is a constant

c) If $Lu = 0$, $u \neq$ constant then in Ω,

$$|u| < \max_{\partial\Omega} |u|$$

d) If $Lu = 0$, $b \equiv 0$, $u \neq$ constant then in Ω

$$\min_{\partial\Omega} u < u < \max_{\partial\Omega} u$$

Proof

a) follows immediately from the theorem

b) follows by a modification of the proof of the theorem

c) follows from the continuity of u in $\bar{\Omega}$ and from a)

d) follows from the continuity of u in $\bar{\Omega}$ and from b)

From c) we can infer the uniqueness of any solution of the boundary value problem

$$Lu = f \qquad u|_{\partial\Omega} = \phi \tag{3,3}$$

with $b \leq 0$. Indeed, if there were 2 solutions, say u_1 and u_2 their difference would satisfy the equation

$$L(u_1 - u_2) = 0$$

and the boundary condition $(u_1 - u_2)|_{\partial\Omega} = 0$. From c) $u_1 - u_2 \equiv 0$ so $u_1 \equiv u_2$. There are known examples for which uniqueness cannot in general be inferred if $b > 0$.

Another useful inference from c) is a statement about the continuous dependence of solutions to (3,3) on the boundary data, ϕ. Thus, if we perturb ϕ by ε so that $\bar{u}|_{\partial\Omega} = \phi + \varepsilon$, $L\bar{u} = f$ then $|\bar{u} - u| < \varepsilon|_{\partial\Omega}$, from c). This is of manifest significance for the problem of obtaining approximate solutions to (3,3).

We shall employ the results of Corollary 3.1 subsequently, when they will be used to establish the uniqueness of a solution of a special quasi-linear equation. This equation is introduced in the next section.

1.4 The equation $\operatorname{div}[\rho(|\nabla u|^2)\cdot\nabla u] = f(x)$

We shall, from now on confine our attention mostly to the equation

$$\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left[\rho(|\nabla u|^2)\frac{\partial u}{\partial x_i}\right] = f(x) \qquad (4,1)$$

with various boundary conditions of the type $(2,5)$. The function $\rho(t)$ where $t \geq 0$ is a scalar variable, will be assumed to have a continuous first derivative with respect to t. An example of a function $\rho(t)$ would be $(1+t)^{-\frac{1}{2}}$ which occurs in the theory of minimal surfaces. In this case, $n=2$ and $(4,1)$ is the equation

$$\frac{\partial}{\partial x}\left[\frac{u_x}{(1+u_x^2+u_y^2)^{1/2}}\right] + \frac{\partial}{\partial y}\left[\frac{u_y}{(1+u_x^2+u_y^2)^{1/2}}\right] = 0$$

Another example of a function ρ is given by $\rho = 1$ for which $(4,1)$ is just Poissons equation. Other examples of the occurence of $(4,1)$ in mathematical physics are given below with an appropriate density function

Occurence	density function	
[3] minimal surfaces	$(1+t)^{-1/2}$	$(4,2)$
[5] Gas dynamics	$(1-\frac{1}{2}c_1 t)^{1/c_1}$	$(4,3)$
[7] Plasticity	t^{-c_2} $0 \leq c_2 < \frac{1}{2}$	$(4,4)$
[8] Magnetostatics	$\frac{c_3+t}{1+t}$	$(4,5)$

where c_1, c_2, c_3 are various constants characteristic of the continua in which the phenomena are being studied. In each case, the boundary condition $(2,5)$ occurs. In the first 2 cases $f \equiv 0$.

We shall discuss the ellipticity of the operator on the left of $(4,1)$ for the case $n=2$. We have now for any $u \in C^{(2)}$

$$\frac{\partial}{\partial x_1}\left[\rho(|\nabla u|^2)u_{x_1}\right] = \rho(|\nabla u|^2)u_{x_1x_1} + u_{x_1}\left[2u_{x_1}u_{x_1x_1} + 2u_{x_2}u_{x_1x_1}\right]\rho'[|\nabla u|^2]$$

with a corresponding expression for $\frac{\partial}{\partial x_2}[\rho u_{x_2}]$ and we can arrange the coefficients of the second order terms as the elements of a (2×2) matrix A whose element a_{ij} say, is given by the coefficient of $u_{x_i x_j}$. This matrix is

$$\begin{bmatrix} \rho + 2u_{x_1}^2 \rho' & 2u_{x_1}u_{x_2}\rho' \\ 2u_{x_1}u_{x_2}\rho' & \rho + 2u_{x_2}^2\rho' \end{bmatrix} \qquad (4,6a)$$

and to apply the definition of ellipticity, it is sufficient for us to examine the eigenvalues of it. The eigenvalues are easily seen to be

$$\lambda_1 = \rho$$
$$\lambda_2 = \rho + 2|\nabla u|^2\rho' \qquad (4,6b)$$

and their positivity, or otherwise, with respect to the given function u determines the ellipticity of the operator. For the functions $(4,2)-(4,5)$ we get the ellipticity conditions (using $t = |\nabla u|^2 \geq 0$ for brevity)

$$(1+t)^{-1/2} > 0\,, \qquad (1+t)^{-1/2}\left(1 - \frac{t}{1+t}\right) > 0 \qquad (4,7)$$

for $(4,2)$. For $(4,3)$, $(4,6b)$ becomes

$$(1-\tfrac{1}{2}c_1 t)^{\frac{1}{c_1}} > 0\,, \qquad (1-\tfrac{1}{2}c_1 t)\left[1 + \frac{t}{1-\frac{1}{2}c_1 t}\right] > 0 \qquad (4,8)$$

whilst for $(4,4)$ and $(4,5)$ we find

$$t^{c_2} > 0 \qquad (1+2c_2)t^{c_2} > 0 \qquad (4,9)$$

$$\frac{c_3+t}{1+t} > 0 \qquad \frac{c_3+t}{1+t}\left[1 + 2t\left(\frac{1}{c_3+t} - \frac{1}{1+t}\right)\right] > 0 \qquad (4,10)$$

Thus, in the case of $(4,2)$ we have ellipticity everywhere that $t = |\nabla u|^2 < \infty$. For $(4,3)$ we have ellipticity whenever $t < 2/c_1$ (this implies in the physical problem that the flow is subsonic at each point). For $(4,4)$ it is necessary that $t > 0$ and finally, for $(4,5)$, $\forall \; t \geq 0$ we have ellipticity.

Now let us consider the solution of $(4,1)$ with $n = 2$ in a region Ω with the boundary condition $(2,5)$ Then for $(4,2)$ we can say that $(4,1)$ is elliptic with respect to any function u which has bounded 1st derivatives locally in Ω (of course $u \in C^{(1)}(\Omega)$). In particular, if the 1st derivatives of any solution are in $C^{(1)}(\bar{\Omega})$ the equation is elliptic with respect to this solution. A corresponding statement holds for $(4,3)$ and $(4,4)$. In the case of $(4,5)$ we can make the stronger statement that the equation is elliptic in Ω with respect to <u>any</u> $u \in C^{(2)}(\Omega)$.

1.5 <u>Uniqueness results</u>

We shall consider here the uniqueness of any given solution, $u(x)$, to the problem

$$\sum_{i=1}^{2} \frac{\partial}{\partial x_i}\left[\rho(|\nabla u|^2) \frac{\partial u}{\partial x_i} \right] = f(x) \qquad (5,1)$$

$$u|_{\partial\Omega} = \phi \qquad (5,2)$$

We replace $(5,1)$ with the slightly more general equation

$$a_{11}(u_{x_1}, u_{x_2}) u_{x_1x_1} + 2a_{12}(u_{x_1}, u_{x_2}) u_{x_1x_2} + a_{22}(u_{x_1}, u_{x_2}) = f(x) \qquad (5,3)$$

with the boundary condition $(5,2)$

Lemma 5.1

Let $(5,3)$ be elliptic in Ω with respect to the solution $u^{(1)}$, let $u^{(2)}$ be another solution and let the coefficients $a_{ij}(p_1,p_2) \in C^{(1)}(R_2)$. Then the function $V(x_1,x_2) = u^{(1)}(x_1,x_2) - u^{(2)}(x_1,x_2)$ satisfies the elliptic equation

$$a_{11}(u^{(1)}_{x_1}, u^{(1)}_{x_2}) V_{x_1x_1} + 2a_{12}(u^{(1)}_{x_1}, u^{(1)}_{x_2}) V_{x_1x_2} + a_{22}(u^{(1)}_{x_1}, u^{(1)}_{x_2}) V_{x_2x_2} +$$

$$+ b_1 V_{x_1} + b_2 V_{x_2} = 0 \qquad (5,4)$$

(b_i $i=1,2$ are defined below), $\forall$ $x \in \Omega$.

Proof

We shall use the notation $a^{(k)}_{ij} = a_{ij}(u^{(k)}_{x_1}, u^{(k)}_{x_2})$ for shortness. Then by assumption

$$\begin{aligned} a^{(1)}_{11} u^{(1)}_{x_1x_1} + 2a^{(1)}_{12} u^{(1)}_{x_1x_2} + a^{(1)}_{22} u^{(1)}_{x_2x_2} &= f(x) \\ a^{(2)}_{11} u^{(2)}_{x_1x_1} + 2a^{(2)}_{12} u^{(2)}_{x_1x_2} + a^{(2)}_{22} u^{(2)}_{x_2x_2} &= f(x) \qquad (5,5) \\ u^{(1)}\big|_{\partial\Omega} \equiv u^{(2)}\big|_{\partial\Omega} &= \phi \end{aligned}$$

By subtraction in $(5,5)$ it follows that

$$a^{(1)}_{11} V_{x_1x_1} + 2a^{(1)}_{12} V_{x_1x_2} + a^{(1)}_{22} V_{x_2x_2} +$$

$$+ (a^{(1)}_{11} - a^{(2)}_{11}) u^{(2)}_{x_1x_1} + 2(a^{(1)}_{12} - a^{(2)}_{12}) u^{(2)}_{x_1x_2} + (a^{(1)}_{22} - a^{(2)}_{22}) u^{(2)}_{yy} = 0 \qquad (5,6)$$

Now

$$a^{(1)}_{ij} - a^{(2)}_{ij} = a_{ij}(\nabla u^{(1)}) - a_{ij}(\nabla u^{(2)})$$

$$= \int_0^1 \frac{d}{dt} a_{ij}\left[t\nabla u^{(1)} + (1-t)\nabla u^{(2)}\right] dt$$

$$= \int_0^1 \frac{\partial a_{ij}}{\partial p_1}\left(t\nabla u^{(1)} + (1-t)\nabla u^{(2)}\right) V_{x_1}\, dt \quad +$$

$$+ \int_0^1 \frac{\partial a_{ij}}{\partial p_2}\left(t\nabla u^{(1)} + (1-t)\nabla u^{(2)}\right) V_{x_2}\, dt$$

where the derivatives exist and are continuous by assumption.

Substituting these expressions in (5,6) we achieve (5,4) with

$$b_i = u^{(2)}_{x_1x_1} \int_0^1 \frac{\partial a_{11}}{\partial p_i}\left(t\nabla u^{(1)} + (1-t)\nabla u^{(2)}\right) dt \quad +$$

$$+ 2u^{(2)}_{x_1x_2} \int_0^1 \frac{\partial a_{12}}{\partial p_i}(\cdot)\, dt + u^{(2)}_{x_2x_2} \int_0^1 \frac{\partial a_{22}}{\partial p_i}(\cdot)\, dt \qquad i = 1, 2$$

The lemma is proved.

The uniqueness theorem is obtained by an application of Corollary 3.1 d), and noting that $V \equiv 0$ on the boundary. However, it is necessary to impose sufficiently strong conditions on the functions a_{ij} and $\nabla u^{(i)}$ in order to ensure that the conditions of the Corollary are satisfied. One such set of conditions is the following

a) the operator

$$Lw = a_{11}(u^{(1)}_{x_1}, u^{(1)}_{x_2})\, w_{x_1x_1} + 2a_{12}(u^{(1)}_{x_1}, u^{(1)}_{x_2})\, w_{x_1x_2} + a_{22}(u^{(1)}_{x_1}, u^{(1)}_{x_2})\, w_{x_2x_2}$$

is uniformly elliptic

b) for all $t \in [0,1]$, $x \in \Omega$

$$\| t\nabla u^{(1)} + (1-t)\nabla u^{(2)} \|_2 < \infty$$

The second condition, b) ensures the boundedness of the coefficients

$b_i \quad i = 1, 2$.

1.6 Direct methods in the calculus of variations.

In this section we shall outline an approach, known in the literature on the calculus of variations as a direct method, to solving the problem of extremizing a functional defined in an ∞ dimensional function space. Some acquaintance with the calculus of variations is assumed. We have to introduce the following terminology: a function u, defined in a region Ω and on its boundary $\partial\Omega$ is said to have piecewise continuous first derivatives in $\bar{\Omega}$ if

a) u is continuous in $\bar{\Omega}$

b) $\bar{\Omega}$ can be divided into a finite number of subregions Ω_i $i = 1, 2, \ldots, M(u)$ in such a way that

i) In the k^{th} subregion the derivatives u_{x_i} $i = 1, 2, \ldots, n$ exist and are continuous, for $k = 1, 2, \ldots, M$

ii) In the k^{th} subregion the derivatives u_{x_i} $i = 1, 2, \ldots, n$ can be continuously extended onto the boundary $\partial\Omega_k$.

The condition b) (ii) imposes conditions on both the smoothness of $\partial\Omega_k$ and on the derivatives u_{x_i}.

It is clear from the above definition that discontinuities in the derivatives of u can only occur on the boundaries of the subregions Ω_i. At such points the derivatives may fail to exist. In order to have derivatives defined at each point of $\bar{\Omega}$ we shall define u_{x_i} at a point where boundaries $\partial\Omega_{k_1}, \partial\Omega_{k_2}, \ldots, \partial\Omega_{k_s}$ intersect to be the average value of all the derivatives (continuously extended) at the point. This means that if in Ω_{k_i} the derivative with respect to x_j is denoted by $u_{x_j}^{(k_i)}$ and its continuous extension onto $\bar{\Omega}_{k_i}$ by $\hat{u}_{x_j}^{(k_i)}$ then at any point x of the intersection $\partial\Omega_{k_1} \cap \ldots \cap \partial\Omega_{k_s}$ we define u_{x_j} to be $\frac{1}{s} \sum_{i=1}^{s} \hat{u}_{x_j}^{(k_i)}(x)$: the derivative u_{x_j}

defined in this way is defined at each point of $\bar{\Omega}$, but is generally discontinuous, with bounded discontinuities.

For the purposes of illustrating the use of direct methods we shall consider the following problem:

minimize

$$I(u) = \int_{\Omega} F(x, u, \nabla u)\, d\Omega \qquad (6,1)$$

in that class of functions $u \in L_2(\Omega)$ which have piecewise continuous 1st derivatives and are continuous in $\bar{\Omega}$. and satisfy the boundary condition

$$u|_{\partial\Omega} = \phi \qquad (6,2)$$

where ϕ is a given function, continuous on $\partial\Omega$. Ω is a bounded region of R_2. The above defined class of functions will be denoted by $\mathcal{B}_\phi$ and the problem just stated will be referred to as problem $(I, \mathcal{B}_\phi)$.

Let us first note that there has to exist at least one function $u \in \mathcal{B}_\phi$ for which $(6,1)$ is defined. Given $(6,2)$ this immediately places restrictions on the smoothness of $\partial\Omega$. We shall always assume that these restrictions are met. Secondly, it is necessary that

$$\inf_{u \in \mathcal{B}_\phi} I(u) > -\infty \qquad (6,3)$$

this will be the case, for example, if the function $F(x_1, x_2; y_1; z_1, z_2) \geq 0$ for $(x_1, x_2; y_1; z_1, z_2) \in R_5$.

$(6,3)$ will also be assumed always to hold. Then there will exist a greatest lower bound to the set

$$\{ I(u),\ u \in \mathcal{B}_\phi \}$$

and denoting this bound by m there exists a sequence of elements $u_1, u_2, \ldots, \in \mathfrak{B}_\phi$ such that

$$\lim_{n\to\infty} I(u_n) = m \tag{6,4}$$

Any sequence $\{u_n\}$ for which (6,4) holds is called a minimizing sequence. In connection with the minimizing sequence, two questions arise:

1) Does $\lim_{n\to\infty} \{u_n\}$ exist? (6,5)

2) If so, can it be said that if $\{u_n\} \to u \in \mathfrak{B}_\phi$ then

$$\lim_{n\to\infty} I[u_n] = I[u] \tag{6,6}$$

If the answer to these questions is positive, the problem $(I, \mathfrak{B}_\phi)$ is solved. We shall mention that a positive answer to (6,5) depends on demonstrating various compactness properties of the class $\mathfrak{B}_\phi$ whilst the possibility of the limit operation in (6,6) depends on semicontinuity properties of I (lower semicontinuity for the minimization problem, upper semicontinuity for the maximization problem). The relevant proofs can be carried out either in the weak or strong topologies of the space, or even in both if the convergence of the sequence of derivatives is involved. The above approach to the minimization problem is loosely called the direct method of minimization.

The present work is devoted almost entirely to the first aspect of the above discussion, namely to the construction of minimizing sequences. The aspect typified by the questions 1) and 2) requires rather different tools, analytic in character and requiring substantial mathematical preparation.

1.7 Conditions for a sequence to be minimizing

The fundamentals of the Ritz method for the construction of minimizing sequences are as follows:

in the class $\mathfrak{F}_\phi$ a family of functions, $\mathcal{Y}$, is taken, each member ω of which depends on a number s of numerical parameters (for example, the coefficients of a polynomial). Changing the values of the parameters changes the particular member of the family but not the fact of its membership. The members also have the property that among the members depending on $s+t$ parameters, $t \geq 0$, is included every member containing s parameters; i.e. if $\mathcal{Y}_s$ denotes the subclass of $\mathcal{Y}$ containing s parameters, then

$$\mathcal{Y}_0 \subset \mathcal{Y}_1 \subset \mathcal{Y}_2 \subset \ldots \subset \mathcal{Y}$$

the inclusions being strict. $\mathcal{Y}_i$ may, for example be polynomials of degree i, for $i = 0, 1, \ldots$.

Now let N be a given integer, and suppose that $\phi(a_1, \ldots, a_N; x) \equiv \phi_N$ (a_i are the parameters) generates the class $\mathcal{Y}_N$. Substituting ϕ_N into $I(u)$ we get

$$I(\phi(a_1, \ldots, a_N; x)) = I_N$$

which is a function of the N variables $a_1, \ldots, a_N$. These variables are now determined in such a way that the function of N variables, I_N is as small as possible. This will usually involve solving the simultaneous (non-linear) equations

$$\frac{\partial I_N}{\partial a_i} = 0 \qquad i = 1, 2, \ldots, N$$

Let I_N take the value $\bar{I}_N$ at the minimum (assumed to exist). Repeating the calculations with functions $\phi_{N'}$, $N' > N$ it follows, in view of $\mathcal{Y}_N \subset \mathcal{Y}_{N'}$ that $\bar{I}_N \geq \bar{I}_{N'}$. In this way we can generate a sequence $\{\bar{I}_N\}$ $N = 0, 1, \ldots, \ldots$ strictly non-increasing and bounded below. The sequence $\bar{\phi}_N(x)$ is called the Ritz sequence (for the family $\mathcal{Y}$). Now there exists ℓ such that $\{\bar{I}_N\} \to \ell$. We cannot conclude

$$\ell = \inf_{u \in \mathcal{B}_\phi} I(u) = m$$

In order to draw this conclusion we have to impose conditions on both $I(u)$ and $\mathcal{Y}$. A quite general result in this direction will be given. The following definition is needed.

Definition

A functional I defined on a metric space $(\mathcal{B}, \sigma)$ (with metric σ) is said to be upper semicontinuous (lower semicontinuous) at $u \in \mathcal{B}$ if, given any $\varepsilon > 0$, $\exists$ $\delta(\varepsilon)$ such that

$$I(\omega) < I(u) + \varepsilon \qquad (I(\omega) > I(u) - \varepsilon)$$

$\forall$ ω satisfying $\sigma(\omega, u) < \delta(\varepsilon)$.

Clearly a functional continuous at u is both upper and lower semicontinuous there, and conversely.

Theorem 7.1

Let functional I attain its infimum at a point $\bar{u}$ of a metric space $(\mathcal{B}, \sigma)$. If I is upper semicontinuous at $\bar{u}$ and $\mathcal{Y}$ is a set dense in $(\mathcal{B}, \sigma)$ the Ritz sequence is minimizing.

Proof

Let $\varepsilon > 0$ be given. By upper semicontinuity, $\exists$ $\delta(\varepsilon)$ such

that

$$I(v) < I(\bar{u}) + \varepsilon \qquad \text{for} \qquad \sigma(v, \bar{u}) < \delta(\varepsilon)$$

Choose v such that $\sigma(v, \bar{u}) < \frac{1}{2}\delta(\varepsilon)$ and then a $u_N \in \mathcal{Y}$ such that $\sigma(u_N, v) < \frac{1}{2}\delta(\varepsilon)$. Then

$$\sigma(u_N, \bar{u}) \leq \sigma(u_N, v) + \sigma(v, \bar{u})$$
$$= \delta(\varepsilon)$$

Consequently,

$$I(u_N) \leq I(\bar{u}) + \varepsilon \tag{7,1}$$

If now $\bar{u}_N$ is the member of the Ritz sequence corresponding to u_N in the sense of containing the same number of parameters, then $I(\bar{u}_N) \leq I(u_N)$ and from (7,1)

$$I(\bar{u}_N) \leq I(u_N) < I(\bar{u}) + \varepsilon$$

and since ε is arbitrary, the theorem is proved.

Later on, we shall apply this theorem to problems of interest to us, but it is first necessary to examine how we can formulate boundary value problems in a variational form.

1.8 On the variational formulation of boundary value problems

In this paragraph we shall investigate the senses in which the problem of solving

$$\operatorname{div}\left(\rho(|\nabla u|^2)\,\nabla u\right) = f(x) \qquad u|_{\partial\Omega} = \phi$$

can be reformulated as a problem of type $(I, \mathcal{B}_\phi)$.

In addition to $\mathcal{B}_\phi$ we shall introduce a class $\mathcal{B}_{2,\phi}$ consisting of that subset of $\mathcal{B}_\phi$ whose 2nd derivatives are continuous in Ω. Our interest in variational methods arises from the following observation:

consider the functional

$$I(u) = \int_\Omega \left[\int_a^{|\nabla u|^2} \rho(t)\, dt \right] d\Omega - 2\int_\Omega f \cdot u \, d\Omega \qquad (8,1)$$

where Ω is now a bounded region in R_n, ρ is the function considered in paragraph 4 and $f \in C^{(0)}(\bar{\Omega})$.

Considering $(8,1)$ in the form $(6,1)$ we find as a consequence of the hypotheses on $\rho(t)$ that $F \in C^{(2)}(R_{2n+1})$ where $(x; u; p) \in R_{2n+1}$ For $(8,1)$ we pose the problem $(I, \mathcal{B}_{2,\phi})$. By a well known argument [4] a necessary condition for u to be a minimum in $\mathcal{B}_{2,\phi}$ is that

$$\frac{d}{d\alpha}\left[I(u+\alpha\eta)\right]_{\alpha=0} = 0 \qquad (8,2)$$

where $\eta \in \mathcal{B}_{2,0}$. ($\mathcal{B}_{2,0}$ is that class of functions satisfying the same smoothness conditions as functions of $\mathcal{B}_{2,\phi}$ but vanishing on $\partial\Omega$). In view of the smoothness conditions on F and $u+\alpha\eta$ the derivative $(8,2)$ can be found by differentiating under the integral sign, and $(8,2)$ takes the form

$$\int_\Omega \left(\sum_{i=1}^n u_{x_i}\eta_{x_i} \right) \rho(|\nabla u|^2) - f\cdot\eta \; d\Omega = 0 \qquad (8,3)$$

Once again using the smoothness properties of u and η we can integrate the term

$$\int_\Omega \left(\sum_{i=1}^n u_{x_i}\eta_{x_i} \right) \rho(|\nabla u|^2)\, d\Omega = \int_\Omega \sum_{i=1}^n \frac{\partial}{\partial x_i}\left(\rho(|\nabla u|^2 u_{x_i}\right)\cdot\eta \; d\Omega \qquad (8,4)$$

and from $(8,3)-(8,4)$ since η is arbitrary, follows

$$\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(g(|\nabla u|^2)u_{x_i}\right) = f(x) \qquad (8,5)$$

$$u|_{\partial\Omega} = \phi\ , \qquad u \in \mathcal{B}_{2,\phi} \qquad (8,6)$$

On the other hand, multiplying $(8,5)$ by η and integrating by parts we can arrive back at $(8,2)$. From this we conclude the following:

a) at every stationary point of $I(u)$ in $\mathcal{B}_{2,\phi}$ $(8,5)-(8,6)$ are satisfied.

b) If u satisfies $(8,5)-(8,6)$ it is a stationary point for $I(u)$ in $\mathcal{B}_{2,\phi}$;

thus the stationary points of $(8,1)$ in $\mathcal{B}_{2,\phi}$ and the solutions of $(8,5)-(8,6)$ can be placed in a one to one correspondence.

What this means is that solutions to $(8,5)-(8,6)$ may be found as stationary points of $(8,1)$ in $\mathcal{B}_{2,\phi}$. In particular, if it was known a priori that $(8,5)-(8,6)$ had a unique solution, then $(8,1)$ would have a unique stationary point in $\mathcal{B}_{2,\phi}$: in such a case it is often possible to decide the character of the stationary point, (e.g. a minimum) on physical grounds.

A number of objections can be raised against considering $(8,1)$ in the class $\mathcal{B}_{2,\phi}$ as opposed to, say, the class $\mathcal{B}_{\phi}$. Let us mention two of them, one theoretical and one practical. The theoretical objection is that $I(u)$ in no way requires the functions for which it is defined to have 2nd derivatives. This requirement was imposed by us, in order that the connection with the differential equation could be seen. The practical objection is simply that it is very difficult, in general, to

find families of functions in $\mathfrak{B}_{2,\phi}$ with which to carry out the Ritz method. Neither of these objections applies to the class $\mathfrak{B}_\phi$. For this reason, our main concern will be with problem $(I, \mathfrak{B}_\phi)$.

We shall briefly explain certain new difficulties that arise in connection with the differential equation, and the solutions to problem $(I, \mathfrak{B}_\phi)$. Certainly, for $(I, \mathfrak{B}_\phi)$ the necessary condition (8,2) can be derived in the form (8,3),

$$\int_\Omega \Big(\sum_{i=1}^{n} u_{x_i} \eta_{x_i} \Big) \rho(|\nabla u|^2) - f\eta \; d\Omega = 0 \qquad (8,7)$$

however, the partial integration cannot now be carried out, since it is not known, for example, that

$$\frac{\partial}{\partial x_i}\left(u_{x_i}\right) \qquad i = 1, 2, \ldots, n$$

exists. Thus, a relation like (8,5) cannot be derived, and the connection with the differential equation cannot be established without further analysis. Indeed, in the absence of certain restrictions on the functional, it cannot be established at all - widening the domain of definition of the functional actually changes the character of the original minimization problem.

Faced with these difficulties, those persons interested in the theory of the existence of solutions to differential equations evolved the following approach: (8,7) is to be considered as the starting point for the investigation of the existence of solutions to the equation

$$\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\rho(|\nabla u|^2)\, u_{x_i}\right) = f(x) \qquad (8,8)$$

$$u|_{\partial\Omega} = \phi \qquad (8,9)$$

Functions u contained in a class of functions $\bar{\mathcal{B}}_\phi$ (actually wider than $\mathcal{B}_\phi$) satisfying $(8,9)$ and for which the integral identity $(8,7)$ is satisfied for all $\eta \in \bar{\mathcal{B}}_0$ are called generalized solutions of $(8,8)-(8,9)$. Finding conditions under which generalized solutions exist is the first (and easiest) problem to solve. Having given these conditions, there remains the question of their smoothness up to and on the boundary $\partial\Omega$.

