Lawrence Berkeley National Laboratory
Recent Work

Title

MIXED EXPLICIT-IMPLICIT INTERATIVE FINITE ELEMENT SCHEME FOR DIFFUSION-TYPE PROBLEMS: I. THEORY

Permalink

https://escholarship.org/uc/item/2d990802

Author

Neuman, S.P.

Publication Date

1975-08-01

eScholarship.org Powered by the California Digital Library
University of California

0 0 0 0 4 4 0 5 4 7 2 1

Submitted to International Journal for
Numerical Methods in Engineering

LBL-4405
Preprint

MIXED EXPLICIT-IMPLICIT ITERATIVE FINITE ELEMENT SCHEME FOR DIFFUSION-TYPE PROBLEMS: I. THEORY

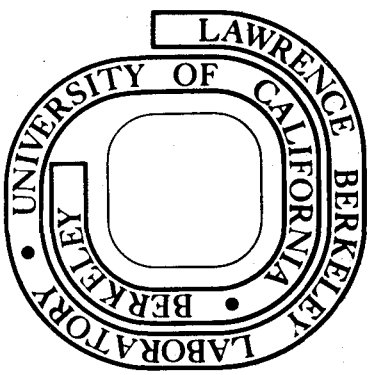
S. P. Neuman and T. N. Narasimhan

August 1975

Prepared for the U. S. Energy Research and
Development Administration under Contract W-7405-ENG-48

For Reference

Not to be taken from this room



LBL-4405

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

0 0 0 0 4 4 0 4 7 2 2

MIXED EXPLICIT-IMPLICIT ITERATIVE FINITE ELEMENT SCHEME FOR DIFFUSION-TYPE PROBLEMS: I. THEORY

by

S. P. Neuman[1,2] and T. N. Narasimhan[2]

Summary

A Galerkin finite element formulation of diffusion processes based on a diagonal capacity matrix is analyzed from the standpoint of local stability and convergence. The theoretical analysis assumes that the conductance matrix is locally diagonally dominant, and it is shown that one can always construct a finite element network of linear triangles satisfying this condition. Time derivatives are replaced by finite differences, leading to a mixed explicit-implicit system of algebraic equations which can be efficiently solved by a point iterative technique. In this work the accelerated point iterative method is adopted and is shown to converge when the conductance matrix is locally diagonally dominant. Several examples are included in Part II of this paper to demonstrate the efficiency of the new approach.

---

[1]Lawrence Berkeley Laboratory, Berkeley, California 94720.

[2]Department of Civil Engineering, University of California, Berkeley, California 94720.

Introduction

Finite element formulations of parabolic equations such as those governing chemical diffusion, heat conduction, and fluid flow through porous media often lead to a system of first-order linear differential equations of the form[1]

$$[\underline{\underline{A}}]\ \{\underline{h}\ (t)\} + [\underline{\underline{D}}]\ \{\dot{\underline{h}}\ (t)\} = \{\underline{Q}\ (t)\} \qquad (1)$$

where $[\underline{\underline{A}}]$ is the conductance or stiffness matrix, $[\underline{\underline{D}}]$ is the capacity matrix, $\{\underline{h}\}$ is the dependent variable vector (e.g., hydraulic head in a groundwater system), and $\{\underline{Q}\}$ is a flux vector representing sources or sinks.

In general, the capacity matrix $[\underline{\underline{D}}]$ includes non-zero off-diagonal terms (i.e., it is non-diagonal) and there is evidence in the literature suggesting that this may lead to conceptual as well as numerical difficulties. For example, a recent analysis by Narasimhan[2] indicates that a non-diagonal $[\underline{\underline{D}}]$ matrix may upset the maintenance of local mass or energy balance, although overall balance over the entire region may still be preserved. This may perhaps explain why as our experience indicates, equation (1) sometimes yields physically unrealistic values of $\{\underline{h}\}$ when there is a sudden and drastic change in $\{Q\}$, and why this can be remedied by diagonalizing the $[\underline{\underline{D}}]$ matrix as has been done by Wilson,[3] Emery and Carson,[4] and others. Narasimhan's theory may also explain why Neuman[5,6] was forced to diagonalize the $[\underline{\underline{D}}]$ matrix in dealing with the highly nonlinear problem of saturated-unsaturated groundwater flow; otherwise, the finite element scheme would not converge. Obviously, the use of a diagonal capacity matrix also results in increased computational efficiency as compared to a non-diagonal

0 0 0 0 4 4 0 4 7 2 3

matrix owing to lesser storage requirements and fewer algebraic operations in calculating the matrix terms.

The purpose of this paper is to discuss an alternative form for equation (1) with a diagonal capacity matrix which we believe has many advantages over the traditional approach. The resulting differential equations are discretized in time by finite differences which enables one to treat them either explicitly, or implicitly, or by an optimum combination of both schemes. The implicit equations are solved by a point iterative technique rather than by a direct method such as Gaussian elimination. Part I of the paper examines local stability and convergence criteria for the explicit-implicit scheme as well as convergence of the proposed point iterative technique. The theoretical analysis assumes that the matrix $[\underline{\underline{A}}]$ is locally diagonally dominant; at the end of the text it is shown that one can always construct a finite element mesh satisfying this condition. Part II describes various aspects of the solution strategy including questions related to choice of time step size, choice of relative weights of explicit and implicit terms to be assigned during any given time step, initial guess for iterative scheme, treatment of nonlinearities, etc.; and provides examples to illustrate the capabilities of the new approach.

Explicit-Implicit Formulation

Consider the particular diffusion-type equation for fluid flow in an anisotropic porous medium

$$\nabla \cdot (\underline{\underline{K}} \nabla h) = C \frac{\partial h}{\partial t} \qquad (2)$$

where h is hydraulic head, $\underline{\underline{K}}$ is hydraulic conductivity tensor, and C is specific storage or fluid capacity (defined as volume of fluid instantaneously released from storage per unit bulk volume of porous medium when

h is lowered by one unit). To discretize equation (2) in space we adopt a network of triangular elements for plane flow and of concentric rings of constant triangular cross-section for axisymmetric problems. In each individual element, the hydraulic head is described approximately in terms of linear shape functions and the values h_n of head at the corner (node) points. The next step is to apply the Galerkin method[1,5] to equation (2). However, since the Galerkin method is applicable only at a given instant of time, the time derivative $\partial h/\partial t$ must be determined independently of the Galerkin orthogonalization process. Thus, instead of replacing h in the time derivative by the approximating Galerkin sequence as is usually done in the finite element approach, one is justified to define the nodal values of $\partial h/\partial t$ as averages over the exclusive domains associated with each node (every nodal point is associated with one third of each adjoining element;[3] for further details the reader is referred to Narasimhan[7]). This leads to a system of first-order linear differential equations of the form

$$[\underline{\underline{A}}]\ \{\underline{h}\ (t)\} + [\underline{\underline{D^*}}]\ \{\dot{\underline{h}}\ (t)\} = \{\underline{Q}\ (t)\} \tag{3}$$

which is identical with (1) except for the capacity matrix, $[\underline{\underline{D^*}}]$, which does not include any non-zero off-diagonal terms. The individual terms of the matrices $[\underline{\underline{A}}]$ and $[\underline{\underline{D^*}}]$ are given by [3,5]

$$A_{nm} = \sum_e \frac{\delta}{4S} \left[K_{xx}\ b_n\ b_m + K_{xy}\ (b_n\ c_m + b_m\ c_n) + K_{yy}\ c_n\ c_m\right] \tag{4}$$

$$D^*_{nm} = \begin{cases} \sum_e \frac{\delta S}{3}\ C & \text{if } n = m \\ 0 & \text{if } n \neq m \end{cases} \tag{5}$$

