Iterative procedures for solving finite-difference approximations to separable partial differential equations

By M. R. Osborne

We show that, for a class of separable partial differential equations of elliptic type in two independent variables, the optimum value of the over-relaxation parameter ω can be calculated when the eigenvalue of maximum modulus of a certain double eigenvalue problem is known. This generalizes the well-known case of Laplace's equation for a rectangle (for which explicit results are known), and provides a convenient method for the calculation of the optimum value of ω. We extend some of our results to partial differential equations in three independent variables, and we also use our formalism to discuss the Peaceman–Rachford iteration.

Introduction

In this paper we consider certain classes of iterative methods for solving finite-difference approximations to elliptic partial differential equations having the form

$$\frac{\partial^2 Q}{\partial x^2} + \frac{\partial^2 Q}{\partial y^2} + f(x)\frac{\partial Q}{\partial x} + g(y)\frac{\partial Q}{\partial y} + [p(x) + q(y)]Q = c(x, y) \qquad (1)$$

with the boundary conditions given on a rectangle with sides parallel to the x and y axes, respectively. Our analysis will be worked for the case Q prescribed on the boundary but applies (with obvious modifications) to the case $AQ + B\frac{\partial Q}{\partial n}$ prescribed, where A and B are constant on any one side. Typical equations to which our results are applicable are Laplace's equation for the rectangle and (in polar coordinates) the circular annulus, and Reynolds equation* for a rectangle.

By restricting our attention to rectangular regions and separable partial differential equations, we find that it is possible to represent the unknowns in our finite-difference equations as the components of a rectangular matrix (in contrast to the usual representation as the components of a vector), and that the finite-difference equations then take a very simple form [equation (5)]. This representation is given in section 1 where it is applied to analyze the iterative scheme of successive over-relaxation by points. Successive over-relaxation by lines is considered in section 2, and the Peaceman–Rachford iteration in section 3. Successive over-relaxation by points for three independent variables is considered in section 4. In section 5 we give a computational scheme for solving the eigenvalue problems derived in the preceding sections, and in section 6 we consider the relation between our method and that normally used to analyze the iterative procedures discussed here.

1. Iteration by points

If we write $\boldsymbol{\phi}_i$ for the vector whose components make up the ith line ($x = x_i$, $i = 1, 2, \ldots, n$) of the solution, then the standard five-point difference approximation to (1) takes the form

$$h^{-2}\left(1 - \frac{hf_i}{2}\right)\boldsymbol{\phi}_{i-1} + [\boldsymbol{G} - (2h^{-2} - p_i)\boldsymbol{I}]\boldsymbol{\phi}_i + h^{-2}\left(1 + \frac{hf_i}{2}\right)\boldsymbol{\phi}_{i+1} = \boldsymbol{C}_i \qquad (2)$$

where $\boldsymbol{G}$ is an $(m \times m)$ tridiagonal matrix whose jth row is $k^{-2}\left(1 - \frac{kg_j}{2}\right)$, $-2k^{-2} + q_j$, $k^{-2}\left(1 + \frac{kg_j}{2}\right)$, and where h and k are the mesh spacings in the x and y directions, respectively.

Writing $\boldsymbol{G} - (2h^{-2} - p_i)\boldsymbol{I} = \boldsymbol{L} + \boldsymbol{X}_i + \boldsymbol{U}$ as the sum of a lower-triangular, a diagonal, and an upper-triangular matrix, we can represent the Young–Frankel over-relaxation scheme with iteration parameter ω as

$$h^{-2}\left(1 - \frac{hf_i}{2}\right)\boldsymbol{\phi}_{i-1}^{(s+1)} + (\boldsymbol{L} + \omega^{-1}\boldsymbol{X}_i)\boldsymbol{\phi}_i^{(s+1)} + [(1 - \omega^{-1})\boldsymbol{X}_i + \boldsymbol{U}]\boldsymbol{\phi}_i^{(s)} + h^{-2}\left(1 + \frac{hf_i}{2}\right)\boldsymbol{\phi}_{i+1}^{(s)} = \boldsymbol{C}_i \qquad (3)$$

where the upper suffix s indicates the progress of the iteration.

To determine the rate of convergence of this iteration we require a knowledge of the eigenvalue of maximum modulus of the problem (as in Forsythe and Wasow, 1960, p. 214 and p. 247)

$$\lambda\left\{h^{-2}\left(1 - \frac{hf_i}{2}\right)\boldsymbol{\phi}_{i-1} + (\boldsymbol{L} + \omega^{-1}\boldsymbol{X}_i)\boldsymbol{\phi}_i\right\} + \{(1 - \omega^{-1})\boldsymbol{X}_i + \boldsymbol{U}\}\boldsymbol{\phi}_i + h^{-2}\left(1 + \frac{hf_i}{2}\right)\boldsymbol{\phi}_{i+1} = 0. \qquad (4)$$

* This equation, which is important in the theory of film lubrication, has the form

$$\frac{\partial}{\partial x}\left[h^3(x)\frac{\partial P}{\partial x}\right] + h^3(x)\frac{\partial^2 P}{\partial y^2} = \frac{dh}{dx}.$$

Before reducing (4) we note that the left-hand side of equation (2) can be described by an operator combining together consecutive $\boldsymbol{\phi}_i$ plus an operation $\boldsymbol{G}$ on the individual $\boldsymbol{\phi}_i$. If we write Φ for the $m \times n$ matrix with columns $\boldsymbol{\phi}_i$ then (2) becomes

$$\boldsymbol{G}\Phi + \Phi\boldsymbol{F} = \boldsymbol{C} \qquad (5)$$

where $\boldsymbol{F}$ is the $(n \times n)$ tridiagonal matrix whose ith row is

$$h^{-2}\left(1 + \frac{hf_{i-1}}{2}\right),\ -2h^{-2} + p_i,\ h^{-2}\left(1 - \frac{hf_{i+1}}{2}\right).$$

The representation of the difference equation in this form has been given before by Bickley and McNamee (1960).

