and T. R. Scott, "Network analysis using a sparse tableau with tree selection to increase sparseness," *in Proc. IEEE Int. Symp. Circuit Theory* (Toronto, Ont., Canada, Apr. 1973), pp. 165–168.
[12] *IBM Advanced Statistical Analysis Program*, ASTAP manual GH20-1271-0.
[13] A. E. Ruehli, "Electrical analysis of interconnections in a solid-state circuit environment," in *Dig. IEEE Int. Solid-State Circuit Conf.* (New York, 1972), pp. 64–65 and p. 216.
[14] E. C. Jordan and K. G. Balmain, *Electro-Magnetic Waves and Radiating Systems*. Englewood Cliffs, N. J.: Prentice-Hall, 1968, ch. 14.
[15] A. Gopinath and P. Silvester, "Calculation of inductance of finite-length strips and its variation with frequency," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-21, pp. 380–386, June 1973.
[16] R. F. Harrington, *Field Computation by Moment Methods*. New York: Macmillan, 1968, ch. 1 and 4.

An Iterative Approach to the Finite-Element Method in Field Problems

W. KINSNER, MEMBER, IEEE, AND EDWARD DELLA TORRE, SENIOR MEMBER, IEEE

***Abstract*—An iterative approach to the finite-element method is presented. Several finite-element formulations are presented for the Laplace, Poisson, and Helmholtz equations. These formulations permit iterative solutions. The convergence of the vector sequences generated by the iterative method is accelerated using successive extrapolation and other methods. Accuracy and convergence of the solutions are discussed.**

I. Introduction

THE THEORETICAL background of the finite-element method has been given by Aubin [1]. Other authors [2], [3] have introduced some practical aspects of the method as applied to structural mechanics. Silvester [4] and others [5], [16], [20]–[22] discussed the method as applied to the electromagnetic field problems. Convergence of the method, as a function of the relative size of the discretizing elements and the order of the approximating polynomials, is discussed in many recent mathematical and technical journals [6]. In particular, explicit discretization errors are given in [1], [2], [6], and [7], and some experimental results are given in [5] and [8].

The variational formulation of waveguide problems, using complete polynomials, leads to the general eigenvalue problem

$$A\boldsymbol{x} = \lambda B\boldsymbol{x} \tag{1}$$

where A and B are symmetric positive-definite $n \times n$ band matrices, λ is the eigenvalue(s) and $\boldsymbol{x}$ the eigenvector(s) associated with the particular eigenvalue(s). It is noted that for the H modes A is only semidefinite. The finite-difference formulation of these problems also leads to (1), however, the matrix B is the identity matrix, and A is not necessarily symmetric.

A variety of methods for solving (1) have been presented (e.g., [9], [10], and [12]). The finite-difference solutions most frequently employ iterative techniques and the finite-element solutions almost exclusively use direct methods.

Iterative methods for solving eigenvalue problems may be divided into two categories. The first methods use the fact that eigenvectors of a system form a linearly independent set which spans an n dimensional space. The methods in the second category use the property that the generalized Rayleigh quotient

$$Q = \frac{\boldsymbol{x}^T A\boldsymbol{x}}{\boldsymbol{x}^T B\boldsymbol{x}} \tag{2}$$

is equal to an eigenvalue and is stationary when $\boldsymbol{x}$ is the corresponding eigenvector. A method using this property with Fletcher–Powell iteration has been described [13]. These methods combined with the deflation or orthogonalization yield, however, only partial eigensolutions, i.e., the dominant and several closest eigenvectors. An iterative method for the complete eigensolution shall be presented.

II. Finite-Element Formulation

Let R be either a simply or multiply-connected bounded region in an n dimensional space V^n with boundary Γ. The boundary Γ consists of a finite number of closed, nonintersecting hypersurfaces $\Gamma_h (h = 0, \cdots, \tau)$ such

Manuscript received June 26, 1973; revised November 7, 1973. This work was supported by the National Research Council of Canada.

The authors are with the Group on Simulation, Optimization, and Control, Department of Electrical Engineering, McMaster University, Hamilton, Ont., Canada.

that $\Gamma_h \subseteq \Gamma$. Fig. 1 shows a two-dimensional nonconvex multiply-connected region R.

It can be shown [14] that the solution of the inhomogeneous Helmholtz equation $\nabla^2\rho + \lambda\rho = g$, subject to the associated natural boundary conditions, $\rho(\partial\rho/\partial n)$ equal to zero [18, p. 208], [19], is equivalent to minimizing the following functional:

$$F = \frac{1}{2}\iint_R (|\nabla\rho|^2 + \lambda\rho^2)\, dR - \iint_R \rho g\, dR \qquad (3)$$

where ρ is the function that is sought, g is the corresponding source density function, and λ is a constant. Instead of finding the true ρ over R the finite-element formulation employs, in general, the subdomains E_e such that each $E_e \subset R$. Any proper assembly of elements into a connected model $\bar{R}$ gives an approximate global solution ϕ over the region $\bar{R}$. It is possible to develop different finite-element schemes for the same problem. In the iterative solutions, a proper embedding of the elements expedites the iteration processes.

