Numerical Determination of Fundamental Modes

Donald A. Flanders and George Shortley*
Argonne National Laboratory, Chicago, Illinois

(Received July 12, 1950)

A convenient and practical method of numerical determination of the fundamental eigenfunction and eigenvalue in a class of linear eigenvalue problems has been developed and applied in two and three dimensions. The method is based on use of a network and of difference equations, but departs from previous methods in that it is not iterative. Rather, a polynomial operator is applied to a trial function just once, to accomplish a determinable degree of reduction in all eigenfunctions other than the fundamental that are contained in the trial function. In the case of the diffusion equation, the polynomial operator is a Tschebyscheff polynomial of a simple averaging operator. It is shown that this operator, when of degree *m*, is "better" than any other polynomial operator of this degree and much "better" than *m* iterations of a simple averaging operator—"better" in the sense of accomplishing to a greater degree the elimination of *all* unwanted eigenfunctions. Techniques for the use of computing equipment for application of the polynomial operator are discussed. By orthogonalization, the method can be applied to modes other than the fundamental.

INTRODUCTION

In connection with work on complex problems occurring in certain applications of diffusion theory, a convenient and practical method of numerical determination of the fundamental eigenfunction and eigenvalue in linear eigenvalue problems has been developed and applied in two and three dimensions.[1]

While the methods we have developed are applicable to a wide class of systems of linear partial differential equations and a wide variety of linear boundary conditions, we feel that the new ideas and techniques will be most useful if they are first presented in the simplest possible setting. Hence, in this paper we shall first consider in detail the numerical determination of the lowest eigenvalue and corresponding fundamental eigenfunction of the equation

$$\Delta u+\alpha u=0, \tag{1}$$

which occurs in the theories of heat conduction, diffusion, vibration, and elsewhere. For simplicity we shall fix attention on a connected two-dimensional region bounded by straight-line segments parallel to the x- and y-coordinate axes, and shall consider the simplest boundary condition

$$u=0 \quad \text{on the boundary.} \tag{2}$$

The methods discussed can be generalized so as to apply to more general linear differential equations of elliptic type, to systems of such equations, to more complex geometrical configurations, to more general linear boundary and interface conditions and to more dimensions. Systems of equations will be discussed briefly in the last section.

We replace the continuum by a square lattice of mesh h, and equation (1) by a difference equation, in the usual way. Again for simplicity we assume that the dimensions are such that the boundaries fall along rows of lattice points, and that a lattice point falls at the origin of coordinates. We further assume that the mesh is so small that every pair of interior points can be joined by a polygonal path consisting of segments of the network lying wholly in the interior of the region; we shall say that the *interior* points are *connected* by the net. We denote the lattice points by P_{ij} and the corresponding values of u by u_{ij}, where $i=x/h$, $j=y/h$.

If we use the notation ωu_{ij} to represent the average of the values of u at the four nearest neighbors of the point P_{ij}, i.e.,

$$\omega u_{ij}=\tfrac{1}{4}(u_{i+1,\,j}+u_{i-1,\,j}+u_{i,\,j+1}+u_{i,\,j-1}),$$

the second-difference approximation[2] to Δu at the point P_{ij} is

$$\Delta u_{ij}\approx\frac{\omega u_{ij}-u_{ij}}{\tfrac{1}{4}h^2}.$$

The difference equation that approximates (1) is then

$$\omega u_{ij}-u_{ij}+\tfrac{1}{4}h^2\alpha u_{ij}=0,$$

or

$$\omega u_{ij}=\lambda u_{ij}, \tag{3}$$

where

$$\lambda=1-\tfrac{1}{4}h^2\alpha. \tag{4}$$

* Operations Research Office, The Johns Hopkins University, on leave of absence from The Ohio State University; consultant at the Argonne National Laboratory.

[1] For a review of previous work on numerical methods in two and three dimensions (mostly work on boundary-value rather than eigenvalue problems) see Thomas J. Higgins, *Numerical Methods of Analysis in Engineering* (The Macmillan Company, New York, 1949), Chapter 10. This chapter contains an extensive bibliography.

[2] Nothing we say in this paper will be changed except in minor details if one chooses to use the higher-order approximation to Δu that is obtained if we define the averaging operator ω as

$$\omega u_{ij}=\tfrac{1}{5}(u_{i+1,\,j}+u_{i-1,\,j}+u_{i,\,j+1}+u_{i,\,j-1})+\tfrac{1}{20}(u_{i+1,\,j+1}+u_{i-1,\,j+1}+u_{i+1,\,j-1}+u_{i-1,\,j-1}),$$

in which case

$$\Delta u_{ij}\approx(\omega u_{ij}-u_{ij})/\tfrac{3}{10}h^2$$

and (4) becomes

$$\lambda=1-\tfrac{3}{10}h^2\alpha.$$

This expression for Δu_{ij} is valid for any seventh-order polynomial passing through the nine points, whereas the expression in the text is valid for a cubic polynomial passing through the five points. For a detailed discussion of approximations to the Laplacian, see Shortley and Weller, J. App. Phys. 9, 345 (1938).

The highest eigenvalue λ_1 of (3) will give an approximation to the lowest eigenvalue α_1 of (1) according to the relation (4), and the corresponding eigenvector of (3) will approximate to the fundamental eigenfunction of (1). We shall not discuss in this paper the accuracy of these approximations; it is well known that the solutions of the difference equation converge to solutions of the differential equation as $h \to 0$.

