A Ph.D. Thesis of Historical Importance

Iterative Methods for Solving Partial Difference Equations of Elliptic Type

by

David M. Young, Jr.

Preface

David Young's thesis is one of the monumental works of modern numerical analysis. His creation, development and analysis of the Successive Overrelaxation (SOR) method has been fundamental in our understanding of iterative methods for solving linear equations. He cleverly abstracted the linear algebra problem from the origin of the matrix and by studying matrices with 'Property A' , he could analyze the SOR method for large classes of problems. The discussion of matrices with the red/black ordering has been of great importance in developing methods for parallel architectures. The many ideas generated in this thesis have had lasting impact.

The thesis was originally typed by a typist in Cambridge, MA who was typing three other theses at the time — all of which were past due, Young recalls. David and his wife, Mildred, wrote in all the equations and symbols in ink in three copies of the thesis. In 1975, Dorothy Baker re-typed this thesis at The University of Texas at Austin.

Barbara Morris LaTeXed the current version, reconciling the two versions of the typed thesis. Some editorial choices and changes were made throughout using modern typographical style since modern typography is a bit different from when Young wrote his thesis 50 years ago!

David Kincaid compared this type-set version to two copies of the original thesis and to Dorothy's version. David Young and Gene Golub clarified and resolved any questions. Some minor changes have been made. For instance, Lemmas 8.4 and 8.5 are taken from an early revision; but, we have tried to keep to the original as far as possible.

Gene H. Golub and David R. Kincaid

Iterative Methods for Solving Partial Difference Equations of Elliptic Type

by
David M. Young, Jr.

Thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Department of Mathematics

Harvard University
Cambridge, Mass,
May 1, 1950

Foreword

The author is greatly indebted to Professor Garrett Birkhoff for his helpful suggestions and his guidance in the present investigation. The interest and encouragement of Professor Birkhoff were stimulating and valuable.

The author gratefully acknowledges the support of the Office of Naval Research, extended through Contract N5-ori-07634, with Harvard University.

Table of Contents

Introduction

Finite difference methods afford a powerful tool for obtaining approximate numerical solutions for many differential equations whose analytic solutions are not known. The differential equation is replaced by a difference equation which must be satisfied by the values of the unknown function u at a finite set of points in the domain, Ω, of the independent variable. This set of points usually consists of the nodes, or net points, of a square network Ω_h contained in Ω. The mesh size is denoted by h (> 0).

It is shown in Chapter I that if N is the number of net points and $u = (u_1, u_2, \ldots, u_N)$ is the unknown function, corresponding to a linear self adjoint second order partial differential equation of elliptic type with prescribed boundary values, u must satisfy a system of linear equations of the form

$$\sum_{j=1}^{N} a_{i,j} u_j + d_i = 0 \qquad (i = 1, 2, ..., N) \tag{0.1}$$

where the coefficients $a_{i,j}$ are real and where

$$\begin{cases} \text{(a)} \quad a_{ii} > 0 & (i = 1, 2, \ldots, N) \\ \text{(b)} \quad a_{i,j} \leq 0 & (i \neq j;\ i, j = 1, 2, \ldots, N) \\ \text{(c)} \quad a_{ii} \geq \sum\limits_{\substack{j=1 \\ j \neq i}}^{N} |a_{i,j}| & (i = 1, 2, \ldots, N) \quad \text{and for some } i \\ \qquad\quad a_{ii} > \sum\limits_{\substack{j=1 \\ j \neq i}}^{N} |a_{i,j}| & \\ \text{(d)} \quad \text{The } N \times N \text{ matrix } (a_{i,j}) \text{ is } \textbf{irreducible},^{\dagger} \text{ that is, given any two} & \\ \quad \text{non empty complementary subsets, } \mathcal{S}_{\mathcal{E}} \text{ and } \mathcal{S}_{\prime} \text{ of the set of the first} & \\ \quad N \text{ integers there exists } a_{i,j} \neq 0 \text{ such that } i \in \mathcal{S}_E \text{ and } j \in \mathcal{S}_0. & \end{cases} \tag{0.2}$$

Definition 0.1 A difference equation is **elliptic** if, when reduced to the form (0.1), the coefficients of $(a_{i,j})$ satisfy conditions (0.2).

Definition 0.2 An elliptic difference equation is **self adjoint** if

$$\text{(e)} \quad a_{i,j} = a_{j,i} \qquad (i, j = 1, 2, \ldots, N). \tag{0.3}$$

Although the finite difference analogue of a self adjoint partial differential equation is self adjoint, if the mesh size is altered, as for example near the boundary, (e) will no longer be fulfilled. We shall avoid using (e) whenever possible, and shall always state when it is used.

Geiringer [12] has shown that any system of equations, fulfilling the conditions (0.2) has a unique solution. The proof will be given in Chapter I. Actually obtaining the solution, however, may be very laborious.

It is the purpose of this thesis to consider the practicability of the various methods for solving these equations, with special emphasis on those methods which are adapted to large automatic computing machines. I shall be particularly concerned with the Dirichlet Problem.

[†] Geiringer [12]. Numbers in brackets, [ ], refer to the bibliography at the end of the thesis.

Direct methods such as the use of determinants and elimination for solving (0.1) do not appear to be very practical when N is large, and various methods of successive approximation are usually employed, including the iterative methods of Kormes [19] and Liebmann [21] (modified by Shortley and Weller [30]) and the relaxation methods of Southwell [32].

For these methods an arbitrary initial approximation $u^{(0)} = (u_1^{(0)}, u_2^{(0)}, \ldots, u_N^{(0)})$ is chosen and successively improved. One obtains a sequence $\{u^{(m)}\}$ such that under the conditions on $(a_{i,j})$

$$\lim_{m\to\infty} u_i^{(m)} = u_i \qquad (i = 1, 2, \ldots, N) \tag{0.4}$$

(a) For the **Kormes Method**, which was applied by Kormes to a special case of (0.1), the sequence $\{u^{(m)}\}$ is defined by

$$u_i^{(m+1)} = -\sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{a_{i,j}}{a_{ii}}\, u_j^{(m)} - \frac{d_i}{a_{ii}} \qquad (i = 1, 2, \ldots, N) \tag{0.5}$$

or

$$u^{(m+1)} = \mathcal{K}\,[u^{(m)}] + c\ , \tag{0.5a}$$

where $\mathcal{K}$ is a linear operator on V_N the vector space of N-tuples of complex numbers, and where the vector c is given by

$$c = (-\frac{d_1}{a_{11}}, -\frac{d_2}{a_{22}}, \ldots, -\frac{d_N}{a_{NN}}). \tag{0.6}$$

(b) For the **Liebmann Method**, which is actually a special case of the Gauss-Seidel Method [28], the equations are taken in a prescribed order, σ, and the sequence $\{u^{(m)}\}$ is defined by

$$u_i^{(m+1)} = -\sum_{j=1}^{i-1} \frac{a_{i,j}}{a_{i,i}}\, u_j^{(m+1)} - \sum_{j=i+1}^{N} \frac{a_{i,j}}{a_{i,i}}\, u_j^{(m)} - \frac{d_i}{a_{ii}} \qquad (i = 1, 2, \ldots, N) \tag{0.7}$$

or

$$u^{(m+1)} = \mathcal{L}_\sigma[u^{(m)}] + c. \tag{0.7a}$$

Clearly, $\mathcal{L}_\sigma$ is a linear operator on V_N.

In Chapter I, a proof, due to Geiringer [12] is given for the convergence of the Kormes and Liebmann Methods, using the fact that the matrix $(a_{i,j})$ satisfies the conditions (0.2).

(c) For the **relaxation methods**, one first computes the residuals of $u^{(0)}$ by the formula

$$Z_i^{(0)} = -\left[\sum_{j=1}^{N} a_{i,j} u_j^{(0)} + d_i\right] \qquad (i = 1, 2, \ldots, N)\ . \tag{0.8}$$

If $Z_i^{(0)} = 0$ for all i, $u^{(0)}$ satisfies (0.1). Otherwise, the values of $u_i^{(0)}$ are adjusted and the residuals are recomputed. This process is continued until all residuals are reduced to a negligible amount. No exact instructions are usually given for doing this; "the former (the relaxation methods) challenges one's intellect at each step to make the best possible guess ... " [11]. On the other hand, high intellectual powers are not required; by adjusting the coordinates of $u^{(0)}$ one at a time, just enough

to remove the corresponding residual, and so that the coordinates are adjusted cyclically in a prescribed order, one has the Liebmann Method. Any one residual $Z_i^{(m)}$ can be removed as follows

$$\begin{cases} u_j^{(m+1)} = u_j^{(m)} & (j \neq i) \\ u_i^{(m+1)} = u_i^{(m)} + \dfrac{Z_i^{(m)}}{a_{ii}} \end{cases} \tag{0.9}$$

as can be verified directly. This is called **relaxing** the residual $Z_i^{(m)}$. One logical way of relaxing the residuals is to relax the residual of largest magnitude at each stage. In Chapter I, a proof due to Temple [35], is given for the convergence of this procedure when $(a_{i,j})$ satisfies (0.2) and (0.3). If only (0.2) is assumed, the method is shown to converge under a weak assumption on the order of relaxing.

Various devices have been introduced to accelerate the convergence, such as **block relaxation** ([32] page 55) where a group of values $u_i^{(m)}$ are modified simultaneously, and **overrelaxation** and **underrelaxation** ([32] page 65), where one modifies $u_i^{(m)}$ by more or less, respectively, than indicated by (0.9). These will be discussed later in this section.

Emmons [11] estimates that by proper use of the relaxation methods, the labor required by the Liebmann Method can be reduced by a factor of 5. However, the effective use of relaxation methods requires scanning of the residuals, a process which is easy for a human computer but which cannot be done efficiently by any large automatic computing machine in existence or being built.

Iterative methods, on the other hand, appear to be best suited for large automatic computing machines. In Chapter II, the rates of convergence of the Kormes and Liebmann Methods are studied in detail. The rates of convergence depend on the eigenvalues of the operators $\mathcal{K}$ and $\mathcal{L}_\sigma$ respectively. For the study of these eigenvalues, we make use of the fact that the matrix $(a_{i,j})$ has property (A_q) for some q. This property is now defined.

Definition 0.3 An $N \times N$ matrix $(a_{i,j})$ has **property** $(\mathbf{A_q})$ if there exist non empty disjoint subsets $T_1, T_2, \ldots, T_q$ of $\mathcal{T}$, the set of the first N integers, such that $\bigcup_{\ell=1}^{q} T_\ell = \mathcal{T}$, and such that the T_l can be labeled so that

$$a_{i,j} = 0 \qquad \text{unless } i = j \qquad \text{or} \qquad i \in T_\ell \quad \text{and} \quad j \in T_{\ell-1} \bigcup T_{\ell+1}. \tag{0.10}$$

(By convention T_0 and T_{q+1} denote the empty set.)

Since $(a_{i,j})$ has property (A_q) for some q, it is easy to show that if μ is an eigenvalue of $\mathcal{K}$, $(-\mu)$ is also an eigenvalue. There exist certain orderings, σ, of the equations, called **consistent orderings**, such that to each such pair $(\mu, -\mu)$ of eigenvalues of $\mathcal{K}$ corresponds an eigenvalue $\lambda = \mu^2$ of $\mathcal{L}_\sigma$. The rate of convergence of $\mathcal{L}_\sigma$ is exactly twice that of $\mathcal{K}$. This has been shown to be true *asymptotically* in N by Shortley and Weller [30]. An explicit expression of the eigenvectors of $\mathcal{L}_\sigma$ in terms of the eigenvectors of $\mathcal{K}$ is also given. If $a_{i,j} = a_{j,i}$ all eigenvalues of $\mathcal{K}$ and $\mathcal{L}_\sigma$ are real and the Jordan normal form of the corresponding matrices are diagonal with the possible exception of the subspace associated with $\lambda = 0$ for $\mathcal{L}_\sigma$.

If, also, $a_{ii} =$ constant, as for the Dirichlet Problem, the eigenvectors of $\mathcal{K}$ are orthogonal. For the Dirichlet Problem the eigenvalues and eigenvectors of $\mathcal{K}$ and $\mathcal{L}_\sigma$ can be computed exactly for a rectangular region. For one ordering, σ_2, the normal form of the matrix of $\mathcal{L}_{\sigma_2}$ is diagonal and if the coordinates of an arbitrary vector in V_N referred to the basis of eigenvectors of $\mathcal{K}$ are known, the coordinates of that same vector referred to the basis of eigenvectors of $\mathcal{L}_{\sigma_2}$ can be computed at once.

For the symmetric case, a conjecture that by the Liebmann Method, the rate of convergence can not be increased by using an ordering which is not consistent is proved in one special case. Some numerical studies bear out a conjecture by Shortley and Weller [30] that for large N (hence small h) the rate of convergence of $\mathcal{L}_\sigma$ is practically independent of σ.

It is also shown that for those methods the number of iterations required to reduce the norm of the initial error function

$$\|e^{(0)}\| = \left[\sum_{i=1}^{N} {e_i^{(0)}}^2\right]^{1/2} \tag{0.11}$$

where

$$e_i^{(0)} = u_i^{(0)} - u_i \qquad (i = 1, 2, \ldots, N) \tag{0.12}$$

to a definite fraction of itself is asymptotically for small h proportional to h^{-2}. For some problems, the time required to obtain an acceptable degree of accuracy, even with a fast large automatic computing machine such as the UNIVAC, is prohibitive.

In Chapter III, it is shown that the required number of iterations can be greatly reduced by using the Successive Overrelaxation Method, where the idea of *systematic overrelaxation*, first used by L. F. Richardson, [26], is combined with the Liebmann Method. The idea of overrelaxation itself has been used in connection with the relaxation methods by such authors as Hartree, [14], page 120, Southwell, [32], page 65, and Emmons [11]. Overrelaxation is an attempt to "anticipate at each step the effect of later steps in the process," [14]. By Richardson's Method, the values of an approximate solution, $u^{(m)}$ of (0.1) are modified *simultaneously* by the formula

$$u_i^{(m+1)} = u_i^{(m)} - \omega\left\{\sum_{j=1}^{N} a_{i,j}u_j^{(m)} + d_i\right\} \qquad (i = 1, 2, \ldots, N) \tag{0.13}$$

or

$$u^{(m+1)} = \mathcal{R}_\omega[u^{(m)}] + \omega d \tag{0.13a}$$

where ω is the **relaxation factor**. If a_{ii} is constant for $i = 1, 2, \ldots, N$ and $\omega = 1/a_{ii}$ we have the Kormes improvement formula (0.5). In general, for the Kormes Method combined with overrelaxation we have the improvement formula

$$u_i^{(m+1)} = \omega\left[\sum_{\substack{j=1\\ j\neq i}}^{N} -\frac{a_{i,j}}{a_{i,i}}\, u_j^{(m)} - \frac{d_i}{a_{i,i}}\right] - (\omega - 1)u_i^{(m)} \quad (i = 1, 2, \ldots, N), \tag{0.14}$$

or

$$u^{(m+1)} = \mathcal{K}_\omega[u^{(m)}] + \omega c \tag{0.14a}$$

where c is given in (0.6). Again if $a_{i,i}$ is constant for $i = 1, 2, \ldots, N$ (0.14) is equivalent to (0.13) for suitable choice of ω.

The gain in convergence rate using a *fixed* relaxation factor is in general very slight indeed for either (0.13) or (0.14). Richardson used different values of ω for each iteration, but it appears doubtful that a gain of a factor of greater than 5 in the rate of convergence can in general be realized unless one is extremely fortunate in the choice of the values of ω. Shortley and Weller [30], are of the opinion that the gain is even less. This is discussed further in Chapter III.

However, by the simple device of improving the values $u_i^{(m)}$ *successively* in a cyclic order σ and using *new values* as soon as they are available, one can use a *fixed* relaxation factor, and, if this single factor is suitably chosen, a large gain in the rate of convergence is possible. For the Dirichlet

Problem, the gain is of the order of h^{-1} and for the general self adjoint case, if the required number of iterations with the Liebmann Method is of the order of h^{-k} that number is of the order of $h^{-k/2}$ if this new method, the **Successive Overrelaxation Method**, is used with the proper value of ω.

The improvement formula for the Successive Overrelaxation Method is

$$(0.15)\quad u_i^{(m+1)} = \omega\Big[-\sum_{j=1}^{i-1}\frac{a_{i,j}}{a_{i,i}}\,u_j^{(m+1)} - \sum_{j=i+1}^{N}\frac{a_{i,j}}{a_{i,i}}\,u_j^{(m)} - \frac{d_i}{a_{i,i}}\Big] - (\omega-1)u_i^{(m)} \quad (i=1,2,\ldots,N)$$

or

$$(0.15a)\qquad u^{(m+1)} = \mathcal{L}_{\sigma,\omega}[u^{(m)}] + \omega c$$

where the subscripts σ and ω of $\mathcal{L}_{\sigma,\omega}$ denote the *ordering* and the *relaxation factor* respectively. Clearly $\mathcal{L}_{\sigma,\omega}$ is a linear operator, and if $\omega = 1$ we have the Liebmann Method. The Successive Overrelaxation Method is included in a more general class of iterative methods considered by H. Geiringer [12].

It is by using the fact that the matrix $(a_{i,j})$ has property (A_q) that it can be shown, for the first time, that one can obtain the above mentioned remarkable gain in the rate of convergence. We show that, always assuming σ consistent, there is an exact algebraic relation between the eigenvalues and eigenvectors of $\mathcal{K}$ and $\mathcal{L}_{\sigma,\omega}$.

If μ is any eigenvalue of $\mathcal{K}$ there exists an eigenvalue $\widehat{\lambda}$ of $\mathcal{L}_{\sigma,\omega}$ such that

$$(0.16)\qquad \mu\omega\widehat{\lambda}^{1/2} = \widehat{\lambda} + (\omega - 1)$$

and conversely every eigenvalue of $\mathcal{L}_{\sigma,\omega}$ can be determined by (0.16) for some μ. The eigenvectors of $\mathcal{L}_{\sigma,\omega}$ can also be expressed explicitly in terms of the eigenvectors of $\mathcal{K}$. If $(a_{i,j})$ is symmetric then the optimum relaxation factor ω_b is given by

$$(0.17)\qquad \mu_1^2\omega_b^2 = 4(\omega_b - 1)$$

where μ_1 is the largest eigenvalue of $\mathcal{K}$. For all $\omega \geq \omega_b$ every eigenvalue of $\mathcal{L}_{\sigma,\omega}$ has absolute value $(\omega - 1)$ and the Jordan normal form of the matrix of $\mathcal{L}_{\sigma,\omega}$ is diagonal unless $\omega = \omega_b$. In this case, the normal matrix form contains precisely one non diagonal element.

For the Dirichlet Problem for small h, μ_1 is very nearly equal to one. μ_1 can be calculated for a rectangle and can be estimated for other regions by comparison theorems. For the general self adjoint case, provided μ_1 is not underestimated (for the Dirichlet Problem a non trivial upper bound for μ_1 can always be found) the relative decrease in the rate of convergence if $\omega' \geq \omega_b$ is used is approximately

$$(0.18)\qquad \zeta^{-1/2} - 1$$

where

$$(0.19)\qquad (1 - \mu_1') = \zeta(1 - \mu_1) \qquad (1 < \zeta \leq 1)$$

and where ω' is determined from (0.17) but with μ_1 replaced by the estimated value μ_1'. Thus a relatively large error in the estimation of $(1 - \mu_1)$ can be allowed, and the improvement over the Liebmann Method will not suffer appreciably. This also suggests that the Successive Overrelaxation Method can be successfully applied to self adjoint equations other than Laplace's Equation.

The Successive Overrelaxation Method can be used with any large automatic computing machine for which the Liebmann Method can be used. The machine time per iteration would not be increased by more than 10%. It is expected that the use of the Successive Overrelaxation Method will considerably augment the class of problems for which the use of large automatic computing machines is practical.

Chapter I

Partial Difference Equations of Elliptic Type

Let Ω be a closed bounded region in Euclidean n-space, with interior R and boundary S. Consider the self adjoint elliptic partial differential equation of the second order

$$(I.0.1) \qquad (L[\bar{u}] + G)(x) = 0 \qquad x \in R$$

$$\bar{u}(x) = g(x) \qquad x \in S$$

where

$$(I.0.2) \qquad L[\bar{u}] = \sum_{k=1}^{n} \frac{\partial}{\partial x_k} \left(\alpha^{(k)} \frac{\partial \bar{u}}{\partial x_k} \right) + F\ \bar{u}$$

and $\bar{u}$, $\alpha^{(k)}$ $(k = 1, 2, \ldots, n)$, F, and g are functions of the vector x whose components, (referred to an orthogonal basis $e_1, e_2, \ldots, e_n$ of unit coordinate vectors), are $(x_1, x_2, \ldots, x_n)$. We assume F, $\alpha^{(k)}$ and G are continuous functions of x with continuous first and second partial derivatives in Ω, and that g is a continuous function of x on S. $\alpha^{(k)}$ and F satisfy the conditions

$$(I.0.3a) \qquad \alpha^{(k)} > 0 \qquad (k = 1, 2, \ldots, n)$$

$$(I.0.3b) \qquad F \leq 0\ .$$

If f is any function defined on Ω we write

$$(I.0.4) \qquad f(x) = f(x_1, x_2, \ldots, x_n)\ .$$

An important special case of (I.0.2) is the **Dirichlet Problem**, Kellogg [18] page 236. In this case

$$(I.0.5) \qquad L[\bar{u}] = \nabla^2[\bar{u}] = \sum_{k=1}^{n} \frac{\partial^2 \bar{u}}{\partial {x_k}^2}\ .$$

§1. Basic Facts

For the Dirichlet Problem, the existence of a unique solution has been proved for a wide class of regions, see for example Lichtenstein [20] and Kellogg [18], Chapter XI. In [18] page 329 Zaremba's criterion is given for $n = 3$:

> a unique solution exists if each point P of S is the vertex of a right circular cone, which has no points in the portion of R in any sphere about P however small.

Phillips and Wiener [25] gave a proof, based on the use of finite difference methods, for the existence of a unique solution of the Dirichlet Problem for any region Ω with the following property:

> Whenever a point P belongs to S there are positive numbers a and b such that if $r < a$ and an n-sphere Γ of radius r is drawn with P as center, and if Σ is the set of points in Γ but not in R, then the projection of Σ on at least one of the $(n-1)$ dimensional spaces determined by a set of n perpendicular axes exceeds br^{n-1} in content.

Courant, Friedrichs and Lewy, [8] have given a proof for the more general problem (I.0.1) based on finite difference methods. The existence of a unique solution of (I.0.1) is proved for regular regions ([18] page 113), such that

$$\lim_{r\to 0} \frac{1}{r} \int_{S_r} u^2 \, dx_1 \, dx_2 \, \ldots \, dx_n = 0 \tag{1.1}$$

implies $u = 0$ on S where S_r is the set of all points of Ω at a distance less than r from some point of S and where u is any continuous function. This condition is always fulfilled for $n = 2$, [8].

The analytic solution is not known, however, except for the simplest regions. Consequently if one desires numerical results, approximate methods must ordinarily be used. Included among these methods are variational methods [16], the use of orthogonal functions [1] (for $n = 2$), conformal mapping (for $n = 2$ and for the Dirichlet Problem), and finite difference methods [16].

In this dissertation, I shall be concerned with finite difference methods. The region Ω is covered by a network Ω_h and the differential equation is replaced by a difference equation involving the values of the unknown function at the nodes (net points) of Ω_h. The solution of the difference equation differs from the solution of the differential equation by an amount which depends on the fineness of the mesh and on the type of finite difference approximation which is used. Consider a square network* whose nodes are the set

$$x = (x_1, x_2, \ldots, x_n)$$

such that

$$x_k = p_k h \qquad (k = 1, 2, \ldots, n)$$

where h is the mesh size and the p_k are integers.

Two net points x and x' are *adjacent* if $p_k = p'_k$ for all k except for a single i

$$|p_i - p'_i| = 1$$

where

$$x = (p_1 h, p_2 h, \ldots, p_n h)$$
$$x' = (p'_1 h, p'_2 h, \ldots, p'_n h)$$

If x is adjacent to x' we write $x \wedge x'$.

*Triangular and hexagonal networks have been used by Southwell and his collaborators [32]. Square networks are simpler, and have been used almost exclusively in the United States.

Ω_h is the set of all such net points contained in Ω. R_h, the *interior* of Ω_h is the set of all points x such that all net points adjacent to x belong to Ω_h. All other points of Ω_h belong to S_h the *boundary* of Ω_h. A *segment* is a straight line joining two adjacent net points. Ω_h is *connected* provided any two points of R_h can be joined by an unbroken line consisting of segments contained in R_h. We assume that h is sufficiently small so that Ω_h is connected.

Let N be the number of points of R_h and N_B be the number of points of S_h. The general point of Ω_h may be denoted by

$$x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}), \qquad (i = 1, 2, \ldots, N + N_B) \tag{1.2}$$

where

$$x_k^{(i)} = p_k^{(i)} h\,.$$

We assume that if $i \leq N$, $x^{(i)} \in R_h$. Let V_{N+N_B} be the space of all complex valued functions defined on Ω_h and let V_N be the space of all functions on V_{N+N_B} vanishing on S_h.

A function (vector), $f \in V_{N+N_B}$ has coordinates

$$f = (f_1, f_2, \ldots, f_{N+N_B})$$

referred to a basis of unit functions which equal one at one point of Ω_h and vanish elsewhere, where

$$f_i = f(x^{(i)}) \qquad (i = 1, 2, \ldots, N + N_B)\,. \tag{1.3}$$

We define the operators ([33] page 4)

$$\begin{cases} E_{x_k} f(x) &= f(x + he_k) \\ \\ E_{-x_k} f(x) &= f(x - he_k) \end{cases} \qquad (k = 1, 2, \ldots, n) \tag{1.4}$$

To derive the finite difference analogue of (I.0.1), we replace the derivatives by difference quotients. Thus as in [8], the differential operator

$$\frac{\partial}{\partial x_k}\left(\alpha^{(k)}\ \frac{\partial}{\partial x_k}\right)$$

is replaced by

$$h^{-2}\ \left\{(1 - E_{-x_k})[\alpha^{(k)}\ (E_{x_k} - 1)]\right\} \tag{1.5}$$

and the operator (L) is replaced by

$$-\ell h^{-2} = h^{-2}\ \left\{\sum_{k=1}^{n} \alpha^{(k)} E_{x_k} - \left[\alpha^{(k)} + E_{-x_k}\alpha^{(k)} + \left(E_{-x_k}\alpha^{(k)}\right) E_{-x_k}\right] + h^2 F\right\} \tag{1.6}$$

The finite difference analogue of (I.0.1) becomes

$$\text{(1.7)}\qquad \begin{cases} (\ell[u])_i - h^2 G_i = 0 & (i = 1, 2, \ldots, N) \\ u_i = g_i & (i = N+1, \ldots, N+N_B) \end{cases}$$

or

$$\text{(1.8)}\qquad \begin{cases} \sum_{j=1}^{N+N_B} a_{i,j} u_j - h^2 G_i = 0 & (i = 1, 2, \ldots, N) \\ u_i = g_i & (i = N+1, \ldots, N+N_B)^{\dagger} \end{cases}$$

where the general element of the $N \times (N + N_B)$ matrix $(a_{i,j})$ is

$$\text{(1.9)}\qquad a_{i,j} = \begin{cases} \sum_{k=1}^{n} [\alpha_i^{(k)} + E_{-x_k}\alpha_i^{(k)} - h^2 F_i] & (i = j) \\ -\alpha_i^{(k)} & (x^{(j)} = x^{(i)} + he_k \text{ for some } k) \\ -E_{-x_k}\alpha_i^{(k)} & (x^{(j)} = x^{(i)} - he_k \text{ for some } k) \\ 0 & (i \neq j \text{ and } x^{(i)} \text{ not adjacent to } x^{(j)}) \end{cases}$$

For the Dirichlet problem,[‡] we have

$$\text{(1.10)}\qquad -(\ell[u])_i = \left(\left[\sum_{k=1}^{n} (E_{x_k} + E_{-x_k}) - 2n \right] [u] \right)_i = 0 \qquad (i = 1, 2, \ldots, N)$$

and

$$\text{(1.10a)}\qquad a_{i,j} = \begin{cases} 2n & (i = j) \\ -1 & (x^{(i)} \wedge x^{(j)}) \\ 0 & (i \neq j \text{ and } x^{(i)} \text{ not adjacent to } x^{(j)}). \end{cases}$$

Since $\mu_i = g_i \quad (i = N+1, \ldots, N+N_B)$, we have

$$\text{(1.11)}\qquad \sum_{j=1}^{N} a_{i,j}\, u_j + d_i = 0 \qquad (i = 1, 2, \ldots, N)$$

where

$$\text{(1.12)}\qquad d_i = \sum_{j=N+1}^{N+N_B} a_{i,j}\, g_j - h^2 G_i \qquad (i = 1, 2, \ldots, N).$$

We may also at times write (1.11) in the form

$$\text{(1.13)}\qquad u_i = \sum_{j=1}^{N} b_{i,j}\, u_j + c_i \qquad (i = 1, 2, \ldots, N)$$

[†]Here g_i is taken as the value $g(x)$ where $x \in S$ and x is the nearest point of S to $x^{(i)} \in S_h$. Alternatively, one might extend the function g to R and take g_i to be the value of this extended function. Another possibility is to change the mesh size near R whenever $x^{(i)} \in S_h$ does not belong to S, [30]. This does not make the problem essentially more difficult from a practical point of view but does spoil some of the mathematical relationships to be given later. Further discussion of this question will be given later.