The discrete model of R is constructed in several steps. First, a finite number N of global nodes is identified in R and labelled $P_g(g = 1, 2, \cdots, N)$. These points may lie in R or on Γ, and are contained in another region $\bar{R}$, with boundary $\bar{\Gamma}$, which approximates R. Then, a finite number M of elements E_e is defined on the global points. Now, each element is a closed region, and the elements are disjoint. The global nodes associated with a given element are called the local nodes $P_l(l = 0,1, \cdots, L)$. Assembling the elements into $\bar{R}$ completes the process. This assembly employs a grouping of m elements adjacent to a specific node. There may be J adjoining nodes. The procedure for developing a finite-element model function ϕ of a continuous function ρ is closely related to that for $\bar{R}$ since $\bar{R}$ is the domain of the function ϕ.

For simplicity, the paper will consider only two-dimensional regions and triangular elements in homogeneous isotropic media. The finite-element model function ϕ_e within an element shall be represented by the first-order complete polynomial of the form

$$\phi_e = \alpha_1 + \alpha_2 x + \alpha_3 y \qquad (4)$$

where x and y are the local Cartesian coordinates, and the α's are coefficients to be determined by the finite-element formulation.

A. Laplace's Equation

When $\lambda = 0$ and $g = 0$, the differential equation reduces to the Laplace equation. The functional (3) is then reduced to

$$F = \frac{1}{2}\iint_{\bar{R}} |\nabla\phi|^2\, d\bar{R}. \qquad (5)$$

In this section, a formula is obtained which minimizes (5) and depends only on the angles of the triangular elements.

Let E_e be a triangle in $\bar{R}$ with vertices and their associated angles, as shown in Fig. 2. The subscripts of all the entities in Fig. 2 refer to a triangle in $\bar{R}$ and, at the same time, to a triangle in a group of adjoining triangles (see Fig. 1). The local coordinates are located with the x coordinate coincident with l_{0j}, and y coordinate intersecting node P_0. When ϕ_0 is perturbed and all other nodal values are kept constant, the functional changes only in the m triangles adjoining node P_0. If one denotes by F_0 the part of the functional associated with these triangles then

$$F_0 = \sum_{j=1}^{m} [|\nabla\phi_x|^2 + |\nabla\phi_y|^2]A_j \qquad (6)$$

where $\nabla\phi_i$ is the component of the gradient in the i direction $(i = x,y)$, and A_j is the area of the triangle. For the specific nodal point P_0, one can write

$$\frac{\partial F}{\partial\phi_0} = \frac{\partial F_0}{\partial\phi_0} = \sum_{j=1}^{m} 2A\frac{\partial\nabla\phi_y}{\partial\phi_0}\nabla\phi_y \overset{\text{set}}{=} 0. \qquad (7)$$

Note that the x component of the gradient vanishes in (7)

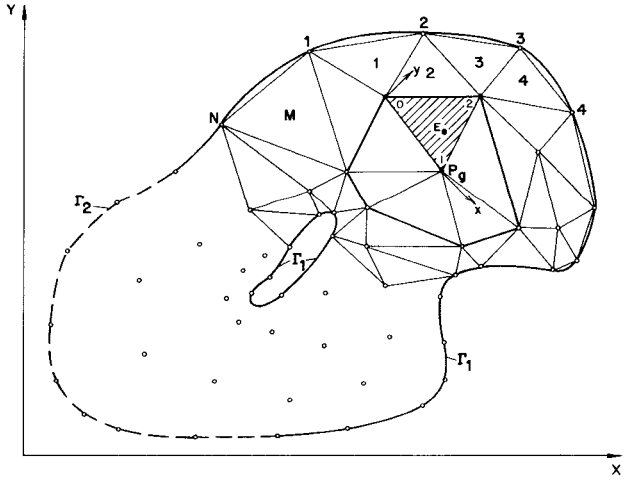

Fig. 1. A general nonconvex two-dimensional multiply connected region with two types of boundary, placed in global coordinates X, Y. Triangulation of the region with M elements and N global nodes. Local coordinates x, y are connected with an element E_e.

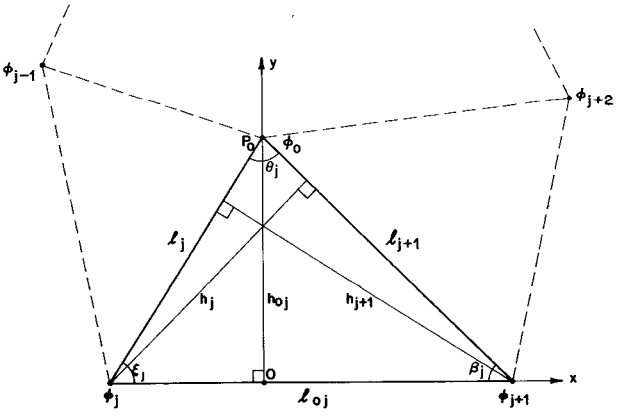

Fig. 2. One of m triangles adjacent to a node P_0. The origin of local coordinates lies on $\overline{P_1P_2}$.

since it is constant for this perturbation. This leads to

$$\phi_0 = \frac{\sum_{j=1}^{m} (\phi_j \cot \beta_j + \phi_{j+1} \cot \xi_j)}{\sum_{j=1}^{m} (\cot \beta_j + \cot \xi_j)}$$

$$= \frac{1}{k_0} \sum_{j=1}^{m} (\phi_j k_{j+1} + \phi_{j+1} k_j) \quad (8)$$

where k_0 is a constant for a collection of adjoining triangles (all k's may be computed once for a given geometry), and ϕ_j, ϕ_{j+1} are the nodal values of the approximating function ϕ_e.

