JOURNAL OF COMPUTATIONAL PHYSICS 42, 20–52 (1981)

Physical Constraints in Numerical Calculations of Diffusion

G. J. PERT

Department of Applied Physics, University of Hull, Hull HU6 7RX, England

Received February 26, 1980; revised February 20, 1981

The physical behaviour of the diffusion equation is examined, and shown to be a consequence of appropriate mathematical properties of the diffusion operator. Amongst these, the familiar decay of extrema, a consequence of the maximum principle, is given particular attention. The development of spatial and temporal differencing to preserve this property, to be called extremal, yields solutions which preserve positivity and converge uniformly to the steady state. The general construction of extremal algorithms is described for use in a two-level system. The use of weights to improve the accuracy of temporal integration is discussed.

INTRODUCTION

The inclusion of diffusion processes into the body of large hydro-dynamic codes has stimulated the use of implicit solutions to the diffusion equations which converge in some well-behaved fashion onto the equilibrium (uniform) solution. Such solutions are clearly of value since it frequently happens that regions in which diffusion processes are extremely rapid are of limited physical interest. One is thus faced with the partial differential equation equivalent of the stiff equation problem in ordinary differential equations. In a similar fashion one seeks solutions that are stable, and convergent for the fast processes of little interest, yet accurate for the slower ones. The restrictions of computer store and CPU time imposed by the mesh of a large fluid code necessitate that no more two time levels of data be stored simultaneously. In consequence one is forced to consider two-step integration schemes in which the data is advanced from one time level to the next. In the course of this time-step the characteristic rate coefficients are assumed to take constant values varying in space, and from time-step to time-step. As a result of its convergence property for large time-steps, fully implicit—or Laasonen—temporal differencing [1] is usually preferred to the split-time-step—or Crank–Nicholson scheme—despite the loss in small order accuracy introduced thereby, to avoid the spurious oscillation inherent in the latter procedure.

The spatial differencing of the diffusion equation is usually required to satisfy various constraints inherent within the governing differential equation. These are usually consistency [1] and conservatism. More recently Kershaw [2] has drawn attention to the need for the matrix representing this difference—the diffusion matrix—to be non-positive definite. We shall show that this condition essentially

0021-9991/81/070020-33$02.00/0

expresses the second law of thermodynamics, and in consequence the mathematical stability of diffusion [3]. A hitherto neglected (in this context) physical property of diffusion is contained in the maximum principle [3], namely, that extrema decay in time [4]. We shall examine the form of spatial and temporal differencing necessary for the numerical solution to satisfy this condition, i.e., to be extremal. Laasonen [1] has drawn attention to this property for fully implicit differencing in an isotropic medium on an orthogonal mesh, but in the more general case the properties of an extremal form have not been considered.

The recent major advance in tackling implicit problems of this type, the development of the ICCG sparse matrix solving routine by Kershaw [5] has generated a need for improvements in the existing finite difference representation in a generalised geometry. In particular two problems exist:

(a) Convergence for large rates requires fully implicit temporal differencing, with relatively poor accuracy. On the other hand the improved accuracy given by split-time-step schemes necessitates the use of limited time-steps.

(b) The spurious generation of occasional negative values, which violate essential physical constraints, is an unavoidable occurence in calculations with existing methods. These must be removed by an ad hoc procedure (which normally violates conservation) such as "reset to zero."

The application of extremal principles is shown to generate algorithms to overcome both these difficulties.

In this paper we first examine the mathematical properties of the fundamental diffusion equation. In this our approach is similar to that of Kershaw [2], who derived conditions for the less general symmetric case. The close relationship of this approach with the theory of stiff equations is emphasided by considering the spatial differencing (thereby yielding the diffusion matrix), independently of the temporal, and allowing the conditions on the diffusion matrix necessary to re-produce those of the exact equations to be determined. Implicit temporal differencing, and its relation to the extremal principle, is examined separately.

To examples of extremal schemes are constructed. In each case the simplest finite difference representation is not extremal. In one case resourse to the known physical behaviour allows the diffusion matrix to be made extremal. In the second no general consistent extremal representation of the diffusion matrix exists; under restricted conditions it is therefore necessary to introduce an extremum limited flux term to treat this problem. It is believed that the two methods outlined in these examples are sufficient to cover all such problems.

Finally the use of extremum limited solutions to generate weights for the temporal integration is shown to allow split-time-step differencing for small rates and fully implicit for large. This approach is very similar to that devised for integrating positivity maintaining rate equations [6].

The Diffusion Equation and Its Properties

The diffusion of some intrinsic (real) quantity, ε, for example, specific energy, is governed by an equation of the form

$$\frac{\partial}{\partial t}(a\varepsilon) = \mathscr{D}(\varepsilon), \tag{1}$$

where a is a real, positive valued (often time independent) function, and $\mathscr{D}(\varepsilon)$ a differential operator—the diffusion operator—of the form

$$\mathscr{D}(\varepsilon) = \nabla \cdot \mathbf{q}, \tag{2}$$

where $\mathbf{q}$ is the diffusive flux, given by

$$\mathbf{q} = -\underline{\boldsymbol{\kappa}} \cdot \nabla\varepsilon, \tag{3}$$

and $\underline{\boldsymbol{\kappa}}$ is the diffusivity tensor.

We consider the behaviour of the above equation within an irreducible isolated enclosure, such that there is no flux through the walls:

$$\underline{\boldsymbol{\kappa}} \cdot \nabla\varepsilon\,|_n = 0, \tag{4}$$

where n is normal to the bounding surface s.

In general, the diffusion operator may be considered as the sum of symmetric, $\mathscr{D}_s$, and anti-symmetric, $\mathscr{D}_A$, components, such that $\mathscr{D}_s$ is self-adjoint, and $\mathscr{D}_A$ equal to minus its adjoint. For example, in the case of diffusion in a magnetic field:

$$\underline{\boldsymbol{\kappa}} \cdot \nabla\varepsilon = \kappa_{\parallel}\nabla_{\parallel}\varepsilon + \kappa_{\perp}\nabla_{\perp}\varepsilon + \boldsymbol{\kappa}_{\wedge} \wedge \nabla\varepsilon, \tag{5}$$

where the suffixes $\parallel$, $\perp$ and $\wedge$ have their usual meaning, parallel, perpendicular and cross product to the magnetic field [7]. The first two terms are clearly symmetric, and the third anti-symmetric.

The diffusion operator is *differential*, i.e., if ε is constant throughout space, of value ε_0,

$$\mathscr{D}\varepsilon_0 = 0 \tag{6}$$

and ε is constant in time. The diffusion equation is *conservative*:

$$\frac{\partial}{\partial t}\int_V a\varepsilon\, dV = -\int_s d\mathbf{s} \cdot \underline{\boldsymbol{\kappa}} \cdot \nabla\varepsilon = 0. \tag{7}$$

Two further important constraints on the diffusion operator are discussed in

Appendix 1. It is a consequence of the second law of thermodynamics that the operator must be *non-positive definite* in the sense:

$$\int_V \varepsilon \mathscr{D}(\varepsilon)\, dV \leqslant 0. \tag{8}$$

Furthermore the familiar picture of diffusion in terms of a decay of extrema is contained in the *extremal condition*: namely, that if a is independent of time

$$\mathrm{Max}[\varepsilon(t_0), \varepsilon_b] \geqslant \mathrm{Max}[\varepsilon(t)] \geqslant \mathrm{Min}[\varepsilon(t)] \geqslant \mathrm{Min}[\varepsilon(t_0), \varepsilon_b], \tag{9}$$

where $t > t_0$, Max and Min are the largest and smallest values of the set denoted, respectively, and ε_b are the boundary values of ε over the interval t_0 to t. This condition ensures the essential positivity of the physical quantity ε during diffusion.

The extremal condition is stronger than the previous one, for an extremal operator has eigenvalues whose real part must be non-positive, and therefore if not defective must be non-positive definite. We may remark that one eigenvalue is necessarily zero since the operator is differential and corresponds to the steady state.

Thus far our discussion of the properties of diffusion has been quite general in that arbitrary forms of the system parameters a and κ are allowed. In a finite difference scheme the integration of Eq. (1) follows a step-by-step procedure integrating from time-step to time-step, the system parameters a and κ being held constant during this time interval. In the following we assume that such a procedure is to be adopted, and therefore investigate the integration properties under the assumption that a and κ are temporal constants, i.e., we are considering integration over one time-step. The constants a and κ will of course change their values from time-step to time-step.

The Diffusion Matrix

Since the diffusion equation is linear in ε, the spatial finite difference of the diffusion equation takes a linear (matrix) form:

$$A_{ii} \frac{d\varepsilon_i}{dt} = \sum_j D_{ij} \varepsilon_j, \tag{10}$$

where ε_i is the value of ε at some point $\mathbf{r}_i$, A is a real positive diagonal matrix, independent of time, whose components $A_{ii} = a(\mathbf{r}_i)$. D is the real diffusion matrix, a consistent representation of the diffusion operator $\mathscr{D}$. Since $\mathscr{D}$ is defined within an irreducible space, D is itself irreducible.