[‡]Other finite difference analogues of the Dirichlet Problem, with a square net, have been considered by Milne [23] and Bickley [2].

where

$$b_{i,j} = \begin{cases} -\dfrac{a_{i,j}}{a_{i,i}} & (i \neq j) \\ 0 & (i = j) \end{cases} \tag{1.14}$$

$$c_i = -\frac{d_i}{a_{i,i}} \qquad (i = 1, 2, \ldots, N)\ . \tag{1.15}$$

Clearly by (1.8) and (I.0.3a,b), the coefficients of $(a_{i,j})$ have the following properties

$$\begin{cases} \text{(a)} & a_{i,i} > 0 & (i = 1, 2, \ldots, N) \\ \text{(b)} & a_{i,j} \leq 0 & (i, j = 1, 2, \ldots, N) \\ \text{(c)} & a_{i,i} \geq \sum\limits_{\substack{j=1 \\ j \neq i}}^{N} |a_{i,j}| & (i = 1, 2, \ldots, N), \text{ and for some } i \ ^{\dagger} \\ & a_{i,i} > \sum\limits_{\substack{j=1 \\ j \neq i}}^{N} |a_{i,j}| & \\ \text{(d)} & a_{i,j} < 0 & (x^{(i)} \wedge x^{(j)}) \\ \text{(e)} & a_{i,j} = 0 & (i \neq j \text{ and } x^{(i)} \text{ not adjacent to } x^{(j)}) \end{cases} \tag{1.16}$$

$$\text{(f)} \quad a_{i,j} = a_{j,i} \qquad (i, j = 1, 2, \ldots, N)^{\ddagger}\ . \tag{1.17}$$

Theorem 1.1 The matrix $(a_{i,j})$ satisfying (1.16d), (1.16e) is irreducible and has property (A_2).

Proof: (a) By (1.16d), given any two non empty complementary subsets of R_h there must exist two adjacent points not belonging to the same subset since Ω_h is connected. The irreducibility follows

(b) Let T_2 be the set of integers such that $\sum\limits_{k=1}^{N} p_k^{(i)}$ is even, and let T_1 be the set of integers such that $\sum\limits_{k=1}^{N} p_k^{(i)}$ is odd.

Then T_1 and T_2 are complementary subsets of $\mathcal{T}$ and by (1.16e), if $i \neq j$

$$a_{i,j} = 0$$

unless $i \in T_1$ and $j \in T_2$ or $i \in T_2$ and $j \in T_1$.
Thus $(a_{i,j})$ has property (A_2).

Q. E. D.

Corollary 1.1 (1.7) is an elliptic self adjoint difference equation.

The existence of a unique solution of (1.11) has been proved by Gerschgorin [13], without requiring that $(a_{i,j})$ be symmetric but using (1.16d). We shall give a proof for a somewhat more general case, which is due to Geiringer [12].

---

†i.e. for all i such that $x(i)$ is adjacent to a point of S_h.

‡As already stated we avoid using this fact wherever possible. For if the mesh size is varied (see (1.8) footnote), this condition is no longer fulfilled.

Theorem 1.2 If the matrix $(a_{i,j})$ satisfies (0.2), then the determinant of $(a_{i,j})$ does not vanish, and there exists a unique solution of (0.1).

Proof: To prove the theorem we need only show that $d_i = 0, \quad (i = 1, 2, \ldots, N)$ implies $u_i = 0, \quad (i = 1, 2, \ldots, N)$ is the only solution of (0.1). By linear equation theory, this will imply that the determinant of $(a_{i,j})$ is not zero, and that there exists a unique solution of (0.1).

Thus, let $d_i = 0, \quad (i = 1, 2, \ldots, N)$ and suppose $u_i > 0$ for some i. Then u must assume a maximum value $u_{i_0} = M$ for some i_0. By (0.2) this is not possible unless $u_j = M$ for $j \in \mathcal{S}_1$ where $\mathcal{S}_1$ is the set of all integers such that $a_{i_0,j} \neq 0$. Similarly, by induction $u_j = M$ for all $j \in \mathcal{S}_t$, where $\mathcal{S}_t$ is the set of all integers such that $a_{i,j} \neq 0 \quad i \in \mathcal{S}_{t-1}$. If we let $\mathcal{S}_0 = \{i_0\}$ we have $\mathcal{S}_0 \subset \mathcal{S}_1 \subset \mathcal{S}_2 \ldots$ where the inclusion relation is strict by the irreducibility of $(a_{i,j})$. Therefore we have $u_i = M, \quad (i = 1, 2, \ldots, N)$ and by (0.2) this is impossible. Therefore $u_i \leq 0, \quad (i = 1, 2, \ldots, N)$. Similarly $u_i \geq 0, \quad (i = 1, 2, \ldots, N)$; therefore $u_i = 0, \quad (i = 1, 2, \ldots, N)$.

Q. E. D.

Phillips and Wiener [25] proved that, for the Dirichlet Problem, under the assumptions on Ω stated above $u \to \bar{u}$ as $h \to 0$. Courant, Friedrichs and Lewy [8] proved the same result for the general problem under their assumptions on Ω already stated.

Gerschgorin [13] derived an error bound for $|u - \bar{u}|$ under the assumption that $\bar{u}$ has continuous partial derivatives of all orders up to and including the fourth. He also obtained an error bound for $|u' - u|$ where u' is any approximate solution of (1.11). For the Dirichlet Problem the proof is relatively simple and is given below.

Theorem 1.3 If

$$(\nabla^2[\bar{u}])(x) = \sum_{k=1}^{n} \frac{\partial^2 \bar{u}(x)}{\partial x_k^2} = 0 \qquad x \in R$$

and if

$$M_4 = \max_{k=1,2,\ldots,n} \left\{ \max_{x \in \Omega} \left| \frac{\partial^4 \bar{u}(x)}{\partial x_k^4} \right| \right\}$$

then

$$|(\ell[\bar{u}])_i| \leq \frac{h^4}{6} M_4 \qquad (i = 1, 2, ..., N). \tag{1.18}$$

Proof: By (1.10), we have

$$-(\ell[\bar{u}])_i = \left(\sum_{k=1}^{n} E_{x_k} + E_{-x_k} - 2n \right) \bar{u}(x^{(i)}).$$

By Taylor's Theorem,

$$E_{x_k}(\bar{u}_i) = \bar{u}(x^{(i)} + he_k) = \bar{u}_i + h \left(\frac{\partial \bar{u}}{\partial x_k} \right)_i + \frac{h^2}{2!} \left(\frac{\partial^2 \bar{u}}{\partial x_k^2} \right)_i + \frac{h^3}{3!} \left(\frac{\partial^3 \bar{u}}{\partial x_k^3} \right)_i + \frac{h^4}{4!} \left(\frac{\partial^4 \bar{u}}{\partial x_k^4} \right) (\xi_k^+)$$

where ξ_k^+ is a point on the segment joining $x^{(i)}$ and $x^{(i)} + he_k$. We have

$$-(\ell[\bar{u}])_i = \frac{h^2}{2!} (\nabla^2 \bar{u})_i + \frac{h^4}{4!} \left\{ \sum_{k=1}^{n} \left[\frac{\partial^4 \bar{u}}{\partial x_k^4} (\xi_k^+) + \frac{\partial^4 \bar{u}}{\partial x_k^4} (\xi_k^-) \right] \right.$$

Hence,

$$|(\ell[\bar{u}])_i| \leq \frac{h^4}{6} M_4.$$

Q. E. D.

Theorem 1.4 For the Dirichlet Problem, if r is the radius of any n-sphere containing Ω and if

$$M_1 = \max_{k=1,2,\ldots,n} \left\{ \max_{x\in\Omega} \left| \frac{\partial \bar{u}(x)}{\partial x_k} \right| \right\}$$

then[§]

$$|u_i - \bar{u}_i| = \frac{M_4}{24} h^2 r^2 + \sqrt{n}\ M_1 h \qquad (i = 1, 2, \ldots, N). \tag{1.19}$$

Proof:
Lemma 1.1 If $(\ell[u])_i \geq 0, \quad (i = 1, 2, \ldots, N)$ and if $u_i \geq 0, \quad (i = N+1, \ldots, N+N_B)$, then $u_i \geq 0 \quad (i = 1, 2, \ldots, N)$.

Lemma 1.2 If $|(\ell[u])_i| \leq (\ell[v])_i, \quad (i = 1, 2, \ldots, N)$, and if $|u_i| \leq v_i, \quad (i = N+1, \ldots, N+N_B)$, then $|u_i| \leq v_i, \quad (i = 1, 2, \ldots, N)$.

Lemma 1.3 Let $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}) \in \mathcal{R}$ and let $\Omega \subseteq C_r$ where C_r is the closed n-sphere

$$\sum_{k=1}^{n} \left(x_k - x_k^{(0)} \right)^2 \leq r^2 \ .$$

Let A be a non negative real constant and let

$$\phi(x) = A \left[1 - \frac{\sum_{k=1}^{n} \left(x_k - x_k^{(0)} \right)^2}{r^2} \right] + \sqrt{n}\ M_1 h$$

then

$$(\ell[\phi])_i = \frac{4h^2}{r^2} \ .$$

We have by the mean value theorem

$$|u_i - \bar{u}_i| \leq \sqrt{n} M_1 h \qquad (i = N+1, \ldots, N+N_B)$$

since for $x^{(i)} \in S_h$, u_i is taken as the value of $\bar{u}(x)$ on the nearest point of S.

Also, by Theorem 1.4 and Lemma 1.3,

$$A = \frac{h^2}{24}\ r^2 M_4$$

$$|(\ell[\bar{u}])_i| = |(\ell[\bar{u} - u])_i| \leq (\ell[\phi])_i \qquad (i = 1, 2, \ldots, N)$$

and

$$|u_i - \bar{u}_i| \leq \phi_i \qquad (i = N+1, \ldots, N+N_B) \ .$$

The theorem follows by Lemma 1.2. Q. E. D.

Corollary 1.2 For the Dirichlet Problem, if u' is an approximate solution of (1.11) and if

$$\max_{i=1,2,\ldots,N} \{|Z_i'|\} = Z_0'$$

[§]Collatz [6] has shown that by a modification of the difference equations near the boundary, the second term can be made to approach zero with h^2.

where

$$Z_i' = -(\ell[u'])_i$$

and u is the solution of (1.11), then

$$|u_i' - u_i| \le \frac{r^2}{4h^2}\, Z_0' \qquad (i = 1, 2, \ldots, N). \tag{1.20}$$

Proof: We first note that $|u_i' - u_i| = 0 \qquad (i = N+1, \ldots, N+N_B)$.
Also, if we let

$$\phi'(x) = A\left[1 - \frac{\sum_{k=1}^{n} (x_k - x_k^{(0)})^2}{r^2}\right],$$

then

$$(\ell[\phi'])_i = \frac{4Ah^2}{r^2}\ .$$

If

$$A = \frac{r^2 Z_0'}{4h^2}\ \text{, then } (\ell[\phi'])_i \ge |(\ell[u'])_i| = |(\ell[u' - u])_i|\ .$$

Also

$$\phi_i' \ge |u_i' - u_i| = 0 \qquad (i = N+1, \ldots, N+N_B)$$

therefore the result follows by Lemma 1.3.

Q. E. D.

A disadvantage of (1.19) is that the derivatives of $\bar{u}$ are not known, in general. However, we may estimate the derivatives by difference quotients of u.

Thus for $n = 2$, we replace

$$\frac{\partial^4 \bar{u}}{\partial x_1^4} = \frac{\partial^4 \bar{u}}{\partial x_2^4} = \frac{\partial^4 \bar{u}}{\partial x_1^2 \partial x_2^2} \qquad \text{by}$$

$$D_4(x) = h^{-4}\left[(E_{x_1} + E_{-x_1})(E_{x_2} + E_{-x_2}) - 4\right] u(x).$$

If $S_h \subseteq S$, we have by (1.19)

$$|u_i - \bar{u}_i| \lessapprox \frac{h^2 r^2}{24}\, \frac{\overline{D}_4}{h^4} = \frac{r^2 \overline{D}_4}{24}\ \P$$

where

$$\overline{D}_4 = \max_{x \in R_h} D_4(x).$$

For the unit square with I intervals on a side, we have

$$h = I^{-1} \qquad \text{and} \qquad r = \frac{1}{2}\sqrt{2}\ .$$

then

$$\frac{r^2}{24}\, \overline{D}_4 h^{-2} = \frac{I^2}{2}\, \frac{\overline{D}_4}{24} = .0833\ I^2 \left(\frac{\overline{D}_4}{4}\right).$$

Shortley and Weller [30] give

$$.06\ I^2 \left(\frac{\overline{D}_4}{4}\right)$$

as an asymptotic estimate for the error.

¶Here the symbol $\lessapprox$ means "approximately less than or equal to."

§2. Numerical Methods

The existence and uniqueness of a solution of (1.11) has already been proved. Thus, from the standpoint of pure mathematics, the problem might be considered solved. From the practical, numerical point of view, however, actually solving the equations is the hardest part of the problem.

a. Direct Methods

The formal solution of (1.11) in terms of determinants is of little practical value except for very small N. Direct solution by elimination methods might be practical for large automatic computing machines but the number of operations required increases very rapidly with N.

Runge [27] proposed a method by which one can reduce the number of linear equations by a factor which is of the order of h^{-1}. The simplicity of the original equations is sacrificed. To apply the method one first expresses the u_i , $(x^{(i)} \in R_h)$, in terms of the u_i , $(x^{(i)} \in (\zeta_k^+) \cup S_h)$ where (ζ_k^+) is the set of points $x^{(i)}$ of R_h such that

$$(x^{(i)} - he_h) \in S_h \ .$$

For example, consider the Dirichlet Problem for a plane region (Figure 1.1).

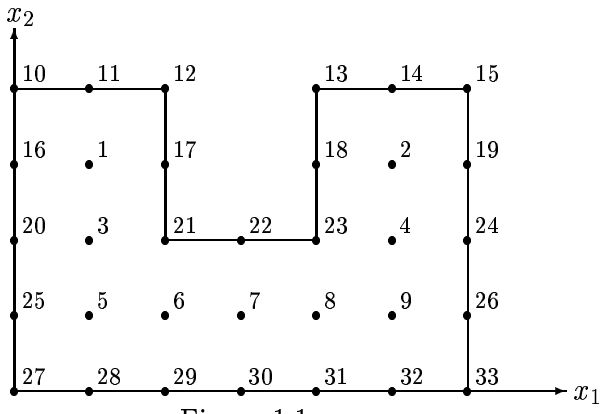

Figure 1.1

Here (ζ_k^+), $(k = 1)$, consists of the points 1, 3, 5, 2, 4 and (ζ_k^-) consists of the points 1, 3, 2, 4, 9 where, in general, (ζ_k^-) is the set of points $x^{(i)}$ of R_h such that

$$(x^{(i)} + he_k) \in S_h \ .$$

The u_i, $(x^{(i)} \in R_h)$ can be expressed in terms of u_i, $(x^{(i)} \in S_h \cup (\zeta_k^+))$ by means of the difference equation. Thus, in Figure 1.1, we have

$$u_6 + u_3 + u_{27} + u_{28} - 4u_5 = 0.$$

Solving for u_6 we have

$$u_6 = 4u_5 - u_3 - u_{27} - u_{28}.$$

If $\mathcal{N}_k$ is the number of points of (ζ_k^+), $\mathcal{N}_k$ is also the number of points of (ζ_k^-). The condition that the difference equation must be satisfied at the points of (ζ_k^-) gives $\mathcal{N}_k$ equations and $\mathcal{N}_k$ unknowns. Solving these we have u_i, $(x^{(i)} \in (\zeta_k^+))$, and hence all u_i, $(x^{(i)} \in R_h)$.

The method does not appear well adapted to large automatic computing machines. The complicated equations are not easy to solve either directly or by iterative methods. Moreover, the

determination of the coefficients is laborious, especially for large N. In practice, one would obtain the non-homogeneous part of the equations by assuming $u_i = 0$, $(x^{(i)} \in (\zeta_k^+))$, and determining u_i, $(x^{(i)} \in (\zeta_k^-))$, from the boundary values alone. Then, letting the boundary values equal zero, the coefficients of the homogeneous part of the equations are determined as follows:

> Assume $u_i = 1$ at one point of (ζ_k^+) and zero at all other points of (ζ_k^+). Then calculate the values of u at the points of (ζ_k^-).

High accuracy in all calculations is very necessary because of the instability of the method. Thus, in the example (Figure 1.1), an error of ϵ in the final determination of u_5 would result in the following errors:

$$u_6 = 4\epsilon, \quad u_7 = 15\epsilon, \quad u_8 = 56\epsilon, \quad u_9 = 209\epsilon,$$

If I_k is the average number of intervals in the x_k direction, the accumulated error would be about $\epsilon\ 4^{I_k}$.

Thomas [36] suggests a method for direct solution which he asserts is at least as good as the ordinary methods of iteration for $n = 2$. However, the method does not appear to be suitable for large automatic computing machines and is certainly not as good as The Successive Overrelaxation Method of Iteration (Chapter III).

For the Dirichlet Problem and for a rectangle with commensurable sides, an explicit solution in terms of the boundary values is possible. Thus let Ω be a rectangular region with sides $\tau_1, \tau_2, \ldots, \tau_n$ where $\tau_k = I_k h \quad (k = 1, 2, \ldots, n)$ and where $I_k \quad (k = 1, 2, \ldots, n)$ are integers.

We assume that on $S, u = 0$ except on $x_n = \tau_n$ where $u = g(x_1, x_2, \ldots, x_{n-1})$. Obviously the general solution can be obtained by linear superposition of such solutions. Using the method of separation of variables for (1.10), as for the continuous case (see Jackson [17] page 95) and noting that

$$\sum_{p_k=1}^{I_k-1} \sin\Big(\frac{\nu_k \pi}{\tau_k}\ p_k h\Big) \sin\Big(\frac{\nu_k' \pi}{\tau_k}\ p_k h\Big) = \begin{cases} I_k/2 & \nu_k = \nu_k' \\ 0 & \nu_k \neq \nu_k' \end{cases} \qquad (k = 1, 2, \ldots, n)$$

we obtain by the methods of Phillips and Wiener [25] *

$$u(x_1, x_2, \ldots, x_n) =$$

$$= \frac{2^{n-1}}{I_1 I_2 \cdots I_{n-1}} \sum_{\nu_1=1}^{I_1-1} \cdots \sum_{\nu_{n-1}=1}^{I_{n-1}-1} \Big\{ \sum_{p_1'=1}^{I_1-1} \cdots \sum_{p_{n-1}'=1}^{I_{n-1}-1} \prod_{k=1}^{n-1} \sin\Big(\frac{\nu_k \pi}{\tau_k}\ p_k h\Big) \prod_{k=1}^{n-1} \sin\Big(\frac{\nu_k \pi}{\tau_k}\ x_k\Big) \frac{\sinh \beta\pi x_n}{\sinh \beta\pi \tau_n} \Big\}$$

where

$$\sinh^2\Big(\frac{\beta h}{2}\Big) = \sum_{k-1}^{n-1} \sin^2\Big(\frac{\nu_k \pi}{\tau_k}\ \frac{h}{2}\Big)\ .$$

I have used the above formulas to obtain elementary solutions, (solutions where $g = 1$ at one point of S_h and 0 elsewhere), for squares with 4, 6 and 8 intervals on a side. The solution for any boundary values is clearly a linear combination of these solutions. For a rectangle with $I_1 \times I_2$ intervals on a side, $2(I_1-1)(I_2-1)(I_1+I_2-2)$ multiplications are required, assuming the elementary functions are known.

However, the computation of elementary functions for other regions is in general more difficult, and even for rectangular regions, unless several problems are to be done for the same region, the labor of computing the elementary functions would ordinarily not be justified.†

*Phillips and Wiener gave the solution for a unit cube $n = 3$.

†Moskovitz [24] has published tables for various rectangular regions whose shortest side contains at most four intervals. Liebmann [21] has given tables to 4 decimals for square regions with 4 and 6 intervals on a side.

For large N and for a region of general shape it seems more practical to proceed by successive approximation methods.

b. Iterative Methods

The Kormes, Liebmann and the Successive Overrelaxation Methods of systematic iteration have been described in the Introduction for general systems of linear equations. We shall give a rather general proof of the convergence of the Kormes and Liebmann Methods which is due to Geiringer [12]. From this the convergence of the Successive Overrelaxation Method follows as will be shown in Chapter III, where this method is discussed.

Theorem 2.1 Given the system of linear equations

$$\sum_{j=1}^{N} a_{i,j}u_j + d_i = 0 \qquad (i = 1, 2, \ldots, N) \tag{0.1}$$

where $(a_{i,j})$ is irreducible and where the elements of $(a_{i,j})$ satisfy (0.2), then the sequence of vectors given by the Kormes Method, (0.5), and the Liebmann Method, (0.6), each converge to the solution of (0.1).

Proof: (a) For the Liebmann Method, if $u^{(m)}$ is the m-th approximation to u defined by (0.7), then the m-th error function

$$e^{(m)} = u^{(m)} - u$$

is given by

$$\begin{cases} e_i^{(m)} = -\sum_{j=1}^{i-1} \dfrac{a_{i,j}}{a_{i,i}}\, e_j^{(m)} - \sum_{j=i+1}^{N} \dfrac{a_{i,j}}{a_{i,i}}\, e_j^{(m-1)} & (i = 1, 2, \ldots, N) \\ e^{(0)} = u^{(0)} - u \end{cases} \tag{2.1}$$

Let $$e_0^{(m)} = \max_{i=1,2,\ldots,N} |e_i^{(m)}|$$

For all i, we have by (0.2) part (c) $\quad |e_i^{(1)}| \le e_0^{(0)}$

and for some i, by (0.2) part (c) $\quad |e_{i_1}^{(1)}| < e_0^{(0)}$.

By the irreducibility of $(a_{i,j})$, we have $\quad |e_{i_1}^{(2)}| < e_0^{(0)}$

$$|e_{i_2}^{(2)}| < e_0^{(0)} \qquad \text{for some } i_2 \neq i_1.$$

$$|e_i^{(2)}| \le e_0^{(0)} \qquad (i = 1, 2, \ldots, N).$$

Continuing this process, we have finally $\quad |e_i^{(N)}| < e_0^{(0)} \qquad (i = 1, 2, \ldots, N).$

Letting $\quad \nu_1 = \max_{i=1,2,\ldots,N} \dfrac{|e_i^{*(N)}|}{e_0^{(0)}}$ where $\quad e_i^{*(0)} = e_0^{(0)} \qquad (i = 1, 2, \ldots, N)$,

we have

$$e_0^{(N)} \le \nu_1\, e_0^{(0)}, \qquad (\nu_1 < 1).$$

Similarly $$e_0^{(mN)} \leq \nu_1^m e_0^{(0)}$$

and hence $$\lim_{m\to\infty} e_0^{(m)} = 0$$

and $$\lim_{m\to\infty} u^{(m)} = u.$$

(b) For the Kormes Method, the argument is practically the same.

Q. E. D.

Corollary 2.1 For (1.11), the Kormes Method and the Liebmann Method each converge.

We remark that the Liebmann Method is convergent for any ordering of the equations, and that the ordering may be changed after each total step. [12]

As we shall see in Chapter II, although the Liebmann Method converges exactly twice as fast as the Kormes Method, the rate of convergence in either case is very slow. Shortley and Weller [30] have introduced two modifications to accelerate the convergence of the Liebmann Method for the Dirichlet Problem for $n = 2$:
(a) The first is the use of formulas by means of which the values of $u_i^{(m)}$ for an entire block are improved simultaneously. Square blocks with 4, 9 and 25 blocks are used. Ordinarily, the gain in convergence would seem to be offset by the extra work involved in the use of more complicated formulas. However, by a slight modification of the difference equations this extra work is reduced to about 20% using a square block with nine points, while the rate of convergence is improved by a factor of about 3.5.
(b) The second modification is the use of an extrapolation procedure after some iterations have been performed. In most cases the required number of iterations can be reduced by a factor of two or three. Recently Shanks [29] has developed a method which may permit a sufficiently accurate extrapolation after fewer iterations and hence increase the gain somewhat.

Nevertheless the factor of saving, effected by these modifications does not increase appreciably with h^{-1}. The Successive Overrelaxation Method described in Chapter III, on the other hand, affords a gain in convergence which does increase with h^{-1} and for the Dirichlet Problem is proportional to h^{-1}.

c. Relaxation Methods

Relaxation methods have been described briefly in the Introduction. We shall give below a proof, due to Temple [35], of the convergence of the relaxation methods when applied to any linear system (0.1) assuming that $(a_{i,j})$ is symmetric and positive definite.

Theorem 2.2 For the system of linear equations (0.1), if $(a_{i,j})$ is symmetric and positive definite, and if at each step a largest residual (in absolute value) is relaxed, then the relaxation method converges to the solution of (0.1).

Proof: Obviously, since $(a_{i,j})$ is positive definite a unique solution of (0.1) exists. Let $u^{(1)}$ be an approximate solution of (0.1) and let a largest residual (in absolute value) occur at $x^{(i)}$.

For any two vectors $u, v \in V_N$, we define the following symmetric bilinear form

$$\mathcal{Q}(u,v) = \sum_{i,j=1}^{N} a_{i,j}\ u_i\ v_j\ .$$

Since $(a_{i,j})$ is positive definite, $\mathcal{Q}(u,v) \geq 0$ and $\mathcal{Q}(u,u) = 0$ if and only if $u = 0$.

