UNCLASSIFIED

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES
3 4456 0361033 1

CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION

ORNL-2146
Physics cy. 4

I. ITERATIVE METHODS FOR THE APPROXIMATE SOLUTION OF LINEAR ALGEBRAIC SYSTEMS

II. SELF-ADJOINTNESS IN ONE-GROUP MULTI-REGION DIFFUSION PROBLEMS

T. W. Hildebrandt

CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION

LIBRARY LOAN COPY

DO NOT TRANSFER TO ANOTHER PERSON

If you wish someone else to see this document, send in name with document and the library will arrange a loan.

OAK RIDGE NATIONAL LABORATORY
OPERATED BY
UNION CARBIDE NUCLEAR COMPANY
A Division of Union Carbide and Carbon Corporation
UCC
POST OFFICE BOX P · OAK RIDGE, TENNESSEE

UNCLASSIFIED

Printed in USA. Price 40 cents. Available from the

Office of Technical Services
U. S. Department of Commerce
Washington 25, D. C.

LEGAL NOTICE

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, express or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission to the extent that such employee or contractor prepares, handles or distributes, or provides access to, any information pursuant to his employment or contract with the Commission.

UNCLASSIFIED

ORNL 2146

Contract No. W-7405-eng-26

MATHEMATICS PANEL

I. ITERATIVE METHODS FOR THE APPROXIMATE SOLUTION OF LINEAR ALGEBRAIC SYSTEMS

II. SELF-ADJOINTNESS IN ONE-GROUP MULTI-REGION DIFFUSION PROBLEMS

Theodore W. Hildebrandt

DATE ISSUED AUG 24 1956

OAK RIDGE NATIONAL LABORATORY
Operated by
UNION CARBIDE NUCLEAR COMPANY
A Division of Union Carbide and Carbon Corporation
Post Office Box P
Oak Ridge, Tennessee

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES
3 4456 0361033 1

UNCLASSIFIED

INTERNAL DISTRIBUTION

1. C. E. Center
2. Biology Library
3. Health Physics Library
4-5. Central Research Library
6. Reactor Experimental Engineering Library
7-34. Laboratory Records Department
35. Laboratory Records, ORNL R.C.
36. A. M. Weinberg
37. L. B. Emlet (K-25)
38. J. P. Murray (Y-12)
39. E. H. Taylor
40. E. D. Shipley
41-42. A. S. Householder
43. C. P. Keim
44. M. L. Nelson
45. W. H. Jordan
46. G. E. Boyd
47. S. C. Lind
48. F. L. Culler
49. R. S. Livingston
50. A. H. Snell
51. A. Hollaender
52. M. T. Kelley
53. J. H. Frye, Jr.
54. K. Z. Morgan
55. T. A. Lincoln
56. C. S. Harrill
57. D. W. Cardwell
58. C. E. Winters
59. E. M. King
60. R. W. Johnson
61. D. D. Cowen
62. J. A. Lane
63. R. A. Charpie
64. S. J. Cromer
65. M. J. Skinner
66. R. R. Dickison
67. N. M. Dismuke
68. A. C. Downing
69. C. L. Gerberich
70. J. W. Givens
71. J. H. Alexander
72. G. W. Medlin
73. C. H. Griffith
74. U. H. Bents
75. ORNL - Y-12 Technical Library, Document Reference Section

EXTERNAL DISTRIBUTION

76. R. F. Bacher, California Institute of Technology
77. Division of Research and Development, AEC, ORO
78-177. Oak Ridge Institute of Nuclear Studies
178-880. Given distribution as shown in TID-4500 under Physics category (300 copies - OTS)

DISTRIBUTION PAGE TO BE REMOVED IF REPORT IS GIVEN PUBLIC DISTRIBUTION

NOTATION

In Part I, the following conventions of notation are observed: matrices are designated by capital letters, vectors by lower-case Roman letters, and scalars by lower-case Greek letters except for indices, and for certain other integers and bounds, such as the dimensionality of a space.

In Part II, these conventions are relaxed, so that, for example, scalar functions are represented by lower-case Roman letters. The notation of Part II should be self-explanatory.

TABLE OF CONTENTS

PART I. ITERATIVE METHODS FOR THE APPROXIMATE SOLUTION OF LINEAR ALGEBRAIC SYSTEMS

I. INTRODUCTION 1

A. Scope of the study 1
B. Definitions and examples 1
C. The three classical iterative methods. 5

II. NORMS AND CONVERGENCE. 13

A. Introduction 13
B. Vector and matrix norms. 13
C. Rate of convergence. 20

III. THE δ^2-PROCESS 24

A. The scalar δ^2-process. 24
B. The vector δ^2-process. 26

IV. THE STEEPEST DESCENT FOR SINGLE STEP PROCESS 33

A. Generation of algorithm for (SDS). 33
B. An estimate for the rate of convergence. 38
C. A method for computation using (SDS) 43

PART II. SELF-ADJOINTNESS IN ONE-GROUP MULTI-REGION DIFFUSION PROBLEMS

V. SELF-ADJOINTNESS IN ONE-GROUP MULTI-REGION DIFFUSION PROBLEMS 45

A. Introduction 45
B. Definitions and statement of problem 46
C. A sufficient condition for self-adjointness. 49
D. Remarks. 53

APPENDIX. 55

Computation of approximate solution of a Dirichlet Problem. Application of δ^2-process

BIBLIOGRAPHY. 58

PART I

ITERATIVE METHODS FOR THE APPROXIMATE SOLUTION OF LINEAR ALGEBRAIC SYSTEMS

I. INTRODUCTION

A. Scope of the Study

This part of the dissertation is concerned with iterative techniques for the approximate solution of non-singular linear algebraic systems, and in particular with a new modification of the so-called Seidel method analogous to the classical method of steepest descent. In the course of the investigation, a vector process analogous to the well-known scalar δ^2-process was defined. It was found that repeated application of this vector δ^2-process to the classical Jacobi iteration for linear systems produces an iteration identical to the classical method of steepest descent. On the other hand, it is shown that the vector δ^2-process does not have the property of the scalar δ^2-process of increasing the order of an iteration.

B. Definitions and Examples

The linear algebraic system under discussion has the matrix form

$$Au = b \tag{1.1}$$

where u is the n-dimensional unknown vector, A is an $n \times n$ symmetric and positive definite* real matrix, and $b \neq 0$ is an n-dimensional constant

---

*The condition that A be symmetric and positive definite is not particularly restrictive, since any linear system with non-singular matrix can be transformed to an equivalent system with symmetric positive definite matrix by multiplying by the transpose of the given matrix. Thus, for non-singular F, the solution of $Fu = g$ is identical with that of $F^T Fu = F^T g$.

real vector. In component form the system is

$$\sum_{k=1}^{n} \alpha_{jk} \eta_k = \beta_j , \qquad j = 1, \ldots, n . \tag{1.2}$$

We shall find it convenient to use the matrix notation wherever possible.

It is also convenient to make the further assumption that the matrix A has unit main diagonal, that is, that all the diagonal elements

$$\alpha_{jj} = 1 , \qquad j = 1, \ldots, n .$$

It is possible to transform any system with symmetric positive definite matrix to this form by the following method: Given the system

$$Fv = g , \tag{1.3}$$

with symmetric, positive definite F. Let D be the matrix of positive (F being positive definite) diagonal elements of F, and $D^{1/2}$ be the diagonal matrix whose non-zero elements are the positive square roots of the corresponding elements of D. Let $D^{-1/2} = (D^{1/2})^{-1}$. Then setting

$$A = D^{-1/2} FD^{-1/2} , \quad u = D^{1/2} v , \quad b = D^{-1/2} g , \tag{1.4}$$

the system

$$Au = b \tag{1.1}$$

has a positive definite matrix A with unit main diagonal.*

In the following we use two decompositions of A,

$$A = I - B \tag{1.5}$$

where $-B$ is the symmetric matrix of non-diagonal elements of A, and has zero main diagonal; and

$$A = I - B_1 - B_1^T \quad , \tag{1.6}$$

where $-B_1$ is the lower triangular matrix of A exclusive of the main diagonal.

Note that

$$B = B_1 + B_1^T \quad . \tag{1.7}$$

If it is possible by a conjugate permutation P of rows and columns of A to reduce A to the block form

---

*When all the diagonal elements of F are equal, $\varphi_{jj} = \delta$, the matrix D becomes the scalar matrix δI; then we may set

$$A = \delta^{-1} F \ , \quad u = v \ , \quad b = \delta^{-1} g \ , \tag{1.4*}$$

since δI commutes with any matrix.

$$(1.8) \qquad PAP^T = \left(\begin{array}{c|c} I_1 & -P_1^T \\ \hline -P_1 & I_2 \end{array}\right)$$

then A is said to have property (A) [Young (1954a), Householder (1955a)].* The transformed matrix will be called a "matrix canonical with respect to property (A)".

It is to be noted that when A is a general symmetric positive definite matrix, the lower triangular submatrix B_1 is nil-potent of index at most n, i.e., there exists a $k \leqslant n$ such that

$$(1.9) \qquad B_1^k = 0 \quad .$$

If A is canonical with respect to property (A), then the lower triangular submatrix B_1 is nil-potent of index 2, whatever the order of the matrix A:

$$(1.10) \qquad B_1 = \begin{pmatrix} 0 & 0 \\ P_1 & 0 \end{pmatrix} \neq 0$$

$$(1.11) \qquad B_1^2 = \begin{pmatrix} 0 & 0 \\ P_1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ P_1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = 0 \quad .$$

This property of matrices canonical with respect to property (A) leads to important simplifications, to be mentioned later, in the theory of Seidel iterations.

---

*Throughout the paper, bibliographic references are given by the date in parentheses following the author's name. Where the references are parenthetical, they are enclosed in square brackets.

Of the several ways in which linear systems (1.1) may arise in applied mathematics, one of the most important is by the construction of the finite difference analog of a boundary value problem with elliptic partial differential equation. Examples of methods by which these analogs may be constructed are found in Collatz (1951), and (for a restricted class of self-adjoint equations) in Young (1954a); there are of course many other similar discussions in the literature.

In many cases of physical interest, the matrices for these finite difference analogs are symmetric and positive definite. This is indeed the case when the given partial differential equation is the Laplace equation (or its generalization discussed by Young (1954a)) and the boundary value problem is the Dirichlet problem, that is, the solution function itself is prescribed at all points of the boundary. In the case of the Dirichlet problem, the matrix A in (1.1) reflects the properties of the differential operator and the shape of the boundary, and the constant vector b contains the boundary information. Each component of the unknown vector u represents the value of the solution of the finite difference equation at a point of the difference grid interior to the region over which the solution is sought.