The matrix D can be split into two components corresponding to the equivalent parts of $\mathscr{D}$. Thus if V is a real positive diagonal matrix whose component V_{ii} is the volume of the cell surrounding the point $\mathbf{r}_i$, and the matrix:

$$B = VD. \tag{11}$$

The representation B_s of the symmetric parts of $\mathscr{D}$ must, in the limit as the mesh is refined, satisfy

$$\int \delta \mathscr{D}_s \varepsilon \, dV \Rightarrow \sum_{i,j} \delta_i B_{ij} \varepsilon_j = \int \varepsilon \mathscr{D}_s \delta \, dV \Rightarrow \sum_{i,j} \varepsilon_i B_{ij} \delta_j \tag{12}$$

for all vectors δ and ε. Hence if the representation is consistent B_s must be symmetric to terms of lowest order of the mesh spacing. In practice this implies that B_s must be symmetric. Similarly the anti-symmetric term $\mathscr{D}_A$ must have an anti-symmetric representation B_A.

The representation must be *differential*, i.e., if $\varepsilon_i = \varepsilon_j = \varepsilon_0$ for all i and j then

$$\sum_j D_{ij} \varepsilon_j = 0$$

and (13)

$$\sum_j D_{ij} = \sum_j B_{ij} = 0.$$

which ensures the existence of a steady-state solution, and that D is singular. (Appendix 2, Lemma 1.) We shall call any matrix satisfying (13) a differential matrix.

The matrix representation is *conservative* if

$$\frac{d}{dt} \sum_i A_{ii} \varepsilon_i V_{ii} = \sum_{i,j} B_{ij} \varepsilon_j = 0 \tag{14}$$

for all ε_j. Therefore

$$\sum_i B_{ij} = 0. \tag{15}$$

This result is clearly equivalent to (13) if the matrix is purely symmetric [2]. Any matrix satisfying (15) is called a conservative matrix.

The matrix equation (10) is clearly stable if the eigenvalues of $(A^{-1}D)$, namely, λ, have real parts satisfying $\mathrm{Re}(\lambda) \leqslant 0$. This condition is ensured if the matrix B is *non-positive definite*,[1] a condition equivalent to that for the operator $\mathscr{D}$. Let $\Lambda = \Phi + i\Psi$ be an eigenvector of $(A^{-1}D)$ with eigenvalue $\lambda = \phi + i\psi$. Consider:

$$\begin{aligned}(\Lambda^*, B\Lambda) &= (\Phi, B\Phi) + (\Psi, B\Psi) + i[(\Phi, B\Psi) - (\Psi, B\Phi)] \\ &= \lambda(\Lambda^*, (VA)\Lambda) = (\phi + i\psi)[(\Phi, (VA)\Phi) + (\Psi, (VA)\Psi)]. \end{aligned} \tag{16}$$

[1] We note that these definitions of "definiteness" involve a generalisation of the usual form to asymmetric matrices.

But (VA) is a positive real diagonal matrix and is therefore positive definite. Hence

$$\mathrm{Re}(\lambda) = \phi = [(\Phi, B\Phi) + (\Psi, B\Psi)]/[(\Phi, (VA)\Phi) + (\Psi, (VA)\Psi)] \tag{17}$$

and is non-positive if B is non-positive definite. If $(A^{-1}D)$ is not defective this is also a necessary condition.

We defer a discussion of the extremal properties until later, since its form is more appropriately derived in connection with temporal differencing. We may, however, remark that it may be proved by a similar analysis that Eq. (10) is extremal if and only if the matrix D is differential and non-negativity maintaining (i.e., if $\varepsilon(0) \geqslant 0$, then $\varepsilon(t) \geqslant 0$ for all $t \geqslant 0$), i.e., D is the negative of an M-matrix form[2] [6].

The Temporal Finite Difference

The set of linear equations (10), forms a stiff system. If the matrix D is defined physically we have seen that the eigenvalues of this matrix $(A^{-1}D)$ have non-positive real parts, so that the solutions are decaying. A general two-step form of Eq. (10) is

$$\varepsilon = [A - \theta D\,\Delta t]^{-1}\,[A + (1-\theta)D\,\Delta t]\,\varepsilon^0, \tag{18}$$

where ε^0 is the value of ε at time $(t - \Delta t)$, and θ is an implicitness parameter. Assuming the matrix $(A^{-1}D)$ is not defective we may solve this equation formally in terms of the projections C_l of the vector ε onto the eigenvectors Λ_l of $(A^{-1}D)$.

$$C_l = (1 - \theta\lambda_l \Delta t)^{-1}\,[1 + (1-\theta)\,\lambda_l\,\Delta t]\,C_l^0. \tag{19}$$

For stability we require that the projections C_l be bounded as the number of repetitions of Eq. (18) tends to infinity. Thus if D is time independent

$$\theta \geqslant \frac{1}{2} + \frac{\phi_l}{(\phi_l^2 + \psi_l^2)\,\Delta t}, \tag{20}$$

where ϕ_l and ψ_l are the real and imaginary parts of the eigenvalues λ_l, respectively. This may be recognised as a generalisation of the usual stability condition for diffusion. Noting that the above inequality, cannot be satisfied as $\Delta t \to 0$ if $\mathrm{Re}(\lambda) > 0$ (anti-diffusion) we conclude that the two-step system is unconditionally stable if and only if $(A^{-1}D)$ has only non-positive real part eigenvalues and $\frac{1}{2} \leqslant \theta \leqslant 1$.

It is clear that unless $\theta = 1$, the solution C_l does not decay uniformly, as required, by the exact solution, but may oscillate for large values of Δt. Such oscillations will be inhibited if the extremal condition can be applied.

[2] An M-matrix is by definition non-singular [8] but D is singular. We therefore use the above term to describe such matrices.

Extremal Forms

We define an extremal operator, G, as one whose operation on a vector set ε^0, yields a set ε:

$$\varepsilon = G\varepsilon^0 \tag{21}$$

whose values are extremal (9) with respect to the set ε^0.

Theorem. *A linear operator is extremal if and only if it is both differential (i.e., if ε^0 is a uniform set, then $\varepsilon = \varepsilon^0$) and non-negativity maintaining (i.e., if $\varepsilon^0 \geqslant 0$ then $\varepsilon \geqslant 0$).*

Consider two vector sets ε^0 and δ^0 such that $\varepsilon^0 \geqslant \delta^0$, therefore since G is non-negativity maintaining:

$$\varepsilon = G\varepsilon^0 \geqslant G\delta^0 = \delta.$$

Let δ^0 be a uniform vector whose components are all equal to the smallest component of ε^0, $\delta^0 = \mathrm{Min}(\varepsilon^0)$, and γ^0 a uniform vector equal to the largest components of ε^0, $\gamma^0 = \mathrm{Max}(\varepsilon^0)$. If G is a differential operator:

$$\gamma = G\gamma^0 = \gamma^0 = \mathrm{Max}(\varepsilon^0),$$
$$\delta = G\delta^0 = \delta^0 = \mathrm{Min}(\varepsilon^0).$$

Therefore since $\gamma^0 \geqslant \varepsilon^0 \geqslant \delta^0$

$$\gamma = \mathrm{Max}(\varepsilon^0) \geqslant \varepsilon \geqslant \mathrm{Min}(\varepsilon^0) = \delta, \tag{22}$$

i.e., the solution is extremal.

It is clear that an operator, which is extremal, must be differential and non-negativity maintaining, and the above condition is proven.

For a two-step finite difference representation (18), G has a matrix form

$$G = F^{-1}E \tag{23}$$

when E is an explicit operation:

$$E = I + \{(1-\theta)A^{-1}D\,\Delta t\} \tag{24}$$

and F an implicit one:

$$F = I - \{\theta A^{-1}D\,\Delta t\}, \tag{25}$$

where I is the identity matrix. It is readily shown that G is differential if $(E-F) = A^{-1}D\,\Delta t$ is a differential representation, i.e., if D satisfies Eq. (13). If G is non-negativity preserving and E is non-negative (i.e., if $\varepsilon^0 \geqslant 0$, then $\varepsilon = E\varepsilon^0 \geqslant 0$), then

F is monotone [9] (i.e., if $\varepsilon^0 \geqslant 0$, then $\varepsilon = F^{-1}\varepsilon^0 \geqslant 0$); and similarly since E and F commute, if F is monotone then E must be non-negative. This reciprocity occurs naturally within the matrices E and F, for if E is non-negative, all its components, $E_{ij} \geqslant 0$ and

$$D_{ij} \geqslant 0, \qquad i \neq j,$$
$$D_{ii} = -\sum_{j \neq i} D_{ij} \geqslant -A_{ii}/\{(1-\theta)\Delta t\}. \tag{26}$$

D is therefore the negative of an M-matrix form. When D has this form, F is an M-matrix and therefore monotone [8].

In general, D is a local matrix such that it has non-zero components between only a restricted set of mesh points, which we call neighbours. If D is a local M-matrix form. then the operation G is locally extremal. Thus if the set of values ε^0 and ε at the neighbours of i is ξ_i, then

$$\operatorname{Min}(\xi_i, \varepsilon_i^0) \leqslant \varepsilon_i \leqslant \operatorname{Max}(\xi_i, \varepsilon_i^0), \tag{27}$$

a condition closely related to the more general maximum principle of the governing differential equation. This result follows since

$$F_{ii}\varepsilon_i = E_{ii}\varepsilon_i^0 + \sum_{j \neq i} (E_{ij}\varepsilon_j^0 + |F_{ij}|\,\varepsilon_j)$$
$$\leqslant \left\{E_{ii} + \left(\sum_{j \neq i} E_{ij} + |F_{ij}|\right)\right\} \operatorname{Max}(\xi_i, \varepsilon_i^0)$$

and

$$F_{ii} = E_{ii} + \sum_{j \neq i} (E_{ij} + |F_{ij}|).$$

An extremal operator is never divergent, and therefore the associated matrix $(A^{-1}D)$ has non-positive real part eigenvalues, i.e., it is derived from a Lyapunov semi-stable matrix.

