Numerical Solution of the Two-Group Diffusion Equations in x-y Geometry*

R. S. VARGA†

Summary—The problem studied in this paper is the numerical solution of the two-group diffusion equations describing the reactivity and power distribution of a nuclear power reactor. The problem is treated in two dimensions (Cartesian coordinates). The method of solution by replacement of the differential equations by finite difference equations is outlined. The properties of the resulting matrices are studied in detail. The method of successive overrelaxation is described and the theory developed. The convergence properties of the method and its application is indicated.

I. Introduction

THE PROBLEM we are concerned with here is the numerical solution of the two-group diffusion equations for a heterogeneous reactor in a rectangular region R of the x-y plane with external boundary Γ. These equations are of the form:

$$\begin{cases} \nabla\cdot[D_s(x, y)\nabla\phi(x, y)] - \sigma_s(x, y)\phi(x, y) \\ \qquad\qquad + \Sigma_s^f(x, y)\psi(x, y) = 0 \\ \nabla\cdot[D_f(x, y)\nabla\psi(x, y)] - \sigma_f(x, y)\psi(x, y) \\ \qquad\qquad + \eta\Sigma_{25}(x, y)\phi(x, y) = 0. \end{cases} \tag{1}$$

The functions $\phi(x, y)$, $\psi(x, y)$, $D_s\nabla\phi$, and $D_f\nabla\psi$, where defined, are to be continuous interior to the region R. With homogeneous boundary conditions specified on Γ such as the vanishing of both ϕ, ψ on Γ, we seek to find the *pair* of functions $\phi(x, y)$, $\psi(x, y)$ satisfying (1) interior to R and the specified homogeneous boundary conditions on Γ, corresponding to the *smallest* (in modulus) eigenvalue η.

The problem stated above is what may be called the two-group, steady-state diffusion problem in x-y geometry for the continuous case. In order to solve the problem numerically, we shall replace the problem for the continuous case by what is called the discrete case, making certain simplifying assumptions in the passage from one problem to the other. The conditions and assumptions which we shall make are just those used in the machine code *QED*.[1] The subsequent sections contain a complete mathematical analysis for the discrete case and the relevant mathematical discussion of the iterative procedures used to solve this discrete case.

II. The Continuous and Discrete Problems

A. Statement of the Problem in the Continuous Case—Passage to the Discrete Case

To state precisely the problem in the continuous case, we assume we have a finite set of regions R_i and C_i and internal interfaces γ_i, which separate the various regions (see Fig. 1). Together, the regions R_i and C_i plus the interfaces γ_i constitute R, our rectangular region of interest. All internal interfaces are composed of horizontal or vertical line segments only. The regions R_i are called *diffusion regions* and interior to any diffusion region R_i the functions $\phi(x, y)$, $\psi(x, y)$, called respectively the slow group flux and the fast group flux, are defined and satisfy (1). The quantities D_s, D_f, σ_s, σ_f, Σ_s^f, and Σ_{25} are assumed to be region-wise constant, and they are all positive except for Σ_{25} which is nonnegative. For the region R_i, (1) reduces to

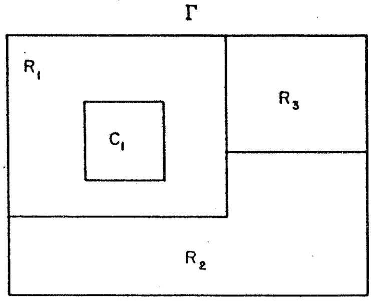

Fig. 1.

$$\begin{cases} D_{s,i}\nabla^2\phi(x, y) - \sigma_{s,i}\phi(x, y) \\ \qquad\qquad + \Sigma_{s,i}^f\psi(x, y) = 0 \\ D_{f,i}\nabla^2\psi(x, y) - \sigma_{f,i}\psi(x, y) \\ \qquad\qquad + \eta\Sigma_{25,i}\phi(x, y) = 0, \qquad x, y\,\varepsilon\, R_i\,. \end{cases} \tag{2}$$

Across any internal interface γ_i between two diffusion regions, the quantities ϕ, ψ, $D_s\nabla\phi$, and $D_f\nabla\psi$ are continuous. These restrictions are referred to, respectively, as continuity of flux and current.

For the regions C_i with boundary γ_i the slow group flux ϕ is *not* defined interior to C_i but satisfies

$$\frac{D_s}{\phi(x, y)}\frac{\partial\phi}{\partial n}\bigg|_{x,y\varepsilon\gamma_i} = -c, \tag{3}$$

where c is a *positive* constant, and the derivative is taken perpendicular to γ_i in the direction of C_i. For the fast

* Manuscript received by the PGNS, September 20, 1956. Presented at the Third Annual Meeting of the PGNS. This report was prepared from work performed under AEC Contract AT-11-1 Gen-14.

† Bettis Atomic Power Div., Westinghouse Elec. Corp., Pittsburgh, Pa.

[1] This is a Bettis Atomic Power Division (of Westinghouse) code for the IBM-704. It had as its predecessor the Cuthill code for UNIVAC, which was developed at the David Taylor Model Basin by Drs. Cuthill and Davis.

group flux ψ, C_i is a diffusion region with $\Sigma_{25} = 0$ in C_i, so that $\psi(x, y)$ satisfies, interior to C_i, the Helmholtz equation:

$$D_f \nabla^2 \psi(x, y) - \sigma_f \psi(x, y) = 0. \tag{4}$$

On the exterior boundary Γ of R, $\psi(x, y)$, and $\phi(x, y)$ where defined, satisfy the same boundary conditions on each segment of Γ. The boundary conditions which may be used are

$$\psi(x, y) = \phi(x, y) = 0, \quad \text{or} \quad \frac{\partial \phi}{\partial n} = \frac{\partial \psi}{\partial n} = 0. \tag{5}$$

Having stated the problem for the continuous case, we now proceed to the problem for the discrete case. Since all internal interfaces γ_i and the external boundary Γ are, by assumption, composed of horizontal and vertical line segments only, we impose a mesh Λ of horizontal and vertical lines on $R + \Gamma$ in such a way that all internal interfaces and external boundaries lie exactly on mesh lines. The mesh spacings in the x and y directions need not be constant. With the mesh Λ the unknowns in the discrete case are defined to be the values of ϕ, ψ at the intersections of the horizontal and vertical line segments of Λ. By replacing the differential conditions of (2)–(5) with difference equations in the unknowns of the discrete case, we will have defined the discrete problem. The next section gives the derivation of the difference equations for the discrete case.

B. Derivation of the Difference Equations

For an arbitrary mesh point (x_0, y_0) which does *not* lie on an interface γ_i defining a region C_i, each of the regions in the neighborhood of the point (x_0, y_0) is, therefore, a diffusion region. The following group-independent derivation is standard, but it is included for the sake of completeness.

For each of the regions R_i surrounding the point (x_0, y_0), the diffusion equation for that region is

$$D_i \nabla^2 u(x, y) - \sigma_i u(x, y) + \Sigma_i S(x, y) = 0, \qquad i = 1, 2, 3, 4, \tag{6}$$

where the Laplacian operator is, of course, in Cartesian coordinates. Integrating each of these diffusion equations over the appropriate rectangle, say r_i of Fig. 2, we have, since the constants D_i, σ_i and Σ_i are region-wise constant,

$$D_i \iint_{r_i} \nabla^2 u(x, y)\, dx\, dy - \sigma_i \iint_{r} u(x, y)\, dx\, dy + \Sigma_i \iint_{r_i} S(x, y)\, dx\, dy = 0, \qquad i = 1, 2, 3, 4. \tag{7}$$

By Green's theorem, the first term of (7) can be reduced to a line integral about the circumference d_i of the rectangle r_i, which gives

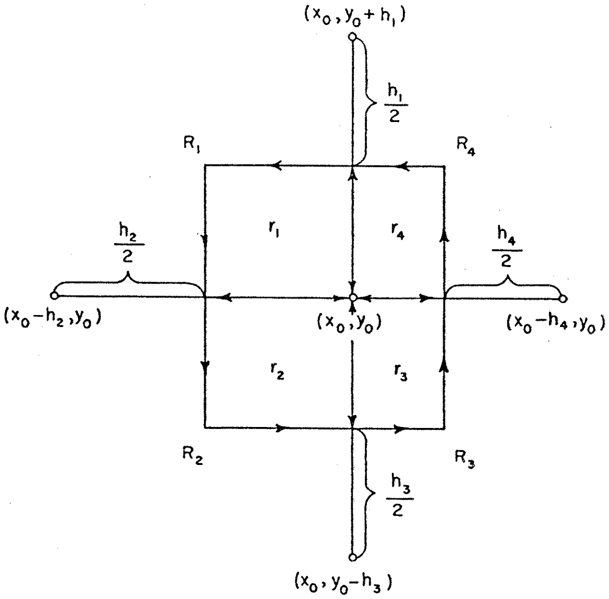

Fig. 2.

$$D_i \int_{d_i} \frac{\partial u}{\partial n}(x, y)\, ds - \sigma_i \iint_{r_i} u(x, y)\, dx\, dy + \Sigma_i \iint_{r_i} S(x, y)\, dx\, dy = 0, \qquad i = 1, 2, 3, 4, \tag{8}$$

where $\partial u(x, y)/\partial n$ is the derivative in the direction along the outward normal to d_i, the line integration being performed in the usual counterclockwise manner, as indicated by the arrows in Fig. 2.