0 0 0 0 4 4 0 4 7 2 4

where S is area of triangle, $\delta = 1$ for plane flow and $\delta = 2\pi\bar{r}$ for axisymmetric flow ($\bar{r}$ being average radius of triangle), b and c are geometric coefficients defined in Appendix A, and the summation sign applies to all elements adjacent to nodal point n. It can be shown[3,7] that $-A_{nm}$ represents the rate of fluid transfer into the exclusive subdomain of node n (one third of each adjacent triangle) due to unit difference in head between nodes m and n. On the other hand, D^*_{nn} is the fluid capacity of the exclusive subdomain associated with node n.

If we replace the time derivatives in equation (3) by finite differences and introduce a weighting factor, θ, we obtain a system of simultaneous linear algebraic equations of the form

$$[\underline{\underline{A}}]\,[\theta\,\{\underline{h}^{k+1}\} + (1-\theta)\,\{\underline{h}^{k}\}] + [\underline{\underline{D}}^*]\,\frac{\{\underline{h}^{k+1}\} - \{\underline{h}^{k}\}}{\Delta t} = \{\underline{Q}\} \qquad (6)$$

where $0 \leq \theta \leq 1$, Δt is time increment, and k indicates number of time steps. Defining a new term

$$\lambda_{nm} = \frac{\Delta t \cdot A_{nm}}{D^*_{nn}} \qquad (7)$$

and recognizing from equation (47) in Appendix A that

$$\lambda_{nn} = -\sum_{m \neq n} \lambda_{nm} \qquad (8)$$

we can rewrite (6) as

$$h_n^{k+1} - h_n^{k} = \theta \sum_{m \neq n} \lambda_{nm} \left(h_n^{k+1} - h_m^{k+1}\right) + (1-\theta) \sum_{m \neq n} \lambda_{nm} \left(h_n^{k} - h_m^{k}\right) + Q_n \Delta t / D^*_{nn}$$

$$n = 1, 2, \cdots, N \qquad (9)$$

where N is total number of nodes and the summation is taken over all values of m = 1, 2, . . . , N other than m = n. When $\theta = 0$, all the values of h_n^{k+1} can be calculated explicitly from the system of equations (9), which now corresponds to a forward difference scheme in time. When $\theta = 1$, the result is a fully implicit backward difference scheme, whereas $\theta = \frac{1}{2}$ corresponds to a time-centered or Crank-Nicholson[8] scheme. It is important to note that equation (9) does not include any diagonal terms of the matrix $[\underline{\underline{A}}]$, a fact which may save a considerable amount of storage and computer time, especially in dealing with nonlinear problems where the matrix must be recomputed at each time step.

Local Stability Criteria

The local stability of (9) at any given node, n, can be conveniently analyzed by using the von Neumann harmonic approach.[8,9,10] According to this approach, the solution h_n^k can be viewed as the sum of an exact solution of (9), h_n^{*k}, and an error, ε_n^k. The initial error is expressed as a complex Fourier series

$$\varepsilon_n^o = \sum_{p=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} B_{pr} \exp\,(ipx_n + iry_n) \tag{10}$$

where p and r are integers, i is $\sqrt{-1}$, and B_{pr} are Fourier coefficients. The error at any later time is expressed as

$$\varepsilon_n^k = \sum_{p=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} B_{pr}\, \xi_{pr}^{(k)} \exp\,(ipx_n + iry_n) \tag{11}$$

where the growth (or decay) factor, ξ_{pr}, is raised to the k-th power. Since h_n^k as well as h_n^{*k} satisfy (9), it follows that ε_n^k satisfies a similar equation but one which does not include sinks or sources, Q_n. If we substitute a typical term from (11) into (9) without Q_n and solve for ξ_{pr},

0 0 0 0 4 4 0 4 7 2 5

we obtain

$$\xi_{pr} = \frac{1 + (1-\theta) \sum_{m \neq n} \lambda_{nm} [1 - \exp (ip \Delta x_m + ir \Delta y_m)]}{1 - \theta \sum_{m \neq n} \lambda_{nm} [1 - \exp (ip \Delta x_m + ir \Delta y_m)]} \tag{12}$$

where $\Delta x_m = x_m - x_n$ and $\Delta y_m = y_m - y_n$. Let us assume that the matrix $[\underline{\underline{\lambda}}]$ is locally diagonally dominant, i.e., that $\lambda_{nn} \geq \sum_{m \neq n} |\lambda_{nm}|$ for some n; later we will show that one can always construct a finite element mesh which satisfies this requirement at any given node. As is shown in Appendix A, this means that all the off-diagonal terms of λ_{nm} are non-positive and therefore ξ_{pr} can be rewritten as

$$\xi_{pr} = \frac{1 - (1-\theta) \sum_{m \neq n} |\lambda_{nm}| [1 - \exp (ip \Delta x_m + ir \Delta y_m)]}{1 + \theta \sum_{m \neq n} |\lambda_{nm}| [1 - \exp (ip \Delta x_m + ir \Delta y_m)]} \tag{13}$$

Decomposing ξ_{pr} in (13) into its real and imaginary parts yields

$$\begin{aligned} \text{Re } (\xi_{pr}) = \Bigg[& 1 - (1-2\theta) \sum_{m \neq n} |\lambda_{nm}| [1 - \cos (p \Delta x_m + r \Delta y_m)] \\ & - \theta (1-\theta) \left\{ \sum_{m \neq n} |\lambda_{nm}| [1 - \cos (p \Delta x_m + r \Delta y_m)] \right\}^2 \\ & - \theta (1-\theta) \left\{ \sum_{m \neq n} |\lambda_{nm}| \sin (p \Delta x_m + r \Delta y_m) \right\}^2 \Bigg] / \text{den} \end{aligned} \tag{14a}$$

$$\text{Im } (\xi_{pr}) = 1/\text{den} \tag{14b}$$

where

$$\begin{aligned} \text{den} = & \left\{ 1 + \theta \sum_{m \neq n} |\lambda_{nm}| [1 - \cos (p \Delta x_m + r \Delta y_m)] \right\}^2 \\ & + \theta^2 \left\{ \sum_{m \neq n} |\lambda_{nm}| \sin (p \Delta x_m + r \Delta y_m) \right\}^2 \end{aligned}$$