If we set
$$\boldsymbol{F} = \boldsymbol{L}_1 + \boldsymbol{D}_1 + \boldsymbol{U}_1,$$
$$\boldsymbol{G} = \boldsymbol{L}_2 + \boldsymbol{D}_2 + \boldsymbol{U}_2, \qquad (6)$$

then equation (4) becomes

$$[\lambda\boldsymbol{L}_2 + (\omega^{-1}\lambda + (1 - \omega^{-1}))\boldsymbol{D}_2 + \boldsymbol{U}_2]\Phi$$
$$+ \Phi[\boldsymbol{L}_1 + (\lambda\omega^{-1} + (1 - \omega^{-1}))\boldsymbol{D}_1 + \lambda\boldsymbol{U}_1] = 0. \qquad (7)$$

Equation (7) can be further transformed by using the result (due to Friedman; see Forsythe and Wasow, 1960, p. 249) that there exist diagonal matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ such that

$$\lambda\boldsymbol{L}_2 + \zeta\boldsymbol{D}_2 + \boldsymbol{U}_2 = \sqrt{\lambda}\boldsymbol{A}^{-1}\left\{\boldsymbol{L}_2 + \frac{\zeta}{\sqrt{\lambda}}\boldsymbol{D}_2 + \boldsymbol{U}_2\right\}\boldsymbol{A},$$

$$\text{and } \boldsymbol{L}_1 + \zeta\boldsymbol{D}_1 + \lambda\boldsymbol{U}_1 = \sqrt{\lambda}\boldsymbol{B}\left\{\boldsymbol{L}_1 + \frac{\zeta}{\sqrt{\lambda}}\boldsymbol{D}_1 + \boldsymbol{U}_1\right\}\boldsymbol{B}^{-1}. \qquad (8)$$

If we now set $\bar{\Phi} = \boldsymbol{A}\Phi\boldsymbol{B}$ we have reduced our problem to that of finding the eigenvalues of

$$\{\boldsymbol{L}_2 + \sigma\boldsymbol{D}_2 + \boldsymbol{U}_2\}\bar{\Phi} + \bar{\Phi}\{\boldsymbol{L}_1 + \sigma\boldsymbol{D}_1 + \boldsymbol{U}_1\} = 0 \qquad (9)$$

$$\text{where} \qquad \sigma = \frac{\omega^{-1}\lambda + (1 - \omega^{-1})}{\sqrt{\lambda}}. \qquad (10)$$

The values of σ which satisfy (9) can also be characterized as the solutions of the double eigenvalue problem

$$[\boldsymbol{L}_1 + \sigma\boldsymbol{D}_1 + \boldsymbol{U}_1 + \gamma\boldsymbol{I}]\boldsymbol{v}_1 = 0,$$
$$[\boldsymbol{L}_2 + \sigma\boldsymbol{D}_2 + \boldsymbol{U}_2 - \gamma\boldsymbol{I}]\boldsymbol{v}_2 = 0. \qquad (11)$$

This follows from a result of Bickley and McNamee (1960, p. 118), who showed that the eigenvalue problem

$$\boldsymbol{R}\Phi + \Phi\boldsymbol{S} = \lambda\Phi \qquad (12)$$

where $\boldsymbol{R}$ and $\boldsymbol{S}$ can be diagonalized by similarity transformations, has its eigenvalues in the form $\lambda = \lambda_1 + \lambda_2$, where λ_1 is an eigenvalue of $\boldsymbol{R}$, and λ_2 is an eigenvalue of $\boldsymbol{S}$ (the eigenmatrix being the dyad formed from the corresponding eigenvectors of $\boldsymbol{R}$ and $\boldsymbol{S}$). In particular there is a non-trivial solution with $\lambda = 0$ only if λ_1 is an eigenvalue of $\boldsymbol{R}$ and $-\lambda_1$ is an eigenvalue of $\boldsymbol{S}$. Thus (11) just expresses the condition for (9) to have a non-trivial solution. The condition that the matrices operating on $\bar{\Phi}$ in equation (9) can be diagonalized by similarity transformations is certainly satisfied in many cases of interest. A sufficient condition is that h and k can be chosen small enough to make $h|f_i| < 2$, $k|g_j| < 2$, $h^2|p_i| < 2$, and $k^2|q_j| < 2$ for all i, j, for then $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are negative definite while the off-diagonal elements in $\boldsymbol{G}$ and $\boldsymbol{F}$ are one-signed.

In special cases the equations (11) can be further simplified. Assuming that the conditions in the previous paragraph are satisfied, we have for each of the matrix operators in (9)

$$\boldsymbol{U} + \boldsymbol{L} = \boldsymbol{D}\boldsymbol{V}\boldsymbol{M}\boldsymbol{V}^{-1} \qquad (13)$$

where $\boldsymbol{V}$ is the matrix of (right) eigenvectors of $\boldsymbol{U} + \boldsymbol{L}$, and $\boldsymbol{M}$ is the diagonal matrix of the eigenvalues μ taken with respect to the density matrix $\boldsymbol{D}$. Therefore (in an obvious notation) (9) becomes

$$\boldsymbol{D}_2\boldsymbol{V}_2\{\sigma\boldsymbol{I} + \boldsymbol{M}_2\}\boldsymbol{V}_2^{-1}\bar{\Phi}$$
$$+ \bar{\Phi}\boldsymbol{D}_1\boldsymbol{V}_1\{\sigma\boldsymbol{I} + \boldsymbol{M}_1\}\boldsymbol{V}_1^{-1} = 0. \qquad (14)$$

Case 1 Let $\boldsymbol{D}_1$ be a constant multiple (a_1 say) of the unit matrix $\boldsymbol{I}$. Then $\boldsymbol{D}_1$ commutes with $\boldsymbol{V}_1$ and equation (14) can be written (setting $\bar{\bar{\Phi}} = \bar{\Phi}\boldsymbol{V}_1$)

$$\{\boldsymbol{L}_2 + \sigma\boldsymbol{D}_2 + \boldsymbol{U}_2\}\bar{\bar{\Phi}} + a_1\bar{\bar{\Phi}}\{\sigma\boldsymbol{I} + \boldsymbol{M}_1\} = 0. \qquad (15)$$

From (15) we see that the ith column of $\bar{\bar{\Phi}}$ (say $\boldsymbol{\theta}_i$) satisfies the equation

$$\{\boldsymbol{L}_2 + a_1\mu_{1i}\boldsymbol{I} + \boldsymbol{U}_2 + \sigma(\boldsymbol{D}_2 + a_1\boldsymbol{I})\}\boldsymbol{\theta}_i = 0, \qquad (16)$$
$$i = 1, 2, \ldots, n.$$

Equation (16) is an eigenvalue problem which determines the possible values of σ in this case.

Case 2 Let both $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ be constant multiples of the unit matrix ($\boldsymbol{D}_2 = a_2\boldsymbol{I}$), and let $\bar{\bar{\Phi}} = \boldsymbol{V}_2^{-1}\bar{\Phi}\boldsymbol{V}_1$; then equation (14) can be written

$$a_2\{\sigma\boldsymbol{I} + \boldsymbol{M}_2\}\bar{\bar{\Phi}} + a_1\bar{\bar{\Phi}}\{\sigma\boldsymbol{I} + \boldsymbol{M}_1\} = 0. \qquad (17)$$

In this case $\boldsymbol{\theta}_i$ satisfies

$$\{\sigma(a_1 + a_2)\boldsymbol{I} + (a_2\boldsymbol{M}_2 + a_1\mu_{1i}\boldsymbol{I})\boldsymbol{\theta}_i = 0, \qquad (18)$$

and as the matrices within the brackets are diagonal we have

$$-\sigma_{ij} = \frac{a_2\mu_{2j} + a_1\mu_{1i}}{a_1 + a_2},$$
$$i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m. \qquad (19)$$

A special case of equation (19) is well known. This is the case of Laplace's equation for a rectangle (Forsythe and Wasow, 1960, p. 255). Also in this reference (pp. 250–57) it is shown that the optimum value of ω depends only on the maximum value of σ. A method of calculating this maximum value of σ from equation (11) is given in section 5. No discussion of equation (11) other than that necessary to characterize the maximum value of σ and its attendant value of γ is attempted.