Adding the four expressions of (8) we obtain

$$\sum_{i=1}^{4} \left\{ D_i \int_{d_i} \frac{\partial u}{\partial n}(x, y)\, ds - \sigma_i \iint_{r_i} u(x, y)\, dx\, dy + \Sigma_i \iint_{r_i} S(x, y)\, dx\, dy \right\} = 0. \tag{9}$$

If T denotes the circumference of the union of the rectangles r_1, r_2, r_3, and r_4, then using the continuity of current of Section II-A, we have an exact theoretical cancellation of all line integrals which do not coincide with T, and (9) reduces to

$$\sum_{i=1}^{4} \left\{ D_i \int_{\tilde{d}_i} \frac{\partial u}{\partial n}(x, y)\, ds - \sigma_i \iint_{r_i} u(x, y)\, dx\, dy + \Sigma_i \iint_{r_i} S(x, y)\, dx\, dy \right\} = 0, \tag{10}$$

where $\tilde{d}_i$ is that part of d_i common to T.

In order to reduce this specifically to a five-point formula, we now make numerical approximations to the integrals above. Basically, the approximations made are

$$\int_{b_1}^{b_2}\int_{a_1}^{a_2} g(x, y)\, dx\, dy \doteq g(a_1 , b_1)[(a_2 - a_1)(b_2 - b_1)];$$

$$\int_{a_1}^{a_2} g(x)\, dx = g(a_1)[(a_2 - a_1)]. \qquad (11)$$

Furthermore, derivatives such as $\partial(x_0 + h_1/2)/\partial y$ are approximated by the central difference formula:

$$\frac{\partial u}{\partial y}\left(x_0 , y_0 + \frac{h_1}{2}\right) = \frac{u(x_0 , y_0 + h_1) - u(x_0 , y_0)}{h_1}. \qquad (12)$$

The final numerical approximation to (10) which will be our five-point formula can be written in the form:

$$\begin{aligned}
&\left(\frac{D_1h_2 + D_4h_4}{2h_1}\right)u(x_0 , y_0 + h_1)\\
&\quad + \left(\frac{D_1h_1 + D_2h_3}{2h_2}\right)u(x_0 - h_2 , y_0)\\
&\quad + \left(\frac{D_2h_2 + D_3h_4}{2h_3}\right)u(x_0 , y_0 - h_3)\\
&\quad + \left(\frac{D_3h_3 + D_4h_1}{2h_4}\right)u(x_0 + h_4 , y_0)\\
&\quad - u(x_0 , y_0)\left\{\frac{D_1h_2 + D_4h_4}{2h_1} + \frac{D_1h_1 + D_2h_3}{2h_2}\right.\\
&\quad + \frac{D_2h_2 + D_3h_4}{2h_3} + \frac{D_3h_3 + D_4h_1}{2h_4}\\
&\quad \left. + \tfrac{1}{4}(\sigma_1h_1h_2 + \sigma_2h_2h_3 + \sigma_3h_3h_4 + \sigma_4h_4h_1)\right\}\\
&\quad + S_0(x_0 , y_0)\cdot\\
&\left(\frac{\Sigma_1h_1h_2 + \Sigma_2h_2h_3 + \Sigma_3h_3h_4 + \Sigma_4h_4h_1}{4}\right) = 0. \qquad (13)
\end{aligned}$$

We observe that the coefficient of $u(x_0, y_0)$ is strictly greater, in absolute value, than the sum of the coefficients of the other $u(x, y)$'s since σ_i, a physical cross section, is positive. This will be of importance in Section II-C.

If the chosen mesh point (x_0, y_0) *does* lie on an interface γ_i defining a region C_i and the group considered is the slow group, the above derivation of the five-point formula is not valid. We shall show, for example, how the five-point formula for the slow group is derived at the corner (x_0, y_0) of the region C_1 of Fig. 3. The other cases follow in a similar manner.

In regions R_1, R_2, R_3, the diffusion equations are

$$D_i\nabla^2u(x, y) - \sigma_iu(x, y) + \Sigma_iS(x, y) = 0,$$
$$i = 1, 2, 3. \qquad (14)$$

Integrating over the three rectangles r_1, r_2, r_3, we have, as before:

$$D_i \iint_{r_i} \nabla^2u(x, y)\, dx\, dy - \sigma_i \iint_{r_i} u(x, y)\, dx\, dy$$
$$+ \Sigma_i \iint_{r_i} S(x, y)\, dx\, dy = 0, \qquad r = 1, 2, 3. \qquad (15)$$

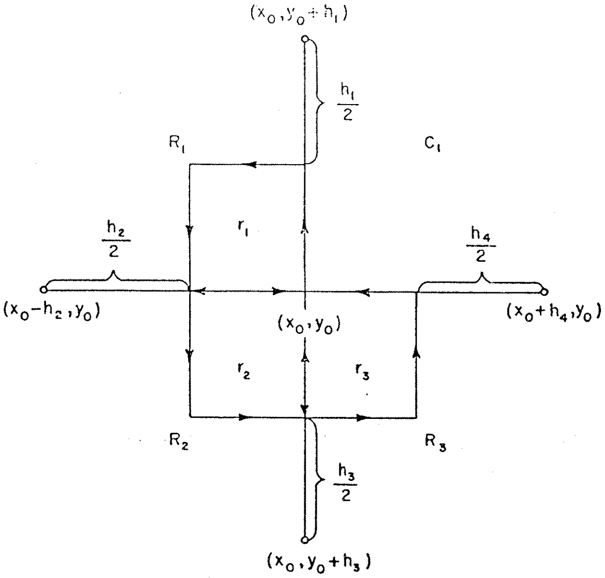

Fig. 3.

Using Green's theorem and the continuity of current, we have analogously after adding the three expressions above

$$\sum_{i=1}^{3}\left\{D_i \int_{\bar{a}_i} \frac{\partial u}{\partial n}(x, y)\, ds - \sigma_i \iint_{r_i} u(x, y)\, dx\, dy \right. \qquad (16)$$
$$\left. + \Sigma_i \iint_{r_i} S(x, y)\, dx\, dy\right\} = 0.$$