If we now let

$$W(u) = \frac{1}{2}\mathcal{Q}(u,u) + \sum_{i=1}^{N} d_i \; u_i$$

then

$$\frac{\partial W}{\partial u_i} = \sum_{j=1}^{N} a_{i,j} \; u_j + d_i = 0 \qquad (i = 1, 2, \ldots, N)$$

and

$$\frac{\partial^2 W}{\partial u_i \partial u_j} = a_{i,j} \qquad (i, j = 1, 2, \ldots, N).$$

As is well known, see, for example, Widder ([37] pages 109–112), if u satisfies (0.1), then, since $(a_{i,j})$ is positive definite, W assumes a relative minimum, which we denote by W_0. If we choose a new orthogonal basis of V_N by an orthogonal transformation ([3] page 247), we have

$$\mathcal{Q}(u,u) = \sum_{i=1}^{N} \lambda_i u_i^{*2} \qquad (\lambda_i > 0)$$

where $u_i^*, u_2^*, \ldots, u_N^*$ are the coordinates of u referred to the new basis. It is clear that as

$$\sum_{i=1}^{N} {u_i}^2$$

becomes very large, $W(u)$ is also very large. Hence, W_0 is an *absolute minimum* for $W(u)$ for all $u \in V_N$.

We have

$$\begin{aligned} W(u^{(1)} + te_i) &= \frac{1}{2}\mathcal{Q}(u^{(1)} + te_i \; , u^{(1)} + te_i) + \sum_{j=1}^{N} d_j[u_j^{(1)} + te_i] \\ &= \frac{1}{2}\mathcal{Q}(u^{(1)}, u^{(1)}) + t \; \mathcal{Q}(u^{(1)}, e_i) + \frac{1}{2}t^2\mathcal{Q}(e_i, e_i) + \sum_{j=1}^{N} d_j[u_j^{(1)} + te_i] \\ W(u^{(1)} + te_i) &= W(u^{(1)}) + t \left[\mathcal{Q}(u^{(1)}, e_i) + \sum_{j=1}^{N} d_j\right] + \frac{1}{2}t^2\mathcal{Q}(e_i, e_i) \; . \end{aligned}$$

This expression is minimized by setting

$$t = -\frac{\mathcal{Q}(u^{(1)}, e_i) + d_i}{\mathcal{Q}(e_i, e_i)} = \frac{-\left[\sum\limits_{j=1}^{N} a_{i,j}u_j^{(1)} + d_i\right]}{a_{i,i}} \; .$$

If, as in (0.9), we obtain $u^{(2)}$ from $u^{(1)}$ by relaxing the residual at $x^{(i)}$, we have

$$\begin{cases} u_i^{(2)} = u_i^{(1)} + \dfrac{-\left[\sum\limits_{j=1}^{N} a_{i,j}u_j^{(1)} + d_i\right]}{a_{i,i}} \\ u_j^{(2)} = u_j^{(1)} \qquad (i \neq j) \end{cases}$$

and

$$W^{(2)} = W(u^{(2)}) = W(u^{(1)}) - \frac{(Z_i^{(1)})^2}{a_{i,i}} \ .$$

If this process is repeated we get a decreasing sequence of values of $W^{(m)}$ which approach a limit since $W^{(m)}$ is bounded below by W_0. Since $a_{ii} > 0 \quad (i = 1, 2, \ldots, N)$ (because $(a_{i,j})$ is positive definite), the largest residual of $u^{(m)}$ approaches zero; hence for all $i = 1, 2, \ldots, N$

$$\lim_{m \to \infty} \{Z_i^{(m)}\} = 0.$$

Now let $a^{i,j}$ be the cofactor of $a_{i,j}$ in the determinant of $(a_{i,j})$. We have

$$\begin{aligned} u_i &= \sum_{j=1}^{N} -a^{i,j} d_j \qquad (i = 1, 2, \ldots, N) \\ u_i^{(m)} &= \sum_{j=1}^{N} -a^{i,j} (d_j + Z_j^{(m)}) \qquad (i = 1, 2, \ldots, N) \ . \end{aligned}$$

Therefore, for $i = 1, 2, \ldots, N$,

$$\lim_{m \to \infty} u_i^{(m)} = \lim_{m \to \infty} \sum_{j=1}^{N} -a^{i,j} (d_j + Z_j^{(m)}) = -\sum_{j=1}^{N} -a^{i,j} d_j = u_i \ .$$

Q. E. D.

Theorem 2.3 If $(a_{i,j})$ is an $N \times N$ matrix satisfying (0.2) and (0.3) then $(a_{i,j})$ is positive definite.

Proof: It is well known that if $(a_{i,j})$ is symmetric, then $(a_{i,j})$ is positive definite if and only if

$$\sum_{j=1}^{N} a_{i,j} \ u_j^* = \lambda u_i^* \qquad (\lambda \leq 0) \qquad (i = 1, 2, \ldots, N)$$

implies $u_i{}^* = 0, \quad (i = 1, 2, \ldots, N)$: see for instance [3] pages 245 and 305. Thus, suppose for some $\lambda \leq 0$

$$\sum_{j=1}^{N} a_{i,j} \ u^*{}_j = \lambda u^*{}_i \qquad (i = 1, 2, \ldots, N).$$

Then

$$(a_{ii} - \lambda) u_i^* + \sum_{\substack{j=1 \\ j \neq i}}^{N} a_{i,j} \ u_j^* = 0 \qquad (i = 1, 2, \ldots, N)$$

and by (0.2)

$$(a_{ii} - \lambda) \geq \sum_{\substack{j=1 \\ j \neq i}}^{N} |a_{i,j}| \qquad (i = 1, 2, \ldots, N)$$

and for some i

$$(a_{ii} - \lambda) > \sum_{\substack{j=1 \\ j \neq i}}^{N} |a_{i,j}| \ .$$

By Theorem 1.2, since $a_{i,j} - \lambda I$ satisfies (0.2), the determinant of $(a_{i,j} - \lambda I)$, (where I is the identity matrix), does not vanish. Therefore

$$u_i^* = 0 \qquad (i = 1, 2, \ldots, N).$$

Q. E. D.

Corollary 2.2 If $(a_{i,j})$ satisfies (0.2) and (0.3) then the relaxation method converges to the solution of (0.1) if, at each step a largest residual is relaxed.

By the method of Theorem 2.1, it is clear that the relaxation method converges if $(a_{i,j})$ satisfies (0.2) if the residuals are relaxed in any order subject to the following restriction:

There exists a fixed number M such that for all i and m the residual at $x^{(i)}$ is relaxed at least once, on an iteration after the m-th and before the $(m + M)$-th.

We have already observed that relaxation methods are not well suited for large automatic computing machines. However, an electric computing board developed at the Watertown Arsenal Laboratory [5], affords a convenient method of relaxing residuals for the Dirichlet Problem with a plane region Ω_h which can be inscribed in a square region with 21 intervals on a side. A machine capable of being used with larger regions could be built.

Mounted on a large board is a square lattice of electric terminals, each corresponding to a net point. Each net point is connected to the four adjacent points by fixed resistances. The terminals corresponding to points of $\mathcal{S}_h$ are grounded. The residuals of the trial solution $u^{(0)}$ are computed and currents, proportional to these residuals are introduced into the corresponding terminals. The voltages, which are then measured, are approximately to within the accuracy of the electric measuring devices, proportional to the corrections to be added to the $u_i^{(0)}$. The residuals of $u^{(1)}$, the first improved approximation thus obtained, are then computed and the process repeated if necessary. Any desired accuracy may be obtained.

Through the courtesy of the Watertown Arsenal Laboratory, I was able to use the computing board to solve several problems each involving about 150 net points. I found that the maximum residual was reduced by a factor of from 10 to 50. Each iteration required about 3 hours, with the greatest part of the time being devoted to computing the residuals (by a desk machine), adjusting the currents and reading the voltages. An experienced operator can perform about three complete iterations with a 20×20 region in two eight-hour days.

Chapter II

Rates of Convergence of the Kormes Method and of the Liebmann Method

Shortley and Weller [30] investigated the rapidity of convergence of $\mathcal{L}_\sigma$, the transformation defined by the Liebmann Method, see (0.7), for the case of the Dirichlet Problem. They considered the eigenvalues of $\mathcal{L}_\sigma$ and derived asymptotic estimates for them, for small h. In this chapter, I will derive an exact relation, for certain σ, between the eigenvalues and eigenvectors of $\mathcal{L}_\sigma$ and those of $\mathcal{K}$, the transformation defined by the Kormes Method, see (0.5). The analysis of $\mathcal{K}$ is relatively easy, and using these relations an exact analysis is made of $\mathcal{L}_\sigma$. From this analysis, estimates for the number of iterations required to obtain a specified accuracy are derived.

§3. The Rate of Convergence of a Linear Transformation

We shall consider first the rate of convergence of a linear transformation T of an N dimensional complex vector space, V_N, onto itself. If $f = (f_1, f_2, \ldots, f_N)$ is any vector of V_N, we define the **norm** of f by

$$\|f\| = \left[\sum_{i=1}^{N} |f_i|^2 \right]^{1/2} . \tag{3.1}$$

Also, with the inner product

$$(f, g) = \sum_{i=1}^{N} f_i \, \bar{g}_i \tag{3.2}$$

V_N is a Euclidean vector space.

Definition 3.1 T is a **convergent** transformation if for all $f \in V_N$

$$\lim_{m \to \infty} \|T^m(f)\| = 0$$

Definition 3.2 The **rate of convergence** of a convergent transformation, T, is

$$\phi(T) = -\log \psi(T) \tag{3.3}$$

where

$$\psi(T) = \lim_{m \to \infty} \left[\operatorname*{lub}_{\substack{f \in V_N \\ m_1 > m}} \sqrt[m_1]{\frac{\|T^{m_1}(f)\|}{\|f\|}} \right] . \tag{3.4}$$

As we shall see later, the number of times T must be applied to a vector f to reduce $\|f\|$ to a specified fraction of itself is, approximately, inversely proportional to the rate of convergence.

It is well known, see for example MacDuffee [22], Chapter VII, that T has M distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_M$ with multiplicities $\kappa_1, \kappa_2, \ldots, \kappa_M$ where

$$\sum_{i=1}^{M} \kappa_i = N .$$

Definition 3.3 λ_1 is the **dominant eigenvalue** of T if

$$|\lambda_1| \geq |\lambda|$$

where λ is any eigenvalue of T.

We shall show below that $|\lambda_1| = \psi(T)$. The Jordan normal matrix form is, written in block form, ([22] pages 236 and 241)

$$\begin{pmatrix} J_1 & 0 & 0 & . & . & . & 0 \\ 0 & J_2 & 0 & . & . & . & 0 \\ 0 & 0 & J_3 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & J_{M_1} \end{pmatrix} \tag{3.5}$$

where $M_1 \geq M$ and $J_i \quad (i = 1, 2, \ldots, M_1)$ is a square Jordan matrix of the form

$$\begin{pmatrix} \lambda_i & 0 & 0 & . & . & . & 0 \\ 1 & \lambda_i & 0 & . & . & . & 0 \\ 0 & 1 & \lambda_i & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & \lambda_i \end{pmatrix} . \tag{3.6}$$

Here, not all the λ_i need be distinct. Let the number of rows and columns of the matrix J_i be $\nu_i, \quad (i = 1, 2, \ldots, M_1)$.

Associated with the normal matrix form (3.5) is a basis

$$\{v_{i,k}\} \qquad \begin{pmatrix} i = 1, 2, \ldots, M_1 \\ k = 1, 2, \ldots, \nu_i \end{pmatrix}$$

with associated eigenvalue λ_i.

We can normalize the $\{v_{i,k}\}$ so that

$$\|v_{i,k}\| = 1 \qquad \begin{pmatrix} i = 1, 2, \ldots, M_1 \\ k = 1, 2, \ldots, \nu_i \end{pmatrix} . \tag{3.7}$$

For any $f \in V_N$ we have

$$f = \sum_{i=1}^{M_1} \sum_{k=1}^{\nu_i} A_{i,k} \; v_{i,k} \tag{3.8}$$

and

$$T^m(v_{i,k}) = \sum_{s=0}^{k-1} {}_mC_s \; \lambda_i^{m-s} \; v_{i,k-s} \qquad \begin{pmatrix} i = 1, 2, \ldots, M_1 \\ k = 1, 2, \ldots, \nu_i \end{pmatrix} \tag{3.9}$$

where

$${}_mC_s = \frac{m!}{s!(m-s)!} .$$

By using (3.8) and (3.9), Dresden proved that if $|\lambda_1| < 1$, T is a convergent transformation. The following theorem is obtained in a similar way. It undoubtedly has appeared in the literature although I have been unable to find it.

Theorem 3.1 If T is a convergent transformation and if λ_1 is the dominant eigenvalue of T then

$$\psi(T) = |\lambda_1| \ . \tag{3.10}$$

Proof: If we let

$$x_m = \operatorname*{lub}_{\substack{f \in V_N \\ m_1 > m}} \sqrt[m_1]{\frac{\|T^{m_1}(f)\|}{\|f\|}}$$

the sequence $\{x_m\}$ is monotone decreasing, and is bounded below by zero. Therefore, $\lim_{m\to\infty} x_m$ exists.

By the homogeneity of T, we can assume $\|f\| = 1$. By (3.8) and (3.9) we have

$$T^{m_1}(f) = \sum_{i=1}^{M_1} \sum_{k=1}^{\nu_i} \sum_{s=0}^{k-1} A_{i,k\ m_1} C_s \lambda_i^{m_1 - s}\ v_{i,k} \ . \tag{3.11}$$

But $s \le N$; hence, ${}_{m_1}C_s \le m_1^s \le m^N$.

By the linear independence of the $v_{i,k}$, the coefficients $A_{i,k}$ of (3.8) are bounded for all $\|f\| = 1$. Let $A = \max_{i,k} |A_{i,k}|$. We have

$$\|T^{m_1}(f)\| \le NA\ m_1^N |\lambda_1|^{m_1 - N}$$

and

$$x_m \le \operatorname*{lub}_{m_1 > m} \left[\left\{ NA\ m_1^N |\lambda_1|^{m_1 - N} \right\}^{1/m_1} \right]$$

$$x_m \le \operatorname*{lub}_{m_1 > m} \left[(NA)^{1/m_1}\ m_1^{N/m_1}\ |\lambda_1|^{1 - N/m_1} \right].$$

But for all $m_1 > e$, $\quad m_1^{N/m_1}$ is a decreasing function of m_1. Hence †

$$x_m \le (NA)^{1/m}\ m^{N/m}\ |\lambda_1|^{1 - N/m} \ .$$

Therefore,

$$\psi(T) = \lim_{m\to\infty} x_m \le |\lambda_1|.$$

On the other hand, if $\quad T(f) = \lambda_1 f, \quad x_m = |\lambda_1|$ for all m. Therefore $\psi(T) \ge |\lambda_1|$ and the theorem follows.

Q. E. D.

Corollary 3.1

$$\phi(T) = -\log |\lambda_1|. \tag{3.12}$$

Corollary 3.2 T is a convergent transformation if and only if all eigenvalues of T are smaller than one in absolute value.

Theorem 3.2 By similar methods we can prove

$$\psi(T) = \lim_{m\to\infty} \sqrt[m]{\|T^m\|} \tag{3.13}$$

where

$$\|T^m\| = \operatorname*{lub}_{f \in V_N} \ \frac{\|T^m(f)\|}{\|f\|} \ . \tag{3.14}$$

† We note that $|\lambda_1| < 1$ since T is convergent. For if f is chosen so that $T(f) = \lambda_1 f$ then $\|T^m(f)\| = |\lambda_1|^m \|f\|$ and would not approach zero unless $|\lambda_1| < 1$.

We now consider the number of iterations necessary to reduce $\|f\|$ to a specified fraction of itself.

Definition 3.4 If T is a convergent transformation on V_N

$$\mathcal{N}_f(T, \rho, V_N)$$

denotes the smallest number m such that

$$\|T^m(f)\| \leq \rho \|f\| \ , \qquad (0 < \rho < 1).$$

Definition 3.5

$$\mathcal{N}(T, \rho, V_N) = \operatorname*{lub}_{f \in V_N} \mathcal{N}_f(T, \rho, V_N) \ . \tag{3.15}$$

When no confusion will arise we may drop the third argument for $\mathcal{N}(T, \rho, V_N)$.

Theorem 3.3

$$\lim_{\rho \to 0} \left[\mathcal{N}(T, \rho) \Big/ \frac{-\log \rho}{-\log |\lambda_1|} \right] = 1 \ . \tag{3.16}$$

Proof: For all ρ we have

$$\mathcal{N}_{v_1}(T, \rho) = \frac{-\log \rho}{-\log |\lambda_1|}$$

where $T(v_1) = \lambda_1 \ v_1$. Hence

$$\mathcal{N}(T, \rho) \geq \frac{-\log \rho}{-\log |\lambda_1|} \ .$$

The requirement that $\|f\|$ be reduced to a fraction ρ of itself is equivalent to requiring that

$$\|T^m\| = \operatorname*{lub}_{f \in V_N} \frac{\|T^m(f)\|}{\|f\|} = \rho \ .$$

Obviously as $\rho \to 0, \ m \to \infty$.

By Theorem 3.2,

$$|\lambda_1| + \epsilon(m) = \sqrt[m]{\|T^m\|}$$

where

$$\lim_{m \to \infty} \epsilon(m) = 0 \ .$$

The condition $\|T^m\| = \rho$ requires that

$$[\ |\lambda_1| + \epsilon(m)\ ]^m = \|T^m\| = \rho$$

or

$$m = \frac{-\log \rho}{-\log[\ |\lambda_1| + \epsilon(m)\ ]}$$

$$\frac{m}{\dfrac{-\log \rho}{-\log |\lambda_1|}} = \frac{-\log |\lambda_1|}{-\log[\ |\lambda_1| + \epsilon(m)\ ]} \ .$$

By the continuity of the logarithmic function for $|\lambda_1| > 0$

$$\lim_{\rho \to 0} \frac{m}{\dfrac{-\log \rho}{-\log |\lambda_1|}} = 1 \ .$$

Q. E. D.

§4. The Kormes Method

The improvement formula for the Kormes Method of iteration is given by (0.5). Using (1.14) and (1.15), we have

$$u_i^{(m+1)} = \sum_{j=1}^{N} b_{i,j}\ u_j^{(m)} + c_i \qquad (i = 1, 2, \ldots, N) \tag{4.1}$$

or

$$u^{(m+1)} = \mathcal{K}[u^{(m)}] + c \tag{4.1a}$$

where c denotes the vector $(c_1, c_2, \ldots, c_N)$. The general element of $(\mathcal{K}_{i,j})$, the matrix of $\mathcal{K}$, is given by

$$\mathcal{K}_{i,j} = b_{i,j} \qquad (i, j = 1, 2, \ldots, N)\ . \tag{4.2}$$

By (1.16) and (1.14),

$$b_{i,i} = 0 \qquad (i = 1, 2, \ldots, N)$$
$$b_{i,j} \geq 0 \quad i \neq j \quad (i, j = 1, 2, \ldots, N)$$

and the matrix $(b_{i,j})$ is irreducible.

By Theorem 3.1, the rate of convergence of the Kormes Method depends on the dominant eigenvalue of $\mathcal{K}$.

We note that $r = (r_1.r_2, \ldots, r_N)$ is an *eigenvector* of $\mathcal{K}$ with eigenvalue μ if and only if

$$\mathcal{K}[r] = \mu r \tag{4.3}$$

or, equivalently

$$\sum_{j=1}^{N} b_{i,j}\ r_j = \mu r_i \qquad (i = 1, 2, \ldots, N)\ . \tag{4.3a}$$

Theorem 4.1 If $(a_{i,j})$ is a symmetric $N \times N$ matrix satisfying conditions (0.2), then the normal form of the matrix of $\mathcal{K}$ is a real diagonal matrix; if μ_1 denotes the largest eigenvalue of $\mathcal{K}, \mu_1$ is simple (not repeated).

Lemma 4.1 The eigenvalues of $\mathcal{K}$ are the same as the eigenvalues of $\mathcal{K}'$ where the general element of $(\mathcal{K}'_{i,j})$ is

$$\mathcal{K}'_{i,j} = b'_{i,j} = \begin{cases} \dfrac{-a_{i,j}}{\sqrt{a_{i,i}}\ \sqrt{a_{i,j}}} & (i \neq j) \\ 0 & (i = j) \end{cases}$$

Proof: If

$$\mathcal{K}(r) = \mu r \tag{4.4}$$

then

$$\sum_{j=1}^{N} b_{i,j}\ r_j = \mu r_i \qquad (i = 1, 2, \ldots, N)$$
$$\sum_{j=1}^{N} b_{i,j}\sqrt{a_{i,i}}\ r_j = \mu\sqrt{a_{i,i}}\ r_i \qquad (i = 1, 2, \ldots, N)\ .$$

Let

$$r'_i = \sqrt{a_{i,i}}\ r_i \qquad (i = 1, 2, \ldots, N) \tag{4.5}$$
$$\sum_{j=1}^{N} b_{i,j}\ \frac{\sqrt{a_{i,i}}}{\sqrt{a_{j,j}}}\ r'_j = \mu r'_i$$
$$= -\sum_{j=1, j\neq i}^{N} \frac{a_{i,j}}{\sqrt{a_{j,j}}\ \sqrt{a_{i,i}}}\ r'_j = -\sum_{j=1}^{N} b'_{i,j}\ r'_j\ .$$

Thus

$$\mathcal{K}'[r'] = \mu r'. \tag{4.6}$$

Conversely if μ is an eigenvalue of $\mathcal{K}'$ we can show that μ is an eigenvalue of $\mathcal{K}$. This proves the lemma.

Q. E. D.

Since $(\mathcal{K}'_{i,j})$ is symmetric by [3] pages 305–6, there exists an orthonormal basis of eigenvectors

$$r^{(1)'}, r^{(2)'}, \ldots, r^{(N)'}$$

such that

$$\mathcal{K}'[r^{(i)'}] = \mu_i r^{(i)'} \qquad (i = 1, 2, \ldots, N)$$

where

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_N \ . \tag{4.7}$$

We show that the corresponding vectors $r^{(i)}$ satisfying (4.4) are linearly independent even if some μ is repeated k times. In this case let $r^{(1)'}, r^{(2)'}, \ldots, r^{(k)'}$ be the linearly independent eigenvalues of $\mathcal{K}'$ such that for $j = 1, 2, \ldots, k$

$$\mathcal{K}'[r^{(j)'}] = \mu r^{(j)'} \ .$$

Obviously $r^{(j)}, \quad (j = 1, 2, \ldots, k)$ are also linearly independent where

$$r_i^{(j)} = \frac{r_i^{(j)'}}{\sqrt{a_{ii}}} \qquad \begin{pmatrix} i = 1, 2, \ldots, N \\ j = 1, 2, \ldots, k \end{pmatrix}$$

and $a_{ii} > 0$.

Hence the normal form of $(\mathcal{K}_{i,j})$ is diagonal.

Lemma 4.2 Let $(a_{i,j})$ be a real symmetric $N \times N$ matrix.
(a) μ_1 is the largest eigenvalue of $(a_{i,j})$ if and only if μ_1 is the maximum value of

$$\mathcal{Q}[w] = \sum_{i,j=1}^{N} a_{i,j} w_i w_j \tag{4.8}$$

subject to the condition

$$\|w\|^2 = \sum_{i=1}^{N} w_i^2 = 1 \tag{4.9}$$

(b) If $\|w^{(1)}\|^2 = 1$, then $w^{(1)}$ is an eigenvector associated with μ_1 if and only if

$$\mathcal{Q}[w^{(1)}] = \mu_1.$$

Proof: (a) It is well known that any real symmetric matrix is orthogonally equivalent to a real diagonal matrix $(c_{i,j})$, the largest of whose elements is the maximum of $\mathcal{Q}[w]$ subject to the requirement $\|w\|^2 = 1$, (see for instance [3] pages 247–250). The maximum clearly exists and is attained because $\mathcal{Q}[w]$ is a continuous function on a compact set. The diagonal elements of $(c_{i,j})$ are eigenvalues of $(a_{i,j})$ and conversely (see [3] page 305). Thus, (a) is proved.

(b) As in Widder [37] pages 113–115, the maximum problem may be formulated in terms of Lagrange's multipliers. We have the conditions

$$\text{(4.10)} \qquad \begin{cases} \dfrac{\partial}{\partial w_i}\left\{\displaystyle\sum_{i,j=1}^{N} a_{i,j}w_iw_j - \mu\Big[\sum_{i=1}^{N} w_i^2 - 1\Big]\right\} = 0 \qquad (i = 1, 2, \ldots, N) \\ \displaystyle\sum_{i=1}^{N} w_i^2 = 1 \end{cases}$$

where μ is a Lagrange's multiplier. Upon simplification these conditions become

$$\text{(4.11)} \qquad \begin{cases} \displaystyle\sum_{j=1}^{N} a_{i,j}w_j = \mu w_i \qquad (i = 1, 2, \ldots, N) \\ \displaystyle\sum_{i=1}^{N} w_i^2 = 1 \ . \end{cases}$$

If w satisfies (4.11), we have

$$\text{(4.12)} \qquad \begin{aligned} \mathcal{Q}[w] \ &= \sum_{i,j=1}^{N} a_{i,j}w_iw_j = \sum_{i=1}^{N} w_i\Big\{\sum_{j=1}^{N} a_{i,j}w_j\Big\} \\ &= \sum_{i=1}^{N} w_i\mu w_i = \mu \ . \end{aligned}$$

Thus, if $w^{(1)}$, where $\|w^{(1)}\|^2 = 1$, is an eigenvector of $(a_{i,j})$ with eigenvalue μ_1, $\mathcal{Q}[w^{(1)}] = \mu_1$. Conversely, if $\mathcal{Q}[w^{(1)}] = \mu_1$ and $\|w^{(1)}\|^2 = 1$, since μ_1 is the maximum of $\mathcal{Q}[w]$ subject to $\|w\|^2 = 1$, by (4.11), $w^{(1)}$ is an eigenvector of $(a_{i,j})$ with eigenvalue μ_1.

Q. E. D.

Lemma 4.3 Let $(a_{i,j})$ be a real symmetric $N \times N$ matrix such that

$$a_{i,i} = 0 \qquad (i = 1, 2, \ldots, N)$$

$$a_{i,j} \geq 0 \qquad (i, j = 1, 2, \ldots, N)$$

and $(a_{i,j})$ is irreducible. If μ_1 is the largest eigenvalue of $(a_{i,j})$ and if $w^{(1)}$ is any eigenvector associated with μ_1, then either

$$w_i^{(1)} > 0 \qquad (i = 1, 2, \ldots, N)$$

or

$$w_i^{(1)} < 0 \qquad (i = 1, 2, \ldots, N) \ .$$

Moreover, μ_1 is simple (not repeated).