2. Block relaxation by lines

The iterative scheme for successive over-relaxation by lines (Forsythe and Wasow, 1960, pp. 266–71) can be written

$$h^{-2}\left(1-\frac{hf_i}{2}\right)\boldsymbol{\phi}_{i-1}^{(s+1)}+\omega^{-1}[\boldsymbol{G}-(2h^{-2}-p_i)\boldsymbol{I}]\boldsymbol{\phi}_i^{(s+1)}$$
$$+(1-\omega^{-1})[\boldsymbol{G}-(2h^{-2}-p_i)\boldsymbol{I}]\boldsymbol{\phi}_i^{(s)}$$
$$+h^{-2}\left(1+\frac{hf_i}{2}\right)\boldsymbol{\phi}_i^{(s)}=\boldsymbol{C}_i, \quad (20)$$

and the eigenvalue problem determining the rate of convergence is

$$[\lambda\omega^{-1}+(1-\omega^{-1})]\boldsymbol{G}\Phi$$
$$+\Phi\{\lambda\boldsymbol{U}_1+[\lambda\omega^{-1}+(1-\omega^{-1})]\boldsymbol{D}_1+\boldsymbol{L}_1\}=0. \quad (21)$$

Applying a Friedman transformation to the post-multiplier in (21) gives us

$$\frac{\lambda\omega^{-1}+(1-\omega^{-1})}{\sqrt{\lambda}}\boldsymbol{G}\overline{\Phi}+\overline{\Phi}$$
$$\left\{\boldsymbol{L}_1+\frac{\lambda\omega^{-1}+(1-\omega^{-1})}{\sqrt{\lambda}}\boldsymbol{D}_1+\boldsymbol{U}_1\right\}=0 \quad (22)$$

so that the eigenvalues are determined when the values of σ for which the set of equations

$$\sigma\boldsymbol{G}\overline{\Phi}+\overline{\Phi}\{\boldsymbol{L}_1+\sigma\boldsymbol{D}_1+\boldsymbol{U}_1\}=0 \quad (23)$$

have non-trivial solutions $\overline{\Phi}$ are known.

As before we have σ determined by the double eigenvalue problem

$$\{\sigma\boldsymbol{G}-\nu\boldsymbol{I}\}\boldsymbol{v}_2=0,$$
$$\{\boldsymbol{L}_1+\sigma\boldsymbol{D}_1+\boldsymbol{U}_1+\nu\boldsymbol{I}\}\boldsymbol{v}_1=0. \quad (24)$$

In the case $\boldsymbol{D}_1=a_1\boldsymbol{I}$ we have, substituting $\boldsymbol{L}_1+\boldsymbol{U}_1=\boldsymbol{D}_1\boldsymbol{V}_1\boldsymbol{M}_1\boldsymbol{V}_1^{-1}$ into the second equation (24),

$$\boldsymbol{V}_1\{a_1\boldsymbol{M}_1+(a_1\sigma+\nu)\boldsymbol{I}\}\boldsymbol{V}_1^{-1}\boldsymbol{v}_1=0,$$

and this gives the eigenvalues ν in terms of σ as

$$-\nu_i=a_1(\mu_{1i}+\sigma),\ i=1,2,\ldots,n. \quad (25)$$

Substituting for ν from (25) into the first equation of (24) we have

$$\{\sigma(\boldsymbol{G}+a_1\boldsymbol{I})+a_1\mu_{1i}\boldsymbol{I}\}\boldsymbol{v}_2=0. \quad (26)$$

Finally, if $\boldsymbol{D}_2=a_2\boldsymbol{I}$ we have

$$\boldsymbol{G}=\boldsymbol{L}_2+\boldsymbol{D}_2+\boldsymbol{U}_2=\boldsymbol{V}_2\{a_2\boldsymbol{I}+a_2\boldsymbol{M}_2\}\boldsymbol{V}_2^{-1},$$

and the eigenvalues σ of equation (26) are given by

$$\sigma_{ij}=\frac{-a_1\mu_{1i}}{a_1+a_2+a_2\mu_{2j}}. \quad (27)$$

Again (27) is well known for the case of Laplace's equation for a rectangle (Forsythe and Wasow, 1960, p. 270).

3. The Peaceman–Rachford iteration

Our methods can also be applied to analyze the implicit alternating direction methods for solving (5), and here we consider the Peaceman–Rachford method (Forsythe and Wasow, 1960, pp. 272–82). The iteration scheme for the error matrix is

$$\Phi^{(s-\frac{1}{2})}=\Phi^{(s-1)}-\alpha_s(\boldsymbol{G}\Phi^{(s-1)}+\Phi^{(s-\frac{1}{2})}\boldsymbol{F}),$$
$$\Phi^{(s)}=\Phi^{(s-\frac{1}{2})}-\alpha_s(\boldsymbol{G}\Phi^{(s)}+\Phi^{(s-\frac{1}{2})}\boldsymbol{F}). \quad (28)$$

Eliminating $\Phi^{(s-\frac{1}{2})}$ gives

$$\Phi^{(s)}=\{\boldsymbol{I}+\alpha_s\boldsymbol{G}\}^{-1}\{\boldsymbol{I}-\alpha_s\boldsymbol{G}\}\Phi^{(s-1)}\{\boldsymbol{I}+\alpha_s\boldsymbol{F}\}^{-1}\{\boldsymbol{I}-\alpha_s\boldsymbol{F}\},$$
$$=\prod_{r=1}^{s}\{\boldsymbol{I}+\alpha_r\boldsymbol{G}\}^{-1}\{\boldsymbol{I}-\alpha_r\boldsymbol{G}\}\Phi^{(0)}\prod_{r=1}^{s}\{\boldsymbol{I}+\alpha_r\boldsymbol{F}\}^{-1}$$
$$\{\boldsymbol{I}-\alpha_r\boldsymbol{F}\}. \quad (29)$$