We cannot go into these questions here: the reasons are the same as indicated at the end of paragraph 5. The problems discussed there are essentially connected with those we briefly introduced in the above paragraph. Accounts of this work, (which leads to a very complete existence theory) may be found, along with further references in [1] and [4].

As a step towards establishing a connection between solutions to $(8,5)-(8,6)$ and solutions to problem $(I, \mathcal{B}_\phi)$ we shall prove the following result.

Theorem 8.1

Let $\bar{u}$ satisfy $(8,5)-(8,6)$ and let $(8,5)$ be elliptic. Let $I(u)$ be given by $(8,1)$. Then

$$\inf_{u \in \mathcal{B}_{2,\phi}} I(u) = \inf_{v \in \mathcal{B}_\phi} I(v)$$

Proof

Let $n = 2$; the general case is similar. Consider any $\omega \in \mathcal{B}_\phi$ and put $h = \omega - \bar{u} \in \mathcal{B}_0$. Then

$$\int_a^{|\nabla\omega|^2} g(t)\,dt = \int_a^{(\bar{u}_{x_1}+h_{x_1})^2+(\bar{u}_{x_2}+h_{x_2})^2} g(t)\,dt = \psi\left(\bar{u}_{x_1}+h_{x_1},\ \bar{u}_{x_2}+h_{x_2}\right)$$

Then

$$\psi(\bar{u}_{x_1}+h_{x_1},\ \bar{u}_{x_2}+h_{x_2}) = \psi(\bar{u}_{x_1},\bar{u}_{x_2}) + h_{x_1}\psi_{\bar{u}_{x_1}}(\bar{u}_{x_1},\bar{u}_{x_2}) + h_{x_2}\psi_{\bar{u}_{x_2}}(\bar{u}_{x_1},\bar{u}_{x_2}) +$$

$$+\frac{1}{2}\left\{\psi^*_{\bar{u}_{x_1}\bar{u}_{x_1}} h_{x_1}h_{x_1} + 2\psi^*_{\bar{u}_{x_1}\bar{u}_{x_2}} h_{x_1}h_{x_2} + \psi^*_{\bar{u}_{x_2}\bar{u}_{x_2}} h_{x_2}h_{x_2}\right\}$$

where $\psi^*_{\bar{u}_{x_1}\bar{u}_{x_1}}$ represents a mean value for $\psi_{\bar{u}_{x_1}\bar{u}_{x_1}}$ etc. Inte grating over Ω

$$I(\omega) - I(u) = \int_\Omega h_{x_1}\psi_{\bar{u}_{x_1}}(\nabla\bar{u}) + h_{x_2}\psi_{\bar{u}_{x_2}}(\nabla\bar{u}) - 2fh\ d\Omega +$$

$$+\frac{1}{2}\int_\Omega \psi^*_{\bar{u}_{x_1}\bar{u}_{x_1}} h_{x_1}h_{x_1} + 2\psi^*_{\bar{u}_{x_1}\bar{u}_{x_2}} h_{x_1}h_{x_2} + \psi^*_{\bar{u}_{x_2}\bar{u}_{x_2}} h_{x_2}h_{x_2}\ d\Omega$$

The 1st integral can be integrated by parts since $\bar{u} \in C^{(1)}(\Omega)$. and $\rho \in C^{(1)}[0,\infty]$. The result of this is

$$\int_\Omega \left(\frac{\partial}{\partial x_1}(\rho(|\nabla\bar{u}|^2)\bar{u}_{x_1} + \frac{\partial}{\partial x_2}(\rho(|\nabla\bar{u}|^2)\bar{u}_{x_2} - f\bar{u}\right)h\ d\Omega,$$

which is zero by (8,5)

To evaluate the 2nd integral, note that

$$\psi_{u_{x_1}u_{x_1}} = 2\left[\rho(|\nabla u|^2) + 2u_{x_1}^2\rho'(|\nabla u|^2)\right]$$

$$\psi_{u_{x_1}u_{x_2}} = 2\left[2u_{x_1}u_{x_2}\rho'(|\nabla u|^2)\right]$$

$$\psi_{u_{x_2}u_{x_2}} = 2\left[\rho(|\nabla u|^2) + 2u_{x_2}^2\rho'(|\nabla u|^2)\right]$$

so that the coefficients in the quadratic form under the integral sign are precisely the elements of the matrix (4,6a). Since the eigenvalues are positive for an elliptic operator we conclude that the second integral is positive. Thus,

$$I(\omega) - I(\bar{u}) = I(\bar{u}+h) - I(\bar{u}) > 0$$

But ω is arbitrary in $\mathcal{B}_\phi$ and the theorem follows from this.

We are therefore assured that enlarging the domain of definition of functional

$$\int_\Omega \left(\int_a^{|\nabla u|^2} p(t)\,dt - 2fu \right) d\Omega$$

from $\mathcal{B}_{2,\phi}$ to $\mathcal{B}_\phi$ does not bring in any lower minimum, in the case that $(8,5)-(8,6)$ holds. It is at least necessary that such a result should hold if we wish to solve the differential equation by extremizing the above functional in a class of functions wider than that for which the differential equation is defined. But we have <u>not</u> proved that the extremal elements in the wider class actually lie in the smaller class.

1.9 Statement of the variational problem

We shall be concerned in the sequel with the following problem:

Minimize the functional

$$J(u) = \int_\Omega \left[\int_a^{|\nabla u|^2} p(t)\,dt - 2f(x)u \right] d\Omega \qquad (9,1)$$

in the class of functions $\mathcal{B}'_\phi$ continuous and having piecewise continuous first derivatives in $\bar{\Omega}$ and taking on the value ϕ on $\partial\Omega_1$ where $\partial\Omega = \partial\Omega_1 + \partial\Omega_2$.

The only new point here is that the boundary function ϕ need be specified on only part of the boundary, $\partial\Omega_1$. If the functional was minimized in $\mathcal{B}'_{2,\phi}$ it is well known that a minimizing element would satisfy the condition

$$\frac{\partial u}{\partial n} = 0 \quad \text{on} \quad \partial\Omega_2 \qquad (9,2)$$

where $\hat{n}$ is the outward normal to $\partial\Omega_2$. This condition is often referred to by saying that $(9,2)$ is a "natural" boundary condition for $(9,1)$. Its computational significance originates with the fact that members of the class in which $(9,1)$ is to be minimized do not have to be constrained to satisfy $(9,2)$. The minimizing element automatically satisfies this condition.

1.10 Semicontinuity of $J(u)$

The present section is given over to a discussion of the semicontinuity of $J(u)$.

In the class $\mathcal{B}'_\phi$ we introduce the metric

$$\sigma(u,v) = \max_{\bar{\Omega}} |u-v| + \sum_{i=1}^{n} \max_{\bar{\Omega}} |u_{x_i} - v_{x_i}|$$

We shall prove that at the point $\bar{u}$ defined in theorem 7.1 we have for any $\varepsilon > 0$

$$J(\omega) < J(\bar{u}) + \varepsilon \qquad \forall \quad \sigma(\omega, \bar{u}) < \delta(\varepsilon)$$

This is equivalent to the upper semicontinuity of J at $\bar{u}$. For the proof, consider for any fixed $x \in \bar{\Omega}$ and $\omega \in \mathcal{B}'_\phi$

$$\left(\int_a^{|\nabla w|^2} g(t)\,dt - 2f\omega \right) - \left(\int_a^{|\nabla \bar{u}|^2} g(t)\,dt - 2f\bar{u} \right) =$$

$$= \int_{|\nabla \bar{u}|^2}^{|\nabla w|^2} g(t)\,dt - 2f(\omega - \bar{u}) \qquad (10,1)$$

$$\leq \left| \int_{|\nabla \bar{u}|^2}^{|\nabla w|^2} g(t)\,dt \right| + |2f| \cdot |\omega - \bar{u}|$$

$$\leq |g(t^*)| \cdot \left| |\nabla w|^2 - |\nabla \bar{u}|^2 \right| + |2f| \cdot |\omega - \bar{u}|$$

where $\xi^* \in (|\nabla\omega|^2, |\nabla\bar{u}|^2)$. Recalling that $\rho \in C^1[0,\infty)$ and that $|f| \leq B$ we have that this expression

$$\leq B_1 \left| |\nabla\omega|^2 - |\nabla\bar{u}|^2 \right| + 2B|\omega - \bar{u}|$$

$$= B_1 \sum_{i=1}^{n} |(\omega_{x_i} + \bar{u}_{x_i})| \cdot |\omega_{x_i} - \bar{u}_{x_i}| + 2B|\omega - \bar{u}| \qquad (10,2)$$

It is evident that $\exists\ r_0$ such that the constant B_1 remains valid for all ω such that $\sigma(\omega, \bar{u}) < r_0$. Furthermore, for all such ω, $|\omega_{x_i} + \bar{u}_{x_i}| \leq B_2 \quad i = 1, 2, \ldots, n$ since $\bar{u}$ is fixed. Hence $(10,2)$ satisfies

$$\leq B_3 \sum_{i=1}^{n} |\omega_{x_i} - \bar{u}_{x_i}| + 2B|\omega - \bar{u}|$$

$$\leq B_4\, \sigma(\omega, \bar{u}) \qquad (10,3)$$

From $(10,1)$ and $(10,3)$ we have then

$$\left(\int_a^{|\nabla\omega|^2} \rho(t)\,dt - 2f\omega \right) - \left(\int_0^{|\nabla\bar{u}|^2} \rho(t)\,dt - 2f\bar{u} \right) \leq B_4\, \sigma(\omega, \bar{u})$$

and integrating over Ω,

$$J(\omega) - J(\bar{u}) \leq B_5\, \sigma(\omega, \bar{u})$$

Hence, by taking $\sigma(\omega, \bar{u}) < \frac{\varepsilon}{B_5}$ we will have

$$J(\omega) < J(\bar{u}) + \varepsilon$$

and the uppersemicontinuity is proved.

Now consider the metric in $\mathcal{B}'_\phi$,

$$\sigma'(u, v) = \left(\int_\Omega |u - v|^2\, d\Omega + \sum_{i=1}^{n} \int_\Omega |u_{x_i} - v_{x_i}|^2\, d\Omega \right)^{1/2} \qquad (10,4)$$

If it is assumed that ρ is a bounded function, $\rho(t) \leq M$ then it can be proved that J is uppersemicontinuous in the metric space (B_4', σ') The proof follows the lines of the previous proof, using the integral metric (10,4) instead of the 'pointwise' metric $\sigma(u,v)$. The bound on ρ is needed to ensure the existence of a constant to play the role of B_1 in (10,2) but the rest of the proof is not different in principle from the proof already given.

1.11 Resume of Part 1

The purposes of this part have been threefold; to present salient theoretical facts about ellipticity, to introduce a class of non-linear elliptic equations of practical importance, and to indicate certain difficulties which arise when boundary value problems are reformulated as variational problems.

The first of these matters was considered in sections 1.2, 1.3 and 1.5, the second in section 1.4 and the third in sections 1.6 - 1.10. In 1.2 the notions of ellipticity were discussed and in 1.3 an important consequence of the definition of ellipticity was given. Application of the results of 1.3 was made in 1.5, in proving the uniqueness theorem for quasilinear equations. Additional general references for these three sections are [9] and [10].

In section 1.4 we introduced a class of equations of importance in mathematical physics. The literature contains a great many equations of this type. It may perhaps be said that the given equation will occur whenever it is necessary to take into account physical non-linearities in describing phenomena which, in the first approximation, are governed by the Laplace (or Poisson) equation. Furthermore, there is a plasticity theory in which a functional of the type (8,1) occurs directly, and for which it is difficult to write down a governing differential equation on account of the lack of smoothness in the function ρ. An example from this plasticity theory is given in [11].

In sections 1.6 and 1.7 various problems concerning variational methods were introduced. For some of them we were able to provide

reasonable solutions, as for example Theorem 7.1. It is probable that related theorems to this are known, but they do not seem to be easily available. The mathematical theory alluded to at the end of section 1.6 is contained in [4]. In section 1.8 some problems concerning the extension of functionals (to wider domains of definition) were exposed. These problems cannot be regarded as solved, although it should be mentioned that for a class of quadratic functionals they have been more or less completely solved. The solutions form part of the Friedrichs-Mikhlin theory of semibounded operators.

Finally, section 1.9 contained a statement of the variational problem to be considered in the sequel and section 1.10 consisted of a demonstration of the semicontinuity of the functional involved in the variational problem.

PART 2: Piecewise polynomial trial functions

1. Scope and purpose of this section.

From the last few sections of Part 1 emerges the importance, for the construction of minimizing sequences, of having available functions depending on parameters and possessing certain smoothness properties (trial functions). The present section is devoted to an examination of a class of functions of this type. In order to motivate the following analysis we shall outline in a heuristic way what kind of results are needed.

The principles are clearly seen in the case in which $\Omega \in R_1$. Let Ω be the open interval (x_0, x_n) and $\bar{\Omega} = \Omega + \partial\Omega$ the closed interval $[x_0, x_n]$. Subdivide $\bar{\Omega}$ into n not necessarily equal intervals $[x_0, x_1]$, $[x_1, x_2]$, . . . , $[x_{n-1}, x_n]$. It has been found empirically that a trial function $\psi(x, a_1, \dots a_N)$ of a computationally convenient kind is one satisfying the following conditions

1) For any given set of a_i's, $\psi(x, a)$ is continuous at each point $x \in [x_0, x_n]$
2) In each subinterval $[x_K, x_{K+1}]$, $\psi(x, a)$ reduces to a polynomial $P_{K+1}(x)$, of some specified degree $m \geq 1$, the same for each subinterval.

Additionally, it may be required that $\psi(x, a)$ satisfies some boundary conditions. It will be assumed here that

$$\psi(x_0, a) \equiv \psi(x_n, a) = 0 \qquad \forall \quad a \in R_N$$

It is now necessary to recall some basic facts about Lagrange interpolation: let $[y_0, y_m]$ be an interval and let it be sub-

divided by points $y_0 \leq y_1 < \ldots < y_{m-1} < y_m$. Let $Q(x)$ by any polynomial of degree m. Then Q can be represented in the form

$$Q(x) = \sum_{i=0}^{m} Q(y_i)\, \ell_i(x) \qquad (1,1)$$

where the functions $\ell_i(x)$, the Lagrange polynomials, are m^{th} degree polynomials which satisfy the conditions

$$\ell_i(y_j) = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases} = \delta_{ij} \qquad (1,2)$$

This representation $(1,1)$, $(1,2)$ is unique. On the other hand, if Q_i $i=0,1,\ldots,m$ are arbitrary numbers then the expression

$$Q^{(m)}(x) = \sum_{i=0}^{m} Q_i\, \ell_i(x) \qquad (1,3)$$

represents a polynomial of degree m which takes on the values Q_i at the points y_i for $0 \leq i \leq m$. It therefore follows that the class of all polynomials of degree m can be generated by varying the <u>parameters</u> Q_i in $(1,3)$.

These ideas can be employed in the construction of the functions $\psi(x, a)$ satisfying 1) and 2) above. For this purpose, each interval $[x_k, x_{k+1}]$ $k=0,1,\ldots,n-1$ has introduced into it $m-1$ points $x_i^{(k+1)}$ $i=1,2,\ldots,m-1$ not coinciding with each other or the end points. For convenience it will be supposed that $[x_k, x_{k+1}]$ is subdivided into m equal parts. Moreover, it will also be assumed, again for convenience only, that the intervals $[x_k, x_{k+1}]$ are of equal length.

With the interval $[x_k, x_{k+1}]$ is now associated the Lagrange polynomials $\ell_i^{(k+1)}(x)$ $i=0,1,\ldots,m$ such

that $\ell_i^{(k+1)}(x_i^{(k+1)}) = 1, \quad i = 1, 2, \ldots, m \; ; \quad \ell_0^{(k+1)}(x_k) = \ell_m^{(k+1)}(x_{k+1}) = 1 \; ;$

with $x_i^{(k+1)}$ let parameters $a_i^{(k+1)}$ $i = 1, 2, \ldots, m-1$ be associated, and with x_k and x_{k+1} let parameters b_k, b_{k+1} be associated. Consider the polynomials

$$P_{k+1}(x) = b_k \ell_0^{(k+1)}(x) + \sum_{j=1}^{m-1} a_j^{(k+1)} \ell_j^{(k+1)}(x) + b_{k+1} \ell_m^{(k+1)}(x) \qquad (1,4)$$

$$P_k(x) = b_{k-1} \ell_0^{(k)}(x) + \sum_{j=1}^{m-1} a_j^{(k)} \ell_j^{(k)}(x) + b_k \ell_m^{(k)}(x) \qquad (1,5)$$

In the interval $[x_{k-1}, x_{k+1}]$ let us define a function

$$\bar{\Psi}(x) = \begin{cases} P_k(x) & x \in [x_{k-1}, x_k) \\ \\ P_{k+1}(x) & x \in (x_k, x_{k+1}] \end{cases}$$

$\bar{\Psi}$ is not defined at x_k at the moment, and we are also assuming that neither x_{k-1} or x_{k+1} is an endpoint of Ω. Both $(1,4)$, $(1,5)$ are defined at $x = x_k$. They take the values

$$P_{k+1}(x_k) = b_k \ell_0^{(k+1)}(x_k) = b_k$$

$$P_k(x_k) = b_k \ell_m^{(k)}(x_k) = b_k$$

by the definition of $\ell_0^{(k+1)}(x)$, $\ell_m^{(k)}(x)$. Extending the definition of $\bar{\Psi}(x)$ to take in x_k a new function $\tilde{\Psi}(x)$ is obtained;

$$\tilde{\psi}(x) = \begin{cases} P_k(x) & x \in [x_{k+1}, x_k] \\ P_{k+1}(x) & x \in [x_k, x_{k+1}] \end{cases}$$

$\tilde{\psi}$ is continuous in $[x_{k-1}, x_{k+1}]$, reduces to an m^{th} degree polynomial in each of the subintervals $[x_{k-1}, x_k]$, $[x_k, x_{k+1}]$ and has both these properties regardless of the manner in which the parameters are assigned. It is evident that this procedure can be extended to each of the subintervals in $[x_0, x_n]$; the only point is that in order to satisfy the imposed boundary condition, we have to take $b_0 = b_n = 0$. In view of the meaning attached to the parameters, namely as values of the polynomial at the nodes, the boundary condition will thereby be satisfied. The function $\psi(x, a)$ defined in this way is a piecewise $C^{(1)}[a, b]$ function. It has been explained previously that functions of this type can be used for the construction of minimizing sequences.

The objective of the next few sections is the development of a theory of Lagrange interpolation in n dimensions in such a way that trial functions analogous to those just considered in R_1 can be constructed in a manner suitable for computation. A range of generalizations are possible, depending, amongst other things, on how an n - dimensional interval is defined. The class of polynomials which can be accomodated in the theory is closely bound up with this definition. Within the

scope of the definitions of interval and polynomial considered by us, we shall develop results which enable the inference to be made that each polynomial is uniquely representable in a form analogous to (1,1) and that, conversely, a form analogous to (1,3) generates the entire class of polynomials within our definition (which is a general one). This converse inference is important. The methods by which these results are obtained are interesting in themselves. It should be mentioned that the question of satisfying the conditions set out in Part 1, section 7 is considered subsequently.

2. Simplexes and polynomials

We begin by studying certain properties of simplexes. Throughout, $\bar{S}_n$ will mean a nondegerate simplex (i.e. a simplex whose n measure is nonzero) whose $n+1$ vertices X_i $i=0,1,\ldots,n$ are situated in R_n. As is well known, if x is any point in R_n it is uniquely representable in the barycentric co-ordinates λ_i which satisfy

$$\sum_{i=0}^{n} \lambda_i = 1 \quad , \qquad x = \sum_{i=0}^{n} \lambda_i X_i \tag{2,1}$$

the following geometric interpretation of the λ_i is available: let $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T$, so that (2,1) is then equivalent to

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{01} & x_{11} & \cdots & x_{n1} \\ x_{02} & x_{12} & \cdots & \\ \vdots & & & \vdots \\ x_{0n} & \cdots & & x_{nn} \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \tag{2,2}$$

where $x = (x_1, x_2, \dots, x_n)^T$. Now the determinant of the matrix on the left is $n!$ x the volume of $\bar{S}_n$; by Cramer's rule we see that λ_i is the ratio of the n volumes of the simplexes whose vertices are $X_0, X_1, \dots, X_{i-1}, x, X_{i+1}, \dots X_n$ and $X_0, X_1, \dots, X_n$. This is particularly important when $x \in \bar{S}_n$ so that $\lambda_i \geq 0 \quad i = 0, 1, \dots, n$.

We can imagine (2,2) solved for the vector $\Lambda = (\lambda_0, \lambda_1, \dots, \lambda_n)$ We should then get vectors $u^{(i)}$ and scalars b_i such that

$$\lambda_i = \lambda_i(x) = (u^{(i)}, x) + b_i \qquad i = 0, 1, \dots, n .$$

Consider the set of points x for which $\lambda_i(x) = 0$. Evidently, these points lie in the hyperplane

$$(u^{(i)}, x) + b_i = 0$$

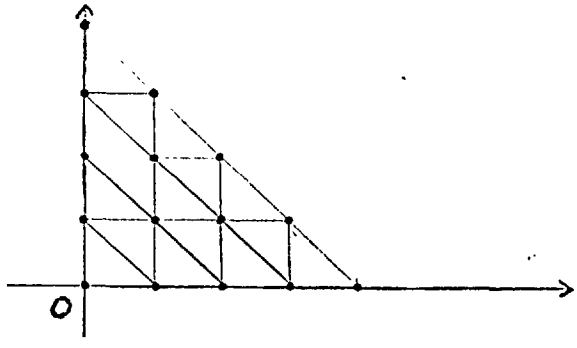
In accordance with the geometrical meaning of λ_i it follows that the hyperplane $\lambda_i(x)$ contains the face of $\bar{S}_n$ not containing the vertex X_i or, to put it another way, x is a linear combination of the vectors $X_0, X_1, \dots, X_{i-1}, X_{i+1}, \dots X_n$.

Each simplex contains an important collection of discrete point sets, here called principal lattices of various orders. Their definition is this: let m be a fixed positive inte ger. The discrete point set

$$B(m,n) = \left\{ x \in R_n \;\middle|\; x = \sum_{i=0}^{n} \lambda_i X_i ,\ \lambda_i = j/m ,\ (j \text{ integral}),\ \sum_{i=0}^{n} \lambda_i = 1 \right\}$$

is called the m^{th} order principal lattice of $\bar{S}_n$. It is easy to prove that $B(m,n)$ contains $\binom{m+n}{m} \equiv \binom{m+n}{n}$ members.

In order to clarify the significance of this definition, a simple example will be considered. We shall show how to construct $B(4,2)$ for $\overline{S}_2$. we take the triangle with vertices $(0,0)$, $(1,0)$, $(0,1)$. Divide each side into 4 equal parts (m in general). Draw in the lines, parallel to the sides of the triangles, through the points of subdivision. The points of intersection of these lines, together with the points on the boundary of the triangle constitute $B(4,2)$



Note that it has $\binom{4+2}{2} = 15$ members.
This simple example of course does not prejudice the generality of the definition of $B(m,n)$ already given.

We pass on to a consideration of certain properties of a class of polynomials which will be used later for interpolation. Let us define the symbol.

$$P_{\tau,s} = \sum_{|\alpha| \le \tau} a_{\alpha_1 \alpha_2 \ldots \alpha_s} x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_s^{\alpha_s} \qquad (2,3)$$

where α_i $i = 1,2,\ldots,s$ is a non-negative integer $|\alpha| = \sum_{i=1}^{s} \alpha_i$ and $x = (x_1, x_2, \ldots, x_s)^T$ is an s dimensional vector. A notable property of $(2,3)$ is that each component x_i is 'equally represented' The class of polynomials $(2,3)$ of degree τ and with $x \in R_s$ will be denoted by $V_{\tau,s}$. If $P \in V_{\tau,s}$ then $\mathcal{R}(P, \Pi_q)$ will mean the restriction of P to the hyperplane Π_q where

Π_q

is defined in

$$\Pi_q = \left\{ x \in R_n \,\middle|\, x = \sum_{i=0}^{q} \lambda_i X_{k_i}, \quad \sum_{i=0}^{q} \lambda_i = 1 \right\}$$

for $q+1$ vertices X_{k_i} $i=0,1,\cdots,q$. Also, $V_{t,s}(\Pi_q)$ will mean the class of functions obtained by restricting each element of $V_{t,s}$ to Π_q.

It will now be proved that, roughly speaking, a polynomial of type (2,3) retains its characteristics if its domain of definition is limited to linear manifolds in R_n.

Lemma 2.1

Let $Q \in V_{m,n}(\Pi_{n-d})$ for some given Π_{n-d}; then $Q \in V_{m,n-d}$.

Proof

By definition Q is got by restricting a certain element P of $V_{m,n}$ to Π_{n-d}

We can select $n-d+1$ points X_i $i=0,1,..,n-d$ in Π_{n-d} in such a way that each $x \in \Pi_{n-d}$ is uniquely representable as

$$x = \sum_{i=0}^{n-d} \lambda_i X_i, \qquad \sum_{i=0}^{n-d} \lambda_i = 1 .$$

For such an x,

$$P(x) = P(x(\lambda)) = \sum_{|\alpha| \le m} a_{\alpha_1 \ldots \alpha_n} x_1^{\alpha_1} \ldots x_n^{\alpha_n}$$

$$= \sum_{|\alpha| \le m} a_{\alpha_1 \ldots \alpha_n} \left(X_{10} + \sum_{i=1}^{n-d} \lambda_i (X_{1i} - X_{10}) \right)^{\alpha_1} \times \ldots \ldots \times$$

$$\times \left(X_{n0} + \sum_{i=1}^{n-d} \lambda_i (X_{ni} - X_{n0}) \right)^{\alpha_n} \qquad (2,4)$$

where $X_i = (x_{1i}, x_{2i}, \ldots x_{ni})^T$. It is clear that the quantity on the right is a polynomial in the λ_i and that this polynomial is of degree m. Since also the λ_i appear symmetrically in (2,4) we can write it as

$$P = \sum_{|\alpha| \leq m} a^i_{\alpha_1 \ldots \alpha_{n-d}} \lambda_1^{\alpha_1} \ldots \lambda_{n-d}^{\alpha_{n-d}}$$

on rearrangement. But the λ_i's $i = 1, 2, \ldots, n-d$ can be chosen arbitrarily and this amounts to the assertion of the Lemma.

The next two lemmas are needed for the main theorem of this section. This theorem may be regarded as a generalization of the fact that if α is a zero of a polynomial $P(x)$ in R_1 then $P(x) = (x-\alpha) \times$ $\times P_-(x)$, $P_-(x)$ being of 1 degree lower than $P(x)$.

We shall denote by $\sigma_{n-d}(x_0, \rho)$ an open sphere of radius $\rho > 0$ and centre $x_0 \in \Pi_{n-d}$ and lying entirely in Π_{n-d}.

Lemma 2.2

Let $P \in V_{m,n}$ and suppose that $P(x) = 0$ whenever $x \in \sigma_{n-d}(x_0, \rho)$. Then for each $y \in \Pi_{n-d}$ $P(y) = 0$.

Proof

Let $x_1 \in \Pi_{n-d}$. Since Π_{n-d} is convex it contains the segment $\overline{x_0 x_1}$. Consider

$$\Pi_1 = \{ x \mid x = (1-\mu) x_0 + \mu x_1 \}$$

By Lemma 2.1 , $\mathcal{R}(P, \Pi_1) \in V_{m,1}$. Selecting $m+1$ points in $\Pi_1 \cap \sigma_{n-d}(x_0; \mathcal{P})$ defined by μ_i $i=0,1,\dots,m$ we see that $\mathcal{R}(P, \Pi_1) \equiv 0$. Since x_1 is arbitrary, the lemma follows.

Lemma 2.3

Let $P_1 \in V_{m,n}$ and suppose that for any fixed but arbitrary hyperplane Π_{n-1} we have $\mathcal{R}(P_1, \Pi_{n-1}) = 0$. Then $P_1(x)$ can be written in the form

$$P_1(x) = \lambda(x) . P(x)$$

where $\lambda \in V_{1,n}$ and for each $y \in \Pi_{n-1}$ $\lambda(y) = 0$.