We may illustrate these results by the simple example of one dimensional diffusion on a uniform Cartesian mesh with constant diffusivity, κ, for which the only non-zero elements are

$$(A^{-1}D)_{i,i+1} = (A^{-1}D)_{i,i-1} = \kappa, \qquad (A^{-1}D)_{i,i} = -2\kappa. \tag{28}$$

An explicit calculation ($\theta = 0$) is extremal only if $\kappa\Delta t \leqslant \frac{1}{2}$, which is also the well-known stability condition [1]. For a split-time-step calculation ($\theta = \frac{1}{2}$), the operation is unconditionally stable, (20), but extremal only if $\kappa\Delta t \leqslant 1$, reflecting the well-known "overshoot" of this algorithm. Fully implicit schemes ($\theta = 1$) are unconditionally extremal, and, of course, stable.

The M-matrix form of D is clearly a sufficient condition that G be extremal for a restricted range of the implicitness parameter θ. It is not, however, necessary; for example, a monotone matrix is not necessarily an M-matrix [9]; indeed we shall consider later a case where the diffusion matrix, D, cannot be cast into M-matrix form but yields a monotone operation, F. In this case the explicit form, E, is not positively maintaining, and the extremal form must be fully implicit ($\theta = 1$). The M-matrix form of D is however a very useful "necessary" condition in a general sense providing a least restrictive condition on the form of D suitable for rapid test during repetitive calculation, whilst still remaining extremal.

Some general properties of the implicit operator, F, may be derived from the theorems in Appendix 2. In particular it follows from Theorem 1 that if D is non-positive definite F is non-singular, and a solution of (21) exists. Furthermore from Theorem 2 it follows that as $\Delta t \to \infty$, the matrix F is monotone, and converges to the steady state, provided D is differential and conservative. Thus the fully implicit form is always extremal provided the time-step, Δt, is sufficiently large.

The matrices E and F may be defined in a number of ways equivalent to Eqs. (24) and (25), of which the following is particularly useful:

$$E = M + \{(1-\theta) B\,\Delta t\} \tag{24a}$$

and

$$F = M - \{\theta B\,\Delta t\}, \tag{25a}$$

where B is given by (11) and $M = AV$ is the diagonal cell "mass" matrix.

Extremal Finite Difference Representations in Orthogonal Two Dimensional Geometries

We consider the finite difference representation of the diffusion operation with both symmetric and anti-symmetric terms in a Cartesian (r, z) space with either planar or cylindrical symmetry. The diffusion flux may be written:

$$\mathbf{q} = -\kappa_0 \nabla\varepsilon - \kappa_1 \hat{\mathbf{n}} \wedge \nabla\varepsilon, \tag{29}$$

where $\hat{\mathbf{n}}$ is a unit vector perpendicular to the plane of the co-ordinates. The differential equation (1), takes the form

$$\begin{aligned} a\frac{\partial\varepsilon}{\partial t} &= \frac{\partial}{\partial z}\left(\kappa_0 \frac{\partial\varepsilon}{\partial z}\right) + \frac{1}{\tilde{r}}\frac{\partial}{\partial r}\left(\tilde{r}\kappa_0 \frac{\partial\varepsilon}{\partial r}\right) \\ &\quad + \frac{1}{\tilde{r}}\left\{\frac{\partial}{\partial r}(\tilde{r}\kappa_1)\frac{\partial\varepsilon}{\partial z} - \frac{\partial}{\partial z}(\tilde{r}\kappa_1)\frac{\partial\varepsilon}{\partial r}\right\}, \end{aligned} \tag{30}$$

where $\tilde{r} = 1$ or r for planar or cylindrical geometry, respectively.

The difference form may be found by a standard centred-difference about the centre $(r_{i,j}z_{i,j})$ of the cell (i, j). Thus, for example,

$$\frac{\partial}{\partial z}\left(\kappa_0\frac{\partial\varepsilon}{\partial z}\right)=\left[\left(\kappa_0\frac{\partial\varepsilon}{\partial z}\right)_{(i,j+1/2)}-\left(\kappa_0\frac{\partial\varepsilon}{\partial z}\right)_{(i,j-1/2)}\right]\{z_{i,j+1/2}-z_{i,j-1/2}\},$$

$$\left(\kappa_0\frac{\partial\varepsilon}{\partial z}\right)_{i,j+1/2}=(\kappa_0)_{i,j+1/2}\,(\varepsilon_{i,j+1}-\varepsilon_{i,j})/\{z_{i,j+1}-z_{i,j}\},\tag{31}$$

where $\varepsilon_{i,j}$ is the value of ε at the mesh point $(r_{i,j}, z_{i,j})$: the values of terms at intermediate points being defined by a suitable interpolation.

Similarly:

$$\frac{\partial\varepsilon}{\partial z}=\frac{(\varepsilon_{i,j+1/2}-\varepsilon_{i,j-1/2})}{(z_{i,j+1/2}-z_{i,j-1/2})}=\frac{1}{2}\,\frac{(\varepsilon_{i,j+1}-\varepsilon_{i,j-1})}{(z_{i,j+1/2}-z_{i,j-1/2})}.\tag{21}$$

Since the volume of the cell $V_{(i,j),(i,j)}=\tilde{r}_{i,j}(r_{i+1/2,j}-r_{i-1/2,j})(z_{i,j+1/2}-z_{i,j-1/2})$ we obtain the matrix, B, whose only non-zero components are

$$\begin{aligned}
B_{(i,j),(i+1,j)}&=(\kappa_0)_{i+1/2,j}\tilde{r}_{i+1,j}(z_{i,j-1/2}-z_{i,j-1/2})/(r_{i+1,j}-r_{i,j})\\
&\quad-\tfrac{1}{4}\{(\tilde{r}\kappa_1)_{i,j+1}-(\tilde{r}\kappa_1)_{i,j-1}\},\\
B_{(i,j),(i,j+i)}&=(\kappa_0)_{i,j+1/2}\,\tilde{r}_{i,j}(r_{i+1/2,j}-r_{i-1/2,j})/(z_{i,j+1}-z_{i,j})\\
&\quad+\tfrac{1}{4}\{(\tilde{r}\kappa_1)_{i-1,j}-(\tilde{r}\kappa_1)_{i-1,j}\},\\
B_{(i,j),(i-i,j)}&=(\kappa_0)_{i-1/2,j}\,\tilde{r}_{i-1,j}(z_{i,j+1/2}-z_{i,j-1/2})/(r_{i,j}-r_{i-1,j})\\
&\quad+\tfrac{1}{4}\{(\tilde{r}\kappa_1)_{i,j+1}-(\tilde{r}\kappa_1)_{i,j-1}\},\\
B_{(i,j),(i,j-1)}&=(\kappa_0)_{i,j-1/2}\,\tilde{r}_{i,j}(r_{i+1/2,j}-r_{i-1/2,j})/(z_{i,j}-z_{i,j-1})\\
&\quad-\tfrac{1}{4}[(\tilde{r}\kappa_1)_{i+1,j}]-[(\tilde{r}\kappa_1)_{i-1,j}\},\\
B_{(i,j),(i,j)}&=-(z_{i,j+1/2}-z_{i,j-1/2})\{\tilde{r}_{i+1,j}(\kappa_0)_{i+1/2,j}/(r_{i+1,j}-r_{i,j})\\
&\quad+\tilde{r}_{i-1,j}(\kappa_0)_{i-1/2,j}/(r_{i,j}-r_{i-1,j})]\\
&\quad-\tilde{r}_{i,j}(r_{i+1/2,j}-r_{i-1/2,j})\{(\kappa_0)_{i,j+1/2}/(z_{i,j+1}-z_{i,j})\\
&\quad+(\kappa_0)_{i,j+1/2}/(z_{i,j}-z_{i,j-1})\}.
\end{aligned}\tag{33}$$

The matrix B is clearly both differential (13) and conservative (15), and from the derivation also consistent. It is not, however, an M-matrix, and therefore not necessarily extremal. This defect is readily identified as associated with the antisymmetric terms in Eq. (29).