It is readily seen that, if $\beta_j = \xi_j$ for all j, then (8) reduces to

$$\phi_0 = \frac{1}{m} \sum_{j=1}^{m} \phi_j \quad (9)$$

which is identical to the finite-difference formula for an equilateral polygonal mesh. For the square mesh shown in Fig. 3, (8) becomes the ordinary 5-point Laplacian operator for all nodes no matter how the triangulation of R is performed, and the value ϕ_0 is the average of the four adjacent values. This implies that the optimum interpolation formula for a finite-difference solution using the 5-point operator is the linear interpolation formula. It is noted that there are many different linear interpolations for a given set of nodes as shown in Fig. 3, however, the optimum values for the nodes is the same in all these cases. For a solution of a given problem, a consistent manner of interpolation must be used over the whole region $\bar{R}$.

B. Poisson's Equation

For Poisson's equation, the functional (3) to be minimized for natural boundary conditions reduces to

$$F = \frac{1}{2} \iint_{\bar{R}} |\nabla \phi|^2 \, d\bar{R} - \iint_{\bar{R}} \phi g \, d\bar{R}. \quad (10)$$

Since a variation of the function ϕ_e at a specific node can only effect the energy in the m adjoining triangles then, to minimize the functional, it is sufficient to consider these triangles only. Thus

$$\frac{\partial F}{\partial \phi_0} = \sum_{j=1}^{m} \frac{\partial}{\partial \phi_0} (r_j + s_j) = 0 \quad (11)$$

where r_j and s_j correspond to the first and the second integral. It can be shown that

$$\sum_{j=1}^{m} \frac{\partial r_j}{\partial \phi_0} = \phi_0 \sum_{j=1}^{m} k_{j0} + \sum_{j=1}^{m} (\phi_1 k_{j1} + \phi_2 k_{j2}) \quad (12)$$

where the constants k_{jl}, which are different for each triangle, are given by

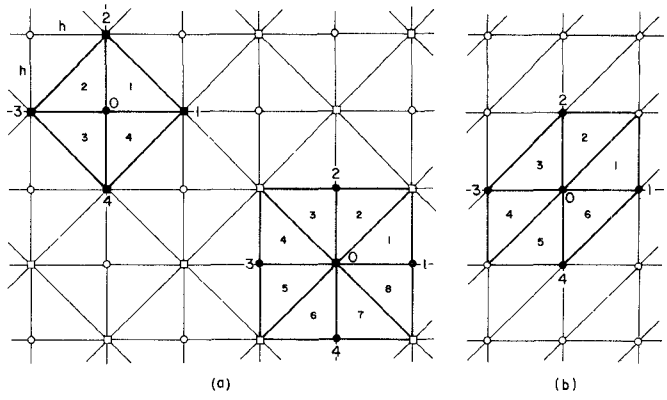

Fig. 3. Five-point Laplacian operators resulting from various triangulations of the region. (a) Two types of nodes with 4 adjacent triangles ○ and 8 adjacent triangles □. (b) One type of node with 6 adjacent triangles.

$$k_{j0} = \frac{1}{4A_j} [(y_1 - y_2)^2 + (x_2 - x_1)^2] \quad (13a)$$

$$k_{j1} = \frac{1}{4A_j} [(y_1 - y_2)(y_2 - y_0) + (x_2 - x_1)(x_0 - x_2)] \quad (13b)$$

$$k_{j2} = \frac{1}{4A_j} [(y_1 - y_2)(y_0 - y_1) + (x_2 - x_1)(x_1 - x_0)] \quad (13c)$$

where A_j is the area of a triangle. Also

$$\sum_{j=1}^{m} \frac{\partial s_j}{\partial \phi_0} = k_0 \quad (14)$$

where k_0 is a constant which may be computed once for a group of adjoining triangles in numerical solutions. Substituting (12) and (14) into (11) gives

$$\phi_0 = \frac{-1}{\sum_{j=1}^{m} k_{j0}} \left[\sum_{j=1}^{m} (\phi_1 k_{j1} + \phi_2 k_{j2}) + k_0 \right]. \quad (15)$$

For a square grid with a mesh length h, (15) reduces to a 5-point operator

$$\phi_0 = \tfrac{1}{4}[\phi_1 + \phi_2 + \phi_3 + \phi_4 + h^2 g]. \quad (16)$$

C. Helmholtz's Equation

Similarly, for Helmholtz's equation it can be shown that the potential ϕ_g at an arbitrary global node P_g expressed in terms of both the potentials at the adjacent nodes and the local (or global) coordinates of these nodes is given by

$$\phi_0 = \frac{1}{\sum_{j=1}^{m} (k_{j0} - \lambda k_{j3})} \cdot \sum_{j=1}^{m} [\lambda(\phi_1 + \phi_2) k_{j12} - (\phi_1 k_{j1} + \phi_2 k_{j2})] \quad (17)$$