Proof

Speaking loosely, we may say that if P_1 vanishes in a linear manifold of dimension $n-1$ then P_1 contains the equation of this manifold as a (linear) factor.

For the proof, we introduce into R_n the simplex $\bar{S}_n$ where vertices $X_1, \dots, X_n$ lie in Π_{n-1} and the distance from X_0 to Π_{n-1} is positive. If $x \in R_n$ then

$$x = \sum_{i=0}^{n} \lambda_i X_i \qquad \sum_{i=0}^{n} \lambda_i = 1$$

and so

$$\begin{aligned} P_1(x) &= \sum_{j=0}^{m} \sum_{|\alpha|=j} a_{\alpha_1 \dots \alpha_n} \left(\sum_{i=0}^{n} \lambda_i x_{1i} \right)^{\alpha_1} \cdots \left(\sum_{i=0}^{n} \lambda_i x_{ni} \right)^{\alpha_n} \\ &= \sum_{j=0}^{m} \left(\sum_{i=0}^{m} \lambda_i \right)^{m-j} \sum_{|\alpha|=j} a_{\alpha_1 \dots \alpha_n} \left(\sum_{i=0}^{n} \lambda_i x_{1i} \right)^{\alpha_1} \cdots \left(\sum_{i=0}^{n} \lambda_i x_{ni} \right)^{\alpha_n} \\ &= \sum_{|\alpha|=m} b_{\alpha_0 \dots \alpha_n} \lambda_0^{\alpha_0} \cdots \lambda_n^{\alpha_n} \end{aligned}$$

where $b_{\alpha_0 \dots \alpha_n}$ $(|\alpha| = m)$ are certain combinations of the

components of the X_i and the $a_{\alpha_1 \cdots \alpha_n}$.

Now let $y \in \Pi_{n-1}$; then

$$y = \sum_{i=0}^{n} \mu_i X_i , \quad \sum_{i=0}^{n} \mu_i = 1, \quad \mu_0 = 0 .$$

For such a y, P_1 takes the form

$$P_1(y) = \sum_{|\alpha|=m} b_{\alpha_0 \cdots \alpha_n} \mu_0^{\alpha_0} \cdots \mu_n^{\alpha_n}$$

$$= \sum_{|\alpha|=m} b_{0\alpha_1 \cdots \alpha_n} \mu_1^{\alpha_1} \cdots \mu_n^{\alpha_n} , \quad \text{since } \mu_0 = 0 .$$

But this is zero, since $P_1(y) = 0$ for $y \in \Pi_{n-1}$ so that

$$\sum_{|\alpha|=m} b_{0\alpha_1 \cdots \alpha_n} \mu_1^{\alpha_1} \cdots \mu_n^{\alpha_n} = 0$$

We shall now consider the homogeneous polynomial

$$\sum_{|\alpha|=m} b_{0\alpha_1 \cdots \alpha_n} \zeta_1^{\alpha_1} \cdots \zeta_n^{\alpha_n}$$

and we shall show that <u>whatever</u> the ζ_i it is zero. Indeed, for $\sum_{i=1}^{n} \zeta_i = 1$ this statement is already true. The general case follows simply from this when $\sum_{i=1}^{n} \zeta_i \neq 0$. For then,

$$\sum_{|\alpha|=m} b_{0\alpha_1 \cdots \alpha_n} \zeta_1^{\alpha_1} \cdots \zeta_n^{\alpha_n} = s^m \sum_{|\alpha|=m} b_{0\alpha_1 \cdots \alpha_n} (\zeta_1/s)^{\alpha_1} \cdots (\zeta_n/s)^{\alpha_n}$$

where $s = \sum_{i=1}^{n} \zeta_i$. If $s = 0$ then a continuity argument is used.

Now write, generally

$$\sum_{|\alpha|=m} b_{\alpha_0 \cdots \alpha_n} \lambda_0^{\alpha_0} \cdots \lambda_n^{\alpha_n} = \sum_{|\alpha|=m} b_{0\alpha_1 \cdots \alpha_n} \lambda_1^{\alpha_1} \cdots \lambda_n^{\alpha_n} +$$

$$+ \lambda_0 \sum_{|\beta|=m-1} b'_{\beta_0 \cdots \beta_n} \lambda_0^{\beta_0} \cdots \lambda_n^{\beta_n}$$

$$= \lambda_0 \sum_{|\beta|=m-1} b'_{\beta_0 \cdots \beta_n} \lambda_0^{\beta_0} \cdots \lambda_n^{\beta_n}$$

in view of what we just proved. But as we saw at the beginning of this section, the λ_i can be regarded as (linear) functions of position and therefore,

$$P_1(x) = \lambda_0(x) \sum_{|\beta|=m-1} b'_{\beta_0 \cdots \beta_n} \lambda_0(x)^{\beta_0} \cdots \lambda_n(x)^{\beta_n}$$

This is the statement of the lemma. We should point out that the function $\lambda(x)$ of the lemma is determined only up to an arbitrary multiplicative constant, since if $\lambda(y) = 0$ so does $K\lambda(y) = 0$.

Theorem 2.1

Let positive integers m, s be given, $s \leq m$ and let $\Pi_{n-1,j}$ $j = 0, 1, \ldots, s-1$ be s distinct hyperplanes. Let there be given s functions, $\eta_j(x)$ $j = 0, 1, \ldots, s-1$ linear in x and such that $x \in \Pi_{n-1,j}$ implies $\eta_j(x) = 0$. Suppose $P_1 \in V_{m,n}$ and that $\mathcal{R}(P, \Pi_{n-1,j}) = 0$ $j = 0, 1, \ldots, s-1$. Then

$$P_1(x) = \left(\prod_{i=0}^{s-1} \eta_i(x) \right) \cdot P_2(x)$$

where $P_2 \in V_{m-s,n}$.

Proof

The proof is by induction. Assume that the theorem is true for some $1 \leq t < s$ so that P_1 vanishes in the hyperplanes $\Pi_{n-1,i}$ $i = 0, 1, \ldots, t-1$ and is representable as

$$P_1(x) = P_3(x) \cdot \left(\prod_{i=0}^{t-1} \eta_i(x) \right), \qquad P_3 \in V_{m-t,n}.$$

Consider $\Pi_{n-1,t}$. By hypothesis, $\Pi_{n-1,t}$ contains a $\sigma(x_0, \rho)$ not containing any point of the intersection $\Pi_{n-1,0} \cap \cdots \cap \Pi_{n-1,t}$. Let $x_1 \in \sigma(x_0, \rho)$. Then

$$P(x_1) = P_3(x_1) \cdot \left(\prod_{i=0}^{t-1} \eta_i(x_1) \right) = 0$$

and therefore, $P_3(x_1) = 0$. By lemma 2.2, $\mathcal{R}(P_3, \Pi_{n-1,t}) \equiv 0$ and by lemma 2.3

$$P_3(x) = P_4(x) \cdot \eta_t(x).$$

therefore, we have proved that

$$P_1(x) = P_4(x) \left(\prod_{i=0}^{t} \eta_i(x) \right)$$

But, again by lemma 2.3 the result is true for $t = 1$ and so the lemma is proved, by mathematical induction.

The following corollary holds.

Corollary

Let $P \in V_{m,n}$ and vanish in $s > m$ distinct hyperlanes $\Pi_{n-1,i}$ $i = 0, 1, \ldots, s-1$. Then $P \equiv 0$

Proof

Select any m hyperlanes, say $\Pi_{n-1,i}$ $i = 0, 1, \ldots, m-1$

Then

$$P(x) = C_1 \prod_{i=0}^{m} \eta_i(x) \qquad C_1 = \text{const.}$$

Since P vanishes in a further hyperplane, C_1 must be zero and the result is proved.

Before continuing the development, let us indicate an extension of this theorem.

Considering again R_1 we recall that if a polynomial $P(x)$ has the form $(x-\alpha)^s P_-(x)$ then P and its derivatives up to $(s-1)$th order vanish at α, and conversely. If we replace $(x-\alpha)$ by $\eta_i(x)$ we can make a related statement in the general case. Since this result

is not required for our present purposes we shall omit the rigourous statement and proof.

3. Interpolation

The material in section 2 can now be used to prove the following theorem.

Theorem 3.1

Let $x \in R_n$, let $\bar{S}_n$ be given. Let the points of the principal m^{th} order lattice of $\bar{S}_n$ be ordered and called $q_i \quad i=1,2,\ldots,\binom{m+n}{m}$. Suppose that $f_i \quad i=1,2,\ldots,\binom{m+n}{m}$ are given arbitrary numbers. Then there is exactly one $P \in V_{m,n}$ satisfying $P(q_i) = f_i \quad i=1,2,\ldots,\binom{m+n}{m}$.

Proof

Put another way, the theorem asserts the unique solvability of the interpolation problem in the class $(2,3)$ and in $B(m,n)$.

To prove this result we use induction on n. Let us note the following important fact: Let $0 \le j \le m$ be a fixed integer and consider the hyperplane

$$\Pi_{n-1,j} = \left\{ x \;\middle|\; x = \frac{j}{m} X_0 + \sum_{i=1}^{n} \lambda_i X_i, \; \frac{j}{m} + \sum_{i=1}^{n} \lambda_i = 1 \right\}. \qquad (3,0)$$

Speaking geometrically, this is a certain hyperplane parallel to that face of the simplex opposite vertex X_0. Its intersections with the edges of $\bar{S}_n$, $(n+1)$ of them, are the lattice points defined by

$$\lambda_0 = \frac{j}{m} \qquad \lambda_i = 1 - \frac{j}{m} \qquad i=1,2,\ldots,n \; ;$$

i.e. if these points are $Y_i \quad i=1,2,\ldots,m$ then

$$Y_i = \frac{j}{m} X_0 + \left(1 - \frac{j}{m}\right) X_i \qquad i=1,2,\ldots,n \qquad (3,1)$$

These n points generate an $n-1$ dimensional simplex, $\bar{S}_{n-1}$. We shall show that the lattice $B(m-j, n-1)$ in $\bar{S}_{n-1}$ conincides with that part of $B(m, n)$ of $\bar{S}_n$ which lies in $\Pi_{n-1,j}$. Indeed, $B(m-j, n-1)$ is

$$\left\{ x \mid x = \sum_{i=1}^{n} \mu_i V_i , \quad \mu_i = \frac{I_i}{m-j} \quad 0 \le I_i \le m-j , \quad \sum_{i=1}^{n} \mu_i = 1 \right\}$$

where I_i is an integer, so that $x \in B(m-j, n-1)$ implies, by (3,1)

$$\begin{aligned} x &= \sum_{i=1}^{n} \mu_i \left[\frac{j}{m} X_0 + \left(1 - \frac{j}{m}\right) X_i \right] \\ &= \frac{j}{m} X_0 + \sum_{i=1}^{n} \mu_i \left(1 - \frac{j}{m}\right) X_i \\ &= \frac{j}{m} X_0 + \sum_{i=1}^{n} \frac{I_i}{m-j} \left(\frac{m-j}{m} \right) X_i \\ &= \frac{j}{m} X_0 + \sum_{i=1}^{n} \frac{I_i}{m} X_i \end{aligned}$$

which is, by definition, a point of $B(m,n)$. It is easy then to see that $B(m-j, n-1)$ and that part of $B(m, n)$ in $\Pi_{n-1,j}$ are identical. The proof uses this fact. We shall now show that the homogeneous interpolation problem (i.e. $f_i = 0$ $i = 1, 2, \ldots, \binom{m+n}{m}$) has only the zero solution. This will imply our result. Let $1 \le N < n$ and suppose that the homogeneous problem in $B(m, N)$ has any solution $P_0 \in V_{m,N}$: then $Q_0 = R(P_0, \Pi_{N-1,0}) \in V_{m,N-1}$. But Q_0 vanishes at the lattice points of $\Pi_{n-1,0}$ and by the induction hypothesis the only element of $V_{m,N-1}$ with this property is the zero element, and thus, $R(P_0, \Pi_{N-1,0}) \equiv 0$. By

Theorem 2.1 then

$$P_0(x) = P_1(x)\,\eta_0(x)$$

where $\eta_0(x)$ is as given in that Theorem and $P_1 \in V_{m-1,N}$.

Now consider $\Pi_{N-1,1}$. Since $\Pi_{N-1,1} \cap \Pi_{N-1,0} \equiv 0$ we can apply a similar reasoning again to $P_1(x)$ and $\Pi_{N-1,1}$ to get

$$P_1(x) = P_2(x)\,\eta_1(x) \qquad P_2 \in V_{m-2,N}$$

and continuing this, we eventually arrive at

$$P_0(x) = k \prod_{i=0}^{m-1} \eta_i(x) \qquad k = \text{const}$$

and it remains to consider $\Pi_{N-1,m}$. Now $P_0(x)$ must vanish at X_0 and this means that $k \equiv 0$. Therefore, if the homogeneous problem has a solution, that solution is necessarily zero. Since zero is a solution it is unique. The homogeneous problem is uniquely solvable for the N dimensional problem if the $N-1$ dimensional problem is. Obviously for $N = 1$ we have this result, and the theorem is proved.

Note that (2,3) contains $\binom{m+n}{m}$ coefficients. The method of proof of this theorem is rather more general than may appear initially and it can be extended to cover some more general distributions of interpolation points.

4. Representation results

So far we have not given any method for finding the interpolation polynomial. In the present part we shall give explicit formulas for it. In fact we shall construct $\binom{m+n}{m}$ functions, $\ell_i(x)$ such that

$$\ell_i(p_j) = \delta_{ij} \qquad i = 1, 2, \ldots, \binom{m+n}{m} \tag{4,1}$$

where p_j are points of $B(m, n)$ and δ_{ij} is the Kronecker delta. Consider any vertex, say X_0 of $\bar{S}_n$ and consider the parallel hyperplanes $\Pi_{n-1,j}$ $j = 0, 1, \ldots, m-1$ of $(3, 0)$. Let $\eta_0^{(j)}(x)$ be linear and such that

$$\eta_0^{(j)}(x) = 0 \qquad \forall \quad x \in \Pi_{n-1,j}$$

Thus $\eta_0^{(j)}(x) = 0$ is the equation of the hyperplane $\Pi_{n-1,j}$; moreover it differs by only a multiplicative constant from $\eta_i(x)$ of Theorem 2.1. We shall associate a similar set of functions with each vertex. Thus with X_i is associated $\eta_i^{(j)}(x)$ for $i = 0, 1, \ldots, n$ and $\eta_i^{(j)}(x) = 0$ whenever $x \in \Pi_{n-1,j}^{(i)}$

Assume now that $\bar{x}$ is any definite member of $B(m, n)$

$$\bar{x} = \sum_{i=0}^{n} \bar{\lambda}_i X_i \qquad \sum_{i=0}^{n} \bar{\lambda}_i = 1$$

where $\bar{\lambda}_i = a_i/m$ and a_i is a non-negative integer. By definition of the $\eta_i^{(j)}(x)$ we have $\eta_i^{(a_i)}(\bar{x}) = 0$ $i = 0, 1, \ldots, n$. This is because $\bar{x}$ lies at the intersection of the hyperplanes $\lambda_k(x) = a_k/m$ $k = 0, 1, \ldots, n$ and so lies in each of them: therefore, since $\lambda_k(x) - \frac{a_k}{m} = c\,\eta_k^{(a_k)}(x)$ for some constant c, $\eta_i^{(a_i)}(\bar{x}) = 0$.

Consider the expression

$$P_{\bar{x}} = \prod_{i=0}^{n} \left(\prod_{j=0}^{a_i - 1} \eta_i^{(j)}(x) \right) \qquad (\eta_i^{(-1)} = 1 \quad \forall \ i)$$

and observe that $P_{\bar{x}} \in V_{m,n}$

Let $\bar{y}$ $(\neq \bar{x}) \in B(m, n)$ and

$$\bar{y} = \sum_{i=0}^{n} \bar{\mu}\, \chi_i \qquad \sum_{i=0}^{n} \bar{\mu}_i = 1 \qquad \mu_i = \frac{b_i}{m} \quad i = 0, 1, \ldots, m$$

As with $\bar{x}$ we have $\eta_i^{(b_i)}(\bar{y}) = 0 \quad i = 0, 1, \ldots, n$. Now, there is at least one index K, $0 \leq K \leq n$ for which $b_K - a_K \leq -1$. For if not, since b_j, a_j are integers $b_j - a_j \geq 0$ for each j and this contradicts $\sum_{j=0}^{n} (a_j - b_j) = 0$, since $\bar{x} \neq \bar{y}$. So, let K be such that $b_K \leq a_K - 1$.

Now consider

$$P_{\bar{x}}(\bar{y}) = \prod_{i=0}^{n} \left(\prod_{j=0}^{a_i - 1} \eta_i^{(j)}(\bar{y}) \right)$$

$$= \prod_{\substack{i=0 \\ i \neq k}}^{n} \left(\prod_{j=0}^{a_i - 1} \eta_i^{(j)}(\bar{y}) \right) \cdot \prod_{j=0}^{a_K - 1} \eta_K^{(j)}(\bar{y})$$

We have just shown that $b_K \leq a_K - 1$ and therefore the factor $\eta_K^{(b_K)}(\bar{y})$ appears in the rightmost product. But this is zero, as we have seen, so $P_{\bar{x}}(\bar{y}) = 0$. Thus, $P_{\bar{x}}(x)$ vanishes whenever $x \neq \bar{x} \in B(m, n)$. It follows that the function

$$\frac{P_{\bar{x}}(x)}{P_{\bar{x}}(\bar{x})} \qquad (4,3)$$

satisfies the condition (4,1) and is the desired function. Since $\bar{x}$ was any point of $B(m, n)$ the entire family (4,1) can be generated in this way.

We shall write (4,3) in a more tractable form. We take $\eta_i^{(0)}(x) = \lambda_i(x)$; then as already indicated, $\eta_i^{(j)}(x) = \lambda_i(x) - j/m$. Inserting this in (4,3), and (4,1) we get

$$\ell_{\bar{x}}(x) = \frac{1}{a_0! \cdot \ldots \cdot a_n!} \prod_{i=0}^{n} \prod_{j=0}^{a_i - 1} [m\lambda_i(x) - j] \qquad (4,4)$$

Expressions which are essentially identical to these were given in [12]. where they were obtained on the basis of their possession of certain formal properties of the one dimensional Lagrange polynomials.

For future reference, we note at this point that if $x \in \bar{S}_n$ $\exists$ $K'(m,n)$, independent of $\bar{S}_n$ but dependent on m and n such that

$$|\ell_{\bar{x}}(x)| \leq K'(m,n)$$

This follows immediately from (4,4), since $0 \leq \lambda_i(x) \leq 1$. Obviously, if $\ell_i(x)$ is the polynomial of type (4,4) associated with $p_i \in B(m,n)$ then $\exists$ $K(m,n)$ such that

$$\sum_{B(m,n)} |\ell_i(x)| \leq K(m,n) \qquad (4,5)$$

Let us observe that the interpolation process described gives for any $P \in V_{m_1,n}$, $0 \leq m_1 \leq m$ an alternative representation, as follows from a consideration of the difference between P and its interpolant. In view of this we can get certain algebraic identities from the relation

$$P(x) = \sum_{B(m,n)} \ell_i(x) P(p_i) \qquad P \in V_{m_1,n};\ m_1 \leq m$$

In particular,

$$\sum_{B(m,n)} \ell_i(x) = 1$$

which is well known when $n = 1$.

5. Examples

Some simple examples of the application of the previous results will now be given:

Example 1

Let us construct the interpolation polynomial of degree l in n dimensional Euclidean space. For our simplex, we shall take the simplex T_1 whose vertices are

$$\begin{aligned} X_0 &= (0, 0, \dots, 0)^T \\ X_1 &= (1, 0, \dots, 0)^T \\ X_2 &= (0, 1, \dots, 0)^T \\ &\vdots \\ X_n &= (0, 0, \dots, 1)^T \end{aligned}$$

this set of points coincides also with the principal lattice $B(1,n)$.

It follows from the properties of the $\lambda_i(x)$ that

$$\lambda_i(x) = x_i \qquad i = 1, 2, \dots, n$$

$$\lambda_0(x) = -\sum_{i=1}^{n} x_i + 1$$

since these functions are linear and satisfy

$$\lambda_i(X_j) = \delta_{ij}$$

From the formula $(4,4)$, it follows that they are the functions $l_i(x)$ $i=0,1,\dots,n$ of $(4,4)$ associated with the lattice. If now we associate

with each X_i a number a_i then the linear function

$$P_1(x) = \sum_{i=0}^{n} a_i \lambda_i(x)$$

is the (only) linear function which assumes the given values a_i at the vertices X_i. These linear functions find wide use in numerical analysis. They are closely related to the 'pyramid' functions of Synge [13] they also appear in certain versions of the multidimensional secant method; this is most clearly shown in the algorithm of Wolfe [14]. Finally, they make another appearance in the simplest form of the finite element method.

Example 2

It is well known that some very useful integration formulas can be got by integrating the interpolating polynomial in 1 dimension. The same is true of the present interpolation polynomials.

We shall indicate the basis of this procedure. It was developed by Sylvester [12]. Suppose we want to integrate a function f over the simplex T_1. Let us interpolate f with an m^{th} degree polynomial,

$$f = \sum_{\beta(m,n)} f_i \ell_i(x) + R_m$$

Then

$$\int_{T_1} f \, d\bar{T}_1 \sim \sum_{\beta(m,n)} f_i \int_{T_1} \ell_i(x) \, d\bar{T}_1$$

$$= \sum_{i=1}^{\binom{m+n}{m}} \omega_i f_i$$

where $\omega_i = \int_{T_1} \ell_i(x) \, d\bar{T}_1$. It is therefore a question of integrating the functions $\ell_i(x)$ over T_1 in order to generate the weights ω_i.

To do this we can make use of a known formula,

$$\int_{T_1} \lambda_1^{m_1} \cdots \lambda_{n+1}^{m_{n+1}} \, dT_1 = \frac{m_1!\, m_2! \cdots m_{n+1}!\, n!}{(m_1 + m_2 + \cdots + m_{n+1} + n)!}$$

This is proved e.g. in [15]. Since all the functions $\ell_i(x)$ are expressed in the form of polynomials in λ_k they can be integrated to get the ω_i. Sylvester carried out this programme for polynomials of degree ≤ 8 for $n=2$ and has given detailed instructions for using the resulting formulas.

We might also point out that if f is itself a polynomial of degree $\leq m$ then (from the interpolation theory) the integration is exact.

6. Triangulations and continuous trial functions

We have seen the possibility of defining polynomials on terms of their values at the points of a principal lattice of a simplex. In this section the previous results will be used to construct trial functions of the piecewise polynomial type which are continuous on a closed region $\bar{\Omega}$. Before this can be done, some definitions concerning the regions $\bar{\Omega}$ are needed.

Definition (Wallace [16])

A (simplicial) complex is a subset of R_n consisting of the union of simplexes of various dimensions such that the intersection of any 2 of them is empty, or an entire k dimensional face, $k=0,1,\ldots,n-1$.

The dimension of complex is the dimension of its largest constituent simplex.

We shall restrict $\bar{\Omega}$ by requiring that $\partial\Omega$ is a complex of

dimension $(n-1)$.

Definition

$\bar{\Omega}$ is triangulated if it is expressed as the union of a finite number of nondegenerate n dimensional simplexes, in such a way that every face of every simplex either defines part of $\partial\Omega$ or is also a face of another simplex.

The use of complexes to define $\partial\Omega$ and the above definition were given in [17].

Let us now assume that $\bar{\Omega} \in R_n$ has been triangulated, and let $\bar{S}_1, \ldots, \bar{S}_t$ be the components of the triangulation. Let an integer $m \geq 1$ be given. In $\bar{S}_i$ $i=1,2,\ldots,t$ we construct the principal lattice of order m. Let $\bar{S}_i$ and $\bar{S}_j$ share a common face, $\bar{S}_{ij}$ of dimension $n-1$. By the definition of a principal lattice, those lattice points of $\bar{S}_i$ lying in $\bar{S}_{ij}$ constitute a principal lattice in the $n-1$ dimensional simplex $\bar{S}_{ij}$. Similarly, those of $\bar{S}_j$ also constitute a principal lattice in $\bar{S}_{ij}$. Since $\bar{S}_{ij}$ has only 1 m^{th} order principal lattice we conclude that that part of the principal lattice of $\bar{S}_i$ lying in $\bar{S}_{ij}$ coincides with that part of the principal lattice of $\bar{S}_j$ lying in $\bar{S}_{ij}$ and that this lattice in $\bar{S}_{ij}$ in fact constitutes its m^{th} order principal lattice. This argument does not depend essentially on the fact that dimension $(\bar{S}_{ij}) = n-1$, and holds whatever the dimension of the intersection of $\bar{S}_i$ and $\bar{S}_j$. Whatever the dimension of this intersection, the intersection itself is always part of a linear manifold.

We shall now show how to construct trial functions continuous across the interface of two simplexes of a triangulation.. Let parameters $a_K^{(i)}$ be associated with the points of the lattice in $\bar{S}_i$ not in $\bar{S}_{ij}$ and

parameters $a_K^{(j)}$ with the corresponding points of $\bar{S}_j$. With the (common) boundary points of $\bar{S}_i$ and $\bar{S}_j$ lying in $\bar{S}_{ij}$ associate parameters $b_s^{(ij)}$. The subscripts K and S range over all the points in the classes of points indicated for them. In $\bar{S}_i$ denote the m^{th} order Lagrange polynomials, defined in (4,4), by $\ell_s^{(i)}(x)$ and in $\bar{S}_j$ by $\ell_s^{(j)}(x)$. Let $B_1^{(i)}(m,n)$, $B_2^{(i)}(m,n)$ denote the lattice points in $\bar{S}_i$ respectively in $\bar{S}_{ij}$ and not in $\bar{S}_{ij}$; similarly for $B_1^{(j)}(m,n)$ and $B_2^{(j)}(m,n)$

Consider the function

$$P_m(x) = \begin{cases} \sum_{B_1^{(i)}(m,n)} b_s^{(ij)} \ell_s^{(i)}(x) + \sum_{B_2^{(i)}(m,n)} a_K^{(i)} \ell_K^{(i)}(x) & x \in \bar{S}_i \qquad (6,1) \\ \sum_{B_1^{(j)}(m,n)} b_s^{(ij)} \ell_s^{(j)}(x) + \sum_{B_2^{(j)}(m,n)} a_K^{(j)} \ell_K^{(j)}(x) & x \in \bar{S}_j \qquad (6,2) \end{cases}$$

We shall prove that it is continuous in $\bar{S}_i \cup \bar{S}_j$. Obviously this is true if $x \notin \bar{S}_{ij}$. Suppose $x \in \bar{S}_{ij}$. By lemma 2.1 in $\bar{S}_{ij}$ (6,1) and (6,2) are both polynomials of degree m in $n-d$ variables, where $n-d$ is the dimension of $\bar{S}_{ij}$. It is easy to verify that if $\bar{S}_{ij}$ is of dimension $n-d$ then it contains $\binom{m+n-d}{n-d}$ lattice points which, as we have seen, also constitute its principal lattice. Subtracting (6,1) from (6,2), we again get an m^{th} degree polynomial in $n-d$ variables. Consider its behaviour at the lattice points in $\bar{S}_{ij}$. It vanishes at them, by the definition of the $\ell^{(i)}(x)$ and $\ell^{(j)}(x)$ and the fact that parameters $b_s^{(ij)}$ are common to (6,1) and (6,2). By theorem 3.1, in $\bar{S}_{ij}$, (6,1) is identically equal to (6,2). Thus, the continuity of P_m is established.

It is quite clear how these arguments can be extended to the whole of $\bar{\Omega}$. The resulting function, depending on parameters is our sought trial

function. It is continuous on $\bar{\Omega}$; if it is required to satisfy a boundary condition, for example with an m^{th} degree polynomial function given on the boundary it is only necessary to assign to those parameters associated with boundary points the values of the given boundary polynomial at the points. In this way, whatever values are ascribed, or computed for the remaining parameters the resulting function, continuous on $\bar{\Omega}$ will satisfy the imposed boundary condition.