Examination of these terms shows that they have a similar form to those describing advection, and suggests that they may be conveniently differenced by forward or backward differencing depending on the direction of the flux: thus,

$$\begin{aligned}\frac{\partial\varepsilon}{\partial z} &= \frac{(\varepsilon_{i,j+1}-\varepsilon_{i,j})}{(z_{i,j+1/2}-z_{i,j-1/2})} \quad \text{if} \quad \frac{\partial}{\partial r}(\tilde{r}\kappa_1) > 0 \\ &= \frac{(\varepsilon_{i,j}-\varepsilon_{i,j-1})}{(z_{i,j+1/2}-z_{i,j-1/2})} \quad \text{otherwise.}\end{aligned} \tag{34}$$

The modified form of B is then

$$\begin{aligned}B_{(i,j),(i+1,j)} &= (\kappa_0)_{i+1/2,j}\,\tilde{r}_{i+1,j}(z_{i,j+1/2}-z_{i,j+1/2})/(r_{i+1,j}-r_{i,j}) \\ &\quad -\tfrac{1}{2}\,\mathrm{Min}(0,\,[(\tilde{r}\kappa_1)_{i,j+1}-(\tilde{r}\kappa_1)_{i,j-1}]), \\ B_{(i,j),(i,j+1)} &= (\kappa_0)_{i,j+1/2}\,\tilde{r}_{i,j}(r_{i+1/2,j}-r_{i-1/2,j})/(z_{i,j+1}-z_{i,j}) \\ &\quad +\tfrac{1}{2}\,\mathrm{Max}(0,\,[(\tilde{r}\kappa_1)_{i+1,j}-(\tilde{r}\kappa_1)_{i\ \ 1,j}]), \\ B_{(i,j),(i-1,j)} &= (\kappa_0)_{i-1/2,j}\,\tilde{r}_{i-1,j}(z_{i,j+1/2}-z_{i,j-1/2})/(r_{i,j}-r_{i-1,j}) \\ &\quad +\tfrac{1}{2}\,\mathrm{Max}(0,\,[(\tilde{r}\kappa_1)_{i,j+1}-(\tilde{r}\kappa_1)_{i,j-1}]), \\ B_{(i,j),(i,j-1)} &= (\kappa_0)_{i,j-1/2}\,r_{i,j}(r_{i+1/2,j}-r_{i-1/2,j})/(z_{i,j}-z_{i-1,j}) \\ &\quad -\tfrac{1}{2}\,\mathrm{Min}(0,\,[(\tilde{r}\kappa_1)_{i+1,j}-(\tilde{r}\kappa_1)_{i-1,j}]), \\ B_{(i,j),(i,j)} &= -\,[B_{(i,j),(i+1,j)}+B_{(i,j),(i,j+1)}+B_{(i,j),(i-1,j)} \\ &\quad +B_{(i,j),(i,j-1)}].\end{aligned} \tag{35}$$

As before this term is differential and conservative, and can also be shown to be consistent. The error terms, are of lower order than those of the set (33). Against this, however, the modified set converges properly to a uniform state for large time-steps. It is therefore suggested that the set (33) be used unless one of the non-diagonal elements of B is negative, in which case the appropriate forward or backward differencing for that term be used (with care to ensure that conservation is maintained). This algorithm, in a slightly modified correctly symmetrised form, has been extensively used by the author to treat diffusion in magnetic field [10]. The advantages of the M-matrix form have been clearly demonstrated by the removal (with no ill effects) of a "fix to zero" check, which was formerly necessary to avoid occasional negative values arising in the calculation using the centred-difference form (33).

Extremal Finite Difference Representations in Non-Orthogonal Two Dimensional Geometries

We consider the finite difference representation of the diffusion operator in a general non-orthogonal two dimensional co-ordinate system (k, l). The transformation between this system and a Cartesian (R, Z) system is accomplished by means of the Jacobian:

$$j = R_k Z_l - R_l Z_k, \tag{36}$$

where partial derivatives are denoted by sub-scripts; thus, for example, $R_k = \partial R/\partial k$. The derivatives of a function f are given by

$$\frac{\partial f}{\partial R} = (f_k Z_l - f_l Z_k)/j,$$

$$\frac{\partial f}{\partial Z} = -(f_k R_l - f_l R_k)/j. \tag{37}$$

For simplicity we restrict our study to the case of an isotropic medium in which the diffusion operator is symmetric, and κ is a scalar. The generalisation to non-isotropic systems is accomplished in a similar manner.

We consider the flux balance of a cell centred at (k, l) with faces δk and δl parallel to the local k and l axes, respectively. The area of the faces parallel to the l axis (k face) is

$$S_k = \tilde{R}\,|\mathbf{R}_l|\,\delta l, \tag{38}$$

where $\tilde{R} = 1$ or R for planar or cylindrical geometries, respectively, and $\mathbf{R}$ is the position vector:

$$\mathbf{R} = R\hat{\mathbf{R}} + Z\hat{\mathbf{Z}}. \tag{39}$$

The unit normal to the k face in the direction of increasing k is

$$\hat{\mathbf{N}}_k = (Z_l\hat{\mathbf{R}} - R_l\hat{\mathbf{Z}})/\{R_l^2 + Z_l^2\}. \tag{40}$$

The total flux through the k face at $(k + \delta k/2)$ in the direction of increasing k is

$$\begin{aligned} Q_{k+} &= -\,[\kappa \mathbf{S}_k \cdot \nabla\varepsilon]_{(k+\delta k/2)} \\ &= -\left[\frac{\kappa\tilde{R}}{j}\,\{\mathbf{R}_l \cdot \mathbf{R}_l \varepsilon_k - \mathbf{R}_k \cdot \mathbf{R}_l \varepsilon_l\}\right]_{(k+\delta k/2)} \delta l. \end{aligned} \tag{41}$$

Similarly the total flux through the l face at $(l + \delta l/2)$ in the direction of increasing l is

$$Q_{l+} = -\left[\frac{\kappa\tilde{R}}{j}\,\{\mathbf{R}_k \cdot \mathbf{R}_k \varepsilon_l - \mathbf{R}_k \cdot \mathbf{R}_l \varepsilon_k\}\right]_{(l+\delta l/2)} \delta k. \tag{42}$$

The volume of the cell is $\tilde{R}j\,\delta k\,\delta l$ so that the flux balance for the cell (k, l) is

$$\tilde{R}ja\frac{\partial\varepsilon}{\partial t}\,\delta k\,\delta l = -\,(Q_{k+} - Q_{k-}) - (Q_{l+} - Q_{l-}). \tag{43}$$

Allowing δk and δl to proceed to the limit we obtain the differential form

$$a\frac{\partial\varepsilon}{\partial t}=\frac{1}{\tilde{R}j}\left\{\frac{\partial}{\partial k}\left[\frac{\tilde{R}\kappa}{j}(\mathbf{R}_l\cdot\mathbf{R}_l\varepsilon_k-\mathbf{R}_k\cdot\mathbf{R}_l\varepsilon_l)\right]\right.$$
$$\left.+\frac{\partial}{\partial l}\left[\frac{\tilde{R}\kappa}{j}(\mathbf{R}_k\cdot\mathbf{R}_k\varepsilon_l-\mathbf{R}_k\cdot\mathbf{R}_l\varepsilon_k)\right]\right\} \qquad (44)$$

is agreement with a direct calculation.

We establish a finite difference representation in a similar manner, using centred-differences on the faces. Thus if a mesh is defined in (k,l) space such that $k=K\,\delta k$ and $l=L\,\delta l$, we define a cell (K,L) with centre integer values K and L, faces $(K+\frac{1}{2},L)$, $(K,L+\frac{1}{2})$, $(K-\frac{1}{2},L)$ and $(K,L-\frac{1}{2})$, and corners $(K+\frac{1}{2},L+\frac{1}{2})$, $(K-\frac{1}{2},L+\frac{1}{2})$, $(K-\frac{1}{2},L-\frac{1}{2})$ and $(K+\frac{1}{2},L-\frac{1}{2})$. The quantities $\varepsilon_{K,L}$ are defined at the cell centre. Thus,

$$\left[\frac{\kappa\tilde{R}}{j}\mathbf{R}_l\cdot\mathbf{R}_l\varepsilon_k\right]_{(k+\delta k/2)}\to\frac{1}{\delta k}\left[\frac{\kappa\tilde{R}}{j}\mathbf{R}_l\cdot\mathbf{R}_l\right]_{K-1/2,L}(\varepsilon_{K+1,L}-\varepsilon_{K,L}) \qquad (45)$$

and

$$\left[\frac{\kappa\tilde{R}}{j}\mathbf{R}_k\cdot\mathbf{R}_l\varepsilon_l\right]_{(k+\delta k/2)}\to\frac{1}{2}\left\{\left[\frac{\kappa\tilde{R}}{j}\mathbf{R}_k\cdot\mathbf{R}_l\varepsilon_l\right]_{K+1/2,L-1/2}\right.$$
$$\left.+\left[\frac{\kappa\tilde{R}}{j}\mathbf{R}_k\cdot\mathbf{R}_l\varepsilon_l\right]_{K+1/2,L-1/2}\right\}, \qquad (46)$$

where

$$\left[\frac{\kappa\tilde{R}}{j}\mathbf{R}_k\cdot\mathbf{R}_l\varepsilon_l\right]_{K+1/2,L+1/2}$$
$$=\frac{1}{2}\delta l\left[\frac{\kappa\tilde{R}}{j}\mathbf{R}_k\cdot\mathbf{R}_l\right]_{K+1/2,L+1/2}\lfloor\varepsilon_{K+1,L+1}+\varepsilon_{K,L+1}-\varepsilon_{K+1,L}-\varepsilon_{K,L}\rfloor. \qquad (47)$$

The difference (46) gives weight across the face $(l+\delta l/2)$ to quantities evaluated at the corners. In view of this it would be consistent to evaluate the coefficient term in (45), using corner values, i.e.,

$$\left[\frac{\kappa\tilde{R}}{j}\mathbf{R}_l\mathbf{R}_l\right]_{K+1/2,L}=\frac{1}{2}\left\{\left[\frac{\kappa\tilde{R}}{j}\mathbf{R}_l\cdot\mathbf{R}_l\right]_{K+1/2,L+1/2}+\left[\frac{\kappa\tilde{R}}{j}\mathbf{R}_l\cdot\mathbf{R}_l\right]_{K+1/2,L-1/2}\right\}. \qquad (48)$$