Since the denominator is always greater than or equal to 1, it is obvious from (14b) that $|\text{Im}\ (\xi_{pr})| \leq 1$ under all circumstances. On the other hand, Re (ξ_{pr}) can never exceed 1 and therefore the system will be locally stable (i.e., the error at node n will not grow) whenever Re $(\xi_{pr}) \geq -1$. The most negative value of Re (ξ_{pr}) occurs when cos $(p\ \Delta x_m + r\ \Delta y_m) = -1$ and sin $(p\ \Delta x_m + r\ \Delta y_m) = 0$, in which case the above requirement reduces to

$$\text{Re}\ (\xi_{pr}) = \frac{1 + 2\ (\theta-1) \sum_{m \neq n} |\lambda_{nm}|}{1 + 2\theta \sum_{m \neq n} |\lambda_{nm}|} \geq -1 \qquad (15)$$

Again, the most negative value occurs when the sum in equation (15) tends to infinity, in which case

$$\lim_{\sum_{m \neq n} |\lambda_{nm}| \to \infty} \text{Re}\ (\xi_{pr}) = \frac{\theta - 1}{\theta} \geq -1 \qquad (16)$$

Equation (16) is satisfied whenever $\theta \geq 0.5$, indicating that our scheme is unconditionally stable at node n for all values of θ which are not less than 0.5.

If $\theta < 0.5$, stability is conditioned upon equation (15). Since $[\underline{\underline{\lambda}}]$ is diagonally dominant near n, (8) implies that

$$\lambda_{nn} = \sum_{m \neq n} |\lambda_{nm}| \qquad (17)$$

and therefore the local stability criterion given by equation (15) can be expressed as

$$\lambda_{nn} \leq \frac{1}{1 - 2\theta} \qquad \text{or} \qquad \Delta t \leq \frac{D_{nn}^{*}}{(1-2\theta)\ A_{nn}}\ ;\ \theta < 0.5 \qquad (18)$$

0 0 0 0 4 4 0 4 7 2 6

For example, the explicit version of equation (9) is stable at any node at which the ratio between capacity and conductance is large enough so that $\Delta t \leq D^*_{nn}/A_{nn}$ for any given Δt. Conversely, the explicit scheme can be made stable at all nodes by choosing a sufficiently small Δt.

The stability criterion $\Delta t \leq D^*_{nn}/A_{nn}$ for the explicit scheme can also be derived by physical reasoning. According to equation (47) in Appendix A, A_{nn} is the sum of all fluxes entering into the exclusive subdomain of node n when heads at all adjacent nodes, h_m, exceed h_n by unity. On the other hand, D^*_{nn} is the capacity of the subdomain to absorb fluid when h_n changes by one unit. Thus, the above stability criterion merely states that the amount of fluid entering into the subdomain of n must not exceed the capacity of the subdomain to absorb fluid. A value of Δt in excess of what is prescribed by the stability criterion would imply that h_n must change by more than unity, which is contrary to physics (recall that in the explicit scheme, the values of h_m remain fixed during a time step) and may therefore lead to uncontrolled local oscillations in the values of h_n.

Local Convergence Criteria

We now turn our attention to an equally important question: Under what conditions does the approximate solution obtained from the numerical scheme converge to the exact solution of the differential equation as the mesh is made finer and finer? Many finite element formulations are known to converge in the mean; however, they seldom guarantee convergence at a point. The purpose of this section is to demonstrate that the explicit-implicit scheme in (9) converges to the exact solution of the partial differential equation (2) at every node satisfying the following requirements: a) The node is not a sink or a source; b) the node does not lie on a material boundary; c) the

node lies in an isotropic domain; and d) all elements in the immediate neighborhood of the node form a mesh having at least two orthogonal axes of symmetry with respect to the positions of the nodes (for example, a mesh consisting locally of equilateral triangles such as in Figure 1a, or a rectangular nodal pattern such as in Figure 1b). Later in the text we will show how these criteria can be extended to anisotropic domains.

To prove this, consider a homogeneous isotropic region with a superimposed finite element mesh consisting of equilateral triangles (see Figure 1a), with a spacing a^1 between the nodes, and a constant time increment Δt^1. Suppose that one wishes to improve the calculation at a given time t by repeating it with successively finer and finer meshes. The value of $k = t/\Delta t$ corresponding to the fixed time t will, of course, tend to infinity as $\Delta t \to 0$, and it is therefore obvious that the mesh must be refined in such a manner as to constantly maintain stability. Therefore, consider another mesh with increments given by $a = a^1/J$, $\Delta t = \Delta t^1/J^2$ (J = integer) and assume that a^1 and Δt^1 have been chosen so that

$$\lambda_{nn} = \frac{\Delta t^1 \cdot A^1_{nn}}{D^{*1}_{nn}} = \frac{\Delta t \cdot A_{nn}}{D^{*}_{nn}} \leq \frac{1}{1 - 2\theta} \tag{19}$$

for all $\theta < 0.5$, in accordance with the stability criterion in (18) (for $\theta \geq 0.5$, the values of a^1 and Δt^1 can be chosen arbitrarily). Then, we can prove the following

Theorem: Let h_n^k denote the approximate value of h (x,y,t) obtained at a fixed node n at time $t = k\,\Delta t$ by solving the finite element equations (9). Then h_n^k converges to the exact solution of the partial differential equation (2), h, as $J \to \infty$.