Introducing the spectral decompositions of these right and left multiplying matrices we may write (29) as

$$\Phi^{(s)}=\boldsymbol{V}_2\boldsymbol{D}_2\boldsymbol{V}_2^{-1}\Phi^{(0)}\boldsymbol{V}_1\boldsymbol{D}_1\boldsymbol{V}_1^{-1} \quad (30)$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are the diagonal matrices of the eigenvalues. Setting $\overline{\Phi}=\boldsymbol{V}_2^{-1}\Phi\boldsymbol{V}_1$ we have

$$\overline{\Phi}^{(s)}=\boldsymbol{D}_2\overline{\Phi}^{(0)}\boldsymbol{D}_1, \quad (31)$$

and separating out the ith column $\boldsymbol{\phi}_i^{(s)}$ we see that it satisfies

$$\boldsymbol{\phi}_i^{(s)}=d_{1i}\boldsymbol{D}_2\boldsymbol{\phi}_i^{(0)} \quad (32)$$

where d_{1i} is the ith diagonal element of $\boldsymbol{D}_1$. Thus the rate at which $\overline{\Phi}^{(s)}$ tends to zero is determined by the maximum value of $|d_{2j}d_{1i}|$. If we call the eigenvalues of $\boldsymbol{F}\lambda_1,\ldots,\lambda_n$, and those of $\boldsymbol{G}\mu_1,\ldots,\mu_m$

then $$d_{1i}d_{2j}=\prod_{r=1}^{s}\frac{1-\alpha_r\lambda_i\,1-\alpha_r\mu_j}{1+\alpha_r\lambda_i\,1+\alpha_r\mu_j}. \quad (33)$$

Equation (33) is well known for Laplace's equation in a rectangle. An alternative derivation of (33) is quoted in Martin and Tee (1961).

4. Separable equations in three dimensions

Here we indicate the way in which our results extend to problems with more than two independent variables. We consider the partial differential equation

$$\nabla^2\phi+f(x)\frac{\partial\phi}{\partial x}+g(y)\frac{\partial\phi}{\partial y}+e(z)\frac{\partial\phi}{\partial z}$$
$$+(p(x)+q(y)+r(z))\phi=c(x,y,z) \quad (34)$$

with the boundary conditions ϕ specified on the surface of a rectangular box with faces parallel to the coordinate planes. The standard seven-point finite-difference approximation to this equation can be written [using equation (5)]

$$\alpha^{-2}\left(1-\frac{\alpha e_k}{2}\right)\Phi_{k-1}-(2\alpha^{-2}-r_k)\Phi_k$$
$$+\alpha^{-2}\left(1+\frac{\alpha e_k}{2}\right)\Phi_{k+1}+\boldsymbol{G}\Phi_k+\Phi_k\boldsymbol{F}=\boldsymbol{C}_k \quad (35)$$

where Φ_k is the matrix of solution points ϕ_{ijk} for fixed k (constant z), $k = 1, 2, \ldots, t$, and α is the spacing of the mesh in the z direction.

The formula for the iterative solution of (35) by successive over-relaxation by points, assuming each plane Φ_k is relaxed before proceeding to Φ_{k+1}, is

$$\begin{aligned}\alpha^{-2}\left(1 - \frac{\alpha e_k}{2}\right)\Phi_{k-1}^{(s)} - \omega^{-1}(2\alpha^{--2} - r_k)\Phi_k^{(s)}\\ - (1 - \omega^{-1})(2\alpha^{-2} - r_k)\Phi_k^{(s-1)}\\ + \alpha^{-2}\left(1 + \frac{\alpha e_k}{2}\right)\Phi_{(k+1)}^{s-1} + [\boldsymbol{L}_2 + \omega^{-1}\boldsymbol{D}_2]\Phi_k^{(s)}\\ + \Phi_k^{(s)}[\boldsymbol{U}_1 + \omega^{-1}\boldsymbol{D}_1] + [(1 - \omega^{-1})\boldsymbol{D}_2 + \boldsymbol{U}_2]\Phi_k^{(s-1)}\\ + \Phi_k^{(s-1)}[\boldsymbol{L}_1 + (1 - \omega^{-1})\boldsymbol{D}_1] = \boldsymbol{C}_k. \qquad (36)\end{aligned}$$

The rate of convergence of the iteration (36) depends on the eigenvalue of maximum modulus of the problem

$$\begin{aligned}\lambda\alpha^{-2}\left(1 - \frac{\alpha e_k}{2}\right)\Phi_{k-1} - (\lambda\omega^{-1} + (1 - \omega^{-1}))\\ (2\alpha^{-2} - r_k)\Phi_k + \alpha^{-2}\left(1 + \frac{\alpha e_k}{2}\right)\Phi_{k+1}\\ + [\lambda\boldsymbol{L}_2 + (\omega^{-1}\lambda + (1 - \omega^{-1}))\boldsymbol{D}_2 + \boldsymbol{U}_2]\Phi_k\\ + \Phi_k[\boldsymbol{L}_1 + (\lambda\omega^{-1} + (1 - \omega^{-1}))\boldsymbol{D}_1 + \lambda\boldsymbol{U}_1] = 0.\\ (37)\end{aligned}$$

If we now apply the transformations (8) to equation (37) we have, writing $\overline{\Phi}_k = \boldsymbol{A}\Phi_k\boldsymbol{B}$,

$$\begin{aligned}\lambda\alpha^{-2}\left(1 - \frac{\alpha e_k}{2}\right)\overline{\Phi}_{k-1} - \sqrt{\lambda}\sigma(2\alpha^{-2} - r_k)\overline{\Phi}_k\\ + \alpha^{-2}\left(1 + \frac{\alpha e_k}{2}\right)\overline{\Phi}_{k+1} + \sqrt{\lambda}\{[\boldsymbol{L}_2 + \sigma\boldsymbol{D}_2 + \boldsymbol{U}_2]\overline{\Phi}_k\\ + \overline{\Phi}_k[\boldsymbol{L}_1 + \sigma\boldsymbol{D}_1 + \boldsymbol{U}_1]\} = 0. \qquad (38)\end{aligned}$$

Now let the normalized, right and left, bi-orthogonal eigenvectors, and the corresponding eigenvalues of $\boldsymbol{L}_1 + \sigma\boldsymbol{D}_1 + \boldsymbol{U}_1$ be $\boldsymbol{v}_{1q}$, $\boldsymbol{v}_{1q}^*$, and μ_{1q} respectively, $q = 1, 2, \ldots, n$. Also let the normalized, right and left, bi-orthogonal eigenvectors and the corresponding eigenvalues of $\boldsymbol{L}_2 + \sigma\boldsymbol{D}_2 + \boldsymbol{U}_2$ be $\boldsymbol{v}_{2p}$, $\boldsymbol{v}_{2p}^*$, and μ_{2p}, respectively; $p = 1, 2, \ldots, m$. The conditions stated following equation (12) are sufficient to guarantee the existence of these, and also to guarantee that we can represent Φ_k, $k = 1, \ldots, t$ in the form

$$\Phi_k = \sum_{p,q} A_{kpq}\boldsymbol{v}_{2p}\boldsymbol{v}_{1q}^{*T}. \qquad (39)$$