7. Error of interpolation

So far, we have constructed polynomials in simplexes and shown how they can be defined across interfaces of pairs of simplexes in a continuous manner. We have also seen that trial functions can be constructed on this basis. With a view to applying Theorem 7.1 of Part 1, we are now going to derive bounds for the error of interpolation of a function by the methods set out in sections 2,3 & 4 and also bounds on the error incurred by using the derivatives of the interpolation polynomial to approximate the derivatives of the function. Apart from the use to which they will be put by us, these results complement the results of sections 2,3 & 4. The combined results of sections 2,3,4,7, and 8 form a natural generalization to n dimensions of the 1 dimensional Lagrange interpolation theory for equidistant nodes. As such, it has several applications in numerical analysis.

An exact representation for the error of interpolation will now be obtained. Let $f \in C^{(m+1)}(\bar{S}_n)$ and let p, q be any 2 points of $\bar{S}_n$. With these points is associated a function $\phi_{qp}(\lambda)$ of the scalar variable λ, defined by

$$\phi_{qp}(\lambda) = f[q + \lambda(p-q)] \qquad (7,1)$$

and for $p_i \in B(m,n)$ we shall write

$$\phi_{pp_i} = \phi_i(\lambda) \qquad i = 1,2,\dots,\binom{m+n}{m}$$

Furthermore, put

$$R_{m+1,i} = \frac{1}{m!}\int_0^1 \phi_i^{(m+1)}(\lambda)(1-\lambda)^m d\lambda \tag{7,2}$$

If $p_i \in B(m,n)$ then that m^{th} degree polynomial taking the values 1 at p_i and 0 at every other lattice point will be denoted by $\ell_i(x)$.

Theorem 7.1

Let $p \in \bar{S}_n$. With the above notations, there holds

$$f(p) = \sum_{B(m,n)} [f(p_i) - R_{m+1,i}]\,\ell_i(p) \tag{7,3}$$

and

$$|f(p)| \le K_1(m,n) \max_i |f(p_i)| + K_2(m,n) M_{m+1} h^{m+1} \tag{7,4}$$

where $|D^i f| \le M_{m+1}$ in $\bar{S}_n$, $|i| = m+1$; h is the greatest linear dimension in the simplex and $K_i(m,n)$ depend on m,n but not on $\bar{S}_n$

Proof

Let $q \in \bar{S}_n$ and consider $\phi_{qp}(\lambda)$. By Taylors Theorem,

$$\phi_{qp}(1) = \sum_{k=0}^{m} \frac{1}{k!}\phi_{qp}^{(k)}(0) + \frac{1}{m!}\int_0^1 \phi_{qp}^{(m+1)}(\lambda)(1-\lambda)^m d\lambda \tag{7,5}$$

Let $(p-q)_k$ be the k^{th} component of the vector $(p-q)$. Then from (7,1), for $j = 0,1,\dots,m+1$

$$\phi_{qp}^{(j)}(0) = \sum_{|i|=j} \frac{\partial^j f(q)}{\partial x_1^{i_1}\cdots\partial x_n^{i_n}} (p-q)_1^{i_1}\cdots(p-q)_n^{i_n}$$

so that (7,5) is equivalent to

$$f(P) = \sum_{k=0}^{m} \frac{1}{k!} \sum_{|i|=k} \frac{\partial^k f(q)}{\partial x_1^{i_1} \cdots \partial x_n^{i_n}} (p-q)_1^{i_1} \cdots (p-q)_n^{i_n} +$$

$$+ \frac{1}{m!} \int_0^1 \phi_{qp}^{(m+1)}(\lambda)(1-\lambda)^m d\lambda \qquad (7,6)$$

The 1st term on the right is a polynomial in the components of P, of degree m. Put it equal to $T_m(p, q)$ so that

$$T_m(p,q) = \sum_{k=0}^{m} \frac{1}{k!} \sum_{|i|=k} \frac{\partial^k f(q)}{\partial x_1^{i_1} \cdots \partial x_n^{i_n}} (p-q)_1^{i_1} \cdots (p-q)_n^{i_n}$$

By Theorem 3.1

$$T_m(p,q) = \sum_{B(m,n)} T_m(p_i, q)\, \ell_i(p)$$

and from (7, 6)

$$T_m(p_i, q) = f(p_i) - \frac{1}{m!} \int_0^1 \phi_{qp_i}^{(m+1)}(\lambda)(1-\lambda)^m d\lambda$$

so that

$$T_m(p,q) = \sum_{B(m,n)} \left[f(p_i) - \frac{1}{m!} \int_0^1 \phi_{qp_i}^{(m+1)}(\lambda)(1-\lambda)^m d\lambda \right] \ell_i(p)$$

Substituting this into (7,6), we get

$$f(p) = \sum_{B(m,n)} \left[f(p_i) - \frac{1}{m!} \int_0^1 \phi_{qp_i}^{(m+1)}(\lambda)(1-\lambda)^m d\lambda \right] \ell_i(p) +$$

$$+ \frac{1}{m!} \int_0^1 \phi_{qp}^{(m+1)}(\lambda)(1-\lambda)^m d\lambda$$

Now recalling that q is arbitrary in $\bar{S}_n$ take $q = p$. The last integral is zero and the sum remaining takes the form

$$f(p) = \sum_{B(m,n)} \left[f(p_i) - R_{m+1,i} \right] \ell_i(p)$$

which proves (7,3). From this follows immediately

$$|f(p)| \le \left(\max_i |f(p_i)| + \max_i |R_{m+1,i}| \right) \sum_{B(m,n)} |\ell_i(p)| \qquad (7,7)$$

and it has already been indicated in (4,5) that $\sum_{\beta(m,n)} |\ell_i(p)|$ can be bounded by a constant dependent on m and n but not on $\bar{S}_n$. Also,

$$R_{m+1,i} = \frac{1}{m!}\int_0^1 \phi_i^{(m+1)}(\lambda)(1-\lambda)^m d\lambda \qquad (7,8)$$

$$|R_{m+1,i}| \leq \frac{1}{m!} \max_{\lambda\in[0,1]} |\phi_i^{(m+1)}(\lambda)| \cdot \left|\int_0^1 (1-\lambda)^m d\lambda\right|$$

Now,

$$\phi_i^{(m+1)}(\lambda) = \sum_{|i|=m+1} \frac{\partial^{(m+1)} f[p+\lambda(p_i-p)]}{\partial x_1^{i_1}\cdots\partial x_n^{i_n}} (p_i-p)_1^{i_1}\cdots(p_i-p)_n^{i_n}.$$

and hence $|\phi_i^{(m+1)}| \leq K_1'(m,n) M_{m+1} h^{m+1}$, $K_1' = \text{const}$; (7,9)

whilst $\int_0^1 (1-\lambda)^m d\lambda = (m+1)^{-1}$. (7,10)

Substituting these results in (7,7) there follows

$$|f(p)| \leq \left[\max_i |f(p_i)| + K_2'(m,n) M_{m+1} h^{m+1}\right] \cdot K_1(m,n)$$

which is essentially equivalent to (7,4) and the theorem is proved. The bound on the interpolation error follows from the observation that the error vanishes at the points of $\beta(m,n)$ on the assumption that the function being interpolation has bounded derivatives $(\leq M_{m+1})$ of order $m+1$. Indeed, if the error is $\varepsilon(x)$ then by (7,4) we have:

$$|\varepsilon(x)| \leq K_2(m,n) M_{m+1} h^{m+1} \qquad (7,11)$$

8. Error of derivative interpolation

The result of this section is based on the differentiation of equation (7,3), followed by estimation of the resulting expressions. It is convenient to carry this out in stages. We shall use the following

notations for the derivatives:

$$\frac{\partial^t f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} = D_\alpha^t f \qquad |\alpha| = t \tag{8,0}$$

and the values of f at the lattice points p_i will be denoted by f_i. The following lemma is required.

Lemma 8.1

Under the assumptions of theorem 7.1 there holds the estimate

$$\left| \frac{\partial R_{m+1,i}(\bar{p})}{\partial x_j} \right| \leq C_3(m,n) M_{m+1} h^m, \qquad i = 1,2,\ldots,\binom{m+n}{m};\ j = 1,2,\ldots,n$$

where $C_3(m,n)$ is independent of $\bar{S}_n$ and $\bar{p}$ is any point of $\bar{S}_n$

Proof

By the definition,

$$R_{m+1,i} = \frac{1}{m!} \int_0^1 \phi_i^{(m+1)}(\lambda)(1-\lambda)^m \, d\lambda \tag{8,1}$$

where

$$\phi_i(\lambda) = f(\bar{p} + \lambda(p_i - \bar{p})) \tag{8,2}$$

and

$$\phi_i^{(k)}(\lambda) = \sum_{|i|=k} \frac{\partial^k f(\bar{p} + \lambda(p_i - \bar{p}))}{\partial x_1^{i_1} \cdots \partial x_n^{i_n}} (p_i - \bar{p})_1^{i_1} \cdots (p_i - \bar{p})_n^{i_n} \tag{8,3}$$

Integrating (8,1) by parts,

$$R_{m+1,i} = -\frac{1}{m!} \phi_i^{(m)}(0) + \frac{1}{(m-1)!} \int_0^1 \phi_i^{(m)}(\lambda)(1-\lambda)^{m-1} \, d\lambda$$

and from (8,3)

$$\phi_i^{(m)}(0) = \sum_{|i|=m} \frac{\partial^m f(\bar{p})}{\partial x_1^{i_1} \cdots \partial x_n^{i_n}} (p_i - \bar{p})_1^{i_1} \cdots (p_i - \bar{p})_n^{i_n}$$

Differentiating with respect to $\bar{x}_j$ and using (8,0)

$$\frac{\partial \phi_i^{(m)}(0)}{\partial \bar{x}_j} = \sum_{|\alpha|=m} \left[D_{\alpha+\varepsilon_j}^{m+1} f(\bar{p})(p_i - \bar{p})^\alpha + D_\alpha^m f(\bar{p})(p_i - \bar{p})^{\alpha-\varepsilon_j}(-\alpha_j) \right] \tag{8,4}$$

where $\alpha \pm \varepsilon_j = (\alpha_1, \ldots, \alpha_j \pm 1, \ldots, \alpha_n)$. Next consider

$$\int_0^1 \phi_i^{(m)}(\lambda)(1-\lambda)^{m-1}d\lambda = \sum_{|\alpha|=m}(\rho_i-\bar{\rho})^{\alpha}\int_0^1 D_\alpha^m f(\bar{\rho}+\lambda(\rho_i-\bar{\rho}))(1-\lambda)^{m-1}\,d\lambda .$$

Differentiating with respect to $\bar{x}_j$

$$\frac{\partial}{\partial \bar{x}_j}\int_0^1 \phi_i^{(m)}(\lambda)(1-\lambda)^{m-1}\,d\lambda = \sum_{|\alpha|=m}\Big[(-\alpha_j)(\rho_i-\bar{\rho})^{\alpha-\varepsilon_j}\int_0^1 D_\alpha^m f($$

$$(\bar{\rho}+\lambda(\rho_i-\bar{\rho}))(1-\lambda)^{m-1}d\lambda + (\rho_i-\bar{\rho})^{\alpha}\int_0^1 D_{\alpha+\varepsilon_j}^{m+1} f(\bar{\rho}+\lambda(\rho_i-\bar{\rho}))(1-\lambda)^{m}\,d\lambda\Big] \qquad (8,5)$$

Consider the 1st member of the above expression,

$$\sum_{|\alpha|=m}(-\alpha_j)(\rho_i-\bar{\rho})^{\alpha-\varepsilon_j}\int_0^1 D_\alpha^m f(\bar{\rho}+\lambda(\rho_i-\bar{\rho}))(1-\lambda)^{m-1}\,d\lambda \qquad (8,6)$$

Integrating by parts in the integral, we find

$$\int_0^1 D_\alpha^m f(\bar{\rho}+\lambda(\rho_i-\bar{\rho}))(1-\lambda)^{m-1}\,d\lambda = \frac{1}{m}D_\alpha^m f(\bar{\rho}) +$$

$$+\frac{1}{m}\int_0^1\Big[\sum_{k=1}^{n}\Big(\frac{\partial}{\partial x_k}D_\alpha^s f(\bar{\rho}+\lambda(\rho_i-\bar{\rho}))\Big)(\rho_i-\bar{\rho})_k\Big](1-\lambda)^m\,d\lambda$$

so that (8,6) is equivalent to

$$\sum_{|\alpha|=m}(-\alpha_j)(\rho_i-\bar{\rho})^{\alpha-\varepsilon_j}\Big\{\frac{1}{m}D_\alpha^m f(\bar{\rho}) + \frac{1}{m}\int_0^1\Big[\sum_{K=1}^{n}\Big(\frac{\partial}{\partial \bar{x}_K}D_\alpha^m f(\bar{\rho}+$$

$$+\lambda(\rho_i-\bar{\rho}))\Big)(\rho_i-\bar{\rho})_K\Big](1-\lambda)^m\,d\lambda\Big\} \qquad (8,7).$$

Collecting together (8,4), (8,5), (8,7), it follows that

$$\frac{\partial R_{m+1,i}}{\partial \bar{x}_j} = -\frac{1}{m!}\sum_{|\alpha|=m}D_{\alpha+\varepsilon_j}^{m+1}f(\bar{\rho})(\rho_i-\bar{\rho})^{\alpha}+\frac{1}{(m-1)!}\sum_{|\alpha|=m}(\rho_i-\bar{\rho})^{\alpha}\int_0^1 D_{\alpha+\varepsilon_j}^{m+1}f($$

$$(\bar{\rho}+\lambda(\rho_i-\bar{\rho}))(1-\lambda)^{m-1}\,d\lambda + \frac{1}{m!}\sum_{|\alpha|=m}(-\alpha_j)(\rho_i-\bar{\rho})^{\alpha-\varepsilon_j}\int_0^1\Big[\sum_{k=1}^{n}\frac{\partial}{\partial \bar{x}_k}D_\alpha^m f(\bar{\rho}+\lambda($$

$$(\rho_i-\bar{\rho}))\Big)(\rho_i-\bar{\rho})_k\Big](1-\lambda)^m\,d\lambda$$

the terms containing derivatives of order m cancelling. It is evident that each of the 3 terms in the above expression is bounded in modulus by expressions of the form $K_i(m,n)\, h^m M_{m+1}$ $\quad i = 1, 2, 3$ and the lemma is proved.

Now let $h_i \equiv$ distance from X_i to its opposite face. The following lemma will be used later.

Lemma 8.2

There exists a constant, $K(m,n)$ independent of $\bar{S}_n$ such that

$$\sum_{i \in B(m,n)} \left| \frac{\partial \ell_i}{\partial x_K} \right| \leq K(m,n) \sum_{i=0}^{n} \frac{1}{h_i} \qquad K = 1, 2, \ldots, n$$

Proof

By (4,4) the point $\bar{x} = \sum_{i=0}^{n} \frac{a_i}{m} X_i$ contained in $B(m,n)$ has associated with it the polynomial

$$\ell_{\bar{x}}(p) = \frac{1}{a_0! \cdots a_n!} \prod_{i=0}^{n} \prod_{j=0}^{a_i - 1} [m\lambda_i(p) - j]$$

For $p \in \bar{S}_n$, $0 \leq \lambda_i(p) \leq 1$, $i = 0, 1, \ldots, n$. Differentiating with respect to x_K

$$\frac{\partial \ell_{\bar{x}}(p)}{\partial x_K} = \sum_{j=0}^{n} \frac{\partial \lambda_j}{\partial x_K} P_j^{(a_j - 1)}(p)$$

where $P_j^{(a_j-1)}$ is a polynomial of degree $(a_j - 1)$ which is bounded in modulus by a quantity dependent only on m & n when $p \in \bar{S}_n$. Then,

$$\left| \frac{\partial \ell_{\bar{x}}}{\partial x_K} \right| \leq K'(m,n) \sum_{j=0}^{n} \left| \frac{\partial \lambda_j}{\partial x_K} \right|$$

Let $S_K^{(j)}$ $\quad K = 1, 2, \ldots, n$ be the direction cosines of the normal to the face of $\bar{S}_n$ opposite X_j. Then

$$\sum_{j=0}^{n} \left| \frac{\partial \lambda_j}{\partial x_K} \right| = \sum_{j=0}^{n} \left| \frac{S_K^{(j)}}{h_j} \right| < \sum_{j=0}^{n} \frac{1}{h_j}$$

so

$$\left| \frac{\partial \ell_{\bar{x}}}{\partial x_K}(p) \right| \leq K'(m,n) \sum_{j=0}^{n} \left| \frac{1}{h_j} \right|$$

and

$$\sum_{B(m,n)} \left| \frac{\partial \ell_i}{\partial x_K}(p) \right| \leq K(m,n) \sum_{j=0}^{n} \frac{1}{h_j}$$

which proves the lemma.

We now prove the main result.

Theorem 8.1

Let $p \in \bar{S}_n$. With the previous notations and assumptions, there holds the estimate

$$\left| \frac{\partial f(p)}{\partial x_K} \right| \leq K_4(m,n) M_{m+1} h^m + \left[K_5(m,n) \max_i |f_i| + K_6(m,n) \Pi_{m+1} h^{m+1} \right] \sum_{i=0}^{n} \frac{1}{h_i}$$

Proof

By theorem 7.1

$$f(p) = \sum_{B(m,n)} \left[f(p_i) - R_{m+1,i} \right] \ell_i(p)$$

and

$$\frac{\partial f(p)}{\partial x_K} = \sum_{B(m,n)} \left[- \frac{\partial R_{m+1,i}}{\partial x_K} \right] \ell_i(p) + \sum_{B(m,n)} \left[f(p_i) - R_{m+1,i} \right] \frac{\partial \ell_i(p)}{\partial x_K}$$

so that

$$\left| \frac{\partial f(p)}{\partial x_K} \right| \leq \max_i \left| \frac{\partial R_{m+1,i}}{\partial x_K} \right| \cdot \sum_{B(m,n)} |\ell_i(p)| + \left(\max_i |f_i| + \max_i |R_{m+1,i}| \right) \sum_{B(m,n)} \left| \frac{\partial \ell_i(p)}{\partial x_K} \right|$$

Since $\sum_{B(m,n)} |\ell_i(p)|$ can be estimated in terms of m, n only, using lemma 8.1 to estimate the 1st term and using (7,9) and (7,10) and lemma 8.2 to estimate the second,

$$\left| \frac{\partial f(p)}{\partial x_K} \right| \leq K_4(m,n) M_{m+1} h^m + \left[K_5(m,n) \max_i |f_i| + K_6(m,n) M_{m+1} h^{m+1} \right] \sum_{i=0}^{n} \frac{1}{h_i}$$

and this is the assertion of the theorem.

Applying this to the error of interpolation $(f(P_i) = 0)$ there follows,

$$\left| \frac{\partial f(P)}{\partial x_K} \right| \leq K_4(m,n) M_{m+1} h^m + K_7(m,n) M_{m+1} h^{m+1} \sum_{i=0}^{n} \frac{1}{h_i} \qquad (8,8)$$

If the maximum side length of $\bar{S}_n \to 0$ in such a way that e.g. $h_i = k_i h$ where k_i does not depend on h we can get the asymptotic error formula

$$\left| \frac{\partial f(P)}{\partial x_K} \right| = O(h^m) \qquad K = 1, 2, \ldots, n$$

whilst we have seen previously that

$$|f(P)| = O(h^{m+1})$$

independently of the manner in which $\bar{S}_n \to 0$.

In the case $m = 1, 2$ similar estimates were given in [17]. The factor $\sum_{i=0}^{n} \left| \frac{1}{h_i} \right|$ is there replaced by $\frac{1}{h'}$ where h' is the radius of the inscribed sphere of $\bar{S}_n$ We can easily prove that

$$\sum_{i=0}^{n} \frac{1}{h_i} \leq (n+1) \frac{1}{h'}$$

so that we can use h' in the estimates instead of $\sum_{i=0}^{n} \frac{1}{h_i}$, if necessary.

9. Additional remarks on the interpolation theory.

The interpolation theory given in sections 2,3,4,7 and 8 was developed by the author at various times during 1969/70/71. During this period, various results which are mostly special cases of the results given here appeared in the literature. In the present section we shall indicate the relationship between some of these papers and the present work. The papers to be considered are all associated with the finite

element method or motivated by its results.

The first paper which was essentially a mathematical paper on the finite element method appears to be [18] . This well known and widely quoted paper is concerned with the case of 2 space dimensions, $n = 2$. The results that concern us here are that error formulae of the type $(7,11)$ and $(8,8)$ were derived for polynomials of degree $m = 1$ and $m = 2$. The existence and uniqueness of polynomials taking on prescribed values at the points of $\delta(m,2)$ $m = 1,2$ were demonstrated. Such a result is, however, deficient in a certain respect, namely it does not imply the following: if the data (prescribed values) are those of a polynomial of degree $m = 1$ or 2 then the interpolating polynomial coincides with the polynomial from which the data were derived. Let us call this the reproducing property of the interpolating polynomial. The reproducing property does in fact follow from the error formulas of Zlamal, which were established by him without recourse to the reproducing property. This follows on account of the fact that the error bound contains multiplying the factor M_2 (M_3) which would be zero for $m = 1$ $(m = 2)$

The approach followed by us was that the reproducing property was established first, being a consequence of the material in section 2. It was then used in sections 7 and 8 in the derivation of the error bounds.

A more recent paper [17] generalizes the results of Zlamal. The relevant results of these authors are that the existence and uniqueness result for the cases n arbitrary, $m = 1, 2$ Again without using the reproducing property, the appropriate error bounds are obtained, both on

the function and derivative interpolation. The method used bears a superficial resemblance to the mthod used here, but it is actually rather different as the reproducing property is not available. This latter paper contains a number of interesting references.

10. Denseness of trial functions

In the present paragraph we shall consider the trial functions, whose construction has been explained, in the light of the results of Part 1 section 7.

Let $\bar{\Omega}$ be given and let it satisfy the conditions set out in section 6. It will be assumed that an imposed boundary condition $\phi(\partial\Omega) = 0$ is to be satisfied.

Let T_1 and T_2 be 2 triangulations of $\bar{\Omega}$. We shall say that $T_1 \subset T_2$ if T_2 is obtained by triangulating every component simplex of T_1. If $\bar{\tau}_{11}$ and $\bar{\tau}_{12}$ are any 2 simplexes of T_1, it is not required that they should be triangulated according to the same rule in the passage from T_1 to T_2. Now let a sequence of triangulations $\{T_i\}$ be given, such that

$$T_1 \subset T_2 \subset T_3 \text{ ------}$$

With each of these triangulations T_i we associate a trial function $f^{(i)}(a_1^{(i)}, \ldots, a_{s_i}^{(i)}; x)$ satisfying

(i) $f^{(i)}$ is an m^{th} degree polynomial in each simplex of T_i

(ii) $f^{(i)}$ is of class $C^{(0)}(\bar{\Omega})$

(iii) $f^{(i)}$ vanishes on $\partial\Omega$

The parameters in the trial function are the $a_j^{(i)}$ and m is a given fixed integer.

The following result amounts to a verification of the fact that if the family of functions $\mathcal{F}_i$ of Part 1 section 7 is identified with the family $f^{(i)}$ defined above, then

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots\cdots$$

in the sense of that section. Thus, consecutive 'trial functions' are obtained by refining the triangulation in the manner described above.

Lemma 10.1

Let $a_j^{(i)}$ $j = 1, 2, \ldots, s_i$ be any given numbers. Let $\ell \geq i$ Then $\exists$ a set of numbers $a_j^{(\ell)}$ $j = 1, 2, \ldots, s_\ell$ such that

$$f^{(\ell)}(a_1^{(\ell)}, \ldots, a_{s_\ell}^{(\ell)}; x) = f^{(i)}(a_1^{(i)}, \ldots, a_{s_i}^{(i)}; x)$$

whenever $x \in \bar{\Omega}$.

Proof

Consider any simplex of T_i, say $\bar{\tau}_{ij}$. In it, $f^{(i)}$ reduces to an m^{th} degree polynomial $Q_{ij}^{(m)}(x)$ which can be expressed in terms of its values at the points of $\mathcal{B}(m,n)$ in $\bar{\tau}_{ij}$. Let $\bar{\tau}_{ij}$ be triangulated in any manner, into simplexes $\bar{\tau}_{ijk}$, $k = 1, 2, \ldots, t$. Constructing in each of the $\bar{\tau}_{ijk}$ its principal lattice $\mathcal{B}_k(m,n)$ we can represent $Q_{ij}^{(m)}(x)$ in terms of its values at $\mathcal{B}_k(m,n)$ in $\bar{\tau}_{ijk}$. Now $T_\ell \supset T_i$ and is obtained from it by triangulating the components of T_i. Therefore, those $a_i^{(\ell)}$'s associated with points in $\bar{\tau}_{ij}$ are simply the values of $Q_{ij}^{(m)}$ at these points. Both $f^{(\ell)}$ and $f^{(i)}$ satisfy the imposed

boundary condition, by definition, and the existence of the numbers

$a_j^{(i)} \quad j = 1, 2, \ldots, s_\varepsilon$ is proved.

It is evident that this set of numbers is unique.

Next we shall attend to the question of the denseness of the collection of functions $f^{(i)}$ in the metric space (B, σ') defined below. The need for this result arises in Theorem 7.1 of Part 1. At this point, in order to avoid circumlocution, we shall make use of certain results of higher analysis that we have been able to avoid so far. The introduction of these results can actually be 'avoided', although in order to do so they have actually to be proved on an ad hoc basis as the work proceeds.

For the set B we shall take the class of functions u continuous on $\bar{\Omega}$ taking the value zero on $\partial\Omega$ and having piecewise continuous 1st derivatives in $\bar{\Omega}$. For the metric σ' we shall take the expression

$$\sigma'(u,v) = \left(\int_\Omega |u-v|^2 \, d\Omega + \sum_{i=1}^{n} \int_\Omega |u_{x_i} - v_{x_i}|^2 \, d\Omega \right)^{1/2}$$

We have to prove that for any fixed $u \in B$ and $\varepsilon > 0$ there is an i such that the family $f^{(i)}$ (of course, $f^{(i)} \in B$) contains a member $\bar{f}^{(i)}$ such that $\sigma'(u, \bar{f}^{(i)}) < \varepsilon$

To prove this, we shall make use of the known result that for any u in (B, σ') and $\varepsilon' > 0$ there is a function $u_{\varepsilon'}$ infinitely differentiable in Ω and vanishing in a boundary layer (whose width depends on ε') such that $\sigma'(u, u_{\varepsilon'}) < \varepsilon'$. We shall not prove this, but simply note that the function $u_{\varepsilon'}$ can be constructed

by an averaging operation on u. The details can be found in [23].

Our result now follows by interpolation of the function $u_{\varepsilon'}$. Denoting the interpolant by $\bar{u}_{\varepsilon'}$, in any simplex we have the (pointwise) estimates (7,11) and (8,8)

$$|u_{\varepsilon'} - \bar{u}_{\varepsilon'}| \leq K_1 M_{m+1} h^{m+1} \tag{10,1}$$

$$|u_{\varepsilon', x_i} - \bar{u}_{\varepsilon', x_i}| \leq K_2 M_{m+1} h^m + K_3 M_{m+1} h^{m+1} \sum_{i=1}^{n} \frac{1}{h_i}$$

where M_{m+1} is an upper bound on the $(m+1)$th derivatives of $u_{\varepsilon'}$ and the K_i are not dependent on the particular simplex. Without loss of generality we can suppose the estimates (10,1) to be written for those simplexes for which the right hand sides are as great as possible. Now let the greatest linear dimension h of all the simplexes go to zero in such a way that in all of the numbers h_i go to zero like $c.h$ where $c < 1$ is a constant. Then integrating (10,1) we can say that

$$\sigma(u_{\varepsilon'}, \bar{u}_{\varepsilon'}) \leq K h^m$$

where K is independent of the triangulation. Therefore,

$$\begin{aligned} \sigma(u, \bar{u}_{\varepsilon'}) &\leq \sigma(u, u_{\varepsilon'}) + \sigma(u_{\varepsilon'}, \bar{u}_{\varepsilon'}) \\ &\leq \varepsilon' + K h^m \\ &\leq \varepsilon \end{aligned}$$

for a fine enough triangulation. Since $\bar{u}_{\varepsilon'}$ is a member of $f^{(i)}$ for some i our result is proved.