C. The Three Classical Iterative Methods

By the three classical iterative methods for approximate solution of linear algebraic systems, we mean the Jacobi or "total-step" method, the Seidel or "single-step" method, and the steepest descent method. Of these, the first two differ from each other in the decomposition of the matrix A employed, and the last is an optimized version of the first.

(1) The Jacobi method, was, according to Forsythe (1953), first published by Jacobi (1845). We shall hereafter use the symbol (T), standing for the descriptive term "total-step" (Gesamtschritt), as a designation for this method.

The method (T) is based on the decomposition (1.3) of the matrix A. Starting with (1.1), we have

$$Au = (I - B)u = b$$

$$u = Bu + b$$

which leads to the iteration

$$u_{i+1} = Bu_i + b \tag{1.12}$$

where u_i stands for the ith approximation to the exact solution u.

In studying this and the other iterations discussed in this paper, it is often convenient to write the algorithm in terms of the error or <u>remainder</u>

$$s_i \equiv u_i - u \tag{1.13}$$

and/or the <u>residue</u>

$$r_i \equiv As_i \equiv Au_i - b \quad . \tag{1.14}$$

One should observe that for practical purposes s_i is unavailable, since u is unknown, while r_i is available. We then have for (T)

$$(1.15) \qquad \begin{aligned} s_{i+1} &= Bs_i \\ &= (I - A)s_i \\ &= s_i - r_i \quad . \end{aligned}$$

Since the matrix B is constant, the iteration can also be written

$$(1.16) \qquad s_{i+1} = B^{i+1} s_o$$

where s_o corresponds to the (arbitrary) starting vector u_o .

Clearly, (1.16) converges to zero for symmetric A if and only if the eigenvalues μ_j of B $(= 1 - \lambda_j$, where λ_j are the corresponding eigenvalues of A) lie in the interval

$$(1.17) \qquad -1 < \mu_j < +1 \quad , \qquad j = 1, \ldots, n \quad .$$

Hence we have the

<u>Theorem</u> [Stein and Rosenberg (1948)] <u>For symmetric</u> A , <u>the iteration</u> (T) <u>converges if and only if both</u> $A = I - B$ <u>and</u> $I + B$ <u>are positive definite</u>.

The rate of convergence* is given by

$$(1.18) \qquad R(T) = - \ell n \left(\max_j |\mu_j|\right) .$$

---

*For a discussion of the notion of "rate of convergence" in matrix iterations, see Chapter II.

(2) The method commonly known as the Seidel method was apparently first advocated by Nekrasov (1884) who attributed it to Seidel [Forsythe (1953), p. 5 footnote] . We shall hereafter use the symbol (S) , standing for the descriptive term "single-step" (Einzelschritt), as a designation for this method.

The method (S) is based on the decomposition (1.6) of the matrix A . Starting with (1.1) we have

$$Au = (I - B_1 - B_1^T)\, u = b$$

$$(I - B_1)u = B_1^T u + b$$

which leads to the iteration

$$(I - B_1)\, u_{i+1} = B_1^T u_i + b \tag{1.19a}$$

or

$$u_{i+1} = (I - B_1)^{-1} (B_1^T u_i + b) \tag{1.19b}$$

or, adding and subtracting u_i on the right hand side,

$$u_{i+1} = u_i - (I - B_1)^{-1} (Au_i - b) \quad . \tag{1.19c}$$

In terms of s_i and r_i , we have

$$(1.20) \qquad s_{i+1} = s_i - (I - B_1)^{-1} As_i = s_i - (I - B_1)^{-1} r_i$$

$$= (I - B_1)^{-1} B_1^T s_i \quad .$$

Setting $C \equiv (I - B_1)^{-1} B_1^T$, we have also

$$(1.21) \qquad s_{i+1} = C^{i+1} s_o \quad .$$

Reich (1949) has shown that positive definiteness of the matrix A is a necessary and sufficient condition for the convergence of (S) .

As far as the rate of convergence is concerned, there is a theorem of Stein and Rosenberg (1948): Let γ be the maximum modulus of the eigenvalues of C , and μ be the maximum modulus of the eigenvalues of B . If the elements of B , and hence of B_1 and B_1^T , are non-negative, then either $\gamma < \mu < 1$ or $\gamma = \mu = 1$ or $\gamma > \mu > 1$. That is, (S) and (T) converge or diverge together; if (T) converges, then (S) converges more quickly.

In case the matrix A is canonical with respect to property (A) , (S) converges if and only if (T) converges, and if both converge, (S) converges twice as fast as (T) [Young (1954a)] . The proof is based on the fact that B_1 is nil-potent of index 2. This permits the immediate computation of the exact inverse of $I - B_1$:

$$(1.22) \qquad (I - B_1)^{-1} = I + B_1 \quad ,$$

for

$$\begin{pmatrix} I_1 & 0 \\ P_1 & I_2 \end{pmatrix}\begin{pmatrix} I_1 & 0 \\ -P_1 & I_2 \end{pmatrix} = \begin{pmatrix} I_1 & 0 \\ 0 & I_2 \end{pmatrix} = \begin{pmatrix} I_1 & 0 \\ -P_1 & I_2 \end{pmatrix}\begin{pmatrix} I_1 & 0 \\ P_1 & I_2 \end{pmatrix} .$$

Then

$$(1.23) \qquad C = (I + B_1)B_1^T = \begin{pmatrix} I_1 & 0 \\ P_1 & I_2 \end{pmatrix}\begin{pmatrix} 0 & P_1^T \\ 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & P_1^T \\ 0 & P_1 P_1^T \end{pmatrix}$$

and the characteristic equation is

$$(1.24) \qquad \det(\gamma I - C) = \det \begin{pmatrix} \gamma I_1 & -P_1^T \\ 0 & \gamma I_2 - P_1 P_1^T \end{pmatrix} = 0 .$$

But if μ is a non-zero eigenvalue of B ,

$$0 = \det \begin{pmatrix} \mu I_1 & -P_1^T \\ -P_1 & \mu I_2 \end{pmatrix} = \det \left[\begin{pmatrix} I_1 & 0 \\ P_1 & \mu I_2 \end{pmatrix}\begin{pmatrix} \mu I_1 & -P_1^T \\ -P_1 & \mu I_2 \end{pmatrix} \right]$$

$$= \det \begin{pmatrix} \mu I_1 & -P_1^T \\ 0 & \mu^2 I_2 - P_1 P_1^T \end{pmatrix}$$

since $\mu \neq 0$ implies

$$\det \begin{pmatrix} I_1 & 0 \\ P_1 & \mu I_2 \end{pmatrix} \neq 0 \quad .$$

Therefore, the non-zero eigenvalues of C are squares of the non-zero eigenvalues of B . [Young (1954a), Householder (1955a)] .

(3) As indicated previously, the classical method of steepest descent is a modification of (T) . Forsythe (1953) finds the earliest exposition in Cauchy (1847). Because of its relation to (T) , we shall use the obvious designation (SDT) for this method.

The development of (SDT) may be traced as follows: Introduce the scalar parameter α_i into (1.13) to obtain

$$s_{i+1} = s_i - \alpha_i r_i \tag{1.25}$$

or the equivalent

$$u_{i+1} = u_i - \alpha_i (Au_i - b) \quad . \tag{1.26}$$

Geometrically, (1.25) locates s_{i+1} as a point on the line in the direction of r_i through the point s_i . The parameter α_i is now selected to fix s_{i+1} as the point at which the line is tangent to an

ellipsoidal surface of the family $s^T As = k$; that is, to minimize the quadratic form

$$(1.27) \qquad s_{i+1}^T As_{i+1} = s_i^T As_i - 2\alpha_i s_i^T Ar_i + \alpha_1^2 r_i^T Ar_i \quad .$$

The minimum is obtained by taking

$$(1.28) \qquad \alpha_i = \frac{s_i^T Ar_i}{r_i^T Ar_i} = \frac{r_i^T r_i}{r_i^T Ar_i} \quad .$$

The algorithm (1.25), (1.28) converges for every positive definite symmetric A [Householder (1955a)] and does so at an average rate over m steps

$$(1.29) \qquad R_{av}(\mathrm{SDT}) \geqslant \ell n \frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1} - \frac{1}{2m} \ell n \frac{\lambda_n}{\lambda_1}$$

where $0 < \lambda_1 \leqslant \ldots \leqslant \lambda_n$ are the eigenvalues of A [Kantorovič (1948, 1948a)] .

II. NORMS AND CONVERGENCE

A. Introduction

As indicated in Chapter I, the notion of rate of convergence must be precisely defined. In order to do this we must first introduce the notions of vector and matrix norms, and also define what the concept "convergence" is to mean.

The material of this chapter is not, for the most part, new, but largely follows the exposition of Householder (1954b, 1955a). However, it is felt that it has been possible to put certain aspects of norms and convergence in a finite dimensional vector space into a clearer light, and the chapter is included for this reason.

B. Vector and Matrix Norms

For convenience of reference, we collect here the definitions and some of the properties of vector and matrix norms. For a more complete discussion, see Householder (1954b, 1955a).

Definition. In a vector space X, a (vector) norm is a bounded real valued function $\|x\|$ of the elements $x \in X$ satisfying the following three axioms:

I. $\|x\| \geqslant 0$; $\|x\| = 0$ if and only if $x = 0$.

II. $\|\alpha x\| = |\alpha| \, \|x\|$, where α is a (real or complex) scalar.

III. $\|x_1 + x_2\| \leqslant \|x_1\| + \|x_2\|$.

We are concerned here with the convergence of sequences in an n-dimensional vector space, and in particular, whether convergence in an arbitrary norm is sufficient to insure convergence. It is well known

that the introduction of a notion of convergence or of a notion of norm defines a topology in a space; either notion may be taken as basic, and leads to a "natural" definition of the other, at least in a finite-dimensional space. In other words, either the notion of convergence, or the notion of a basic norm must serve as an undefined term.

In considering the iterative solution of a linear algebraic system, one is doubtless concerned that each component of the approximating vector u_i can be made arbitrarily close to the corresponding component of the solution vector u. Hence it is natural to take as basic the notion of convergence by components.