We shall consider the function u_{ij} to be defined only at interior points of the region. At a point adjacent to the boundary ωu_{ij} then will be taken to represent $\frac{1}{4}$ the sum of the values at the three, two, or one neighboring interior points, since one or more zero boundary values occur in the averaging process. We assume that there are N interior points, and for convenience number them from 1 to N, replacing the double index i,j by a single index k. The set of values u_k will then be denoted by a column vector $\mathbf{u}=(u_k)$. For the sake of brevity two components of $\mathbf{u}$ will be called neighboring if they represent values of u at neighboring points of the lattice.

Let $\omega_{kl}=\frac{1}{4}$ or 0 according as u_k and u_l are neighbors or not, and let ω denote the square matrix with elements ω_{kl}. The matrix equation

$$\omega\mathbf{u}=\lambda\mathbf{u} \tag{5}$$

then represents the result of incorporating the boundary conditions into the system of Eqs. (3). Thus Eq. (5) is the complete finite-difference analog of the analytic system (1), (2). It is the algebraic eigenvalue problem set by (5) with which we shall be concerned henceforth.

BASIC PROPERTIES OF THE SOLUTIONS OF THE MATRIX EQUATION

The general theory of equations such as (5) is well known. We set down here such general facts as we shall need, and we derive certain further useful properties specific to the particular equation.

The matrix ω is real and symmetric. Hence, it has N real eigenvalues and a complete orthonormal system of eigenvectors.[3] We denote the eigenvalues by

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N,$$

and the corresponding eigenvectors of an arbitrary (but fixed) complete orthonormal system by

$$\mathbf{u}^1, \mathbf{u}^2, \cdots, \mathbf{u}^N. \tag{6}$$

(Since $-\mathbf{u}$ is an eigenvector belonging to the eigenvalue λ if $\mathbf{u}$ is, and is normalized if $\mathbf{u}$ is, we shall assume throughout that whenever $\mathbf{u}$ denotes an eigenvector, the maximum of the absolute values of the components of $\mathbf{u}$ is equal to the value of some component u_k of $\mathbf{u}$.)

The averaging operation performed by ω cannot increase the maximum of the absolute values of the components of any eigenvector $\mathbf{u}$, so that every $|\lambda_n| \leq 1$. In fact one can readily see that only the inequality can hold because of the zero boundary condition. (It is clear that the $\leq$ sign will hold for more general operators than ω provided the corresponding matrix contains no negative elements and the sum of the elements in every row is ≤ 1, while the inequality may be asserted if in addition the sum of the elements of some row is <1.)

The highest eigenvalue, λ_1, *is non-degenerate* (i.e., $\lambda_1 > \lambda_2$ or, equivalently, there are not two linearly independent eigenvectors with eigenvalue λ_1), *and* $\mathbf{u}^1$ *is everywhere positive.* These facts may be proved by using the variational analog of the eigenvalue problem, which may be stated thus: If $\mathbf{u}^0$ is a normalized eigenvector of (5) with eigenvalue λ_0, then $\mathbf{u}^0$ is a vector that gives the quadratic form

$$(\mathbf{u}, \omega\mathbf{u}) \equiv \sum_{k,l} \omega_{kl} u_k u_l$$

the stationary value λ_0, subject to the normalization condition

$$(\mathbf{u}, \mathbf{u}) \equiv \sum_k u_k^2 = 1;$$

and conversely.

Let $\mathbf{v}$ be any normalized eigenvector belonging to λ_1 (i.e., with eigenvalue λ_1), and let $\mathbf{w}=(|v_k|)$. Then $\mathbf{w}$ is normalized, and we shall show that it is also an eigenvector belonging to λ_1. For since no $\omega_{kl}<0$ it follows that $\omega_{kl}w_k w_l \geq \omega_{kl} v_k v_l$ in every case, and hence that $(\mathbf{w}, \omega\mathbf{w}) \geq \lambda_1$. Since λ_1 is the maximum value that $(\mathbf{u}, \omega\mathbf{u})$ can assume for any normalized vector $\mathbf{u}$, we

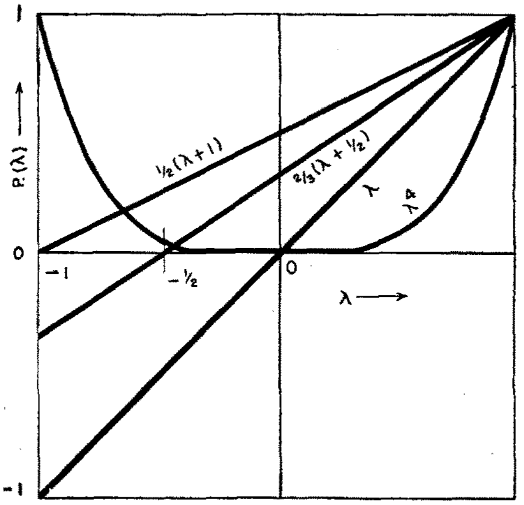

FIG. 1. The factors $P(\lambda)$ corresponding to certain polynomial operators $P(\omega)$.

[3] If ω is a real but unsymmetric matrix, and weight factors γ_i exist such that $\gamma_i\omega_{ij}=\gamma_j\omega_{ji}$, ω will have N real eigenvalues and a complete system of eigenvectors orthonormal with respect to the weight factors γ_i, the direct product being defined by

$$(\mathbf{u}, \mathbf{v}) = \Sigma_i \gamma_i u_i v_i.$$

In this case $(\mathbf{u}, \omega\mathbf{u})$ is a symmetric quadratic form that is an extremum for an eigenvector. This case is of importance when different choices of net spacing are made in different regions, the weight factor representing essentially the area of continuum associated with each net point. The weight factors for change of net spacing are given by Kimball and Shortley, Phys. Rev. **45**, 815 (1934). The techniques of the present paper are readily applicable to matrix operators of this type, and in particular to the Schrödinger problem discussed by Kimball and Shortley.