Using the bi-orthogonality of the eigenvectors we have

$$A_{kpq} = \boldsymbol{v}_{2p}^{*T}\Phi_k\boldsymbol{v}_{1q}. \qquad (40)$$

If we now substitute (39) into (38) we obtain, after premultiplying by $\boldsymbol{v}_{2p}^{*T}$, postmultiplying by $\boldsymbol{v}_{1q}$, and using the bi-orthogonality of the eigenvectors

$$\begin{aligned}\lambda\alpha^{-2}\left(1 - \frac{\alpha e_k}{2}\right)A_{(k-1)pq} + \sqrt{\lambda}(-\sigma(2\alpha^{-2} - r_k)\\ + \mu_{2p} + \mu_{1q})A_{kpq} + \alpha^{-2}\left(1 + \frac{\alpha e_k}{2}\right)A_{(k+1)pq} = 0 \quad (41)\end{aligned}$$

If we denote by $\boldsymbol{A}$ the vector with components A_{kpq}, $k = 1, 2, \ldots, t$, and if we define $\boldsymbol{H}$ as the tridiagonal matrix whose kth row is

$$\alpha^{-2}\left(1 - \frac{\alpha e_k}{2}\right), \ -(2\alpha^{-2} - r_k), \ \alpha^{-2}\left(1 + \frac{\alpha e_k}{2}\right),$$

and set $\boldsymbol{H} = \boldsymbol{L}_3 + \boldsymbol{D}_3 + \boldsymbol{U}_3$, then equation (41) is clearly the kth row of the matrix equation

$$\{\lambda\boldsymbol{L}_3 + \sqrt{\lambda}(\sigma\boldsymbol{D}_3 + (\mu_{2p} + \mu_{1q})\boldsymbol{I}) + \boldsymbol{U}_3\}\boldsymbol{A} = 0. \quad (42)$$

Applying a Friedman transformation to equation (42) gives us

$$\{\boldsymbol{L}_3 + \sigma\boldsymbol{D}_3 + \boldsymbol{U}_3 + (\mu_{2p} + \mu_{1q})\boldsymbol{I}\}\boldsymbol{A} = 0. \qquad (43)$$

From equation (43) we see that σ is determined by the solutions of a triple eigenvalue problem which is the natural extension of equation (11). We may write this as

$$\begin{aligned}\{\boldsymbol{L}_1 + \sigma\boldsymbol{D}_1 + \boldsymbol{U}_1 - \lambda\boldsymbol{I}\}\boldsymbol{V}_1 = 0,\\ \{\boldsymbol{L}_2 + \sigma\boldsymbol{D}_2 + \boldsymbol{U}_2 - \mu\boldsymbol{I}\}\boldsymbol{V}_2 = 0,\\ \text{and} \quad \{\boldsymbol{L}_3 + \sigma\boldsymbol{D}_3 + \boldsymbol{U}_3 - \nu\boldsymbol{I}\}\boldsymbol{V}_3 = 0,\\ \text{where} \quad \lambda + \mu + \nu = 0. \qquad (44)\end{aligned}$$

The reduction of (44) in the case where some or all of $\boldsymbol{D}_1$, $\boldsymbol{D}_2$, $\boldsymbol{D}_3$ are scalar multiples of the corresponding unit matrices follows as before.

5. Calculation of the maximum value of σ

In this section we suggest a method for solving the double eigenvalue problem given by equation (11). Our procedure is a straightforward application of the initial value techniques often used in solving finite-difference approximations to the eigenvalue problems of ordinary differential equations (see, for example, Fox, 1960). The method can be applied with obvious modifications to equations (24) and (44).

We write equation (11) in the form

$$\begin{aligned}[\boldsymbol{L}_1 + \sigma\boldsymbol{D}_1 + \boldsymbol{U}_1 - \mu_1\boldsymbol{I}]\boldsymbol{v}_1 = 0,\\ [\boldsymbol{L}_2 + \sigma\boldsymbol{D}_2 + \boldsymbol{U}_2 - \mu_2\boldsymbol{I}]\boldsymbol{v}_2 = 0,\\ \mu_1 + \mu_2 \quad = 0. \qquad (45)\end{aligned}$$

Taking trial values of σ, μ_1, and μ_2 we fix the scale of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ by setting $(\boldsymbol{v}_1)_1 = (\boldsymbol{v}_2)_1 = 1$, and we use each matrix equation as a three-term recurrence for the components of the corresponding $\boldsymbol{v}$. By this means we solve the equations

$$\begin{aligned}[\boldsymbol{L}_1 + \sigma\boldsymbol{D}_1 + \boldsymbol{U}_1 - \mu_1\boldsymbol{I}]\boldsymbol{v}_1 = \beta_1\boldsymbol{e}_n,\\ \text{and} \quad [\boldsymbol{L}_2 + \sigma\boldsymbol{D}_2 + \boldsymbol{U}_2 - \mu_2\boldsymbol{I}]\boldsymbol{v}_2 = \beta_2\boldsymbol{e}_m \qquad (46)\end{aligned}$$

where $\boldsymbol{e}_n$ is the n-dimensional unit vector with one in the nth place, and $\boldsymbol{e}_m$ is defined similarly. The required values of σ and μ can now be characterized as solutions of the system

$$\begin{aligned}\beta_1(\sigma, \mu_1) = 0,\\ \beta_2(\sigma, \mu_2) = 0,\\ \mu_1 + \mu_2 = 0. \qquad (47)\end{aligned}$$

Corrections to our trial values of σ and μ can now be calculated by applying Newton's method to (47). This gives

$$\frac{\partial\beta_1}{\partial\sigma}\Delta\sigma + \frac{\partial\beta_1}{\partial\mu_1}\Delta\mu_1 + \beta_1(\sigma, \mu_1) = 0,$$

$$\frac{\partial\beta_2}{\partial\sigma}\Delta\sigma + \frac{\partial\beta_2}{\partial\mu_2}\Delta\mu_2 + \beta_2(\sigma, \mu_2) = 0,$$

$$\mu_1 + \mu_2 + \Delta\mu_1 + \Delta\mu_2 = 0. \tag{48}$$

The partial derivatives of β_1 can be computed from the the final rows of variational equations

$$[\boldsymbol{L}_1 + \sigma\boldsymbol{D}_1 + \boldsymbol{U}_1 - \mu_1\boldsymbol{I}]\frac{\partial\boldsymbol{v}_1}{\partial\sigma} = \frac{\partial\beta_1}{\partial\sigma}\boldsymbol{e}_n - \boldsymbol{D}_1\boldsymbol{v}_1,$$

$$[\boldsymbol{L}_1 + \sigma\boldsymbol{D}_1 + \boldsymbol{U}_1 - \mu_1\boldsymbol{I}]\frac{\partial\boldsymbol{v}_1}{\partial\mu_1} = \frac{\partial\beta_1}{\partial\mu_1}\boldsymbol{e}_n + \boldsymbol{v}_1, \tag{49}$$

with the initial conditions $\left(\frac{\partial\boldsymbol{v}_1}{\partial\sigma}\right)_1 = \left(\frac{\partial\boldsymbol{v}_1}{\partial\mu_1}\right)_1 = 0$, and there is a similar set of equations for the partial derivatives of β_2. We mention that by making a different choice of the scale in the calculation of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ it is possible to by-pass the solution of the variational equations in calculating the partial derivatives of β_1 and β_2. However, this may lead to a computation which is numerically unstable (Osborne, 1962).