Proof: By Lemma 4.2, for all $w \in V_N$,

$$\mathcal{Q}[w^{(1)}] \geq \mathcal{Q}[w] \ .$$

On the other hand, since $a_{i,j} \geq 0 \qquad (i, j, = 1, 2, \ldots, N)$,

$$\mathcal{Q}[w^{(1)}] \leq \mathcal{Q}[\widetilde{w}^{(1)}] \ ,$$

where

$$\widetilde{w}^{(1)} = (\ |w_1^{(1)}|, |w_2^{(1)}|, \ldots, |w_N^{(1)}|\ )\ . \tag{4.13}$$

Equality can only occur if, whenever $a_{i,j} \neq 0$

$$w_i^{(1)} w_j^{(1)} \geq 0\ . \tag{4.14}$$

By Lemma 4.2, $\widetilde{w}^{(1)}$ is an eigenvalue of $(a_{i,j})$. Suppose that for some $i, w_i^{(1)} = 0$. Then $\widetilde{w}_i^{(1)} = 0$ and

$$\sum_{j=1}^{N} a_{i,j} \widetilde{w}_j^{(1)} = \mu_1 \widetilde{w}_i^{(1)} = 0$$

and since $a_{(i,j)} \geq 0$ we have $\widetilde{w}_j^{(1)} = 0$ for all j such that $a_{i,j} \neq 0$. Using the irreducibility of $(a_{i,j})$ it can be proved that[‡]

$$\widetilde{w}_i^{(1)} = 0 \qquad (i = 1, 2, \ldots, N)\ .$$

This contradicts the condition

$$\|w^{(1)}\|^2 = \|\widetilde{w}^{(1)}\|^2 = 1$$

Therefore, $w_i^{(1)} \neq 0 \qquad (i = 1, 2, \ldots, N)$.
Again, by the irreducibility of $(a_{i,j})$, since $w_i^{(1)} w_j^{(1)} \geq 0$ whenever $a_{i,j} \neq 0$ we can show that

$$w_i^{(1)} > 0 \qquad (i = 1, 2, \ldots, N)$$

or

$$w_i^{(1)} < 0 \qquad (i = 1, 2, \ldots, N)\ .$$

To show that μ_1 is *simple* we first note that there exists an orthonormal basis of N eigenvectors for $(a_{i,j})$. If μ_1 is repeated k times then there would exist k linearly independent orthogonal eigenvectors each associated with μ_1. But since each is non vanishing and one signed it is impossible that they should be orthogonal.

Q. E. D.

From Lemma 4.3, it follows that the largest eigenvalue of $\mathcal{K}'$ is simple. The theorem follows at once by Lemma 4.1.

Q. E. D.

Theorem 4.2 Let $(a_{i,j})$ be a real symmetric $N \times N$ matrix, $(N > 1)$ such that

$$a_{ii} = 0 \qquad (i = 1, 2, \ldots, N)$$

$$a_{i,j} \geq 0 \qquad (i = 1, 2, \ldots, N)$$

and $(a_{i,j})$ is irreducible. Let $(a_{i,j}^*)$ be the $N' \times N'$ matrix obtained from $(a_{i,j})$ by deleting the elements in $(N - N')$ rows and the corresponding columns, where $N' < N$. If μ_1 and μ_1^* denote the largest eigenvalues of $(a_{i,j})$ and $(a_{i,j}^*)$ then $\mu_1 < \mu_1^*$.

[‡]See the proof of Theorem 1.2.

Proof: Let
$$\mathcal{Q}^*[w] = \sum_{i,j=1}^{N'} a^*_{i,j} w_i w_j$$
and
$$\mathcal{Q}[w] = \sum_{i,j=1}^{N} a_{i,j} w_i w_j \ .$$

If $w^{*(1)}$ where $\|w^{*(1)}\|^2 = 1$ is an eigenvector of $(a^*_{i,j})$ then
$$\mu^*_1 = \mathcal{Q}^*[w^{*(1)}] = \mathcal{Q}[w^{*(1)}]$$
where $w^{*(1)}_i$ in the second expression is taken to be zero if the ith row and column have been deleted. Hence $\mu_1 \geq \mu^*_1$. By Lemma 4.3, Theorem 4.1, $w^{*(1)}$ is not an eigenvector of $(a_{i,j})$ associated with the largest eigenvalue of $(a_{i,j})$. Therefore, $\mu_1 > \mu^*_1$.

Q. E. D.

Corollary 4.1 If $R^{(1)}_h \subset R^{(2)}_h$ where $R^{(1)}_h \neq R^{(2)}_h$ and $R^{(2)}_h$ is connected, and if $\mu^{(1)}_1$ and $\mu^{(2)}_1$ denote the largest eigenvalue of $\mathcal{K}$ applied to (1.11), for the networks $\Omega^{(1)}_h$ and $\Omega^{(2)}_h$ with interiors $R^{(1)}_h$ and $R^{(2)}_h$ respectively, then
$$\mu^{(1)}_1 < \mu^{(2)}_1 \ .$$

We have already seen, (Theorem 1.1), that $(a_{i,j})$ has property (A_2). We define the vector $\gamma \in V_N$ such that

(4.15)
$$\begin{aligned} &\gamma_i = 1 \qquad i \in T_1, \text{ the set where } \sum_{k=1}^{m} p^{(i)}_k \text{ is odd} \\ &\gamma_i = 0 \qquad i \in T_2, \text{ the set where } \sum_{k=1}^{m} p^{(i)}_k \text{ is even.} \end{aligned}$$

Theorem 4.3 If $\mathcal{K}[r] = \mu r$ and if

(4.16)
$$r^* = (-1)^\gamma r$$

then

(4.17)
$$\mathcal{K}[r^*] = (-\mu) r^* \ .$$

(We do not use the fact that $(a_{i,j})$ is symmetric).

Proof:
$$\sum_{j=1}^{N} b_{i,j} r^*_j = \sum_{j=1}^{N} b_{i,j} (-1)^{\gamma_j} r_j \qquad (i = 1, 2, \ldots, N) \ .$$

If $b_{i,j} \neq 0, \quad i \in T_1$, and $j \in T_2 \quad$ or $i \in T_2$ and $j \in T_1$. We have
$$\sum_{j=1}^{N} b_{i,j} r^*_j = -(-1)^{\gamma_i} \sum_{j=1}^{N} b_{i,j} r_j$$
$$= -(-1)^{\gamma_i} \mu r_i = (-\mu) r^*_i \qquad (i = 1, 2, \ldots, N) \ .$$

Hence, by (4.3a),
$$\mathcal{K}[r^*] = \mu r^* \ .$$

Q. E. D.

Corollary 4.2 The eigenvector r^* may be expressed in the form

$$r^* = (1 - 2\gamma)r. \tag{4.18}$$

Now let N_1, N_2 be the number of integers in T_1, T_2 respectively.

Theorem 4.4 Let $\bar{s}$ be the number of non-zero eigenvalues of $\mathcal{K}$ and let

$$s' = \min(N_1, N_2) \qquad s'' = \max(N_1, N_2)\ . \tag{4.19}$$

If $(a_{i,j})$ is symmetric, then
(a) There are $\bar{s}$ eigenvectors associated with non-zero eigenvalues where $0 \leq \bar{s} \leq 2s'$, and $\bar{s}$ is even. These associated eigenvalues do not vanish identically on either T_1 or T_2.
(b) The $(N - \bar{s})$ dimensional null space V_0 of $\mathcal{K}$ can be referred to a basis of $(N - \bar{s})$ vectors such that $(N_2 - \bar{s}/2)$ of these vectors vanish identically on T_1 and $(N_1 - \bar{s}/2)$ vanish identically on T_2.

Proof:
(i) If $\mathcal{K}[r] = \mu r$, where r is not identically zero but r vanishes identically on either T_2 or T_1, then $\mu = 0$. For, $\mathcal{K}[r]$ vanishes identically everywhere. Thus, if $\mu \neq 0$, r must not vanish identically on either T_2 or T_1.

(ii) If $r^{(1)}$ and $r^{(2)}$ are any two linearly independent eigenvectors of $\mathcal{K}$ with non-zero eigenvalues it is obvious that $r^{*(1)}(\neq r^{(1)})$ and $r^{*(2)}(\neq r^{(2)})$ are linearly independent. Therefore, $\bar{s}$ is even. Now consider the sets $\mathcal{X}_1 = \{r^{(j)} - r^{*(j)}\}$ and $\mathcal{X}_2 = \{r^{(j)} + r^{*(j)}\}$, $(j = 1, 2, \ldots, \bar{s}/2)$. The vectors in $\mathcal{X}_1$ vanish in T_2 and those in $\mathcal{X}_2$ vanish in T_1. On the other hand, the $\bar{s}$ vectors of $\mathcal{X}_1 \cup \mathcal{X}_2$ are linearly independent. Hence $\bar{s}/2 \leq \bar{s}'$.

(iii) If $\mathcal{K}[r] = 0$ we may write

$$r = r' + r''$$

where r' vanishes identically on T_1 and r'' vanishes identically on T_2 and both r', r'' lie in V_0. The set of all vectors obtained in this way span V_0 and hence there exists a linearly independent subset spanning V_0. By (ii), $(N_1 - \bar{s}/2)$ of these vectors vanish identically on T_2 and $(N_2 - \bar{s}/2)$ vanish identically on T_1.

Q. E. D.

Given an arbitrary error vector, $e^{(0)}$ in V_N we may write

$$e^{(0)} = \sum_{j=1}^{N} A_j r^{(j)} \tag{4.20}$$

and by Theorem 4.4, we have

$$e^{(0)} = \sum_{j=1}^{\bar{s}/2} A_j r^{(j)} + A_j^* r^{*(j)} + \sum_{j=\bar{s}/2+1}^{N-\bar{s}/2} A_j r^{(j)} \tag{4.21}$$

where

$$\begin{aligned}
&\mathcal{K}[r^{(j)}] = \mu_j r^{(j)} && (j = 1, 2, \ldots, \bar{s}/2)\\
&\mu_j > 0 && (j = 1, 2, \ldots, \bar{s}/2)\\
&r^{(N-j)} = r^{*(j)} && (j = 1, 2, \ldots, \bar{s}/2)\\
&A_j^* = A_{N-j} && (j = 1, 2, \ldots, \bar{s}/2)\\
&\mathcal{K}[r^{(j)}] = 0 && (j = \bar{s}/2 + 1, \ldots, N - \bar{s}/2)\ .
\end{aligned}$$

After m iterations, $(m > 1)$, we have

$$e^{(m)} = \sum_{j=1}^{\bar{s}/2} \mu_j^m A_j r^{(j)} + (-\mu_j)^m A_j^* r^{*(j)} \ . \tag{4.22}$$

For those cases where $a_{ii} = 1$ as in the Dirichlet Problem, the $r^{(j)}$ are orthogonal. If we assume the $r^{(j)}$ have been normalized we have

$$A_j = \sum_{i=1}^{N} e_i^{(0)} r_i^{(j)} \qquad (j = 1, 2, \ldots, N) \tag{4.23}$$

and

$$\|e^{(0)}\| = \sum_{j=1}^{N} A_j^2 \quad . \tag{4.24}$$

For the Dirichlet Problem, we will now show that

$$\|e'\|^2 \leq \frac{\|Z'\|^2}{(1-\mu_1)^2} \tag{4.25}$$

where $e' = u' - u$, $\quad \mu'$ is an approximate solution of (1.11), and where $\|Z'\|^2$ is the sum of the squares of the residuals.[§] We have already, in Section 2, given a bound for $|u_i' - u_i|, \qquad (i = 1, 2, \ldots, N)$ due to Gerschgorin.

To prove (4.25), we use (4.23) to express e' in terms of the eigenvectors of $\mathcal{K}$. We have

$$e' = \sum_{j=1}^{N} A_j r^{(j)}$$

and

$$\|e'\|^2 = \sum_{j=1}^{N} A_j^2 \ .$$

Also

$$(\ell[u'])_i = \sum_{j=1}^{N} a_{i,j} u_j' + d_i \qquad (i = 1, 2, \ldots, N)$$

and

$$0 = \sum_{j=1}^{N} a_{i,j} u_j + d_i \qquad (i = 1, 2, \ldots, N).$$

Hence

$$(\ell[u'])_i = \sum_{j=1}^{N} a_{i,j} e_j' \qquad (i = 1, 2, \ldots, N) \ .$$

But since

$$a_{ii} = 1 \qquad (i = 1, 2, \ldots, N),$$

[§] The sum of the squares of the residuals has been used as a measure of convergence by Bowie [4].

and $$b_{i,j} = -a_{i,j} \qquad (i \neq j)$$

$$\begin{aligned}(\ell[u'])_i &= e'_i - (\mathcal{K}[e'])_i \\ &= e'_i - \sum_{j=1}^{N} \mu_j A_j r_i^{(j)} \qquad (i = 1, 2, \ldots, N) \\ (\ell[u'])_i &= \sum_{j=1}^{N} (1 - \mu_j) A_j r_i^{(j)} \qquad (i = 1, 2, \ldots, N)\end{aligned}$$

$$\|Z'\|^2 = \|(-\ell[u'])\|^2 = \sum_{j=1}^{N} (1 - \mu_j)^2 A_j^2 \geq (1 - \mu_1)^2 \|e'\|^2 \ .$$

Hence, (4.25) is proved.

For the Dirichlet Problem with a rectangular region, the eigenvalues and eigenvectors of $\mathcal{K}$ can be expressed explicitly. Let I_k be the number of intervals on the kth side. We have

$$r_i^{(j)} = \prod_{k=1}^{n} \sin \frac{\ell_k^{(j)} \pi p_k^{(i)}}{I_k} \tag{4.26}$$

where

$$\begin{aligned}\ell_k^{(j)} &= 1, 2, \ldots, I_{k-1} \\ j &= 1, 2, \ldots, N.\end{aligned}$$

The corresponding eigenvalue is

$$\mu_j = \frac{1}{n} \sum_{k=1}^{n} \cos \frac{\ell_k^{(j)} \pi}{I_k} \ . \tag{4.27}$$

This may be readily verified by (0.5) and (1.10).

In particular

$$\mu_1 = \frac{1}{n} \sum_{k=1}^{n} \cos \frac{\pi}{I_k} \tag{4.28}$$

For large $I_k \qquad (k = 1, 2, \ldots, n)$

$$\mu_1 \sim 1 - \frac{\pi^2}{4} \sum_{k=1}^{n} \frac{1}{I_k^2} \tag{4.29}$$

and

$$\phi(\mathcal{K}) \sim \frac{\pi^2}{4} \sum_{k=1}^{n} \frac{1}{I_k^2} \ . \tag{4.30}$$

The rate of convergence is therefore very slow for large I_k. We remark that μ_1 and hence the rate of convergence can be estimated for non rectangular regions by using Theorem 4.2, Corollary 4.1.

Number of Iterations

Theorem 4.5 For the Dirichlet Problem

$$\mathcal{N}(\mathcal{K}, \rho) = \frac{-\log \rho}{-\log \mu_1} \tag{4.31}$$

where $\mathcal{N}(\mathcal{K}, \rho)$ is defined by Definition 3.4, (the third argument has been omitted).

Proof: For all $f \in V_N$ by (4.22) and (4.24)

$$\| \mathcal{K}^m[f] \| \leq \mu_1^m \| f \| \ .$$

But

$$\mu_1^m \| f \| \leq \rho \| f \|$$

provided

$$m \geq \frac{-\log \rho}{-\log \mu_1} \ .$$

Hence

$$\mathcal{N}(\mathcal{K}, \rho) \leq \frac{-\log \rho}{-\log \mu_1} \ .$$

On the other hand, if $f = r^{(1)}$

$$\| \mathcal{K}^m[f] \| = \| u_1^m f \| = \mu_1 \| f \| \ .$$

Therefore

$$\mathcal{N}(\mathcal{K}, \rho) \geq \frac{-\log \rho}{-\log \mu_1}$$

and the theorem follows.

Q. E. D.

Theorem 4.6 Let Ω_h enclose a bounded region Ω, and let Ω_h be connected. Let $\Omega_{h_1}, \Omega_{h_2}, \ldots$ be nets covering Ω and obtained from Ω_h by subdividing the mesh by a factor of $2^1, 2^2, \ldots,$ then for the Dirichlet Problem, there exist positive constants m_1 and M_1 and a mesh size h_0 such that for all $h = h_0 2^{-i}$ with $0 < h \leq h_0$

$$m_1 \leq h^2 \mathcal{N}(\mathcal{K}, \rho, V_{N_h}) \leq M_1 \tag{4.32}$$

where V_{N_h} denotes the vector space of all functions defined on R_h.

Proof: Let $\square_2$ and $\square_1$ be rectangular regions circumscribing and inscribed in Ω. Let the sides of $\square_1$ and $\square_2$ be $\gamma_k^{(1)}$ and $\gamma_k^{(2)}$ respectively, $(k = 1, 2, \ldots, n)$ and let the $\gamma_k^{(j)}$ be chosen so they are multiples of h. Let $R_h^{(1)}$ and $R_h^{(2)}$ be the set of net points interior to $\square_1$ and $\square_2$ respectively. If a $\mu_1^{(1)}(h), \mu_1(h)$ and $\mu_1^{(2)}(h)$ denote the largest eigenvalue of $\mathcal{K}$ for $R_h^{(1)}, R_h$ and $R_h^{(2)}$ respectively we have by Theorem 4.2, Corollary 4.1,

$$\mu_1^{(1)}(h) \leq \mu_1(h) \leq \mu_1^{(2)}(h). \tag{4.33}$$

By (4.28),

$$\mu_1^{(1)} = 1 - \frac{\pi^2 h^2}{4} \sum_{k=1}^{n} (\gamma_k^{(1)})^{-2} + O(h^4)$$

$$-\log \mu_1^{(1)} = \frac{\pi^2 h^2}{4} \sum_{k=1}^{n} (\gamma_k^{(1)})^{-2} + O(h^4)$$

$$-\log \mu_1^{(2)} = \frac{\pi^2 h^2}{4} \sum_{k=1}^{n} (\gamma_k^{(2)})^{-2} + O(h^4)$$

where $O(h^4)$ vanishes with h^4.
Therefore, by (4.31),

$$\frac{-\log\rho}{-h^2\frac{\pi^2}{4}\sum_{k=1}^{n}(\gamma_k^{(1)})^{-2}+O(h^4)} \le \mathcal{N}(\mathcal{K},\rho,V_{N_h}) \le \frac{-\log\rho}{-h^2\frac{\pi^2}{4}\sum_{k=1}^{n}(\gamma_k^{(2)})^{-2}+O(h^4)}$$

and the theorem follows.

Q. E. D.

In the above sense we may state that as $h \to 0$ the required number of iterations is asymptotically proportional to h^{-2}. In symbols

(4.34) $$\mathcal{N}(\mathcal{K},\rho) = O(h^{-2}) \ .$$

§5. The Liebmann Method with a Consistent Ordering

The improvement formula for the Liebmann Method has been given in (0.7). Using (1.14) and (1.15), we have

$$\mu_i^{(m+1)} = \sum_{j=1}^{i-1} b_{i,j} u_j^{(m+1)} + \sum_{j=i+1}^{N} b_{i,j} u_j^{(m)} + c_i \tag{5.1}$$

or $$u^{(m+1)} = \mathcal{L}_\sigma[u^{(m)}] + c \tag{5.1a}$$

where σ denotes the ordering of the equations and c denotes the vector $(c_1, c_2, \ldots, c_N)$.

By Theorem 3.1, the rate of convergence of the Liebmann Method depends on the dominant eigenvalue of $\mathcal{L}_\sigma$. The following has been noted by Geiringer [12] and by Stein and Rosenberg [34]

Theorem 5.0

$$\mathcal{L}_\sigma[v] = \lambda v \qquad \text{if and only if}$$

$$\sum_{j=1}^{i-1} b_{i,j}\lambda v_j + \sum_{j=i+1}^{N} b_{i,j} v_j = \lambda v_i \qquad (i = 1, 2, \ldots, N)\ . \tag{5.2}$$

Moreover, if the matrix $(\xi_{i,j})$ is given by

$$\xi_{i,j} = \begin{cases} b_{i,j} & j > i \\ \lambda b_{i,j} & j < i \\ -\lambda & j = i \end{cases} \tag{5.3}$$

then $|\mathcal{L}_\sigma - \lambda I|$, the characteristic determinant of $\mathcal{L}_\sigma$, is equal to the determinant of $(\xi_{i,j})$.

Clearly, the determinant of $(\xi_{i,j})$ can be obtained from $|\mathcal{K} - \lambda I|$ by multiplying the elements of the latter determinant which are below the main diagonal by λ.

Consistent Orderings

Because $(a_{i,j})$ has property (A_q), it is possible to order the equations in such a way that there is an exact relation between the eigenvalues and eigenvectors of $\mathcal{K}$ and $\mathcal{L}_\sigma$. Such orderings will be called **consistent orderings**. Before defining them, however, we shall first define and show that there exists an **ordering vector**.

Theorem 5.1 An $N \times N$ matrix $(a_{i,j})$ has property (A_q) if and only if there exists a vector γ in V_N with integral coordinates such that $a_{i,j} \neq 0$ and $i \neq j$ imply $|\gamma_i - \gamma_j| = 1$, and such that γ assumes q distinct values.

Proof:

(a) If $(a_{i,j})$ has property (A_q), there exist non empty disjoint sets $T_1, T_2, \ldots, T_q$ such that $\bigcup_{\ell=1}^{q} T_\ell = \mathcal{T}$. We assume the sets T_ℓ ordered so that $i \in T_\ell$ and $i \neq j$ implies $a_{i,j} = 0$ or $j \in T_{\ell-1} \bigcup T_{\ell+1}$. (As before, T_0 and T_{q+1} are empty sets). We can define γ such that $i, j \in T_\ell$ implies $\gamma_i = \gamma_j$ and so that if $i \in T_\ell$ and $j \in T_{\ell+1} \bigcup T_{\ell-1}$, $|\gamma_i - \gamma_j| = 1$. γ will have the required properties. We note that γ is not uniquely determined.

(b) If γ exists, on the other hand, we denote by $T_1, T_2, \ldots, T_{\widetilde{q}}$ sets on which γ_i is a constant arranged in ascending order, and where $\widetilde{q} = \max_i \gamma_i - \min_i \gamma_i$. Clearly if $a_{i,j} \neq 0$, $i \in T_\ell$ and $i \neq j$ we have $|\gamma_i - \gamma_j| = 1$ and $j \in T_{\ell-1} \bigcup T_{\ell+1}$. Hence, if q is the number of distinct values of γ, q of the sets $T_1, T_2, \ldots, T_{\widetilde{q}}$ will be non empty. Thus, $(a_{i,j})$ has property (A_q).

Q. E. D.

Corollary 5.1 If $(a_{i,j})$ has property (A_q), $(q \geq 2)$, it has property $(A_{q'})$, where $2 \leq q' \leq q$.

Proof: If $(a_{i,j})$ has property (A_q) we can let

$$\gamma(T_1) = 1,\ \gamma(T_2) = 2, \ldots,\ \gamma(T_{q'}) = q'$$

$$\gamma(T_{q'+1}) = q' - 1,\ \gamma(T_{q'+2}) = q' \qquad \text{etc.}$$

getting only q' distinct values $2 \leq q' \leq q$. ($\gamma(T_\ell)$ denotes γ_i for $i \in T_\ell$). Hence new sets $T'_1, T'_2, \ldots, T'_{q'}$ are obtained as in the Theorem and $(a_{i,j})$ has property $(A_{q'})$.

Q. E. D.

Corollary 5.2 If $(a_{i,j})$ is *irreducible* and has property (A_q), then

$$q = \widetilde{q} = \max_i \gamma_i - \min_i \gamma_i \ .$$

Proof: We need only prove that if $\gamma_i = a$, $\gamma_j = a + 2$ then there exists i' such that $\gamma_{i'} = a + 1$. Let J_1 be the set of all i such that $\gamma_i \leq a$, and J_2 be the set of all i such that $\gamma_i \geq a + 2$. By the irreducibility of $(a_{i,j})$ there exists $i \in J_1$, $j \in J_2$ such that $a_{i,j} \neq 0$. But this implies $|\gamma_i - \gamma_j| = 1$, which is impossible if $J_1 \cup J_2 = \mathcal{T}$.

Q. E. D.

Definition 5.1 γ is called the **ordering vector** of $(a_{i,j})$.

Definition 5.2 An *ordering* σ of the rows and columns of a matrix $(a_{i,j})$ with property (A_q) is **consistent**, if for some ordering vector γ, $\gamma_i > \gamma_j$ implies $i > j$ under σ,[†] for all $i, j \in T$.

When we speak of a *consistent ordering*, it is implied that $(a_{i,j})$ has property (A_q) for some q. It is obvious that if a matrix $(a_{i,j})$ has property (A_q) then at least one consistent ordering of the rows and columns of $(a_{i,j})$ exists.

Corollary 5.3 If σ is a consistent ordering and if $a_{i,j} \neq 0$ and $i > j$, then $\gamma_i > \gamma_j$.

Examples of Orderings
We first define equivalent orderings.

Definition 5.3 Two orderings, σ, σ' are **equivalent** if wherever $b_{i,j} \neq 0$ and $i \neq j$

$$i > j \text{ under } \sigma \text{ if and only if } i > j \text{ under } \sigma'.$$

Corollary 5.4 If σ is equivalent to σ', σ is consistent if and only if σ' is consistent.

Clearly, if σ is consistent one may interchange the ith and the jth row of $(a_{i,j})$, where $\gamma_i = \gamma_j$ and the new ordering obtained will be consistent.

For R_k, we need only define the ordering relation for adjacent net points by (1.16e). We now define some typical orderings.

[†] "$i > j$ under σ" means that under the ordering the ith row follows the jth row of $(a_{i,j})$.

Definition 5.4

$$\text{(a)}\quad \sigma_0 \qquad x > x' \qquad \text{if} \begin{cases} p_n > p'_n \text{ or} \\ p_n = p'_n, \text{ and } p_{n-1} > p'_{n-1} \text{ or} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ p_n = p'_n, \ldots, p_2 = p'_2,\ p_1 > p'_1 \end{cases}$$

$$\text{(b)}\quad \sigma_1 \qquad x > x' \qquad \text{if } \sum_{i=1}^{n} p_i > \sum_{i=1}^{n} p'_i$$

$$\text{(c)}\quad \sigma_2 \qquad x > x' \qquad \text{if } \sum_{i=1}^{n} p_i \text{ is odd and } \sum_{i=1}^{n} p'_i \text{ is even}$$

$$\text{(d)}\quad \sigma_3 \qquad \begin{matrix} x > x' \\ (n=2) \end{matrix} \qquad \text{if} \begin{cases} p_2 \text{ is even and } p'_2 \text{ is odd, or} \\ p_2 = p'_2 \text{ and } p_1 > p_2 \end{cases}$$

$$\text{(e)}\quad \sigma_4 \qquad \begin{matrix} x > x' \\ (n=2) \end{matrix} \qquad \text{if} \begin{cases} p_2 > p'_2 \\ p_2 \text{ is odd and } p'_2 > p'_1 \\ p_2 \text{ is even and } p_1 < p'_1\ . \end{cases}$$

It is evident that σ_0 is equivalent to σ_1 . σ_0 is the most common ordering although σ_2 has certain advantages.

Theorem 5.2 The orderings $\sigma_0, \sigma_1, \sigma_2$ and σ_3 are consistent. The ordering σ_4 is not consistent for all networks.

Proof: (a) We have already noted that σ_0 is equivalent to σ_1. We define the ordering vector by

$$\gamma_i = \sum_{k=1}^{n} p_k^{(i)} \qquad (i = 1, 2, \ldots, N)\ . \tag{5.4}$$

Clearly $a_{i,j} \neq 0$ and $i \neq j$ implies $|\gamma_i - \gamma_j| = 1$. Hence by Theorem (5.1), $(a_{i,j})$ has property (A_q) where q is the number of distinct values of γ_i.