TABLE I. Reduction factor $1/T_m(d)$, where $d=(2/r)-1$.

	$r=0.9$ $d=1.22222$	0.925 1.16216	0.95 1.10526	0.96 1.08333	0.97 1.06186
$m=$ 1	0.81818	0.86047	0.90476	0.92308	0.94175
2	0.50311	0.58781	0.69290	0.74227	0.79676
3	0.27499	0.35816	0.47961	0.54477	0.62361
4	0.14490	0.20884	0.31588	0.38023	0.46502
6	0.03393	0.06890	0.12996	0.17425	0.24138
8	0.01061	0.02229	0.05251	0.07792	0.12123
12	0.00077	0.00238	0.00852	0.01542	0.03001
16	0.00006	0.00025	0.00138	0.00305	0.00740
24			0.00004	0.00012	0.00045
32					0.00003
	$r=0.98$ $d=1.04082$	0.985 1.03046	0.99 1.02020	0.9925 1.01511	0.995 1.01005
$m=$ 1	0.96078	0.97044	0.98020	0.98511	0.99005
2	0.85719	0.88993	0.92454	0.94259	0.96117
3	0.72066	0.77799	0.84264	0.87814	0.91603
4	0.58076	0.65559	0.74637	0.79932	0.85846
6	0.35076	0.43397	0.55043	0.62750	0.72281
8	0.20285	0.27373	0.38606	0.46941	0.58347
12	0.06555	0.10396	0.17853	0.24514	0.35360
16	0.02101	0.03892	0.08052	0.12382	0.20514
24	0.00215	0.00543	0.01620	0.03098	0.06686
32	0.00022	0.00076	0.00325	0.00772	0.02149
40	0.00002	0.00011	0.00065	0.00193	0.00692
48		0.00001	0.00013	0.00048	0.00223
56			0.00003	0.00012	0.00072
64			0.00001	0.00003	0.00023
72					0.00008
80					0.00002

must have $(\mathbf{w}, \boldsymbol{\omega}\mathbf{w})=\lambda_1$, and hence $\mathbf{w}$ is an eigenvector belonging to λ_1. Thus $\omega w_k \equiv \sum_l \omega_{kl} w_l = \lambda_1 w_k$ for every k.

Now $\mathbf{v}$ (and hence $\mathbf{w}$) can have no zero component. For if $\mathbf{w}$ has any zero values, there must be at least one point P_r for which $w_r=0$ that has a neighboring point P_s for which $w_s>0$. This follows because $\mathbf{w}$ does not vanish identically and because the interior points are connected by the net. But we have arrived at a contradiction since on the one hand $\omega w_r = \lambda_1 w_r = 0$, while on the other ωw_r is a sum of non-negative terms $\omega_{rl} w_l$ of which at least one, $\omega_{rs} w_s$, is not zero.

Further, no component of $\mathbf{v}$ is negative. As before one could find neighboring values $v_r<0, v_s>0$ ($v_k=0$ having been excluded). But now we have $(\mathbf{w}, \boldsymbol{\omega}\mathbf{w}) > (\mathbf{v}, \boldsymbol{\omega}\mathbf{v})$ since $\omega_{rs} w_r w_s = -\omega_{rs} v_r v_s > 0 > \omega_{rs} v_r v_s$.

Finally, no two linearly independent eigenvectors can belong to λ_1. For if $\mathbf{v}^1$ and $\mathbf{v}^2$ were such eigenvectors, normalized and with only positive components, their difference would be an (unnormalized) eigenvector belonging to λ_1 that would have both positive and negative values, which is impossible. Thus $\mathbf{u}^1$ is unique and is everywhere positive. Since every eigenvector belonging to a different eigenvalue must be orthogonal to $\mathbf{u}^1$, $\mathbf{u}^1$ must be the only everywhere positive eigenvector in the set $\mathbf{u}^1, \cdots, \mathbf{u}^N$.

We have thus proved that

$$1 > \lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_N > -1. \tag{7}$$

An arbitrary vector $\mathbf{v}$ has a unique expansion in terms of the set (6):

$$\mathbf{v} = \sum_n c_n \mathbf{u}^n, \tag{8}$$

where c_n is the inner product of $\mathbf{v}$ and $\mathbf{u}^n$, that is

$$c_n = (\mathbf{v}, \mathbf{u}^n) = \sum_k v_k u^n{}_k.$$

If $\mathbf{v}$ is everywhere positive the coefficient c_1 must also be positive, and if furthermore $\mathbf{v}$ is reasonably pillow-shaped, c_1 will probably be larger than the magnitude of any other coefficient.

The effect of applying the operator $\boldsymbol{\omega}$ to an arbitrary vector is to produce the vector

$$\boldsymbol{\omega}\mathbf{v} = \sum_n \lambda_n c_n \mathbf{u}^n.$$

More generally, if $P(\boldsymbol{\omega})$ is any polynomial in $\boldsymbol{\omega}$,

$$P(\boldsymbol{\omega})\mathbf{v} = \sum_n P(\lambda_n) c_n \mathbf{u}^n,$$

that is to say that each coefficient c_n is multiplied by the value of the polynomial $P(\lambda)$ at $\lambda=\lambda_n$. Figure 1 shows the graphs of the factors $P(\lambda)$ corresponding to the operators $\boldsymbol{\omega}$, $\boldsymbol{\omega}^4$ and $(\boldsymbol{\omega}-a)/(1-a)$ for $a=-1$ and for $a=-\frac{1}{2}$.