The numerical approximations of (11) and (12) are used on all terms of (16) with the exception of the integrals:

$$D_1 \int_{y_0}^{y_0+h_1/2} \frac{\partial u}{\partial x}(x_0 , y)\, dy,$$

and

$$D_3 \int_{x_0}^{x_0+h_4/2} \frac{\partial u}{\partial y}(x, y_0)\, dx.$$

From (3) we have that

$$\frac{\partial u}{\partial n}(x, y)\bigg|_{\gamma_i} = -\frac{c}{D}u(x, y)$$

at all points of γ_i. Thus,

$$D_1 \int_{y_0}^{y_0+h_1/2} \frac{\partial u}{\partial x}(x_0 , y)\, dy = -c \int_{y_0}^{y_0+h_1/2} u(x_0 , y)\, dy,$$

and

$$D_3 \int_{x_0}^{x_0+h_4/2} \frac{\partial u}{\partial y}(x, y_0)\, dx = -c \int_{x_0}^{x_0+h_4/2} u(x, y_0)\, dx;$$

and these integrals are approximated numerically, respectively, by $-c(h_1/2)u(x_0, y_0)$, and $-c(h_4/2)u(x_0, y_0)$. The final numerical approximation to (14) can then be written in the form

$$\frac{D_1h_2}{2h_1}u(x_0, y_0 + h_1) + \left(\frac{D_1h_1 + D_2h_3}{2h_2}\right)u(x_0 - h_2, y_0)$$
$$+ \left(\frac{D_2h_2 + D_3h_4}{2h_3}\right)u(x_0, y_0 - h_3)$$
$$+ \frac{D_3h_3}{2h_4}u(x_0 + h_4, y_0)$$
$$- u(x_0, y_0)\left\{\frac{D_1h_2}{2h_1} + \frac{D_1h_1 + D_2h_3}{2h_2}\right.$$
$$+ \frac{D_2h_2 + D_3h_4}{2h_3} + \frac{D_3h_3}{2h_4} + c\,\frac{h_1 + h_4}{2}$$
$$\left. + \frac{\sigma_1h_1h_2 + \sigma_2h_2h_3 + \sigma_3h_3h_4}{4}\right\}$$
$$+ S(x_0, y_0)\cdot$$
$$\left(\frac{\Sigma_1h_1h_2 + \Sigma_2h_2h_3 + \Sigma_3h_3h_4}{4}\right) = 0. \qquad (17)$$

Again we remark that the coefficient of $u(x_0, y_0)$ is, in absolute value, greater than the sum of the coefficients of the other $u(x, y)$'s since c and the σ's are positive.

Although the slow group flux ϕ theoretically is not defined interior to the regions C_i, a trivial five-point formula with coefficients all zero is, nevertheless, designated for ϕ in the discrete problem.

C. Properties of the Matrices Defined by the Discrete Approximations

If the total number of interior mesh points is N, then with the previous section for each group we have N linear equations in N unknowns. As yet, no systematic numbering or ordering of the N unknowns for each group has been given, but since the heterogeneous reactor problem has been stated only for rectangular regions, the most natural ordering of the points is to number the unknowns by consecutive rows, as shown by the example in Fig. 4. For the example, both ϕ, ψ vanish on Γ. Thus, having ordered the unknowns, we may write the N equations in N unknowns for the two-group problem in the matrix form:

$$\begin{aligned} A_1\vec{\phi} &= B_1\vec{\psi} \\ A_2\vec{\psi} &= \eta B_2\vec{\phi}. \end{aligned} \qquad (18)$$

Here A_1, A_2, B_1, B_2 are $N \times N$ real matrices, and the column vectors $\vec{\phi}$, $\vec{\psi}$ represent, respectively, the slow group flux and the fast group flux for the discrete case.

If in the problem to be solved there exist regions C_i, then since the slow flux is *not* defined in the interior of each C_i while the fast flux is defined there, we see that the total number of *essential* unknowns in the discrete numerical approximation for the slow group flux is less than that for the fast group flux. In such problems we see by the closing remark of Section II-B that B_1 must have rows of zeros and is, therefore, singular. For mathematical convenience then, we shall renumber the unknowns for the slow group by skipping such points within regions C_i, and the matrix equations are now

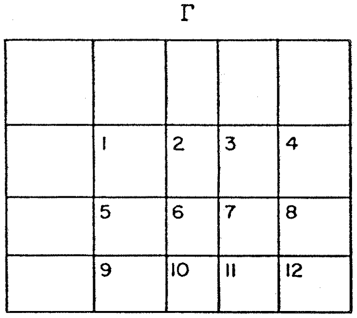

Fig. 4.

$$\tilde{A}_1\vec{\Phi} = \tilde{B}_1\vec{\Psi}; \qquad A_2\vec{\psi} = \eta B_2\vec{\phi} \qquad (19)$$

where $\tilde{A}_1$, $\tilde{B}_1$ are $N' \times N'$ real matrices with $N' \leq N$.

Theorem 1: For all heterogeneous reactors of the type described in Section II-A and for all positive values of the mesh increments in both x, y directions, we have:

> $\tilde{A}_1$, A_2 are symmetric and positive definite, the diagonal elements of both being positive, whereas the off-diagonal elements of $\tilde{A}_1$, A_2 are nonpositive.
>
> B_1, B_2 are diagonal matrices, $\tilde{B}_1$ having only positive diagonal elements, whereas B_2 has nonnegative diagonal elements.

Proof: Everything except the positive definite nature of $\tilde{A}_1$ and A_2 follows immediately from the discussion in the previous section. Let D_1 and D_2 be positive diagonal matrices such that $D_1\tilde{A}_1$ and D_2A_2 are matrices with one's on their main diagonals. Thus, we may write $D_1\tilde{A}_1 = I - M_1$, and $D_2A_2 = I - M_2$; M_1 is an $N' \times N'$ matrix with zero diagonal elements, and M_2 is an $N \times N$ matrix with zero diagonal elements. Furthermore, M_1 and M_2 have all their elements nonnegative and, by virtue of the preceding section, the sum of the elements in any row of M_1, M_2 is less than unity. Thus, by Corollary 1, (Section IV-A) the spectral norms of M_1, M_2 are less than unity. Thus[2] $\tilde{A}_1$, A_2 are positive definite. (*QED*).

We shall refer to the matrices M_1, M_2 as the *initial iteration matrices*.

Theorem 2: $\tilde{A}_1$ and A_2 are ordered consistently and satisfy property (A) in the sense of Young [1].

Proof: This is an easy consequence of the definitions given by Young [1].

The results of the above theorems will be useful when the iterative procedure is discussed in Section III.

D. The Well-set Nature of the Discrete Case for Heterogeneous Reactors

In the initial statement of the reactor problem for the continuous case, we stated that we sought the solution

[2] See Young, [1], p. 94.

pair $\phi(x, y)$, $\psi(x, y)$ corresponding to the smallest (in modulus) eigenvalue. For heterogeneous reactors, it has not been shown for the steady-state case, with any number of groups, that the smallest (in modulus) eigenvalue for the continuous (space) case is *positive* and has multiplicity one.

For the discrete space case, with or without time dependence, the problem is considerably easier to answer; it is known [2] that for two groups, or more generally for n groups, for every set of positive mesh increments the smallest eigenvalue in modulus is positive, has multiplicity one, and the corresponding neutron fluxes can be taken to have positive components. The proof holds in any dimension. With this fact, we can actually *prove* that the iterative procedure to be described converges. Thus, by virtue of the fact that in this paper we are interested in solutions obtained from computing machines we are guaranteed of a well-set problem for the machines'

III. The Iterative Procedure Used to Solve the Discrete Problem

A. The Power Method of Solving the Eigenvalue Problem in the Discrete Case

From Section II-B, the matrix equations to be solved are

$$\begin{cases} A_1\vec{\phi} = B_1\vec{\psi} \\ A_2\vec{\psi} = \eta B_2\vec{\phi}, \end{cases} \tag{20}$$

where all matrices are $N \times N$. By Theorem 1), A_2 is positive definite and therefore nonsingular. We may then reduce the above set of equations to the single equation

$$\frac{1}{\eta} A_1\vec{\phi} = B_1 A_2^{-1} B_2\vec{\phi} \equiv L\vec{\phi}. \tag{21}$$

By definition, the product $L = B_1 A_2^{-1} B_2$ is an $N \times N$ matrix. Since the $N \times N$ diagonal matrix B_1 has diagonal elements zero and, therefore, entire rows of zero elements (corresponding to mesh points interior to the regions C_i, if they exist), the same will be true of the product matrix L. As in Section II-B, we now may renumber our unknowns for the slow-group flux, giving

$$\frac{1}{\eta}\tilde{A}_1\vec{\Phi} = \tilde{L}\vec{\Phi}. \tag{22}$$

By Theorem 1, $\tilde{A}_1$ is positive definite and thus nonsingular. If we denote the product matrix $\tilde{A}_1^{-1}\tilde{L}$ by T, (22) can be written in the form:

$$\frac{1}{\eta}\vec{\Phi} = T\vec{\Phi}. \tag{23}$$

To solve the above eigenvalue problem, we use the well-known power method, whose convergence is assured by the well-set nature of the discrete problem.