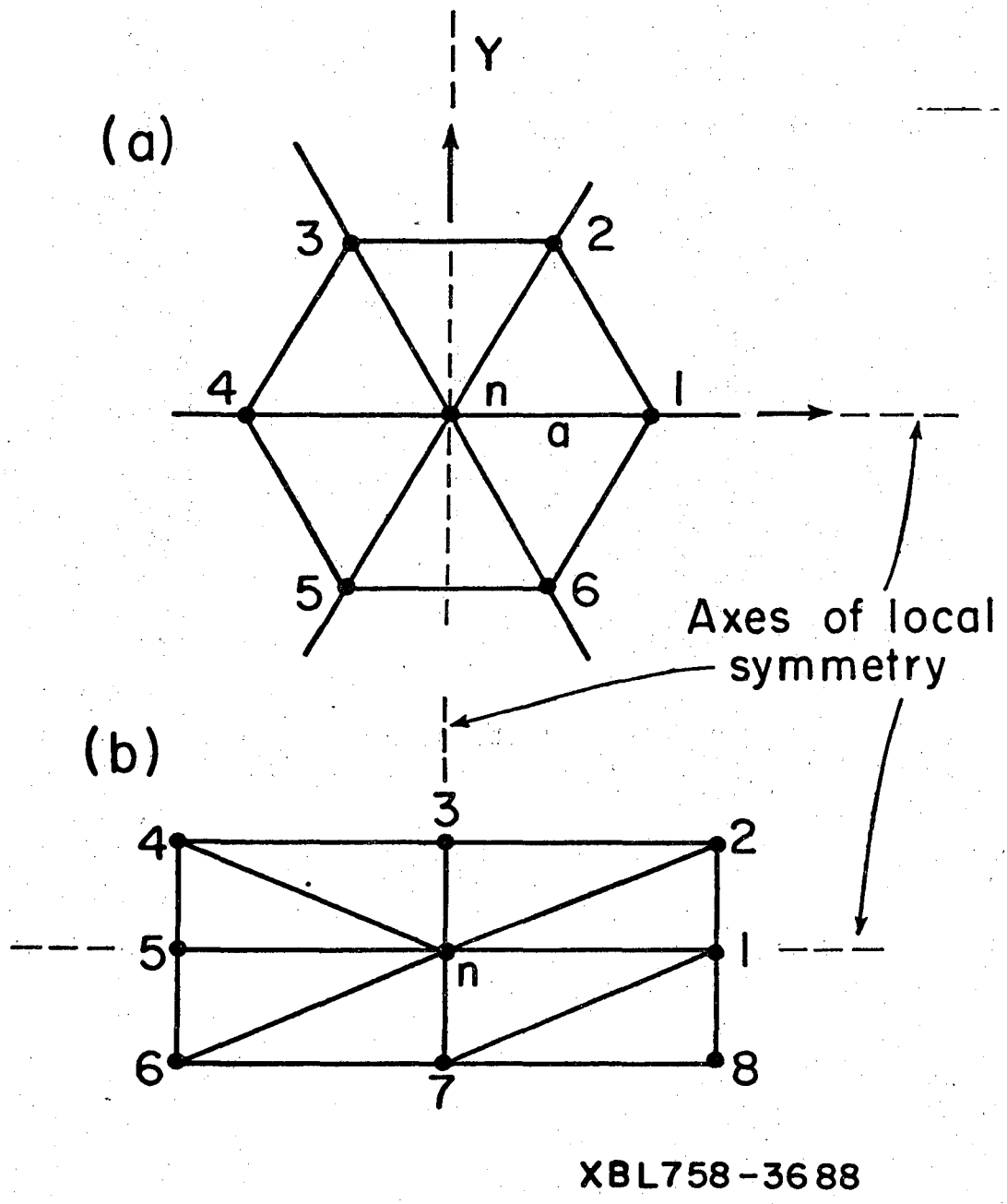


Figure 1. Examples of nodal patterns with two orthogonal axes of local symmetry.

Proof: The solution of (2) when $\underline{\underline{K}}$ is a scalar can be expressed in the form

$$h = \sum_{p=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} H_{pr} \exp\left[ipx + iry - (p^2 + r^2)\,\alpha\, t\right] ;\ t \geq 0 \qquad (20a)$$

where $\alpha = K/C$ and H_{pr} are Fourier coefficients which can be determined from the known initial conditions (the initial conditions are expressed by the Fourier series obtained from (20a) upon setting $t = 0$). Since node n is not a sink or a source, the solution of the algebraic equations (9) can be written, in analogy to (11), as

$$h_n^k = \sum_{p=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} H_{pr}\, \xi_{pr}^{(k)} \exp\left(ipx_n + iry_n\right) \qquad (20b)$$

where ξ_{pr} is given by (12). Let p_o and r_o be arbitrary positive integers. Then, according to (20a) and (20b), we can write for a given J,

$$|h_n^k - h| = \left|\left(\sum_{|p|\leq p_o} \sum_{|r|\leq r_o} + \sum_{|p|>p_o \cup |r|>r_o}\right) H_{pr} \exp\left(ipx_n + iry_n\right) \cdot \left(\xi_{pr}^{t/\Delta t} - \eta^{t/\Delta t}\right)\right| \leq |\Sigma_1| + |\Sigma_2| \qquad (21)$$

where Σ_1 and Σ_2 stand for the first and second sum, respectively, and $\eta = \exp\left[-(p^2 + r^2)\,\alpha\,\Delta t\right]$. The second sum satisfies

$$|\Sigma_2| \leq 2 \sum_{|p|>p_o \cup |r|>r_o} |H_{pr}| \qquad (22)$$

because $|\xi_{pr}| \leq 1$ due to stability and $|\eta| \leq 1$ due to positivity of α and Δt; therefore, $|\Sigma_2|$ can be made arbitrarily small by choosing sufficiently large values of p_o and/or r_o, since the Fourier series in (11) is known to be absolutely convergent.

0 0 0 0 4 4 0 4 7 2 8

To estimate $|\Sigma_1|$ we first note that

$$\left|\xi_{pr}^{t/\Delta t} - \eta^{t/\Delta t}\right| = \left|\xi_{pr}^{k} - \eta^{k}\right| = \left|\left(\xi_{pr} - \eta\right)\left(\xi_{pr}^{k-1} + \xi_{pr}^{k-2}\eta + \cdots + \eta^{k-1}\right)\right|$$

$$\leq \left|\xi_{pr} - \eta\right| k \tag{23}$$

Furthermore, η can be expanded as

$$\eta = 1 - (p^2 + r^2)\,\alpha\,\Delta t + \frac{1}{2!}(p^2 + r^2)^2\,(\alpha\,\Delta t)^2 - \cdots \tag{24}$$

and, according to equation (57) in Appendix B,

$$\xi_{pr} = 1 - (p^2 + r^2)\,\alpha\,\Delta t + \left(\frac{1}{4\lambda_{nn}} + 2\theta\right)(\alpha\,\Delta t)^2\,(p^2 + r^2)^2 + \cdots \tag{25}$$

Thus, remembering that λ_{nn} is constant, we find that

$$\frac{|\xi_{pr} - \eta|}{(p^2 + r^2)^2\,(\Delta t)^2}$$

is an analytic function of $(p^2 + r^2)\,\Delta t$ in some neighborhood of the origin, and is therefore bounded for all non-negative $(p^2 + r^2)\Delta t$. Let this bound be M. Then

$$|\Sigma_1| \leq \sum_{|p|\leq p_o} \sum_{|r|\leq r_o} (p^2 + r^2)^2\,(\Delta t)^2\,M\,\frac{t}{\Delta t}\,|H_{pr}|$$

$$\leq (p_o^2 + r_o^2)\,\Delta t\,M\,t \sum_{p=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} |H_{pr}| \tag{26}$$

Having chosen p_o and/or r_o large to make $|\Sigma_2|$ small, we can now choose a sufficiently small Δt to make $|\Sigma_1|$ arbitrarily small. Thus, the error in (21) can be made as small as one wishes by choosing J sufficiently large.