where

$$k_{j0} = \frac{1}{4A_j}[(x_2 - x_1)^2 + (y_2 - y_1)^2], \qquad \forall E_j \tag{18a}$$

$$k_{j1} = \frac{-1}{4A_j}[(x_2 - x_1)x_2 + (y_2 - y_1)y_2], \qquad \forall E_j \tag{18b}$$

$$k_{j2} = \frac{-1}{4A_j}[(x_1 - x_2)x_1 + (y_1 - y_2)y_1], \qquad \forall E_j \tag{18c}$$

$$k_{j3} = \tfrac{1}{6}A_j \tag{18d}$$

$$k_{j12} = \tfrac{1}{2}k_{j3} \tag{18e}$$

and the area of the triangle is

$$A_j = \tfrac{1}{2}(x_1y_2 - x_2y_1), \qquad \forall E_j. \tag{19}$$

Equations (18a)–(18c) are analogous to (13a)–(13c) in the simplified positioning of the triangle.

For clarity, (17) may be rewritten as

$$\phi_0 = \frac{\lambda k_3 - k_2}{k_1 - \lambda k_0} \tag{20}$$

where

$$k_1 = \sum_{j=1}^{m} k_{j0} \qquad k_0 = \sum_{j=1}^{m} k_{j3} \tag{21}$$

$$k_2(\phi) = \sum_{j=1}^{m} (\phi_1 k_{j1} + \phi_2 k_{j2}) \qquad k_3(\phi) = \sum_{j=1}^{m} (\phi_1 + \phi_2)_j k_{j12}.$$

For an internal node of a periodic lattice, as shown in Fig. 4, (21) reduces to

$$k_1 = 4 \tag{22a}$$

$$k_2(\phi) = -(\phi_1 + \phi_2 + \phi_3 + \phi_4) \tag{22b}$$

$$k_0 = \tfrac{1}{2}h^2 \tag{22c}$$

$$k_3(\phi) = \tfrac{1}{2}(\phi_1 + \phi_2 + \phi_3 + \phi_4 + \phi_5 + \phi_6)k_0 \tag{22d}$$

and (20) then yields

$$\phi_0 = \frac{1}{4 - \frac{1}{2}h^2\lambda}\left[\frac{h^2}{12}\lambda(\phi_1 + \phi_2 + \phi_3 + \phi_4 + \phi_5 + \phi_6) + \phi_1 + \phi_2 + \phi_3 + \phi_4\right]$$

$$= \frac{1}{8 - \lambda h^2}\left[\left(2 - \frac{\lambda h^2}{6}\right)(\phi_1 + \phi_2 + \phi_3 + \phi_4) + \frac{\lambda h^2}{6}(\phi_5 + \phi_6)\right]. \tag{23}$$

It is seen that if $\sum_{j=1}^{6} \phi_j$ is approximated by $6\phi_0$ then (22) becomes identical to the 5-point finite-difference operator

$$\phi_0 = \frac{1}{4 - \lambda h^2}(\phi_1 + \phi_2 + \phi_3 + \phi_4) \tag{24}$$

which in turn is a special case of a more general finite-difference approximation when $(\phi_1 + \phi_2)k_{j12} \approx \phi_0 k_{j3}$.

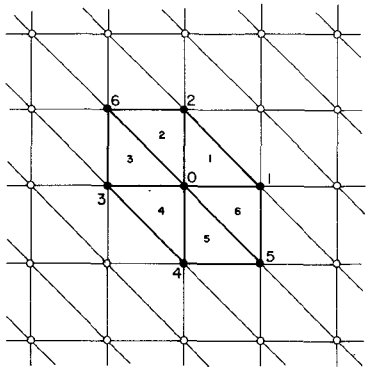

Fig. 4. A seven-point finite-element operator for the dominant solution of Helmholtz's equation.

III. Acceleration of Convergence

Two general approaches exist to the problem of accelerating the convergence of the finite-element method to reduce the computation time required for a complete or partial eigensolution to a given accuracy. The first one refers to accelerating the direct solution of $A\phi = \lambda B\phi$ as given in existing programs [8] and the other refers to accelerating the iterative methods as described in the previous section.

It is noted that convergence is understood here as the diminishing of some norm of an error between the initial solution and the approximate solution ϕ on $\bar{R}$ rather than between ϕ and the true solution ρ on R when either the largest side or the minimum angle of all of the triangles is varied.

A. Wilkinson's Algorithm

In the first approach, the most efficient algorithm [17] for obtaining eigenvalues of the generalized matrix eigenvalue problem, in which the matrices are of a band type and only specific eigenvalues are required, should be used in the direct methods. The eigenvectors are then computed by the Wielandt inverse iteration [11, p. 321].

B. Successive Extrapolation

The other technique relates to accelerating convergent vector sequences generated by iterative methods. Such sequences have been investigated [24], [25], and a successive extrapolated relaxation technique has been proposed for accelerating their convergence [23]. After every third iteration, a modification of Aitken's δ^2 process is used to extrapolate the linearly convergent vector sequence to its limit. Extrapolation is improved if the rate of convergence of the original sequence is known. Therefore, more efficient extrapolation can be employed if the rate is estimated during the iteration process. This scheme of nested iteration is terminated when a specific error criterion is met.