Theorem 3: If $\vec{\Phi}_0$ is an arbitrary vector with positive components, define

$$\vec{\Phi}_{i+1} = \frac{\vec{S}_i}{\lambda_i},$$

where $T\vec{\Phi}_i = \vec{S}_i$, and

$$\lambda_i = \frac{(\vec{S}_i\ , \vec{S}_i)}{(\vec{S}_i\ , \vec{\Phi}_i)}, \quad \text{for} \quad i = 0, 1, 2, \cdots .$$

Then, the sequence $\{\vec{\Phi}_i\}_{i=0}^{\infty}$ converges to the discrete slow-group flux corresponding to the smallest (in modulus) eigenvalue η. Moreover, $\eta = \lim_{i\to\infty} (1/\lambda_i) > 0$, and if $\lim_{n\to\infty} \vec{\Phi}_n \equiv \vec{\Phi}$, then $\vec{\Phi}$ has positive components.

Proof: The power method clearly converges since the smallest (in modulus) eigenvalue η has multiplicity one, on the basis of the results of Birkhoff and Varga [2]. We also know from [2] that $\eta = \lim_{n\to\infty} 1/\lambda_i$ is positive, and since the product matrix T is nonnegative and semi-transitive [2], each vector $\vec{\Phi}_i$ has positive components, and therefore, the sequence converges to a vector with positive components.

The successive reciprocals $1/\lambda_i \equiv \eta_i$ are called the successive eigenvalue estimates of η. The iteration defined in terms of (23) can be interpreted via (20). Specifically, let $\eta_0 > 0$ and $\vec{\Phi}_0$, a vector with positive components, be initial estimates of the smallest eigenvalue and the corresponding slow-group flux vector of the discrete problem. Solve the second matrix equation of (20) for $\vec{\psi}_1$. With $B_1\vec{\psi}_1$, solve for $\vec{\phi}_1$ in the first matrix equation of (20). With $\vec{\phi}_1$, η_0, $\vec{\phi}_0$, form

$$\eta_1 = \eta_0\left(\frac{\vec{\phi}_1\ , \vec{\phi}_0}{\vec{\phi}_1\ , \vec{\phi}_1}\right).$$

Now we are in a position to repeat the process with η_1, $\vec{\phi}_1$ replacing η_0, $\vec{\phi}_0$, respectively. By Theorem 3, we are assured that $\lim_{i\to\infty} \vec{\phi}_i$, $\lim_{i\to\infty} \vec{\psi}_i$, and $\lim_{i\to\infty} \eta_i$ exist, the limits giving rise to the sought-for solutions of the discrete form of the two-group diffusion problem for a heterogeneous reactor.

B. The Successive Overrelaxation Method [1], [15]

In the previous section, we found it necessary, in order to find the largest of the quantities $1/\eta$ by the power method, to solve intermediate matrix equations of the form:

$$\tilde{A}_1\vec{\Phi} = \vec{k}_1\ , \quad \text{and} \quad A_2\vec{\psi} = \vec{k}_2\ , \tag{24}$$

where $\tilde{A}_1$, A_2, $\vec{k}_1$, $\vec{k}_2$ are assumed to be known. The method that is used in the machine code *QED*-1 to solve these matrix equations is the Young-Frankel method [1], [15]. If we first take the matrix equation $A_2\vec{\psi} = \vec{k}_2$, then, by Section II-B, there is a positive diagonal matrix D_2 such that $D_2A_2 = I - M_2$, where M_2 is a matrix whose diagonal elements are zero and whose off-diagonal elements are nonnegative. From $A_2\vec{\psi} = \vec{k}_2$, we have

$$D_2A_2\vec{\psi} = (I - M_2)\vec{\psi} = \vec{\psi} - M_2\vec{\psi} = D_2\vec{k}_2\ , \tag{25}$$

or

$$\vec{\psi} = M_2\vec{\psi} + \vec{f}, \quad \text{where} \quad \vec{f} = D_2\vec{k}_2\ . \tag{26}$$

If $M_2 = \| m_{i,j}^{(2)} \|$ and $\vec{f}$ has components f_i, we may write (26) as

$$\psi_i = \sum_{j=1}^{i-1} m_{i,j}^{(2)} \psi_j + \sum_{j=i+1}^{N} m_{i,j}^{(2)} \psi_j + f_i \,, \qquad i = 1, 2, \cdots, N. \tag{27}$$

The iterative method of successive overrelaxation is defined as

$$\psi_i^{(n+1)} = \omega_2 \left\{ \sum_{j=1}^{i-1} m_{i,j}^{(2)} \psi_j^{(n+1)} + \sum_{j=i+1}^{N} m_{i,j}^{(2)} \psi_j^{(n)} + f_i - \psi_i^{(n)} \right\} + \psi_i^{(n)}, \qquad i = 1, \cdots, N. \tag{28}$$

Similarly, for the first matrix equation of (24),

$$\Phi_i^{(n+1)} = \omega_1 \left\{ \sum_{j=1}^{i-1} m_{i,j}^{(1)} \Phi_j^{(n+1)} + \sum_{j=i+1}^{N'} m_{i,j}^{(1)} \Phi_j^{(n)} + g_i - \Phi_i^{(n)} \right\} + \Phi_i^{(n)}, \qquad i = 1, 2, \cdots, N', \tag{29}$$

where $D_1 \tilde{A}_1 = I - M_1$, and $M_1 = \| m_{i,j}^{(1)} \|$, and $\vec{g} = D_1 \vec{k}_1$. The parameters ω_1, ω_2 are called, respectively, the successive overrelaxation factors for the slow-group flux and the fast-group flux. The convergence of these iterative procedures is guaranteed by:

Theorem 4: The iterative procedures of (28) and (29) converge for any ω_1, ω_2 with $1 \leq \omega_1, \omega_2 < 2$. Furthermore, the optimum rates of convergence for these iterative procedures are obtained by selecting ω_1 and ω_2 according to the formulas

$$\omega_1 = \frac{2}{1 + \sqrt{1 - [\bar{\mu}(M_1)]^2}} \,,$$

$$\omega_2 = \frac{2}{1 + \sqrt{1 - [\bar{\mu}(M_2)]^2}} \,,$$

where $\bar{\mu}(M_1)$, $\bar{\mu}(M_2)$ are, respectively, the spectral norms of the definition in Section IV-B, for the matrices M_1, M_2.

Proof: From Young [1], to prove this theorem, it is sufficient to prove that A_1, A_2 are symmetric and positive definite, satisfy property (A), and are consistently ordered. But this is known from Theorems 1 and 2.

A priori, it is just as difficult to estimate ω_1 and ω_2 as it is to estimate $\bar{\mu}(M_1)$ and $\bar{\mu}(M_2)$ for a general discrete problem, so that Theorem 4, by itself, offers little to the problem of estimating the ω's. To make the problem even more difficult, it is known that to achieve an optimum rate of convergence it is necessary that the ω's be chosen with great care. This has been demonstrated both theoretically [1] and experimentally [14]. In Section IV, we shall show how this factor can be estimated accurately by a theoretically sound technique, whose practical application results in an efficient means of estimating the factors ω.

C. Inner and Outer Iterations

Two remarks concerning the iterative scheme of Section III-B are now in order. First, an *infinite* number of iterations of the type in (28) and (29) are required, in general, to obtain exact answers to the matrix equations of (24). Second, the number of iterations required to achieve a certain accuracy in (28) and (29) is very much dependent on the initial flux estimates $\vec{\psi}^{(0)}$, $\vec{\Phi}^{(0)}$, a remark that will be useful in Section V.