Q. E. D.

A similar proof for the equivalent one-dimensional finite difference equation has been given by Hildebrand[11] and has also been outlined by Richtmyer and Morton.[8]

It will be noted that the above proof rests on the fact that $|\xi_{pr} - \eta|$ is of order $(\Delta t)^2$. Following the procedure outlined in Appendix B, one can easily verify that this will also hold true for the rectangular mesh shown in Figure 1b, as well as for other meshes having at least two orthogonal axes of local symmetry (reflecting symmetry of nodal locations) passing through node n. However, $|\xi_{pr} - \eta|$ will be of order Δt for all other meshes that we can think of, and therefore our proof will no longer apply.

Point Iterative Scheme

The iterative scheme adopted in this work is known as the point acceleration method and has been developed originally by Evans et al.[10] in 1954. As will be seen below, it differs from the more familiar point successive over-relaxation technique; the latter can be viewed as an extension of the Gauss-Seidel method, whereas the former is more closely related to the point Jacobi method. The acceleration method is readily amenable to an analysis of pointwise convergence and is therefore ideally suited for the mixed explicit-implicit scheme proposed in this work.

The system of equations (9) can be rewritten as

$$\Delta h_n = \sum_{m \neq n} \lambda_{nm} \left[\underbrace{(h_n^k - h_m^k)}_{\text{Explicit part}} + \theta \underbrace{(\Delta h_n - \Delta h_m)}_{\text{Implicit part}} \right] + Q_n \Delta t / D^*_{nn}$$

$$n = 1, 2, \cdots, N \qquad (27)$$

where $\Delta h_n = h_n^{k+1} - h_n^k$ and it is seen that the implicit part vanishes when $\theta = 0$. The acceleration method consists of introducing the following

0 0 0 0 4 4 0 4 7 2 9

substitutions into (27):

$$\Delta h_n \text{ (left side)} \rightarrow \Delta h_n^{j+1}$$

$$\Delta h_n \text{ (right side)} \rightarrow (1 + g)\, \Delta h_n^{j+1} - g\, \Delta h_n^{j}$$

$$\Delta h_m \text{ (right side)} \rightarrow \Delta h_m^{j}$$

where j is number of iterations and g is acceleration factor. Solving for Δh_n^{j+1}, the acceleration algorithm takes the form

$$\Delta h_n^{j+1} = \frac{\sum_{m \neq n} \lambda_{nm} (h_n^k - h_m^k) - \theta \sum_{m \neq n} \lambda_{nm} (g\, \Delta h_n^j + \Delta h_m^j) + Q_n \Delta t / D^*_{nn}}{1 - \theta\, (1 + g) \sum_{m \neq n} \lambda_{nm}} \quad (28)$$

The reader may easily recognize the fact that when g is set equal to zero, equation (28) reduces to the point Jacobi algorithm. As a matter of contrast, when the relaxation factor in the point successive over-relaxation algorithm is set equal to unity, it reduces to the Gauss-Seidel algorithm.

As will be shown below, the acceleration method converges at any node at which $[\underline{\lambda}]$ is locally diagonally dominant (we mentioned earlier that it is always possible to construct a mesh which satisfies this requirement), provided that $g \geq 0$. For optimum results, g should not exceed 1 and should usually be less than 0.5. Experience indicates that near-zero values of g may cause difficulties and the optimum value tends to be in the vicinity of 0.2.

Convergence of Iterative Scheme

Let us define the iteration error at a node as $\Delta\varepsilon_n^{j+1} = \Delta h_n^{j+1} - \Delta h_n^{j}$ and introduce it into equation (28) by utilizing (8), with the result

$$\Delta\varepsilon_n^{j+1} = \frac{-\theta \sum_{m \neq n} \lambda_{nm} \Delta\varepsilon_m^{j} + \theta\, g\, \lambda_{nn} \Delta\varepsilon_n^{j}}{1 + \theta\,(1+g)\,\lambda_{nn}} \tag{29}$$

Using again von Neumann's harmonic analysis in a manner similar to Evans et al.,[10] we replace $\Delta\varepsilon_n^{j}$ by $B_{pr}\,\xi_{pr}^{(j)} \exp(ipx_n + ir\,\Delta y_n)$ where the growth factor is raised to the j-th power and thus obtain from equation (29)

$$\xi_{pr} = \frac{-\theta \sum_{m \neq n} \lambda_{nm} \exp(ip\,\Delta x_m + ir\,\Delta y_m) + \theta\, g\, \lambda_{nn}}{1 + \theta\,(1+g)\,\lambda_{nn}} \tag{30}$$

The growth factor can be decomposed into its real and imaginary parts,

$$\mathrm{Re}(\xi_{pr}) = \frac{-\theta \sum_{m \neq n} \lambda_{nm} \cos(p\,\Delta x_m + r\,\Delta y_m) + \theta\, g\, \lambda_{nn}}{1 + \theta\,(1+g)\,\lambda_{nn}} \tag{31}$$

$$\mathrm{Im}(\xi_{pr}) = \frac{-\theta \sum_{m \neq n} \lambda_{nm} \sin(p\,\Delta x_m + r\,\Delta y_m) + \theta\, g\, \lambda_{nn}}{1 + \theta\,(1+g)\,\lambda_{nn}} \tag{32}$$

Assuming that $[\underline{\lambda}]$ is locally diagonally dominant, all the terms λ_{nm} $(m \neq n)$ are negative according to Appendix A. This implies that the most extreme values of $\mathrm{Re}(\xi_{pr})$ are obtained when the cosines in (31) are replaced by ± 1, in which case (see (8))

$$\mathrm{Re}(\xi_{pr}) = \frac{\theta\,\lambda_{nn}\,(\pm 1 + g)}{1 + \theta\,\lambda_{nn}\,(1+g)} \tag{33}$$

For convergence we require that $|\mathrm{Re}(\xi_{pr})| < 1$ or, according to (33), that

0 0 0 0 4 4 0 4 7 3 0

$$-1 < \frac{\theta \lambda_{nn} (\pm 1 + g)}{1 + \theta \lambda_{nn} (1 + g)} < 1 \tag{34}$$

A similar result is obtained by considering $\mathrm{Im}(\xi_{pr})$. Since θ and λ_{nn} are non-negative, equation (34) is satisfied for all $g \geq 0$. If the positive sign is chosen in the numerator, the smallest absolute value of ξ_{pr} (i.e., the fastest rate of convergence) is achieved with $g = 0$. If the negative sign is chosen, the smallest absolute value of ξ_{pr} is achieved with $g = 1$. Thus, optimum rate of convergence is obtained when $0 \leq g \leq 1$.

The above criterion was obtained considering the most extreme situations that may arise. In general, however, the sine and cosine terms in (31) and (32) will attain the values ± 0.5 twice as often as ± 1. If we replace ± 1 by ± 0.5 in equation (34) and remember that g must not be negative, we find that the optimum rate of convergence is achieved when $0 \leq g \leq 0.5$.