When a linearly convergent vector sequence oscillates globally or locally then the effectiveness of the successive extrapolation method decreases. However, it has been shown [25] that if this method cannot improve the con-

vergence of the original sequence then it does not decelerate it. Both oscillations and their frequencies of the vector sequences depend on the eigenvalues of matrices involved in the processes generating these sequences. If the eigenvalues are complex then a vector sequence oscillates. The successive extrapolation process [24] improves convergence of such sequences providing the period of the oscillations is longer than 5–10 iterations. These frequencies are often lower in the iterative processes [26].

C. Chebyshev Acceleration

When some information concerning the distribution of the eigenvalues μ_i of the iteration matrix $\mathcal{L}$ is available, and the eigenvalues are real then the Chebyshev polynomials [15, p. 301] can be used to accelerate the convergence of different methods such as the Jacobi, successive overrelaxation (SOR), Richardson's, and various second-degree iterative methods [15]. This method assumes the convergence of a vector sequence upon which it should act. The strategy is to find a function which would not only describe the convergence property of the vector sequence based on the eigenvalues of $\mathcal{L}$ but, at the same time, would cause some acceleration of the convergence by projecting the terms of the sequence towards its limit. Since for a convergent sequence the spectral radius of $\mathcal{L}$ has to be less than one, then the Chebyshev polynomials offer the means of weighting the projections. When the eigenvalues μ_i are distributed between $0 \leq |\mu_i| \leq |a| < 1$ where a is a constant then the appropriate linear combinations of some terms of the vector sequence $\{\phi\}$ may be obtained by means of the three-term recurrence relation

$$\phi^{(n+1)} = \phi^{(n)} + \frac{T_{n-2}(1/a)}{T_n(1/a)} [\phi^{(n)} - \phi^{(n-2)}] \tag{25}$$

where $\phi^{(n)}$ is the vector obtained by the iterative process and T_n is the Chebyshev polynomial of degree n in x defined by $T_n(x) = \cos(n \cos^{-1} x)$ in the range $|x| < 1$.

The Chebyshev acceleration of Richardson's method

$$\phi^{(n+1)} = \phi^{(n)} + \beta^{(n)} r^{(n)} \tag{26}$$

where $\delta^{(n)}$ is the residual vector at nth iteration, has the form of the second-degree Richardson's method

$$\phi^{(n+1)} = \phi^{(n)} + \beta^{(n)} r^{(n)} + \gamma^{(n)} \delta^{(n)} \tag{27}$$

where $\phi^{(n)} = \phi^{(n)} - \phi^{(n-1)}$ is the displacement vector and the coefficients are given by

$$\beta^{(n)} = \frac{4}{b - a} \frac{\cosh nt}{\cosh (n+1)t} \qquad \gamma^{(n)} = \frac{\cosh (n-1)t}{\cosh (n+1)t} \tag{28}$$

where $\cosh t = (b + a)/(b - a)$, and a, b are the smallest and largest eigenvalues of $\mathcal{L}$. This process is nonstationary since the values of β and γ depend on n.

The Chebyshev acceleration of SSOR when applied for a biharmonic problem causes its convergence to be twice as fast as SOR [27]. Young [15] estimates an order faster convergence when this type of acceleration is applied to some methods. A serious limitation on this method is posed by the lower bound a for μ_i [14, p. 228]. It is easy to find a close estimate to the upper bound b but determination of a is very difficult. It appears that the effectiveness of the method depends mainly on a.

D. Gradient Methods

Let $f(\phi)$ be a quadratic function defined by

$$f(\phi) \triangleq \phi^T A \phi - 2\phi^T b + b^T A^{-1} b \tag{29}$$

where ϕ^T denotes transpose of ϕ, and A and b correspond to the quantities in $Ax = b$. When A is symmetric and positive definite then the $f(\phi)$ takes its minimum value, zero, at $A^{-1}b$. Various minimization methods of $f(\phi)$ have been discussed [28]. Interesting combinations of iterative methods (accelerated Jacobi, SSOR) with some simple gradient methods and with the Chebyshev acceleration have been discussed, and corresponding numerical results on the biharmonic operator have been presented [27].

The major feature of some of the gradient iterative methods is that they lead to the solution in a finite number of steps. One can use this powerful feature of the gradient methods by combining the best minimization techniques (see [28]) with the finite-element method for Laplace's and Poisson's equations or several eigensolutions of the Helmholtz equation. Such a combined method should start from a direct or iterative solution of the finite-element equations. The process then could proceed with an iterative refinement by a finite-difference method on a fine mesh, however, this intermediate step is not necessary. Finally, the process could use one of the efficient gradient methods to find the minimum of (29) or another objective function associated with the given problem. These minimization iterations would start from an initial vector close to the solution thus the minimum would be obtained quickly and the process would be reliable since (29) represents, in this case, a set of concentric hypersurfaces. It is noted, however, that the very efficient gradient methods require existence of the first and second derivatives of a function. The model function ϕ constructed in the finite element is continuous, but the first derivative is discontinuous at the boundary between the elements, thus the second derivative may be indeterminate at these boundaries. This restriction can be removed by redefining the function to be continuous with continuous derivatives. This model of ϕ describes the actual ϱ more accurately than the finite-element model, and, therefore, the gradient vector containing the first partial derivatives and the Hessian matrix with second partial derivatives can be constructed.