σ_1 is consistent since $\gamma_i > \gamma_j$ implies $x^{(1)} > x^{(j)}$ under σ_1.

(b) We have already seen that $(a_{i,j})$ has property (A_2) where T_1, T_2 are the sets where $\sum_{k=1}^{n} p_k^{(i)}$ is odd or even, respectively.

The ordering vector is given by

$$\gamma_i = \begin{cases} 1 & \text{if } \sum_{k=1}^{n} p_k^{(i)} \text{ is odd} \\ 0 & \text{if } \sum_{k=1}^{n} p_k^{(i)} \text{ is even}\ . \end{cases} \tag{5.5}$$

Clearly $a_{i,j} \neq 0$, $i \neq j$ implies $|\gamma_i - \gamma_j| = 1$. Also $\gamma_i > \gamma_j$ implies $x^{(1)} > x^{(j)}$ under σ_2. Hence σ_2 is consistent.

(c) For σ_3, we let the ordering vector be given by

$$\gamma_i = \begin{cases} p_i^{(i)} & \text{if } p_2^{(i)} \text{ is odd} \\ p_i^{(i)} - 1 & \text{if } p_2^{(i)} \text{ is even}\ . \end{cases} \tag{5.6}$$

Again $a_{i,j} \neq 0$ and $i \neq j$ imply $|\gamma_i - \gamma_j| = 1$, and $\gamma_i > \gamma_j$ implies $x^{(i)} > x^{(j)}$ under σ_3. Thus, σ_3 is consistent. We note that $(a_{i,j})$ has property (A_2) where q is the number of distinct values of γ.
(d) Let R_h consist of the vertices of a square

$$\begin{array}{ll} a & b \\ & \\ d & c \end{array}$$

Under σ_4 the ordering is

$$a, b, c, d.$$

If σ_4 were consistent we could define an ordering vector γ such that $\gamma(a) = 1$. By the required properties of γ we would have

$$\gamma(b) = 2 \qquad \gamma(c) = 3 \qquad \gamma(d) = 4.$$

Therefore, $|\gamma(d) - \gamma(a)| > 1$, a contradiction.

Q. E. D.

Theorem 5.3 Let $(a_{i,j})$ be an $N \times N$ matrix with property (A_q) and where the ordering of the rows and columns of $(a_{i,j})$ is consistent. Let the general element of $(a'_{i,j})$ be given by

$$a'_{i,j} = a_{i,j} \qquad (i = j \quad \text{or} \quad i < j)$$
$$a'_{i,j} = \lambda a_{i,j} \qquad (i > j) \ .$$

Then, the determinant $|(a'_{i,j})|$ equals $|(a''_{i,j})|$ where the general element of $(a''_{i,j})$ is given by

$$a''_{i,j} = a_{i,j} \qquad (i = j)$$
$$a''_{i,j} = \sqrt{\lambda}\, a_{i,j} \qquad (i \neq j) \ .$$

Proof: Each term of $|(a'_{i,j})|$ is of the form

$$t(j(i)) = \prod_{i=1}^{N} (a'_{i,j(i)})$$

where $j(i)$ is a permutation of the first N integers.

$$t(j(i)) = \prod_{\substack{i=1 \\ j(i)=i}}^{N} a_{i,j(i)} \prod_{\substack{i=1 \\ j(i)<i}}^{N} \lambda a_{i,j(i)} \prod_{\substack{i=1 \\ j(i)>i}}^{N} a_{i,j(i)} \ .$$

But since $(a_{i,j})$ and hence $(a'_{i,j})$ has property (A_q) if $a_{i,j} \neq 0$, $i < j$ implies $\gamma_{j(i)} - \gamma_i = 1$, and $i > j(i)$ implies $\gamma_i - \gamma_{j(i)} = 1$, where the ordering vector γ exists by Theorem (5.1). We have

$$t(j(i)) = \prod_{\substack{i=1 \\ j(i)=i}}^{N} a_{i,j(i)} \prod_{\substack{i=1 \\ \gamma_i > \gamma_{j(i)}}}^{N} \lambda a_{i,j(i)} \prod_{\substack{i=1 \\ \gamma_i < \gamma_{j(i)}}}^{N} a_{i,j(i)} \ .$$

Let β_1 be the number of factors with $\gamma_i > \gamma_{j(i)}$ and β_2 be the number of factors with $\gamma_i < \gamma_{j(i)}$.

$$\beta_1 = \sum_{\substack{i=1 \\ \gamma_i > \gamma_{j(i)}}}^{N} \gamma_i - \gamma_{j(i)}$$

$$\beta_2 = \sum_{\substack{i=1 \\ \gamma_i < \gamma_{j(i)}}}^{N} \gamma_{j(i)} - \gamma_i$$

$$\beta_1 - \beta_2 = \sum_{\substack{i=1 \\ \gamma_i \neq \gamma_{j(i)}}}^{N} \gamma_i - \gamma_{j(i)} = 0 \ .$$

Therefore

$$t(j(i)) = \prod_{\substack{i=1 \\ j(i)=i}}^{N} a_{i,j(i)} \prod_{\substack{i=1 \\ j(i)\neq i}}^{N} \sqrt{\lambda}\, a_{i,j(i)}$$

which is a general term of $|(a''_{i,j}|$.

Q. E. D.

Theorem 5.4 If μ is a k-fold non-zero eigenvalue of $\mathcal{K}$ and if σ is a consistent ordering, then $\lambda = \mu^2$ is a k-fold eigenvalue of $\mathcal{L}_\sigma$. If $\bar{s}$ is the number of non-zero eigenvalues of $\mathcal{K}$, then *zero* is an $(N - \bar{s}/2)$-fold eigenvalue of $\mathcal{L}_\sigma$.

Proof: By Theorem 5.0 and Theorem 5.3 the characteristic determinant of $\mathcal{L}_\sigma$ equals

$$|\mathcal{L}_\sigma - \lambda I| = |\sqrt{\lambda}\mathcal{K} - \lambda I| = \lambda^{N/2} |\mathcal{K} - \sqrt{\lambda} I|.$$

By Theorem 4.3,

$$|\mathcal{K} - \sqrt{\lambda} I| = \prod_{i=1}^{\bar{s}/2} (\lambda - \mu_i^2)(\sqrt{\lambda})^{N-\bar{s}} \ ,$$

hence

$$|\mathcal{L}_\sigma - \lambda I| = \lambda^{N-\bar{s}/2} \prod_{i=1}^{\bar{s}/2} (\lambda - \mu_i^2) \ .$$

Q. E. D.

Corollary 5.5 If σ is consistent, then the rate of convergence of $\mathcal{L}_\sigma$ is exactly twice the rate of convergence of $\mathcal{K}$.

Corollary 5.6 If $(a_{i,j})$ is symmetric and if σ is consistent, then the eigenvalues of $\mathcal{L}_\sigma$ are real and non negative.

Theorem 5.5 If σ is a consistent ordering and if $\mathcal{K}[r] = \mu r$ and if

$$\lambda = \mu^2 \tag{5.7}$$

$$v = \lambda^{\gamma/2} r \tag{5.8}$$

then

$$\mathcal{L}_\sigma[v] = \lambda v \ . \tag{5.9}$$

Proof: By Theorem 5.4, $\lambda = \mu^2$ is an eigenvalue of $\mathcal{L}_\sigma$. We have

$$\sum_{j=1}^{i-1} b_{i,j}\lambda\ v_j + \sum_{j=i+1}^{N} b_{i,j}\ v_j = \sum_{j=1}^{i-1} b_{i,j}\lambda^{(\gamma_j/2)+1} r_j + \sum_{j=i+1}^{N} b_{i,j}\lambda^{\gamma_j/2}\ r_j\ .$$

But for $b_{i,j} \neq 0$

$$\gamma_j = \begin{cases} \gamma_i + 1 & j > i \\ \gamma_i - 1 & j < i \end{cases}$$

since the ordering is consistent, by Definition 5.2, Corollary 5.3. The last expression becomes

$$\lambda^{(\gamma_i+1)/2} \sum_{\substack{j=1 \\ j\neq i}}^{N} b_{i,j} r_j = \lambda^{(\gamma_i+1)/2}\mu r_i = (\lambda^{1/2}\mu)v_i = \lambda v_i\ .$$

By Theorem 5.0, $\qquad \mathcal{L}_\sigma[v] = \lambda\ v\ .$

Q. E. D.

Theorem 5.6 If $(a_{i,j})$ is symmetric and if σ is consistent, the normal form of the matrix of $\mathcal{L}_\sigma$ is diagonal except possibly for the submatrix associated with the eigenvalue zero.

Proof: If $\lambda \neq 0$ is a k-fold root of $|\mathcal{L}_\sigma - \lambda I| = 0$, $\mu = \sqrt{\lambda}$ is a k-fold root of $|\mathcal{K} - \lambda I| = 0$ by Theorem 5.4. Since $(a_{i,j})$ is symmetric, the normal form of the matrix of $\mathcal{K}$ is diagonal by Theorem 4.1 and hence there exist k linearly independent eigenvectors of $\mathcal{K}$

$$r^{(j)} \qquad (j = 1, 2, \ldots, k).$$

The eigenvectors

$$v^{(j)} \qquad (j = 1, 2, \ldots, k)$$

of $\mathcal{L}_\sigma$ given by (5.8), are also linearly independent. Otherwise there would exist constants $\nu_1, \nu_2, \ldots, \nu_k$ not all zero such that

$$\sum_{j=1}^{k} \nu_j\ r^{(j)} = 0.$$

For all $i = 1, 2, \ldots, N$

$$\sum_{j=1}^{k} \nu_j\ v_i^{(j)} = \sum_{j=1}^{k} \nu_j\ \lambda_j^{\gamma_i/2}\ r_i^{(j)} = 0.$$

Since $\lambda_j \neq 0$ this implies

$$\sum_{j=1}^{k} \nu_j\ r_i^{(j)} = 0 \qquad (i = 1, 2, \ldots, N)$$

or

$$\sum_{j=1}^{k} \nu_j\ r^{(j)} = 0.$$

This contradicts the fact that the $r^{(j)}$ are linearly independent.

Q. E. D.

Theorem 5.7 If $(a_{i,j})$ is symmetric, the normal form of $\mathcal{L}_{\sigma_2}$ is diagonal.

Proof: By Theorems 5.4 and 5.6, we need only show that there exist $N - \bar{s}/2$ linearly independent vectors in the null space of $\mathcal{L}_{\sigma_2}$. There are N_2 linearly independent vectors in the null space of $\mathcal{L}_{\sigma_2}$ whose components are zero except for one $i \in T_2$ as may be seen easily. By Theorem 4.4, there are $N_1 - \bar{s}/2$ linearly independent vectors in the null space of $\mathcal{K}$ which vanish identically on T_2. Obviously, these vectors are also in the null space of $\mathcal{L}_{\sigma_2}$. Thus, the total number of linearly independent vectors in the null space of $\mathcal{L}_{\sigma_2}$ is

$$N_2 + (N_1 - \frac{\bar{s}}{2}) = N - \frac{\bar{s}}{2} \ .$$

Q. E. D.

That Theorem 5.7 cannot be extended to all consistent orderings can be shown easily. Thus, consider a region with four points $x^{(1)} = (1,1), \quad x^{(2)} = (1,2), \quad x^{(3)} = (2,1) \quad$ and $x^{(4)} = (2,2)$ with the ordering $\sigma_0 : x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}$. We have for the Dirichlet Problem

$$\mathcal{L}_{\sigma_0}(0,2,2,-1) = (1,0,0,0)$$

where $\mathcal{L}_{\sigma_0}(1,0,0,0) = 0$. Thus, a nilpotent vector exists. It can be easily shown [30] that the normal form of the matrix of $\mathcal{L}_{\sigma_0}$ is

$$\begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

with corresponding basis

$$\begin{array}{rrrr} (4, & 2, & 2, & 1) \\ (1, & 0, & 0, & 0) \\ (0, & 2, & 2, & -1) \\ (0, & 1, & -1, & 0). \end{array}$$

In the following,[‡] we assume $(a_{i,j})$ is symmetric. Given an arbitrary error vector

$$e^{(0)} = u^{(0)} - u \tag{5.10}$$

we can expand $e^{(0)}$ in eigenvectors and nilpotent vectors of $\mathcal{L}_\sigma$ (σ consistent), by Theorem 5.6. We have

$$e^{(0)} = \sum_{j=1}^{\bar{s}/2} B_j \ v^{(j)} + \eta \tag{5.11}$$

where

$$\mathcal{L}_\sigma[v^{(j)}] = \lambda_j \ v^{(j)}$$

[‡]I have not yet been able to prove the following conjectures which are suggested by numerical results with square regions with 3, 4 and 5 intervals on a scale.

(a) For the Dirichlet Problem and for a square region with I intervals on a side, there are $(I-t)$ nilpotent vectors of degree $t \quad (I-1 > t \geq 1)$, for the ordering σ_0.

(b) For any region, the highest degree of nilpotency does not exceed the number of distinct values of γ, for σ_0.

and where η is a suitable linear combination of nilpotent vectors. After m iterations, the error vector becomes

$$e^{(m)} = \sum_{j=1}^{\bar{s}/2} \lambda_j^m B_j \; v^{(j)} + \mathcal{L}_\sigma^m[\eta] \tag{5.12}$$

where for some $M \leq N, \quad m > M$ implies $\mathcal{L}_\sigma^m[\eta] = 0$.

If the coefficients (A_j) of (4.23) for the expansion of $e^{(0)}$ in eigenvectors of $\mathcal{K}$ are known, the B_j of (5.11) may be computed easily for the ordering σ_2. We have by (5.8 and 5.9)

$$v^{(j)} = \lambda^{\gamma/2} \; r^{(j)} \qquad (j = 1, 2, \ldots, \bar{s}/2)$$

where

$$\mathcal{K}[r^{(j)}] = \mu_j \; r^{(j)} \quad ; \; \mu_j > 0 \qquad (j = 1, 2, \ldots, \bar{s}/2)$$

and where

$$\mathcal{L}_{\sigma_2}[v^{(j)}] = \lambda_j \; v^{(j)} = \mu_j^2 \; v^{(j)} \qquad (j = 1, 2, \ldots, \bar{s}/2).$$

From (5.5), we have, since

$$\begin{aligned} \lambda_j^{\gamma/2} &= (1-\gamma) + \lambda_j^{1/2} \; \gamma \qquad (j = 1, 2, \ldots, \bar{s}/2) \\ v^{(j)} &= ((1-\gamma) + \gamma \; \lambda_j^{1/2}) \; r \qquad (j = 1, 2, \ldots, \bar{s}/2). \end{aligned} \tag{5.13}$$

If we set

$$v^{*(j)} = (1-\gamma) \; r^{(j)} \qquad (j = 1, 2, \ldots, \bar{s}/2) \tag{5.14}$$

it can be easily verified that

$$\mathcal{L}_{\sigma_2}[v^{*(j)}] = 0 \qquad (j = 1, 2, \ldots, \bar{s}/2).$$

By (4.18), we have

$$\left\{ \begin{array}{rcl} \gamma \; r^{(j)} &=& \dfrac{1}{2} \; [r^{(j)} - r^{*(j)}] \\ (1-\gamma) \; r^{(j)} &=& \dfrac{1}{2} \; [r^{(j)} + r^{*(j)}] \end{array} \right. \qquad (j = 1, 2, \ldots, \bar{s}/2) \tag{5.15}$$

therefore

$$\left\{ \begin{array}{rcl} v^{(j)} &=& \dfrac{1+\lambda_j^{1/2}}{2} \; r^{(j)} \quad + \quad \dfrac{1-\lambda_j^{1/2}}{2} \; r^{*(j)} \\ v^{*(j)} &=& \dfrac{1}{2} \; r^{(j)} \quad + \quad \dfrac{1}{2} \; r^{*(j)} \end{array} \right. \qquad (j = 1, 2, \ldots, \bar{s}/2). \tag{5.16}$$

Solving these equations for $r^{(j)}$ and $r^{*(j)}$, we get

$$\left\{ \begin{array}{l} r^{(j)} = \lambda_j^{-1/2}[v^{(j)} + (1-\lambda_j^{1/2}) v^{*(j)} = \lambda_j^{-1/2} \; v^{(j)} + \eta_1 \\ r^{*(j)} = \lambda_j^{-1/2}[-v^{(j)} + (1+\lambda_j^{1/2}) v^{*(j)} = -\lambda_j^{-1/2} \; v^{(j)} + \eta_2 \end{array} \right. \qquad (j = 1, 2, \ldots, \bar{s}/2) \tag{5.17}$$

where

$$\mathcal{L}_{\sigma_2}[\eta_1] = \mathcal{L}_{\sigma_2}[\eta_2] = 0.$$

By (4.21), an arbitrary error vector $e^{(0)}$ can be written in the form

$$e^{(0)} = \sum_{j=1}^{\bar{s}/2} A_j r^{(j)} + A_j^* r^{*(j)} + \sum_{j=\bar{s}/2+1}^{N-\bar{s}/2} A_j r^{(j)}.$$

Using (5.17) and the fact that

$$\mathcal{K}[r^{(j)}] = 0 \quad \text{implies} \quad \mathcal{L}_{\sigma_2}[r^{(j)}] = 0 \ ,$$

we have

$$e^{(0)} = \sum_{j=1}^{\bar{s}/2} \lambda_j^{-1/2} \ (A_j - A_j^*) \ v^{(j)} + \eta \tag{5.18}$$

where $\mathcal{L}_{\sigma_2}[\eta] = 0$. In Section 7, we shall use this expansion to get an upper bound for the number of iterations required to solve the finite difference analogue of the Dirichlet Problem using $\mathcal{L}_{\sigma_2}$.

For a general consistent ordering, σ, however, even if the A_j and $r^{(j)}$ of (4.20) are known, the determination of the B_j of (5.11) could only be done by solving a set of linear equations. This is true even though the eigenvectors, $v^{(j)}$ of $\mathcal{L}_\sigma$ are known explicitly in terms of the $r^{(j)}$.

§6. Other Orderings

We have seen that the rate of convergence of the Liebmann Method with a consistent ordering is exactly twice the rate of convergence of the Kormes Method. It is logical to ask whether the rate of convergence of the Liebmann Method could not be increased by using orderings which are not consistent.

Shortley and Weller [30] state that, except for small N, the rate of convergence of the Liebmann Method applied to the Dirichlet Problem is practically independent of the order in which the points are traversed. Numerical evidence, given in Table 6.1, indicates that the effect of the ordering on the rate of convergence is small, not exceeding 12% in the most extreme case treated, and decreases rapidly as N increases. Furthermore, in these cases, the consistent orderings were found to give the best convergence. It is my conjecture that the consistent orderings *always* give the best convergence. I have not found a proof of this except for one special case. If the ordering relationship between any one given point and one or more adjacent points is changed, a consistent ordering will in general be made non consistent; it is proved for the self adjoint case that in any case the rate of convergence is not increased.

Before giving the proof of this fact, we shall first give an example to show that the results of Section 5 are not valid for all orderings.

6A. An Ordering for Which the Results of Theorem 5.4 Are Not Valid

Consider the Dirichlet Problem for a square region with four interior points. If the order of improvement is $\sigma_0 \begin{bmatrix} .1 & .2 \\ .3 & .4 \end{bmatrix}$ the characteristic equation for the Liebmann Method is

$$\frac{1}{256}\,[256\lambda^4 - 64\lambda^3] = 0 \tag{6.1}$$

and the eigenvalues are $\lambda = 0, 0, 0, 1/4$.

For the Kormes Method the eigenvalues are

$$\mu = 0, 0, 1/2, -1/2 \ .$$

This agrees with Theorem 5.4. On the other hand, with the ordering $\sigma_4 \begin{bmatrix} .1 & .2 \\ .4 & .3 \end{bmatrix}$ the characteristic equation for the Liebmann Method is

$$\frac{1}{256}\,[256\lambda^4 - 65\lambda^3 + 2\lambda^2 - \lambda] = 0, \tag{6.2}$$

and the eigenvalues are $\lambda = 0, \quad -.0115 + .1185i, \quad -.0115 - .1185i, \quad .277$.

Since $.277 > 1/4$, the rate of convergence of $\mathcal{L}_{\sigma_4}$ is less than the rate of convergence of $\mathcal{L}_{\sigma_0}$ for this region.

We observe that the rate of convergence is independent of a cyclic permutation of the ordering. Hence, the essentially different orderings are

$$\sigma_0 \begin{bmatrix} .1 & .2 \\ .3 & .4 \end{bmatrix} \qquad \sigma_2 \begin{bmatrix} .1 & .3 \\ .4 & .2 \end{bmatrix} \qquad \sigma_2' \begin{bmatrix} .1 & .4 \\ .3 & .2 \end{bmatrix}$$

$$\sigma_4 \begin{bmatrix} .1 & .2 \\ .4 & .3 \end{bmatrix} \qquad \sigma_0' \begin{bmatrix} .1 & .3 \\ .2 & .4 \end{bmatrix} \qquad \sigma_4' \begin{bmatrix} .1 & .4 \\ .2 & .3 \end{bmatrix}$$

The orderings $\sigma_0, \sigma_0', \sigma_2$ and σ_2' are consistent. By symmetry, it is evident that σ_4 is equivalent to σ_4'. Thus for this region, we have shown that for no ordering can the rate of convergence of the Liebmann Method exceed that for a consistent ordering.

6B. The Effect of Changing the Relative Position of One Equation in a Consistent Ordering

We now prove a special case of the conjecture that for $(a_{i,j})$ symmetric, the rate of convergence of the Liebmann Method for any ordering does not exceed that for a consistent ordering.

Theorem 6.1 Let σ be a consistent ordering of the rows and columns of the symmetric $N \times N$ matrix $(a_{i,j})$ satisfying (0.2). If σ' is a new ordering obtained by changing the relative position of any single row and the corresponding column under σ, then the rate of convergence of $\mathcal{L}_{\sigma'}$ does not exceed the rate of convergence of $\mathcal{L}_\sigma$.

Proof: By Theorem 5.0

$$|\mathcal{L}_\sigma - \lambda I| = |(\xi_{i,j})|$$

where

$$\xi_{i,j} = \begin{cases} b_{i,j} & (j > i) \\ \lambda b_{i,j} & (j < i) \\ -\lambda & (j = i) \end{cases}$$

and where, by (1.4)

$$b_{i,j} = \begin{cases} 0 & (i = j) \\ -\dfrac{a_{i,j}}{a_{i,i}} & (i \neq j)\ . \end{cases}$$

It is easily verified that

$$|(\xi_{i,j})| = |(\xi'_{i,j})|\ ,$$

where

$$\xi'_{i,j} = \begin{cases} b'_{i,j} & (j > i) \\ \lambda b'_{i,j} & (j < i) \\ -\lambda & (j = i) \end{cases}$$

and where

$$b'_{i,j} = \begin{cases} 0 & (i = j) \\ \dfrac{-a_{i,j}}{\sqrt{a_{i,i}}\sqrt{a_{j,j}}} & (i \neq j)\ . \end{cases}$$

We note that $\xi'_{i,j}$ is symmetric.

Lemma 6.1 Let N be even and let λ_1 be the largest eigenvalue of $\mathcal{L}_\sigma$. Then all subdeterminants, obtained from the matrix $(\xi'_{i,j})$ by deleting an even number of rows and the corresponding columns are positive, when $\lambda = \lambda_1$.

Proof: We first remark that, since σ is consistent for $(b'_{i,j})$, it is also consistent for any submatrix of the above type. By Theorem 4.2, the largest eigenvalue of the Kormes Method for the submatrix is smaller than for the entire matrix. But by Theorem 5.4, the same is true for $\mathcal{L}_\sigma$. Since N is even, $|\mathcal{L}_\sigma - \lambda I| > 0$ for the submatrix and hence the matrix is positive for sufficiently large λ; therefore $|\mathcal{L}_\sigma - \lambda_1 I| > 0$ for the submatrix.

Q. E. D.

Now let the rows and columns of $|\mathcal{L}_{\sigma'} - \lambda I| = |(\widetilde{\xi}'_{i,j})|$ be arranged so that they are in the order σ'. If the i_0-th equation was originally displaced we have

$$\widetilde{\xi}'_{i,j} = \begin{cases} b'_{i,j} & (j > i) \\ \lambda b'_{i,j} & (j < i) \\ -\lambda & (j = i) \end{cases} \tag{6.3}$$

except that for some k we may have

$$\begin{cases} \begin{cases} \widetilde{\xi}'_{k,i_0} = b'_{k,i_0} \quad \text{and} \\ \widetilde{\xi}'_{i_0,k} = \lambda b'_{i_0,k} \end{cases} & (i_0 < k) \\ \text{or} & \\ \begin{cases} \widetilde{\xi}'_{k,i_0} = \lambda b'_{k,i_0} \quad \text{and} \\ \widetilde{\xi}'_{i_0,k} = b'_{i_0,k} \end{cases} & (i_0 > k) \ . \end{cases} \tag{6.4}$$

If no such k exists

$$|\mathcal{L}_{\sigma} - \lambda I| = |\mathcal{L}_{\sigma'} - \lambda I|$$

and the theorem follows at once.

By (6.3) and (6.4), the determinants $|(\xi'_{i,j})|$ and $|(\widetilde{\xi}'_{i,j})|$ can differ only in terms containing the factors $(\xi'_{i_0,k}\ \xi'_{j,i_0})$ where $i_0 \neq j, k$ and $k \neq j$.

Typical sets of such terms for $|(\xi'_{i,j})|$ are

$$\tau(j(i)) = -\xi'_{i_0,i_1}\xi'_{i_1,i_2}\cdots\xi'_{i_s,i_0}\overline{\alpha}(j(i))$$

and the complementary set

$$\tau(j^*(i)) = -\xi'_{i_1,i_0}\xi'_{i_2,i_1}\cdots\xi'_{i_0,i_s}\overline{\alpha}(j(i))$$

where $i_1, i_2, \ldots, i_s, i_0$ are distinct integers, and $\overline{\alpha}(j(i))$ is the subdeterminant obtained from $(\xi'_{i,j})$ by removing the rows and columns containing $i_0, i_1, \ldots, i_s$. Such subdeterminants are positive for $\lambda = \lambda_1$ by Lemma 6.1. Similar formulas can be given for the corresponding sets of terms $\widetilde{\tau}(j(i))$ and $\widetilde{\tau}(j^*(i))$ for $|(\widetilde{\xi}'_{i,j})|$.

We note that since $(a_{i,j})$ has property A_2, s must be odd. Moreover for each set of terms, the set of second subscripts is obtained from the set of first subscripts by an *odd* number of transpositions, hence we have a negative sign.

By Theorem 5.3,

$$\tau(j(i)) = \tau(j^*(i)).$$

Also for $\lambda = \lambda_1$, we have

$$\tau(j(i)) \leq 0.$$

On the other hand, by (6.4)

$$\widetilde{\tau}(j(i)) = \lambda^{\overline{r}}\ \tau(j(i))$$
$$\widetilde{\tau}(j^*(i)) = \lambda^{-\overline{r}}\ \tau(j(i))$$

for some integer $\overline{r}$. Therefore

$$[\widetilde{\tau}(j(i)) + \widetilde{\tau}(j^*(i)) - 2\tau(j(i))]\ \Big|_{\lambda=\lambda_1} = [(\lambda_1^{\overline{r}/2} - \lambda_1^{-\overline{r}/2})^2\ \tau(j(i))]\ \Big|_{\lambda=\lambda_1} \leq 0\ .$$

Summing over all such sets of terms, we get

$$|\mathcal{L}_{\sigma'} - \lambda_1 I| \leq |\mathcal{L}_{\sigma} - \lambda_1 I| = 0 \ .$$

Since N is even, $|\mathcal{L}_{\sigma'} - \lambda I|$ is positive for sufficiently large λ. Therefore, for some $\lambda_1' \geq \lambda_1$, we have

$$|\mathcal{L}_{\sigma'} - \lambda_1' I| = 0.$$

A similar argument can be used when N is odd.