The numerical solution of the discrete case is obtained by forming sequence of vectors $\{\vec{\Phi}_i\}_{i=0}^{\infty}$ of Theorem 3 and truncating this sequence after a certain number of steps in order to achieve a certain accuracy. By Section III-B, we see that to form each new $\vec{\Phi}_i$ an infinite number of iterations by the successive overrelaxation method would be required to solve (24). This situation is replaced by a more practical one in which only a *finite* number of iterations are performed in each group. The iterations in each group are called *inner iterations.*[3] When sufficiently many inner iterations have been performed in each group, the new eigenvalue estimate η_{i+1} is calculated from the inner product:

$$\eta_{i+1} = \eta_i \frac{(\vec{\Phi}_i \,, \vec{\Phi}_{i+1})}{(\vec{\Phi}_{i+1} \,, \vec{\Phi}_{i+1})} . \tag{30}$$

Also, the quantities $\bar{\eta}_{i+1}$ and $\underline{\eta}_{i+1}$ are calculated from

$$\begin{aligned} \underline{\eta}_{i+1} &= \eta_i \min_j \left\{ \frac{(\vec{\Phi}_i)_j}{(\vec{\Phi}_{i+1})_j} \right\} , \\ \bar{\eta}_{i+1} &= \eta_i \max_j \left\{ \frac{(\vec{\Phi}_i)_j}{(\vec{\Phi}_{i+1})_j} \right\} , \end{aligned} \tag{31}$$

where $(\vec{\Phi}_i)_j$ is the j-th component of $\vec{\Phi}_i$. The transition from η_i, $\vec{\Phi}_i$ to η_{i+1}, $\vec{\Phi}_{i+1}$ is called an *outer iteration.*

The pairs of numbers $\{\underline{\eta}_i, \bar{\eta}_i\}_{i=1}^{\infty}$ are pair-wise nested [2] in the sense of Corollary 2, Section IV-A, if sufficiently many inner iterations are performed in each group. Furthermore, it is obvious that $\underline{\eta}_i \leq \eta_i \leq \bar{\eta}_i$ and all three are precisely equal only if $\vec{\Phi}_i$ is the unique slow-group flux solution of (20). Thus,

$$E_i = \frac{\bar{\eta}_i - \underline{\eta}_i}{\eta_i} \tag{32}$$

is a measure of the accuracy of $\vec{\Phi}_i$ as an eigenvector in (24).

Definition 1: If $\vec{X}$ has components x_i, $i = 1, 2, \cdots, n$, then $\| X \| \equiv \sum_{i=1}^{n} | x_i |$ is called the *norm* of X.

Definition 2: $\vec{R}_{m+1}^{(1)} \equiv \vec{\psi}_{m+1} - \vec{\psi}_m$, $\vec{R}_{m+1}^{(2)} = \vec{\Phi}_{m+1} - \vec{\Phi}_m$ are called, respectively, the residual vectors after m iterations for the fast group flux and the slow group flux.

It is clear that the values $\underline{\eta}_i$, η_i, $\bar{\eta}_i$ depend on the *number* of inner iterations performed in each group. If too few inner iterations are performed, there is the possibility of a type of *pseudo-convergence* where E_i becomes very small without $\vec{\Phi}_i$ being nearly an eigenvector of (24). To circumvent this difficulty, a positive number ϵ is prescribed, where ϵ is associated with the desired accuracy

[3] Much of the terminology which accompanies the numerical solution of multigroup diffusion problems was introduced via the Cuthill code.

of the E_i's, and enough inner iterations m_1, m_2 are performed in each group so that

$$\| \vec{R}^{(1)}_{m_1} \| \leq \epsilon \| \vec{R}^{(1)}_1 \|, \qquad \| \vec{R}^{(2)}_{m_2} \| \leq \epsilon \| \vec{R}^{(2)}_1 \|.$$

Here m_1, m_2 are functions of the number of outer iterations. The problem is terminated when

$$E_i \leq 2\epsilon^2.$$

IV. Computation of the Successive Overrelaxation Factor for Each Group

A. Determination of the Spectral Norm of the Matrices M

The essential fact used in the determination of the optimum overrelaxation factor ω is that $N \times N$ matrices $M = \| m_{i,j} \|$ of Section III have *nonnegative* elements. We begin with:

Definition 3: The $n \times n$ matrix $A = \| a_{i,j} \|$ is a Perron matrix $\langle = \rangle$ $a_{i,j} > 0$ for all $i, j = 1, \cdots, n$.

The following theorem is known [4].

Theorem 5: If A is a Perron matrix, then the largest (in modulus) eigenvalue of A is positive has multiplicity one and its corresponding eigenvector has components which may be taken to be positive.

The particular eigenvalue given by Theorem 5 is called the *Perron root of A* and is denoted by λ_A [3], [4].

Lemma 1: if A is a Perron matrix,

$$\vec{x} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \text{and} \quad A\vec{x} = \vec{y}, \quad \text{and}$$

$$\overline{m} = \max_i (y_i), \; \underline{m} = \min_i (y_i),$$

then $\underline{m} \leq \lambda_A \leq \overline{m}$.

Proof: We include the proof for completeness. It is clear that $y_i = \sum_{j=1}^n a_{i,j}$ is the sum of the elements of the i-th row of A, and $\overline{m}$, $\underline{m}$ are, respectively, the greatest, least of the sums of the elements of each row of A. If $\vec{v}$ is the eigenvector associated with λ_A, then $v_i > 0$ for $i = 1, 2, \cdots, n$, and

$$\sum_{j=1}^n a_{i,j} v_j = \lambda_A v_i \; .$$

If we normalize v so that $\max_i v_i = 1$, we have for at least one value of the index i, say r, that

$$\lambda_A = \sum_{j=1}^n a_{r,j} v_j \leq \sum_{j=1}^n a_{r,j}$$

since the $a_{i,j} > 0$, and, therefore,

$$\lambda_A \leq \sum_{j=1}^n a_{r,j} \leq \max_i \sum_{j=1}^n a_{i,j} = \overline{m}.$$

The other inequality follows similarly. *QED.*

Theorem 6: If A is a Perron matrix, and $\vec{u}$ is an arbitrary vector with positive components, then

a)
$$\min_i \left[\frac{\sum_{j=1}^n a_{i,j} u_j}{u_i} \right] \leq \lambda_A \leq \max_i \left[\frac{\sum_{j=1}^n a_{i,j} u_j}{u_i} \right]$$

b)
$$\max_{\vec{u} \epsilon R} \left\{ \min_i \left[\frac{\sum_{j=1}^n a_{i,j} u_j}{u_i} \right] \right\} = \lambda_A = \min_{\vec{u} \epsilon R} \left\{ \max_i \left[\frac{\sum_{j=1}^n a_{i,j} u_j}{u_i} \right] \right\},$$

where R is the set of all vectors $\vec{u}$ with positive components.

Proof: Part a) is a special case of results by Stein [5] and Collatz [6]. If B) is the positive diagonal matrix

$$\begin{bmatrix} \frac{1}{u_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{u_n} \end{bmatrix},$$

then $C = BAB^{-1}$ has the same eigenvalues as A and, moreover, C is a Perron matrix. Applying Lemma 1 to the matrix C, we have part a). Thus

$$\max_{\vec{u} \epsilon R} \left\{ \min_i \left[\frac{\sum_{j=1}^n a_{i,j} u_j}{u_i} \right] \right\} \leq \lambda_A \leq \min_{\vec{u} \epsilon R} \left\{ \max_i \left[\frac{\sum_{j=1}^n a_{i,j} u_j}{u_i} \right] \right\},$$

and since the eigenvector $\vec{v}$ of A, associated with λ_A, is an element of R, then only equality is possible, proving part b). *QED.*

element of R, then only equality is possible, proving b).

This min-max nature of Perron roots has been utilized in a recent paper of Bellman [3].

Definition 4: The *spectral norm* $\bar{\mu}(B)$ of the $n \times n$ matrix B is given by

$$\bar{\mu}(B) = \max_{k=1,2,\cdots,n} | \lambda_k |, \text{ where } \lambda_k \text{ is an eigenvalue of } B.$$

Corollary 1: If the $n \times n$ matrix $B = \| b_{i,j} \|$ has nonnegative elements and $\vec{u}$ is an arbitrary vector with positive components, then

$$\min_i \left(\frac{\Sigma b_{i,j} u_j}{u_i} \right) \leq \bar{\mu}(B) \leq \max_i \left(\frac{\Sigma b_{i,j} u_j}{u_i} \right).$$

Proof: By replacing the zero elements of B by ϵ, $\epsilon > 0$, the new $n \times n$ matrix $B^* = \| b^*_{i,j} \|$ is a Perron matrix, and for $\vec{u}$ a positive vector,

$$\min_i \left(\frac{\Sigma b^*_{i,j} u_j}{u_i} \right) \leq \lambda_{B^*} \leq \max_i \left(\frac{\Sigma b^*_{i,j} u_j}{u_i} \right).$$

Keeping the positive vector $\vec{u}$ fixed, we have by Theorem 5 that $\lambda_B{}^* = \bar{\mu}(B^*)$ for every $\epsilon > 0$, and clearly $\lim_{\epsilon \downarrow 0} \bar{\mu}(B^*) = \mu(B)$. Letting $\epsilon \downarrow 0$, we have the result.