The number of iterations can be reduced considerably when using the gradient methods but the amount of work involved in computing the gradient vector and the Hessian matrix may actually decrease the efficiency of the method below that of other simpler methods.

E. Mesh Refinement

Another technique for accelerating the convergence is mesh halving. In general, an application of bisection to all the sides of triangles or rectangles as shown in Fig. 5 leads to a finer mesh. Since the convergence depends on the starting values at the nodal points then a quick solution on a coarse mesh with a small number of nodes can serve as the starting values for the next cycle of iterations. This mesh halving can be repeated, leading to a better first approximation to ϱ (see next section). The solution obtained by mesh halving is faster than that carried on the finer mesh only.

The advantage of using bisection rather than other techniques of the mesh refinement is that the bisection maintains the same angles in the new elements. This fact removes the necessity of recomputing all the coefficients k in (8). The "old" nodes have exactly the same k's, and the coefficients for the "new" nodes are chosen from the old k's.

Suppose the discretization error of the particular algorithm used in the finite-element solution of ϱ is of order $\| h \|^p$ where $\| h \|$ is the Chebyshev or some other norm of the side length of all the triangles, and p is some number. Let $\{\phi_k\}$ be a sequence of solutions, generated by the same algorithms, using a corresponding sequence $\{h_k\}$ of decreasing h on the same region $\bar{R}$. When $\{h_k\} \to 0$ then $\{\phi_k\} \to \varrho$. Since $\{\phi_k\}$ converges monotonically to ϱ the following relation holds:

$$\frac{\| e_{k+1} \|}{\| e_k \|} = \frac{\| h_{k+1} \|^p}{\| h_k \|^p} \tag{30}$$

where the discretization error vector e_k at kth subdivision of the elements is given by $e_k = \phi_\infty - \phi_k$.

Any norm of e_k, including the vector itself, satisfies (30) since p depends on the choice of the norm. Hence one can extrapolate the sequence $\{\phi_k\}$ from two or more successive solution vectors ϕ_k obtained on two or more triangulations with decreasing size of the elements. For only two solutions ($k = 1,2$) one obtains the following extrapolation formula

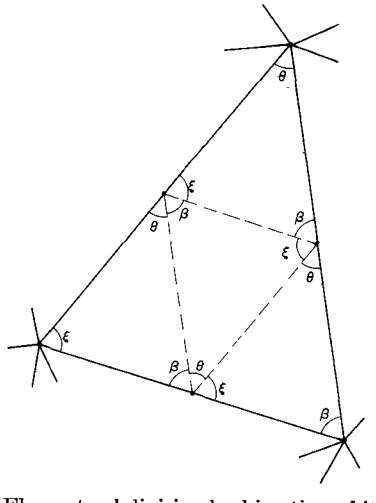

Fig. 5. Element subdivision by bisection of its sides.

$$\phi_E = \frac{\phi_1 \| h_2 \|^p - \phi_2 \| h_1 \|^p}{\| h_2 \|^p - \| h_1 \|^p} \tag{31}$$

where ϕ_E indicates the extrapolated vector, not necessarily ϱ. This difference is caused by various factors affecting $\{\phi_k\}$ such as the cumulative roundoff errors, and the residual error of ϕ_k. A better extrapolation formula results from three solutions ($k = 1,2,3$)

$$\phi_E = \frac{\phi_1 s_1 + \phi_2 s_2 + \phi_3 s_3}{s_1 + s_2 + s_3} \tag{32}$$

where

$$s_l = \frac{\| h_m \|^p - \| h_n \|^p}{\| h_l \|^p} \tag{33}$$

and l, m, n are even permutation of 1,2,3.

The extrapolated values ϕ_E do not, in general, result in ϱ even though p is known in advance. However, when the cumulative roundoff error is small and when the change in $\| h_k \|$ is appropriately chosen, then a significant improvement in accuracy is achieved by this extrapolation. An assessment of p will be given in the following section.

IV. Accuracy and Convergence

At the beginning of the previous section, two concepts of convergence were introduced: 1) the convergence of ϕ to ϱ, and 2) the convergence of successive estimates of ϕ to the model function solution ϕ. Similarly, there are two types of errors: 1) the accuracy of the model function ϕ, related to the original function ϱ, i.e., discretization error, and 2) the accuracy of the estimate solution $\phi^{(n)}$ with respect to the actual function ϕ, i.e., residual error. This distinction has been neglected in some cases and the finite-element results presented were not as accurate as expected [29].

A. Discretization Error

Recall that, by assumption, the function ρ is continuous with continuous partial derivatives in a region R bounded by Γ. When triangulation of $\bar{R}$ is employed and the first-order polynomial is used to approximate ρ within each triangle then the model polyhedral function ϕ is continuous with continuous first derivatives within each triangle and discontinuous at junctions between the triangles. The second partial derivatives vanish within each triangle and may be, in general, infinite at the junctions. We shall now show that a given function ρ and its first partial derivatives may be approximated arbitrarily closely by a polyhedral function ϕ based on a suitable traingulation of R.