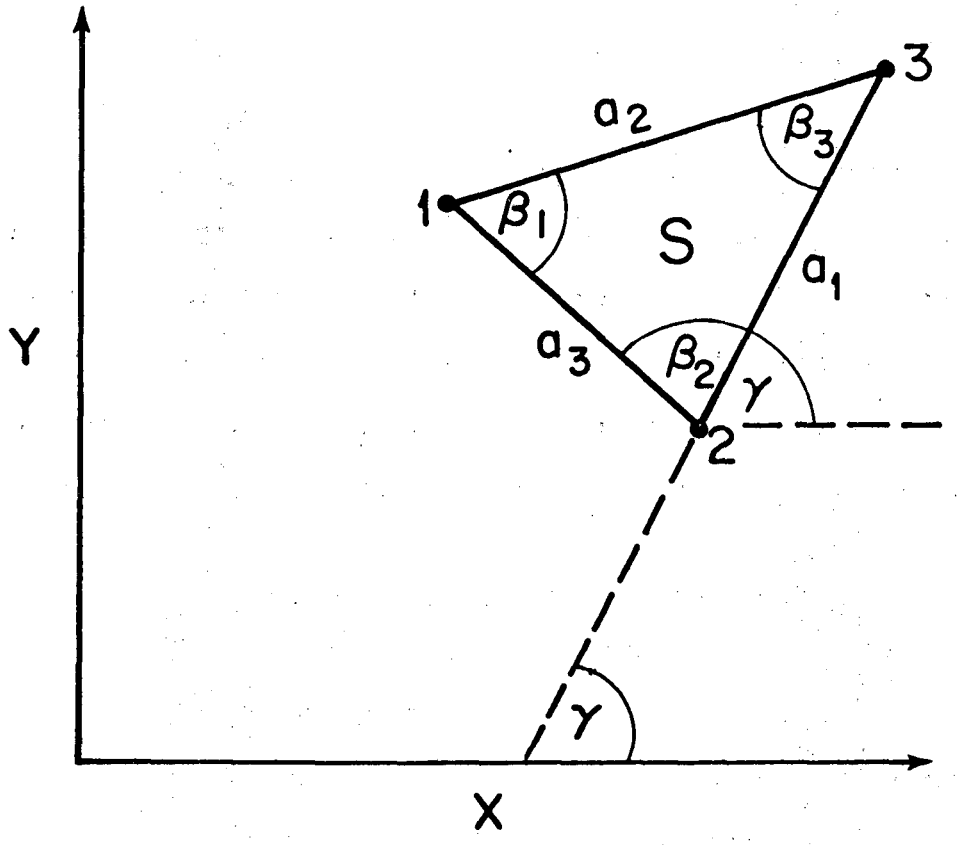
Ensuring Local Diagonal Dominance

The purpose of this section is to show that one can always construct a finite element network of triangles so as to guarantee that $[\underline{\underline{\lambda}}]$ will be diagonally dominant. To do so for isotropic domains, we will follow an approach suggested earlier by Gambolati.[12] However, we will show that Gambolati's analysis for anisotropic domains is in error.

Consider a triangular element in a plane described by the coordinate system x,y as shown in Figure 2. Then it is easy to show that

$$\begin{aligned} b_1 &= -a_1 \sin\gamma & c_1 &= a_1 \cos\gamma \\ b_2 &= a_2 \sin(\gamma - \beta_3) & c_2 &= -a_2 \cos(\gamma - \beta_3) \\ b_3 &= a_3 \sin(\gamma + \beta_2) & c_3 &= -a_3 \cos(\gamma + \beta_2) \end{aligned} \tag{35}$$

where b and c are geometric coefficients defined in Appendix A and all



XBL758-3689

Figure 2. Triangular element in a fixed coordinate system x, y.

0 0 0 0 4 4 0 4 7 3 1

other terms are defined in Figure 2. If $[\underline{\underline{A}}^e]$ is the contribution of this triangle to the global matrix $[\underline{\underline{A}}]$ then equations (4) and (35) imply that, in an isotropic domain,

$$[\underline{\underline{A}}^e] = \frac{K}{4S}\begin{bmatrix} a_1^2 & -a_1 a_2 \cos\beta_3 & -a_1 a_3 \cos\beta_2 \\ -a_1 a_2 \cos\beta_3 & a_2^2 & -a_2 a_3 \cos\beta_1 \\ -a_1 a_3 \cos\beta_2 & -a_2 a_3 \cos\beta_1 & a_3^2 \end{bmatrix} \quad (36)$$

where S is area of triangle and K is the scalar equivalent of $[\underline{\underline{K}}]$. From Appendix A we know that

$$A_{nn}^e = -\sum_{m \neq n} A_{nm}^e \quad (37)$$

indicating that $[\underline{\underline{A}}^e]$ is diagonally dominant if and only if all the off-diagonal terms in (36) are negative. This means that none of the angles β may exceed 90^o, and therefore a sufficient (though not always necessary) condition for the global matrix $[\underline{\underline{A}}]$ to be diagonally dominant at any node is that the adjacent network consists entirely of right and/or acute angled triangles. The same, of course, holds for $[\underline{\underline{\lambda}}]$.

Next, consider an anisotropic domain with principal hydraulic conductivities K_1 and K_2 oriented parallel to the x and y coordinates, respectively. Then, according to equation (4),

$$A_{nm}^e = \frac{1}{4S}\left(K_1 b_n b_m + K_2 c_n c_m\right) \quad (38)$$

We can now define a new set of coordinates $x' = x/(K_1/K_2)^{\frac{1}{2}}$ and $y' = y$, so that in the transformed domain of x' and y', equation (2) will take the form

$$(K_1K_2)^{\frac{1}{2}} \nabla^2 h = C\left(\frac{K_1}{K_2}\right)^{\frac{1}{2}} \frac{\partial h}{\partial t} \tag{39}$$

In other words, the original anisotropic domain in the x,y plane has been transformed into an equivalent isotropic domain with conductivity $(K_1K_2)^{\frac{1}{2}}$ in the x', y' plane merely by expanding or contracting one of the coordinate axes. If b' and c' are the equivalents of b and c in the transformed domain, then it is easy to verify that equation (38) can also be written as

$$A_{nm}^{e} = \frac{(K_1K_2)^{\frac{1}{2}}}{4S'} (b_n' b_m' + c_n' c_m') \tag{40}$$

Let a, β, γ in Figure 2 transform into a', β', γ' in the x', y' plane. Then it is immediately obvious from (35), (36), and (40) that

$$[\underline{\underline{A}}^e] = \frac{(K_1K_2)^{\frac{1}{2}}}{4S'} \begin{bmatrix} (a_1')^2 & -a_1' a_2' \cos\beta_3' & -a_1' a_3' \cos\beta_2' \\ -a_1' a_2' \cos\beta_3' & (a_2')^2 & -a_2' a_3' \cos\beta_1' \\ -a_1' a_3' \cos\beta_2' & -a_2' a_3' \cos\beta_1' & (a_3')^2 \end{bmatrix} \tag{41}$$

Thus, following the same line of reasoning as before, it is evident that the global matrix $[\underline{\underline{A}}]$ is locally diagonally dominant whenever the local network in the transformed domain consists entirely of right and/or acute angled triangles. Again, so is $[\underline{\underline{\lambda}}]$.