Henceforth, then, let us restrict our attention to an n-dimensional linear space X spanned by n independent basis vectors f_j. In terms of the basis, any vector $u \in X$ has a representation

$$u = \sum_{j}^{n} \varphi_j f_j \tag{2.1}$$

where the scalars φ_j are the components of the vector u. Our notion of convergence is defined in terms of convergence of scalar sequences as follows:

<u>Definition</u>. <u>A sequence</u> $\{u_i\} \subset X$ <u>is said to converge to an element</u> $u \in X$ <u>if and only if</u> $\lim \varphi_{ji} = \varphi_j$, <u>all</u> $j = 1, \ldots, n$; <u>i.e., for every</u> $\epsilon > 0$ <u>there exists an</u> $N > 0$ <u>such that</u> $m > N$ <u>implies</u> $|\varphi_{jm} - \varphi_j| < \epsilon$, <u>all</u> $j = 1, \ldots, n$.

Convergence in norm also is defined in terms of convergence of a sequence of scalars:

Definition. Suppose a norm $\| \ \|$ is defined in the space X. A sequence $\{u_i\}$ is said to converge in norm to an element $u \in X$ if and only if for every $\epsilon > 0$ there exists an $N > 0$ such that $m > N$ implies $\|u_m - u\| < \epsilon$; i.e., $\lim_i \|u_i - u\| = 0$.

It is easy to show that the function $\|u\| = \max_j |\varphi_j|$ is a norm, and that with this norm, convergence in norm is identical with convergence.

Theorem. For whatever norm defined on X, convergence in norm is equivalent to convergence; that is, convergence in norm implies convergence, and conversely.

This theorem is probably a classical one for finite-dimensional normed linear spaces.

Proof*. (a) Suppose $\{u_i\}$ converges to a limit u, i.e., for all j, $\lim_i \varphi_{ji} = \varphi_j$. Then

$$\|u_i - u\| = \|\sum_j^n \varphi_{ji} f_j - \sum_j^n \varphi_j f_j\|$$

$$= \|\sum_j^n (\varphi_{ji} - \varphi_j) f_j\|$$

$$\leqslant \sum_{j=1}^n |\varphi_{ji} - \varphi_j| \, \|f_j\| \quad .$$

Since $\lim_i |\varphi_{ji} - \varphi_j| = 0$, all j , the sum on the right of the inequality can be made as small as desired by taking i sufficiently large. Hence $\lim_i \|u_i - u\| = 0$.

---

*I wish to thank T. H. Hildebrandt for suggesting this proof.

(b) For the converse, we need the following lemma.

<u>Lemma</u>. [Riesz, F. (1917)]. If the sequence $\{\|u_i\|\}$ is bounded, then each component sequence $\{\varphi_{ji}\}$ is bounded.

Suppose at least one component sequence is unbounded. Then by a possible rearrangement, $\lim_i \sum_j^n |\varphi_{ji}| = \infty$. Consider the sequence $\{\psi_{ji}\}$, where $\psi_{ji} = \varphi_{ji}/\sum_j^n |\varphi_{ji}|$. Clearly, $|\psi_{ji}| \leqslant 1$, so $\{\psi_{ji}\}$ is a bounded sequence and, by the Weierstrass-Bolzano theorem, contains a convergent subsequence $\{\psi_{ji_k}\}$, with $\lim_k \psi_{ji_k} = \psi_j$. Since

$$\sum_{j=1}^n |\psi_{ji}| = \sum_{j=1}^n |\varphi_{ji}| \Big/ \sum_{j=1}^n |\varphi_{ji}| = 1, \text{ all } i,$$

it follows that

$$\sum_{j=1}^n |\psi_j| = 1 ,$$

so that not all $\psi_j = 0$. Let $v_i = \sum_{j=1}^n \psi_{ji} f_j$. Since $\{\|u_i\|\}$ is bounded,

$$\lim_k \|v_{i_k}\| = \lim_k \left\{ \|u_{i_k}\| \Big/ \sum_{j=1}^n |\varphi_{ji_k}| \right\} = 0 .$$

Hence $v = \lim_k v_{i_k} = 0$, which, together with the independence of the basis vectors $\{f_j\}$, leads to a contradiction; the lemma follows.

Now suppose $\{u_i\}$ converges to u in norm, i.e.,

$$\lim_i \|u_i - u\| = \lim_i \|s_i\| = 0$$

but that for some j_o , there exists a $\delta > 0$ and an $M > 0$ such that $|\sigma_{j_o i}| > \delta$ for all $i > M$. The sequence $\|s_i\|$ is bounded, hence, by the lemma, the sequence of components $\{\sigma_{ji}\}$ is bounded for all j . Applying the Weierstrass-Bolzano theorem again, it is possible to pick a subset $\{i_k\}$ of the indices i such that the subsequence $\{\sigma_{ji_k}\}$ converges for all j . Let $\lim_k \sigma_{ji_k} = \sigma_j$. The corresponding limit vector $s' = \sum_j^n \sigma_j f_j$ has $\|s'\| = 0$, since a subsequence of a convergent sequence has the same limit. It follows that all $\sigma_j = 0$, by the independence of the basis, and in particular $\sigma_{j_o} = 0$. This yields a contradiction, and completes the proof of the theorem.

For the matrix norm, we have the following definition

<u>Definition</u>. <u>A matrix norm is real-valued function</u> $\|A\|$ <u>of a square matrix</u> A , <u>bounded in the sense that if the elements of the matrix are bounded, then the norm is bounded; and satisfying the following four axioms</u> (the first three are identical to the three for a vector norm):

I. $\|A\| \geqslant 0$; $\|A\| = 0$ if and only if $A = 0$ (where 0 is the matrix consisting of zero elements.)

II. $\|\alpha A\| = |\alpha| \, \|A\|$, where α is any scalar.

III. $\|A + B\| \leqslant \|A\| + \|B\|$.

IV. $\|AB\| \leqslant \|A\| \, \|B\|$.

A matrix norm is said to be <u>consistent</u> with a given vector norm if for every matrix A and vector u ,

$$\|Au\| \leqslant \|A\| \, \|u\| \quad .$$

A matrix norm is said to be <u>subordinate</u> to a given vector norm if it is consistent, and, further, for every A there exists a $u \neq 0$ such that

$$\|Au\| = \|A\| \, \|u\| \quad .$$

In particular, if the matrix norm is subordinate, then $\|I\| = 1$. The notions of subordination and consistency seem to be first introduced by Fadeeva (1950, p. 61), (1950a, p. 85). See also Householder, loc. cit.

Householder has proved a number of theorems concerning matrix norms of which we may need the following:

1) For the sequence of powers A^p of a matrix to converge to zero, it is sufficient that $\|A\| < 1$, however the norm may be defined.

2) Let λ be an eigenvalue of largest modulus of the matrix A . Then for any $\epsilon > 0$, there exists a norm such that $\|A\| \leqslant |\lambda| + \epsilon$. Moreover, if all eigenvalues of modulus $|\lambda|$ are simple roots of the characteristic equation, then there exists a norm such that $\|A\| = |\lambda|$.

The Euclidean norm of a numerical vector x , or a vector referred to a unit orthogonal basis, is defined by

$$\|x\| = (x^* x)^{1/2}$$

where x^* is the conjugate transpose of the vector x. The matrix norm subordinate to the Euclidean vector norm is the <u>spectral</u> norm*: the largest singular value of the matrix, i.e., the largest positive square root of the eigenvalues of A^*A, where A^* is the conjugate transpose of the matrix A. Clearly, in case A is complex Hermitian, or real symmetric, the spectral norm is the maximum modulus of its eigenvalues**, and thus is the limiting case referred to in 2) above. On the other hand, for more general classes of matrices, the spectral norm may exceed the maximum modulus of the eigenvalues: consider for example a matrix

---

*This definition differs from that used by Young (1954a). To form a matrix norm subordinate to a given vector norm following Fadeeva (1950, 1950a) one may proceed as follows: Given some vector norm $\|x\|$. Let

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} .$$

It is easy to show that this definition satisfies the axioms for matrix norm. That it is subordinate follows directly from the definition.

In particular, when $\|x\| = (x^* x)^{1/2}$,

$$\|A\| = \max_{x} \frac{(x^* A^* Ax)^{1/2}}{(x^* x)^{1/2}} = \max \mu^{1/2}$$

where μ is an eigenvalue of the non-negative semi-definite Hermitian matrix A^*A; thus the spectral norm defined above is obtained.

**When A is itself Hermitian with real eigenvalues λ_i, then there exists a unitary transformation U such that

$$A = U^* \Lambda U , \quad \text{where} \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) .$$

But

$$A^*A = U^* \Lambda U U^* \Lambda U = U^* \Lambda^2 U .$$

Hence

$$\max \mu = \max \lambda^2 , \text{ or } \|A\| = \max |\lambda| .$$

whose Jordan normal form has all diagonal elements equal to λ and all elements of the super-diagonal equal to one. Thus, with $n = 2$, let

$$A = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}, \quad \text{with real} \quad \lambda > 0 .$$

$$A^* A = A^T A = \begin{pmatrix} \lambda^2 & 1 \\ \lambda & 1 + \lambda^2 \end{pmatrix} .$$

The eigenvalues of $A^T A$ are

$$\mu = \frac{1}{2} (2 \lambda^2 + 1 \pm \sqrt{4 \lambda^2 + 1})$$

and

$$\mu_{max} \geqslant \frac{1}{2} (2\lambda^2 + 2\lambda + 1) > \lambda^2$$

or

$$\|A\| > \max |\lambda| .$$

C. Rate of Convergence

An iteration of the types mentioned in Chapter I may be represented by the formula

(2.2) $$s_{i+1} = T_i s_i$$

where the transformation T_i may vary with the index i through dependence on s_i as in (SDT), or may be constant as in (T) and (S). This formulation is written in terms of the remainder $s_i = u_i - u$, where u is the solution of the given system (1.1). Clearly, any iteration can be expressed in terms of s_i, since only a shift in the origin is involved. And the sequence $\{u_i\}$ converges to the solution u if and only if $\{s_i\} \longrightarrow 0$, a fact we have already made use of in (B).

In general, in studying the convergence of a process, we consider its effect on a measure, such as the remainder s_i, of the deviation of an element in the sequence from the limit of the sequence. However, in considering the approximate solution of linear systems the remainder is useless, because the solution u is unknown. As an alternative, we may multiply (2.2) by the matrix A of the system (1.1), and, <u>providing</u> T_i <u>and</u> A <u>commute</u>, we have

$$r_{i+1} = As_{i+1} = AT_i\, s_i = T_i\, As_i = T_i r_i \quad . \tag{2.3}$$

The residue r_i is computable, as has been pointed out in Chapter I.