2.11 Resume and notes

In the part just finished we have constructed an interpolation theory of some utility. Sections 1 - 5 were concerned with what we might call the algebraic aspect of the theory, whilst sections 6 - 9 were concerned rather more with analytic problems relating to the various interpolation errors.

We shall make some additional remarks on this material, firstly on the algebraic aspect. This theory can be presented in another form: the starting point of it is the representation of the interpolating polynomial given in Section 4. We have then $\binom{m+n}{m}$ functions $\ell_i(x)$ satisfying the condition

$$\ell_i(x_j) = \delta_{ij}$$

where x_j are the lattice points. It is easy to prove on the basis of (11,1) that the functions $\ell_i(x)$ $i=1, 2, \ldots, \binom{m+n}{m}$ are linearly independent and from this soon follows that any function f has a unique interpolant

$$\bar{f} = \sum_{B(m,n)} f_i \ell_i(x)$$

Moreover, from the linear independence of the $\binom{m+n}{m}$ functions $\ell_i(x)$ follows the fact that they generate a finite dimensional vector space of dimension $\binom{m+n}{m}$. We recall at this point that $\ell_i(x)$ is an m^{th} degree polynomial. Now what we should like to conclude is that the class of all polynomials of degree m (in R_n) is also a vector space of dimension $\binom{m+n}{m}$ for in that case every member of it could be generated by varying the numbers p_i in the expression $\sum_{B(m,n)} p_i \ell_i(x)$. However, the linear independence of the $\ell_i(x)$ is not, by itself, a powerful

enough result to enable this to be proved. The 'reproducing property', proved by Theorem 3.1 is the required result. In fact it can be looked at as a statement of the equivalence of the dimension of the space spanned by the $\ell_i(x)$ and the dimension of the space of M^{th} degree polynomials. Thus, a theory which starts with the functions $\ell_i(x)$ still has to prove a theorem like 3.1. It should also be mentioned that the results of sections 1 - 4, and section 7 are given in a forthcoming paper [19] This material was also given in [20]. Secondly, we shall add some remarks about the analytic aspects. The most awkward part of these results was the differentiation of the remainder term in (8.1.) It is possible by a slight modification of the present arguments to replace this differentiation with a simpler one. Since the author was not aware of this at the time the present results were constructed it was thought best to adhere to the original treatment. The modified treatment was communicated to the author by G. Strang, and as far as is known will appear in a forthcoming book by him on the finite method.

The results of sections 7 and 8 form the contents of a paper [21] submitted for publication. This material was also given in [22]

Part 3 : Derivation and properties of the algebraic systems

3.1 Introduction

In part 3 we investigate in detail the systems of algebraic equations (generally not linear) resulting from the use of the piecewise polynomial trial functions of part 2 in the Ritz method applied to the problem set out in part 1 section 9.

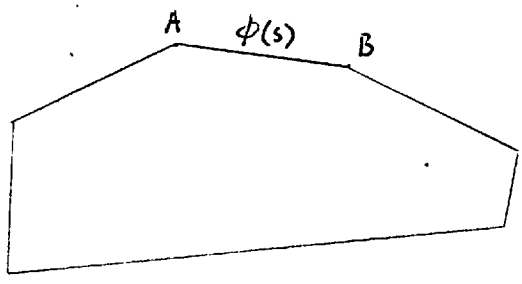
Section 2 consists of simple generalisations of results in wide use (e.g. [24], [25]) for the case of a quadratic functional, $\Omega \subset R_2$ (or perhaps R_3), and polynomials of low degree. Section 3 contains results on the existence and uniqueness of solutions to the algebraic problems, and section 4 contains an explicit derivation of the system of nonlinear equations arising in a 'model' problem.

Let Ω be triangulated. We shall impose a condition on this triangulation in the situation where both 'natural' and 'essential' (or imposed) boundary conditions are present. Namely, it is required that if $\bar{S}_{n-1}$ is a simplex of the boundary complex, then the boundary condition in S_{n-1} is <u>either</u> natural <u>or</u> essential, but not both.

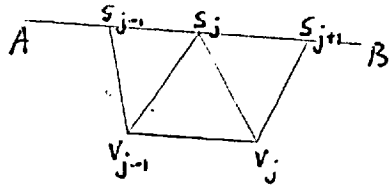
It is also necessary to say something about constraining the trial functions to satisfy the essential boundary condition ϕ. Suppose to begin with that ϕ is, in each boundary simplex, a polynomial of degree $\leq m$. Then from the interpolation theory it follows that if trial functions of degree m are used, and their values at the boundary points of the principal lattices taken to be the values of ϕ at these points, then the trial functions will satisfy the essential boundary condition at

all points where ϕ is defined. On the other hand if ϕ is not a polynomial we are confronted with the problem of making the trial function satisfy the boundary condition. To circumvent this difficulty we may suppose that a function ϕ_0 defined in $\bar{\Omega}$ and coinciding with ϕ can be found. If $P^{(m)}$ is a trial function vanishing on $\partial\Omega$, then $\phi_0 + P^{(m)}$ can be used as a trial function satisfying the imposed boundary condition.

It must be pointed out however that finding a ϕ_0 is generally not easy. Although the following idea has not been tried out in practice it is always possible, using the interpolation theory, to construct a piecewise polynomial function, defined on $\bar{\Omega}$ vanishing <u>outside</u> of a boundary layer which approximates an arbitrarily piecewise smooth boundary function with any prescribed accuracy. The basis of this procedure can best be illustrated with a simple example. Let $n=2$ and consider the region Ω shown below:



Let $\phi(s)$ be given on AB where s is the distance from A to B; $s_A = 0$, $s_B = 1$ say. Divide AB into $p+1$ equal parts at points $s_A, s_1, \dots, s_p, s_B$; consider the intervals $[s_{j-1}, s_j]$ $[s_j, s_{j+1}]$,



and construct the triangles $S_{j-1}V_{j-1}S_j$, $S_jV_{j-1}V_j$ and $S_jV_jS_{j+1}$. Let m be any integer >0. In each of these triangles construct the lattices $B(m,2)$. We can now define in each triangle an m^{th} degree polynomial taking the values zero at points of $B(m,n)$ not on AB and the values of ϕ at points of $B(m,n)$ on AB. The resulting piecewise polynomial defined in the quadrilateral $S_{j-1}V_{j-1}V_jS_{j+1}$ has the following properties: a) it is continuous b) it reduces to an m^{th} degree polynomial in $S_{j-1}S_j$ and to another m^{th} degree polynomial in S_jS_{j+1} c) it approximates ϕ in these 2 intervals with an accuracy $O\left(\frac{1}{\rho+1}\right)^{m+1}$

The construction of ϕ_0 is now clear: we construct a small boundary layer whose inner boundary is also polygonal and triangulate it reasonably. In each triangle we take an m^{th} degree polynomial as explained above. We then arbitrarily triangulate the inner polygon and formally define m^{th} degree polynomials identically zero in each triangle. The resulting piecewise polynomial function, ϕ_0 has all the properties we set out to achieve.

Some remarks are in order. Firstly, since ϕ is known on AB its derivatives are also known. This may be useful in deciding how big to take ρ and m. Secondly, let us remark that the above procedure is logically independent of the actual solution of the minimization problem. In particular, the above triangulation and the triangulation of Ω for the purpose of minimization are logically independent, as are the degrees of the polynomials used for the two procedures. (However, if they can be combined, certain computational economies will follow.)

In the future, unless indicated to the contrary it will be assumed that a ϕ_0 is known and it will be added into the trial function as shown previously.

Finally, we shall mention some special nomenclature it will be convenient to use in this part. This is the following: let τ_K be a simplex in a triangulated region. With τ_K will be associated an m^{th} degree polynomial trial function containing, for an interior simplex $\binom{m+n}{m}$ parameters (coefficients) $a_i^{(k)}$ $i = 1, 2, \ldots, \binom{m+n}{m}$. An adjacent simplex τ_j similarly will have associated with it parameters $a_i^{(j)}$ $i = 1, 2, \ldots, \binom{m+n}{m}$. Some of the $a_i^{(k)}$ will coincide with some of the $a_i^{(j)}$ since the trial function is continuous in $\tau_j \cap \tau_K$. We shall call the parameters $a_i^{(j)}$ $i = 1, 2, \ldots, \binom{m+n}{m}$ <u>local</u> parameters (they are local to τ_j). On the other hand, the entire trial function will be dependent on, say, M parameters, $a_1, \ldots, a_M$. Some of these are the $a_i^{(j)}$ and $a_i^{(k)}$. In the context of the entire (or globally defined) trial function, they will be referred to as <u>global</u> parameters.

3.2 General form of the equations

The present section contains a derivation of the general form of the algebraic equations.

We consider the functional

$$J(u) = \int_{\Omega} G(|\nabla u|^2) - f(x)u \, d\Omega \,, \qquad G(|\nabla u|^2) = \int_c^{|\nabla u|^2} g(t)\, dt \tag{2,1}$$

along with

$$u|_{\partial\Omega_1} = \phi \;;$$

the regions $\bar{\Omega}$ allowed here are those specified in part 2, section 6.

Let T be a triangulation of $\bar{\Omega}$, and let τ_o be any simplex in the triangulation. Let any positive integer m be given: this is the degree of the polynomials that will be used. Into τ_o, we introduce the lattice $\beta(m,n)$ and define in $\bar{\tau}_o$ the polynomial $p^{(m)}(x)$ of degree m

$$p^{(m)}(x) = \sum_{\beta(m,n)} a_i^{(o)} \ell_i(x) \qquad x \in \bar{\tau}_o$$

dependent on the local parameters $a_i^{(o)}$. The functions $\ell_i(x)$ were defined in Part 2, section 4. If $\bar{\tau}_o$ has nonzero intersection with $\partial\Omega_1$ those parameters associated with lattice points lying in $\partial\Omega_1$ are to be set to zero as indicated in para.1. Let ϕ_o be defined on $\bar{\Omega}$ and satisfy the imposed boundary condition and let $\phi_o^{(o)}$ be the restriction to $\bar{\tau}_o$ of ϕ_o The local trial function for is given by

$$u^{(o)}(x) = \phi_o^{(o)}(x) + p^{(m)}(x)$$

so that

$$\nabla u^{(o)}(x) = \nabla \phi_o^{(o)}(x) + \nabla p^{(m)}(x)$$

$$= \nabla \phi_o^{(o)}(x) + \sum_{\beta(m,n)} a_i^{(o)} \nabla \ell_i(x)$$

$$|\nabla u_o|^2 = \left| \nabla \phi_o^{(o)}(x) + \sum_{\beta(m,n)} a_i^{(o)} \nabla \ell_i(x) \right|^2$$

$$= \sum_{j=1}^{n} \left\{ \phi_{o,j}^{(o)} + \sum_{\beta(m,n)} a_i^{(o)} \ell_{i,j}(x) \right\}^2 \quad ; \left(\phi_{o,j}^{(o)} = \frac{\partial \phi_o^{(o)}}{\partial x_j} \right)$$

therefore,

$$\frac{\partial}{\partial a_\kappa^{(o)}} |\nabla u^{(o)}|^2 = \sum_{j=1}^{n} 2 \left\{ \phi_{o,j} + \sum_{\beta(m,n)} a_i^{(o)} \ell_{i,j} \right\} \ell_{\kappa,j}$$

Put $\phi_K^{(0)}(x) = \sum_{j=1}^{n} \phi_{0,j}\, \ell_{K,j}(x)$ so that

$$\frac{\partial}{\partial a_K^{(0)}} |\nabla u^{(0)}|^2 = 2 \sum_{j=1}^{n} \sum_{\beta(m,n)} a_i^{(0)}\, \ell_{i,j}(x)\, \ell_{K,j}(x) + \phi_K^{(0)}(x) \qquad (2,2)$$

Now consider

$$J_{\tau_0}(u^{(0)}) = \int_{\tau_0} \int_{c}^{|\nabla u^{(0)}|^2} \rho(t)\, dt - f u^{(0)}\, d\tau_0$$

Differentiating with respect to a_K in this,

$$\frac{\partial J_{\tau_0}}{\partial a_K} = \int_{\tau_0} \rho(|\nabla u^{(0)}|^2)\, \frac{\partial}{\partial a_K} |\nabla u^{(0)}|^2 - f \frac{\partial u^{(0)}}{\partial a_K}\, d\tau_0$$

and substituting (2,2) into it,

$$= \int_{\tau_0} \rho_0 \left[2 \sum_{j=1}^{n} \sum_{\beta(m,n)} a_i^{(0)}\, \ell_{i,j}(x)\, \ell_{K,j}(x) + 2 b_K^{(0)} \right] - f \ell_K(x)\, d\tau_0$$

where $\rho_0 = \rho(|\nabla u^{(0)}|^2)$

$$= 2 \sum_{\beta(m,n)} a_i^{(0)} \int_{\tau_0} \rho_0 \sum_{j=1}^{n} \ell_{i,j}\, \ell_{K,j}(x)\, d\tau_0 + \int_{\tau_0} 2 b_K^{(0)} \rho_0 - f \ell_K(x)\, d\tau_0 .$$

We observe immediately that this expression is in the form

$$\frac{\partial J_{\tau_0}}{\partial a_K^{(0)}} = \sum_{\beta(m,n)} \omega_{iK}(a^{(0)})\, a_i^{(0)} - f_K \qquad (2,3)$$

where $a^{(0)} = (a_1^{(0)}, a_2^{(0)}, \ldots, a_N^{(0)})^T$ $\quad N = \binom{m+n}{m}$.

$$\tfrac{1}{2} \omega_{iK}(a^{(0)}) = \int_{\tau_0} \rho_0 \sum_{j=1}^{n} \ell_{i,j}\, \ell_{K,j}\, d\tau_0 = \int_{\tau_0} \rho_0 (\nabla \ell_i, \nabla \ell_K)\, d\tau_0 \qquad (2,4)$$

and

$$f_K = \int_{\tau_0} 2 b_K^{(0)} \rho_0 - f \ell_K(x)\, d\tau_0 \qquad (2,5)$$

Suppose that τ_0 is an interior simplex of the triangulation. Then in (2,3) $K = 1, 2, \ldots, N$ $\quad N = \binom{m+n}{m}$, and using the notation $\nabla_{a^{(0)}} J_{\tau_0}$ to denote the vector $\left(\frac{\partial J_{\tau_0}}{\partial a_1^{(0)}}, \ldots, \frac{\partial J_{\tau_0}}{\partial a_N^{(0)}} \right)$, (2,3) is the (nonlinear) matrix equation

$$\nabla_{a^{(0)}} J_{\bar{\tau}_0} = W(a^{(0)}) \cdot a^{(0)} - F^{(0)}$$

where $[W(a^{(0)})]_{ij} = w_{ij}(a^{(0)})$ and $F^{(0)} \equiv (f_1, \ldots, f_N)^T$. In the finite element method, (when usually W = constant $\rightarrow W^{(0)}$) $W^{(0)}$ is called the stiffness matrix of the element (simplex) $\bar{\tau}_0$. Continuing with the case of an interior simplex $\bar{\tau}_0$ we can easily see that no matter what values are assigned to the $a_i^{(0)}$ the matrix $W(a^{(0)})$ is necessarily singular. To prove this, consider (2,4) and

$$\sum_{j=1}^{N} w_{ij}(a^{(0)}) = \int_{\tau_0} \rho_0 \left(\nabla \ell_i(x), \nabla \sum_{j=1}^{n} \ell_j(x) \right) d\tau_0 \qquad i = 1, 2, \ldots, N$$

$$= 0, \quad \text{since} \quad \sum_{j=1}^{N} \ell_j(x) = 1 \qquad (2,5a)$$

from which it immediately follows that $\det W(a^{(0)})$ is zero. We shall mention that the singularity of $W(a^{(0)})$ is bound up with the possibility of making translations and rotations of the domain $\bar{\tau}_0$ along and about the axis of the dependent variable u. (These are "rigid body"-or, "Newtonian" motions). This result holds for all m and n. On the other hand, if $\bar{\tau}_0$ intersects $\partial\Omega_1$, in general $W(a^{(0)})$ is not singular; this can be explained by the remark that $\bar{\tau}_0$ cannot now have Newtonian motions imparted to it, without an alteration in the boundary conditions.

In principle, (2,4) and (2,5) are all the results needed for the minimization of (2,1) with respect to the (global) parameters $a_1, a_2, \ldots, a_M$.

To see this, we return to (2,1) which can be written in the form

$$J(u) = \sum_{i=1}^{S} \int_{\bar{\tau}^{(i)}} G(|\nabla u|^2) - f(x)u \, d\tau_i = \sum_{i=1}^{S} J_{\tau_i}(u) \qquad (2,6)$$

where $\bar{\Omega} = \bigcup_{i=1}^{S} \tau^{(i)}$. In the manner previously described for $\bar{\tau}^{(i)}$ a trial function $u^{(i)}$ is constructed in $\bar{\tau}^{(i)}$ for each $i = 1, 2, \ldots, S$. The continuous global trial function, denoted by u_M reduces to $u^{(i)}$ whenever $x \in \tau^{(i)}$ and depends globally on parameters $a_1, a_2, \ldots, a_M$. Then from (2,6) there follows

$$J(u_M) = \sum_{i=1}^{S} J_{\tau_i}(u^{(i)}) = \psi_M(a_1, \ldots, a_M) \qquad (2,6a)$$

To minimize $J(u_M) \equiv \psi_M(a_1, \ldots a_M)$ with respect to the parameters we have to solve the system

$$\frac{\partial \psi_M}{\partial a_K} = 0 \qquad K = 1, 2, \ldots, M$$

These equations, written in terms of (2,6) are the same as

$$\sum_{i=1}^{S} \frac{\partial J_{\tau_i}}{\partial a_K} = 0 \qquad K = 1, 2, \ldots, M \qquad (2,7)$$

Now assume that the global parameter a_K is local to simplexes $\tau_{t_1} \ldots \tau_{t_K}$. Then the only nonzero terms in sum (2,7) are those in the form

$$\frac{\partial J_{\tau_i}}{\partial a_K} \qquad i = t_1, t_2, \ldots, t_K \qquad (2,8)$$

and the equation (2,7) takes the form

$$\sum_{i=t_1}^{t_K} \frac{\partial J_{\tau_i}}{\partial a_K} = 0 \qquad K = 1, 2, \ldots, M \qquad (2,9)$$

It will now be supposed that the function ρ depends only on x, or is constant. It follows from (2,3) – (2,5) that (2,9) is a system of equations, linear in the a_K's. It can be written in

the form of a matrix equation, say

$$\bar{W}\, a = \bar{F} \qquad (2,10)$$

where the elements of $\bar{W}$ are various expressions of the type (2,4) $\bar{W}$ is an $M \times M$ matrix.

It is clear that in the nonlinear case, with $\rho = \rho(|\nabla u|^2)$ a system of equations of the type

$$\bar{W}(a).a = \bar{F}(a) \qquad (2,10a)$$

will result from the above operations. $\bar{W}(a)$ is again an $M \times M$ matrix, whose elements are integrals (which depend on a) of the form (2,4). The right hand side is also dependent on a.

Returning to the linear case, we can say that from the stage (2,3) – (2,8) onwards, the problem of the minimization of (2,6) is reduced to a programming problem (albeit, a complex one). The programming problem has 3 clearly distinguishable aspects;

a) computation of the coefficients of $\bar{W}$ from (2,4)

b) organizational problems concerning the interrelationships between the local and global parameters

c) solution of the algebriaic equations.

In the nonlinear case, $\rho = \rho(|\nabla u|^2)$ we cannot compute the coefficients once and for all, since they are dependent on the parameters. In general, they have to be continually re-computed, with a frequency determined by the particular algorithm chosen for the solution of the nonlinear equations.

We shall now indicate a useful simplification that can be made in the case of a quadratic functional: (a quadratic functional arises if φ is not dependent on $|\nabla u|^2$ but at most dependent on x the space variables). For the case $n=2$, $m=3$ it is a known device [26] and the result given is obtained as a straightforward generalisation.

This is a device applicable when m and n are such that $\beta(m,n)$, for any given simplex $\bar{\tau}_K$, contains points interior to $\bar{\tau}_K$. For example with $n=2$, $m=3$ $\beta(m,n)$ contains the centroid of $\bar{\tau}_K$. Similarly, with $n=3$, $m=4$ $\beta(m,n)$ contains the centroid of $\bar{\tau}_K$ and in general, it is easy to prove that in R_n the lowest degree of polynomials for which $\beta(m,n)$ has interior points is $m=n+1$. The simplification is based on the observation that global parameters associated with interior points can be put into biunique correspondence with local parameters associated with interior points. This means that the equations (2,9) for interior points are obtained as the derivatives of exactly one of the J_{τ_i}. In other words, the sum (2,9) reduces to a single term.

Let us then consider m,n to be such that each simplex has interior points in its lattice $\beta(m,n)$. Consider any simplex, say τ_0 which, for simplicity we take to be an interior simplex of the triangulation. We agree to list the local parameters in such a way that all parameters associated with lattice points on the boundary appear first in the list and are followed by parameters associated with interior lattice points. From (2,3) there holds the equation

$$\frac{\partial J_{\tau_0}}{\partial a_K^{(\omega)}} = \sum_{\beta(m,n)} \omega_{iK}\, a_i^{(\omega)} - f_K$$

which is written in matrix form as

$$\nabla_{a^{(0)}} J_{\tau_0} = W^{(0)} a^{(0)} - F^{(0)}$$

Let $a^{(0)}$ be partitioned into two blocks, the first of them consisting of boundary parameters and the second consisting of interior parameters. In an obvious notation then,

$$\left[\begin{array}{c|c} W_{11}^{(0)} & W_{12}^{(0)} \\ \hline (W_{12}^{(0)})^T & W_{22}^{(0)} \end{array}\right] \left[\begin{array}{c} a_B^{(0)} \\ \hline a_I^{(0)} \end{array}\right] = \left[\begin{array}{c} F_B^{(0)} \\ \hline F_I^{(0)} \end{array}\right]$$

from which follows

$$a_I^{(0)} = (W_{22}^{(0)})^{-1} F_I^{(0)} - (W_{22}^{(0)})^{-1} (W_{12}^{(0)}) a_B^{(0)} \qquad (2,11)$$

The possibility of this equation depends upon the possibility of inverting the matrix $W_{22}^{(0)}$. The fact that the inverse exists is a consequence of a more general existence theorem which will be proved later. Now, (2,11) expresses the interior parameters as linear combinations of the boundary parameters. To use (2,11) to carry out the above mentioned simplification, every equation in the set (2,10) associated with points of lattices inside simplexes are simply omitted from (2,10) whilst wherever interior parameters appear in equations associated with lattice boundary points their values in terms of boundary parameters, determined uniquely from (2,11) are substituted. The order of the matrix in (2,10) will be substantially reduced by this device. An idea of the magnitude of this reduction may be gauged from the following. We take for $\bar{\Omega}$ the square $0 \le x, y \le 1$

A triangulation of $\bar{\Omega}$ is established by dissecting $\bar{\Omega}$ into smaller squares by means of equally spaced lines drawn parallel to the coordinate axes, and then drawing in each square its diagonal of positive slope. Some notations will now be introduced: let P be the number of points on the x axis where the vertical mesh lines intersect it (including $x=0$ and $x=1$). Denote by V_I and V_B the numbers of vertices of the triangles lying respectively in Ω and on $\partial\Omega$, and let S_I and S_B denote the number of sides of the triangles lying respectively in Ω and on $\partial\Omega$. Let T be the number of triangles; m as usual, represents the degree of the polynomials and M represents the number of global parameters (we suppose for simplicity that ϕ is given on all $\partial\Omega$ as a boundary condition). By the methods of e.g. [27] it is easy to prove the following:

$$2S_I + S_B = 3T$$

$$2V_I + V_B = T+2$$

$$V_B = S_B$$

and

$$T = 2(P-1)^2$$

$$M = (3P-5)(P-1)(m-1) + \binom{m-1}{2}T + V_I$$

The first group of these are geometrical relationships and are independent of m. In the last equation above, the number $\binom{m-1}{2}T$ gives the total number of parameters which can be eliminated using $(2,11)$. Denoting this number by M_e we can easily show that

$$\frac{M_e}{M} = \frac{2\binom{m-1}{2}(P-1)^2}{(3P-5)(P-1)(m-1) + 2\binom{m-1}{2}(P-1)^2 + (P-2)^2}$$

For $m = 3$ we have

$$\frac{M_e}{M} = \frac{2(p-1)^2}{2(3p-5)(p-1) + 2(p-1)^2 + (p-2)^2}$$

and for $p = 5$, $\frac{M_e}{M} = \frac{32}{121} \triangleq .26$ and for $p = 10$, $\frac{M_e}{M} \triangleq .24$
For $m = 4$ the corresponding figures are $\frac{M_e}{M} \triangleq .43$
and $\frac{M_e}{M} \triangleq .40$. Thus, when $m = 3$ we get about a 25% reduction in the number of variables. This effects a worthwhile saving in computer time at the expense of more complex programming. It should be emphasised that these figures are for a quadratic functional.

Finally, one more observation can be made concerning the coefficient matrix in $(2, 10)$. It would obviously simplify the programming operations if the integrals $(2, 4)$ could be explicitly evaluated. In the important case $\rho \equiv 1$ this would have to be done on the basis of the expressions given in part 2 section 4. Unfortunately it has not been able to carry out these computations for arbitrary m and n although attempts were made.

3. Existence and uniqueness of solutions

In the present section the existence and uniqueness of solutions to the equations (whose construction was discussed in the previous section) is considered. We shall prove the following theorem:

Theorem 3.1

Let $\rho(t)$ satisfy the condition

$$\rho(t) \geq \frac{K_1}{(1 + K_2 t)^{1-k_3}} \qquad (3,0)$$

with $K_1 > 0$, $K_2 \geq 0$ and $K_3 \geq \frac{1}{2}\varepsilon$ if $\|f\|_{L_2(\Omega)} = 0$ or $\geq \frac{1}{2}(1+\varepsilon)$ if $\|f\|_{L_2(\Omega)} > 0$ where $\varepsilon > 0$ is arbitrary (but fixed). Let ϕ be prescribed everywhere on $\partial\Omega$ Then equations (2,7) have at least one solution.

Proof

The proof consists of a demonstration of the fact that (2,6a) is an increasing function of $\|a\|_2$ The theorem will follow from this.

We shall make use of the following facts:

1) For any 2 vectors $p, q \in R_M$ there holds

$$\|p\|_2^2 + \|q\|_2^2 - 2\|p\|_2\|q\|_2 \leq \|p+q\|_2^2 \leq \|p\|_2^2 + \|q\|_2^2 + 2\|p\|_2\|q\|_2 \quad . \quad (3,1)$$

2) Let $\alpha_i(x)$ $i = 1, 2, \cdots M$ be linearly independent functions defined and continuous in a region Ω of R_n Let $a \in R_M$. Then the expression

$$|||a|||^2 = \int_\Omega \Big(\sum_{i=1}^{M} a_i \alpha_i(x) \Big)^2 d\Omega$$

is a norm on R_M. There exist constants k, K not dependent on a such that

$$k|||a||| \leq \|a\|_2 \leq K|||a||| \qquad (3,2)$$

With regard to the proofs of these, 1) is obvious whilst 2) may be proved by observing that

$$\Big\| \sum_{i=1}^{M} a_i \alpha_i(x) \Big\|_{L_2(\Omega)} = \int_\Omega \Big(\sum_{i=1}^{M} a_i \alpha_i(x) \Big)^2 d\Omega$$

and recalling that all norms on a finite dimensional space satisfy a relation of the type (3,2).