It is clear from Fig. 1 that the operator $\boldsymbol{\omega}$ will reduce the magnitude of every coefficient c_n in proportion to the nearness of the corresponding eigenvalue to 0, but that if $|\lambda_n|$ is close to λ_1 the reduction relative to c_1 will be small, or there may even be a relative increase if $|\lambda_n|>\lambda_1$. Iteration of the operator $\boldsymbol{\omega}$, represented by the operator $\boldsymbol{\omega}^p$, is seen to have the same properties, except that the relative effect increases with increasing p. On the other hand the operator $(\boldsymbol{\omega}-a)/(1-a)$ can be chosen with a negative so that every $c_n(\neq c_1)$ is reduced in magnitude, not only absolutely, but relatively to the magnitude of c_1.

TECHNIQUES FOR THE SOLUTION OF THE MATRIX EQUATION

By iterative application to a trial vector (8) of a polynomial operator $P(\boldsymbol{\omega})$ of the type shown in Fig. 1 it is possible to obtain a sequence of vectors $\mathbf{v}^0(=\mathbf{v})$, $\mathbf{v}^1, \cdots, \mathbf{v}^j, \cdots$, such that the ratios $|c_n{}^j/c_1{}^j|$ are successively reduced. Past procedures for obtaining the fundamental solution of such a matrix equation have been of this type. p iterations of a polynomial operator $P_d(\boldsymbol{\omega})$ of degree d are equivalent to the operator $[P_d(\boldsymbol{\omega})]^p$, which is of degree pd. We shall now show that for given degree m there exists a "best" polynomial operator $P_m(\boldsymbol{\omega})$. Since this polynomial is not a power of any polynomial of lower degree, it is therefore "better" than the result of p iterations of any operator of degree d such that $pd \leq m$.

The criterion for the "best" polynomial operator $P(\boldsymbol{\omega})$ of degree m may be rigorously expressed in the following form: We assume that the eigenvalues of the problem lie in an interval $(-a, c)$ and that there is an interval $(-a, b)$, with $b<c$, within which it is desired that the maximum of $|P(\lambda)/P(c)|$ shall be a minimum. By limiting consideration to those polynomials for which $P(c)=1$ we then seek among such polynomials those for which the maximum of $|P(\lambda)|$ is a minimum throughout the interval $(-a, b)$. If such a polynomial exists for each choice of m, a, b, c, then by taking $a=-\lambda_N$, $b=\lambda_2$, $c=\lambda_1$ we can preserve the magnitude of c_1 in the expansion of an arbitrary trial vector (8), and obtain the maximum reduction in magnitude of the remaining coefficients.[4] In practice, we do not know in advance what values should be assigned to a, b, and c. In later sections we shall show how such assignments may reasonably be made.

The existence of the desired "best" polynomial is given by the following theorem, in which, for convenience, we have made the changes of variable

$$\Omega=\frac{2}{a+b}\omega+\frac{a-b}{a+b}, \qquad \mu(\lambda)=\frac{2}{a+b}\lambda+\frac{a-b}{a+b} \tag{9}$$

[so that $\mu(-a)=-1$, $\mu(b)=+1$], and have set $d=\mu(c)\equiv 2(a+c)/(a+b)-1>1$.

Theorem: Among all polynomials of degree m in μ having the value $+1$ at $\mu=d>1$, there is just one having the minimum maximum absolute value throughout the interval $(-1, +1)$, namely the polynomial

$$S_m(\mu)=T_m(\mu)/T_m(d), \tag{10}$$

where $T_m(\mu)$ is the mth-order Tschebyscheff polynomial,[5] obtained by expanding

$$T_m(\mu)=\cos(m \arccos\mu)$$

in powers of μ.

Proof: $T_m(\mu)$ is alternately $+1$ and -1 at the $m+1$ points $\mu_k=\cos(k\pi/m)$ $(k=0, 1, \cdots, m)$. Consequently $|S_m(\mu)|$ has $1/T_m(d)$ as its maximum at these $m+1$ values of μ. Suppose that $R(\mu)$ is a polynomial of degree m that is $=1$ at $\mu=d$ and has equal or smaller maximum absolute value in the interval $(-1, +1)$. Let $Q(\mu)=S_m(\mu)-R(\mu)$. Then $Q(\mu)$ is of degree m or less, $Q(d)=0$, and $Q(\mu_k)\geq 0$ or ≤ 0 according as k is even or odd. We proceed to show that $Q(\mu)$ has at least $m+1$ zeros in the closed interval $[-1, d]$, which is possible only if $Q\equiv 0$, i.e., if $R\equiv S_m$. Note that if $0<k<m$ and $Q(\mu_k)=0$ then μ_k is a multiple zero since both S_m and R must have natural extrema at μ_k, and hence $S_m'(\mu_k)-R'(\mu_k)=Q'(\mu_k)=0$. Now consider the interval $[\mu_k, \mu_{k+1}]$ $(k=1, \cdots, m-2)$. Either Q is positive at one end of this interval and negative at the other, in which case there is a zero in the interior of the interval; or Q is zero at one or both ends of the interval. In the latter case Q has at least two zeros in the closed interval—let us assign *one* of these zeros to the interval $[\mu_k, \mu_{k+1}]$ and one to the adjoining interval that has the double zero at its end point. In this way we see that Q has at least one zero assignable to each of the $m-2$ intervals mentioned above. The intervals $[\mu_0, \mu_1]$ and $[\mu_{m-1}, \mu_m]$ have (at least) either an interior zero, a double zero at μ_1 or μ_{m-1}, or a single zero at μ_0 or μ_m. Hence at least one zero is assignable to each of the m intervals $[\mu_k, \mu_{k+1}]$ $(k=0, \cdots, m-1)$. Q also has a zero at d. Hence it has at least $m+1$ zeros, which proves that $Q\equiv 0$ and $R\equiv S_m$.