Let us now assume that a suitable measure of deviation of the ith iterate from the true solution exists, say r_i, and that the iteration is defined by

$$r_{i+1} = T_i\, r_i \tag{2.3'}$$

(where T_i may differ from that in (2.2) and (2.3) if the matrices in

(2.3) do not commute). Define the "instantaneous" rate of convergence at the ith step by

$$(2.4) \qquad R(T_i) = -\ln \frac{\|r_{i+1}\|}{\|r_i\|} = -\ln \frac{\|T_i\, r_i\|}{\|r_i\|}$$

where the norm is any vector norm. The introduction of a matrix norm subordinate to the vector norm leads to

$$(2.5) \qquad R(T_i) \geqslant -\ln \frac{\|T_i\|\ \|r_i\|}{\|r_i\|} = -\ln \|T_i\| .$$

The term on the right of the inequality obviously provides a lower bound for the rate of convergence at a given step, which depends on the choice of matrix norm. We may then use it as an estimate for the rate of convergence, and designate it

$$(2.6) \qquad R_N(T_i) = -\ln \|T_i\|$$

where the index N designates the norm chosen. In view of property (2) of norms quoted above, the best estimate R_N for a given T_i is obtained by using the minimum norm for the matrix, which can be made as close as desired to the maximum modulus of the eigenvalues of T_i. When the transformation is symmetric, this is the spectral norm, as indicated earlier.

The <u>average</u> rate of convergence over m steps can be defined

$$(2.7) \qquad R_{av} = \frac{1}{m} \sum_{i=0}^{m-1} R(T_i) = -\frac{1}{m} \ell n \frac{\|r_m\|}{\|r_o\|} = -\frac{1}{m} \ell n \frac{\left\|\left(\prod_{i=0}^{m-1} T_i\right) r_o\right\|}{\|r_o\|}$$

$$\geqslant -\frac{1}{m} \ell n \left\|\prod_{i=0}^{m-1} T_i\right\| \geqslant -\frac{1}{m} \sum_{i=0}^{m-1} \ell n \|T_i\| = R_{Nav} .$$

If T is independent of i, and the minimum matrix norm assumes the maximum modulus Λ of the eigenvalues of T, then

$$(2.8) \qquad R_{Nav}(T) = -\ell n \Lambda$$

which is the definition used by Young (1954a) for the rate of convergence of T.

Note that the second equality of (2.7) furnishes an algorithm for the direct computation of the average rate of convergence, useful particularly in experimental studies of iterative techniques, and requiring no prior knowledge of the eigenvalues of the transformations T_i. When the measure of the deviation of the approximate solution from the true solution is chosen to be the residue r_i, there is the added convenience that most iterative methods require the computation of r_i at each step. The choice of the norm is not restricted, although it may affect the value for the rate. One possible norm is the e-norm [Householder (1955a)], the maximum absolute value of the components of the given vector; another is the e'-norm, the sum of the absolute values of the components. Both of these norms have an advantage over the Euclidean norm in requiring less computation for their evaluation.

III. THE δ^2-PROCESS

A. The Scalar δ^2-Process

The scalar δ^2-process was first applied in general numerical analysis by Aitken (1926), but was apparently used in astronomical computations by Delaunay about 1860 [Andoyer (1902)] . The δ^2-process and its generalizations have been extensively studied in two recent papers by Lubkin (1952) and Shanks (1954, 1955).

The δ^2-process is a transformation involving the first and second differences of a sequence which, under certain conditions, has the property of transforming a convergent (or divergent) sequence into a more quickly convergent (or more slowly divergent) sequence. It also has the property of increasing the order of a scalar iterative process [Householder (1953)] .

Let $\{\xi_i\}$ be a sequence of scalars. The first difference

$$\delta \xi_i = \xi_{i+1} - \xi_i \tag{3.1}$$

and the second difference

$$\begin{aligned} \delta^2 \xi_i &= \delta \xi_{i+1} - \delta \xi_i \\ &= \xi_{i+2} - 2\xi_{i+1} + \xi_i \ . \end{aligned} \tag{3.2}$$

Then the δ^2-process defines a derived sequence $\{\xi_i^*\}$:

$$\xi_i^* = \xi_i - \frac{(\delta\,\xi_i)^2}{\delta^2\,\xi_i} \quad . \tag{3.3}$$

Equation (3.3) may also be written in a form involving a determinant, and permitting the simplification

$$\xi_i^* = \frac{1}{\delta^2\,\xi_i}\begin{vmatrix} \xi_i & \delta\xi_i \\ \delta\xi_i & \delta^2\,\xi_i \end{vmatrix} = \frac{1}{\delta^2\,\xi_i}\begin{vmatrix} \xi_i & \xi_{i+1} \\ \xi_{i+1} & \xi_{i+2} \end{vmatrix} \tag{3.4}$$

by row and column operations.

In applying the δ^2-process to a scalar iteration of the form

$$\xi_{i+1} = f(\xi_i) \tag{3.5}$$

one must compute at least two successive iterates ξ_{i+1} , ξ_{i+2} . Then ξ_i^* computed from ξ_i , ξ_{i+1} , ξ_{i+2} may be taken as a new starting value ξ_{i+3} and the process may be repeated [Householder (1953)] .

The δ^2-process has been used with varying degrees of success to accelerate matrix iterations (T) or (S) . In the case at least of the iteration (T) it is possible to justify its application [Householder (1954a)] . The method of application is as follows: The matrix iteration

$$u_{i+1} = F(u_i) \tag{3.6}$$

where u_i is a vector, and F is a matrix function, defines a sequence of vectors $\{u_i\}$. Each component sequence $\left\{\eta_j^{(i)}\right\}$, $j = 1, \ldots, n$, is treated as an independent scalar sequence, and the scalar δ^2-process used accordingly.

Of course, the component sequences are <u>not</u> actually independent, being related by the iterative algorithm, which takes the component form

$$\eta_j^{(i+1)} = \varphi_j(\eta_1^{(i)}, \ldots, \eta_j^{(i)}, \ldots, \eta_n^{(i)}) \quad . \tag{3.7}$$

For the report of an experimental study of the application of the scalar δ^2-process to the iteration (S) for a system of order 361, which arises in a Dirichlet problem for the Laplace equation, see the Appendix.

B. The Vector δ^2-Process

Since matrix iterations deal with vectors rather than independent sequences of scalars, it seemed desirable to define a process analogous to the scalar δ^2-process which takes account of the vector nature of the iteration.

The vector δ^2-process can be simply derived by introducing the notion of a scalar product to account for the fact that the reciprocal of a vector cannot be defined: For a sequence of vectors $\{u_i\}$

$$u_i^* = u_i - \frac{(\delta u_i)^T \delta u_i}{(\delta^2 u_i)^T \delta u_i} \delta u_i \tag{3.8}$$

where

$$\delta u_i = u_{i+1} - u_i \tag{3.9}$$

$$\delta^2 u_i = \delta u_{i+1} - \delta u_i = u_{i+2} - 2u_{i+1} + u_i \quad . \tag{3.10}$$

It is clear that for a sequence of scalars, this vector δ^2-process is identical with the scalar δ^2-process.

<u>Theorem</u>. <u>The application of the vector δ^2-process to the iteration</u> (T) <u>yields an algorithm identical to that for</u> (SDT).

That is, given a starting vector u_i . If one proceeds by (T) and the vector δ^2-process to find u_i^* , and by (SDT) to obtain v_{i+1} , then $u_i^* = v_{i+1}$.

<u>Proof</u>. From (1.12), (T) becomes

$$\begin{aligned} u_{i+1} &= u_i - (A u_i - b) \\ &= u_i - r_i \end{aligned} \tag{3.11}$$

and from (1.15) $r_{i+1} = (I - A)r_i$. Then

$$\delta u_i = -r_i$$

$$\delta^2 u_i = -(r_{i+1} - r_i) = Ar_i$$

and

$$u_i^* = u_i - \frac{r_i^T r_i}{r_i^T Ar_i} r_i \quad . \tag{3.12}$$

But, from (1.26), (1.28),

$$v_{i+1} = u_i - \frac{r_i^T r_i}{r_i^T A r_i} r_i = u_i^* .$$

<u>Theorem</u>. <u>The application of the vector δ^2-process to (S) yields a new iteration which converges provided the matrix $EA \equiv (I - B_1)^{-1} A$ has a positive definite quadratic form.</u>

<u>Proof</u>. From (1.19c), (S) can be written

$$u_{i+1} = u_i - Er_i \quad , \text{ where } \quad E \equiv (I - B_1)^{-1}$$

or

(3.13a) $$u_{i+1} = u_i - t_i \quad , \quad \text{where} \quad t_i \equiv E\, r_i$$

and from (1.20)

$$t_{i+1} = t_i - EAt_i \quad .$$

Then

(3.13b) $$\delta\, u_i = - t_i$$

(3.13c) $$\delta^2\, u_i = - (t_{i+1} - t_i) = EAt_i$$

and

$$(3.14) \qquad u_i^* = u_i - \frac{t_i^T t_i}{t_i^T EAt_i} t_i \quad .$$

The process defined by

$$(3.15) \qquad \bar{u}_{i+1} = \bar{u}_i - \frac{\bar{t}_i^T \bar{t}_i}{\bar{t}_i^T EA\bar{t}_i} \bar{t}_i$$

where $\bar{u}_o = u_o$, $\bar{t}_o = t_o$, $\bar{u}_1 = u_o^*$ in (3.14) is a new iteration based on (S).

The proof of convergence is deferred to Chapter IV, where it follows as Corollary 1 of Theorem 4.1.

The question arises whether the vector δ^2-process increases the order of a matrix iteration. That it does not do so, in general, can be deduced from a study of (SDT), which we have shown to be the result of applying the vector δ^2-process to the basic iteration (T).

For the purposes of this discussion, the notations and definitions of Ludwig (1954b) are in order:

Given an iterative algorithm

$$(3.16a) \qquad x_{k+1} = \bar{f}(x_k) \text{ , where } x_k = \begin{pmatrix} \xi_1^{(k)} \\ \vdots \\ \xi_n^{(k)} \end{pmatrix}$$

and

$$\bar{f}(x_k) = \begin{pmatrix} \varphi_1(\xi_1^{(k)}, \ldots, \xi_n^{(k)}) \\ \vdots \\ \varphi_n(\xi_1^{(k)}, \ldots, \xi_n^{(k)}) \end{pmatrix} . \tag{3.16b}$$

Define the column operator

$$d = \begin{bmatrix} \frac{\partial}{\partial \xi_1} \\ \vdots \\ \frac{\partial}{\partial \xi_n} \end{bmatrix} \tag{3.17}$$

and let

$$s_k = x_k - \bar{x} , \quad \text{where} \quad \bar{x} = \bar{f}(\bar{x}) . \tag{3.18}$$

Then the iteration can be written

$$s_{k+1} = (s_k^T d)\, \bar{f}(\bar{x}) + \frac{1}{2!} (s_k^T d)^2\, \bar{f}(\bar{x}) + \ldots + \frac{1}{n!} (s_k^T d)^n\, \bar{f}(\bar{x}) + \ldots \tag{3.19}$$

<u>Definition</u>. <u>If the pth term in the Taylor series (3.19) is the first non-zero term, then the iteration is said to be of the pth order</u>.