The differences for the other faces are evaluated in an identical form. Hence we obtain

$$(AV)_{(K,L),(K,L)}\frac{\partial\varepsilon_{K,L}}{\partial t}=\sum_{(K',L')}B_{(K,L),(K',L')}\varepsilon_{K',L'}, \qquad (49)$$

where the volume of the cell $V_{(K,L),(K,L)} = (\tilde{R}j)_{K,L}\, \delta k\, \delta l$. The matrix B is more conveniently expressed in terms of derivatives with respect to the variables K and L, rather than k and l, whose Jacobian $J = j\, \delta k\, \delta l$:

$$\begin{aligned}
B_{(K+1,L),(K,L)} = B_{(K,L),(K+1,L)} &= \frac{1}{2}\left\{\left[\frac{\kappa\tilde{R}}{J}\mathbf{R}_L\cdot\mathbf{R}_L\right]_{(K+1/2,L+1/2)}\right.\\
&\quad\left.+\left[\frac{\kappa\tilde{R}}{J}\mathbf{R}_L\cdot\mathbf{R}_L\right]_{(K+1/2,L-1/2)}\right\},\\
B_{(K,L-1),(K,L)} = B_{(K,L),(K,L+1)} &= \frac{1}{2}\left\{\left[\frac{\kappa\tilde{R}}{J}\mathbf{R}_K\cdot\mathbf{R}_K\right]_{(K+1/2,L-1/2)}\right.\\
&\quad\left.+\left[\frac{\kappa\tilde{R}}{J}\mathbf{R}_K\cdot\mathbf{R}_K\right]_{(K-1/2,L-1/2)}\right\},\\
B_{(K+1,L+1),(K,L)} = B_{(K,L),(K+1,L+1)} &= -\frac{1}{2}\left[\frac{\kappa\tilde{R}}{J}\mathbf{R}_K\cdot\mathbf{R}_L\right]_{(K+1/2,L+1/2)},\\
B_{(K-1,L+1),(K,L)} = B_{(K,L),(K-1,L+1)} &= \frac{1}{2}\left[\frac{\kappa\tilde{R}}{J}\mathbf{R}_K\cdot\mathbf{R}_L\right]_{(K-1/2,L+1/2)},
\end{aligned}\tag{50}$$

where $\mathbf{R}_K = \partial\mathbf{R}/\partial K$, etc. The matrix B being symmetric, and conservative, the set B is fully specified with

$$B_{(K,L),(K,L)} = -\sum_{(K^1,L^1)\neq(K,L)} B_{(K,L),(K^1,L^1)}\tag{51}$$

and all other elements zero. The matrix B is readily shown to give a consistent representation of Eq. (44).

The matrix B is non-positive definite for the sum (A2.16) of Appendix 2 and may be written:

$$\begin{aligned}
(\varepsilon, B\varepsilon) = -\frac{1}{2}\sum_{(K,L)}&\frac{1}{2}\left[\frac{\kappa\tilde{R}}{J}\mathbf{R}_L\cdot\mathbf{R}_L\right]_{(K+1/2,L+1/2)}\\
&\times\{(\varepsilon_{K+1,L}-\varepsilon_{K,L})^2+(\varepsilon_{K+1,L+1}-\varepsilon_{K,L+1})^2\}\\
&+\frac{1}{2}\frac{\kappa\tilde{R}}{J}\mathbf{R}_K\cdot\mathbf{R}_{K_{(K+1/2,L+1/2)}}\\
&\times\{(\varepsilon_{K,L+1}-\varepsilon_{K,L})^2+(\varepsilon_{K+1,L+1}-\varepsilon_{K+1,L})^2\}\\
&+\frac{1}{2}\frac{\kappa\tilde{R}}{J}\mathbf{R}_K\cdot\mathbf{R}_{L_{(K+1/2,L+1/2)}}\\
&\times\{(\varepsilon_{K,L+1}-\varepsilon_{K+1,L})^2-(\varepsilon_{K+1,L+1}-\varepsilon_{K,L})^2\}.
\end{aligned}\tag{52}$$

Since

$$a(\varepsilon_0-\varepsilon_1)^2+b\varepsilon_1^2+c(\varepsilon_1-\varepsilon_2)^2+b(\varepsilon_0-\varepsilon_2)^2+a\varepsilon_2^2-c\varepsilon_0^2 \leqslant \frac{2(ab-c^2)}{(a+b+2c)}\varepsilon_0^2 \tag{53}$$

provided $(a+b+c)\geqslant|c|$. Hence since

$$(\mathbf{R}_K\cdot\mathbf{R}_K)(\mathbf{R}_L\cdot\mathbf{R}_L)\geqslant(\mathbf{R}_K\cdot\mathbf{R}_L)^2 \tag{53a}$$

and

$$\mathbf{R}_K\cdot\mathbf{R}_K+\mathbf{R}_L\cdot\mathbf{R}_L\geqslant 2\,|\mathbf{R}_K\cdot\mathbf{R}_L|$$

it follows that

$$(\varepsilon, B\varepsilon)\leqslant 0.$$

The matrix B is not an M-matrix since:

$$B_{(K,L)(K+1,L+1)}=-B_{(K+1,L),(K,L+1)}, \tag{54}$$

one of these terms being negative depending on the sign of $(\mathbf{R}_K\cdot\mathbf{R}_L)_{(K+1/2,L-1/2)}$ in agreement with Kershaw's theorem [2] for consistent representations. Examination of Eq. (50) shows that these terms represent flux transfer across the corner of the cell $(K+\frac{1}{2}, L+\frac{1}{2})$ between cells (K, L) and $(K+1, L+1)$, or $(K+1, L)$ and $(K, L+1)$ only, the transfer being such as to reduce the temperature difference if the element is positive, but to increase it if negative, i.e., a negative element $B_{i,j}$ corresponds to a local anti-diffusion. In general, anti-diffusion is not extremal. A general approach to extremum limited anti-diffusion has been devised by Boris and Book [11] to correct spurious diffusion introduced by numerical advection. Thus, given an anti-diffusion flux between two cells i and j, q_{ij}, such that

$$V_{ii}\varepsilon_i=V_{ii}\varepsilon_i^0-\tilde{q}_{ij}, \qquad V_{jj}\varepsilon_j=V_{jj}\varepsilon_j^0+\tilde{q}_{ij} \tag{55}$$

which is required to be locally extremal with respect to the neighbouring set of values ξ_i and ξ_j, respectively. Such an extremum limited flux is

$$q_{ij}=S\,\mathrm{Max}\{0, \mathrm{Min}[V_{ii}\,\mathrm{Max}\{S(\varepsilon_i^0-\xi_i)\}, |\tilde{q}_{ij}|, V_{jj}\,\mathrm{Max}\{S(\xi_j-\varepsilon_j^0)\}]\}, \tag{56}$$

where S is the sign of $\tilde{q}_{ij}$.

As noted earlier it follows from Theorem 2 of Appendix 2 that the form B is extremal if the time-step Δt is sufficiently large. We may obtain a sufficient condition for monotonicity by the use of Theorem 3 of Appendix 2, and by noting that the matrix, B, may be written as the sum of corner matrices of the type discussed in Appendix 3 where

$$\alpha_1 = \frac{1}{2}\frac{\kappa\tilde{R}}{J}\Delta t\mathbf{R}_L \cdot \mathbf{R}_L,$$

$$\alpha_2 = \frac{1}{2}\frac{\kappa\tilde{R}}{J}\Delta t\mathbf{R}_K \cdot \mathbf{R}_K,$$

and (57)

$$\beta_1 = \frac{1}{2}\frac{\kappa\tilde{R}}{J}\Delta t\mathbf{R}_K \cdot \mathbf{R}_L.$$

Each of which yields a monotone implicit form if the conditions given in Appendix 3 are obeyed. In general, these are not restrictive unless Δt is small, in which case the error incurred by using an explicit form for any of the terms is small. We therefore propose the following algorithm which has worked well in practice:

Test each corner matrix for monotonicity. If unsuccessful remove the appropriate anti-diffusion term from B setting appropriate elements to zero, and treat the corresponding flux explicitly using extremum limited flux (56). Solve the resultant matrix equation fully implicitly. This complete operation is clearly extremal.

An equivalent finite difference form for this problem has been elegantly devised by Kershaw [2] using a variational approach, and is also non-positive definite. The present approach appears to have two main advantages. Firstly the use of corner differencing allows a relatively simple condition for monotonicity to be identified, and if not satisfied a self consistent remedy to be adopted; and secondly no square roots are introduced. In essence the above matrix form of B is identical to that of Kershaw differing only in the way in which the interpolated values of the diffusion coefficient are calculated, the form of solution using nine point I.C.C.G. [5] being the same. As a result we have found very little difference between calculations using the above matrix B or Kershaw's form [2] with its well-known advantages [12]. In order to compare these two matrix forms we have performed test calculations on skewed meshes, such as those in Fig. 1, as suggested in Ref. [12]. In Figs. 2 and 3 we show such calculation for a small 18 × 24 mesh in which a "temperature" difference of 1 unit is applied across the mesh, with open boundary conditions in the Z direction, and reflecting ones in R. The diffusivity was uniform of value 1 unit. The mesh forms a square of 1 unit side. The time-step used was 10^{-1}. The initial value of ε within the mesh was 0. Figures 2 and 3 show the contours of ε after 50 time-steps, by which time the steady state is reached. The analytic solution has contours $\varepsilon = Z$. As can be seen both finite difference forms give a remarkably accurate representation of the true result, with the balance slightly in favour of the form B given in Eq. (50).