It is easy to show that the same holds true when the principal conductivities K_1 and K_2 are not parallel to the x and y coordinates. For this purpose, it is sufficient to recognize that the solutions of equations (2) and (9) at any given point in space are independent of the choice of coordinates. Thus, $[\lambda]$ must remain invariant under a rotation of coordinates and therefore, if it is diagonally dominant in a set of coordinates which is

0 0 0 0 4 4 0 4 7 3 2

parallel to K_1 and K_2, it must also remain diagonally dominant in another set of coordinates oriented at an angle to the first one. One can therefore transform any anisotropic domain into an equivalent isotropic domain merely by expanding or contracting it parallel to K_1 by the amount $(K_1/K_2)^{\frac{1}{2}}$. If all the triangles in the transformed domain are constructed without obtuse angles, $[\underline{\underline{\lambda}}]$ will be diagonally dominant.

In a composite material consisting of several segments with different degrees and orientations of anisotropy, each segment must be transformed separately parallel to its own principal direction of conductivity. Here, in addition to ensuring that all the triangles in the transformed domain are free of obtuse angles, one must also make sure that corresponding nodal points at both sides of a material interface will coincide with each other when the meshes are transformed back into the original plane. This is usually not a very difficult task, as will be demonstrated by an example in Part II of this paper.

Gambolati[12] claimed (see his Figure 3 and p. 589) that $[\underline{\underline{A}}]$ cannot be made diagonally dominant when the ratio $(K_2 - K_1)/K_1$ for $K_1 < K_2$ exceeds 2. Our analysis and the examples in Part II of this paper do not support this viewpoint, indicating that Gambolati's analysis must be in error.

We mentioned earlier that in order to achieve pointwise convergence of the explicit-implicit finite element scheme, it is sufficient that the mesh in any isotropic domain conform to one of the two patterns shown in Figure 1 or to certain symmetry requirements. It is now clear that one can approach this ideal very closely in most cases, even if the material is anisotropic, merely by constructing the mesh in the transformed isotropic plane according to this requirement.

Conclusions

The theoretical analysis in Part I of the paper leads to the following conclusions:

1. The explicit-implicit finite element formulation is amenable to an analysis of local stability. If the n-th row or column of the conductance matrix is diagonally dominant (i.e., the matrix is locally diagonally dominant near node n) then the solution at node n is unconditionally stable when $\theta \geq 0.5$ (i.e., the implicit part has a weight equal to or greater than the explicit part). If $\theta < 0.5$, the solution at node n may be stable or not, depending on the local stability criterion in (18).

2. Since stability conditions vary from one node to another in a given finite element mesh, it may be possible to solve for h_n explicitly at some nodes and implicitly at other nodes. We will refer to this as a mixed explicit-implicit solution strategy. The advantages of this solution strategy will be discussed in Part II of the paper.

3. The explicit-implicit finite element formulation is amenable to an analysis of local convergence at any node n which does not lie on a material boundary and does not act as a sink or a source. If the mesh in the immediate neighborhood of n possesses at least two orthogonal axes of symmetry with respect to the positions of the nodes in the transformed isotropic domain, then the numerical solution at n converges to the exact solution of the partial differential equation provided that the stability criteria are not violated.

4. The fact that the explicit-implicit scheme may be shown to converge locally under certain conditions, whereas the more traditional scheme in which the capacity matrix is non-diagonal converges only in the mean, may be significant from the standpoint of mass or energy balance.

0 0 0 0 4 4 0 4 7 3 3

We suspect that the ability of the explicit-implicit scheme to converge locally is closely related to its property of maintaining a local balance of mass or energy.

5. The explicit-implicit equations can be solved by a point iterative method (we use the accelerated method of Evans et al.[10]) which can be shown to converge at any node at which the conductance matrix is locally diagonally dominant. Iterative techniques have certain advantages over direct methods such as Gaussian elimination: Computer storage requirements are less (one need not worry about band widths and proper numbering of nodes) and, if the matrix is properly constructed so as to insure rapid convergence, significant savings in computer time may be achieved. In addition, iterative techniques are ideally suited for the treatment of quasilinear problems in which the conductance and capacity matrices vary with the dependent variable.

6. One can always construct a mesh of triangular elements which will lead to a diagonally dominant conductance matrix. An earlier statement by Gambolati[12] that the conductance matrix cannot be made diagonally dominant when the degree of anisotropy exceeds a certain limit was shown to be incorrect.

7. Our recognition that the finite element equations (9) can be written without the diagonal terms of the conductance matrix leads to a saving in computer time and storage.

Acknowledgment

This work was partly supported by the U. S. Energy Research and Development Administration.

Appendix A

Consider the matrix $[\underline{A}]$ as defined in equation (4) for plane flow (i.e., $\delta = 1$). A typical term contributed by a single triangle such as that shown in Figure 2 has the form

$$A_{nm}^{e} = \frac{1}{4S} \left[K_{xx}\, b_n\, b_m + K_{xy}\, (b_n\, c_m + b_m\, c_n) + K_{yy}\, c_n\, c_m\right] \tag{42}$$

where

$$\begin{aligned} b_1 &= y_2 - y_3 & c_1 &= x_3 - x_2 \\ b_2 &= y_3 - y_1 & c_2 &= x_1 - x_3 \\ b_3 &= y_1 - y_2 & c_3 &= x_2 - x_1 \end{aligned} \tag{43}$$

If K_1 and K_2 are the two principal conductivities, and σ is the angle between K_1 and the x coordinate, then it can be shown with the aid of Darcy's law that

$$\begin{aligned} K_{xx} &= K_1 \cos^2\sigma + K_2 \sin^2\sigma \\ K_{yy} &= K_1 \sin^2\sigma + K_2 \cos^2\sigma \\ K_{xy} &= (K_1 - K_2) \sin\sigma \cos\sigma \end{aligned} \tag{44}$$

Substituting equation (44) into equation (42) and rearranging, we obtain

$$\begin{aligned} A_{nm}^{e} = \frac{1}{4S} \big[&K_1\, (b_n \cos\sigma + c_n \sin\sigma)\, (b_m \cos\sigma + c_m \sin\sigma) \\ &+ K_2\, (b_n \sin\sigma - c_n \cos\sigma)\, (b_m \sin\sigma - c_m \cos\sigma)\big] \end{aligned} \tag{45}$$

indicating that $A_{nn}^{e} \geq 0$, i.e., the diagonal terms of $[\underline{\underline{A}}^{e}]$ and $[\underline{\underline{A}}]$ are always non-negative.