To prove the theorem, note that the function $U_M(x)$ of (2,6)

(the trial function) can be written in the form

$$u_M(x) = \phi_0 + \sum_{i=1}^{M} a_i \theta_i(x)$$

where each $\theta_i(x)$ vanishes outside certain simplexes of the triangulation, is in the form of an m^{th} degree polynomial inside them, and where the function

$$\dot{u}_M(x) = \sum_{i=1}^{M} a_i \theta_i(x)$$

vanishes on $\partial\Omega$. For each $x \in \Omega$

$$\int_0^{|\nabla u|^2} \rho(t)\,dt = |\nabla u|^2 \rho(t^*) \qquad t^* \in (0, |\nabla u|^2)$$

$$\geq \frac{k_1 |\nabla u|^2}{(1 + k_2 t^*)^{1-k_3}} \tag{3,3}$$

We are going to integrate this inequality presently, but before it can be done we need some more results. From the right hand inequality of follows (with $p \equiv \nabla\phi_0(x)$, $q \equiv \nabla\dot{u}_M(x)$),

$$|\nabla u_M|^2 \leq |\nabla\phi_0|^2 + |\nabla\dot{u}_M|^2 + 2|\nabla\phi_0|.|\nabla\dot{u}_M| \tag{3,4}$$

(where $|\nabla\phi_0|^2 = \sum_{i=1}^{n} \left(\frac{\partial\phi_0}{\partial x_i}\right)^2$ etc.), and by Canchy's inequality

$$|\nabla u_M|^2 \leq |\nabla\phi_0|^2 + c_1 \|a\|_2^2 + c_2 \|a\|_2$$

where c_1, c_2 are constants independent of $\|a\|_2$. Now suppose $k_3 \leq 1$: then (3,3) and (3,4) give

$$\int_0^{|\nabla u_M|^2} \rho(t)\,dt \geq \frac{k_1 |\nabla u_M|^2}{\left[1 + \left(c_0^2 + c_1\|a\|_2^2 + c_2\|a\|_2\right) k_2\right]^{1-k_3}} \tag{3,5}$$

putting $c_0^2 = |\nabla\phi_0|^2$. Next, we shall estimate $\int_\Omega |\nabla u_M|^2\,d\Omega$

from below. By the left inequality of (3,1)

$$|\nabla u_M|^2 \geq c_0^2 + |\nabla \dot{u}_M|^2 - 2|\nabla \phi_0||\nabla \dot{u}_M| \tag{3,6}$$

By Cauchy's inequality we have

$$\int_\Omega |\nabla \phi_0| \cdot |\nabla \dot{u}_M| \, d\Omega \leq \left(\int_\Omega |\nabla \phi_0|^2 \, d\Omega \right)^{1/2} \left(\int_\Omega |\nabla \dot{u}_M|^2 \, d\Omega \right)^{1/2} \tag{3,7}$$

and by Friedrichs inequality (see e.g. [28]) we have

$$\int_\Omega |\nabla \dot{u}_M|^2 \, d\Omega \geq c_3 \int_\Omega \dot{u}_M^2 \, d\Omega \tag{3,8}$$

for c_3 independent of u_M. Integrating (3,6) and using (3,7) and (3,8) we arrive at

$$\int_\Omega |\nabla u_M|^2 \, d\Omega \geq c_0'^2 + c_3 \int_\Omega \dot{u}_M^2 \, d\Omega - c_4 \left(\int_\Omega |\nabla \dot{u}_M|^2 \, d\Omega \right)^{1/2} \tag{3,9}$$

where c_4 depends on $\nabla \phi_0$ and not on $\dot{u}_M$. Clearly,

$$\int_\Omega |\nabla \dot{u}_M|^2 \leq c_5 \|a\|_2^2 \tag{3,10}$$

where c_5 depends on $\nabla \theta_i$ $i = 1, 2, \dots, M$ but not on a. Moreover, in view of 2) above

$$\int_\Omega \dot{u}_M^2 \, d\Omega \geq c_6 \|a\|_2^2 \tag{3,11}$$

since the $\theta_i(x)$ are evidently linearly independent.

Using these results in (3,9) we have

$$\int_\Omega |\nabla u_M|^2 \, d\Omega \geq c_0'^2 + c_3' \|a\|_2^2 - c_4' \|a\|_2 \tag{3,12}$$

where it is sufficient to know that c_i' are not dependent on a. Integrating (3,5)

$$\int_\Omega \int_c^{|\nabla u_M|^2} \rho(t)\,dt\,d\Omega \geq \frac{K_1\left[c_0'^2 + c_3'\|a\|_2^2 - c_4'\|a\|_2\right]}{\left[1 + (c_0^2 + c_1\|a\|_2^2 + c_2\|a\|_2)k_2\right]^{1-k_3}}$$

and so

$$J(u_M) \geq \frac{K_1\left[c_0'^2 + c_3'\|a\|_2^2 - c_4'\|a\|_2\right]}{\left[1 + (c_0^2 + c_1\|a\|_2^2 + c_2\|a\|_2)k_2\right]^{1-k_3}} - \|f\|_{L_2(\Omega)}\left[c_5'\|a\|_2 - \|\phi_0\|_{L_2(\Omega)}\right] \qquad (3,13)$$

where the last term is obtained by Cauchy's inequality.

Let us now consider the behaviour of this expression as $\|a\|_2 \to \infty$. The first term behaves like

$$\frac{\|a\|_2^2}{\left(\|a\|_2^2\right)^{1-k_3}} \sim \|a\|_2^{2k_3}$$

whilst the second behaves like $\|a\|_2 \times \|f\|_{L_2(\Omega)}$. It follows from the definition of k_3 that the composite expression $\to \infty$ as $\|a\|_2 \to \infty$. Thus $J(u_M)$ is increasing. On the other hand, $J(u_M)$ is a continuous function of the point a and is therefore bounded below in every closed sphere $\|a\| \leq R$. Therefore in the case $k_3 \leq 1$ the theorem is proved. If $k_3 > 1$ the same arguments hold, except that we can omit the step (3,5) and use (3,12) directly in (3,3). Thus, in all cases the theorem is proved.

Let us note that it can be extended to cover the case where u_M vanishes only on part of $\partial\Omega$. To do this we have to use a more general form of Friedrichs inequality [28] in place of (3,8); the proof is no different in principle.

It is important to observe that the condition (3,0) is <u>not</u> sufficient to insure the existence of a solution to the analytical problem

of finding (smooth) extrema for $\bar{J}(u)$. In other words, it is possible to obtain solutions to the algebraic problems, even though the analytical problem has no solution. For this reason, care must be exercised in the use of Theorem 3.1.

On the basis of (3,13) we can prove uniqueness of any solution when $\rho = \rho(x)$.

Corollary 3.1

Let ρ be independent of $|\nabla u|^2$. Then equations (2,7) have exactly 1 solution.

Proof

First, note that for $\rho(t)$ independent of t, (3,0) reduces to $\rho \equiv \rho(x) \geq K_1 > 0$ since we can take $K_2 = 0$. Then (3,13) takes the form

$$J(u_M) \geq K_{11} \| a \|_2^2 + K_{12} \| a \|_2 + K_{13} \qquad (3,14)$$

where K_{11}, K_{12} and K_{13} are various combinations of c_i, c_i', K_i and $\| \phi_0 \|_{L_2(\Omega)}$, $\| f \|_{L_2(\Omega)}$.

Second, it is immediately clear that $J(u_M)$ is a quadratic form in the variables $a_1, a_2, \ldots, a_M$. Let us carry out a reduction to principal axes on it, and let the variables referred to the principal axes be $A_1, A_2, \ldots, A_M$. Then (3,14) becomes

$$\sum_{i=1}^{M} \lambda_i A_i^2 \geq K_{11} \| A \|_2^2 + K_{12} \| A \|_2 + K_{13}$$

where λ_i's are real numbers. As we have seen, the right hand side tends to ∞ with $\| A \|_2$ and this implies that $\lambda_i > 0$ $i = 1, 2, \ldots M$.

For if say $\lambda_1 \leq 0$ we can put $A_2 = A_3 = \cdots = A_M$ and let A_1 and hence $\|A\|$ tend to ∞. The right hand side approaches ∞ but the left would be negative or zero. Hence $\lambda_i > 0 \quad i = 1, 2, \cdots, M$ this means that the eigenvalues of the coefficient matrix in $(2,7)$ are positive and is nonsingular. The result is therefore proved. Let us remark that the condition $\rho(x) \geq k_1 > 0$ is equivalent to requiring that the Euler equation of $J(u)$ is uniformly elliptic. Other existence and uniqueness proofs for linear problems may be found in [29].

A uniqueness theorem for the general problem, with

will now be proved. A rather more general approach to this problem is possible [30]; the results given here however contain the essence of the more general approach.

Theorem 3.2

Define the function $F; R_n \to R_1$ by

$$F(p_1, \cdots, p_n) = \int_c^{\sum_{i=1}^n p_i^2} g(t)\, dt \equiv F(p)$$

and suppose that the matrix $\left|\frac{\partial^2 F}{\partial p_i \partial p_j}\right|$ is positive definite $\forall\, p \in R_n$. (cf. part 1 section 4). Let ϕ be given on $\partial\Omega$. Then equations $(2,7)$ cannot have more than one solution.

Proof

The positive definiteness of $\left|\frac{\partial^2 F}{\partial p_i \partial p_j}\right|$ implies that F is strictly convex in every closed sphere in R_n. We shall prove that the function of the M variables $a_1, a_2, \ldots, a_M$, $J(u_n)$ also has this property, from which will follow the uniqueness of a solution to

$$\frac{\partial J}{\partial a_i}(u_n) = 0 \qquad i = 1, 2, \cdots, M.$$

Since F is strictly convex, if $p, q \in R_n$

$$F(p) + F(q) \geq 2F\left(\tfrac{1}{2}(p+q)\right) \tag{3,14}$$

where equality holds if and only if $p \equiv q$.

Let u_M and V_M be 2 trial functions

$$u_M = \phi_0 + \sum_{i=1}^{M} a_i \theta_i$$

$$V_M = \phi_0 + \sum_{i=1}^{M} b_i \theta_i$$

Put

$$\left.\begin{aligned} p_j &= \phi_{0,x_j} + \sum_{i=1}^{M} a_i \theta_{i,x_j} \qquad j = 1,2,\dots,n \\ q_j &= \phi_{0,x_j} + \sum_{i=1}^{M} b_i \theta_{i,x_j} \qquad j = 1,2,\dots,n \end{aligned}\right\} \tag{3,15}$$

Substituting this into (3,14) adding $f u_M + f V_M$ to both sides and integrating over Ω we arrive at

$$J(u_M) + J(V_M) \geq 2J\left(\frac{u_M + V_M}{2}\right) \tag{3,16}$$

Now equality cannot hold in (3,16) under our hypothesis of a Dirichlet boundary condition unless $u_M \equiv V_M$. For suppose that $u_M \neq V_M$ and equality holds in (3,16). Then from (3,14) we can say that $p_j \equiv q_j$ in (3,15). But $p_j = u_{M,x_j}$ and $q_j = V_{M,x_j}$ so that $u_{M,x_j} \equiv V_{M,x_j}$ $j = 1,2,\dots,n$. This would have to hold everywhere in Ω and in particular in those simplexes having points in common with $\partial\Omega$. Now in view of $V_{M,x_j} = u_{M,x_j}$

in Ω, u_M and v_M differ by (at most) a constant in Ω; by a limiting argument we see that this would have also to occur on $\partial\Omega$. But u_M and v_M are equal on $\partial\Omega$ and so they are equal throughout $\bar{\Omega}$. This contradiction shows that equality in (3,16) implies that $u_M \equiv v_M$. Since the function $\frac{1}{2}(u_M + v_M)$ satisfies the same boundary conditions as u_M and v_M we can write (3,16) as

$$\psi(a_1,\dots,a_M) + \psi(b_1,\dots b_M) \geq 2\psi\left(\frac{a_1+b_1}{2},\dots,\frac{a_M+b_M}{2}\right)$$

where $\psi(a_1,\dots,a_M) \equiv J(u_M)$ and where there is equality if and only if $a \equiv b$. This is equivalent to the statement that $\psi(a)$ is a strictly convex function of a. Hence it can have at most one point where its gradient vanishes. Thus, the theorem is proved.

It is clear that the same proof holds if ϕ is given only on part of $\partial\Omega$. However if no essential boundary conditions are given we cannot infer uniqueness without further conditions. This is in line with known results for linear problems. We shall not consider this topic further.

3.4. A special case of the equations

We shall examine here a very special and simple case of the algebraic equations. In spite of the restrictions, the theory can be regarded as a good model of what can be done in practical situations. Indeed, its practical significance is great enough for it to be very useful for solving realistic problems. The restrictions, in the notations of para. 2 are

1) $m = 1$, $n = 2$

1a) $f = 0$

2) $p(|\nabla u|^2)$ does not contain the space variables x, y explicitly, and $p \geq 0$.

1a is not an essential condition. It is introduced merely as a technical simplification. In addition to 1) and 2) we shall introduce a further assumption. It is that the imposed boundary function can be approximated by a piecewise linear function to a sufficient degree of accuracy. What this means is that the boundary function can be allowed for in the trial function without the need for explicitly constructing a function ϕ_0 as mentioned in para.1.

In virtue of 2) above and the linearity of the trial function, β_0 of (2,4) is independent of x, y as are $\nabla\ell_i \equiv \nabla\lambda_i$. We shall evaluate the expressions $(\nabla\lambda_i, \nabla\lambda_j)$ in (2,4). To do this, we introduce the following notations: triangle τ_0 is taken initially to to have none of its vertices $(X_0, X_1, X_2) \in \partial\Omega$: the respective local parameters are $a_0^{(0)}$, $a_1^{(0)}$, $a_2^{(0)}$. We shall put $X_i = (x_i\ y_i)^T$ $i=1,2,3$. Introduce now new variables X, Y by

$$x = x_0 + (x_1 - x_0)X + (x_2 - x_0)Y$$

$$y = y_0 + (y_1 - y_0)X + (y_2 - y_0)Y$$

τ_0 maps into the triangle $(0,0)$, $(1,0)$, $(0,1)$ in the XY plane. Inverting (4,2) follows

$$X = \begin{pmatrix} x - x_0 & x_2 - x_0 \\ y - y_0 & y_2 - y_0 \end{pmatrix} \Big/ |J| \quad ; \quad Y = \begin{pmatrix} x_1 - x_0 & x - x_0 \\ y_1 - y_0 & y - y_0 \end{pmatrix} \Big/ |J| \qquad (4,3)$$

where

$$|J| = \det J = \begin{pmatrix} x_1 - x_0 & x_2 - x_0 \\ y_1 - y_0 & y_2 - y_0 \end{pmatrix},$$

and moreover,

$$\frac{\partial X}{\partial x} = \frac{y_2 - y_0}{|J|}, \quad \frac{\partial X}{\partial y} = \frac{x_3 - x_1}{-|J|}, \quad \frac{\partial Y}{\partial x} = \frac{y_2 - y_1}{-|J|}, \quad \frac{\partial Y}{\partial y} = \frac{x_2 - x_1}{|J|} \tag{4,4}$$

Under the transformation (4,2), $\lambda_i(x, y) \rightarrow \sigma_i(X, Y)$ $i = 0, 1, 2$

where

$$\begin{aligned} \sigma_1 &= -X - Y - 1 & \sigma_{1,X} &= \frac{\partial \sigma_1}{\partial X} = -1 & \sigma_{1,Y} &= -1 \\ \sigma_2 &= X & \sigma_{2,X} &= 1 & \sigma_{2,Y} &= 0 \\ \sigma_3 &= Y & \sigma_{3,X} &= 0 & \sigma_{3,Y} &= 1 \end{aligned} \tag{4,5}$$

Now,

$$\frac{\partial \sigma_i}{\partial x} = \frac{\partial \sigma_i}{\partial X}\frac{\partial X}{\partial x} + \frac{\partial \sigma_i}{\partial Y}\frac{\partial Y}{\partial x} = \sigma_{i,x} \qquad i = 0, 1, 2$$

and from (4,3) — (4,5) there follows

$$\begin{aligned} \sigma_{0,x}\,\sigma_{0,x} &= |J|^{-2}\left[(y_2 - y_0)^2 - 2(y_2 - y_0)(y_1 - y_0) + (y_1 - y_0)^2\right] \\ \sigma_{0,x}\,\sigma_{1,x} &= |J|^{-2}\left[-(y_2 - y_0)^2 + (y_1 - y_0)(y_2 - y_0)\right] \\ \sigma_{0,x}\,\sigma_{2,x} &= |J|^{-2}\left[(y_2 - y_0)(y_1 - y_0) - (y_1 - y_0)^2\right] \end{aligned} \tag{4,6}$$

The corresponding expressions for $\sigma_{i,y}\,\sigma_{j,y}$ are got by replacing y_i with x_i $i = 0, 1, 2$ in (4,6).

(4,6) simplify to

$$\begin{aligned} \sigma_{0,x}\,\sigma_{0,x} &= |J|^{-2}(y_2 - y_1)^2 \\ \sigma_{0,x}\,\sigma_{1,x} &= |J|^{-2}(y_2 - y_0)(y_1 - y_2) \\ \sigma_{0,x}\,\sigma_{2,x} &= |J|^{-2}(y_1 - y_0)(y_2 - y_1) \end{aligned}$$

and

$$\sigma_{0,y}\,\sigma_{0,y} = |J|^{-2}(x_2-x_1)^2$$

$$\sigma_{0,y}\,\sigma_{1,y} = |J|^{-2}(x_2-x_0)(x_1-x_0)$$

$$\sigma_{0,y}\,\sigma_{2,y} = |J|^{-2}(x_1-x_0)(x_2-x_1)$$

Collecting these results we find

$$\int_{\tau_0}(\nabla\lambda_0,\nabla\lambda_0)\,d\tau_0 = \int_{\tau_0'}|J|^{-1}\left\{(y_2-y_1)^2+(x_2-x_1)^2\right\}d\tau_0'$$

$$\int_{\tau_0}(\nabla\lambda_0,\nabla\lambda_1)\,d\tau_0 = \int_{\tau_0'}|J|^{-1}\left\{(y_2-y_0)(y_1-y_2)+(x_2-x_0)(x_1-x_2)\right\}d\tau_0'.$$

$$\int_{\tau_0}(\nabla\lambda_0\ \nabla\lambda_2)\,d\tau_0 = \int_{\tau_0'}|J|^{-1}\left\{(y_1-y_0)(y_2-y_1)+(x_1-x_0)(x_2-x_1)\right\}d\tau_0'$$

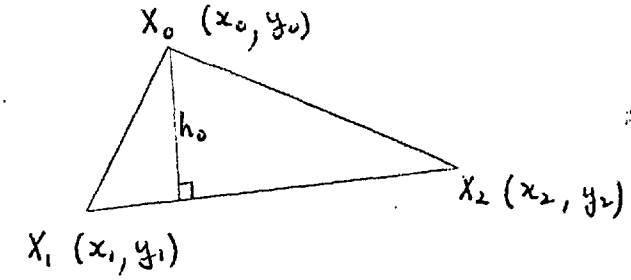
so that, in the notations of (2, 4) with $k=0$

$$\tfrac{1}{2}\omega_{00}(a^{(0)}) = \tfrac{1}{2}\rho_0\cdot\left\{(y_2-y_1)^2+(x_2-x_1)^2\right\}|J|^{-1}$$

$$\tfrac{1}{2}\omega_{10}(a^{(0)}) = \tfrac{1}{2}\rho_0\cdot\left\{(y_2-y_0)(y_1-y_2)+(x_2-x_0)(x_1-x_2)\right\}|J|^{-1}$$

$$\tfrac{1}{2}\omega_{20}(a^{(0)}) = \tfrac{1}{2}\rho_0\cdot\left\{(y_1-y_0)(y_2-y_0)+(x_1-x_0)(x_2-x_1)\right\}|J|^{-1}.$$

The quantities multiplying $\frac{1}{2}\rho_0$ have a simple geometric interpretation: consider the triangle drawn below:

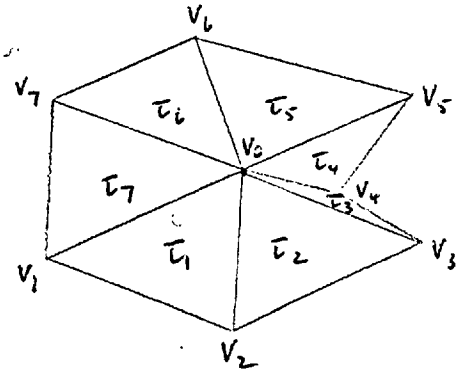


Then the 1st multiplier is just $\ell^2/h_0\ell = \cot X_1 + \cot X_2$ where h_0 is the altitude from X_0 ; the second is $-\cot X_2$ and the third is $-\cot X_1$. To prove these results it is necessary only to observe that the multipliers are unchanged in form by rotations of the axes and that $|J| = 2\times$ area of τ_0. The derivative of the local trial function with respect to $a_0^{(0)}$ therefore is

$$\sum_{j=0}^{2} \omega_{j0}\,(a_1^{(0)}\ a_2^{(0)}\ a_3^{(0)}).\ a_j^{(0)} \qquad (4,9)$$

where the weights ω_{j0} are given by $(4,8)$. If τ_0 has vertices on the boundary, all we have to do is to fix the values $a_j^{(0)}$ associated with these vertices to be the values of the boundary function at the appropriate points. If the vertices lie in parts of the boundary where no condition is imposed the corresponding parameters are determined in the same way as internal parameters.

Using $(4,9)$ we can write out the equation associated with any global parameter; this is done as follows. By the definition of a triangulation, each internal vertex of the triangulation is surrounded by a polygon: with each vertex of the polygon is associated a global parameter as there is with the internal vertex.. The geometric configuration of an internal vertex is shown in the diagram below: the vertices are labelled V_0 and $V_1 - V_7$. In general, the polygon may be an n-gon.



The process of writing down the equation will be illustrated with this diagram. (The principles are completely exposed with an example). Label the triangles in such a way that τ_i has vertices (V_0, V_i, V_{i+1}) $(V_8 \equiv V_1)$.

With V_i is associated a global parameter a_i $i=0,1,\ldots,7$. Consider triangle τ_i ; we associate with each of its vertices a weight, computed from (4,8) and taking V_0 to correspond with $X_0 = (x_0\ y_0)^T$, of (4,8). For brevity we shall introduce the notations $W_0^{(i)}$ $W_i^{(i)}$ $W_{i+1}^{(i)}$ for the weights in τ_i. Applying (4,9) to each triangle, adding the results as discussed in Section 2 equ. (2,9) and equating to zero, we find

$$a_0 \sum_{k=1}^{7} W_0^{(k)} + (W_1^{(7)} + W_1^{(1)})\, a_1 + (W_2^{(1)} + W_2^{(2)})\, a_2 +$$

$$+ (W_3^{(2)} + W_3^{(3)})\, a_3 + (W_4^{(3)} + W_4^{(4)})\, a_4 + (W_5^{(4)} + W_5^{(5)})\, a_5$$

$$+ (W_6^{(5)} + W_6^{(6)})\, a_6 + (W_7^{(6)} + W_7^{(7)})\, a_7 = 0$$

or

$$a_0 \sum_{k=1}^{7} W_0^{(k)} + \sum_{k=1}^{7} (W_k^{(k-1)} + W_k^{(k)})\, a_k = 0 \qquad (W_1^{(0)} \equiv W_1^{(7)}) \qquad (4,10)$$

(4,10) is the basic relationship which exists between a_0 and its surrounding global parameters. In view of

$$W_0^{(k)} + W_k^{(k)} + W_{k+1}^{(k)} = 0$$

(see (2,5a)), the sum of the weights in (4,10) is zero.

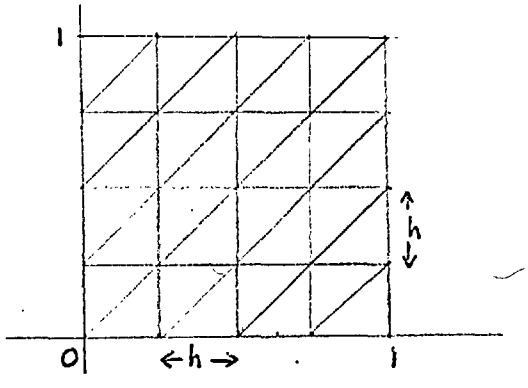
All the above considerations have been based on the assumption that

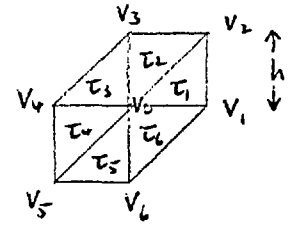
$V_5 - V_7$ do not lie at a point on the boundary at which there is an imposed boundary condition. Let us now suppose that vertex V_i say does lie in such a place. Then all we have to do is to fix a_i to be $\bar{a}_i$ where

$$\bar{a}_i = \phi_0(V_i)$$

If $\overline{V_i V_2}$ lies on the boundary, this corresponds to a linear interpolation (between V_i and V_2) of the boundary condition. If V_i lies on $\partial\Omega$ at a point where there is no boundary condition, a_i still appears in $(4,10)$ as a parameter. There remains the question of the equation for a point V_0 which is not inside a polygon, because it lies on the boundary, but for which a_0 is a parameter (as opposed to being fixed by the boundary conditions). This case can be easily dealt with on the basis of the argument used to derive $(4,10)$ dealing with prescribed boundary conditions as indicated above.

Next, we shall indicate a further simplification which can be made for simple enough regions. It will be illustrated by an example. We take for Ω the square $0 < x, y < 1$ and triangulate it in the manner shown below (and described previously in the example of para. 2)





Consider any internal vertex V_0 and its surrounding hexagon, and let us compute the equation for a_0. This equation takes the form

$$\rho_1\left[a_0-\left(\cot\tfrac{\pi}{4}\right)a_1\right]+\rho_2\left[a_0-\left(\cot\tfrac{\pi}{4}\right)a_3\right]+$$

$$+\rho_3\left[2a_0-\left(\cot\tfrac{\pi}{4}\right)a_3-\left(\cot\tfrac{\pi}{4}\right)a_4\right]+\rho_4\left[a_0-\left(\cot\tfrac{\pi}{4}\right)a_4\right]+$$

$$+\rho_5\left[a_0-\left(\cot\tfrac{\pi}{4}\right)a_6\right]+\rho_6\left[2a_0-\left(\cot\tfrac{\pi}{4}\right)a_6-\left(\cot\tfrac{\pi}{4}\right)a_1\right]=$$

$$=a_0\left(\rho_1+\rho_2+2\rho_3+\rho_4+\rho_5+2\rho_6\right)-$$

$$-a_1\left(\rho_1+\rho_6\right)-a_3\left(\rho_2+\rho_3\right)-a_4\left(\rho_3+\rho_4\right)-a_6\left(\rho_5+\rho_6\right) \qquad (4,10a)$$

Equating this to zero gives the desired equation. Let us note immediately that if $\rho\equiv 1$ this equation reduces to

$$4a_0-a_1-a_3-a_4-a_6=0$$

which is the common difference approximation to the Laplace operator. Moreover, in the general case, the equation does not involve the variables associated with V_2 and V_5 except through the function ρ. Now carrying out the same computation at each vertex of the triangulation, and writing out the equations in matrix form we get

$$W(a)\cdot a=b(a) \qquad (4,11)$$

where $W(a)$ is a matrix whose elements depend on the vector $a=(a_1, a_2, \ldots, a_M)^T$ and $b(a)$ is a vector whose elements depend on a and the boundary conditions. The geometrical properties of W, by which is meant the positions of its nonzero elements, do not depend

on ρ. In particular, if $\rho = 1$ this matrix is known to be 2-cyclic [31], (or have Young's "Property A"). Moreover, it is not hard to see that this feature subsists for regions more general than the square we have taken for the model. It can, to an extent, be exploited for the purposes of solving the equations (4,11). Let us recall that the general equations with m, n arbitrary are also in the form (4,11) (see 2,10a). We cannot, however, say much about the form of the matrix $\bar{W}(a)$ of (2,10a) except that it is 'sparse'.

The following statement is evident:
$W(a)$ of (4,11) is symmetric and positive definite (provided there is at least one vertex of the triangulation lying on part of the boundary with an imposed boundary condition). The proof of it follows immediately in the basis of (4,10a). Furthermore, the corresponding matrix for the case $m = 1$ with an arbitrary triangulation also has these properties. The algebraic problem then arises:- solve the equations

$$W(a)\, a = b(a)$$

where $\bar{W}(a)$ is a symmetric, positive definite matrix (whose positive definiteness holds for any vector a) and $b(a)$ is a vector, dependent on the vector a. We shall consider the solution of this problem in part 4.