We now discuss further some properties of equation (11). We assume that the conditions stated following equation (12) are satisfied and, in addition, that the matrices $\boldsymbol{F}$ and $\boldsymbol{G}$ are negative definite (the most usual case). If we set $\sigma = 1$ in the eigenvalue problems

$$[\boldsymbol{L}_1 + \sigma\boldsymbol{D}_1 + \boldsymbol{U}_1 - \mu_1\boldsymbol{I}]\boldsymbol{v}_1 = 0,$$

and
$$[\boldsymbol{L}_2 + \sigma\boldsymbol{D}_2 + \boldsymbol{U}_2 - \mu_2\boldsymbol{I}]\boldsymbol{v}_2 = 0 \tag{50}$$

it follows from the above assumptions that their maximum eigenvalues μ_1 and μ_2 are negative. We will show that as σ is decreased these eigenvalues increase. It then follows that there is a value of σ for which the maximum eigenvalue of one of the matrices is sufficiently positive to cancel the maximum eigenvalue of the other. This is the value of σ required in the calculation of the optimum value of ω.

The proof is straightforward, and we consider here μ_1 the maximum eigenvalue of the first problem. As σ is decreased we follow the curve $\beta_1(\sigma, \mu_1) = 0$. Using equation (48) we see that our result holds if $\partial\beta_1/\partial\sigma$ and $\partial\beta_1/\partial\mu_1$ have the same sign on this curve. To show this we use that the matrix on the left-hand side of equation (49) is singular on $\beta_1(\sigma, \mu_1) = 0$, so that the right-hand sides must be orthogonal to the eigenvector of the transposed matrix. To derive this vector we use that the tridiagonal matrix $\boldsymbol{L}_1 + \boldsymbol{U}_1$ has its off-diagonal elements onesigned, and so can be made symmetric by premultiplying by a diagonal matrix $\boldsymbol{X}$ with positive elements. The desired eigenvector is $\boldsymbol{X}\boldsymbol{v}_1$ giving the orthogonality conditions

$$\boldsymbol{v}_1^T\boldsymbol{X}\{\partial\beta_1/\partial\sigma\boldsymbol{e}_n - \boldsymbol{D}_1\boldsymbol{v}_1\} = 0,$$

and
$$\boldsymbol{v}_1^T\boldsymbol{X}\{\partial\beta_1/\partial\mu_1\boldsymbol{e}_n + \boldsymbol{v}_1\} = 0. \tag{51}$$

Thus
$$\frac{\partial\beta_1}{\partial\sigma}\bigg/\frac{\partial\beta_1}{\partial\mu_1} = -\, \boldsymbol{v}_1^T\boldsymbol{X}\boldsymbol{D}_1\boldsymbol{v}_1/\boldsymbol{v}_1^T\boldsymbol{X}\boldsymbol{v}_1 \tag{52}$$

and the desired result follows from this as all the elements of $\boldsymbol{D}_1$ are negative.

Another useful property of the solutions to equation (11) can be deduced by noting that the principal minors of the matrices occurring in equation (50) form Sturm sequences with respect to μ for fixed σ. It follows that the eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ associated with the maximum eigenvalues μ_1 and μ_2 have the property that all their components have the same sign. This suggests that a suitable choice of σ and μ to start our iteration would be one which leads to positive vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ (this property can also be used to verify that the solution produced by our iteration is the correct one). A choice compatible with the assumptions made about $\boldsymbol{F}$ and $\boldsymbol{G}$ in this section is $\sigma = 1$, $\mu_1 = \mu_2 = 0$.

To illustrate the calculation consider the case of Laplace's equation for which explicit results are available. Taking $h = k = 1$, $m = 19$, $n = 5$, $\sigma = 1$, and $\mu_1 = \mu_2 = 0$ we found the corrections to σ and μ_1 given in the Table below.

$\Delta\sigma$	$\Delta\mu_1$
$-0{\cdot}0466\,1654$	$-0{\cdot}0781\,9549$
$-0{\cdot}0214\,9499$	$-0{\cdot}0354\,4978$
$-0{\cdot}0048\,0259$	$-0{\cdot}0076\,7722$
$-0{\cdot}0002\,2850$	$-0{\cdot}0003\,3983$
$-0{\cdot}0000\,0052$	$-0{\cdot}0000\,0062$

This gives $\sigma = 0{\cdot}926\,857$, $\mu_1 = -\,0{\cdot}121\,663$ agreeing to all figures with the values of $\frac{1}{2}\left(\cos\frac{\pi}{6} + \cos\frac{\pi}{20}\right)$, and $\cos\frac{\pi}{6} - \cos\frac{\pi}{20}$ taken from Chambers' 6-figure tables.

6. Some connections with the general theory of overrelaxation

In the iterative procedures considered in this paper we may represent the unknowns (*a*) as a column vector, and (*b*) as a rectangular matrix. The representation (*b*) has an advantage for our purposes as separability of the partial differential equation is mirrored by independent operations on the rows and columns of the solution matrix, as in equations (5) and (29) for example. The representation (*a*) has the merit of much greater generality as it applies to non-rectangular regions.

In the general theory of successive over-relaxation (S.O.R.) a key result is that the components of the solution vector can be so arranged that the matrix of the set of linear equations is block-tridiagonal, with the additional property that the blocks on the leading diagonal are diagonal matrices. After this has been done it is fairly easy to show that the study of S.O.R. can be reduced to a study of a simpler iteration called the *method of simultaneous displacements* (Forsythe and Wasow, 1960, pp. 247–50). The essential steps in this development are paralleled in our work in the passage from equation (7) to equations (9) and (10). Equation (9) is the eigenvaiue problem for the method of simultaneous displacements translated into our formalism, and equation (10) which relates the eigenvalues of the S.O.R. iteration with the method of simultaneous displacements is the same in both treatments.

Certainly some at least of the results obtained in this paper can also be obtained by applying known theorems to the matrix of the eigenvalue problem for the method of simultaneous displacements. As example we give a derivation of equation (16), that for (24) being similar.