The quantity $1/T_m(d)$ is the maximum of $|S_m(\mu)|$ in the interval $[-1, +1]$, and is therefore the measure of the reduction of the coefficients of all eigenfunctions with eigenvalue in the range $-a<\lambda<b$, relative to an eigenfunction with eigenvalue c, when $S_m(\Omega)$ is applied to a trial vector (8). Table I gives this reduction factor, and can be used to determine the degree of the polynomial required to remove unwanted eigenfunctions to any given degree of accuracy. The parameter d has been replaced by the quantity $r=(a+b)/(a+c)$, which is more directly related to the spacing of the eigenvalues.

Figure 2 illustrates the effect of the operator $S_6(\Omega)$ for the case $a=1$, $b=0.9$, $c=1$ $(r=0.95)$. All eigenfunctions with eigenvalues in the range from -1 to $+0.9$ are multiplied by a factor of 0.13 or less, as shown by the curve $S_6[(20\lambda+1)/19]$. For comparison the sixth order curve λ^6 is shown, and the range over which this is less than 0.13 is indicated. In an actual computation

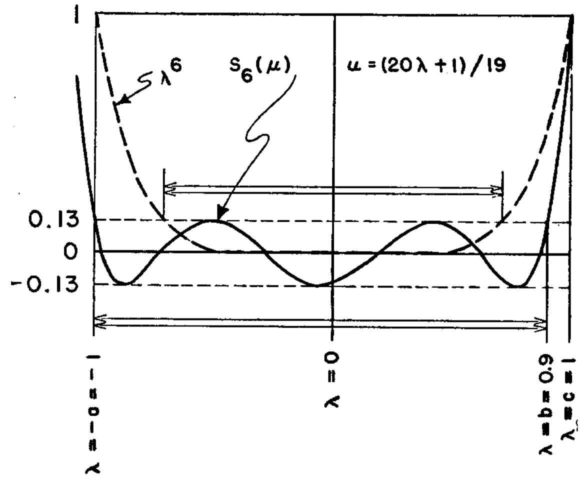

Fig. 2. As shown by the double arrows, the Tschebyscheff operator $S_6(\Omega)$ will accomplish a given reduction over a much wider range than will the operator ω^6. (In the upper right-hand corner, u should be μ.)

[4] In one sense the "best" polynomial would be $(\omega-\lambda_2)(\omega-\lambda_3)\cdots(\omega-\lambda_N)$, since this would eliminate every eigenvector except $\mathbf{u}^1$. But in any practical case such as we envisage, where the number of eigenvectors is large (in the hundreds or thousands), this would not actually be as useful (even assuming the eigenvalues known) as the polynomial we describe, since it would be of much higher degree.

[5] We are indebted to Drs. Tukey and Grosch for suggesting the possible usefulness of Tschebyscheff polynomials, in the course of a discussion of this work at the IBM Seminar on Scientific Computation held in Endicott, New York in November, 1949. We had previously used polynomials with zeros equally spaced. These were fairly efficient, but appreciably less so than the Tschebyscheff polynomials.

a polynomial of order higher than six would always be used.

In the practical application of an operator $S_m(\boldsymbol{\Omega})$ to a trial vector it is often convenient to use the following factor formula for Tschebyscheff polynomials:

$$T_{pq}(x)=\prod_{k=1}^{p}\left[T_q(x)-\cos\frac{2k-1}{p}\frac{\pi}{2}\right]. \tag{11}$$

This follows from the identity

$$\cos(pq\arccos x)\equiv\cos\{p\arccos[\cos(q\arccos x)]\},$$

which may be written

$$T_{pq}(x)=T_p[T_q(x)].$$

Since $T_p(y)$ has p zeros located at $y=\cos[(2k-1)/p](\pi/2)$ $(k=1,\cdots,p)$, (11) results from setting $y=T_q(x)$ in the factored form of $T_p(y)$.

If we have a vector $\mathbf{v}$ that is known to be a linear combination of just the first two eigenvectors, i.e.,

$$\mathbf{v}=c_1\mathbf{u}^1+c_2\mathbf{u}^2, \tag{12a}$$

one can determine λ_1, λ_2, $\mathbf{u}^1$, $\mathbf{u}^2$ by the following "harmonic analysis": First compute $\boldsymbol{\omega}\mathbf{v}$ and $\boldsymbol{\omega}^2\mathbf{v}$,

$$\boldsymbol{\omega}\mathbf{v}=\lambda_1c_1\mathbf{u}^1+\lambda_2c_2\mathbf{u}^2, \tag{12b}$$

$$\boldsymbol{\omega}^2\mathbf{v}=\lambda_1^2c_1\mathbf{u}^1+\lambda_2^2c_2\mathbf{u}^2. \tag{12c}$$