3.5 Resume and notes

In this part we have used the piecewise polynomial trial functions, which were studied in part 2, to reduce the problem stated in part 1 section 9 to a problem dependent on a finite number of variables. We showed when the finite dimensional problems were uniquely solvable in the case of arbitrary triangulations of bounded regions in R_n for all degrees of polynomials. The equations were then set up in detail for a special problem in R_2.

Some remarks concerning the origins of the results will now be made. In section 1 the only item requiring comment is the idea for the construction of ϕ_0. This does not seem to have been previously suggested. As regards section 2, for the linear case at least, the manipulations leading to (2,10a) are well known, especially in connection with the (linear) finite element method. However, it was thought worthwhile to include the derivations here in order that the precise way in which the nonlinearity manifests itself in the algebiaic equations could be seen. (2,10a) brings out this point; this equation does not appear to have been given elsewhere. As was said in the introduction, the 'elimination of internal nodes' considered in the second part of section 2 is a straightforward generalisation of an idea in current use. In section 3 we used (Theorem 3.1) the idea of an increasing functional. The author became aware of this concept and its use in variational methods through works of Mikhlin [32]. (This book is not at present available in English, but a translation is due to appear). Theorem 3.1 is thought to be new. We gave the uniqueness proof of the corollary 3.1 just because of its own interest. It implies that an increasing quadratic form has a positive definite coefficient matrix. Theorem 3.2, which shows that convexity of $F(p_1, \ldots, p_n)$ is transmitted to convexity of $\psi(a_1, \ldots, a_M)$,

was made up by the author before the appearance of the paper [30] referred to in the text. It owes something to the previously mentioned book [32].

In section 4, the derivation of the equations is again justified by the fact that the way that ρ enters the final system of equations is clearly seen. The same system of equations was derived by Greenspan [33] by differencing various differential equations in a particular way. They were also given by Winslow [34] where they were obtained for a problem in magnetostatics by a physical argument. This author also stated that they could be obtained by a variational argument and indicated extensions to other problems. The fact that the coefficient matrix has property A for the case of a right triangulation does not seem to have been pointed out before.

Part 4: Solution of the algebriac equations

4.1 Introduction and objectives

We consider in this part the solution of the equations

$$W(x) \cdot x = b(x) \qquad (1,1)$$

where $W(x')$ is an $n \times n$ matrix whose elements depend on x'. It will be assumed that for any x', $W(x')$ is positive definite and that $(1,1)$ has a unique solution x.

We shall confine attention to methods which in some way exploit various special features of $(1,1)$. Two classes of methods will be distinguished, namely those using derivatives of $(1,1)$ and other methods. These classes will be referred to as containing derivative linked algorithms on the one hand and derivative free algorithms on the other.

The contents of the sections are as follows. Section 2 presents a simple general theory of a wide class of iterative methods. In section 3 two particular algorithms are introduced and it is shown how the results of section 1 can be applied to them. Section 4 contains some general remarks about the choice of certain 'acceleration' parameters for iterative mthods of the type introduced in section 2. Sections 5 and 6 deal with a generalisation of successive overrelaxation methods to nonlinear problems.

Two further remarks are in order. Firstly, especially in sections 4 and 6 the presentation is more discursive than logically continuous. The reason for this is a desire to avoid the presentation of complex synthetic conditions which are difficult (if not impossible) to verify in specific

instances. We attempt to adduce principles rather than general theorems, although usually, sufficient information is given for the construction of such theorems. Secondly, it should be pointed out that the field of the iterative solution of nonlinear equations is continually being added to and refined. There is considerable overlap with its results and those of sections 2,3,5 and 6. In order not to disrupt the treatment, an account of the relations between our results and known results is given in the resume, section 7.

4.2 Iterative methods

We shall consider only algorithms which can be written in the form

$$\left.\begin{aligned} x^{(k+1)} &= x^{(k)} - \alpha^{(k)} L^{(k)} f^{(k)} \\ x^{(0)} &\quad \text{given} \end{aligned}\right\} \qquad k = 0, 1, \ldots \qquad (2,1)$$

where $f^{(k)} \equiv f(x^{(k)})$ and $f(x) = 0$ is a system of n equations in n unknowns to be solved; $\alpha^{(k)}$ are scalars and $L^{(k)}$ are $n \times n$ matrices. (2,1) includes a very great number of known algorithms.

Let $\varepsilon^{(k)} = x^{(k)} - x$ and let $p^{(k)}$ denote <u>any</u> matrix satisfying the equation

$$p^{(k)} \varepsilon^{(k)} = f^{(k)} \qquad (2,2)$$

From (2,1) follows immediately that

$$\begin{aligned} \varepsilon^{(k+1)} &= \left(I - \alpha^{(k)} L^{(k)} p^{(k)}\right) \varepsilon^{(k)} \\ &= \prod_{i=0}^{k} \left(I - \alpha^{(i)} L^{(i)} p^{(i)}\right) \varepsilon^{(0)} \end{aligned} \qquad (2,3)$$

$(2,3)$ is exactly the error after $k+1$ steps of $(2,2)$. If the equations $f(x)$ are linear, so that $f(x') = Ax' - b$ then for $P^{(k)}$ we can clearly take $P^{(k)} \equiv A$. Then $(2,3)$ becomes

$$\varepsilon^{(k+1)} = \prod_{i=0}^{k} (I - \alpha^{(i)} L^{(i)} A)\, \varepsilon^{(0)} \qquad (2,4)$$

which is the governing equation for the error in the case of linear equations. In order to see the uses to which it may be put suppose A is positive definite; now as is well known, many minimization algorithms can be written in the form $(2,2)$ with $f^{(k)} = Ax^{(k)} - b \equiv \nabla F^{(k)}$ say. The parameters $\alpha^{(k)}$ (steplengths) and $L^{(k)}$ define the particular algorithm. Many such algorithms converge to x in n or less steps. From $(2,4)$ we see that a necessary and sufficient condition for this is that the matrices $L^{(k)}$ and parameters $\alpha^{(k)}$ should satisfy the relationship

$$\prod_{i=0}^{k} (I - \alpha^{(i)} L^{(i)} A) \equiv 0$$

Thus e.g. $(2,4)$ may serve as a basis for the construction of algorithms. As another example of its use we cite the derivation of the basic Richardson algorithm in which $L^{(k)} \equiv I$ and $(2,4)$ is used to determine the parameters $\alpha^{(i)}$ to achieve rapid convergence [35]

We are going to use $(2,3)$ the general form of $(2,4)$ to prove a convergence theorem for the iteration $(2,1)$. We have first to choose the matrices $P^{(k)}$ of $(2,2)$ and this will now be done.

Some new vectors will be introduced in order to expedite this process. Let

$$x^{(k)j} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_j^{(k)}, x_{j+1}, \ldots, x_n)$$

where $x^{(k),0} \equiv x$ and $x^{(k),n} \equiv x^{(k)}$; define $f^{(k),j}$ by $f^{(k,j)} \equiv f(x^{(k),j})$ $j=0,1,\dots,n$. We shall assume from now on that in a sufficiently large neighbourhood of x, f_i $i=1,2,\dots,n$ has first partial derivatives which are Lipschitz continuous, with constant μ. Consider the matrix Q_k whose $(i\,j)$th element is

$$\left.\begin{array}{ll} \dfrac{f_i^{(k),j} - f_i^{(k),j-1}}{x^{(k),j} - x_j} & x_j^{(k),j} \neq x_j \\ \dfrac{\partial f_i}{\partial x_j}(x^{(k),j}) & x_j^{(k),j} = x_j \end{array}\right\} \qquad (2,5)$$

If $x_j^{(k)} = x_j$ then $f^{(k),j} = f^{(k),j-1}$. In view of this, we have

$$\sum_{j=1}^{n} [Q_k]_{ij} (x_j^{(k)} - x_j) = \sum_{j=1}^{n} f_i^{(k),j} - f_i^{(k),j-1}$$

$$= f_i^{(k),n} = f_i^{(k)}$$

Thus, for $P^{(k)}$ we can take Q_k. defined in $(2,5)$

On the basis of the Lipschitz continuity of the derivatives $\frac{\partial f_i}{\partial x_j}$ the members of $(2,5)$ can be written as, in either case,

$$[P^{(k)}]_{ij} = \frac{\partial f_i}{\partial x_j}(x) + O(\|\varepsilon^{(k)}\|_\infty) \qquad (2,6)$$

where $O(\|\varepsilon^{(k)}\|_\infty)$ denotes terms $R^{(k)}$ satisfying the inequality

$$|R^{(k)}| \leq \mu \|\varepsilon^{(k)}\|_\infty$$

Next we shall impose some conditions on $\alpha^{(k)}$ and $L^{(k)}$. For $\alpha^{(k)}$ we shall assume that $\alpha^{(k)} = \alpha = \text{const.}$ $L^{(k)}$ will be assumed

to be a matrix of functions each one of which depends only on a vector x' and which are defined everywhere in a neighbourhood of x. The elements of $L(x')$ are assumed to be Lipschitz continuous functions in this neighbourhood. In this case we can write

$$L^{(k)} \equiv L(x^{(k)}) = L(x) + O(\|\varepsilon^{(k)}\|_\infty)$$

where $O(\|\varepsilon^{(k)}\|_\infty)$ now denotes a matrix $S^{(k)}$ satisfying the inequality

$$\|S^{(k)}\|_\infty \leq C_1 \|\varepsilon^{(k)}\|_\infty$$

where C_1 is a number, (constructed from the Lipschitz constants of the elements of $L(x')$) which does not depend on $\varepsilon^{(k)}$.

Let $H(x')$ denote the matrix whose $(i\,j)$th element is $\frac{\partial F_i}{\partial x_j}(x')$. Since these functions are Lipschitz continuous we can also write

$$H^{(k)} \equiv H(x^{(k)}) = H(x) + O(\|\varepsilon^{(k)}\|_\infty)$$

as we did for $L^{(k)}$.

Finally, let $A = I - \alpha L(x) H(x)$ and put $\rho(A) \equiv$ spectral radius of A.

The proof makes essential use of the following fundamental result of Ostrowski:

Lemma 2.1

Let A be an $n \times n$ matrix and $\varepsilon > 0$. There exist two positive constants $\eta_1 > 0$ and $\sigma > 0$ dependent on A and ε only such that if, for a sequence of $n \times n$ matrices U_μ with

$$\| u_\mu \|_\infty < \eta, \qquad \mu = 1, 2, \ldots$$

we form

$$\Pi_m = \prod_{\mu=1}^{m} (A + u_\mu) \qquad (*)$$

then, irrespective of the order of the factors in $(*)$ it holds that

$$\| \Pi_m \|_\infty \leq \sigma (\lambda_A + \varepsilon)^m \qquad m = 1, 2, \ldots$$

where λ_A is the spectral radius of A. In particular, if $\lambda_A < 1$ then $\Pi_m \to 0$ as $m \to \infty$.

Proof

See [36].

Our convergence theorem is this:

Theorem 2.1

Under the stated hypotheses on L, H and α, if $\rho(A) < 1$ and $\| \varepsilon^{(0)} \|_\infty$ is sufficiently small the iteration $(2, 2)$ is defined and generates a sequence convergent to x.

Proof

Select any $\varepsilon > 0$ such that $(\rho(A) + \varepsilon) < 1$. By Lemma 2.1 there are numbers σ, η associated with A such that the conclusions of that lemma hold. Let $\bar{S}(\delta; x)$ be a sphere $\{ y \mid \| y - x \|_\infty \leq \delta \}$ where the Lipschitz conditions on L and P hold for any $y \in S(\delta; x)$; put

$$g(y) = y - \alpha L(y) f(y)$$

We shall first prove that for $y \in S(\delta,x)$ the matrix U_y defined by

$$U_y = (I - \alpha L(y)P(y)) - A$$

satisfies

$$\|U_y\|_\infty = O(\|y-x\|_\infty) \qquad (2,7)$$

Indeed, since $y \in S(\delta, x)$

$$I - \alpha L(y)P(y) = I - \alpha\left(L(y) + O(\|y-x\|_\infty)\right)\left(H(x) + O(\|y-x\|_\infty)\right)$$

$$= A + O(\|y-x\|_\infty)$$

which proves the assertion.

Next we shall prove that if $\gamma > 0$ is any given number, then $\exists \quad \theta(y)$ such that $\|x - g(y)\|_\infty < \gamma$ for all y satisfying $\|x-y\|_\infty < \theta$. We have

$$g(y) - y = (I - \alpha L(y)P(y))(y-x)$$

$$= (A + U_y)(y-x) \qquad (2,8)$$

and by $(2,7)$, $\exists \quad \theta'$ such that $\|U_y\|_\infty < \eta$ if $\|y-x\|_\infty < \theta'$. Hence, choosing y to satisfy this last inequality and applying Lemma 2.1 it follows that

$$\|g(y)-x\|_\infty \leq \sigma(\rho(A)+\varepsilon)\|y-x\|_\infty \quad ;$$

the above statement then holds with $\theta(\gamma) = \min\left(\theta', \frac{\gamma}{\sigma(\rho(A)+\varepsilon)}\right)$.

Now let ν_0+1 be any integer, and $\delta \geq \delta_1 > 0$ any number. It follows from what was just proved that if $\| \varepsilon^{(0)} \|_\omega$ is sufficiently small then $x^{(j)} \in S(\delta_1; x)$ $j = 0, 1, \ldots, \nu_0$.
For to have $x^{(\nu_0)} \in S(\delta_1; x)$ we have only to have $\| \varepsilon^{(\nu_0-1)} \|_\omega$ sufficiently small, $< \theta^{(\nu_0-1)}$ say. To have $\| \varepsilon^{(\nu_0-1)} \|_\omega < \theta^{(\nu_0-1)}$ we have only to have $\| \varepsilon^{(\nu_0-2)} \|_\omega < \theta^{(\nu_0-2)}$ say.
Continuing in this way we infer the existence of a $\theta^{(0)}$ such that if $\| \varepsilon^{(0)} \|_\omega < \theta^{(0)}$ then all the iterates $x^{(j)}$ $j=0,1\ldots, \nu_0$ lie in $S(\delta_1; x)$.

We shall now prove the theorem. Let $\nu+1$ be the smallest integer such that $\sigma(\rho(A)+\varepsilon)^{\nu+1} < 1$; then we shall prove that for any $m \geq \nu+1$ there holds

$$\| \varepsilon^{(m)} \|_\omega \leq \sigma(\rho(A)+\varepsilon)^m \| \varepsilon^{(0)} \|_\omega \qquad (2,9)$$

if $\| \varepsilon^{(0)} \|_\omega$ is sufficiently small. This is equivalent to the statement of the theorem. The proof is by induction on m. First consider $m = \nu+1$; then we have

$$\varepsilon^{(\nu+1)} = \prod_{i=0}^{\nu} \left(I - \alpha L^{(i)} P^{(i)}\right) \varepsilon^{(0)}$$

$$= \prod_{i=0}^{\nu} \left(A + U^{(i)}\right) \varepsilon^{(0)}$$

where $U^{(i)} \equiv \left[I - \alpha L^{(i)} P^{(i)}\right] - A$. Choose $\| \varepsilon^{(0)} \|_\omega$ so small that all the iterates $x^{(0)}, \ldots, x^{(\nu)}$ lie in $S(\theta'; x)$ where θ' has the same significance as above (following (2,8)). Then we have,

by definition of ϑ',

$$\| u^{(i)} \|_\infty < \eta \qquad (2,10)$$

and we can apply Lemma 2.1 to get

$$\| \varepsilon^{(\nu+1)} \|_\infty \leq \sigma \left(\rho(A) + \varepsilon \right)^{\nu+1} \| \varepsilon^{(0)} \|_\infty$$

This establishes the 1st step of the induction. Assume now that (2,9) holds for $\nu+1, \nu+2, \ldots, m-1$. Then

$$\varepsilon^{(m+1)} = \prod_{i=0}^{m-1} \left(I - \alpha L^{(i)} P^{(i)} \right) \varepsilon^{(0)} \qquad (2,11)$$

Since

$$\| \varepsilon^{(j)} \|_\infty \leq \sigma \left(\rho(A)+\varepsilon \right)^{j} \cdot \| \varepsilon^{(0)} \|_\infty \qquad j = \nu+1, \ldots, m-1$$

all the iterates $x^{(j)}$ $j = \nu+1, \ldots, m-1$ lie closer to x than does $x^{(0)}$. Hence the necessary estimates of the form (2,10) hold in (2,11) and the theorem is proved.

In essence, we have also proved that asymptotically, $\| \varepsilon^{(k)} \|_\infty \to 0$ no slower than a geometric progression with ratio $\rho(A)$. This is well known for the case of linear equations.

The question of the choice of the parameter α will be considered briefly in section 4.4.

4.3 Derivative free algorithms

Many algorithms of this type are based on the idea of an 'inner' iteration being carried out on a system of linear equations obtained by some linearization of the nonlinear system. The result of the inner iteration is used to provide a new linearization to which the inner iteration can be

applied, and the whole process is repeated cyclically. The number of algorithms utilizing these principles is very great indeed. Here we shall consider just two of them, both of which have the property that the linearization of the nonlinear system is based on its special form, as given in equation (1,1).

We shall first state the algorithms, then show how they can be written in the standard form (2,1). Finally we shall make some remarks about the application of the convergence theorem to them.

Algorithm 1

Suppose ω, α, $x^{(k)}$ are given

a) Linearize the system (1,1) by putting

$$W^{(k)} \bar{x}^{(k+1)} = b^{(k)} \qquad (3,1)$$

where $W^{(k)} = W(x^{(k)})$ etc.

b) Carry out one iteration of the linear successive overrelaxation algorithm, starting with $x^{(k)}$ with parameter ω on system (3,1) to get $\tilde{x}^{(k+1)}$

c) Define $x^{(k+1)}$ by

$$x^{(k+1)} = x^{(k)} + \alpha (\tilde{x}^{(k+1)} - x^{(k)})$$

d) Set $k := k+1$ and carry on from a)

From a) to c) is one step of Algorithm 1. The motivation for this procedure is clear; if $x^{(k)}$ is very close to x and the equations are linear we would be justified in hoping that $x^{(k+1)}$ is even closer to x. The restriction to a single SOR step can be removed and replaced by any other (fixed) number of steps. We shall consider this no further here. In Algorithm 2 we consider the situation in which

equations $(3,1)$ are solved exactly at each step:

Algorithm 2

Suppose α, $x^{(k)}$ are given

a) Linearize $(1,1)$ by putting

$$W^{(k)} \tilde{x}^{(k+1)} = b^{(k)}$$

b) Solve

$$W^{(k)} \tilde{x}^{(k+1)} = b^{(k)}$$

c) Define

$$x^{(k+1)} = x^{(k)} + \alpha (\tilde{x}^{(k+1)} - x^{(k)})$$

We shall now write both of these algorithms in the form $(2,1)$. Consider first Algorithm 1. Put $W^{(k)} = D^{(k)} + E^{(k)} + F^{(k)}$ where the three matrices on the right are respectively the diagonal, strictly lower and strictly upper triangular parts of $W^{(k)}$. By the linear SOR theory

$$\tilde{x}^{(k+1)} = -(\omega^{-1} D^{(k)} + E^{(k)}) \{ [(1-\omega^{-1}) D^{(k)} + F^{(k)}] x^{(k)} - b^{(k)} \}$$

and

$$\tilde{x}^{(k+1)} - x^{(k)} = -(\omega^{-1} D^{(k)} + E^{(k)})^{-1} \left[\{ \omega^{-1} D^{(k)} + E^{(k)} + (1-\omega^{-1}) D^{(k)} + F^{(k)} \} x^{(k)} - b^{(k)} \right]$$

$$= -(\omega^{-1} D^{(k)} + E^{(k)})^{-1} \{ W^{(k)} x^{(k)} - b^{(k)} \}$$

Putting $f^{(k)} = W^{(k)} x^{(k)} - b^{(k)}$ we have

$$x^{(k+1)} = x^{(k)} - \alpha (\omega^{-1} D^{(k)} + E^{(k)})^{-1} f^{(k)} \qquad (3,2)$$

and this is the desired form.

As regards Algorithm 2, note that we have the relations

$$W^{(k)} \tilde{x}^{(k+1)} - b^{(k)} = 0$$

$$W^{(k)} x^{(k)} - b^{(k)} = f^{(k)}$$

Substracting them

$$W^{(k)} (\tilde{x}^{(k+1)} - x^{(k)}) = - f^{(k)}$$

and as a consequence of this,

$$x^{(k+1)} = x^{(k)} - \alpha (W^{(k)})^{-1} f^{(k)} \qquad (3,3)$$

which is in the form $(2,1)$.

In order to discuss the asymptotic convergence of these algorithms we need to know the Jacobian matrix $H(x)$ of the system $(1,1)$. To compute it, consider the i^{th} equation of the system $(1,1)$,

$$f_i = \sum_{j=1}^{n} w_{ij}(x) . x_j - b_j(x)$$

$$\frac{\partial f_i}{\partial x_j} = \sum_{k=1}^{n} \frac{\partial w_{ik}}{\partial x_j} x_k - \frac{\partial b_i}{\partial x_j} + w_{ij}$$

This expression is $[H(x)]_{ij}$. Let $\frac{\partial W}{\partial x_j}$ denote the matrix obtained by differentiating each element of W with respect to x_j. Then

$$H = \left[\frac{\partial W}{\partial x_1} . x , \frac{\partial W}{\partial x_2} . x , \ldots , \frac{\partial W}{\partial x_n} . x \right] - \left[\frac{\partial b}{\partial x} \right] + W$$

where in the first matrix the column vectors comprising it are shown, and in the second, the $(i\,j)^{th}$ element is $\frac{\partial b_i}{\partial x_j}$. Thus,

$$H - W = \left[\frac{\partial W}{\partial x_1} \cdot x, \ldots, \frac{\partial W}{\partial x_n} \cdot x \right] - \left[\frac{\partial b}{\partial x} \right]$$

and in the case of linear equations this is zero, In this sense, the matrix $H-W$ is a measure of the deviation from linearity of system (1,1). We shall put it equal to Δ,

$$\Delta = H - W = \left[\frac{\partial W}{\partial x_1} \cdot x, \ldots, \frac{\partial W}{\partial x_n} x \right] - \left[\frac{\partial b}{\partial x} \right]$$

so that $H = W + \Delta$.

The matrices $I - \alpha L(x) H(x)$ governing the asymptotic convergence of the two algorithms are

$$I - \alpha (\omega^{-1} D + E)^{-1} \{ W + \Delta \} \equiv A_1 \qquad (3,4)$$

for Algorithm 1, and

$$I - \alpha W^{-1} \{ W + \Delta \} \equiv A_2 \qquad (3,5)$$

for Algorithm 2. These follow immediately from (3,2) and (3,3). The matrices which appear in these expressions are all evaluated at the root x. Both (3,4) and (3,5) reflect the fact that if the equations are linear, so that $\Delta \equiv 0$ the algorithms both reduce to known algorithms. Generally, we will have local convergence if $\rho(A_1) < 1$ or if $\rho(A_2) < 1$; consider Algorithm 1: suppose that SOR applied to the <u>linear</u> problem $Wx = b$ converges. It is intuitively clear that if $\| \Delta \|_\infty$ is sufficiently small then (by 'perturbation' arguments) we can have convergence in the nonlinear case for $\| x - x^{(0)} \|_\infty$ sufficiently small. In the case of Algorithm 2, taking $\alpha = 1$ for convenience we see

that for $\Delta \equiv 0$ the exact solution is obtained in one iteration. Again by applying pertubation arguments we could show that convergence occurs in the nonlinear problem under similar conditions to those just started.

However, we shall not carry out the merchanics of these proofs because the necessary hypothesis (that $\| \Delta \|_\infty$ should be sufficiently small) ignores the origin of the original system of equations. The hypothesis of ellipticity already restricts the size of $\| \Delta \|_\infty$ and it is felt that essential use of this fact should be made. Some investigations along this line were made by the author, but various difficulties encountered could not be surmounted, and therefore we eschew further discussion of this topic.

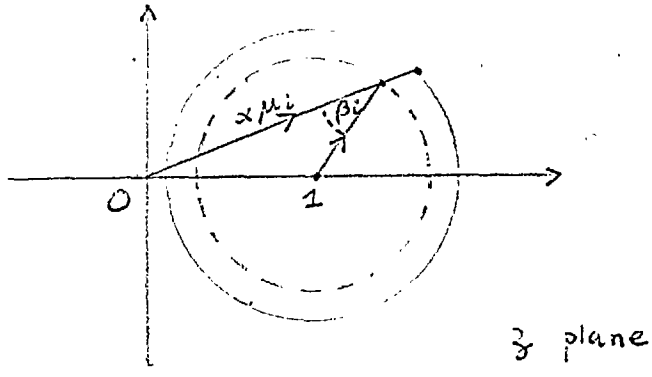
4. Choosing α.

From section 2 and the remark following the proof of Theorem 2.1 it follows that the convergence will be the faster (assymptotically) the smaller $\rho(A(\alpha))$ with $A(\alpha) = I - \alpha L H$. We have α at our disposal to make $\rho(A(\alpha))$ as small as possible. In order to achieve a satisfactory solution to this problem, we should need to have fairly complete information about the spectrum of the matrix LH. Unfortunately, this is not usually available and this fact rather restricts the use of the following remarks. However, some idea of the role played by this parameter can be gained as will be seen.

We shall assume that $\rho(A(1)) < 1$. Let λ_i $i = 1, 2, \ldots, n$ be the eigenvalues of $A(1) \equiv I - LH$. Let μ_i $i = 1, 2, \ldots, n$ be the eigenvalues of LH so that $\lambda_i = 1 - \mu_i$ $i = 1, 2, \ldots, n$. The assumption $|\lambda_i| < 1$ implies that the λ_i all lie inside or on the boundary of a disc C_1 concentric with the circle $|z| \leq 1$ in the

complex (ζ) plane and of radius $\rho(A(1)) < 1$. The corresponding picture for the μ_i is obtained by reflecting each root λ_i in the origin and then translating C_1 so that its centre is at $(1,0)$. The point originally λ_i is now $1-\lambda_i = \mu_i$. These transformations can be reversed to get back to the original picture.

Consider now the eigenvalues of αLH which are just $\alpha\mu_i$. Geometrically this means that the roots μ_i are shifted along the lines joining them to the origin by an amount $(\alpha-1)\mu_i$ (if $\alpha<1$, μ_i moves toward the origin and if $\alpha>1$ it moves away from it). If now β_i is the angle between the vectors μ_i and μ_i-1 it is not difficult to see that the above shift in the position of μ_i is composed of radial and tangential components, relative to the circle $|\zeta-1| = |\mu_i|$ of respective magnitudes $(\alpha-1)|\mu_i|\cos\beta_i$ and $(\alpha-1)|\mu_i|\sin\beta_i$



On the basis of these remarks it would be possible to give general conditions on the dispositions of the μ_i which would ensure the possibility of choosing α to achieve some reduction in $\rho(A(\alpha))$ relative to $\rho(A(1))$. Rather than do this we shall show that there are

situations where no improvement can be obtained by choosing $\alpha \neq 1$. Suppose LH has two real eigenvalues μ_1 and μ_2 equal in magnitude to its spectral radius, and equally spaced about the point $(1, 0)$. Let $\alpha \neq 1$. To obtain the roots $\alpha\mu_i$ $i=1,2$ we have to perform translations on the roots μ_i along the real axis. The size of the disc on $(1, 0)$ required to contain the translates must then increase compared with its size when $\alpha = 1$. Passing back to the roots $1-\alpha\mu_i$ by the reflection/translation mentioned previously we see that the disc required to contain the roots $1-\alpha\mu_i$ has a greater radius than that required to contain the roots $1-\mu_i$. So in this case, $\alpha = 1$ is the best value for α.