Corollary 2: If the $n \times n$ matrix $B = \| b_{i,j} \|$ has nonnegative elements and no row of B has all zero elements and $\vec{u}$ is an arbitrary vector with positive components, then let $\vec{u}_k = B^k\vec{u}$, $k = 0, 1, 2, \cdots$. If $\bar{m}_k$, $\underline{m}_k$ represent, respectively, the upper and lower bounds on $\bar{\mu}(B)$ by using $\vec{u}_k$, then

$$\underline{m}_k \leq \underline{m}_{k+1} \leq \bar{m}_{k+1} \leq \bar{m}_k .$$

Proof: From the assumptions of this corollary, $\vec{u}_k$ has positive components for all $k = 0, 1, 2, \cdots$. It is sufficient, then, to show that the above inequalities are valid for $k = 0$. By definition,

$$\bar{m}_1 = \max_i \left\{ \frac{\sum_{k=1}^{n} b_{i,k}\left(\sum_{l=1}^{n} b_{k,l}u_l\right)}{\sum_{k=1}^{n} b_{i,k}u_k} \right\}$$

$$= \max_i \left\{ \sum_{k=1}^{n} \left[\frac{b_{i,k}u_k}{\sum_{l=1}^{n} b_{i,l}u_l}\right] \left(\frac{1}{u_k} \sum_{l=1}^{n} b_{k,l}u_l\right) \right\}$$

$$\leq \max_i \left\{ \sum_{k=1}^{n} \left[\frac{b_{i,k}u_k}{\sum_{l=1}^{n} b_{i,l}u_l}\right] \bar{m}_0 \right\} = \bar{m}_0 .$$

Similarly, $\underline{m}_0 \leq \underline{m}_1$. *QED.*

The corollaries above then give us a method for computing both upper and lower bounds on $\bar{\mu}(B)$. With the conditions of Corollary 2), we have

$$\lim_{k\to\infty} \underline{m}_k \leq \lim_{k\to\infty} \bar{m}_k .$$

For the matrix M, which is of interest here, one can conclude [1] that the eigenvalues of M occur in ± pairs, so that for some vector $\vec{u}$ with positive components, $\lim_{k\to\infty} \underline{m}_k < \lim_{k\to\infty} \bar{m}_k$.

Theorem 7: Let M be the $n \times n$ matrix described in Section II, and let $\vec{u}$ be an arbitrary vector with positive components. If $\vec{u}_k = (M + \alpha I)^k\vec{u}$, $k = 0, 1, 2, \cdots$ and if $\bar{m}_k$, $\underline{m}_k$ are the upper, lower bounds, respectively, on $\bar{\mu}(M + \alpha I)$, then for every $\alpha > 0$,

a) $$\underline{m}_k \leq \underline{m}_{k+1} \leq \bar{\mu}(M) + \alpha \leq \bar{m}_{k+1} \leq \bar{m}_k ,$$

and

b) $$\lim_{k\to\infty} \underline{m}_k = \lim_{k\to\infty} \bar{m}_k = \bar{\mu}(M) + \alpha.$$

Proof: The matrix M described in Section II satisfies the conditions of Corollaries 1 and 2, as does $M + \alpha I$ for $\alpha > 0$. Since $\bar{\mu}(M + \alpha I) = \bar{\mu}(M) + \alpha$, part a) follows from Corollaries 1 and 2. Since the eigenvalues of M occur in ± pairs, then the largest (in modulus) eigenvalue of $M + \alpha I$ is positive. By the well-known properties of the power method, the vectors $\vec{u}_k$ converge in direction to an eigenvector corresponding to the largest eigenvalue σ of $M + \alpha I$, and thus part b) follows. *QED.*

This theorem gives then a practical means of obtaining $\bar{\mu}(M)$ to any desired accuracy. The selection of the positive constant α will be discussed in Section IV-C. With the spectral norm $\bar{\mu}(M)$, the optimum overrelaxation factor ω is simply computed by means of the formula

$$\omega = \frac{2}{1 + \sqrt{1 - \bar{\mu}^2(M)}}.$$

B. Use of Weighted Inner Products for Improving Bounds on $\bar{\mu}(B)$

We begin here with some elementary lemmas.

Lemma 2: If B is a symmetric $n \times n$ matrix and the eigenvalues of B are σ_k with

$$\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_n$$

then for $\vec{x} \neq \vec{0}$,

$$\sigma_1 \leq \frac{(B\vec{x}, \vec{x})}{(\vec{x}, \vec{x})} \leq \sigma_n .$$

Proof: The eigenvectors $\vec{v}_k$ of B span $V_n(R)$ and can be chosen to be orthonormal. If

$$\vec{x} = \sum_{k=1}^{n} c_k\vec{v}_k , \quad \text{then} \quad \frac{(B\vec{x}, \vec{x})}{(\vec{x}, \vec{x})} = \frac{\sum_{k=1}^{n} c_k^2\sigma_k}{\sum_{k=1}^{n} c_k^2} ,$$

and since this is a weighted average of the σ_k's, the result follows. *QED.*

By way of terminology, the quotient $(B\vec{x}, \vec{x})/(\vec{x}, \vec{x})$ is called the *Rayleigh quotient*, and the quotient $(B\vec{x}, B\vec{x})/(B\vec{x}, \vec{x})$ is called the *modified Rayleigh quotient*, or the *Schwartz quotient*.

Lemma 3: If B is an arbitrary $n \times n$ matrix, $\vec{x} \neq \vec{0}$ and $(B\vec{x}, \vec{x}) \neq 0$, then

$$\left| \frac{(B\vec{x}, \vec{x})}{(\vec{x}, \vec{x})} \right| \leq \left| \frac{(B\vec{x}, B\vec{x})}{(B\vec{x}, \vec{x})} \right| .$$

If B is symmetric and positive definite with eigenvalues σ_k, where $0 < \sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_n$, then

$$\sigma_1 \leq \frac{(B\vec{x}, \vec{x})}{(\vec{x}, \vec{x})} \leq \frac{(B\vec{x}, B\vec{x})}{(B\vec{x}, \vec{x})} \leq \sigma_n .$$

Proof: The first part is just Cauchy's inequality. The second part is known [7], and follows easily from considerations of weighted averages, as in Lemma 1.

The modified Rayleigh quotient is *not* always a conservative estimate of the largest eigenvalue of a symmetric matrix, as is seen by

Lemma 4: If B is a symmetric $n \times n$ matrix with eigenvalues $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_n$, where $\sigma_n > 0$, then

$(B\vec{x}, B\vec{x})/(B\vec{x}, \vec{x}) \leq \sigma_n$ for all vectors $\vec{x} \neq \vec{0}$ if and only if B is symmetric and positive definite.

Proof: By Lemma 3, we may assume that B is not positive definite, so that $\sigma_1 < 0$. If the eigenvectors of B are $\vec{v}_k$ with $B\vec{v}_k = \sigma_k\vec{v}_k$, let $\vec{x} = \vec{v}_n + \beta\vec{v}_1$, where $0 < \beta < (\sigma_n/-\sigma_n)^{1/2}$.