Next form the inner products

$$\left.\begin{aligned}(\mathbf{v},\mathbf{v})&\equiv A=c_1^2+c_2^2,\\(\mathbf{v},\boldsymbol{\omega}\mathbf{v})&\equiv B=\lambda_1c_1^2+\lambda_2c_2^2,\\(\boldsymbol{\omega}\mathbf{v},\boldsymbol{\omega}\mathbf{v})\equiv(\mathbf{v},\boldsymbol{\omega}^2\mathbf{v})&\equiv C=\lambda_1^2c_1^2+\lambda_2^2c_2^2,\\(\boldsymbol{\omega}\mathbf{v},\boldsymbol{\omega}^2\mathbf{v})&\equiv D=\lambda_1^3c_1^2+\lambda_2^3c_2^2.\end{aligned}\right\} \tag{13}$$

When c_1^2 and c_2^2 are eliminated from the equations (13) there result two independent equations for λ_1 and λ_2, which may be written in the form

$$\begin{vmatrix}A&1&1\\B&\lambda_1&\lambda_2\\C&\lambda_1^2&\lambda_2^2\end{vmatrix}=\begin{vmatrix}B&\lambda_1&\lambda_2\\C&\lambda_1^2&\lambda_2^2\\D&\lambda_1^3&\lambda_2^3\end{vmatrix}=0. \tag{14}$$

If these be solved for λ_1 and λ_2, and the possibilities $\lambda_1=0$, $\lambda_2=0$, $\lambda_1=\lambda_2$ (which do not concern us here) be discarded, the two λ's will be found to be the roots of the equation

$$\begin{vmatrix}A&B&1\\B&C&\lambda\\C&D&\lambda^2\end{vmatrix}=0. \tag{15}$$

$c_1\mathbf{u}^1$ and $c_2\mathbf{u}^2$ can then be determined from (12a, b) by

$$\left.\begin{aligned}c_1\mathbf{u}^1&=\frac{1}{\lambda_1-\lambda_2}\boldsymbol{\omega}\mathbf{v}-\frac{\lambda_2}{\lambda_1-\lambda_2}\mathbf{v},\\c_2\mathbf{u}^2&=\frac{-1}{\lambda_1-\lambda_2}\boldsymbol{\omega}\mathbf{v}+\frac{\lambda_1}{\lambda_1-\lambda_2}\mathbf{v}.\end{aligned}\right\} \tag{16}$$

If $\mathbf{v}$ is only approximately represented by the linear combination (12a), so that relations (13) are only approximately true, the solutions of (15) will give approximate values, λ_1' and λ_2', of λ_1 and λ_2. Since λ_1' and λ_2' $(\lambda_1'>\lambda_2')$ satisfy (13) exactly, $\lambda_1'A\equiv\lambda_1'(c_1^2+c_2^2)>\lambda_1'c_1^2+\lambda_2'c_2^2\equiv B$, $\lambda_1'>B/A\equiv(\mathbf{v},\boldsymbol{\omega}\mathbf{v})/(\mathbf{v},\mathbf{v})$, i.e., λ_1' is greater than the estimate of λ_1 that would be obtained directly from the variational principle. In the examples we have tried we have always found $\lambda_1'<\lambda_1$, but we have not been able to prove that this will be true in general.

APPLICATION OF THE TECHNIQUES

In this section we shall discuss in some detail the application of the techniques developed in the preceding section. The general idea is first to do enough preliminary work to determine the rough shape of the eigenvector and approximate values of λ_1 and λ_2; then to make reasonable choices of a, b, and c, compute $r=(a+b)/(a+c)$, and choose an order m from Table I that will give adequate elimination of unwanted eigenfunctions. Then the operator $S_m(\boldsymbol{\Omega})$ given by (9) and (10) is applied to the rough eigenfunction to accomplish this elimination.

The problems for which these techniques were developed were based on relatively complicated regions that necessitated the use of hundreds or thousands of points to get even a moderately acceptable representation of the region and its boundary by the network. This has meant that the computations were possible only with the assistance of fairly high-speed computing machinery with a large memory, such as IBM punched-card equipment, particularly the 604 electronic computer. The following remarks are assumed to be concerned with such large-scale problems computed with comparable equipment.

In starting it is generally worth while to use a rather coarse net, even at the expense of considerably oversimplifying the geometry; e.g., if the region contains holes on whose boundaries the function is 0, it may be worth while to select a mesh so large that such a hole is represented by a single point at which the function is 0. The particular advantages of such a rough first approximation are that (a) the eigenvalues are much less densely spaced, so that approximate values of λ_1 and λ_2 are much more easily determined; (b) relation (4) shows that if λ' and λ'' are the eigenparameters for networks of mesh h' and h'', respectively, then λ_1'' and λ_2'' may be approximated from λ_1' and λ_2' by using $\lambda''=1-(h''/h')^2(1-\lambda')$; (c) the simplest possible interpolation in the vector derived from the coarse net will generally yield a vector that differs from the fundamental for the fine net principally by eigenvectors with eigenvalues near -1 (i.e., high-frequency oscillations), which are easy to eliminate. Thus, one may get a good starting vector for the fine network by treating the solution for the coarse network as a stepfunction in the fine network and smoothing it by a few preliminary operations with $(\boldsymbol{\omega}+1)/2$, about equal in number to the ratio of h' to h''. A better starting function would be obtained by using an interpolation based on the difference equation (3) itself. Thus if 1, 2, 3, 4 are four points at the corners of a square of diagonal $2h'$, the difference equation determines the value u_0 at the center of the square as

$$u_0=(1/\lambda_1')\,\tfrac{1}{4}(u_1+u_2+u_3+u_4),$$

where λ_1' is the eigenvalue appropriate to a net of mesh h'. If we start with a net of mesh h, we can first use $h'=h/\sqrt{2}$ to fill in the center points of the meshes, then $h'=h/2$ to fill in the remaining points in a net of mesh $\frac{1}{2}h$.