We take as our starting point equation (9)

$$[\boldsymbol{L}_2 + \sigma\boldsymbol{D}_2 + \boldsymbol{U}_2]\Phi + \Phi[\boldsymbol{L}_1 + \sigma\boldsymbol{D}_1 + \boldsymbol{U}_1] = 0.$$

We assume that the elements of the ith row of $\boldsymbol{L}_1 + \sigma\boldsymbol{D}_1 + \boldsymbol{U}_1$ are $1_i^{(1)}$, $\sigma d_i^{(1)}$, and $u_i^{(1)}$, and that the ith column of Φ is the vector $\boldsymbol{Q}_i$. With this notation (9) implies the equations

$$1_{i+1}^{(1)}\boldsymbol{Q}_{i+1} + [\boldsymbol{L}_2 + \sigma(\boldsymbol{D}_2 + d_i^{(1)}\boldsymbol{I}) + \boldsymbol{U}_2]\boldsymbol{Q}_i + u_{i-1}^{(1)}\boldsymbol{Q}_{i-1} = 0,$$
$$i = 1, 2, \ldots, n. \qquad (53)$$

Assuming that $\boldsymbol{D}_2 = a_2\boldsymbol{I}$ we may write (53) as

$$\frac{l_{i+1}^{(1)}}{a_2 + d_i^{(1)}}\boldsymbol{Q}_{i+1} + \left\{\frac{\boldsymbol{L}_2 + \boldsymbol{U}_2}{a_2 + d_i^{(1)}} + \sigma\boldsymbol{I}\right\}\boldsymbol{Q}_i + \frac{u_{i-1}^{(1)}}{a_2 + d_i^{(1)}}\boldsymbol{Q}_{i-1} = 0,$$
$$i = 1, 2, \ldots, n. \qquad (54)$$

If we now define an mn-rowed vector by the equation

$$\boldsymbol{v} = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \boldsymbol{Q}_{i-1} \\ \boldsymbol{Q}_i \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \qquad (55)$$

then we see from equation (54) that $-\sigma$ is an eigenvalue, and $\boldsymbol{v}$ an eigenvector of the block tridiagonal matrix

$$\boldsymbol{A} = \begin{bmatrix} \frac{\boldsymbol{L}_2 + \boldsymbol{U}_2}{a_2 + d_1^{(1)}}, & \frac{l_2^{(1)}}{a_2 + d_1^{(1)}}\boldsymbol{I}, & \ldots & & \\ \frac{u_1^{(1)}}{a_2 + d_2^{(1)}}\boldsymbol{I}, & \frac{\boldsymbol{L}_2 + \boldsymbol{U}_2}{a_2 + d_2^{(1)}}, & \frac{l_3^{(1)}}{a_2 + d_2^{(1)}}\boldsymbol{I} & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & \frac{u_{n-1}^{(1)}}{a_2 + d_n^{(1)}}\boldsymbol{I}, & \frac{\boldsymbol{L}_2 + \boldsymbol{U}_2}{a_2 + d_n^{(1)}} \end{bmatrix}. \qquad (56)$$

To calculate the eigenvalues of $\boldsymbol{A}$ we apply a result due to Afriat (Afriat, 1954) which may be stated as follows. Let an $mn \times mn$ matrix $\boldsymbol{M}$ have the property that it can be partitioned into blocks $\boldsymbol{M}_{ij} = R_{ij}(\boldsymbol{B})$, $i, j = 1, 2, \ldots, n$, where $R_{ij}(\boldsymbol{B})$ is a rational function of the $m \times m$ matrix $\boldsymbol{B}$, and let λ be an eigenvalue of $\boldsymbol{B}$, then the eigenvalues of the $n \times n$ matrix with elements $R_{ij}(\lambda)$ are eigenvalues of $\boldsymbol{M}$.

As $\boldsymbol{D}_2 = a_2\boldsymbol{I}$ is a multiple of the unit matrix, we have the eigenvalues of $\boldsymbol{L}_2 + \boldsymbol{U}_2$ given by equation (13) in the form $a_2\mu_{2j}$, $j = 1, 2, \ldots, m$. Applying Afriat's theorem to the matrix $\boldsymbol{A}$ we have the σ given by the eigenvalue problems

$$\left\{\begin{bmatrix} \frac{a_2\mu_{2j}}{a_2 + d_1^{(1)}}, & \frac{l_2^{(1)}}{a_2 + d_1^{(1)}} & & \\ \frac{u_1^{(1)}}{a_2 + d_2^{(1)}}, & \frac{a_2\mu_{2j}}{a_2 + d_2^{(1)}}, & \frac{l_3^{(1)}}{a_2 + d_2^{(1)}} & \\ \cdots & \cdots & \cdots & \cdots \\ & & \frac{u_{n-1}^{(1)}}{a_2 + d_n^{(1)}}, & \frac{a_2\mu_{2j}}{a_2 + d_n^{(1)}} \end{bmatrix} + \sigma\boldsymbol{I}\right\}\boldsymbol{t} = 0. \qquad (57)$$

Equation (57) is equivalent to the eigenvalue problem

$$[\boldsymbol{L}_1 + \sigma(\boldsymbol{D}_1 + a_2\boldsymbol{I}) + \boldsymbol{U}_1 + \mu_{2j}a_2\boldsymbol{I}]\boldsymbol{t} = 0$$

which is precisely the form taken by equation (16) in this case.

Strictly, as $\boldsymbol{A} + \sigma\boldsymbol{I}$ is not diagonally block-tridiagonal, the general theory of S.O.R. would associate a matrix other than $\boldsymbol{A}$ with the eigenvalue problem for the method of simultaneous displacement, the actual form of this matrix depending on the ordering of the components of the solution vector. However, there will be a permutation matrix $\boldsymbol{P}$ such that this matrix is equal to $\boldsymbol{PAP}^T$ (Forsythe and Wasow, 1960, pp. 242–45), and this transformation does not change the eigenvalues of $\boldsymbol{A}$.

7. In conclusion

In this paper we have considered only the five-point difference approximation to partial differential equations in two independent variables. For the more accurate nine-point formula to be applicable we must have $f(x)$ and $g(y)$ equal zero in equation (1). We find that the approach used in this paper works only for the case of line over-relaxation, and here we require the further condition that at least one of $p(x)$ and $q(y)$ must vanish identically.

We wish to acknowledge the assistance of Dr. D. W. Martin of the National Physical Laboratory who, besides providing the derivation of equation (16) given in section 6, has made many suggestions concerning presentation which have greatly improved the paper.

References

Afriat, S. N. (1954). "Composite Matrices," *Quart. J. Math.*, Vol. 5, pp. 81–98.