In practice the failure of the monotonicity conditions, and the inclusion of the extremal limit are only rarely important when the mesh is greatly distorted and Δt is small. The value of this procedure lies in naturally preventing the occasional

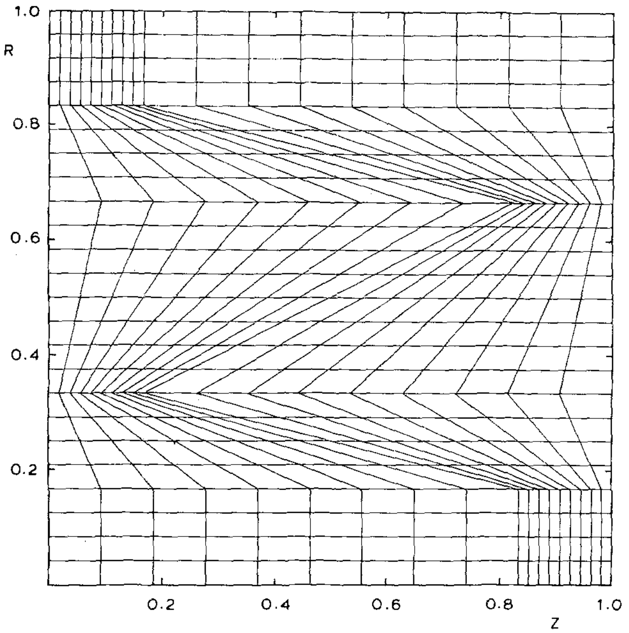

FIG. 1. Plot of the non-orthogonal mesh used in the calculations of Figs. 2–4. The mesh has 18 × 24 cells and is square of side 1 unit. A "temperature" value of 1 unit is applied at the right hand boundary and a value of 0 at the left hand boundary. The upper and lower boundaries are non-conducting.

appearance of negative energies in thermal diffusion problems, which must otherwise be reset by some ad hoc prescription such as a "zero fix-up" to prevent the programme "crashing."

A word of caution concerning the use of the extremum limited anti-diffusion flux is appropriate. If the algorithm is used without the monotonicity check, and the extremum limited flux always used, the resultant scheme is still stable and extremal, but the steady-state solution (i.e., the vector ε which reproduces itself when it is anti- [illegible] correct one. Thus as the time dependent solution approaches the steady state and Δt is allowed to increase the solution gets progressively worse. Indeed on general skewed meshes with $(\Delta t)^{-1}$ small compared to all the terms $(A^{-1}D)_{ij}$ the extremal limited solution departs markedly from the correct one, as shown by the example of Fig. 4. In particular as $\Delta t \rightarrow \infty$ the implicit part of the calculation produces a constant steady-state answer, but $\tilde{q}_{ij} \rightarrow \infty$, so all the ε's get set equal to the largest or smallest (depending on the sign of $\tilde{q}_{ij}$) of their neighbours. In fact the extremum limited solution is only satisfactory if $(\Delta t)^{-1} \gg \operatorname{Max}\{|(A^{-1}D)_{ij}|\}$.[3] This effect is avoided by the monotonicity check, and the use of the explicit anti-diffusion extremum limited flux only when the rates are small.

[3] I am indebted to the referee for this analysis of this condition.

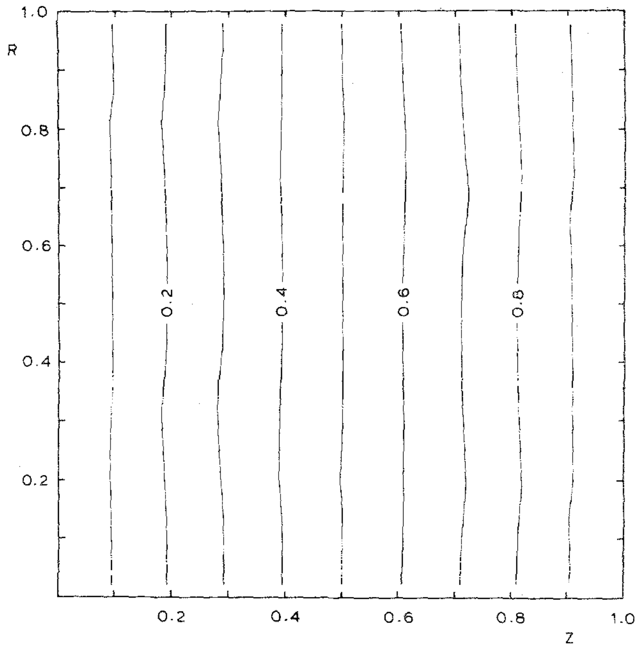

FIG. 2. Contours of constant ε generated in the steady state by the finite difference solution using the matrix B, Eq. (50), after 50 iterations with time-step 0.1. The exact solution is given by the lines $\varepsilon = Z$.

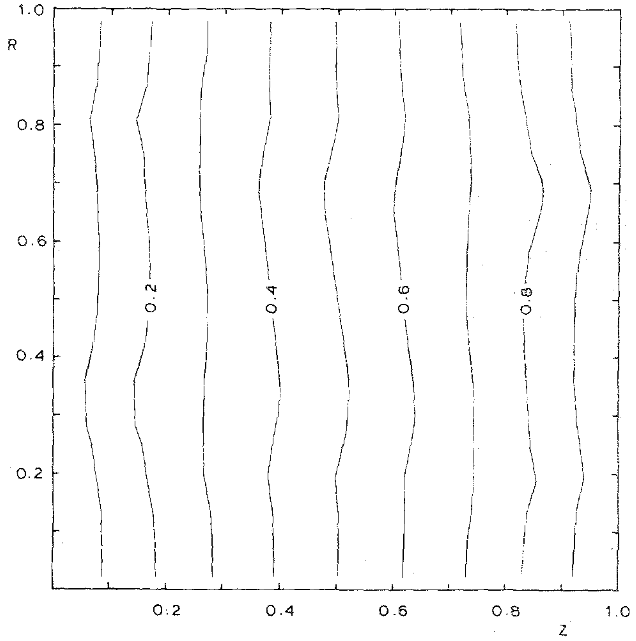

FIG. 3. Isotherms as in Fig. 2 generated by Kershaw's [2] matrix form.

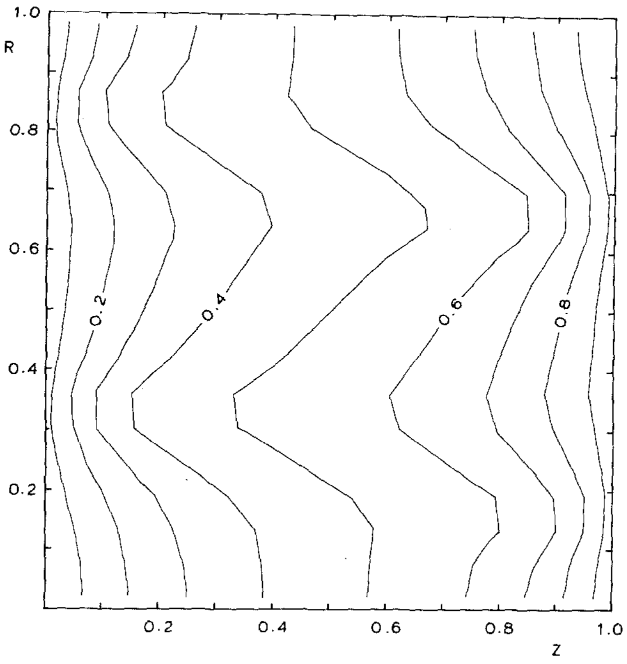

FIG. 4. Isotherms as in Fig. 2 generated by the extremal algorithms with no monotonicity check. Note the strong departure from the correct solution.