The form of the starting function for the coarse network is not too important, provided the generally convex character of the function is observed. Advantage should be taken of any symmetries in the figure to reduce the number of points, which not only reduces the number of eigenfunctions but also decreases the density of the eigenvalues. It should be noted that the matrix resulting from using only a portion of a symmetrical figure is no longer symmetrical. For example, if the full figure is symmetrical along a diagonal and half the figure is used, the neighbors of a diagonal point are the neighbors on one side counted twice. The eigenvectors of the resulting matrix will not then be orthogonal in general. Their orthogonality will be restored if they are treated as eigenvectors of the matrix associated with the complete figure, which is done simply by giving to each point in the subregion actually used a weight equal to the number of points in the full region for which it stands.

In problems of the type considered here one may assume $a=1$ throughout, since $\lambda_N\approx-1$; and $c=1$ initially, since $\lambda_1\approx+1$. The first step is to pick an operator $\mathbf{\Omega}$ and a polynomial $S_m(\mathbf{\Omega})$ of low degree that will enable one to compute λ_1 and λ_2 by the crude harmonic analysis of the previous section with fair accuracy. Thus, with $a=c=1$ the choice of $r=0.9$ and $m=8$ corresponds to $b=0.8$, $d=1.222$, $\mathbf{\Omega}=(10\omega+1)/9$, $1/T_m(d)=0.0106$. This $S_m(\mathbf{\Omega})$ will reduce the coefficients of all eigenvectors in the expansion of $\mathbf{v}$ with eigenvalues lying in the range $-1\leq\lambda\leq0.8$ by at least the factor 0.01, and those with eigenvalues in the range $0.8\leq\lambda\leq1$ at least in proportion to the ratio $(\lambda-0.8)/0.2$. Ordinarily this will permit a harmonic analysis that will determine λ_1 and λ_2 with sufficient accuracy to choose a final operator for this network. However, if the number of points in the coarse network is rather large, it may be worth while to take r and m larger.

In the actual performance of an operation $S_m(\mathbf{\Omega})$ one is faced with a choice of methods. On the one hand $S_m(\mathbf{\Omega})\mathbf{v}$ may be computed by iterating with the operator $\mathbf{\Omega}$ and then forming the polynomial as a linear combination of the results. This has the practical disadvantage that with large m the sizes of the terms are so disparate that a great deal of accuracy is lost in taking differences of large multiples of the results of the iterations. At the other extreme, $S_m(\mathbf{\Omega})$ may be factored into its linear factors,

$$S_m(\mathbf{\Omega})=\prod_{k=1}^{m}\frac{\mathbf{\Omega}-\mu_k}{d-\mu_k},\quad\left(\mu_k=\cos\frac{2k-1}{m}\frac{\pi}{2}\right)\tag{17}$$

and these linear operators applied successively. This method suffers from the disadvantage that the coefficients of the operator must be changed at each step, which is a dangerous source of error in machine operation. A reasonable practical compromise combines the two methods by use of the factor formula (11). Thus if we set

$$S_m(\mathbf{\Omega})=\prod_{k=1}^{p}\frac{T_q(\mathbf{\Omega})-\cos[(2k-1)/p]\pi/2}{T_q(d)-\cos[(2k-1)/p]\pi/2},\tag{18}$$

where q is about 6 or 8[6] and $p=m/q$, we may iterate q times with $\mathbf{\Omega}$, follow this with the computation of the first of the p polynomials, iterate on the result q times with $\mathbf{\Omega}$, compute the second polynomial factor, and so on. Since the cosine term in the denominator is small compared to $T_q(d)$, it may be omitted. This has the effect of changing the scale of the vector slightly, which is of no theoretical importance. If this omission is made, then the operator used in the iterations is always the same, while the coefficients used in the successive polynomials differ only in the constant term. Loss of significant figures is avoided by keeping q relatively small. When q is at least as large as 6 or 8, the change of scale produced by neglecting the cosine term in the denominator is not bothersome. (If the function does grow out of bounds, cutting it in half once will generally be sufficient. Note that this should be done *after* a polynomial factor is computed, not in the course of a set of iterations preparatory to the computation of such a polynomial.)

When a satisfactory solution for the coarse network has been obtained, a starting vector for the fine network may be constructed along the lines suggested above. At this point procedure will depend to some extent on whether the interest in the problem centers in the determination of the fundamental eigenvalue or of the fundamental eigenfunction. If the former, then b may safely be chosen somewhat less than λ_2 and we may determine λ_1 by harmonic analysis. If λ_1 and λ_2 are widely separated this method may also be used for determining the vector $c_1\mathbf{u}^1$. However, the determination of $c_1\mathbf{u}^1$ from (16) generally involves great loss of accuracy because of the nearness of λ_1 and λ_2, so if interest is centered in the eigenfunction, one is forced to choose b close to λ_2 with consequent increase in the size of m.