For (T) , we have by (1.15)

$$s_{k+1} = B\, s_k \quad \text{and} \quad \bar{s} = 0 . \tag{3.20}$$

Written in component form with $\sigma_j^{(k)}$ the components of s_k , e_j the jth basis vector, and using the index notation for summation

$$(3.21) \qquad (s_k^T d)Bs = \sigma_j^{(k)} \frac{\partial}{\partial \sigma_j} \beta_{pq} \sigma_q e_p$$

$$= \sigma_j^{(k)} \beta_{pj} e_p = Bs_k$$

$$(s_k^T d)^2 Bs = \sigma_j^{(k)} \sigma_\ell \frac{\partial}{\partial \sigma_j} \frac{\partial}{\partial \sigma_\ell} \beta_{pq} \sigma_q e_p = 0 .$$

Hence (3.20) describes an iteration of the first order. Incidentally, it is clear that when $\bar{f}(x) = Fx$ for any constant matrix F , the operator $s^T d$ is a substitution operator.

Now, for (SDT) we have algorithm (1.25),(1.28)

$$s_{k+1} = \bar{f}(s_k) = s_k - \alpha_k As_k$$

$$\alpha_k = \frac{s_k^T A^2 s_k}{s_k^T A^3 s_k} .$$

Then

$$(3.22) \qquad (s_k^T d)(s - \alpha As) = s_k - \alpha As_k - As(s_k^T d)\alpha$$

$$= (I - \alpha A)s_k - 2 \frac{s^T A^2 (I - \alpha A)s_k}{s^T A^3 s} As .$$

With $\|x\|_A^2 = x^T Ax$, and letting $I - \alpha A = C_\alpha$

$$(3.23) \qquad \|(s_k^T d)\ (s - \alpha As)\|_A^2 =$$

$$= \|C_\alpha s_k\|_A^2 - 4\frac{s^T A^2 C_\alpha s_k}{s^T A^3 s} s^T A^2 C_\alpha s_k + 4\left(\frac{s^T A^2 C_\alpha s_k}{s^T A^3 s}\right)^2 s^T A^3 s$$

$$= \|C_\alpha s_k\|_A^2 \quad .$$

Thus, with $\alpha_o = \lim_{s \to 0} \alpha \geqslant \lambda_1^2/\lambda_n^3 > 0$

$$(3.24) \qquad \lim_{s \to 0} \|(s_k^T d)(I - \alpha A)s\|_A = \|(I - \alpha_o A)s_k\|_A \neq 0$$

unless $1/\alpha_o$ is an eigenvalue of A and s_k is the corresponding eigenvector. We must conclude that (SDT) is in general a first order process; hence, that the vector δ^2-process does not increase the order of every matrix iteration.

IV. THE STEEPEST DESCENT FOR SINGLE STEP PROCESS

The classical steepest descent process (SDT) optimizes, in a certain sense, the basic process (T). In an analogous fashion, it is possible to find a steepest descent process which optimizes the basic single step process (S); this new process is called steepest descent for single step or (SDS).

A. Generation of Algorithm for (SDS)

The basic iteration (S) as given by (3.13) is modified by the introduction of a scalar parameter β_i .

(4.1a) $$u_{i+1} = u_i - \beta_i t_i$$

or

(4.1b) $$s_{i+1} = s_i - \beta_i t_i$$

where, as before,

$$t_i = Er_i = EAs_i$$

$$E \equiv (I - B_1)^{-1} .$$

Let H be an arbitrary matrix with positive definite quadratic form ($x^T Hx > 0$ for all $x \neq 0$) , and let $\|x\|_H^2 \equiv x^T Hx$. Consider

$$(4.2) \qquad \|s_{i+1}\|_H^2 = \|s_i\|_H^2 - 2\beta_i\, t_i^T Hs_i + \beta_i^2\, t_i^T Ht_i \ .$$

In order that $\|s_{i+1}\|_H$ be a minimum,

$$(4.3) \qquad \beta_i = \frac{t_i^T Hs_i}{t_i^T Ht_i} \ .$$

The algorithm (SDS) is given by (4.1) and (4.3).

When β_i satisfies (4.3),

$$(4.4) \qquad \|s_{i+1}\|_H^2 = \|s_i\|_H^2 - \frac{(t_i^T Hs_i)^2}{t_i^T Ht_i} < \|s_i\|_H^2$$

<u>provided</u> that if $s_i \neq 0$, t_i is not H-orthogonal to s_i ; i.e., provided $t_i^T Hs_i \neq 0$.

There are at least two choices of H which satisfy this condition: (1) $H = EA$, when EA has positive definite quadratic form; this is the algorithm (3.15) obtained by applying the vector δ^2-process to (S) ; (2) $H = A$, the positive definite matrix of the linear system (1.1). Furthermore, in each of these cases, the algorithm depends on the vectors t_i and r_i , which are available, and not on s_i , which is actually unavailable.

<u>Theorem</u> 4.1. <u>If for positive definite</u> H , $s_i \neq 0$ <u>implies</u> $t_i^T Hs_i \neq 0$ <u>then</u> (SDS) <u>converges, and the limit of the sequence</u> $\{u_i\}$ <u>is</u> u , <u>that is, the sequence</u> $\{s_i\}$ <u>converges to</u> 0 .

<u>Proof</u>. Since $\|s_i\|_H$ is a monotonically decreasing sequence of positive numbers, it converges to a unique non-negative limit. We shall

show that the limit is zero, hence $\lim_{t \to \infty} s_i = 0$. For suppose the unique limit of the sequence of norms is $N > 0$. Then there exists a convergent subsequence $\left\{s_{i_k}\right\}$ such that

$$\lim_{i_k \to \infty} s_{i_k} = \bar{s} \quad \text{and} \quad \|\bar{s}\|_H = N .$$

Now consider

$$N^2 = \lim_{i_k \to \infty} \|s_{i_k+1}\|_H^2 = \lim_{i_k \to \infty} \left[\|s_{i_k}\|_H^2 - \frac{(t_{i_k}^T \; Hs_{i_k})^2}{t_{i_k}^T \; H \; t_{i_k}} \right]$$

$$= N^2 - \frac{(\bar{t}^T \; H\bar{s})^2}{\bar{t}^T \; H \; \bar{t}} < N^2$$

since the fraction is a continuous function of s_{i_k} , being a rational function of its components with non-zero denominator, and $\bar{s} \neq 0$ implies $\bar{t} = EA\bar{s} \neq 0$ because the matrix EA , the product of two non-singular matrices, is non-singular. Thus we have a contradiction; hence

$$\lim_{i \to \infty} \|s_i\|_H^2 = 0$$

and therefore

$$\lim_{i \to \infty} s_i = 0 \quad . \tag{4.5}$$

<u>Corollary</u> 1. If EA has positive definite quadratic form, then setting $H = EA$ in (4.3) and (4.1) produces a convergent iteration (3.15).

<u>Proof</u>. If $H = EA$, (4.3) becomes

$$\beta_i = \frac{t_i^T EAs_i}{t_i^T EAt_i} = \frac{t_i^T t_i}{t_i^T EAt_i} \tag{4.6}$$

and (4.4) becomes

$$\|s_{i+1}\|_{EA}^2 = \|s_i\|_{EA}^2 - \frac{(t_i^T t_i)^2}{t_i^T EAt_i} \quad . \tag{4.7}$$

Since $s_i \neq 0$ implies $t_i = EAs_i \neq 0$ and $t_i^T EAs_i = t_i^T t_i > 0$, the conditions of the theorem are satisfied.

Note that the matrix EA does not have positive definite quadratic form for all positive definite A , or even for A canonical with respect to property (A) and having non-positive off-diagonal elements. For consider

$$A = \begin{pmatrix} 1 & -\alpha \\ -\alpha & 1 \end{pmatrix}, \qquad \alpha > 0 \ .$$

This 2 x 2 matrix is canonical with respect to property (A) and has non-positive off-diagonal elements. It is positive definite if the principal minors are positive, i.e., if

$$|\alpha| < 1 \quad .$$

But, from (1.22),

$$E = \begin{pmatrix} 1 & 0 \\ \alpha & 1 \end{pmatrix}$$

$$EA = \begin{pmatrix} 1 & -\alpha \\ 0 & 1 - \alpha^2 \end{pmatrix}$$

and has the quadratic form given by its symmetric part

$$\frac{1}{2}\left(EA + (EA)^T\right) = \begin{pmatrix} 1 & -\alpha/2 \\ -\alpha/2 & 1-\alpha^2 \end{pmatrix}$$

which is positive definite only if

$$\alpha^2 < \frac{4}{5} \quad .$$

Thus, for the example, EA is not positive definite for $\frac{4}{5} \leqslant \alpha^2 < 1$.

<u>Corollary</u> 2. Setting $H = A$ in (4.3) and (4.1) yields a convergent iteration defined by (4.1) and

$$\beta_i = \frac{t_i^T As_i}{t_i^T At_i} = \frac{t_i^T r_i}{t_i^T At_i} = \frac{r_i^T Er_i}{t_i^T At_i} \quad . \tag{4.8}$$

Lemma. The matrix E has a positive definite quadratic form.

Proof*. Since the quadratic form on any real matrix depends only on its symmetric part, we have, for any $x \neq 0$,

$$(4.9) \qquad 2x^T Ex = x^T(E^T + E)x$$

$$= x^T E^T(E^{-1} + E^{T^{-1}})Ex$$

$$= x^T E^T(I + A)Ex$$

$$= x^T E^T Ex + x^T E^T AEx > 0 ,$$

for $E^T E$, being the product of a non-singular matrix by its transpose, is positive definite; the quadratic form $x^T E^T AEx$ is positive definite because A is positive definite; and $x \neq 0$ implies $Ex \neq 0$, E being non-singular.

Proof of Corollary 2. Since $s_i \neq 0$ implies $r_i = As_i \neq 0$, and E has positive definite quadratic form, then $s \neq 0$ implies $t_i^T As_i \neq 0$, so the hypotheses of the theorem are satisfied.

B. A Crude Estimate for the Rate of Convergence of (SDS) when $H = A$.

Theorem 4.2. *If $H = A$, where A is positive definite and is canonical with respect to property (A), then the average rate of*

---

*The method of proof follows from a suggestion of Professor Jack E. McLaughlin.