WEIGHTED FORMS

For small time-steps, Δt, the most accurate integration is given by a centred time difference, $\theta = \frac{1}{2}$, which is second order accurate. On the other hand, for large time-steps centred-difference integration oscillates, and gives rise to appreciable errors. In contrast the fully implicit scheme, $\theta = 1$, converges uniformly to the final steady state and provides a more accurate representation for large time-steps. For intermediate time-steps there exists a set of weights W_{ij} such that

$$\varepsilon = [A - D_1 \Delta t]^{-1} [A + D_2 \Delta t] \varepsilon^0, \tag{58}$$

where D_1 and D_2 are matrices such that

$$\begin{aligned} D_{1ij} &= W_{ij} D_{ij} \qquad \text{for} \quad i \neq j, \\ D_1 + D_2 &= D, \end{aligned} \tag{59}$$

where the weight W_{ij} satisfies $\frac{1}{2} \leqslant W_{ij} \leqslant 1$. In principle the weights could be chosen so as to coerce the finite difference form to the exact integration. Such an approach is extremely difficult, and involves an a priori knowledge of the solution before the weights are determined. We may however note that the term

$$[(VDV^{-1}A^{-1})_{ij}+(VDV^{-1}A^{-1})_{ji}]$$

essentially determines the relaxation rate between the elements ε_i and ε_j. The weight W_{ij} determines which time level of ε_j to use in calculating ε_i. Clearly if ε_i and ε_j rapidly relax a near steady-state distribution will be maintained between them, i.e., $W_{ij}\approx 1$. On the other hand, if the relaxation is weak, the appropriate weight is $W_{ij}\approx\frac{1}{2}$, giving highest order accuracy. Thus the weights are assumed to take the form

$$W_{ij}=F(\lambda_{ij}),$$

where

$$\begin{aligned}\lambda_{ij}&=\{([VDV^{-1}A^{-1}]_{ij}+[VDV^{-1}A^{-1}]_{ji})\,\Delta t\}\\ &[=\{(A^{-1}D)_{ij}+(A^{-1}D)_{ji}\}\,\Delta t \text{ if } (VD)=B \text{ is symmetric}]\end{aligned}\tag{60}$$

and

$$\begin{aligned}F(\lambda)&=\tfrac{1}{2} &&\text{if}\quad \lambda=0\\ &=1 &&\text{as}\quad \lambda\to\infty.\end{aligned}\tag{61}$$

Such a form is obtained for a two element system for which

$$F(\lambda)=1/\{1-\exp(-\lambda)\}-1/\lambda\tag{62}$$

coerces the finite difference form to the exact analytic solution.

In the more general case we may require $F(\lambda)$ to be of a form which maintains the physical properties of the solution. For example, if D can be differenced in an extremal form, we may require the solution of (58) to be extremal. Thus in the case both D_1 and D_2 take an M-matrix form,

$$|(A^{-1}D_2)_{ii}|\,\Delta t\leqslant 1.\tag{63}$$

The element $(D_2)_{ii}$ is evaluated from the condition that (58) be conservative, namely, that both D_1 and D_2 are separately conservative, if the weights W_{ij} are independent, i.e.,

$$(VD_2)_{ii}=-\sum_{j\neq i}(VD_2)_{ji}.\tag{64}$$

Examination of the exact two element weights (62) shows that as $\lambda\to\infty$, the weight $W\to 1-1/\lambda$ and suggests that a similar form

$$W_{ij}=\mathrm{Max}\{\tfrac{1}{2},|1-1/\alpha_{ij}\lambda_{ij}|\}\tag{65}$$

is appropriate in the general case. In comparison with the two element case, or on physical grounds we may expect that the weight is symmetric $W_{ij} = W_{ji}$, in which case Eqs. (63) and (64) show that the minimum value of the constant α_{ij} sufficient to ensure an extremal solution is given by

$$\alpha_{ij} = \alpha_{ji} = \text{Max}(N_i, N_j), \tag{66}$$

where N_i and N_j are the number of neighbours of i and j, respectively.

This term is identical to that proposed in Ref. [6] for weighting the solution in the similar problem of integration of a set of conservative, positive rate equations. Indeed we may establish this equivalence by writing Eq. (10) in the form

$$\frac{d}{dt}(AV\varepsilon) = (VDV^{-1}A^{-1})(AV\varepsilon). \tag{10a}$$

When D is the negative of an M-matrix, these form a set of conservative, positive rate equations in the conserved variables $(AV\varepsilon)$ (Eq. (14)) with transition matrix $(VDV^{-1}A^{-1})$. We note that the matrix is not differential in this form. In view of this equivalence we may expect that the performance of this weighted solution will be the same as that described earlier, a result confirmed by numerical tests. We may therefore refer to Ref. [6] where a detailed discussion and numerical tests of the performance of this weighted solution are given, and its merits compared with those of alternative forms.

As shown in Ref [6] the weights given by Eqs. (65) and (66) allow a marked improvement in the overall accuracy of the finite difference form over more simple differences, whilst retaining the desirable physical properties associated with positivity maintenance. However, the numerical tests in Ref. [6] show that the finite difference solution generally converges more slowly to the steady state, as the step length Δt is increased, than the exact solution, due to weighting too close to the fully implicit solution. In the general case when B is not symmetric no further improvement of the form (65) which maintains positivity can be made.

In the case of a isotropic medium the symmetry of the matrix B introduces a relaxation of the positivity condition (66), namely, that the weight $W_{ij} = \text{Max}\{\frac{1}{2}, [1 - 1/(\alpha_{ij}\lambda_{ij})]\}$ is positivity maintaining if

$$\alpha_{ij} \geqslant N_j/(1 + M_j/M_i), \tag{67}$$

where $M_i = (AV)_{ii}$ is the "mass" of cell i.

Let us consider the related family of weights

$$W_{ij} = 1 - (1 - \beta)\, M_j/[(N + 1)\, B_{ij}\, \Delta t] \tag{68}$$

for a symmetric system in which each cell has N neighbours. The value of α_{ij} corresponding to this weight is

$$\alpha_{ij} = \frac{1}{(1 - \beta)}(N + 1)/(1 + M_j/M_i), \tag{69}$$

β is an adjustable constant. For this set of weights the finite difference equations reduce to

$$\sum_{j \neq i}' B_{ij} \Delta t(\varepsilon_i - \varepsilon_j) = \beta M_i(\varepsilon_i - \varepsilon_i^0). \tag{70}$$

whose solution satisfies

$$\varepsilon(\beta, \Delta t) = \varepsilon(1, \Delta t/\beta), \tag{71}$$

the case $\beta = 1$ corresponding to the fully implicit scheme. In particular the case $\beta = 0$ has the remarkable property that the solution yields the steady state independently of the initial state ε^0. Comparing (69) with (67) we see that the weights (68) are positivity maintaining if $\beta \geqslant -1/N$.

The weights (68) may be put into a form which satisfies (61) and is suitable when N_i is not constant for all elements given by (65) with

$$\alpha_{ij} = \frac{1}{(1-\beta)}(N_j + 1)/(1 + M_j/M_i). \tag{72}$$

This form of weight is not symmetric, and will be referred to as the asymmetric form. The equivalent symmetric form is given by (65) with

$$\alpha_{ij} = \frac{1}{(1-\beta)} \operatorname{Max}\{(N_i + 1)/(1 + M_i/M_j), (N_j + 1)/(1 + M_j/M_i)\}. \tag{73}$$

We expect on physical grounds that the symmetric form should be preferred, as discussed earlier. In addition with symmetric weights, the matrix equation remains symmetric which allows a symmetric form of matrix inversion to be used with a substantial reduction in the overall C.P.U. time.

In order to assess the performance of these weights an extensive series of tests were carried out in which the matrix equation

$$M\frac{d\varepsilon}{dt} = -B'\varepsilon. \tag{74}$$

was solved by a single weighted implicit time-step, Δt, and the results compared with an accurate solution calculated by Gear's method [13]. In these tests B' was a differential, conservative symmetric M-matrix form. In general, as may be expected, the most severe tests were those in which all the masses and rates had the same order of magnitude, and the rate parameters λ of order unity. Figure 5 shows the results of such a typical test under these extreme conditions. B' was a dense 10×10 matrix with off-diagonal elements in the range 0.1–1.0, and cell "masses" in the range 0.1–10.0. The initial conditions were $\varepsilon_1 = 1.0$ and $\varepsilon_i = 0.0$ $(i \neq 1)$. The graphs show the values of ε_1 (which showed the largest error) as functions of the time-step, Δt,

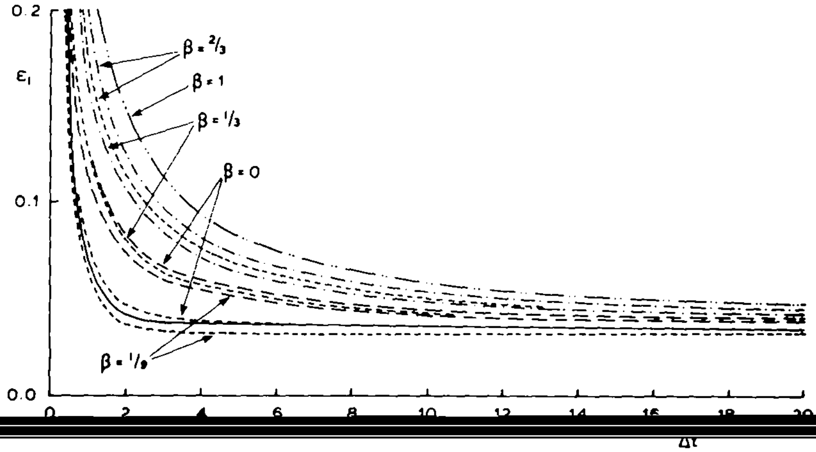

FIG. 5. Comparison of the finite difference solution of the set of conservative, positive rate equations: $M_i(d\varepsilon_i/dt) = -\sum B'_{ij}\varepsilon_j$ for various step lengths, Δt, with the exact solution (full line) calculated by Gear's method [13]. The calculation involves a set of 10 components, with B' a dense symmetric ($B'_{ij} = B'_{ji}$) matrix ($N = 9$) with off-diagonal values in the range 0.1–1.0, and cell "masses" M_i in the range 0.1–10.0. The initial conditions were $\varepsilon_1 = 1.0$, $\varepsilon_i = 0.0$ $(2 \leqslant i \leqslant 10)$. The finite difference calculations were performed with both asymmetric (dotted curves), $W_{ij} = \mathrm{Max}\{\frac{1}{2}, [1 - (1-\beta) M_j/(10B'_{ij}\Delta t)]\}$, and symmetric (dashed curves) weights, $W_{ij} = W_{ji} = \mathrm{Max}\{\frac{1}{2}, [1 - (1-\beta)\,\mathrm{Min}(M_i, M_j)/(10B'_{ij}\Delta t)]\}$, for various values of β. We note that the cases $\beta = 1$, $\beta = 0$, $\beta = -\frac{1}{9}$ correspond to the fully implicit, steady state and limiting positivity maintaining solutions, respectively. The values of the component, ε_1, only have been plotted since this term shows the largest error. We note the improved performance of the asymmetric weights.

calculated for asymmetric (72)—dotted curves—and symmetric (73)—dashed curves—weights for values of β in the range $-1/N$ to 1.