Q. E. D.

The author has so far not been able to extend the proof to the case where more than one change is made in the ordering.

The following numerical results for the Dirichlet Problem are given to show that the effect of a change of ordering is small, in all probability. Except for Examples 1, 2 and 3, the eigenvalues for the non consistent orderings were obtained experimentally, [30]. We note that for any region with no more than three points every ordering is consistent. Hence, on the plausible assumption that the effect of different orderings decreases as N increases, it would appear that it is largest in Example 1.

Table 6.1

	Region and Ordering	λ_1	λ_1[†] (consistent ordering)
1.	1. 2. 4. 3.	.277[†]	.250
2.	1. 2. 3. 5. 4.	.307[†]	.285
3.	1. 2. 3. 6. 5. 4.	.393[†]	.363
4.	1. 2. 3. 6. 5. 4. 7. 8. 9.	$.519 \pm .003$	.500
5.	1. 2. 3. 8. 9. 4. 7. 6. 5.	$.520 \pm .001$	.500
6.	1. 2. 4. 3. 5. 6. 8. 7.	$.445 \pm .004$	.428
7.	1. 2. 3. 6. 5. 4. 7. 8. 9. 12. 11. 10.	$.587 \pm .005$	.5745
8.	1. 2. 3. 4. 8. 7. 6. 5. 9. 10. 11. 12. 16. 15. 14. 13.	$.664 \pm .007$	.655

[†]Exact values to number of decimals given.

§7. The Use of Large Automatic Computing Machines

One might suppose that, if a large automatic computing machine can be used with the Liebmann Method, the slow rate of convergence would be compensated by the speed of the machine. A more careful analysis reveals however that the number of iterations required may be so large that even with a fast machine the time is prohibitive.

For example, in [31], F. Snyder and H. Livingston propose a set of coding instructions for solving a "Laplace Boundary Value Problem" for a plane region by means of the UNIVAC,[†] and using the Liebmann Method.[‡] The authors state that problems for a square region with 300 intervals on a side, and, of course, smaller regions, are easily programmed. On the other hand, the authors make no estimate of the number of iterations required. As we shall see, for the Dirichlet Problem for a 300×300 interval square, the UNIVAC would require about 2 years to reduce the error of the original estimated solution by a factor of 1000.

For the number of iterations required to reduce $\|e^{(0)}\|$ to a definite fraction ρ of its original value, we have by Definition 3.5 and Theorem 3.3.

$$\lim_{\rho \to 0} \left[\mathcal{N}(\mathcal{L}_\sigma, \rho) \Big/ \frac{-\log \rho}{-\log |\lambda_1|} \right] = 1. \tag{7.1}$$

By Theorem 4.6 and Theorem 5.4, Corollary 5.5, we have

$$\mathcal{N}(\mathcal{L}_\sigma, \rho) = O(h^{-2}) \tag{7.2}$$

provided σ is consistent. By the considerations of Section 6, this is almost certainly true for any ordering.

For the ordering σ_2,[§] we can make a more exact statement about the required number of iterations.

Theorem 7.1 For all $\rho \quad (0 < \rho < 1)$

$$\mathcal{N}(\mathcal{L}_{\sigma_2}, \rho) \leq \frac{-\log \rho/2}{-\log \lambda_1} + 1. \tag{7.3}$$

Proof: By (5.18), we have

$$e^{(0)} = \sum_{j=1}^{\bar{s}/2} \left[\frac{A_j - A_j^*}{\lambda_j^{1/2}} \right] v^{(j)} + \eta$$

where $\mathcal{L}_{\sigma_2}[\eta] = 0$. Then for $m \geq 1$

$$e^{(m)} = \sum_{j=1}^{\bar{s}/2} \left(\frac{A_j - A_j^*}{\lambda_j^{1/2}} \right) \lambda_j^m \, v^{(j)} \, .$$

By (5.16)

$$v_i^{(j)} = \frac{1 + \lambda_j^{1/2}}{2} \, r_i^{(j)} + \frac{1 - \lambda_j^{1/2}}{2} \, r_i^{(j)}.$$

[†] When the paper [31] was written, the construction of the UNIVAC had not yet been completed.

[‡] The Liebmann Method is clearly preferable to the Kormes Method. Not only is the rate of convergence twice as large, but the storage problem is simplified.

[§] The ordering σ_0 was used for the coding for the UNIVAC given in [31]. The coding for the ordering σ_2 would be only slightly more difficult and approximately half as many storage registers would be required. This is due to the fact that the values of $\mu_i^{(m)}$ where $\sum_{k=1}^{2} p_k^{(i)}$ is odd completely determine $\mu_i^{(m+1)}, \quad (i = 1, 2, \ldots, N)$.

Hence

$$\begin{aligned} \| \mathcal{L}_{\sigma_2}^m[e^{(0)}] \|^2 &= \sum_{j=1}^{\bar{s}/2} \lambda_j^{2m} \left(\frac{A_j - A_j^*}{\lambda_j^{1/2}} \right)^2 \left(\frac{1+\lambda_j}{2} \right) \\ &\leq 2\lambda_1^{2m-1} \sum_{j=1}^{N} A_j^2 = 2\lambda_1^{2m-1} \|e^{(0)}\|^2 \end{aligned}$$

$$\| \mathcal{L}_{\sigma_2}^m[e^{(0)}] \| \leq \sqrt{2}\, \lambda_1^{m-\frac{1}{2}}\, \|e^{(0)}\| \,. \tag{7.4}$$

Therefore

$$\mathcal{N}(\mathcal{L}_{\sigma_2}, \rho) \leq \frac{-\log \rho/2}{-\log \lambda_1} + 1. \tag{7.5}$$

Q. E. D.

No matter what ordering is used, the rate of convergence of the Liebmann Method is very slow. For $n = 2$ the time per iteration, as well as the number of iterations is of the order of h^{-2}. Hence, the total time is of the order of h^{-4}.

Consider, for example, the unit square with $h^{-1} = 300$. We have by (4.30)

$$\mu_1 \sim 1 - \frac{\pi^2}{2}\, (300)^{-2}$$

$$\lambda_1 \sim 1 - \pi^2 (300)^{-2} \,.$$

Setting $\rho = 10^{-3}$, we have by (7.1)

$$\mathcal{N}(\mathcal{L}_\sigma, 10^{-3}) \sim \frac{-3 \log 10}{\pi^2 300^{-2}} = 63,000 \text{ iterations.}$$

For the UNIVAC, the estimate of the time required (in seconds) per iteration given in [31] is $I_1 I_2/100$ where I_1, I_2 are the number of intervals on the sides of a plane rectangle. In our case, 900 seconds are required per iteration.

The total time required is

$$63,300 \times 900 \text{ seconds} \sim 15,800 \text{ hours} \sim 660 \text{ days.}$$

Numerical estimates for other mesh sizes will be given in Section 9.

Obviously, for large N, a more rapidly converging method is needed. In the next chapter, it is shown that the Liebmann Method can be modified in such a way that the number of iterations required is proportional to h^{-1} instead of h^{-2}.

Chapter III

The Successive Overrelaxation Method

L. F. Richardson's method of systematic overrelaxation has already been described in the Introduction. Using a fixed relaxation factor ω, the rate of convergence cannot in general be appreciably increased. Consider the Dirichlet Problem for example. In the notation of (1.14) and (1.15), the improvement formula (0.13) becomes, (if we replace ωa_{ii} by ω),

$$u_i^{(m+1)} = \omega \sum_{j=1}^{N} b_{i,j} u_j^{(m)} + \omega c_i - (\omega - 1) u_i^{(m)} \qquad (i = 1, 2, \ldots, N) \tag{III.0.1}$$

or

$$u^{(m+1)} = \mathcal{R}_\omega[u^{(m)}] + \omega c. \tag{III.0.1a}$$

We observe that a vector $\widehat{r}$ is an eigenvector of $\mathcal{R}_\omega$ with eigenvalue $\widehat{\mu}$ if and only if

$$\widehat{\mu}\, \widehat{r}_i = \omega \sum_{j=1}^{N} b_{i,j} \widehat{r}_i - (\omega - 1)\widehat{r}_i \qquad (i = 1, 2, \ldots, N). \tag{III.0.2}$$

But if

$$\mathcal{K}[r] = \mu r$$

$$\begin{aligned} \omega \sum_{j=1}^{N} b_{i,j} r_j - (\omega - 1) r_i &= \omega \mu r_i - (\omega - 1) r_i \\ &= r_i[\omega\mu - (\omega - 1)] = r_i[1 - \omega(1 - \mu)] \qquad (i = 1, 2, \ldots, N). \end{aligned} \tag{III.0.3}$$

Therefore

$$\mathcal{R}_\omega[r] = [1 - \omega(1 - \mu)]r \tag{III.0.4}$$

and

$$[1 - \omega(1 - \mu)]$$

is an eigenvalue of $\mathcal{R}_\omega$.

By Theorem 3.1, Corollary 2, $\mathcal{R}_\omega$ will converge if and only if

$$|1 - \omega(1 - \mu)| < 1 \tag{III.0.5}$$

for all eigenvalues μ of $\mathcal{K}$.

For the Dirichlet Problem, all μ are real. Moreover, if μ_1 is the largest eigenvalue of $\mathcal{K}$, $(-\mu_1)$ is the smallest by Theorem 4.3 and the above requirement is equivalent to the condition

$$0 < \omega < \frac{2}{1 + \mu_1}. \tag{III.0.6}$$

Therefore, the rate of convergence cannot be increased by a factor of more than $2/(1 + \mu_1)$ which is very slightly greater than one, since μ_1 is in general very nearly one.

As stated previously, Richardson varied ω on each iteration. For the Dirichlet Problem, the improvement formula may be written

$$u_i^{(m+1)} = \omega_m \sum_{j=1}^{N} b_{i,j} u_j^{(m)} + \omega_m c_i - (\omega_m - 1) u_i^{(m)} \qquad (i = 1, 2, \ldots, N) \tag{III.0.7}$$

or

(III.0.7a) $$u^{(m+1)} = \mathcal{R}_{\omega_m}[u^{(m)}] + \omega_m c.$$

If

(III.0.8) $$e^{(0)} = u^{(0)} - u = \sum_{k=1}^{N} A_k r^{(k)}$$

where $\{r^{(k)}\}$ is the orthonormal set of eigenvectors of $\mathcal{K}$, then it can be easily verified that

(III.0.9) $$e^{(m)} = u^{(m)} - u = \sum_{k=1}^{N} A_k r^{(k)} \prod_{\gamma=1}^{m} (1 - \omega_\gamma(1 - \mu_k))$$

and

(III.0.10) $$\|e^{(m)}\|^2 = \sum_{k=1}^{N} A_k^2 \prod_{\gamma=1}^{m} (1 - \omega_\gamma(1 - \mu_k))^2 \ .$$

Thus, the effect of using $\omega > 2/(1 + \mu_1)$ on one iteration is compensated by using $\omega < 1$ on other iterations. By (III.0.9), we note that if an eigenvalue μ_k is known exactly the coefficient A_k of the corresponding eigenvector can be removed on one iteration by setting $\omega = 1/(1 - \mu_k)$. In fact, if all μ_k, $(k = 1, 2, \ldots, N)$, were known, the error could be reduced to zero after N iterations.

In practice however, all the μ_k are seldom known and even if they were, the number of iterations would be prohibitive by the above method. Richardson chose values of ω approximately evenly spaced in the range

$$0 < \omega < \frac{1}{1 - \mu_1}$$

in an attempt to reduce

$$\prod_{\gamma=1}^{m} (1 - \omega_\gamma(1 - \mu_k))$$

as uniformly as possible for all μ_k.

However, from numerical studies for a square region with 20 intervals on a side (where all μ_k are known), it appears doubtful that the gain in convergence rate could be made to exceed a factor of five, except by a very fortunate choice of the ω_γ.

As stated in the Introduction, one can by successively modifying the values of $u_i^{(m)}$ and using new values as soon as they are available, use just one *fixed* relaxation factor, and yet obtain a large increase in the rate of convergence. The improvement formula for the Successive Overrelaxation Method is, in the notation of (1.14) and (1.15), by (0.15)

(III.0.11) $$u_i^{(m+1)} = \omega \left[\sum_{j=1}^{i-1} b_{i,j} u_j^{(m+1)} + \sum_{j=i+1}^{N} b_{i,j} u_j^{(m)} \right] - (\omega - 1) u_i^{(m)} + \omega c_i$$

or

(III.0.11a) $$u_i^{(m+1)} = \mathcal{L}_{\sigma,\omega}[u^{(m)}] + \omega c \ .$$

§8. Eigenvalues and Eigenvectors

In order to study the rate of convergence of the Successive Overrelaxation Method, we shall analyze the eigenvalues and eigenvectors of $\mathcal{L}_{\sigma,\omega}$. We restrict ourselves to the case where $(a_{i,j})$ has property (A_q) and where the ordering σ is consistent.

Theorem 8.1

$$\mathcal{L}_{\sigma,\omega}[\widehat{v}] = \widehat{\lambda}\,\widehat{v} \tag{8.1}$$

if and only if

$$\omega\left\{\sum_{j=1}^{i-1} b_{i,j}\widehat{\lambda}\,\widehat{v}_j + \sum_{j=i+1}^{N} b_{i,j}\widehat{v}_j\right\} - (\omega-1)\widehat{v}_i = \widehat{\lambda}\,\widehat{v}_i \qquad (i=1,2,\ldots,N)\ . \tag{8.2}$$

The proof is immediate by (III.0.11).

Theorem 8.2 If $\quad \mathcal{K}[r] = \mu r$

$$\widehat{\lambda} + (\omega-1) = \omega\mu\widehat{\lambda}^{1/2} \tag{8.3}$$

$$\widehat{v} = \widehat{\lambda}^{\gamma/2} r \tag{8.4}$$

then

$$\mathcal{L}_{\sigma,\omega}[\widehat{v}] = \widehat{\lambda}\,\widehat{v}, \tag{8.5}$$

where γ is the *ordering vector*, (see Section 5).

Proof: If $\quad \widehat{v}_i = \widehat{\lambda}^{\gamma_i/2} r_i \qquad (i=1,2,\ldots,N) \quad$ and if

$$\widehat{\lambda} + (\omega-1) = \omega\mu\widehat{\lambda}^{1/2}$$

then

$$\omega\left\{\sum_{j=1}^{i-1} b_{i,j}\widehat{\lambda}\,\widehat{v}_j + \sum_{j=i+1}^{N} b_{i,j}\widehat{v}_j\right\} - (\omega-1)\widehat{v}_i$$

$$= \omega\left\{\sum_{j=1}^{i-1} b_{i,j}\widehat{\lambda}\cdot\lambda^{\gamma_i/2} r_j + \sum_{j=i+1}^{N} b_{i,j}\widehat{\lambda}^{\gamma_i/2}\widehat{v}_j\right\} - (\omega-1)\widehat{\lambda}^{\gamma_i/2} r_i\ .$$

But for $b_{i,j} = 0$, we have

$$\gamma_j = \begin{cases} \gamma_i + 1 & (j > i) \\ \gamma_i - 1 & (j < i) \end{cases}$$

Hence, we have for the last expression

$$\omega\widehat{\lambda}^{(\gamma_i+1)/2}\left\{\sum_{\substack{j=1\\ j\neq i}}^{N} b_{i,j} r_j\right\} - (\omega-1)\widehat{\lambda}^{\gamma_i/2} r_i = [\omega\mu\widehat{\lambda}^{1/2} - (\omega-1)]\widehat{\lambda}^{\gamma_i/2} r_i\ ,\ \text{by (4.3a), and}$$

$$= \widehat{\lambda}\,\widehat{v}_i \qquad (i=1,2,\ldots,N)\ ,$$

by (8.3). Hence, $\widehat{v}$ is an eigenvector and $\widehat{\lambda}$ is an eigenvalue of $\mathcal{L}_{\sigma,\omega}$, by Theorem 8.1.

Q. E. D.

Theorem 8.3 If $\mathcal{K}[r] = \mu r$ and

$$\begin{aligned} \mu^2\omega^2 - 4(\omega-1) &= 0 \\ \mathcal{L}_{\sigma,\omega}[\widehat{v}] &= \widehat{\lambda}\,\widehat{v} \\ \widehat{\lambda} + (\omega-1) &= \omega\mu\widehat{\lambda}^{1/2} \\ \widehat{v} &= \widehat{\lambda}^{\gamma/2} r \end{aligned} \tag{8.6}$$

$$\widehat{v}' = \frac{d}{d\widehat{\lambda}}\widehat{v} = (\frac{\gamma}{2} - 1)\widehat{\lambda}^{(\gamma/2)-1} r \tag{8.7}$$

then

$$\mathcal{L}_{\sigma,\omega}[\widehat{v}'] = \widehat{\lambda}\,\widehat{v}' + \widehat{v}. \tag{8.8}$$

We call $\widehat{v}'$ an *invariant vector* of $\mathcal{L}_{\sigma,\omega}$.

Proof:
Lemma 8.1 If $\mathcal{L}_{\sigma,\omega}[\widehat{v}] = \widehat{\lambda}\,\widehat{v}$ then

$$\mathcal{L}_{\sigma,\omega}[\widehat{v}'] = \widehat{\lambda}\,\widehat{v}' + \widehat{v}$$

if and only if

$$\widehat{\lambda}\,\widehat{v}_i' + \widehat{v}_i = \omega\left\{\sum_{j=1}^{i-1} b_{i,j}(\widehat{\lambda}\,\widehat{v}_j' + \widehat{v}_j) + \sum_{j=i+1}^{N} b_{i,j}\widehat{v}_j'\right\} - (\omega-1)\widehat{v}_i', \qquad (i = 1, 2, \ldots, N)\ . \tag{8.9}$$

The proof is immediate from (III.0.11).

We have for all $i = 1, 2, \ldots, N$

$$\omega\left\{\sum_{j=1}^{i-1} b_{i,j}(\widehat{\lambda}\,\widehat{v}_j' + \widehat{v}_j) + \sum_{j=i+1}^{N} b_{i,j}\widehat{v}_j'\right\} - (\omega-1)\widehat{v}_i'$$

$$= \omega\left\{\sum_{j=1}^{i-1} b_{i,j}\widehat{\lambda}^{\gamma_j/2}\left(\frac{\gamma_j}{2}\right) r_j + \sum_{j=i+1}^{N} b_{i,j}\left(\frac{\gamma_j}{2} - 1\right)\widehat{\lambda}^{(\gamma_j/2)-1} r_j\right\} - (\omega-1)\left(\frac{\gamma_i}{2} - 1\right)\widehat{\lambda}^{(\gamma_i/2)-1} r_i\ .$$

Since

$$\gamma_j = \begin{cases} \gamma_i + 1 & j > i \\ \gamma_i - 1 & j < i \end{cases}$$

the last expression becomes

$$\omega\widehat{\lambda}^{(\gamma_i/2)-1}\left(\frac{\gamma_i - 1}{2}\right)\sum_{j=1}^{N} b_{i,j} r_j - (\omega-1)\widehat{\lambda}^{(\gamma_i/2)-1}\left(\frac{\gamma_i}{2} - 1\right) r_i\ .$$

By (4.3a), we have

$$\widehat{\lambda}^{(\gamma_i/2)-1}\left\{\omega\left(\frac{\gamma_i - 1}{2}\right)\mu\ \widehat{\lambda}^{1/2} - (\omega-1)\left(\frac{\gamma_i}{2} - 1\right)\right\} r_i\ .$$

Setting $\omega\mu\widehat{\lambda}^{1/2} = (\omega-1) + \widehat{\lambda}$ and $\mu^2\omega^2 - 4(\omega-1) = 0$ we get

$$\widehat{\lambda} = (\omega-1) \qquad \text{and} \qquad \mu\omega = 2\widehat{\lambda}^{1/2}.$$

Hence, we have

$$\left\{2\widehat{\lambda}^{(\gamma_i/2)-1} - \widehat{\lambda}\left(\frac{\gamma_i}{2}-1\right)\right\}\widehat{\lambda}^{(\gamma_i/2)-1}r_i = \left\{\widehat{\lambda}\left(\frac{\gamma_i}{2}-1\right)+\widehat{\lambda}\right\}\widehat{\lambda}^{(\gamma_i/2)-1}r_i$$
$$= \widehat{\lambda}\,\widehat{v}_i' + \widehat{v}_i' \qquad (i=1,2,\ldots,N)\ .$$

The theorem follows by Lemma 8.1.

Q. E. D.

Corollary 8.1 If $(a_{i,j})$ is symmetric and if μ is a k-fold root of $|\mathcal{K}-\mu I|=0$ and satisfies (8.6), then there exist k linearly independent eigenvectors and k linearly independent invariant vectors of $\mathcal{L}_{\sigma,\omega}$ associated with the eigenvalue $\widehat{\lambda}$ given by (8.3).

Theorem 8.4 If μ_i , $i=1,2,\ldots,N$, are the eigenvalues of $\mathcal{K}$ and if $\omega>1$, then all the eigenvalues of $\mathcal{L}_{\sigma,\omega}$ are given by

(8.10) $$\widehat{\lambda}+(\omega-1)=\omega\mu_i\widehat{\lambda}^{1/2}$$

Multiplicities of the eigenvalues are preserved.

Proof: Let $|\mathcal{L}_{\sigma,\omega}-\widehat{\lambda}I| = |\widehat{\xi}_{i,j}|$, where by Theorem 8.1,

$$\widehat{\xi}_{i,j}' = \begin{cases} -(\omega-1)-\widehat{\lambda} & i=j \\ \omega b_{i,j} & i<j \\ \widehat{\lambda}\omega b_{i,j} & i>j \end{cases}$$

By Theorem 5.3, $|(\widehat{\xi}_{i,j}')| = |(\widehat{\xi}_{i,j}'')|$, where

$$\widehat{\xi}_{i,j}'' = \begin{cases} -(\omega-1)-\widehat{\lambda} & (i=j) \quad (i=1,2,\ldots,N) \\ \widehat{\lambda}^{1/2}b_{i,j} & (i\neq j) \quad (i=1,2,\ldots,N) \end{cases}$$

Therefore

$$|\mathcal{L}_{\sigma,\omega}-\widehat{\lambda}I| = \omega^N\widehat{\lambda}^{N/2}\left|\mathcal{K}-\frac{(\omega-1)+\widehat{\lambda}}{\omega\widehat{\lambda}^{1/2}}\,I\right|$$
$$= \omega^N\widehat{\lambda}^{N/2}\prod_{i=1}^{N}\left(\mu_i-\frac{(\omega-1)+\widehat{\lambda}}{\omega\widehat{\lambda}^{1/2}}\right)$$
$$= \prod_{i=1}^{N}(\omega\mu_i\widehat{\lambda}^{1/2}-[(\omega-1)+\widehat{\lambda}]).$$

Therefore, $|\mathcal{L}_{\sigma,\omega}-\widehat{\lambda}I|$ vanishes if and only if

$$\omega\mu_i\widehat{\lambda}^{1/2} = (\omega-1)+\widehat{\lambda}\ ,$$

with multiplicities preserved.

Q. E. D.

Theorem 8.5 For $\omega > 1$, if $(a_{i,j})$ is symmetric, the normal form of the matrix of $\mathcal{L}_{\sigma,\omega}$ is diagonal except for the submatrix associated with $\widehat{\lambda}(\mu)$ defined by (8.3) where $\mu^2\omega^2 - 4(\omega - 1) = 0$. In the latter case, if μ is a k-fold root of $|\mathcal{K} - \mu I| = 0$ in the normal form, the associated submatrix has $2k$ diagonal elements equal to $\widehat{\lambda}(\mu)$ and has alternating zeros and ones in the diagonal immediately below the main diagonal.

Proof: Solving (8.3) for $\widehat{\lambda}^{1/2}$, we get

$$\widehat{\lambda}(\mu_i)^{1/2} = \frac{\omega\mu_i \pm \sqrt{\omega^2\mu_i^2 - 4(\omega - 1)}}{2} \; . \tag{8.11}$$

If $\omega > 1$ and if $\omega^2\mu_i^2 - 4(\omega - 1) \neq 0$, we get two distinct solutions

$$\widehat{\lambda}'(\mu_i)^{1/2} = \frac{\omega\mu_i + \sqrt{\omega^2\mu_i^2 - 4(\omega - 1)}}{2} \tag{8.12}$$

$$\widehat{\lambda}''(\mu_i)^{1/2} = \frac{\omega\mu_i - \sqrt{\omega^2\mu_i^2 - 4(\omega - 1)}}{2} \; . \tag{8.13}$$

By Theorem (4.3), the non-zero eigenvalues of $\mathcal{K}$ occur in pairs. We note that

$$\widehat{\lambda}'(\mu_i)^{1/2} = -\widehat{\lambda}''(-\mu_i)^{1/2} \tag{8.14}$$

$$\widehat{\lambda}''(\mu_i)^{1/2} = -\widehat{\lambda}'(-\mu_i)^{1/2} \; . \tag{8.15}$$

There are three cases to consider

(a) If $\omega^2\mu_i^2 - 4(\omega - 1) > 0$:

We get two distinct, real positive values $\widehat{\lambda}'(\mu_i)$ and $\widehat{\lambda}''(\mu_i)$ and two distinct eigenvectors defined by (8.4). The pair of eigenvectors associated with μ_i are clearly identical with the pair associated with those associated with $-\mu_i$ by (8.4), (4.16) and (8.11). Thus, for each pair $(\mu_i, -\mu_i)$ satisfying (4.3), we get two linearly independent eigenvectors of $\mathcal{L}_{\sigma,\omega}$. Of course, if $\omega > 1, \mu_i \neq 0$ since $\omega^2\mu_i^2 > 4(\omega - 1)$.

(b) If $\omega^2\mu_i^2 - 4(\omega - 1) < 0$:

In this case, we get two distinct complex eigenvalues $\widehat{\lambda}'$ and $\widehat{\lambda}''$ unless $\mu_i = 0$. If $\mu_i = 0$ is a k-fold eigenvalue of $\mathcal{K}$ we get by (8.4), k linearly independent eigenvectors of $\mathcal{L}_{\sigma,\omega}$ with eigenvalue $\widehat{\lambda} = -(\omega - 1)$. If $\mu_i \neq 0$

$$\widehat{\lambda}'(\mu_i)^{1/2} = \frac{\omega\mu_i + i\zeta_i}{2}, \qquad \widehat{\lambda}''(\mu_i)^{1/2} = \frac{\omega\mu_i - i\zeta_i}{2}$$

where $\zeta_i^2 = 4(\omega - 1) - \omega^2\mu_i^2 > 0$.

As in (a), we obtain exactly two distinct complex eigenvectors of $\mathcal{L}_{\sigma,\omega}$ for each pair of eigenvalues $(\mu_i, -\mu_i)$ of $\mathcal{K}$. If μ_i is a k-fold eigenvalue, we obtain k such pairs.

(c) If $\omega^2\mu_i^2 - 4(\omega - 1) = 0$:

We obtain in this case invariant vectors by Theorem 8.3. If μ_i is repeated k times, there are k linearly independent eigenvalues associated with $\widehat{\lambda}(\mu_i)$ and k invariant vectors by Corollary 8.1 of Theorem 8.3. Q. E. D.