<u>convergence of</u> (SDS) <u>over</u> N <u>steps is bounded from below</u>:

$$(4.10) \qquad R_{av}(SDS) > \frac{1}{2}\,\ell n\,\frac{\lambda_n}{\lambda_n - \frac{1}{2}\lambda_1 + \frac{1}{4}\lambda_1\lambda_n} - \frac{1}{2N}\ell n\,\frac{\lambda_n}{\lambda_1}$$

where λ_1 is the least, λ_n the greatest eigenvalue of A.

<u>Proof</u>. From (4.4), setting $H = A$, we have

$$(4.11) \qquad \|s_{i+1}\|_A^2 = \|s_i\|_A^2 - \frac{(r_i^T Er_i)^2}{r_i^T E^T AEr_i} .$$

We ask for an upper bound on $\dfrac{\|s_{i+1}\|_A^2}{\|s_i\|_A^2}$ for $\|s_i\|_A \neq 0$. From (4.9),

$$r_i^T Er_i = \frac{1}{2} r_i^T E^T (I + A)Er_i = \frac{1}{2} t_i^T (I + A)t_i$$

also

$$r_i^T E^T AEr_i = t_i^T At_i$$

and

$$\|s_i\|_A^2 = s_i^T As_i = r_i^T A^{-1} r_i .$$

Then

$$(4.12) \qquad \max \frac{\|s_{i+1}\|_A^2}{\|s_i\|_A^2} = 1 - \frac{1}{4} \min \frac{(t_i^T (I + A)t_i)^2}{(r_i^T A^{-1} r_i)(t_i^T At_i)} .$$

But

$$(4.13) \qquad \min \frac{(t_i^T (I+A)t_i)^2}{(r_i^T A^{-1} r_i)(t_i^T A t_i)} \geqslant \frac{(1+\lambda_1)^2 (t_1^T t_i)^2}{\lambda_1^{-1} r_i^T r_i \lambda_n t_i^T t_i} = \frac{\lambda_1 (1+\lambda_1)^2 t_i^T t_i}{\lambda_n r_i^T r_i} .$$

Now, under the assumption that A is canonical with respect to property (A),

$$\begin{aligned} t_i^T t_i &= r_i^T E^T E r_i \\ &= r_i^T (I + B_1^T)(I + B_1) r_i \\ &= r_i^T (I + B_1^T + B_1 + B_1^T B_1) r_i \\ &= r_i^T (2I - A) r_i + r_i^T B_1^T B_1 r_i \\ &\geqslant r_i^T (2I - A) r_i \end{aligned}$$

since $B_1^T B_1$, being the product of a singular matrix by its transpose is positive semi-definite. Then

$$(4.14) \qquad t_i^T t_i \geqslant (2 - \lambda_n) r_i^T r_i .$$

Substituting from (4.14) into (4.13), we have

$$\min \frac{(t_i^T (I + A)t_i)^2}{(r_i^T A^{-1} r_i)(t_i^T A t_i)} \geqslant \frac{\lambda_1 (1 + \lambda_1)^2 (2 - \lambda_n)}{\lambda_n}$$

and from (4.12)

$$\max \frac{\|s_{i+1}\|_A^2}{\|s_i\|_A^2} \leqslant 1 - \frac{\lambda_1(1 + \lambda_1)^2 (2 - \lambda_n)}{4\lambda_n}$$

$$\leqslant \frac{\lambda_n - \frac{1}{2}\lambda_1 + \frac{1}{4}\lambda_1 \lambda_n}{\lambda_n} .$$

Then

$$\frac{\|s_{i+1}\|_A}{\|s_i\|_A} < \left(\frac{\lambda_n - \frac{1}{2}\lambda_1 + \frac{1}{4}\lambda_1 \lambda_n}{\lambda_n}\right)^{1/2}$$

or

$$\frac{\|s_N\|_A}{\|s_o\|_A} < \left(\frac{\lambda_n - \frac{1}{2}\lambda_1 + \frac{1}{4}\lambda_1 \lambda_n}{\lambda_n}\right)^{N/2} .$$

To write the inequality in terms of the Euclidean norm $\| \ \|_E$, we note that for any $x \neq 0$,

$$\lambda_1^{1/2} \| x \|_E \leqslant \| x \|_A \leqslant \lambda_n^{1/2} \| x \|_E \quad .$$

Hence

$$\frac{\| s_N \|_E}{\| s_o \|_E} < \frac{\lambda_n^{1/2}}{\lambda_1^{1/2}} \left(\frac{\lambda_n - \frac{1}{2} \lambda_1 + \frac{1}{4} \lambda_1 \lambda_n}{\lambda_n} \right)^{N/2}$$

and

$$R_{av} \text{ (SDS)} > \frac{1}{2} \ln \left(\frac{\lambda_n}{\lambda_n - \frac{1}{2} \lambda_1 + \frac{1}{4} \lambda_1 \lambda_n} \right) - \frac{1}{2N} \ln \frac{\lambda_n}{\lambda_1} \quad .$$

In order to obtain this estimate, we have used some rather crude inequalities, as, for example, in (4.13), where no account is taken of the fact that the numerator and denominator of the fraction are not independent. Because of this crudeness, the estimate lacks sufficient precision to permit a comparison between (SDS) with $H = A$ and the well-known methods, in particular (SDT). In short, the theorem serves at best as an alternate proof of the convergence of the method.

In view of the known facts that when the matrix B of (1.5) has all positive elements, (S) converges faster than (T), and that (SDT) converges at least as fast as (T), one would expect that (SDS) converges at least as fast as (S), and faster than (SDT) under similar circumstances. But the problem remains to find a sufficiently precise estimate for the rate of convergence of (SDS). It is possible that

this problem will be solved by finding a way to express the eigenvalues of $E^T AE$ in terms of those of A, or some other approach may prove to be effective.

C. A Method for Computation using (SDS)

For computation of the vectors appearing in either of the (SDS) algorithms, it is convenient to make use of the differences formed by using (S) and/or (T). We have already seen how the algorithm with $H = EA$ arises naturally by applying the vector δ^2-process to (S). Recall, from (3.13) that (changing notation) if u_i' is formed from u_i by applying (S):

$$u_i' = u_i - E(Au_i - b) \tag{4.15}$$

then

$$t_i = -\delta u_i = u_i - u_i' \tag{4.16}$$

and if t_i' is formed from t_i by the same algorithm with $b = 0$, then

$$EAt_i = -\delta t_i = t_i - t_i' . \tag{4.17}$$

With t_i and EAt_i, it is fairly easy to compute the fraction (4.6).

For the case $H = A$, we need also the vectors r_i and At_i. For their computation one may introduce also the algorithm (T) (1.13) to compute

(4.18) $$u_i'' = u_i - (Au_i - b)$$

whence

(4.19) $$r_i = -\delta u_i = u_i - u_i'' ;$$

similarly,

(4.20) $$t_i'' = t_i - At_i$$

so that

(4.21) $$At_i = -\delta t_i = t_i - t_i'' .$$

PART II

SELF-ADJOINTNESS IN ONE-GROUP MULTIREGION DIFFUSION PROBLEMS

V. SELF-ADJOINTNESS IN ONE-GROUP MULTIREGION DIFFUSION PROBLEMS

A. Introduction

Multi-region diffusion problems are of current interest in nuclear reactor theory, in the theory of biological systems, in heat transfer problems, etc. In its most general form, the problem is concerned with the diffusion of several substances simultaneously, which interact with one another according to certain laws of equilibrium, a situation described by a coupled system of partial differential equations. In nuclear reactor theory, such a problem is known as a "multi-group" problem.

The present discussion is limited to a one-group, multiregion problem, i.e., we assume a single diffusing substance, so that in each subregion the steady-state distribution of the substance satisfies a single partial differential equation. The laws under which the diffusing substance passes from one subregion into another contiguous subregion are described by interface conditions, while the conditions at the outer boundary describe the relations of the system of regions to the rest of the world.

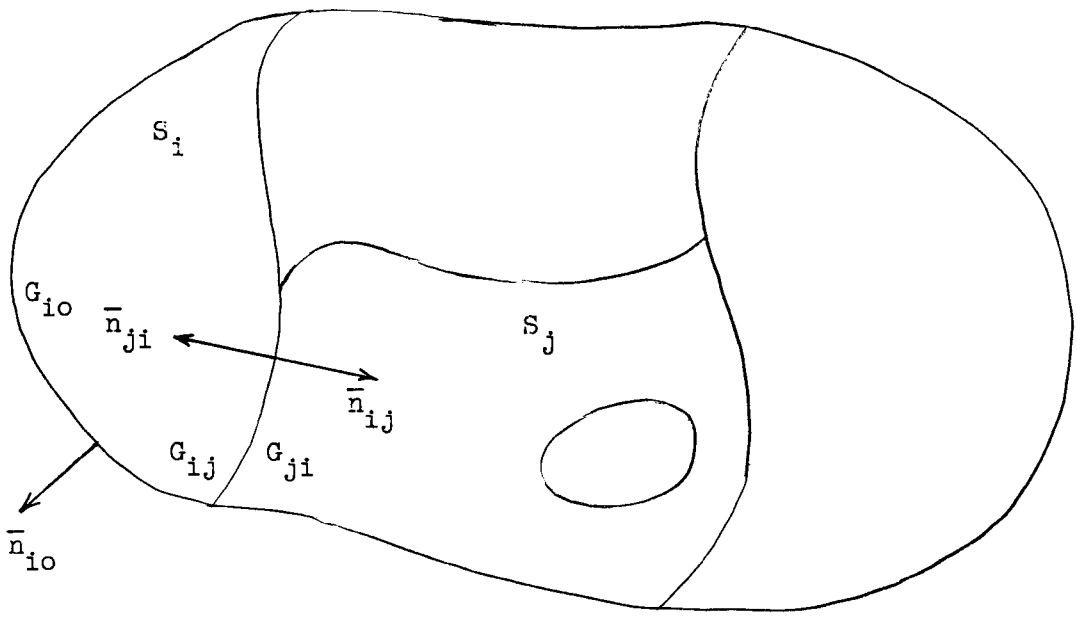
It is of particular interest to know under what circumstances the problem is self-adjoint, so that, for example, its eigenvalues are real, and eigen-functions corresponding to distinct eigenvalues are orthogonal.

In this part of the present paper, we develop a condition which, when satisfied at the interfaces, is sufficient to insure the self-adjointness of a class of multi-region one-group problems. This is an

extension to a more general situation of a result obtained by Sangren (1953); Sangren limited his discussion to the case where the sub-regions are topologically equivalent to the annular regions between concentric spheres, while the discussion here applies to an arbitrary arrangement of interfaces.