The strategy of assembling ϕ (see Section II) guarantees the continuity of ϕ and an appropriate "fitting" of ϕ to ρ. Therefore, only an arbitrary closed triangular element E_e is considered in the proof. Let $E_e \subset \bar{R}$ and the vertices $p_i(x_i,y_i) \in E_e$. Let $\phi_i \in \rho$, that is, the values of the model function at the vertices coincide with the exact function ρ. This result is obtained by minimization of the functional

(5) for each nodal point and for the linear interpolation over E_e. For higher approximations more points have to satisfy this coincidence condition. It is noted that this minimization process is similar to the least squares. It might be possible to construct another functional (objective function) whose least pth minimization could lead to a better approximation.

If one defines h and θ as the largest side and the largest angle of all triangles being associated with one triangle, then it can be shown that the following conclusions hold: the approximation for the derivatives is of order h

$$| \nabla(\phi - \rho) | \leq \frac{1}{\cos \frac{1}{2}\theta} Mh \tag{34}$$

and the approximation for the function is of order h^2:

$$| \phi - \rho | \leq \frac{1}{\cos \frac{1}{2}\theta} Mh^2 \tag{35}$$

where $M \geq 0$ is a bound on ρ''. Equations (34) and (35) indicate that the function ρ and its derivatives ρ' can be approximated by the first-order polynomials based on a triangular mesh arbitrarily close by taking a sufficiently fine triangulation. It is noted that thin triangles give poor approximation, since $\cos \frac{1}{2}\theta \to 0$.

Equation (35) shows that the approximate solution ϕ converges to the exact solution ρ with the same order of the convergence rate as in the case of the finite-difference approximation. However, there is an important difference between the two methods, evident from this analysis, that is, this order of the convergence rate requires a bounded second derivative of ρ whereas, in the finite-difference formulation, a bound on the third-order derivative is required.

Equation (35) can be generalized by removing the assumption that the largest angle θ is associated with the triangle having the longest side. In this case, however, the derivation becomes complicated since the convergence has to be expressed in terms of a norm in the Sobolev space of functions having generalized derivatives up to the order k inclusive, on a compact support [1, p. 138]. Using these norms, following Zlámal's approach [6], it can be shown that for second-order finite-element approximation to the generalized second-order partial differentiation equations the following is true:

$$|| \phi - \rho ||_S \leq k \frac{1}{\sin \nu} M_3 h^2 \tag{36}$$

where the constant k does not depend on the triangulation, ν is the smallest angle of all triangles, h is the largest side of all triangles, and the subscript at M indicates the order of the derivatives that have to be bounded. The subscript S indicates a norm in the Sobolev sense.

The approximation by cubic polynomials leads to the following:

$$|| \phi - \rho ||_S \leq k \frac{1}{\sin \nu} M_4 h^3. \tag{37}$$

For the same problems, the finite-difference formulation gives the convergence rates of the order of h and h^2, respectively.

For the biharmonic equation approximated by the fifth-order polynomial one gets

$$|| \phi - \rho ||_S \leq k \frac{1}{\sin^2 \nu} M_6 h^4. \tag{38}$$

It is noted that the largest angle θ in (35) can describe a triangle better than the smallest angle ν, since the latter indicates one small angle only, whereas the former always implies *two* small angles when $\theta \to 180°$.

B. Residual Error and Error Criteria

In order to terminate an iterative process at some accuracy it is necessary to determine some measure of the error. The simplest measure of the error is the residual measure e_R which indicates the fractional deviation of $\nabla^2\phi^{(n)}$ from zero at nth iteration

$$e_R = \frac{|| \nabla\phi^{(n)} ||_E}{|| \phi^{(n)} ||_E}. \tag{39}$$

This ratio employs the Euclidean norms with sieving function, and it can be shown that it reduces to

$$e_R = \Big[\sum_{g=1}^{N} (\nabla\phi_g^{(n)})^2 / \sum_{g=1}^{N} \phi_g^{(n)} \Big]^{1/2}. \tag{40}$$

Another measure is defined by

$$e_F = \int_\Gamma \left| \frac{\partial\phi^{(n)}}{\partial n} \right|_\Gamma d\Gamma \Big/ \int_{\Gamma_i} \left| \frac{\partial\phi^{(n)}}{\partial n} \right|_{\Gamma_i} d\Gamma_i. \tag{41}$$

It shows the flux imbalance between the flux calculated along the entire boundary Γ and the flux which enters the region across Γ_i.

Since the exact solution ϱ is not known in general, it is impossible to relate the error norms e_R and e_F to an error norm of the exact solution, e.g.,

$$e_\rho = \frac{|| \phi^{(n)} - \varrho ||}{|| \varrho ||} \tag{42}$$

nevertheless some test problems for which the exact solution is known can qualitatively relate (40) and (41) to (42). The knowledge of this relation may be applied to other problems. It has been experimentally found [5] that, for homogeneous and inhomogeneous Dirichlet, Cauchy, and Neumann problems, e_R and e_F are approximately proportional to e_ρ, and they decrease when the order of the approximating polynomial increases.

V. Conclusions

An iterative approach to the finite-element method has been presented in which no matrices and essentially only the solution vector has to be stored. The relationship between the finite-element, the finite-difference, and the gradient methods is discussed.