Bickley, W. S., and McNamee, J. (1960). "Matrix and other Direct Methods for the Solution of Systems of Linear Difference Equations, *Phil. Trans.*, 1005, Vol. 252, pp. 69–131.

Forsythe, S. E., and Wasow, W. R. (1960). *Finite Difference Methods for Partial Differential Equations*, Wiley.

Fox, L. (1960). *Some Numerical Experiments with Eigenvalue Problems in Ordinary Differential Equations, Boundary Problems in Differential Equations* (R. E. Langer, editor), University of Wisconsin Press.

Martin, D. W., and Tee, G. J. (1961). "Iterative Methods for Linear Equations with Symmetric, Positive Definite Matrix," *The Computer Journal*, Vol. 4, pp. 242–54.

Osborne, M. R. (1962). "A Note on Finite-Difference Methods for solving the Eigenvalue Problems of Second-order Differential Equations," *Mathematics of Computation*, Vol. 16, pp. 338–46.

The LL^T and QR methods for symmetric tridiagonal matrices

By James M. Ortega and Henry F. Kaiser

Introduction

It is known (Rutishauser, 1958, 1959 and 1960) that for a positive definite real symmetric matrix A_1 the algorithm:

$$\left.\begin{array}{l}\text{Decompose } \boldsymbol{A}_i \text{ into } \boldsymbol{L}_i\boldsymbol{L}_i^T;\\ \text{Form } \boldsymbol{A}_{i+1} = \boldsymbol{L}_i^T\boldsymbol{L}_i\end{array}\right\} i = 1, 2, \ldots$$

where the $\boldsymbol{L}_i$ are lower triangular matrices and $\boldsymbol{L}_i^T$ is the transpose of L_i, produces a sequence of matrices $\boldsymbol{A}_i$ which are all similar to $\boldsymbol{A}_1$ and which converge to a diagonal matrix $\boldsymbol{A}$.

A similar algorithm due to Francis (1961):

$$\left.\begin{array}{l}\text{Decompose } \boldsymbol{A}_i \text{ into } \boldsymbol{Q}_i\boldsymbol{R}_i;\\ \text{Form } \boldsymbol{A}_{i+1} = \boldsymbol{R}_i\boldsymbol{Q}_i\end{array}\right\} i = 1, 2, \ldots$$

where the $\boldsymbol{Q}_i$ are orthogonal and the $\boldsymbol{R}_i$ are upper triangular, also produces a sequence of matrices which tend to a diagonal matrix. This algorithm has advantage that the similarity transformations involved are orthogonal congruences.

Now when $\boldsymbol{A}_1$ is tridiagonal, that is $a_{ij} = 0$ for $|i - j| > 1$, then the matrices $\boldsymbol{A}_i$ produced by either algorithm are also tridiagonal [Rutishauser (1958) and Francis (1961)] and we might hope that these algorithms would be effective for the important problem of finding the eigenvalues of $\boldsymbol{A}_1$. If we carry out the algorithms in a natural way, however, it would seem that each iteration would require n square roots and consequently would be inefficient compared with the Sturm sequence process now in common use.

It is the purpose of this paper to show that both algorithms can be carried out using no square roots.* Limited experiments have shown that with accelerating techniques involving translations of the origin (see, e.g., Rutishauser, 1959 and 1960), these modified algorithms produce eigenvalues of high accuracy three to ten times as fast as the Sturm process.

The modified LL^T algorithm

Let $\boldsymbol{A}_1$ be a real symmetric positive definite tridiagonal matrix with diagonal elements $a_1, \ldots, a_n$ and off-diagonal elements $b_1, \ldots, b_{n-1}$, and let $A_2\ A_3 \ldots$ be the matrices produced by the LL^T algorithm. Our modified algorithm produces the diagonal elements and squares of the off-diagonal elements of the A_i, and clearly this is sufficient since the signs of the off-diagonal elements do not affect the eigenvalues.

Let $\bar{a}_1, \ldots, \bar{a}_n$ and $\bar{b}_1, \ldots, \bar{b}_{n-1}$ be the diagonal and off-diagonal elements of $\boldsymbol{A}_2$; it will suffice to describe the transition from $a_1, \ldots, a_n$ and $b_1^2, \ldots, b_{n-1}^2$ to $\bar{a}_1, \ldots, \bar{a}_n$ and $\bar{b}_1^2, \ldots, \bar{b}_{n-1}^2$. Now $\boldsymbol{A}_2 = \boldsymbol{L}_1^T\boldsymbol{L}_1$ where $\boldsymbol{L}_1\boldsymbol{L}_1^T = \boldsymbol{A}_1$. If we let $d_1, \ldots, d_n$ and $s_1, \ldots, s_{n-1}$ be the diagonal and sub-diagonal elements of $\boldsymbol{L}_1$ and carry out the indicated matrix multiplications we then obtain:

$$\begin{array}{r}d_1^2 = a_1;\\ \left.\begin{array}{r}s_i^2 = b_i^2/d_i^2\\ \bar{a}_i = d_i^2 + s_i^2\\ d_{i+1}^2 = a_{i+1} - s_i^2\\ \bar{b}_i^2 = d_{i+1}^2 s_i^2\end{array}\right\} i = 1, \ldots, n-1;\\ \bar{a}_n = d_n^2.\end{array}$$

To carry out the algorithm in practice only two temporary storage registers are required to retain the current s_i^2 and d_i^2; the old a_i and b_i^2 are also replaced by $\bar{a}_i^2$ and $\bar{b}_i^2$ immediately, so that a total of only $2n + 1$ storage positions are needed. Each iteration requires no square roots, $n - 1$ divisions, $n - 1$ multiplications and $2(n - 1)$ additions.

Note that we have not included any acceleration procedure in the algorithm, and one must be added to make this a practical process.

The modified QR algorithm

Again let $a_1, \ldots, a_n$ and $b_1, \ldots, b_{n-1}$ be the diagonal and off-diagonal elements of A_1, and $\bar{a}_1 \ldots, \bar{a}_n$ and $\bar{b}_1, \ldots, \bar{b}_{n-1}$ the corresponding elements of A_2 where now

$$A_2 = \boldsymbol{R}_1\boldsymbol{Q}_1$$

and
$$A_1 = \boldsymbol{Q}_1\boldsymbol{R}_1$$

where $\boldsymbol{Q}_1$ is orthogonal and $\boldsymbol{R}_1$ upper triangular. As with the modified LL^T algorithm it will suffice to describe the transition from $a_1, \ldots, a_n$ and $b_1^2, \ldots, b_{n-1}^2$ to $\bar{a}_1, \ldots, \bar{a}_n$ and $\bar{b}_1^2, \ldots, \bar{b}_{n-1}^2$.

* J. H. Wilkinson, of Teddington, England, has also noted that the LL^T algorithm can be carried out without square roots for tridiagonal matrices [private communication].