B. Definitions and Statement of Problem

Consider the bounded connected region S which is subdivided into N subregions S_i, $i = 1, \ldots, N$, by the piecewise smooth surfaces G_{ij}, $i = 1, \ldots, N$; $j = 0, 1, \ldots, N$. The surfaces



G_{ij} and G_{ji} are identical, except that their "outward" normals are oppositely directed:

$$\bar{n}_{ji} = -\bar{n}_{ij} \quad . \tag{5.1}$$

The surface G_{io} is the portion of the outer boundary of S which is common to the boundary of S_i .

Let there be defined in each subregion S_i the linear differential operator with real coefficients

$$L_i(u_i) = \nabla^2 u_i + D_i u_i , \quad i = 1, \ldots, n \tag{5.2}$$

where D_i is a real function of position continuous in S_i , and u_i is any function with first derivatives continuous in $S_i \cup \bigcup_{j=1}^{N} G_{ij}$ and second derivatives continuous in S_i .

The following boundary and interface conditions are to be satisfied on each G_{ij} , $i = 1, \ldots, N$; $j = 0, \ldots, N$:

$$\left\{ \begin{aligned} u_i &= \alpha_{ij} u_j + \beta_{ij} \frac{\partial u_j}{\partial n_{ij}} \\ \frac{\partial u_i}{\partial n_{ij}} &= \gamma_{ij} u_j + \delta_{ij} \frac{\partial u_j}{\partial n_{ij}} \end{aligned} \right. \tag{5.3}$$

where $\alpha_{ij}, \ldots, \delta_{ij}$ are real and

$$\varphi_{ij} \equiv (\alpha_{ij} \delta_{ij} - \gamma_{ij} \beta_{ij}) \neq 0 \quad (j \neq 0) \tag{5.4}$$

$$\varphi_{io} = 0 \quad .$$

The condition on the outer boundary G_{io} contains as special cases the

conditions

$$u_o = 0 \quad , \quad \text{or} \quad \frac{\partial u_o}{\partial n_{io}} = 0 \quad .$$

Note that where two subregions S_i and S_j have no common interface, we may assume that a non-singular $(\varphi_{ij} \neq 0)$ condition is satisfied vacuously, since $G_{ij} = 0$; and also that $\varphi_{ii} = 1$, $G_{ii} = 0$, which is to say that the function u_i is to be continuous with continuous first derivative within any given subregion S_i .

Note also that $\varphi_{ij} \neq 0$ implies

$$\varphi_{ji} = 1/\varphi_{ij} \quad . \tag{5.5}$$

For, by interchanging indices in (5.3), we have

$$u_j = \alpha_{ji}\, u_i + \beta_{ji} \frac{\partial u_i}{\partial n_{ji}} = \alpha_{ji}\, u_i - \beta_{ji} \frac{\partial u_i}{\partial n_{ij}}$$

$$\frac{\partial u_j}{\partial n_{ji}} = \gamma_{ji}\, u_i + \delta_{ji} \frac{\partial u_i}{\partial n_{ji}} = \gamma_{ji}\, u_i - \delta_{ji} \frac{\partial u_i}{\partial n_{ij}} = - \frac{\partial u_j}{\partial n_{ij}}$$

whence we may solve for u_i , $\frac{\partial u_i}{\partial n_{ij}}$, by applying (5.3) to obtain

$$\begin{pmatrix} \alpha_{ij} & \beta_{ij} \\ \gamma_{ij} & \delta_{ij} \end{pmatrix} \begin{pmatrix} \alpha_{ji} & -\beta_{ji} \\ -\gamma_{ji} & \delta_{ji} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

or

$$\varphi_{ij}\,\varphi_{ji} = 1 \quad .$$

C. A Sufficient Condition for Self-Adjointness

A set of functions $\{u_i\}$ satisfying the differentiability conditions in each S_i, and the interface and boundary conditions on each G_{ij}, $i = 1, \ldots, N$; $j = 0, \ldots, N$, are the components of an admissible function U defined on S. The operator L_i defined on each S_i induces an operation $L(U)$ defined on S.

Definition. $L(U)$ is self-adjoint provided there exists a set of N positive constants k_i such that for any two admissible functions $U \neq V$,

$$\sum_{i=1}^{N} k_i \iiint_{S_i} \left[v_i\, L_i(u_i) - u_i\, L_i(v_i)\right] d\tau = 0 \quad . \tag{5.6}$$

Theorem. A sufficient condition that $L(U)$ be self-adjoint is that

$$\begin{aligned} &\varphi_{ih}\,\varphi_{hj}\,\varphi_{ji} = 1 \quad \text{for all } 0 < i, j, h \leq N \\ &\varphi_{ij} > 0 \quad \text{for all } 0 < i, j \leq N \quad . \end{aligned} \tag{5.7}$$

Proof. By applying Green's theorem to the left side of (5.6) we obtain

$$(5.8)\quad \sum_{i=1}^{N} k_i \iiint_{S_i} \left[v_i L_i(u_i) - u_i L_i(v_i)\right] d\tau = \sum_{i=1}^{N} k_i \iiint_{S_i} \left[v_i \nabla^2 u_i - u_i \nabla^2 v_i\right] d\tau$$

$$= \sum_{i=1}^{N} k_i \sum_{j=0}^{N} \iint_{G_{ij}} \left(v_i \frac{\partial u_i}{\partial n_{ij}} - u_i \frac{\partial v_i}{\partial n_{ij}}\right) d\sigma \quad .$$

For each $i < j$, and for $j = 0$ we introduce the boundary and interface conditions (5.3):

$$k_i \iint_{G_{ij}} \left(v_i \frac{\partial u_i}{\partial n_{ij}} - u_i \frac{\partial v_i}{\partial n_{ij}}\right) d\sigma$$

$$= k_i \iint_{G_{ij}} (\alpha_{ij}\delta_{ij} - \beta_{ij}\gamma_{ij}) \left(v_j \frac{\partial u_j}{\partial n_{ij}} - u_j \frac{\partial v_j}{\partial n_{ij}}\right) d\sigma$$

$$= k_i \varphi_{ij} \iint_{G_{ij}} \left(v_j \frac{\partial u_j}{\partial n_{ij}} - u_j \frac{\partial v_j}{\partial n_{ij}}\right) d\sigma \quad .$$

Since $\varphi_{io} = 0$, all terms with $j = 0$ vanish. The right-hand side of (5.8) then becomes (upon interchanging the order of summation)

$$(5.9)\quad \sum_{j=1}^{N} \sum_{i=1}^{j-1} k_i \varphi_{ij} \iint_{G_{ij}} \left(v_j \frac{\partial u_j}{\partial n_{ij}} - u_j \frac{\partial v_j}{\partial n_{ij}}\right) d\sigma$$

$$+ \sum_{j=1}^{N} \sum_{i=j}^{N} k_i \iint_{G_{ij}} \left(v_i \frac{\partial u_i}{\partial n_{ij}} - u_i \frac{\partial v_i}{\partial n_{ij}}\right) d\sigma \quad .$$

The second term of (5.9) can be transformed by interchanging the order of summation, "renaming" indices, and observing that G_{ji} can be replaced by G_{ij} if n_{ji} is replaced by $-n_{ij}$, as follows.

$$\sum_{j=1}^{N} \sum_{i=j}^{N} k_i \iint_{G_{ij}} \left(v_i \frac{\partial u_i}{\partial n_{ij}} - u_i \frac{\partial v_i}{\partial n_{ij}}\right) d\sigma = \sum_{i=1}^{N} \sum_{j=1}^{i-1} k_i \iint_{G_{ij}} \left(v_i \frac{\partial u_i}{\partial n_{ij}} - u_i \frac{\partial v_i}{\partial n_{ij}}\right) d\sigma$$

$$= \sum_{j=1}^{N} \sum_{i=1}^{j-1} k_i \iint_{G_{ji}} \left(v_j \frac{\partial u_j}{\partial n_{ji}} - u_j \frac{\partial v_j}{\partial n_{ji}}\right) d\sigma$$

$$= - \sum_{j=1}^{N} \sum_{i=1}^{j-1} k_j \iint_{G_{ij}} \left(v_j \frac{\partial u_j}{\partial n_{ij}} - u_j \frac{\partial v_j}{\partial n_{ij}}\right) d\sigma \quad .$$

Upon substitution of this term in (5.9) we obtain

$$\sum_{j=1}^{N} \sum_{i=1}^{j-1} (k_i \varphi_{ij} - k_j) \iint_{G_{ij}} \left(v_j \frac{\partial u_j}{\partial n_{ij}} - u_j \frac{\partial v_j}{\partial n_{ij}}\right) d\sigma \quad . \tag{5.10}$$

From (5.10) it follows that

$$k_i \varphi_{ij} - k_j = 0 , \qquad 0 < 1 < j \leqslant N \tag{5.11}$$

is sufficient to insure the vanishing of the right-hand side of (5.8); that is, (5.11) is a sufficient condition for (5.6).

We now observe that (5.11) implies

$$k_j = k_i \varphi_{ij} = k_h \varphi_{hj} = k_i \varphi_{ih} \varphi_{hj} , \qquad i < h < j .$$

Therefore

$$\varphi_{ij} = \varphi_{ih}\,\varphi_{hj} \quad ,$$

and in view of (5.5) and the condition $k_i > 0$, $i = 1, \ldots, N$, (5.7) is equivalent to (5.11).

Theorem 5.2. *Given the eigenvalue problem*

$$L_i(u_i) = \lambda\,\theta_i\,u_i, \quad \theta_i \geqslant 0 \;,\; i = 1, \ldots, N; \quad \sum_{i=1}^{N} \theta_i > 0 \tag{5.12}$$

where λ *is independent of* i , θ_i *is a real number constant in each* S_i , *and the* u_i *are components of an admissible function* U . *If* $L(U)$ *is self-adjoint, then the eigenvalues are real and eigen-functions corresponding to distinct eigenvalues are orthogonal over those subregions* S_i *where* $\theta_i \neq 0$.