Various techniques have been discussed for accelerating

the convergence of the vector sequence to its limit which is the exact solution to the modeling problem (residual error), and accelerating the convergence of the sequence of modeling solutions to its limit which is the exact solution of the field problem (discretization error). Of these techniques successive extrapolation is the most effective in obtaining the limit of a vector sequence. In order to obtain the limit of the sequence of modeling function one must use successive mesh refinement. A technique for element subdivision is presented which introduces new elements with no new angles thereby simplifying the obtaining of the coefficients of the new assembly matrix whose order is twice that of the original assembly matrix.

A discussion of discretization shows that the convergence rate varies as h^2. This permits to obtain an optimal extrapolation technique for estimating the exact solution with finite size elements.

References

[1] J. P. Aubin, *Approximation of Elliptic Boundary-Value Problems*. New York: Wiley, 1972.

[2] J. T. Oden, *Finite Elements of Nonlinear Continua*. New York: McGraw-Hill, 1972.

[3] O. C. Zienkiewicz, *The Finite Element Method in Engineering Science*. London: McGraw-Hill, 1971.

[4] P. Silvester, "High-order polynomial triangular finite elements for potential problems," *Int. J. Eng. Sci.*, vol. 7, pp. 849–861, 1969.

[5] D. J. Richards and A. Wexler, "Finite-element solutions within curved boundaries," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-20, pp. 650–657, Oct. 1972.

[6] M. Zlamal, "On the finite element method," *Numer. Math.*, vol. 12, pp. 394–409, 1968.

[7] Y. Yamamoto and N. Tokuda, "A note on convergence of finite element solutions," *Int. J. Numer. Math. Eng.*, vol. 3, pp. 485–493, 1971.

[8] P. Silvester, "A general high-order finite-element waveguide analysis program," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-17, pp. 204–210, Apr. 1969.

[9] J. B. Davies, "Review of methods for numerical solution of the hollow waveguide problem," *Proc. Inst. Elec. Eng.*, vol. 119, pp. 33–37, Jan. 1972.

[10] F. L. Ng, "Numerical solution of the hollow waveguide problem," *Canterbury Eng. J.*, New Zealand, pp. 17–48, 1972.

[11] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*. London: Oxford, 1965.

[12] G. Cantin, "An equation solver of very large capacity," *Int. J. Numer. Math. Eng.*, vol. 3, pp. 379–388, 1971.

[13] W. W. Bradbury and R. Fletcher, "New iterative methods for solution of the eigenproblem," *Numer. Math.*, vol. 9, pp. 259–267, 1966.

[14] G. E. Forsythe and W. R. Wasow, *Finite-Difference Methods for Partial Differential Equations*. New York: Wiley, 1960, p. 182.

[15] D. M. Young, *Iterative Solution of Large Linear Systems*. New York: Academic, 1971.

[16] S. Ahmed and P. Daly, "Waveguide solutions by the finite-element method," *Radio Electron. Eng.*, vol. 38, pp. 217–223, Oct. 1969.

[17] G. Peters and J. H. Wilkinson, "Eigenvalues of $Ax - \lambda Bx$ with band symmetric A and B," *Comput. J.*, vol. 12, pp. 398–404, 1969.

[18] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, vol. 1, 1st Engl. ed. New York: Interscience, 1953.

[19] T. G. Hazel and A. Wexler, "Variational formulation of the Dirichlet boundary condition," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-20, pp. 385–390, June 1972.

[20] P. Silvester, H. S. Cabayan, and B. T. Browne, "Efficient techniques for finite element analysis of electric machines," *IEEE Trans. Power App. Syst.*, vol. PAS-92, pp. 1274–1281, July/Aug. 1973.

[21] O. W. Anderson, "Iterative solution of finite element equations in magnetic field problems," presented at the IEEE PAS Summer Meeting, T72411-7, San Francisco, Calif., July 1972.

[22] K. Pontoppidan, "Finite-element techniques applied to waveguides of arbitrary cross section, Part I and II," Lab. Electromagnetic Theory, Technical Univ. Denmark, Lyngby, Rep. LD19, Sept. 1971.

[23] E. Della Torre and W. Kinsner, "Solution to waveguide problems by successive extrapolated relaxation," *IEEE Trans. Microwave Theory Tech.* (Short Papers), vol. MTT-21, pp. 490–491, July 1973.

[24] ——, "Convergence properties of the successive extrapolated relaxation (SER) method," *J. Inst. Math. Its Appl.*, vol. 12, pp. 175–185, 1973.

[25] W. Kinsner and E. Della Torre, "Convergence acceleration of vector sequences," submitted to *Numer. Math.*

[26] V. Ashkenazi, "Geodetic normal equations," in *Large Sparse Sets of Linear Equations*, J. K. Reid, Ed. New York: Academic, 1971, pp. 57–73.

[27] M. Engeli, T. Ginsburg, H. Rutishauser, and E. Stiefel, "Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems," in *Mitteilungen ans dem Institut für angewandte Mathematik*. Basel/Stuttgart: Birkhäuser Verlag, 1959, no. 8.

[28] J. W. Bandler, "Optimization methods for computer-aided design," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-17, pp. 533–552, Aug. 1969.

[29] E. F. Fuchs and E. A. Erdelyi, "Nonlinear theory of turboalternators Part II. Load dependent synchronous reactances," *IEEE Trans. Power App. Syst.* (Discussion), vol. PAS-92, pp. 592–598, Mar./Apr. 1973.