Furthermore, recognizing that $b_1 + b_2 + b_3 = 0$ and $c_1 + c_2 + c_3 = 0$, we find that equation (45) can also be written for $n = m = 1$ as

0 0 0 0 4 4 0 4 7 3 4

$$A_{11}^{e} = -\frac{K_1}{4S}(b_1 \cos\sigma + c_1 \sin\sigma)[(b_2 + b_3)\cos\sigma + (c_2 + c_3)\sin\sigma]$$

$$- \frac{K_2}{4S}(b_1 \sin\sigma - c_1 \cos\sigma)[(b_2 + b_3)\cos\sigma - (c_2 + c_3)\sin\sigma]$$

$$= - A_{12}^{e} - A_{13}^{e} \qquad (46)$$

Thus, in general we have

$$A_{nn}^{e} = - \sum_{m \neq n} A_{nm}^{e} \quad \text{and} \quad A_{nn} = - \sum_{m \neq n} A_{nm} \qquad (47)$$

where the summation is taken over all nodes other than n.

Appendix B

Consider a homogeneous isotropic domain with a superimposed mesh of equilateral triangular elements as shown in Figure 1a. Due to equation (8) and the symmetry of the mesh, we have

$$\lambda_{n1} = \lambda_{n2} = \cdots = \lambda_{n6} = -\frac{1}{6}\lambda_{nn} \tag{48}$$

Furthermore, if we orient the coordinates so that the origin is at node n and the x axis points toward node 1, then

$$\begin{aligned} \Delta x_1 &= -\Delta x_4 = a \\ \Delta x_2 &= -\Delta x_3 = -\Delta x_5 = \Delta x_6 = \frac{a}{2} \\ \Delta y_1 &= \Delta y_2 = 0 \\ \Delta y_2 &= \Delta y_3 = -\Delta y_5 = -\Delta y_6 = \sqrt{3}\ \frac{a}{2} \end{aligned} \tag{49}$$

where $\Delta x_m = x_m - x_n$, $\Delta y_m = y_m - y_n$, and a is defined in Figure 1a. Thus, we can expand the term ν_{pr} defined below in the form

$$\begin{aligned} \nu_{pr} &= \sum_{m \neq n} \lambda_{nm} [1 - \exp (ip\, \Delta x_m + ir\, \Delta y_m)] \\ &= -\frac{\lambda_{nn}}{6} \sum_{m \neq n} [1 - \cos (p\, \Delta x_m + r\, \Delta y_m)] \\ &= -\frac{\lambda_{nn}}{6} \sum_{m \neq n} [\tfrac{1}{2!} (p\, \Delta x_m + r\, \Delta y_m)^2 - \tfrac{1}{4!} (p\, \Delta x_m + r\, \Delta y_m)^4 + \cdots] \end{aligned} \tag{50}$$

According to equations (4), (5), (7), (43), and (49), we have

$$\lambda_{nn} = \frac{\alpha\, \Delta t}{8S^2} \sum_{e} (b_n^2 + c_n^2) = 4\, \alpha\, \Delta t / a^2 \tag{51}$$

because $S^2 = 3a^4/16$. Recognizing also that equation (49) implies

0 0 0 0 4 4 0 4 7 3 5

$$\sum_{m \neq n} (p\,\Delta x_m + r\,\Delta y_m)^2 = 3a^2 (p^2 + r^2) \tag{52}$$

$$\sum_{m \neq n} (p\,\Delta x_m + r\,\Delta y_m)^4 = \frac{9a^4}{4} (p^2 + r^2)^2 \tag{53}$$

we can rewrite the term ν_{pr} in equation (50) as

$$\nu_{pr} = -\alpha\,\Delta t\,(p^2 + r^2) + \frac{a^2}{16}\alpha\,\Delta t\,(p^2 + r^2)^2 - \cdots \tag{54}$$

Now ξ_{pr} in equation (12) can be expressed as

$$\xi_{pr} = \frac{1 + (1-\theta)\,\nu_{pr}}{1 - \theta\,\nu_{pr}} \tag{55}$$

and when this is expanded in a Taylor series about $\nu_{pr} = 0$, the result is

$$\xi_{pr} = 1 + \nu_{pr} + 2\theta\,\nu_{pr}^2 + \cdots \tag{56}$$

Substituting (54) into (56), we finally obtain

$$\xi_{pr} = 1 - \alpha\,\Delta t\,(p^2 + r^2) + (\frac{a^2}{16} + 2\theta\,\alpha\,\Delta t)\,\alpha\,\Delta t\,(p^2 + r^2)^2 + \cdots$$

$$= 1 - \alpha\,\Delta t\,(p^2 + r^2) + (\frac{1}{4\lambda_{nn}} + 2\theta)(\alpha\,\Delta t)^2\,(p^2 + r^2)^2 + \cdots \tag{57}$$

Since in the text we showed that λ_{nn} is invariant under a change of coordinates, it is obvious that the above result is not restricted by our particular choice of the x and y axes in Figure 1.

References

1. O. C. Zienkiewicz, The Finite Element Method in Engineering Science, McGraw-Hill, London, 1971.

2. T. N. Narasimhan, "Perspective on numerical analysis of transient groundwater motion," unpublished manuscript, 1975.

3. E. L. Wilson, "The determination of temperatures within mass concrete structures," SESM Rept. No. 68-17, Dept. of Civil Engineering, Univ. of California, Berkeley, 1968.

4. A. F. Emery and W. W. Carson, "An evaluation of the use of the finite-element method in the computation of temperature," Jour. Heat Transfer, Trans. ASME, v. 93, pp. 136-145, 1971.

5. S. P. Neuman, "Saturated-unsaturated seepage by finite elements," Jour. Hydraulic Div., ASCE, v. 99, pp. 2233-2250, 1973.

6. S. P. Neuman, "Galerkin approach to saturated-unsaturated flow in porous media," Chapter 10 in Finite Elements in Fluids, Vol. 1, Taylor et al., ed., Wiley, London, 1975 (in press).

7. T. N. Narasimhan, "A unified numerical model for saturated-unsaturated groundwater flow," Ph. D. dissertation, Dept. of Civil Engineering, Univ. of California, Berkeley, 1975.

8. R. D. Richtmyer and K. W. Morton, Difference Methods for Initial-Value Problems, Interscience, New York, 1967.

9. G. G. O'Brien, M. A. Hyman, and S. Kaplan, "A study of the numerical solution of partial differential equations," Jour. Math. Physics, v. 29, pp. 223-251, 1951.

10. G. W. Evans, R. J. Brousseau, and R. Keirstead, "Instability considerations for various difference equations derived from the diffusion equation," Rept. UCRL-4476, Lawrence Radiation Lab., Livermore, Calif., 1954.

11. F. B. Hildebrand, "On the convergence of numerical solutions of the heat-flow equation," Jour. Math. Physics, v. 31, pp. 35-41, 1952.

12. G. Gambolati, "Diagonally dominant matrices for the finite element method in hydrology," Int. Jour. Num. Meth. Engng., v. 6, pp. 587-591, 1973.

0 0 0 4 4 0 4 7 3 6

LEGAL NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Energy Research and Development Administration, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

TECHNICAL INFORMATION DIVISION
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720