Let us replace $\widehat{\lambda}^{1/2}$ of equation (8.3) by z and μ by w. We obtain the Joukowski transformation

$$z + (\omega - 1)z^{-1} = \omega\ w\ . \tag{8.16}$$

Solving for z, we have

$$z = z_\omega(w) = \frac{\omega\ w \pm \sqrt{\omega^2 w^2 - 4(\omega - 1)}}{2} \tag{8.17}$$

Let

$$\begin{cases} z'_\omega(w) = \dfrac{\omega\ w + \sqrt{\omega^2 w^2 - 4(\omega - 1)}}{2} \\ z''_\omega(w) = \dfrac{\omega\ w - \sqrt{\omega^2 w^2 - 4(\omega - 1)}}{2} \end{cases} \tag{8.17a}$$

$$a_\omega(w) = \max(\ |z'_\omega(w)|, |z''_\omega(w)|\ ). \tag{8.17b}$$

We note that $a_\omega(w) = a_\omega(-w)$.

Theorem 8.6 If A is a positive real constant and if

$$\omega_b = 1 + \left[\frac{1 - \sqrt{1 - A^2}}{A}\ \right]^2 \tag{8.18}$$

then

$$|z_{\omega_b}(w)|^2 = |\omega_b - 1| \qquad \text{for all real } w, (|w| \le A) \tag{8.18a}$$

and

$$\xi_A(\omega) = \operatorname*{lub}_{\substack{-A \le w \le A \\ w \text{ real}}} |a_\omega(w)|^2 > |\omega_b - 1| \tag{8.18b}$$

if $\omega \neq \omega_b$.

Proof: We assume throughout that w is real.

Lemma 8.2 If $$w^2 < \frac{4(\omega - 1)}{\omega^2}$$
$z_\omega(w)$ is complex and $|z_\omega(w)|^2 = |\omega - 1|$.

Proof: The proof follows from (8.17). Q. E. D.

Lemma 8.3 If $$w_1^2 > w_2^2 > \frac{4(\omega - 1)}{\omega^2}$$
then

$$a_\omega(w_1) > a_\omega(w_2)\ .$$

Proof: By (8.17a) and (8.17b), we may assume w_1, w_2 positive, if we take the sign of the radical as plus.

$$\begin{aligned} a_\omega(w_1) - a_\omega(w_2) &= \frac{\omega(w_1 - w_2) + \sqrt{\omega^2 w_1^2 - 4(\omega - 1)} - \sqrt{\omega^2 w_2^2 - 4(\omega - 1)}}{2} \\ &= \frac{\omega}{2}(w_1 - w_2) + \frac{\omega^2(w_1^2 - w_2^2)}{\sqrt{\omega^2 w_1^2 - 4(\omega - 1)} + \sqrt{\omega^2 w_2^2 - 4(\omega - 1)}} \\ &> 0\ . \end{aligned}$$

Q. E. D.

Lemma 8.4 If $2 > \omega_1 > \omega_2 > 0$, then

$$\frac{4(\omega_1 - 1)}{{\omega_1}^2} > \frac{4(\omega_2 - 1)}{{\omega_2}^2} \ .$$

Proof:

$$\frac{4(\omega_1 - 1)}{{\omega_1}^2} - \frac{4(\omega_2 - 1)}{{\omega_2}^2}$$

$$= \frac{4}{{\omega_1}^2{\omega_2}^2}(\omega_1 - \omega_2)\left[\frac{\omega_1}{2}(2 - \omega_2) + \frac{\omega_2}{2}(2 - \omega_1)\right] > 0$$

Q. E. D.

Lemma 8.5 If $2 > \omega_1 > \omega_2 > 0$ and if $w^2 \geq \frac{4(\omega_1 - 1)}{{\omega_1}^2}$, then

$$a_{\omega_1}(w) < a_{\omega_2}(w) \ .$$

Proof: By Lemma 8.4, $w^2 > \frac{4(\omega_2 - 1)}{{\omega_2}^2}$. Now

$$\begin{aligned}
a_{\omega_1}(w) - a_{\omega_2}(w) &= \frac{\omega_1 - \omega_2}{2}w + \frac{\sqrt{{\omega_1}^2w^2 - 4(\omega_1 - 1)} - \sqrt{{\omega_2}^2w^2 - 4(\omega_2 - 1)}}{2} \\
&= \frac{(\omega_1 - \omega_2)}{2}w + \frac{w^2({\omega_1}^2 - {\omega_2}^2) + 4(\omega_2 - \omega_1)}{\sum_{i=1}^{2}\sqrt{\omega_i^2w^2 - 4(\omega_i - 1)}} \\
&= \frac{(\omega_1 - \omega_2)}{2}w + \frac{(\omega_1 - \omega_2)\{w^2(\omega_1 + \omega_2) - 4\}}{\sum_{i=1}^{2}\sqrt{\omega_i^2w^2 - 4(\omega_i - 1)}} \\
&= \frac{(\omega_1 - \omega_2)}{2}\left[\frac{\sum_{i=1}^{2}\sqrt{\omega_i^2w^4 - 4w^2(\omega_i - 1)} + (w^2\omega_i - 2)}{\sum_{i=1}^{2}\sqrt{\omega_i^2w^2 - 4(\omega_i - 1)}}\right] .
\end{aligned}$$

But $\left\{\sqrt{\omega_i^2w^4 - 4w^2(\omega_i - 1)}\right\}^2 = (2 - \omega_iw^2)^2 + 4(w^2 - 1) \qquad (i = 1, 2).$

Since $2 - \omega_iw^2 > 0$, we have

$$\sqrt{\omega_i^2w^4 - 4w^2(\omega_i - 1)} - (2 - \omega_iw^2) < 0 \qquad (i = 1, 2)$$

and $a_{\omega_1}(w) - a_{\omega_2}(w) < 0$. This proves the lemma.

Q. E. D.

The theorem now follows from the lemmas.

Q. E. D.

Theorem 8.7 If $(a_{i,j})$ is symmetric, and if the relaxation factor $\omega = \omega_b$ is used, where

$$\omega_b = 1 + \left[\frac{1 - \sqrt{1 - \mu_1^2}}{\mu_1}\right]^2 \tag{8.19}$$

then the rate of convergence of $\mathcal{L}_{\sigma,\omega}$ is maximized. The rate of convergence of $\mathcal{L}_{\sigma,\omega_b}$ is

$$\phi(\mathcal{L}_{\sigma,\omega_b}) = -\log(\omega_b - 1) \ . \tag{8.20}$$

Moreover, for all i,

$$|\widehat{\lambda}(\mu_i)| = \omega_b - 1 \ . \tag{8.20a}$$

Proof: The proof follows from Theorem 3.1, Corollary 3.1, and Theorem 8.6, and the fact that because $(a_{i,j})$ is symmetric the eigenvalues of $\mathcal{K}$ are real.

Corollary 8.2 $\quad 1 \leq \omega_b < 2 \ .$

Proof: Since $0 \leq \mu_1 < 1$, we can let

$$\mu_1 = \cos\theta \ , \text{ where } \quad 0 < \theta \leq \frac{\pi}{2}$$

$$\omega_b = 1 + \left[\frac{1 - \sin\theta}{\cos\theta}\right]^2$$

Then $$\frac{1 - \sin\theta}{\cos\theta} = \tan(\frac{\pi}{4} - \frac{\theta}{2}) \ .$$

Therefore $$0 \leq \frac{1 - \sin\theta}{\cos\theta} < 1 \ .$$

Q. E. D.

Theorem 8.8 If $(a_{i,j})$ is symmetric, with the value ω_b given by (8.19), then the normal form of $\mathcal{L}_{\sigma,\omega_b}$ contains precisely one non diagonal element associated with the repeated eigenvalue $(\omega_b - 1)$. The submatrix is of the form

$$\begin{pmatrix} \omega_b - 1 & 0 \\ 1 & \omega_b - 1 \end{pmatrix} \ . \tag{8.21}$$

Proof: The proof follows from Theorem 8.4, Theorem 8.5 and Theorem 4.1.

Q. E. D.

Theorem 8.9 If $(a_{i,j})$ is symmetric, then for all $\omega > \omega_b$

$$|\widehat{\lambda}(\mu_i)| = (\omega - 1) \qquad (i = 1, 2, \ldots, N)$$

for all i, and the normal form of the matrix of $\mathcal{L}_{\sigma,\omega}$ is diagonal.

Proof: The proof follows from Theorem 8.5 and Lemma 8.2 of Theorem 8.6.

Q. E. D.

Thus, for $(a_{i,j})$ symmetric, the operator $\mathcal{L}_{\sigma,\omega}$ has been analyzed completely. Even if $(a_{i,j})$ is not symmetric we have shown, Theorem 8.4, how the eigenvalues of $\mathcal{L}_{\sigma,\omega}$ can be determined from the eigenvalues of $\mathcal{K}$ when these are known.

We shall now show that for properly chosen ω the rate of convergence of $\mathcal{L}_{\sigma,\omega}$ is of the order of the square root of the rate of convergence of the Liebmann Method, for $(a_{i,j})$ symmetric.

§9. The Superiority Over the Liebmann Method

A. Rates of Convergence

If N is large, μ_1 is very nearly one. We have

Theorem 9.1

$$\lim_{\mu_1 \to 1} \frac{\phi(\mathcal{L}_{\sigma,\omega_b})}{\phi(\mathcal{L}_\sigma)} \sqrt{1-\mu_1} = \sqrt{2} \ , \tag{9.1}$$

where ϕ is the rate of convergence defined in Definition 3.2, and where ω_b is given by (8.19).

Proof: By (8.20a), using ω_b the optimum value of ω, the dominant eigenvalue of $\mathcal{L}_{\sigma,\omega_b}$ is given by

$$|\widehat{\lambda}(\mu_1)| = \omega_b - 1 = \left[\frac{1-\sqrt{1-\mu_1^2}}{\mu_1} \right]^2 \ . \tag{9.2}$$

But

$$\begin{aligned} -\log(1-\sqrt{1-\mu_1{}^2}) &= \sqrt{2(1-\mu_1)} + (1-\mu_1) &&+ O_1((1-\mu_1)^{3/2}) \\ -\log \mu_1 &= (1-\mu_1) &&+ O_2((1-\mu_1)^{3/2}) \end{aligned}$$

$$-\log \left[\frac{1-\sqrt{1-\mu_1^2}}{\mu_1} \right] = \sqrt{2(1-\mu_1)} \ + \ O((1-\mu_1)^{3/2}) \ .$$

Therefore, by (3.3) and (5.7),

$$\frac{\phi(\mathcal{L}_{\sigma,\omega_b})}{\phi(\mathcal{L}_\sigma)} = \frac{-\log|\widehat{\lambda}(\mu_1)|}{-2\log\mu_1} = \frac{\sqrt{2}}{\sqrt{1-\mu_1}} + O((1-\mu_1)^{1/2})$$

and the theorem follows.

Q. E. D.

Corollary 9.1

$$\phi(\mathcal{L}_{\sigma,\omega_b}) \sim 2 \ [\phi(\mathcal{L}_\sigma)]^{1/2} \ . \tag{9.3}$$

Proof: This follows since

$$\phi(\mathcal{L}_\sigma) = -2\log\mu_1 = -2(1-\mu_1) - 2 \ O_2((1-\mu_1)^{3/2}) \ .$$

Q. E. D.

B. Number of Iterations Necessary to Reduce the Error by a Specified Amount

If the best overrelaxation factor, ω_b is used, the number of iterations necessary to reduce the error to a specified fraction of itself is, by Theorem 3.3, approximately for small ρ

$$\mathcal{N}(\mathcal{L}_{\sigma,\omega_b}, \rho) \sim \frac{-\log\rho}{-\log(\omega_b - 1)} \ . \tag{9.4}$$

For μ_1 nearly equal to one, we have

$$-\log(\omega_b - 1) \sim 2\sqrt{2}\sqrt{1-\mu_1} \ .$$

Therefore

$$\mathcal{N}(\mathcal{L}_{\sigma,\omega_b},\rho) \sim \frac{-\log\rho}{2\sqrt{2}\sqrt{1-\mu_1}} \tag{9.5a}$$

and for the Liebmann Method

$$\mathcal{N}(\mathcal{L}_{\sigma},\rho) \sim \frac{-\log\rho}{2(1-\mu_1)} \ . \tag{9.5b}$$

Hence, as already stated in the Introduction, if the required number of iterations is of the order of h^{-k} by the Liebmann Method, that number is of the order of $h^{-k/2}$ by the Successive Overrelaxation Method. For the Dirichlet Problem, by (7.2), $k = 2$.

For the Dirichlet Problem and for the ordering σ_2, we can derive an *upper bound* for $\mathcal{N}(\mathcal{L}_{\sigma_2,\omega},\rho)$.

Theorem 9.2 For the Dirichlet Problem, if $|\widehat{\lambda}_1| = \omega - 1$, where

$$w \geq \omega_b = 1 + \left[\frac{1-\sqrt{1-\mu_1^2}}{\mu_1}\right]^2 \ ,$$

then $\mathcal{N}(\mathcal{L}_{\sigma_2,\omega},\rho)$ is not greater than the largest solution m of

$$m|\widehat{\lambda}_1|^{m-1} = \rho/5 \ . \tag{9.6}$$

Proof: By (4.21) and (4.24), we have for any $f \in V_N$

$$f = \sum_{j=1}^{\bar{s}/2} A_j r^{(j)} + A_j^* r^{*(j)} + \sum_{\bar{s}/2+1}^{N-\bar{s}/2} A_j r^{(j)}$$

and

$$\|f\|^2 = \sum_{j=1}^{N} A_j^2 \ .$$

By (8.4), since $\widehat{\lambda}_j^{\gamma/2} = (1-\gamma) + \widehat{\lambda}_j^{1/2}\gamma$, we have

$$\begin{cases} \widehat{v}^{(j)} &= [\ (1-\gamma) + \widehat{\lambda}_j^{1/2}\gamma\ ]\ r^{(j)} \\ \widehat{v}^{*(j)} &= [\ (1-\gamma) + \widehat{\lambda}_j^{*1/2}\gamma\ ]\ r^{(j)} \end{cases} \tag{9.7}$$

where

$$\begin{cases} \widehat{\lambda}_j^{1/2} &= \dfrac{\omega\mu_j + \sqrt{\omega^2\mu_j^2 - 4(\omega-1)}}{2} \\ \widehat{\lambda}_j^{*1/2} &= \dfrac{\omega\mu_j - \sqrt{\omega^2\mu_j^2 - 4(\omega-1)}}{2} \end{cases} \tag{9.7a}$$

and

$$\begin{cases} \mathcal{L}_{\sigma_2,\omega}[\widehat{v}^{(j)}] &= \widehat{\lambda}_j \widehat{v}^{(j)} \\ \mathcal{L}_{\sigma_2,\omega}[\widehat{v}^{*(j)}] &= \widehat{\lambda}_j^* \widehat{v}^{*(j)} \ . \end{cases} \tag{9.7b}$$

Since

$$r^{*(j)} = (1-2\gamma) r^{(j)} \qquad (j = 1, 2, \ldots, \bar{s}/2) \tag{9.8}$$

we have

$$\text{(9.9)}\qquad \begin{cases} \widehat{v}^{(j)} &= \dfrac{1+a_j}{2}\, r^{(j)} + \dfrac{1-a_j}{2}\, r^{*(j)} \\ \widehat{v}^{*(j)} &= \dfrac{1+a_j^*}{2}\, r^{(j)} + \dfrac{1-a_j^*}{2}\, r^{*(j)} \end{cases} \qquad (j = 1, 2, \ldots, \bar{s}/2)$$

where, for convenience, we let

$$\text{(9.10)}\qquad \begin{aligned} a_j &= \widehat{\lambda}_j^{1/2} \\ a_j^* &= \widehat{\lambda}_j^{*1/2} \end{aligned} \qquad (j = 1, 2, \ldots, \bar{s}/2)\ .$$

Conversely, we have

$$\text{(9.11)}\qquad \begin{cases} r^{(j)} &= \dfrac{1-a_j^*}{a_j - a_j^*}\, v^{(j)} + \dfrac{1+a_j}{a_j - a_j^*}\, v^{*(j)} \\ r^{*(j)} &= \dfrac{-1-a_j^*}{a_j - a_j^*}\, v^{(j)} + \dfrac{1+a_j}{a_j - a_j^*}\, v^{*(j)} \end{cases} \qquad (j = 1, 2, \ldots, \bar{s}/2)\ .$$

Moreover, if $\dfrac{\bar{s}}{2} < j \le N - \dfrac{\bar{s}}{2}$, $\quad r^{(j)} = r^{*(j)} \quad$ and $\quad \widehat{v}^{(j)}$ and $\widehat{v}^{*(j)} \quad$ are linearly independent.

For $\qquad N - \dfrac{\bar{s}}{2} \ge j > \dfrac{\bar{s}}{2} \qquad \mathcal{K}[r^{(j)}] = 0, \quad$ and $\qquad \mathcal{L}_{\sigma_2,\omega}[r^{(j)}] = -(\omega - 1) r^{(j)}\ .$

Thus, we may let

$$\widehat{v}^{(j)} = \widehat{v}^{*(j)} = r^{(j)} \qquad (j = \frac{\bar{s}}{2} + 1, \ldots, N - \frac{\bar{s}}{2}).$$

We have

$$\text{(9.12)}\qquad \begin{aligned} f = \sum_{j=1}^{\bar{s}/2} \frac{1}{a_j - {a_j}^*} \Big[\Big\{ A_j(1 - {a_j}^*) + A_j^*(-1 - {a_j}^*) \Big\} \widehat{v}^{(j)} \\ + \Big\{ A_j(-1 + a_j) + A_j^*(1 + a_j) \Big\} \widehat{v}^{*(j)} \Big] + \sum_{j=(\bar{s}/2)+1}^{N-\bar{s}/2} A_j \widehat{v}^{(j)} \end{aligned}$$

and

$$\begin{aligned} \mathcal{L}_{\sigma_2,\omega}^m[f] = \sum_{j=1}^{\bar{s}/2} \Bigg[\frac{{a_j}^{2m} - {a_j}^{*2m}}{2(a_j - {a_j}^*)} \Big\{ A_j^*(-1 - {a_j}^*)(1 + a_j) + A_j(1 - a_j {a_j}^*) \Big\} \\ + \frac{{a_j}^{2m} + {a_j}^{*2m}}{2} A_j \Bigg] r^{(j)} + \Bigg[\frac{{a_j}^{2m} - {a_j}^{*2m}}{2(a_j - {a_j}^*)} \Big\{ A_j(1 - {a_j}^*)(1 - a_j) + A_j^*(-1 + a_j {a_j}^*) \Big\} \\ + \frac{{a_j}^{2m} + {a_j}^{*2m}}{2} {A_j}^* \Bigg] r^{*(j)} + \sum_{j=(\bar{s}/2)+1}^{N-\bar{s}/2} {a_j}^{2m} A_j r^{(j)} \\ = \sum_{j=1}^{\bar{s}/2} {A_j}^{(m)} r^{(j)} + {A_j}^{*(m)} r^{*(j)} + \sum_{j=(\bar{s}/2)+1}^{N-\bar{s}/2} {a_j}^{2m} A_j^{(m)} r^{(j)} \end{aligned}$$

By (8.11) and (9.10), either a_j and ${a_j}^*$ are each real and positive or else they are complex conjugates, with positive real parts. Moreover, $|a_j| < 1$ and $|{a_j}^*| < 1$. Thus, we have

$$\begin{cases} \left| \dfrac{{a_j}^{2m} - {a_j}^{*2m}}{a_j - {a_j}^*} \right| \leq 2m|\bar{a}_j|^{2m-1} \\ |1 - a_j {a_j}^*| \leq 1 \\ |(1-a_j)(1-{a_j}^*)| \leq 2 \\ |(1+{a_j}^*)(1+a_j)| \leq 4 \end{cases}$$

where

$$|\bar{a}_j| = \max(|a_j|, |{a_j}^*|).$$

Therefore,

$$\text{(9.12a)} \qquad \begin{cases} |{A_j}^{(m)}| & \leq & m|\bar{a}_j|^{2m-1}(4|{A_j}^*| + |A_j|) + |\bar{a}_j|^{2m}|A_j| \\ |{A_j}^{*(m)}| & \leq & m|\bar{a}_j|^{2m-1}(2|A_j| + |{A_j}^*|) + |\bar{a}_j|^{2m}|{A_j}^*| \ . \end{cases}$$

By the same method, it can be shown that the above inequality holds even if $a_j = {a_j}^*$ which is possible for $\omega = \omega_b$. In this case, $a_1 = a_1^*$ and $\bar{v}^{*(1)}$ is replaced by the invariant vector (see Theorem 8.3)

$$v'^{(1)} = \lim_{a_1^* \to a_1} \frac{\widehat{v}^{(1)} - \widehat{v}^{*(1)}}{a_1 - a_1^*} = \frac{1}{2} a_1^{-1} \gamma r^{(1)}$$

where

$$v^{(1)} = ((1-\gamma) - a_1\gamma) r^{(1)} \ .$$

Thus, the inequality (9.12a) is valid for all $\omega \geq \omega_b$.
We now have

$$\begin{aligned} \| \mathcal{L}^m_{\sigma_2,\omega}[f] \ \|^2 \ \leq \ & \sum_{j=1}^{\bar{s}/2} \Big\{ m^2 |\bar{a}_j|^{2(2m-1)} [17|{A_j}^*| + 5|A_j|^2 \\ & + 12|A_j||{A_j}^*| + |\bar{a}_j|^{4m}[|A_j|^2 + |{A_j}^*|^2] \Big\} + \sum_{j=(\bar{s}/2)+1}^{N-\bar{s}/2} |A_j|^2 |a_j|^{2m} \ . \end{aligned}$$

Therefore

$$\| \mathcal{L}^m_{\sigma_2,\omega}[f] \ \|^2 \leq 20 |\widehat{\lambda}_1|^{2(m-1)} m^2 \|f\|^2$$

$$\text{(9.13)} \qquad \text{or}$$

$$\| \mathcal{L}^m_{\sigma_2,\omega}[f] \ \| \leq 5m |\widehat{\lambda}_1|^{m-1} \|f\| \ .$$

Hence, if $\mathcal{N}$ is the largest solution of

$$m|\widehat{\lambda}_1|^{m-1} = \rho/5$$

for all $\quad m > \mathcal{N}$

$$\| \mathcal{L}^m_{\sigma_2,\omega} \| \leq \rho \|f\| \ .$$

Q. E. D.

The reduction of the number of iterations required using the Successive Overrelaxation Method is seen by (9.6) to be in general not as large as the gain in the rate of convergence. However, when ρ is very small the gains in each case are approximately equal. For small ρ and for μ_1 nearly one

$$\mathcal{N}(\mathcal{L}_{\sigma_2}, \rho) \sim \frac{-\log \rho}{-\log \lambda_1}$$

and by Theorem 9.1

$$\mathcal{N}(\mathcal{L}_{\sigma_2,\omega_b}, \rho) \sim \frac{-\log \rho}{\sqrt{-\log \lambda_1}}$$

For the Dirichlet Problem, for unit square with $h = 1/I$ we have by (4.30) and Theorem 5.4

$$\lambda_1 \sim 1 - \frac{\pi^2}{I^2} \quad ; \quad -\log \lambda_1 \sim \frac{\pi^2}{I^2}$$

By Theorem 8.7

$$\widehat{\lambda}_1|_{\omega=\omega_b} = \omega_b - 1 \sim 1 - \frac{2\pi}{I}$$

$$-\log \widehat{\lambda}_1|_{\omega=\omega_b} \sim \frac{2\pi}{I}$$

$$\frac{-\log \widehat{\lambda}_1|_{\omega=\omega_b}}{-\log \lambda_1} \sim \frac{2I}{\pi}$$

Thus, the gain in the rate of convergence by using the Successive Overrelaxation Method is of the order of I.

C. The Use of Large Automatic Computing Machines.

The large reduction in the required number of iterations makes the Successive Overrelaxation Method much more practical for use with large automatic computing machines. A study of the coding for the UNIVAC for a Dirichlet Problem to be solved by means of the Liebmann Method [31] (see also Section II.4) reveals that only two additional memory locations would be required to use the Successive Overrelaxation Method and that the time per iteration would not be increased by more than 10%. For both methods, the stored values which are used on each operation are the values of u at the same set of net points.

The following table gives a comparison of the estimated number of iterations required using the Liebmann Method and the Successive Overrelaxation Method for the Dirichlet Problem with the unit square with I intervals on a side. Time estimates for the UNIVAC are also given. [See Table next page.]

Table 9.1

	THE LIEBMANN METHOD			THE SUCCESSIVE OVERRELAXATION METHOD			
I	λ_1	$\mathcal{N}(\mathcal{L}_\sigma, \rho)^\dagger$	Time‡	ω_b	$\widehat{\lambda}_1$	$\mathcal{N}(\mathcal{L}_{\sigma,\omega_b}, \rho)^\dagger$	Time‡
20	.9754	280	.31	1.755	.755	45	.05
50	.99605	1750	12.15	1.882	.882	106	.74
100	.999015	7010	195.0	1.940	.940	227	7.95
300	.999890	63300	15,750	1.980	.980	750	186.5

In the above estimates ρ was assumed to be 10^{-3}. For smaller ρ, the advantage of the Successive Overrelaxation Method would be greater.

Since $\mu_1 = \sqrt{\lambda_1}$ is known exactly for a square or rectangular region, the determination of ω_b was easy in the above examples. In the next section, we discuss methods for choosing ω_b for more general regions.

---

† $\mathcal{N}(\mathcal{L}_{\sigma,\omega_b}, \rho)$ is the solution of $\quad m|\omega_b - 1|^{m-1} = \rho/5$.
$\mathcal{N}(\mathcal{L}_\sigma, \rho)$ is the solution of $\quad \lambda_1^m = \rho$.

‡The estimated time is in hours computing time for the UNIVAC. The number of seconds per iteration was assumed to be $I^2/100$, [31].

§10. The Determination of the Optimum Relaxation Factor

By (8.19), the optimum relaxation factor ω_b is given by $\omega_b = 1 + \left[\dfrac{1-\sqrt{1-\mu_1^2}}{\mu_1}\right]^2$ where μ_1 is the largest eigenvalue of the Kormes Method.

For the Dirichlet Problem, μ_1 can be expressed in terms of μ_1^T, the smallest eigenvalue of the finite difference analogue of

$$\text{(10.1)} \qquad \begin{cases} \sum\limits_{k=1}^{n} \dfrac{\partial^2 u(x)}{\partial x_k^2} = -\Lambda u(x) & x \in R \\ u(x) = 0 & x \in S \end{cases}$$

which is given by

$$\text{(10.2)} \qquad \begin{aligned} \left(\left[\sum_{k=1}^{n} E_{x_k} + E_{-x_k} - 2n \right] u \right)_i &= -\mu^T u_i \qquad (x^{(i)} \in R_h) \\ u_i &= 0 \qquad (x^{(i)} \in S_h)\ . \end{aligned}$$

By (10.2) and (4.3a), we have

$$\text{(10.3)} \qquad 2n(1-\mu_1) = \mu_1^T .$$

Also, by [8]

$$\text{(10.4)} \qquad \lim_{h\to 0} \frac{\mu_1^T}{h^2} = \Lambda_1 .$$

where Λ_1 is the smallest eigenvalue of (10.1).