Now consider Algorithm 1, for which the error matrix is given in (3,4). Suppose, in order to gain an insight into what can happen, that $\alpha = 1$ and that W has property A. Suppose further that Δ has the character of a 'perturbation' from the zero matrix. Then (3,4) is approximately the matrix governing the solution by SOR with parameter ω of a linear system $W.x = b$. On the face of it there seems to be little reason for using an ω different from ω_{opt}. But then the roots of the error matrix all lie on a circle of radius $\omega_{opt} - 1$ with real roots at $\omega_{opt} - 1$ and $-\omega_{opt} + 1$. Consequently, the roots μ_i $i=1,2$ corresponding to these are equally spaced about $(1, 0)$. Therefore as we saw above, no improvement can be had by choosing α different from 1. In practice, ω will not be ω_{opt} and $\|\Delta\|_\omega$ may not be very small. But the above discussion shows that we cannot really expect to do more by choosing $\alpha \neq 1$ than prevent divergence of the iteration, rather than induce faster convergence which is what we should like.

With Algorithm 2 the situation is different as we shall now see. The

error matrix is given in (3,5). We have here that μ_i are the eigenvalues of $W^{-1}H$. We shall assume that H is positive definite. This is in line with the interpretation of (1,1) as the gradient of a functional whose minimum is sought. Since W is positive definite by hypothesis it follows [53] that the eigenvalues of $W^{-1}H$ are real and positive. Now $W^{-1}H = I + W^{-1}\Delta$ and so the eigenvalues of $W^{-1}H$ are found by adding 1 to those of $W^{-1}\Delta$. Suppose Δ is a definite matrix: then all the eigenvalues of $W^{-1}H$ will lie on the same side of (1, 0) and it will certainly be possible to speed the convergence by taking $\alpha \neq 1$. It is evident that the best value for α, $\hat{\alpha}$ is that which translates the interval $[\mu_{max}, \mu_{min}]$ so that it is symetrically disposed about (1, 0). This reasoning holds even if Δ is not definite. Also it is easy to see that this value of α is just

$$\hat{\alpha} = \frac{2}{\mu_{max} + \mu_{min}}$$

and the corresponding convergence rate is

$$\frac{\mu_{max} - \mu_{min}}{\mu_{max} + \mu_{min}}$$

It is not easily possible to compare the convergence rates with $\alpha = 1$ and $\alpha = \hat{\alpha}$; this is because for $\alpha = 1$ it depends on only 1 eigenvalue, but for $\alpha = \hat{\alpha}$ it depends on the extreme pair of eigenvalues.

To sum up, we can say that Algorithm 1 is not likely to be (asymptotically) improved by choosing α different from 1, whilst Algorithm 2 will always benefit from the introduction of $\hat{\alpha} \neq 1$ provided the above mentioned equal spacing of the roots μ_{min} and μ_{max} does not occur. It should be emphasised that these remarks apply only to the local convergence of the

algorithms. The question of the global convergence is very much more difficult.

4.5 Derivative linked algorithms

In this section we shall consider the local convergence behaviour of nonlinear block successive overrelaxation (SOR) algorithms (which includes as a special case point SOR). The advantages of these algorithms are well known for linear problems and in certain respects the advantages subsist for the more general problem. We shall not give any theorems explicitly but merely indicate how they can be proved. To begin with we shall indicate some of the history of the generalisations, then we shall formulate the algorithms and finally we shall elucidate the nature of the asymptotic convergence.

The first generalisations of point SOR to nonlinear problems seem to have been given in [37] (for a special problem) and then independently in [38]. The convergence theory was considered from a variational point of view in [39] and subsequently computational studies appeared in [40], [41]. A more general approach to the convergence theory was given in [42] and [43]. Meanwhile material in the spirit of [39] appeared in [44] and [45] in connection with block methods and 'free-steering' algorithms. The results given below, which are for block SOR algorithms generalise those given in [43] for point SOR although the methods used to obtain the results are different. Some computational studies of nonlinear block SOR have been given in [46].

In order to formulate the algorithm some notations are needed. Let x be an n dimensional vector and let a partition of x into m

blocks X_i $i=1,2,\ldots,m$ be given: $x = (X_1^T\ X_2^T \ldots X_m^T)$
The number of components in each block X_i is β_i and $\sum_{i=1}^{m} \beta_i = n$.
Now let A be any given $n \times n$ matrix: the stated partition of x induces a partitioning of A into blocks A_{ij} $i,j = 1,2,\ldots,m$ in which element A_{ij} is of size $\beta_i \times \beta_j$ The diagonal blocks A_{ii} are square matrices. The system of equations

$$f(x) = 0$$

is assumed to have an isolated solution x in a neighbourhood D of which $f \in C^{(1)}(D)$. H is the Jacobian matrix of f. Its (ij)th element is $[H]_{ij} = \frac{\partial f_i}{\partial x_j}$. Its blocks are H_{ij} $i,j = 1,2,\ldots m$ with the above paritioning. The blocks of f are F_i $i=1,2,\ldots,m$. The algorithm to be considered requires a partitioning of H and we assume from now on that it is the one described above. The algorithm is summarized as follows:

Given $x^{(t)}$ and ω compute $x^{(t+1)}$ according to

$$X_i^{(t+1)} = X_i^{(t)} - \omega H_{ii}^{-1}(X_1^{(t+1)}, \ldots X_{i-1}^{(t+1)}, X_i^{(t)}, \ldots X_m^{(t)}) \times$$
$$\times F_i(X_1^{(t+1)}, \ldots, X_{i-1}^{(t+1)}, X_i^{(t)} \ldots, X_m^{(t)}) \qquad (5,1)$$

$i = 1, 2, \ldots, m$

Iterate on t starting with $t=0$. $x^{(0)}$ is a given vector. If $\beta_i = 1$ $i=1,2,\ldots,n$ (so that $m \equiv n$) this algorithm reduces to the nonlinear point SOR algorithm. If the equations $f(x)=0$ are linear it reduces to the well known linear SOR algorithm, with ω as the acceleration parameter. The algorithm $(5,1)$ is not defined if any of the $H_{ii}(x')$ are singular. This cannot arise e.g. if $H(x')$ is positive definite in D.

4.6 Convergence of SOR

The objective of the present section is to discover the factors controlling the asymptotic convergence (i.e. convergence near the root) of algorithm $(5,1)$. In order to do this we shall write $(5,1)$ in another way. Let I_i $i=1,2,\cdots m$ denote that $n\times n$ matrix whose only nonzero elements are 1's in the diagnonal positions occupied by the diagonal elements of $H_{ii}(x')$. Put $y^{(0)} \equiv x^{(0)}$ and consider the algorithm

$$y^{(k)} = y^{(k-1)} - \omega I_k D^{-1}(y^{(k-1)}) f(y^{(k-1)}) \qquad (6,1)$$

$$k = 1, 2, \cdots, m$$

where $D(x')$ is the block diagonal part of $H(x')$ so that the j^{th} block of $D(x')$ is $H_{jj}(x')$. Comparing $(5,1)$ with $t=0$ with $(6,1)$ we see that $x^{(1)}$ of $(5,1)$ and $y^{(m)}$ of $(6,1)$ are identical. $(6,1)$ is just a restatement of the first step of $(5,1)$ for a given initial vector common to each. At the p^{th} step, $1 \le p \le m$ of $(6,1)$ only the p^{th} block of components is altered.

For the purposes of illustrating a point which will arise later, suppose the equations $f(x)=0$ are linear, $f(x) \equiv Ax - b = 0$. From $(6,1)$ we get

$$\begin{aligned} y^{(k)} - x &= y^{(k-1)} - x - \omega I_k D^{-1} A (y^{(k-1)} - x) \\ &= (I - \omega I_k D^{-1} A)(y^{(k-1)} - x) \end{aligned} \qquad (6,2)$$

where now D is the block diagnonal part of A. Thus

$$y^{(m)} - x = x^{(i)} - x = (I - \omega I_m D^{-1} A)(I - \omega I_{m-1} D^{-1} A) \cdots$$

$$\cdots (I - \omega I_i D^{-1} A)(x^{(0)} - x)$$

or

$$\varepsilon^{(i)} = \prod_{i=m}^{1} (I - \omega I_i D^{-1} A)\, \varepsilon^{(0)} \qquad (6,2a)$$

with $x^{(i)} - x = \varepsilon^{(i)}$.

Now, it can be proved that

$$\prod_{i=m}^{1} (I - \omega I_i D^{-1} A) \equiv I - (\omega^{-1} D + E)^{-1} A \qquad (6,3)$$

where E is the (strictly) block lower triangular part of A. The matrix on the right of $(6,3)$ is the known matrix governing successive errors of the block SOR process applied to the system $Ax - b = 0$. Moreover, it is known that if $\rho(I - (\omega^{-1} D + E)^{-1} A) < 1$, convergence of the iteration occurs from an arbitrary starting value.

Let us now return to the nonlinear problem. Consider $(6,1)$. By Taylor's theorem, for $y^{(k-1)} \in \bar{D}$

$$f(y^{(k-1)}) = H(x)(y^{(k-1)} - x) + z^{(k-1)} \qquad (6,4)$$

where $z^{(k-1)}$ is a certain vector satisfying an inequality

$$\| z^{(k-1)} \|_\omega \leq \beta_{k-1} \| y^{(k-1)} - x \|_\omega^2 \qquad (6,4a)$$

where β_{k-1} can be constructed from bounds on the second derivatives of f. Moreover,

$$D^{-1}(y^{(k-1)}) = D^{-1}(x) + R^{(k-1)} \qquad (6,5)$$

where now $R^{(K-1)}$ satisfies an inequality

$$\| R^{(K-1)} \|_\infty \leq C_{K-1} \| y^{(K-1)} - x \|_\infty \qquad (6,5a)$$

and C_{K-1} can be constructed from the same bounds as B_{K-1}.

Using (6,4) and (6,5) we put (6,1) into the form

$$y^{(K)} = y^{(K-1)} - \omega I_K D_x^{-1} (H_x y^{(K-1)} - b) - \omega I_k \left[D_x^{-1} z^{(K-1)} + R^{(K-1)} (H_x y^{(K-1)} - b - z^{(K-1)}) \right]$$

where $H_x \equiv H(x)$, $D_x \equiv D(x)$ and $b \equiv H_x x$. From this,

$$(y^{(K)} - x) = (I - \omega I_K D_x^{-1} H_x)(y^{(K-1)} - x) + \omega I_K \left[D_x^{-1} z^{(K-1)} + R^{(K-1)} H_x (y^{(K-1)} - s) + R^{(K-1)} z^{(K-1)} \right]$$

and the formal similarity between the first term on the right here and in (6,2) can be noticed immediately. Applying this formula recursively we arrive at

$$y^{(m)} - x = x^{(1)} - x = \prod_{i=m}^{1} (I - \omega I_i D_x^{-1} H_x)(x^{(0)} - x) + S^{(1)}$$

$$= \left\{ I - (\omega^{-1} D_x + E_x)^{-1} H_x \right\} (x^{(0)} - x) + S^{(1)}$$

where $S^{(1)}$ is an expression consisting of sums of products of terms of the type $(I - \omega I_K D_x^{-1} H_x)$ and $\omega I_K \left[D_x^{-1} z^{(K-1)} + \right.$ $\left. + R^{(K-1)} H_x (y^{(K)} - x) + R^{(K-1)} z^{(K-1)} \right]$ for various values of K. Putting the errors $x^{(i)} - x = \varepsilon^{(i)}$ we have

$$\varepsilon^{(1)} = \left\{ I - (\omega^{-1} D_x + E_x) H_x \right\} \varepsilon^{(0)} + S^{(1)} \qquad (6,6)$$

and in order to derive information from this we need to know something about $S^{(1)}$.

Suppose, to begin with, that the nonlinear SoR process converges. Then given any small neighbourhood of x there is a T such that all the iterates $x^{(t)}$ with $t \geq T$ lie in it. For such an $x^{(t)}$ we can derive a relationship like (6,6) namely

$$\varepsilon^{(t+1)} = \{ I - (\omega^{-1} D_x + E_x) H_x \} \varepsilon^{(t)} + s^{(t+1)} \qquad (6,7)$$

Recalling the fact that this is derived by considering an iteration of the type (6,1) with $y^{(0)} \equiv x^{(t)}$ and noting that the iterates $y^{(i)}$ $i = 0, 1, \ldots, m$ all lie in the neighbourhood containing $(t \geq T)$ $x^{(t)}$ we can prove that there exists a constant B <u>independent of t</u>, such that

$$\| s^{(t+1)} \|_\omega \leq B \| \varepsilon^{(t)} \|_\omega^2 \qquad (6,8)$$

This follows from (6,4a), (6,5a) and the definition of $s^{(t+1)}$.

For brevity, put $M_\omega = \{ I - (\omega^{-1} D_x + E_x)^{-1} H_x \}$ and assume $\rho(M_\omega) < 1$; from (6,7) we have in view of (6,8).

$$\varepsilon^{(t+1)} = (M_\omega + M'_t) \varepsilon^{(t)}$$

where $\| M'_t \|_\omega \to 0$ as $t \to \infty$. By Lemma 2.1 we have that for all sufficiently large t

$$\| \varepsilon^{(t+1)} \|_\omega \leq \sigma' (\rho(M_\omega) + \varepsilon')^{t+1} \| \varepsilon^{(0)} \|_\omega \qquad (6,9)$$

where $\varepsilon' > 0$ is any positive constant and $\sigma' = \sigma'(\varepsilon')$ is constant. What we have shown by deriving (6,9) is that <u>if</u> the SoR iteration converges, it converges no slower than a geometric progression with ratio $\rho(M_\omega)$. The matrix M_ω is simply the error matrix that would arise if the linear system $H_x x - b = 0$ was solved by SoR using the same blocks as were used in the nonlinear problem.

Although we have omitted some technical details from the above argument it is actually perfectly rigourous.

We have not yet proved that the SOR iteration actually does converge. The proof of this, valid when $\rho(M_\omega) < 1$ rests on a detailed analysis of the perturbed difference equation (6,7). The difficulty arises in proving a relation like (6,8) for the perturbation term. However, a careful consideration of the properties of (6,1) enable a result like (6,8) to be proved, provided $\| \varepsilon^{(0)} \|_\infty$ is sufficiently small. The outcome of this is the following theorem which we shall state, but not prove:

Theorem 6.1

Let f have two continuous derivatives in a neighbourhood of x and suppose $\rho(M_\omega) < 1$. Then there is a neighbourhood of x, $S(\sigma; x)$ such that the sequence $\{x^{(k)}\}$ generated by (5,1) converges to x provided $x^{(0)} \in S(\sigma; x)$. The sequence $\| x - x^{(k)} \|_\infty$ converges to zero no slower than a geometric progression with ratio $\rho(M_\omega)$

Finally, let us observe that the same difficulties arise in the choice of ω as in the linear problem, compounded in the present instance with the fact that H_x is unknown. Furthermore the matrix H_x arising from the boundary value problems considered in part 3 section does not have point property A although there is usually a natural partitioning of it so that it does have block property A.

4.7 Resume and notes

In the part just concluded we have considered the formulation and analysis of several methods of solution of the nonlinear equations arising from our previous work. Examples from two broad classes of methods were considered, and the factors underlying their asymptotic convergence behaviour were isolated.

We shall consider now the relation between the results just given and known results. We shall begin with section 2. The mode of development of the general theory, of this section, is new. However the statement of the Theorem 2.1 was given, in essence, by Ostrowski [47] in 1957. The equivalence of the result given by us and the result of Ostrowski is not immediately evident, but can be proved using the tools of Ortega and Rheinboldt [43] (especially Theorem 10.2.1). This book also contains a wealth of related material.

We turn to section 3. Algorithms 1 and 2 were used independently by the author. However the following must be said; it was pointed out to the author by P. Concus (personal communcation) that Algorithm 1 co-incides with an algorithm proposed by Winslow [48]. However, it does not seem clear from this latter paper that it is this particular algorithm which is being considered. As concerns Algorithm 2, it has been used successfully (with a different discretisation) in [49] to solve the equation of compressible flow. The formulations (3,2) and (3,3) and the subsequent analysis of these algorithms is believed to be new.

In section 4, the question of choosing α is not usually studied in the present generality. But for <u>real</u> roots λ_i the work is standard, [50], [51].

As regards section 6, it was already mentioned in the text that convergence proofs for block nonlinear SOR do not seem easily available. The methods used in this section do not seem to have been used previously. The book [43] contains a more sophisticated approach to the point algorithm, but it appears rather difficult to generalise these methods. The idea of reducing the error matrix to the factored form of (6,3) is allied to the 'splitting' method for time dependent problems as put forward in [52].

Appendix: On computation

1. Introduction

During the course of developing the theory of part 4, various rather unsystematic numerical calculations were carried out by the author in order to obtain an intuitive grasp of the behaviour of the algorithms discussed there. For general interest we give below for each of the algorithms considered a list of a few points that emerged from the calculations. It would of course have been possible to give details of the workings of the programs and detailed numerical results, but the significance of them would be blunted by their highly problem dependent character. It is necessary to emphasize that the following results are put forward for no reason other than general interest. Certainly, no general conclusions should be drawn from them.

2. Test Problems

For the purposes of numerical testing the following problem was constructed and used.

Domain $\mathcal{D}$ (in R_2) $\quad 0 < x < 1 \quad , \quad 0 < y < 1$

Equation: $(1+u_y)^2 u_{xx} - 2u_x u_y u_{xy} + (1+u_x^2) u_{yy} = 0 \qquad (A1)$

Functional associated with (A1): $\int_{\mathcal{D}} (1+u_x^2+u_y^2)^{1/2} \, d\mathcal{D}$

Boundary condition: $\frac{1}{k} \ln\left(\cos \frac{k\pi}{2} y \div \cos \frac{k\pi}{2} x\right)$

The equation is the minimal surface equation. Its physical significance is outlined in e.g. [3]. The function

$$u(x,y) = (k)^{-1} \ln\left(\cos \tfrac{k\pi}{2} y \div \cos \tfrac{k\pi}{2} x\right)$$

satisfies the equation and boundary condition. It is obtained by trivial modifications to one of the few known analytical solutions of (A1) due to Scherk [54]. The parameter k in the boundary conditions was introduced in order to observe the way in which numerical difficulties

increase as u and $|\nabla u|^2 \to \pm\infty$. This behaviour is obtained as $k \to 1$ at the points $(1, 0)$ and $(0, 1)$. For $k \geq 0.9$ the above problem is a reasonable test for the algorithms of part 4.

To solve this problem the triangulation with degree 1 polynomials of part 3 section 4 was used. The appropriate equations were derived there.

The qualitative conclusions of the next section are based on the solution of these equations for triangles whose side lengths in the directions of the axes, (h) were a) $h = 0.1$ b) $h = 0.05$ c) $h = 0.025$. The number of nonlinear equations in each case is 81; 361, and 1521. Each algorithm was used to solve these 3 systems of equations with various values of k taken from $k = 0.5\ (.1)\ 0.9$, and for the SOR algorithms $k = 0.99$, $k = 0.999$, $k = 0.9999$

Starting guesses for the algorithms were obtained by linear interpolation of the boundary conditions.

3. Qualitative behaviour of the algorithms

Algorithm 1

1. This algorithm suffered from instability of the following kind: small changes to the paramters α, ω led to relatively large changes in the required time to solve the problem.
2. In spite of the instability mentioned, there often existed a pair of parameters α', ω' for which the convergence was 'rapid'.
3. When using large values of ω (say $\omega \sim 2$) it was often necessary to have $\alpha < 1$
4. For detailed computational results see [34]

Algorithm 2

1. This was, generally speaking, the slowest of the algorithms, although on the smaller test problems this was less noticible. For the larger test problems and values of k close to $0{\cdot}9$ the algorithm failed to yield solutions in a reasonable period of time.

2. As a rule, it was possible to choose α to achieve reductions in overall computing time by a factor of about $1/3$ relative to the required time with $\alpha = 1$. This factor of $1/3$ did not seem to depend very much on the size of the problem (for fixed k) but appeared to increase as k increased.

3. The linear systems were not solved exactly at each stage as the solution was carried out iteratively. These latter calculations were stopped when the relative change in each component was smaller than 5×10^{-8}

4. For detailed computational results see [49]

Point SOR

1. The nonlinear point SOR algorithm was, overall, the best algorithm for the problems. It never failed to converge in a reasonable period of time, even on problems with $k = 0.9999$. For $k \leq 0{\cdot}9$ it converges between $1\frac{1}{2}$ and 3 times faster that Algorithm 1, which was itself often twice as fast as Algorithm 2.

2. An extrapolation parameter was obtained empirically by carrying out a small number of iterations for a range of values of ω and then working out the area of the surfaces obtained as 'solutions' to the analytical problem. That ω for which the surface area was smallest was chosen for the main calculation. This device was reasonably effective.

3. For detailed computational results see [8] [41].

Block SOR

1. A line of nodes was taken to define a block of the Jacobian.

2. The following (unexplained) effect occurred: for $\kappa \geq 0.9$ as k increased the accuracy of the initial approximation had to increase in order for convergence to occur. In other words, the domain of convergence of the block SOR algorithm decreased as k increased. For large values of k no convergent initial guess could be found. This was only observed with the discretisation considered by us. For another discretisation (due to Concus [8]) it was always possible to obtain convergence with the usual starting guess. It is not known whether the domain of convergence associated with this latter discretisation also decreased as k increased.

3. When convergence did occur, however, the overall computing time was usually only about $\frac{1}{2}$ that required for the point SOR.

4. For detailed computational studies, see [46].

4. Additional remarks.

The above discussions all depend on ρ; e.g. Gelder [49] found that for a different ρ Algorithm 2 was competitive with point SOR. But in making comparisons of this kind we have always to remember that Algorithms 1 and 2 require less information (and less programming) than the SOR algorithms.

It is clear that there is much work to be done in connection with the numerical solution of the equations, and indeed, with setting them up in the first place. It is hoped that the above outlines are a useful guide for further work.

REFERENCES

1. Olga A. Ladyzhenskaya and Nina N. Ural'tseva; Linear and Quasilinear Elliptic Equations. Academic Press, New York and London, 1968

2. J. Serrin; Trans. Roy. Soc. A. 264 (1970) p. 413

3. T. Rado; Problem of Plateau; Chelsea Pub. Co., New York, 1951

4. Charles B. Morrey; Multiple Integrals in the Calculus of Variations, Springer Verlag. Berlin

5. Ascher H. Shapiro; Dynamics and thermodynamics of compressible fluid flow. Ronald Press Co., New York 1953.

6. Murrey H. Protter and Hans F. Weinberger; Maximum Principles in Differential equations. Prentice Hall, Inc., Englewood Cliffs, N.J. 1967.

7. C.W. Cryer; On the numerical solution of a quasi-linear elliptic equation. J. Assoc. Comput. Mach. 14 (1967), 363-375

8. Paul Concus; Numerical solution of the nonlinear magnetostatic - field equation in two dimensions. J. Comp. Phys. 1 (1967), 330-342, (p.339)

9. R. Courant and D. Hilbert; Methods of Mathematical Physics, Interscience Publishers Inc., New York 1962, Volume 2.

10. C. Miranda; Equazioni alle derivate parziali di tipo ellittico; Springer, Berlin 1955.

11. S.G. Mikhlin and K.L. Smolitskiy; Approximate Methods for Solution of Differential and Integral Equations, American Elsevier Pub. Co. Inc, New York, 1967 (p.260)

12. P. Sylvester; symmetric quadrature formulae for simplexes, Math. Comp.24(1969) 95-100.

13. J.L. Synge; Hypercircle in Mathematical Physics, C.U.P. 1957

14. P. Wolfe; The secant method for simultaneous linear equations. Comm ACM. 2 (1959), 12-13

15. A.H. Stroud and Don Secrest; Gaussian Quadrature Formulas, Prentice Hall, Englewood Cliffs, N.J., 1966

16. Andrew H. Wallace; Introduction to Algebraic Topology, Pergamon Press, Oxford, 1957

17. P.G. Ciarlet and C. Wagschal; Multipoint Taylor formulas and applications to the finite element method. Num. Math. 17 (1971) 84-100

18. Milos Zlamal; On the finite element method. Num. Math. 12 (1968) 394-409

19. R.A. Nicolaides; On a class of finite elements generated by Lagrange interpolation. SIAM. Jnl. Num. Analysis. (to appear)

20. R.A. Nicolaides; On Lagrange interpolation in n variables Tech. Note. ICSI 274, Inst. Comp. Sci. London 1970

21. R.A. Nicolaides; On a class of finite elements generated by Lagrange interpolation II. SIAM. Jnl. Num. Analysis (Submitted)

22. R.A. Nicolaides; On a class of finite elements generated by Lagrange interpolation II. Tech. Note. ICSI 329, Inst. Comp. Sci. London 1971.

23. V.I. Smirnov; Course of Higher Mathematics, Pergamon Press, Oxford 1964. (Vol. 5 p.202)

24. O.C. Zienkiewicz; The finite element method in structural and continuum mechanics. McGraw-Hill, London, 1967

25. Milos Zlamal; On some finite element procedures for solving second order boundary value problems. Num. Math 14 (1969) 42-48

26. M.L.J. Hussey; Personal communication

27. Richard Courant and Herbert Robbins; What is Mathematics?, Oxford Univ. Press, London 1941 (p.236)

28. See ref. 23 (Vol 5. p.355)

29. S.G. Mikhlin; Variational Methods in Mathematical Physics, Pergamon Press, Oxford. 1964

30. R. Stepleman ; FINITE DIMENSIONAL ANALOGUES OF VARIATIONAL PROBLEMS IN THE PLANE ; SIAM Jnl. Num. Analysis 8 (1971) p.11

31. R.S. Varga; Matrix Iterative Analysis, Prentice Hall Inc., Englewood Cliffs. N.J., 1962

32. S.G. Mikhlin; Numerical Implementation of Variational Methods, G.T.T.L, Moscow, 1966. (In Russian)

33. D. Greenspan; On approximating extremals of functionals. ICC Bull. 4 (1965) 99-120

34. Alan M. Winslow; Numerical solution of the quasilinear Poisson equation in a nonuniform triangle mesh. J. Comput. Phys. 2 (1967) 149-172

35. G. Forsythe & W. Wasow; Finite Difference Methods for Partial Differential Equations. J. Wiley, New York 1960. (p.227)

36. A.M. Ostrowski; Solution of Equations and Systems of Equations, Academic Press, London and New York 1966. 2nd Ed. (p.153)

37. L. Bers; On mildly nonlinear partial difference equations of elliptic type, J. Res. Nat. Bui. Standards 51 (1953) 229-236

38 H. Lieberstein; Overrelaxation for nonlinear elliptic partial differential equations. M. Res. C., Rept. 80 Univ. of Wisconsin. (1959)

39 Samuel Schecter; Iteration methods for nonlinear problems, Trans. Amer. Math. Soc. 104 (1962) 179-189

40 See ref. 8

41 Paul Concus; Numerical solution of the minimal surface equation. Math. Comp. 21 (1967) 340-350

42 J. Ortega and M. Rockoff; Nonlinear difference equations and Gauss-Seidel type iterative methods. SIAM Jnl. Num. Analysis 3 (1966) 497-513

43 J.M. Ortega and W.C. Rheinboldt; Iterative Solution of Nonlinear equations in several variables; Academic Press, New York 1970.

44. S. Schecter; Relaxation methods for convex problems. SIAM Jnl. Num. Analysis 5 (1968) 601-612

45. S. Schecter; Minimization of a Convex Function by Relaxation; in Integer and Nonlinear Programming ed. J. Abadie Nth. Holland Pub. Co. 1970 (p. 177)

46 P. Concus; Numerical solution of the minimal surface equation by block nonlinear successive overrelaxation. IFIP 1968. 153-158

47 A. Ostrowski; Les points d'attraction et de repulsion pour l'iteration dans l'espace a n dimensions. C.R. Acad. Sci. Paris, 244. 288-289.

48 See ref 34.

49 D. Gelder; Solution of the compressible flow equations. Int. Jnl. Num. Meth. Engrg. 3 (1971) 35-43

50 See ref 35 (p. 225).

51. E. Issacson and H.B. Keller; Analysis of Numerical Methods, John Wiley, New York, 1966. (p.73)

52. N.N. Yanenko; The Method of Fractional Steps, Springer-Verlag Berlin. 1971

53. J. H. Wilkinson; Algebraic Eigenvalue problem, O.U.P., Oxford. 1965

54. D.J. Struik; Differential Geometry, Addison- Wesley Publ. Co. Inc., Mass. 1961 2nd ed. (p. 188 ex.8)