Then, $$\frac{(B\vec{x}, B\vec{x})}{(B\vec{x}, \vec{x})} = \frac{\sigma_n^2 + \beta^2\sigma_1^2}{\sigma_n + \beta^2\sigma_1} > \sigma_n. \qquad QED.$$

We now return to the determination of the spectral norm of the $n \times n$ iteration matrix M associated with the positive definite and symmetric matrix A of Section II-B. Let D be the positive diagonal matrix

$$D = \begin{bmatrix} \frac{1}{a_{1,1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{a_{n,n}} \end{bmatrix},$$

where $A = \| a_{i,i} \|$, and let $B = DA = I - M$. Since the eigenvalues λ_k of M satisfy $\bar{\mu} \leq \lambda_k \leq \bar{\mu}$, then the eigenvalues ν_k of B satisfy $1 - \bar{\mu} \leq \nu_k \leq 1 + \bar{\mu}$. Let $C = D^{-1/2}BD^{1/2} = D^{1/2}AD^{1/2}$. Thus, C is similar to B, and C is obviously symmetric and positive definite. Its eigenvalues c_k satisfy

$$1 - \bar{\mu} \leq c_k \leq 1 + \bar{\mu}.$$

For an arbitrary vector $\vec{y} \neq \vec{0}$, by Lemma 3, we have

$$1 - \bar{\mu} \leq \frac{(C\vec{y}, \vec{y})}{(\vec{y}, \vec{y})} \leq 1 + \bar{\mu}. \tag{33}$$

Setting $\vec{y} = D^{-1/2}\vec{x}$, we have

$$\frac{(C\vec{y}, \vec{y})}{(\vec{y}, \vec{y})} = \frac{(B\vec{x}, D^{-1}\vec{x})}{(\vec{x}, D^{-1}\vec{x})},$$

since D is symmetric. With $B = I - M$, (33) reduces to

$$-\bar{\mu} \leq \frac{(M\vec{x}, D^{-1}\vec{x})}{(\vec{x}, D^{-1}\vec{x})} \leq \bar{\mu}. \tag{34}$$

If we define the weighted inner product as $[\vec{y}, \vec{x}] \equiv \sum_{i=1}^{n} y_i x_i a_{i,i}$, then (34) becomes

$$-\bar{\mu} \leq \frac{[M\vec{x}, \vec{x}]}{[\vec{x}, \vec{x}]} \leq \bar{\mu}. \tag{35}$$

Although M is not in general *symmetric*, the weighted Rayleigh quotient for M gives rise to a *conservative* estimate of the spectral norm of M. This gives us[4]

Theorem 8: If M is the $n \times n$ iteration matrix described above and $\vec{x}$ is an arbitrary vector with positive components, then

$$0 < \frac{[M\vec{x}, \vec{x}]}{[\vec{x}, \vec{x}]} \leq \bar{\mu}.$$

Furthermore, the eigenvectors of M are orthogonal with respect to the weighted inner product.

As a result of this theorem, the best bounds on the spectral norm $\bar{\mu}$ at the i-th iteration will be

$$\frac{[M\vec{x}_i, \vec{x}_i]}{[\vec{x}_i, \vec{x}_i]} \leq \bar{\mu} \leq \max_i \left\{ \frac{(M\vec{x}_i)_i}{(\vec{x}_i)_i} \right\}.$$

At the present time, we are not using the weighted inner products. In *QED*-1, four iterations $\vec{u}_k = (M + \alpha I)^k \vec{u}_0$ with $\vec{u}_0$ having all components unity are performed, and the final upper and lower bounds on $\bar{\mu}$ are, respectively, $\bar{m}_4$ and $\underline{m}_4$. From the sequence of modified Rayleigh quotients $\{\lambda_k\}_{k=1}^{4}$, a δ^2 procedure [9], [10] is used on the last three estimates in the form:

$$\tilde{\sigma}_4 = \lambda_4 - \frac{(\lambda_3 - \lambda_4)^2}{\lambda_2 - 2\lambda_3 + \lambda_4}.$$

Then $\tilde{\sigma}_4$ is the final estimate of $\bar{\mu}$. From the formula

$$\omega(\beta) = \frac{2}{1 + \sqrt{1 - \beta^2}}, \qquad 0 \leq \beta \leq 1,$$

$\omega(\beta)$ is a 1-1 function of β, so that

$\omega(\underline{m}_4) \leq \omega(\tilde{\sigma}_4) \leq \omega(\bar{m}_4)$, provided $\underline{m}_4 \leq \tilde{\sigma}_4 \leq \bar{m}_4$. Performed in this manner, we have upper and lower bounds on, as well as an estimate of, the optimum successive overrelaxation factor.

C. Selection of the Parameter α

In Section IV-A, a positive parameter α was added to the diagonal of the initial iteration matrix M in order to insure convergence of the successive upper and lower bounds $\bar{m}_i$, $\underline{m}_i$ to the same limit. Although the final estimates for the optimum ω after four iterations are quite insensitive to the value of α used, we include here the analysis on the selection of α for the sake of completeness.

Let the eigenvalues of the $n \times n$ iteration matrix M be σ_j, $j = 1, 2, \cdots, n$, where $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_n$. Since the eigenvalues of M occur in ± pairs, then $\sigma_j = -\sigma_{n+1-j}$, $j = 1, 2, \cdots, n$. Also, the eigenvectors $\vec{v}_j$ of M span $V_n(R)$, so that for *any* vector $\vec{x}$ with positive components we have

$$\vec{x} = \sum_{j=1}^{n} c_j \vec{v}_j. \tag{36}$$

Furthermore, using the fact that the matrix M is nonnegative and semitransitive, it can be shown [2] that some power of $(M + \alpha I)$, $\alpha > 0$ is a Perron matrix, so that $\vec{v}_n$ has positive components and that $c_n > 0$ and $\sigma_{n-1} < \sigma_n$. We have

$$\begin{aligned}(M + \alpha I)^m \vec{x} &= \sum_{j=1}^{n} c_j (\sigma_j + \alpha)^m \vec{v}_j \\ &= c_n (\sigma_n + \alpha)^m \left[\vec{v}_n + \sum_{j=1}^{n-1} \frac{c_j}{c_n} \left(\frac{\sigma_j + \alpha}{\sigma_n + \alpha} \right)^m \vec{v}_j \right].\end{aligned} \tag{37}$$

Definition 5: If

$$r(\alpha) \equiv \max_{j=1,2,\cdots,n-1} \left| \frac{\sigma_j + \alpha}{\sigma_n + \alpha} \right|, \quad \text{where} \quad \alpha > 0,$$

then the *rate of convergence* $R(\alpha)$, as a function of α, is given by

$$R(\alpha) = -\ln r(\alpha).$$

Clearly, $r(\alpha)$ gives a measure of the deviation in direction of the vector $\vec{y}_m = (M + \alpha I)^m \vec{x}$ from $\vec{v}_n$, and $R(\alpha)$ is related to the number of iterations required to make the norm of the term

$$\sum_{j=1}^{n-1} \frac{c_j}{c_n} \left(\frac{\sigma_j + \alpha}{\sigma_n + \alpha} \right)^m \vec{v}_j \quad \text{small.}$$

[4] For a discussion of the orthogonality of eigenvectors with respect to the weighted inner product, see Flanders and Shortley [8].

The following lemmas are easily established.

Lemma 5:

$$r(\alpha) = \frac{\sigma_n - \alpha}{\sigma_n + \alpha} \quad \text{for} \quad 0 < \alpha \leq \tfrac{1}{2}(\sigma_n - \sigma_{n-1}),$$

and

$$r(\alpha) = \frac{(\sigma_{n-1} + \alpha)}{(\sigma_n + \alpha)} \quad \text{for} \quad \alpha \geq \tfrac{1}{2}(\sigma_n - \sigma_{n-1}).$$

Lemma 6:

$$\frac{dR(\alpha)}{d\alpha} = \frac{\sigma_{n-1} - \sigma_n}{(\sigma_n + \alpha)(\sigma_{n-1} + \alpha)} < 0 \quad \text{for} \quad \alpha \geq \tfrac{1}{2}(\sigma_n - \sigma_{n-1}) > 0,$$

and

$$\frac{dR(\alpha)}{d\alpha} = \frac{2\sigma_n}{(\sigma_n + \alpha)(\sigma_n - \alpha)} > 0 \quad \text{for} \quad 0 < \alpha \leq \tfrac{1}{2}(\sigma_n - \sigma_{n-1}).$$

With these lemmas, the proof of the following theorem is trivial.

Theorem 9. The optimum rate of convergence is obtained for

$$\alpha^* = \tfrac{1}{2}(\sigma_n - \sigma_{n-1}) > 0.$$

Furthermore, for any δ with $0 < \delta < \alpha^*$, $R(\alpha^* + \delta) > R(\alpha^* - \delta)$.

Thus, Theorem 9 states that it is better to *overestimate* α, rather than to underestimate α. This is very similar to the results on the selection of the optimum successive overrelaxation factor [1]. For similar use of a parameter in optimizing the rate of convergence of solutions to eigenvalue problems, see [11]–[13].