HIGHER MODES

While the method we have outlined is peculiarly adapted to obtaining the fundamental mode, it is pos-

[6] T_6 and T_8, which we have found most useful, are given by

$$2^{-5}T_6(\mathbf{\Omega})=\mathbf{\Omega}^6-\tfrac{3}{2}\mathbf{\Omega}^4+\tfrac{9}{16}\mathbf{\Omega}^2-\tfrac{1}{32}$$
$$2^{-7}T_8(\mathbf{\Omega})=\mathbf{\Omega}^8-2\mathbf{\Omega}^6+\tfrac{5}{4}\mathbf{\Omega}^4-\tfrac{1}{4}\mathbf{\Omega}^2+\tfrac{1}{128}.$$

sibly useful for obtaining a limited number of the higher modes. Once the fundamental has been accurately obtained for the network, a trial vector for the next mode can be chosen that is accurately orthogonal to the fundamental. Application of $P(\omega)$ to such an orthogonal vector should not reintroduce (except as a result of rounding-off errors) any of the fundamental. Hence it should be possible to obtain a higher mode by orthogonalizing the trial vector to the predetermined lower modes, applying a suitable Tschebyscheff operator, and re-orthogonalizing as necessary. One useful application of this technique might be in the solution of Schrödinger's equation for diatomic molecules.

MORE GENERAL PROBLEMS

Several of the problems we have solved by a procedure similar to the above have been of a more general type than the one proposed in the introduction. Since the modifications introduced by this generalization may assist others in applying the procedure, we shall state briefly the most important features of the more general problems and their solution.

The single dependent variable u is replaced by two functions u and v that are to satisfy simultaneous equations

$$\left.\begin{aligned}(\Delta-E)u+Fv&=-\alpha u\\ Gu+(\Delta-H)v&=-\alpha v.\end{aligned}\right\}\qquad(19)$$

The fundamental region is divided into subregions within each of which the coefficients E, F, G, H are constant functions of a parameter ρ. Certain homogeneous conditions are imposed across the interfaces. It is desired to determine ρ so that the fundamental eigenvalue $\alpha_1=0$, and to find the corresponding pair of eigenfunctions u, v.

The matrix equation analogous to (5) is then of the form

$$\begin{pmatrix}\omega-\frac{1}{4}h^2E & \frac{1}{4}h^2F\\ \frac{1}{4}h^2G & \omega-\frac{1}{4}h^2H\end{pmatrix}\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}=\lambda\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix},\qquad(20)$$

where $\lambda=1-\frac{1}{4}h^2\alpha$, and consequently it is desired that $\lambda_1=1$. As in the first problem the introduction of the boundary conditions (and here of the interface conditions as well) results in modification of the interpretation of the operator ω at certain points, but ω still represents a linear combination of neighboring values with non-negative coefficients whose sum is ≤ 1. The two vectors $\mathbf{u}$, $\mathbf{v}$ are considered to form a single vector for the purposes of the computation.

Note that the matrix is no longer symmetric. This means that the eigenvectors may not be assumed to be orthogonal. Hence it is necessary ultimately to compute the adjoint solution in order to form correct inner products if a harmonic analysis is to be carried out. In the early stages this is not necessary since it can be shown that the λ_1 and λ_2 determined by the analog of (15) represent the result of fitting relations (12) by least squares. The λ's thus determined are, however, less accurate than those determined by true inner products.

In these problems h is chosen sufficiently small so that the coefficients h^2E, h^2F, h^2G, h^2H are small compared to 1. Hence the operator is nearly the operator $\begin{pmatrix}\omega 0\\ 0\omega\end{pmatrix}$. Aside from the asymmetry the principal effect of the more general matrix is to increase the range of the eigenvalues. Since it is possible to estimate ρ (on physical grounds) with fair accuracy in advance, λ_1 is still approximately $+1$. λ_N, however, may be considerably less than -1, say -3 or -4 if $\mathbf{u}$ and $\mathbf{v}$ are disparate in size. Our first problem is then to estimate λ_N. This may be done by choosing a sequence of operators of the type $(\omega-a)/(1-a)$ to replace ω in the matrix. If one takes $a=0$ and applies the resulting operator two or three times, the presence of eigenvectors with eigenvalue <-1 will be revealed by their coefficients being multiplied by negative factors of large absolute value, which thus produces oscillations of increasing amplitude in the successive vectors. (That such eigenvectors will appear, provided $\lambda_N<-1$, is practically guaranteed by the rounding-off errors introduced into the computation.) When the existence of such large negative eigenvalues has been established, similar operations may be carried out with $a=-1, -\frac{3}{2}, -2$, etc., until the phenomenon disappears. Thus a practical value of a for use in determining the operator $\mathbf{\Omega}$ in the Tschebyscheff polynomial is achieved. The polynomial operator S_m is now not a function of $\mathbf{\Omega}$ but of the matrix operator in (20) with $\mathbf{\Omega}$ substituted for ω.

A final word about the treatment of the interface conditions. The particular conditions imposed are expressed in the difference equations by assertions of the form: the value of u at an interface point is a weighted average of certain neighboring values of u in the adjacent regions. It would be possible to eliminate all interface points from the vector in a manner similar to that in which the boundary points were eliminated. However, since λ_1 is very nearly $+1$ in these problems it is more convenient to retain the interface points, replacing their values at each stage by the weighted average specified in the interface conditions. This averaging then becomes part of the operator. Since we are doing the averaging and determining the interface values always one step late, this procedure is only exact for an eigenfunction of eigenvalue unity.

We are indebted to many of the members of the Theoretical Physics Division of the Argonne Laboratory, and in particular to Drs. Frank Hoyt and Elmer Eisner, for helpful discussions during the development of these techniques.