The relation between μ_1^T/h^2 and Λ_1 has been investigated by Collatz ([7], pages 280–297) and the agreement has been shown to be remarkably close for some regions even with coarse meshes. Λ_1 can be computed exactly for some regions including the circle ([9] page 260), and the ellipse [15]. Of course, μ_1 itself can be computed exactly for square or rectangular regions, (4.28).

For the case of a rectangular region with side lengths

$$I_k h = \tau_k \qquad (k = 1, 2, \ldots, n)\ ,$$

we have

$$\text{(10.5)} \qquad \Lambda_1 = \frac{\pi^2}{h^2} \sum_{k=1}^{n} \frac{1}{I_k^2} = \pi^2 \sum_{k=1}^{n} \frac{1}{\tau_k^2}\ .$$

By (10.3) and (4.28),

$$\begin{aligned} \frac{\mu_1^T}{h^2} &= \frac{1}{h^2}\left\{ 2n - 2\sum_{k=1}^{n} \cos\frac{\pi}{I_k} \right\} \\ &= \frac{\pi^2}{h^2}\left\{ \sum_{k=1}^{n} \frac{1}{I_k^2} - \frac{1}{12}\sum_{k=1}^{n} \frac{1}{I_k^4} + O(h^6) \right\} \end{aligned}$$

$$\text{(10.6)} \qquad \frac{\mu_1^T}{h^2} = \pi\left\{ \sum_{k=1}^{n} \frac{1}{\tau_k^2} - \frac{h^2}{12}\sum_{k=1}^{n} \frac{1}{\tau_k^4} \right\} + o(h^6)\ .$$

Hence, the convergence of μ_1^T/h^2 to Λ, is, in this case very rapid.

In any case, for the Dirichlet Problem, a lower bound to μ_1^T can be found by computing μ_1^T for a circumscribing rectangular region, because of the Theorem 4.2, Corollary.

We now show that for the general self adjoint case, if $(1-\mu_1)$ is not overestimated, the relative decrease in the rate of convergence is, asymptotically, for small errors in $(1-\mu_1)$ and for small $(1-\mu_1)$, equal to one half the relative error in the estimation of $(1-\mu_1)$.

Definition 10.1 The **relative change**, or **error**, of y' with respect to y is

$$\left\{\frac{|y'-y|}{|y|}\right\} \qquad (y \neq 0)\ .$$

Theorem 10.1 Let μ_1 be the dominant eigenvalue of $\mathcal{K}$ and let μ_1' be the estimated value of μ_1 such that

$$\mu_1 \leq \mu_1' < 1$$

Then

$$\lim_{\mu_1 \to 1} \left\{ \lim_{\mu_1' \to \mu_1} \frac{\left|\frac{\phi(\mathcal{L}_{\sigma,\omega'}) - \phi(\mathcal{L}_{\sigma,\omega})}{\phi(\mathcal{L}_{\sigma,\omega})}\right|}{\left|\frac{(1-\mu_1{}')-(1-\mu_1)}{1-\mu_1}\right|} \right\} = \frac{1}{2} \tag{10.7}$$

where $\omega,\ \omega'$ are determined from (8.19) based on μ_1 and μ_1' respectively, and where $\phi(\mathcal{L}_{\sigma,\omega})$ and $\phi(\mathcal{L}_{\sigma,\omega'})$ are the rates of convergence of $\mathcal{L}_{\sigma,\omega}$ and $\mathcal{L}_{\sigma,\omega'}$ respectively.

Proof: The left member of (10.7) equals

$$\lim_{\mu_1 \to 1} \frac{-\frac{d}{d\mu_1}\left[\log \phi(\mathcal{L}_{\sigma,\omega})\right]}{-\frac{d}{d\mu_1}\left[\log(1-\mu_1)\right]} \tag{10.8}$$

provided the derivatives exist. Now let $\mu_1 = \cos\theta$; by (8.19)

$$(\omega-1)^{1/2} = \frac{1-\sin\theta}{\cos\theta} = \tan(\frac{\pi}{4} - \frac{\theta}{2})$$

$$\frac{d}{d\theta}(\omega-1)^{1/2} = -\frac{1}{2}\sec^2(\frac{\pi}{4} - \frac{\theta}{2})$$

$$\begin{aligned}
-\frac{d}{d\mu_1}\left[\log \phi(\mathcal{L}_{\sigma,\omega})\right] &= -\frac{d}{d\mu_1}\left[\log(-\log(\omega-1))\right] \\
&= \frac{\frac{d}{d\mu_1}(-\log(\omega-1))}{-\log(\omega-1)} = \frac{-2\frac{d}{d\mu_1}(\omega-1)^{1/2}}{-\log(\omega-1)\cdot(\omega-1)^{1/2}} \\
&= \frac{-2\frac{d}{d\mu_1}(\omega-1)^{1/2}\frac{d\theta}{d\mu_1}}{-\log(\omega-1)\cdot(\omega-1)^{1/2}} \\
&= \frac{\sec^2(\frac{\pi}{4}-\frac{\theta}{2})}{-\log\tan^2(\frac{\pi}{4}-\frac{\theta}{2})\cdot\tan(\frac{\pi}{4}-\frac{\theta}{2})}\ \frac{d\theta}{d\mu_1}
\end{aligned}$$

$$-\frac{d}{d\mu_1}\left[\log\phi(\mathcal{L}_{\sigma,\omega})\right] = \frac{1}{-\log\tan(\frac{\pi}{4}-\frac{\theta}{2})\cdot\cos\theta}\,\frac{d\theta}{d\mu_1}\ . \tag{10.9}$$

Also

$$-\frac{d}{d\mu_1}\log(1-\mu_1) = \frac{1}{1-\mu_1}\,\frac{d\mu_1}{d\theta}\,\frac{d\theta}{d\mu_1} = \cot\frac{\theta}{2}\,\frac{d\theta}{d\mu_1}\ . \tag{10.9a}$$

The left member of (10.7) equals

$$\lim_{\theta\to 0}\frac{\tan\frac{\theta}{2}}{-\log\tan(\frac{\pi}{4}-\frac{\theta}{2})\cdot\cos\theta} = \frac{1}{2} \tag{10.10}$$

by l'Hôpital's Rule.

Q. E. D.

Theorem 10.2

$$\lim_{\mu_1\to 1}\frac{\phi(\mathcal{L}_{\sigma,\omega'})}{\phi(\mathcal{L}_{\sigma,\omega})} = \sqrt{\zeta} \tag{10.11}$$

where

$$(1-\mu_1') = \zeta(1-\mu_1) \qquad\qquad (0<\zeta\le 1)\ ,$$

and where ω,ω' are determined from (8.19) using μ_1 and μ_1' respectively.

Proof: $\omega\ge\omega'$, hence all eigenvalues of $\mathcal{L}_{\sigma,\omega'}$ have absolute value $|\omega'-1|$

$$\frac{\phi(\mathcal{L}_{\sigma,\omega'})}{\phi(\mathcal{L}_{\sigma,\omega})} = \frac{-\log(\omega'-1)}{-\log(\omega-1)}\ .$$

Let $\quad \mu_1=\cos\theta\ , \quad \mu_1'=\cos\theta' \quad$ where $\quad \theta'\le\theta$

$$1-\cos\theta' = \zeta(1-\cos\theta)$$

$$\sin^2\frac{\theta'}{2} = \zeta\sin^2\frac{\theta}{2}\ .$$

But, as in Theorem 10.1,

$$(\omega-1)^{1/2} = \frac{1-\sin\theta}{\cos\theta} = \tan(\frac{\pi}{4}-\frac{\theta}{2})\ , \tag{10.12}$$

$$(\omega'-1)^{1/2} = \frac{1-\sin\theta'}{\cos\theta'} = \tan(\frac{\pi}{4}-\frac{\theta'}{2})\ . \tag{10.13}$$

Neglecting terms in θ and θ' of degree higher than the first we have

$$\theta' = \sqrt{\zeta}\theta + o(\theta)\ , \tag{10.14}$$

and

$$-2\log(\omega-1)^{1/2} = 2\theta + o(\theta)$$
$$-2\log(\omega'-1)^{1/2} = 2\theta' + o(\theta)$$

where $o(\theta)$ vanishes with θ. Hence, the theorem follows.

Q. E. D.

From this, it can be seen that even if $(1-\mu_1)$ is underestimated by as much as 50% the rate of convergence will be decreased by less than 30%. Since the gain in the rate of convergence of the Successive Overrelaxation Method is of an order of magnitude, however, such an increase is relatively unimportant.

On the other hand, if $\zeta > 1$ we have

$$\phi(\mathcal{L}_{\sigma,\omega'}) = -\log\left\{\frac{\omega'\mu_1}{2} + \frac{\sqrt{\omega'^2\mu_1^2 - 4(\omega'-1)}}{2}\right\}$$

where $\omega'^2\mu_1'^2 - 4(\omega'-1) = 0$ by (8.19). Therefore

$$\phi(\mathcal{L}_{\sigma,\omega'}) = -\log\left\{\frac{\omega'}{2}\left[\mu_1 + \sqrt{\mu_1^2 - \mu_1'^2}\right]\right\} \ .$$

If $\mu_1' = \cos\theta'$, by (10.13),

$$(\omega'-1)^{1/2} = \frac{1-\sin\theta'}{\cos\theta'} = \tan(\frac{\pi}{4} - \frac{\theta'}{2})$$

$$\omega' = \sec^2(\frac{\pi}{4} - \frac{\theta'}{2}) \ .$$

Neglecting terms in θ' higher than the first and setting $\theta' = \sqrt{\zeta}\theta + o(\theta)$ as in (10.14)

$$\phi(\mathcal{L}_{\sigma,\omega'}) = \theta\left[\sqrt{\zeta} - \sqrt{\zeta-1}\right] + o(\theta)$$

Since $\phi(\mathcal{L}_{\sigma,\omega}) = \theta + o(\theta)$, we have

$$\left[\frac{\phi(\mathcal{L}_{\sigma,\omega'})}{\phi(\mathcal{L}_{\sigma,\omega})}\right] = \sqrt{\zeta} - \sqrt{\zeta-1} + o(\theta) \ . \tag{10.15}$$

Because of the presence of the term $\sqrt{\zeta-1}$, the derivative of the above ratio with respect to ζ becomes large as ζ approaches one. Thus, it is always better to **underestimate** $(1-\mu_1)$, i.e. be sure $\zeta \leq 1$.

As stated in the Introduction, the fact that the gain in rate of convergence is not very sensitive to relative changes in $(1-\mu_1)$ (as long as $(1-\mu_1)$ is not overestimated), suggests that not only for the Dirichlet Problem but for other self adjoint equations the relaxation factor can be chosen so that the problems can be solved much more rapidly. For equations of the type (1.11) where $(a_{i,j})$ is not symmetric, however, the eigenvalues of $\mathcal{K}$ are not in general real and the gain in convergence which could be obtained, even with the best relaxation factor, is much less. Nevertheless the relation between the eigenvalues of $\mathcal{K}$, $\mathcal{L}_\sigma$ and $\mathcal{L}_{\sigma,\omega}$ are valid. The study of the convergence of $\mathcal{L}_\sigma$ and $\mathcal{L}_{\sigma,\omega}$ is therefore simplified.

References

[1] S. Bergmann, *Partial Differential Equations, Advanced Topics*, Brown University, (1941).

[2] W. G. Bickley, "Finite Difference Formulae for the Square Lattice," *Quar. J. Mech. and App. Math.* 1, 253–279, (1948).

[3] G. Birkhoff and S. MacLane, *A Survey of Modern Algebra*, New York, Macmillan, (1946).

[4] O. Bowie, "A Least-Square Application to Relaxation Methods," *J. App. Phys.*, 18, 830–837, (1947).

[5] O. L. Bowie, *Electrical Computing Board for the Numerical Solution of Partial Differential Equations*, Watertown Arsenal Laboratory Report, WAL 790/22

[6] L. Collatz, "Bemerkungen zur Fehlerabschätzung für das Differenzenverfahren bei partiellen Differentialgleichungen," *Zeits. f. Angew. Math. u. Mech.*, 13, 56–57, (1933)

[7] L. Collatz, *Eigenwertprobleme und Ihre Numerische Behandlung*, New York, Chelsea Publishing Co., (1948).

[8] R. Courant, K. Friedrichs and H. Lewy, "Über die Partiellen Differenzengleichungen der Mathematischen Physik," *Math. Ann.* 100, 32–74, (1928).

[9] R. Courant and D. Hilbert, *Methoden der Mathematischen Physik*, I, Berlin, Julius Springer, (1931).

[10] A. Dresden, "On the Iteration of Linear Homogeneous Transformations," *Bull. Amer. Math. Soc.*, 48, 577–579, (1942).

[11] H. Emmons, "Numerical Solution of Partial Differential Equations," *Quar. App. Math.* 2, 173–195, (1945).

[12] H. Geiringer, *On the Solution of Systems of Linear Equations by Certain Iteration Methods*, Ann Arbor, Michigan, Reissner Anniversary Volume, 365–393, (1949).

[13] S. Gerschgorin, "Fehlerabschätzung für das Differenzenverfahren zur Lösung Partieller Differentialgleichungen," *Zeits. f. Angew. Math. u. Mech.*, 10, 373–382, (1930).

[14] D. Hartree, *Calculating Instruments and Machines*, Urbana, Univ. of Illinois Press, (1949).

[15] J. Herriot, *The Principal Frequency of an Elliptic Membrane*, Technical Report No. 3, Navy Contract N6-ORT-106, Task Order 5, California, Stanford University, August, (1949).

[16] T. J. Higgens, "A Survey of the Approximate Solution of Two Dimensional Physical Problems by Variational Methods and Finite Difference Procedures," Chapter 10 of *Numerical Methods of Analysis in Engineering*, New York, Macmillan, (1949).

[17] D. Jackson, *Fourier Series and Orthogonal Polynomials*, No. 6, Carus Mathematical Monographs, Mathematical Association of America, (1941).

[18] O. D. Kellogg, *Foundations of Potential Theory*, Murray Printing Co., (1929).

[19] M. Kormes, "Numerical Solution of the Boundary Value Problem for the Potential Equation by Means of Punched Cards," *Rev. Sci. Inst.* 14, 248–250, (1943).

[20] L. Lichtenstein, "Neuere Entwickelungen der Potentialtheorie, Konforme Abbildung," *Encyklopädie der Mathematischen Wissenschaften*, III. C. 3, 177–377.

[21] H. Liebmann, "Die Angenährte Ermittelung harmonischer Functionen und Konformer Abbildungen," *Sitz-Bayer. Akad. Wiss. Math.–Phys. Klasse*, 385–416 (1918).

[22] C. MacDuffee, *Introduction to Abstract Algebra*, New York, New York, John Wiley and Sons, Inc., (1940).

[23] W. E. Milne, "Numerical Methods Associated with Laplace's Equation," to be published in the *Proceedings of the Symposium on Large-Scale Digital Calculating Machinery*, Cambridge, Harvard University, September, (1949).

[24] D. Moscovitz, "The Numerical Solution of Laplace's and Poisson's Equations," *Quar. App. Math.* 2, 148–163 (1944).

[25] H. B. Phillips and N. Wiener, "Nets and the Dirichlet Problem," *J. Math. and Phys.*, 28, 105–124 (1923).

[26] L. F. Richardson, "The Approximate Arithmetical Solution by Finite Differences of Physical Problems Involving Differential Equations With an Application to the Stresses in a Masonry Dam," *Phil.Trans.*, 210 A 307–357 (1910).

[27] C. Runge, "Über eine Methode die partielle Differentialgleichungen $\triangle u =$ constans numerisch zu integrieren," *Zeits. f. Math. u. Phys.* 56, 225–232, (1908–09).

[28] Ludwig Seidel, *Münchener Akademische Abhandlungen*, 2 Abhandlungen, 81-108, (1874).

[29] D. Shanks, *An Analogy between Transients and Mathematical Sequences and Some Nonlinear Sequence-to-Sequence Transformations Suggested by It*, Naval Ordnance Laboratory Memorandum 9994, White Oak, Maryland, 26 July, (1949).

[30] G. H. Shortley and R. Weller, "The Numerical Solution of Laplace's Equation," *J. App. Phys.* 9, 334–344 (1938).

[31] F. Snyder and H. Livingston, "Coding of a Laplace Boundary Value Problem for the UNIVAC," *Math. Table and other Aids to Computation*, III, 341–350, (1949).

[32] R. V. Southwell, *Relaxation Methods in Theoretical Physics*, Oxford University Press, (1946).

[33] J. F. Steffensen, *Interpolation*, Baltimore, Williams and Wilkins, (1927).

[34] Stein and Rosenberg, "On the Solution of Systems of Simultaneous Equations by Iteration," *Jour. London Math. Soc.*, 23, 111–118, (1946).

[35] G. Temple, "The General Theory of Relaxation Methods Applied to Linear Systems," *Proc. Roy, Soc.* A 169, 476–500 (1938–39).

[36] H. Thomas, "Elliptic Problems in Linear Difference Equations over a Network," Lecture notes, Watson Scientific Laboratory, New York, 1–7.

[37] D. Widder, *Advanced Calculus*, New York, Prentice Hall, Inc., (1947).

Summary

Iterative Methods for Solving Partial Difference Equations of Elliptic Type

David M. Young, Jr.

The finite difference analogue of a boundary value problem involving a linear, second order partial differential equation of elliptic type can be reduced to a system of linear equations of the form

$$\sum_{j=1}^{N} a_{i,j}u_j + d_i = 0 \qquad (i = 1, 2, ..., N) \tag{1}$$

where the coefficients $a_{i,j}$ satisfy the conditions

$$\begin{cases} \text{(a)} \quad a_{i,i} > 0 & (i = 1, 2, \ldots, N) \\ \text{(b)} \quad a_{i,j} \leq 0 & (i \neq j;\ i, j = 1, 2, \ldots, N) \\ \text{(c)} \quad a_{i,i} \geq \sum\limits_{\substack{j=1 \\ j\neq i}}^{N} |a_{i,j}| & (i = 1, 2, \ldots, N) \\ \quad \text{and for some } i \\ \quad a_{i,i} > \sum\limits_{\substack{j=1 \\ j\neq i}}^{N} |a_{i,j}| \\ \text{(d)} \quad \text{The } N \times N \text{ matrix } (a_{i,j}) \text{ is } \textit{irreducible}, \text{ that is, given any two} \\ \quad \text{non empty complementary subsets, } \mathcal{S}_{\mathcal{E}} \text{ and } \mathcal{S}_{\mathcal{I}} \text{ of the set of the first} \\ \quad N \text{ integers there exists } a_{i,j} \neq 0 \text{ such that } i \in \mathcal{S}_E \text{ and } j \in \mathcal{S}_0. \end{cases} \tag{2}$$

For the self adjoint case,

$$a_{i,j} = a_{j,i} \qquad (i, j = 1, 2, \ldots, N). \tag{3}$$

Geiringer [12] has shown that if $(a_{i,j})$ satisfies (2), then (1) has a unique solution. Actually obtaining the solution may be very laborious. In this thesis the practicability of various methods for solving (1) is considered, with special emphasis on those adapted to large automatic computing machines. Particular attention is given to the finite difference analogue of the Dirichlet Problem.

Direct methods such as the use of determinants and elimination do not appear practical for large N, and various methods of successive approximation are usually employed, including iterative methods and relaxation methods. For successive approximation methods, an arbitrary initial approximation $u^{(0)} = (u_1^{(0)}, u_2^{(0)}, \ldots, u_N^{(0)})$ is chosen and successively improved. In Chapter I known proofs are given for the convergence of $u^{(m)}$ to the solution of (1) as $m \to \infty$.

Iterative methods appear to be most suited for large automatic computing machines. The usual methods are the Kormes Method and the Liebmann Method. The sequences $\{u^{(m)}\}$ are defined as follows

(a) The Kormes Method

$$u_i^{(m+1)} = -\sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{a_{i,j}}{a_{i,i}}\, u_j^{(m)} - \frac{d_i}{a_{ii}} \qquad (i = 1, 2, \cdots, N) \tag{4}$$

or

$$u^{(m+1)} = \mathcal{K}\ [u^{(m)}] + c, \tag{4a}$$

where

$$c = (-\frac{d_1}{a_{11}}, -\frac{d_2}{a_{22}}, \ldots, -\frac{d_N}{a_{NN}})\ . \tag{5}$$

(b) The Liebmann Method

$$u_i^{(m+1)} = -\sum_{j=1}^{i-1} \frac{a_{i,j}}{a_{i,i}}\ u_j^{(m+1)} - \sum_{j=i+1}^{N} \frac{a_{i,j}}{a_{i,i}}\ u_j^{(m)} - \frac{d_i}{a_{i,i}} \qquad (i = 1, 2, \cdots, N) \tag{6}$$

or

$$u^{(m+1)} = \mathcal{L}_\sigma\ [u^{(m)}] + c \tag{6a}$$

where σ denotes the ordering of the equations.

The rates of convergence of these methods depend on the dominant eigenvalues of the linear operators $\mathcal{K}$ and $\mathcal{L}_\sigma$, which are linear operators on the vector space V_N of N-tuples of complex numbers. For the study of these, we use the fact that $(a_{i,j})$ has property (A_q) for some integer q.

Definition: $(a_{i,j})$ has **property** $(\mathbf{A_q})$ if there exist non empty disjoint subsets $T_1, T_2, \ldots, T_q$ of $\mathcal{T}$, the set of the first N integers, such that $\bigcup_{\ell=1}^{q} T_\ell = \mathcal{T}$, and such that the T_ℓ can be labeled so that $a_{i,j} = 0$ unless $i = j$ or $i \in T_\ell$ and $j \in T_{\ell-1} \bigcup T_{\ell+1}$. ($T_0$ and T_{q+1} denote the empty set.)

Using this property, we show that the eigenvalues μ of $\mathcal{K}$ occur in pairs $(\mu, -\mu)$. There exist certain orderings[†] of the equations such that to each such pair there corresponds an eigenvalue $\lambda = \mu^2$ of $\mathcal{L}_\sigma$; hence, the rate of convergence of $\mathcal{L}_\sigma$ is exactly twice that of $\mathcal{K}$. An explicit expression of the eigenvectors of $\mathcal{L}_\sigma$ in terms of the eigenvectors of $\mathcal{K}$ is also given. For the symmetric case, all eigenvalues of $\mathcal{K}$ and $\mathcal{L}_\sigma$ are real and the Jordan normal form of the corresponding matrices are diagonal with the possible exception of the subspace asssociated with $\lambda = 0$ for $\mathcal{L}_\sigma$. If, also, a_{ii} = constant, as for the Dirichlet Problem, the eigenvectors of $\mathcal{K}$ are orthogonal. For the Dirichlet Problem the eigenvalues and eigenvectors of $\mathcal{K}$ and $\mathcal{L}_\sigma$ can be computed exactly for a rectangular region. For one ordering, σ_2, the normal form of the matrix of $\mathcal{L}_{\sigma_2}$ is diagonal and if the coordinates of an arbitrary vector in V_N referred to the basis of eigenvectors of $\mathcal{K}$ are known, the coordinates of that same vector referred to the basis of eigenvectors of $\mathcal{L}_{\sigma_2}$ can be computed at once.

For the symmetric case, a conjecture that by the Liebmann Method the rate of convergence can not be increased by using an ordering which is not consistent, is proved in one special case. Some numerical studies bear out a conjecture by Shortley and Weller, [30], that for large N the rate of convergence of $\mathcal{L}_\sigma$ is nearly independent of σ.

The number of iterations required to reduce the norm of the initial error function

$$\|u^{(0)} - u\| = \left[\sum_{i=1}^{N} (u_i^{(0)} - u_i)^2\right]^{1/2}$$

to a definite fraction of itself is asymptotically for small h proportional to h^{-2} (h is the mesh size). For some problems where a very small mesh size must be used, the time required to obtain an acceptable degree of accuracy, even with a fast computing machine such as the UNIVAC, is prohibitive.

[†]Such orderings are called "consistent" orderings.

In Chapter III, it is shown that the required number of iterations can be greatly reduced by using the *Successive Overrelaxation Method*, where the idea of *systematic overrelaxation*, first used by L. F. Richardson, [26], is combined with the Liebmann Method. For the Dirichlet Problem, the reduction is of the order of h^{-1}, and for the general self adjoint case, if the required number of iterations with the Liebmann Method is of the order of h^{-k}, the reduction is of the order of $h^{-k/2}$, if the Successive Overrelaxation Method is used with the proper relaxation factor ω. Moreover, this new improved method can be used with any large automatic computing machine for which the Liebmann Method can be used; the machine time per iteration would not be increased by more than 10%.

The improvement formula is

$$(7)\quad u_i^{(m+1)} = \omega\left[-\sum_{j=1}^{i-1}\frac{a_{i,j}}{a_{i,i}}\,u_j^{(m+1)} - \sum_{j=i+1}^{N}\frac{a_{i,j}}{a_{i,i}}\,u_j^{(m)} - \frac{d_i}{a_{i,i}}\right] - (\omega-1)u_i^{(m)} \qquad (i=1,2,\ldots,N)$$

or

$$(0.7a)\qquad u^{(m+1)} = \mathcal{L}_{\sigma,\omega}[u^{(m)}] + \omega c$$

where the subscripts σ and ω of $\mathcal{L}_{\sigma,\omega}$ denote the ordering, (assumed consistent), of the equations, and the relaxation factor respectively. $\mathcal{L}_{\sigma,\omega}$ is a linear operator on V_N, and if $\omega = 1$, we have the Liebmann Method.

If μ is an eigenvalue of $\mathcal{K}$, there exists an eigenvalue $\widehat{\lambda}$ of $\mathcal{L}_{\sigma,\omega}$ such that

$$(8)\qquad \mu\omega\widehat{\lambda}^{1/2} = \widehat{\lambda} + (\omega - 1)$$

and conversely every eigenvalue of $\mathcal{L}_{\sigma,\omega}$ can be determined by (8), for some μ. The eigenvectors of $\mathcal{L}_{\sigma,\omega}$ can be expressed explicitly in terms of the eigenvectors of $\mathcal{K}$. If $(a_{i,j})$ is symmetric, then the optimum relaxation factor, ω_b, is given by

$$(9)\qquad \mu_1^2\,\omega_b^2 = 4(\omega_b - 1)$$

where μ_1 is the largest eigenvalue of $\mathcal{K}$. For all $\omega = \omega_b$, every eigenvalue of $\mathcal{L}_{\sigma,\omega}$ has absolute value $(\omega - 1)$, and the Jordan normal form of the matrix of $\mathcal{L}_{\sigma,\omega}$ is diagonal unless $\omega = \omega_b$. In this case, the normal matrix form contains precisely one non diagonal element.

For the Dirichlet Problem, if h is small, μ_1 is very nearly one. μ_1 can be calculated exactly for a rectangular region and for other regions can be estimated by comparison theorems. For the general self adjoint case, provided μ_1 is not underestimated, (for the Dirichlet Problem a non trivial upper bound for μ_1 can always be found), the relative decrease in the rate of convergence, if $\omega' \geq \omega_b$ is used, is approximately $\zeta^{-1/2} - 1$, where $(1 - \mu_1') = \zeta(1 - \mu_1)$ $(0 < \zeta \leq 1)$ and where ω' is determined from (9), but with μ_1 replaced by μ_1'. Thus a relatively large error in the estimation of $(1 - \mu_1)$ can be allowed and the improvement over the Liebmann Method will not suffer appreciably. This suggests that the Successive Overrelaxation Method can be applied successfully to self adjoint equations other than Laplace's Equation.