D. A Numerical Example

To illustrate the technique of estimating the optimum overrelaxation factor described in the previous sections, we consider the following numerical example.

$$A = \begin{bmatrix} 1 & -0.5 & 0 & 0 & 0 \\ -0.5 & 1 & -0.4 & 0 & 0 \\ 0 & -0.4 & 1.3333 & -0.6667 & 0 \\ 0 & 0 & -0.6667 & 10.3333 & -0.6667 \\ 0 & 0 & 0 & -0.6667 & 1.6667 \end{bmatrix}.$$

Matrix A is of the type considered in Theorems 1) and 2). The associated initial iteration matrix M is given by:

$$M = \begin{bmatrix} 0 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0.4 & 0 & 0 \\ 0 & 0.3 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0.4 & 0 \end{bmatrix}.$$

The following numerical results were obtained with $\alpha = 0.1$. The quantities $\underline{\lambda}$, $\bar{\lambda}$ are

Iteration No.	$\underline{\lambda}$	$\bar{\lambda}$	λ	$\tilde{\lambda}$	σ	$\tilde{\sigma}$
1	0.4	1.0	0.720	0.785366	0.70526	0.776471
2	0.636364	0.944444	0.747934	0.767771	0.754661	0.775395
3	0.661224	0.882716	0.765365	0.777045	0.759956	0.771826
4	0.690578	0.853619	0.758357	0.765231	0.762577	0.769602.

respectively, the Perron bounds on the spectral norm $\bar{\mu}(M)$ of M; λ, $\tilde{\lambda}$ are, respectively, the Rayleigh quotient and modified Rayleigh quotient estimates of $\bar{\mu}(M)$; and σ, $\tilde{\sigma}$ are, respectively, the Rayleigh quotient and modified Rayleigh quotient estimates of $\bar{\mu}(M)$ obtained by using the weighted inner product of Section IV-B.

If ν_1, ν_2 are, respectively, the results from applying the δ^2 technique to the last three entries in the columns $\tilde{\lambda}$, $\tilde{\sigma}$, then

$$\nu_1 = 0.771850, \qquad \nu_2 = 0.765923;$$

and if ω_1, ω_2 are the corresponding overrelaxation factors associated with ν_1, ν_2, then

$$\omega_1 = 1.222640, \qquad \omega_2 = 1.217336.$$

For this example, the optimum overrelaxation factor is $\omega_b = 1.217985$.

V. Extrapolation and Renormalization

We have seen in Theorem 3 that the sequence of vectors $\{\vec{\Phi}_i\}_{i=0}^{\infty}$ converges to the discrete slow flux corresponding to the smallest (in modulus) eigenvalue of (24). In order to increase the rate of convergence of this iterative procedure, an extrapolation scheme [11]–[13] is used on the sequence $\{\vec{\Phi}_i\}_{i=0}^{\infty}$.

Theorem 10: Let $\vec{\Phi}_0$ be an arbitrary vector with positive components, let $\eta_0 > 0$, and define recursively:

$$\vec{\Phi}_{i+1}^* = \frac{\nu_{i+1}(1 + \beta)\vec{\Phi}_{i+1} - \beta\eta_i\vec{\Phi}_i^*}{\eta_{i+1}},$$

where

$$T(\eta_i\vec{\Phi}_i^*) = \vec{\Phi}_{i+1},$$

$$\eta_{i+1} = \eta_i \frac{(\vec{\Phi}_{i+1}, \vec{\Phi}_i^*)}{(\vec{\Phi}_{i+1}, \vec{\Phi}_{i+1})},$$

and

$$\nu_{i+1} = \frac{\|\eta_i\Phi_i^*\|}{\|\vec{\Phi}_{i+1}\|}.$$

Then, for

$$\frac{1 + \beta}{\beta} > \frac{\bar{\eta}_{i+1}}{\nu_{i+1}},$$

$\vec{\Phi}_{i+1}^*$ has positive components, $i = 0, 1, 2, \cdots$.

Moreover, $\|\eta_{i+1}\vec{\Phi}_{i+1}^*\| = \|\eta_i\vec{\Phi}_i^*\|$.

Proof: It is sufficient to prove the first part for $i = 0$.

Let $\vec{\Phi}_0$ have components $\phi_j > 0$. Then, $\vec{S}_1 \equiv \nu_1(1 + \beta)\vec{\Phi}_1 - \beta\eta_0\vec{\Phi}_0$ has components:

$$S_j = \nu_1(1 + \beta)\psi_j - \beta\eta_0\phi_j \ .$$

For $\beta \geq 0$, S_j is positive $\langle = \rangle$ $(1 + \beta)/\beta > (1/\nu_1)(\eta_0\phi_j/\psi_j)$, $\psi_j > 0$ since the matrix T has nonnegative elements. But $\eta_0\phi_j/\psi_j \leq \bar{\eta}_1$, so that S_j is positive for all j if $(1 + \beta)/\beta > \bar{\eta}_1/\nu_1$. This proves the first part. The second part follows from definition. *QED.*

The above iteration sequence $\{\vec{\Phi}_i^*\}_{i=0}^{\infty}$ can be thought of as an extrapolation and renormalization performed on the sequence $\{\vec{\Phi}_i\}_{i=0}^{\infty}$. The norm-preserving feature of the definition of $\eta_i\vec{\Phi}_i^*$ is important when the last used fast group flux $\vec{\psi}_0$ is used as an *initial* approximation in the matrix equation:

$$A_2\vec{\psi} = B_2(\eta_i\vec{\Phi}_i^*). \qquad (36)$$

Thus, the norm of the vector of the right-hand side of the equation above is kept constant in the course of the iterations. It is fully expected that using the extrapolation and renormalization scheme with the proper β's will result in a reduction of effort in solving discrete matrix problems, the reductions being as much as 50 per cent.

Bibliography

[1] Young, D. "Iterative Methods for Solving Partial Difference Equations of Elliptic Type," *Transactions of the American Mathematical Society*, Vol. 76 (1954), pp. 92-111.

[2] Birkhoff, G. and Varga, R. S. "Reactor Critically and Nonnegative Matrices," Bettis Atomic Power Division of Westinghouse, Pittsburgh, Pa., Report No. WPAD-166 (1957).

[3] Bellman, R., "On an Iterative Procedure for Obtaining the Perron Root of a Positive Matrix," *Proceedings of the American Mathematical Society*, Vol. 6 (1955), pp. 719-725.

[4] Frobenius, S-B. Preuss. Akad. Wiss. (1908), pp. 471-476.

[5] Stein, P. "A Note on the Bounds of the Real Parts of the Characteristic Roots of a Matrix," *Journal of Research National Bureau of Standards*, Vol. 48 (1952), pp. 106-110.

[6] Collatz, L. "Einschliessungssatz für die charakterisischen Zahlen von Matrizen," *Math. Zeit.*, Vol. 48 (1942), pp. 221-226.

[7] Hildebrand, F. B. *Methods of Applied Mathematics*, New York, Prentice-Hall, Inc., 1952.

[8] Flanders, D. A. and Shortley, G. "Numerical Determination of Fundamental Modes, *Journal of Applied Physics*, Vol. 21 (1950), pp. 1326-1332.

[9] Aitken, A. C. "Studies in Practical Mathematics II. The Evaluation of the Latent Roots and Latent Vectors of a Matrix," *Proceedings of the Royal Society of Edinburgh*, Vol. 57 (1937), pp. 269-304.

[10] Shanks, D. "Non-linear Transformations of Divergent and Slowly Convergent Sequences," *Journal of Mathematics and Physics*, Vol. 34 (1955), pp. 1-42.

[11] Hestenes, M. R. "Iterative Computational Methods," Second Symposium on Applied Mathematics, Interscience (1955), pp. 85-95.

[12] *Determination of Eigenvalues and Eigenvectors of Matrices*, National Bureau of Standards Applied Mathematics Series No. 29 (1953), pp. 89-94.

[13] Bilodeau, G. G. "Extrapolation Techniques for Symmetric Matrices." Bettis Atomic Power Division of Westinghouse, Pittsburgh, Pa., Report No. WPAD-TM-52.

[14] Stark, R. H. "Rates of Convergence in Numerical Solution of the Diffusion Equation," *Journal of Associated Computing Machinery*, Vol. 3 (1956), pp. 29-40.

[15] Frankel, S. "Convergence Rates of Iterative Treatments of Partial Differential Equations," *Mathematical Tables and Aids to Computation*, Vol. 4 (1950), pp. 65-75.