Proof. Suppose there exists a complex eigenvalue λ , and a corresponding admissible solution U , such that $\theta_i > 0$ implies $u_i \neq 0$ in S_i . Because all the coefficients in the equation (5.12) and the boundary and interface conditions are real, the complex conjugates $\bar{\lambda}$, $\bar{U}$ must also satisfy, i.e., in each S_i

$$L_i(u_i) = \lambda\,\theta_i\,u_i \tag{5.13a}$$

$$L_i(\bar{u}_i) = \bar{\lambda}\,\theta_i\,\bar{u}_i \quad . \tag{5.13b}$$

Multiplying (5.13a) and (5.13b) by $\bar{u}_i$ and u_i respectively, and integrating the difference, we have

$$(\lambda - \bar{\lambda})\, \theta_i \iiint_{S_i} \bar{u}_i\, u_i d\tau = \iiint_{S_i} (\bar{u}_i \nabla^2 u_i - u_i \nabla^2 \bar{u}_i) d\tau\,, \quad i = 1, \ldots, N.$$

Since $L(U)$ is self-adjoint, there exists a set of $k_i > 0$ such that

$$(\lambda - \bar{\lambda}) \sum_{i=1}^{N} k_i\, \theta_i \iiint_{S_i} u_i\, \bar{u}_i\, d\tau = 0 \quad .$$

Then $\lambda - \bar{\lambda} = 0$, since $u_i\, \bar{u}_i = |u_i|^2 > 0$, and not all $\theta_i = 0$; λ must be real.

The proof of orthogonality follows the same lines, except that $\lambda - \mu \neq 0$ implies

$$\iiint_{S_i} u_i\, v_i\, d\tau = 0 \quad \text{over each } S_i \text{ where } \theta_i > 0 \,.$$

D. <u>Remarks</u>

1. In certain frequently occurring special cases, condition (5.7) is automatically satisfied.

(a) Continuity across the interfaces:

$$\begin{cases} u_i = u_j & \text{on all } G_{ij}\,,\ j \neq 0 \\[2ex] g_i \dfrac{\partial u_i}{\partial n_{ij}} = g_j \dfrac{\partial u_j}{\partial n_{ij}} & g_i > 0\,,\ 0 < i \leqslant N\,. \end{cases}$$

Then $\varphi_{ij} = g_j/g_i > 0$

$$\varphi_{ih}\,\varphi_{hj}\,\varphi_{ji} = (g_h/g_i)\,(g_j/g_h)(g_i/g_j) = 1 \ .$$

(b) Contact resistance at interfaces:

$$p_i \frac{\partial u_i}{\partial n_{ij}} = p_j \frac{\partial u_j}{\partial n_{ij}} = h_{ij}\,(u_i - u_j) \quad \text{on all} \quad G_{ij}\ ,\ j \neq 0$$

$$\left.\begin{array}{l} p_i > 0 \\ \\ h_{ij} \neq 0 \end{array}\right\} \quad \text{all} \quad 0 < i\ ,\ j \leqslant N$$

can be rewritten

$$\left\{\begin{array}{l} u_i = u_j + \dfrac{p_j}{h_{ij}} \dfrac{\partial u_j}{\partial n_{ij}} \\ \\ \dfrac{\partial u_i}{\partial n_{ij}} = \dfrac{p_j}{p_i} \dfrac{\partial u_j}{\partial n_{ij}} \end{array}\right.$$

whence $\varphi_{ij} = p_j/p_i > 0$ and $\varphi_{ih}\,\varphi_{hj}\,\varphi_{ji} = 1$.

2. While Theorem 5.2 gives assurance that the eigenvalues of the operator $L(U)$ will be real when the interface conditions satisfy (5.7), one may well ask whether the result can be extended to the finite difference analog of the problem. Such an extension has been made by Mr. A. C. Downing of the Oak Ridge National Laboratory, in work soon to be published in an ORNL Report.

APPENDIX

Computation of approximate solution of a Dirichlet problem for Laplace's equation. Application of scalar δ^2-process.

The problem was as follows: On the square $0 \leqslant x, y \leqslant 1$, find a function u such that

$$\nabla^2 u = 0 \quad , \tag{A.1}$$

and on the boundaries

$$u(0, y) = \frac{1}{4} \sin \pi y \tag{A.2}$$

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

A square grid with grid-spacing $h = \frac{1}{20}$ was laid on the region so that there were $19^2 = 361$ interior grid-points, and 76 active boundary points (the values at the corner points do not enter into the computation). The simplest finite-difference representation of the Laplace equation was used:

With $\eta_{j,k} = u(jh, kh), \quad 0 < j, k < 20$,

$$\eta_{j,k} = \frac{1}{4} (\eta_{j,k-1} + \eta_{j,k+1} + \eta_{j-1,k} + \eta_{j+1,k}) \quad . \tag{A.3}$$

The basic iteration (S), often called Liebmann's method [Liebmann (1918)] in this special case, was used. It takes the form

$$\eta_{j,k}^{(i+1)} = \frac{1}{4}\left(\eta_{j,k-1}^{(i+1)} + \eta_{j-1,k}^{(i+1)} + \eta_{j,k+1}^{(i)} + \eta_{j+1,k}^{(i)}\right) . \tag{A.4}$$

The computation proceeds successively through all the interior grid-points in the order

$$(j, k) = (1, 1), \ldots, (1, 19), (2, 1), \ldots, (19, 19) \tag{A.5}$$

for each complete iteration.

For the machine computation, the boundary values and initial estimates of the solution $(\eta_{j,k}^{(0)} = 0,\ 0 < j, k < 20)$ were stored in the order

$$\begin{aligned}(j, k) = (0, 0), (0, 1) \ldots (0, 20); (1, 0) \ldots (1, 20);\\ (2, 0) \ldots (20, 20) .\end{aligned} \tag{A.6}$$

The differences

$$\delta\, \eta_{j,k}^{(i)} = \eta_{j,k}^{(i+1)} - \eta_{j,k}^{(i)} \tag{A.7}$$

were computed and stored in the order (A.5), while the newly computed values $\eta_{j,k}^{(i+1)}$ were stored in the positions which originally held the initial estimates. In view of (4.15) the differences (A.7) are the negatives of the corresponding components of the vector t_i .

The differences (A.7) were used first to determine the end-point of the iteration: when the maximum absolute value of the differences (the e-norm of t_i) fell below 2^{-23} , the iteration was stopped. The differences were also used in the application of the scalar δ^2-process.

The δ^2-process was applied after every M^{th} full iteration, and M could be varied to observe the effect upon the convergence of the method. The δ^2-process was actually arranged to coincide with the Mth iteration, so that it was possible to obtain the second difference

$$\delta^2 \eta_{j,k}^{(i)} = \delta \eta_{j,k}^{(i+1)} - \delta \eta_{j,k}^{(i)}$$

at the time it was needed. The second differences were not otherwise computed or stored.

The experiment produced the data given in the table: M = number of iterations between successive applications of δ^2-process; N = total number of iterations required to reduce $\max |\delta \eta|$ to at most 2^{-23} .

M	90	100	110	115	120	125	140	160	180	200	300	387	>387
N	193	185	180	178	178	179	179	180	181	201	301	388	388

The algorithm without the application of the δ^2-process required 388 iterations. A single application of the δ^2-process after about 100 iterations was sufficient to reduce the total number of iterations by more than 200. The optimum point at which to apply the δ^2-process was between 115 and 120 iterations.

BIBLIOGRAPHY

Aitken, A. C. (1926) On Bernoulli's numerical solution of algebraic equations, Proc. Roy. Soc. Edinburgh 46. 289-305.

Andoyer, H. (1902) "La theorie de la lune", Scientia, Paris, ser. math.-phys., p. 59.

Cauchy, A. L. (1847) Méthode générale pour la résolution des systèmes d'équations simultanées. C. R. Acad. Sci. Paris, 25, 536-38.

Collatz, L. (1951) "Numerische Behandlung von Differentialgleichungen". Berlin (Springer).

Fadeeva, V. N. (1950) "Computational methods of linear algebra" (Russian) Moscow-Leningrad.

_____. (1950a) English translation of Chap. I of (1950) available as U. S. Library of Congress RT-1085.

Forsythe, George E. (1953) Tentative classification of methods and bibliography on solving systems of linear equations. U. S. National Bureau of Standards, Applied Math. Series 29, "Simultaneous linear equations and the determination of eigenvalues", 1-28.

Householder, A. S. (1953) "Principles of Numerical Analysis", New York (McGraw-Hill).

_____. (1954a) Sequence Transformations. Mimeographed memorandum, Oak Ridge National Laboratory.

_____. (1954b) On norms of vectors and matrices. Oak Ridge Nat. Lab. Report, ORNL-1756.

_____. (1955a) Convergence of matrix iterations. Oak Ridge Nat. Lab. Report, ORNL-1883.

Jacobi, C. G. J. (1845) Über eine neue Auflösungsart der bei der Methode der Kleinsten Quadrate vorkommenden Lineären Gleichungen - Astr. Nachr. 22, no. 523, 297-306.

Kantorovič, L. V. (1948) Functional Analysis and Applied Mathematics. (Russian) Uspehi Matem. Nauk 3, no. 6: 89.

_____. (1948a) English translation of above available from Library of Congress as RT-334, p. 106.

Liebmann, H. (1918) Die angenäherte Ermittlung harmonischer Funktionen und Konformer Abbildung. Sitzungsber. math. - naturw. K. bayer. Akad. Wiss., München 1918. 385-416.

Lubkin, S. (1952) A method of summing infinite series. J. Res. Nat. Bureau of Standards 48. 228-254.

Ludwig, R. (1954a) Über iterationsverfahren für Gleichungen und Gleichungssystem. I. Zeit. Angew. Math. u. Mech. 34, 210-55.

_____. (1954b) _____. II. Zeit. Angew. Math. u. Mech. 34, 404-16.

Nekrasov, P. A. (1884) Determination of the unknowns by the method of least squares for a very large number of unknowns. Mat. Sbornik 12, 189-204 (Russian).

Reich, E. (1949) On the convergence of the classical iterative method of solving linear simultaneous equations. An. Math. Stat. 20, 448-51.

Riesz, F. (1917) Über lineare Funktionalgleichungen. Acta Math. 41, 71-98.

Sangren, W. C. (1953) Self-adjointness for diffusion problems. Oak Ridge National Laboratory memorandum CF 53-8-40.

Shanks, Daniel (1954) Non-linear transformations of divergent and slowly convergent sequences. Thesis, University of Maryland.

_____. (1955) _____. J. Math. Phys. 34, 1-42.

Stein, P. and R. L. Rosenberg (1948) On the solution of linear simultaneous equations by iteration. J. London Math. Soc. 23, 111-18.

Weissinger, J. (1953) Verallgemeinerung des Seidelschen Iterationsverfahrens. Zeit. Angew. Math. u. Mech. 33, 155-63.

Young, D. (1954a) Iterative methods for solving partial difference equations of elliptic type. - Trans. Am. Math. Soc. 76, 92-111.

TWH:ih
7/25/56