Examination of Fig. 5 shows several results typical of these tests. The asymmetric form is clearly superior to the symmetric, but the latter may be preferred for the reasons given earlier. The value $\beta = 0$ for the asymmetric case gives a remarkably accurate representation, but in the symmetric one the positivity maintaining form is generally superior where

$$\alpha_{ij} = \alpha_{ji} = \mathrm{Max}\{N_i/(1 + M_i/M_j), N_j/(1 + M_j/M_i)\}. \tag{75}$$

The inclusion of these weights into a calculation where the diffusion matrix can be cast in an M-matrix form is extremely simple, and requires little computational effort, yet as the calculation in Fig. 5 can lead to marked improvement in the accuracy of the calculations in the case where Δt is neither large nor small. In the case that B does not have an M-matrix form such weighting is not generally possible since only the fully implicit form can be made extremal. Nonetheless when the M-matrix terms dominate as in the matrix (50) it may be possible to weight these terms, but such a procedure will probably not lead to a significant increase in accuracy, unless the residual terms are all small.

We may remark that the stability analysis given earlier only holds if the weights are all equal. Sets of weights, which are extremal, are by definition bounded, and therefore stable.

Conclusion

The general properties of the diffusion equation have been examined, and it is shown that the diffusion operator is differential, conservative and non-positive definite. Each of these properties is the equivalent of an important physical law; for example, in the case of thermal conduction, the conditions differential, conservative and non-positive definite correspond to the zeroth, first and second laws of thermodynamics, respectively. In addition the diffusion equation is shown to satisfy an extremal condition which is the mathematical expression of the familiar phsyical picture of diffusion involving the progressive smoothing of extrema [4], and furthermore ensures the essential positivity of the phenomenon. The extremal condition is closely related to, but more restrictive than the non-positive definite one. In the past finite difference schemes for the calculation of diffusion have considered the significance of the first three conditions, namely, differential, conservative and non-positive definiteness, but little attention has been paid to that of the extremum, despite its physical significance.

The essential linearity of the diffusion equation implies that the spatial finite difference representation of the diffusion operator must take a matrix form, whose structure is consistent with that of the operator. It is furthermore natural to require that the essential physical properties of the operator be retained by the matrix, namely, that the matrix be differential, conservative, non-positive definite, extremal.... General conditions on the matrix form may be derived. In particular if the matrix differential equation is extremal, the matrix must be negative of an M-matrix form, a condition which may not be possible to satisfy if the representation is consistent [2].

In practice, we shall also perform the temporal integration by means of a finite difference form. The conditions of differential, conservative and stability (corresponding to non-positive definite) may be ensured if the differencing is two-step with an implicitness parameter, θ, in the range $\frac{1}{2} \leqslant \theta \leqslant 1$. Furthermore if the diffusion matrix is the negative of an M-matrix form, the differencing may be extremal. This suggests that the differencing be made extremal, by means of a suitable weight function in such a fashion that for small time-steps the differencing is split-time-step and for large ones fully implicit. Such a weight has been derived for the related problem of a set of positive rate equations, and found to markedly increase the accuracy of solution, with little extra work.

In order that the temporal difference be extremal, it is convenient to ensure that the diffusion matrix take a negative M-matrix form. Two examples are given in which this is not the case. In the first this is associated with advection-like terms $\partial\varepsilon/\partial z$, etc., and an M-matrix form can be recovered by upstream or downstream differencing to ensure convergence to the appropriate neighbouring value of ε [10]. No such simple remedy is available in the second case associated with a term of form $\partial\varepsilon^2/\partial k\,\partial l$. By differencing the problem in terms of fluxes, the departures from an M-matrix form can be identified as fluxes enhencing the term difference, i.e., anti-diffusion fluxes. By retaining the flux form, we may devise a condition under which the solution is extremal, outside this limit we may conservatively treat the monotone terms separately in the usual way, and use an extremal form for the anti-diffusion fluxes, to obtain an operation which is overall extremal. Since the only remaining terms in the diffusion operator are, of the type $\partial\varepsilon^2/\partial z^2$, which readily difference in an M-matrix form, it is believed the above methods will allow any general diffusion operator to be differenced in an extremal form.

It is a direct consequence of the definition (9) that the development of an extremal operator is bounded, and therefore stable [1]. In consequence we may deduce that an extremal form is always stable, irrespective of any temporal or spatial variations of the coefficients $\kappa_{\|}$, $\kappa_{\perp}$ and κ_{Λ}. Furthermore it also follows that repeated application of the operation will converge uniformly (in respect to the maximum norm) onto the uniform equilibrium state [6].

In the present analysis it has been assumed that the boundaries are impermeable. In general, boundaries on which the value of the parameter, ε, is specified may also occur. In this case, the general conclusions reached as to the matrix form of D remain unchanged, although the extremal condition must include the boundary values, as well as those at an earlier time, as in Eq. (9).

APPENDIX 1: Physical Constraints on the Diffusion Operator

Any solution of the diffusion equation must obey the second law of thermodynamics which implies that the total entropy in a closed volume, V, must increase in time. In accordance with Onsager's theory it follows that any diffusion process is described by an appropriate thermodynamic equation of motion [14] relating the flux $\mathbf{q}$ to an appropriate force $\mathbf{X}$: the force $\mathbf{X}$ being related to $\nabla\varepsilon$ by

$$\mathbf{X} = p\nabla\varepsilon, \tag{A1.1}$$

where p is a real, positive function. The entropy production rate per unit volume is then given by

$$T\dot{\Theta} = \mathbf{X}\cdot\mathbf{q} \tag{A1.2}$$

where T is the absolute temperature. Hence since $P = p/T$ is a real positive function:

$$\int_V \nabla\varepsilon\cdot\underline{\boldsymbol{\kappa}}\cdot\nabla\varepsilon P\,dV \geqslant 0. \tag{A1.3}$$

If ∇P is not parallel to $\nabla\varepsilon$, this condition is only satisfied for all fields, ε, if diffusion is non-negative in the sense

$$\int_V \mathbf{A}\cdot\underset{\sim}{\boldsymbol{\kappa}}\cdot\mathbf{A}\,dV \geqslant 0 \tag{A1.4}$$

for all vector fields $\mathbf{A}$. In the case that P is a function of ε only, so that ∇P is parallel to $\nabla\varepsilon$, this condition takes the simpler form that the diffusion operator must be nonpositive definite in the sense

$$-\int \nabla\varepsilon\cdot\underset{\sim}{\boldsymbol{\kappa}}\cdot\nabla\varepsilon\,dV = \int \varepsilon\mathscr{D}(\varepsilon)\,dV \leqslant 0 \tag{A1.5}$$

for all functions ε. This condition is of course necessary in the general case.

Since the eigenvalues, λ, of the operator $\mathscr{D}$ given by

$$\mathscr{D}A = \lambda A \tag{A1.6}$$

have real parts, $\mathrm{Re}(\lambda) \leqslant 0$ when the operator is non-positive definite, the above result ensures that the solutions are mathematically stable for forward going time. The entropy production rate

$$\dot{\Theta} = \int_V \nabla\varepsilon\cdot\underset{\sim}{\boldsymbol{\kappa}}\cdot\nabla\varepsilon P\,dV = \int \nabla\varepsilon\cdot\underset{\sim}{\boldsymbol{\kappa}}_s\cdot\nabla\varepsilon P\,dV \tag{A1.7}$$

depends on the symmetric part of the diffusivity tensor $\underset{\sim}{\boldsymbol{\kappa}}$alone, and therefore on the symmetric part of the operator D. The symmetric component therefore represents an irreversible evolution, the anti-symmetric part being reversible. It follows that the principal components of the symmetric parts of the diffusivity tensor κ , κ must be positive.

Expanding Eq. (1) to (3) into component form we see that Eq. (1) is a parabolic equation in space and time provided $\partial a/\partial t \geqslant 0$ everywhere. In this case it follows from the maximum theorem [3] that the function, ε, cannot have a positive maximum in the space $(\mathbf{r}, t)$ except on a boundary in time, t_0, or space S. In other words if a positive maximum exists in the initial data, or at a spatial boundary it will decay in time. In the important case that a is constant, we may extend the result to include minima by considering the function

$$\varepsilon' = \varepsilon_{\max} - \varepsilon, \tag{A1.8}$$

where $\varepsilon_{\max}$ is the supremum of all values of ε. Clearly in this case ε' also satisfies Eq. (1) and has positive maxima at the minima of ε. Thus if a is constant all spatial extrema decay in time. A slightly weaker condition is

$$\begin{aligned}\mathrm{Max}|\varepsilon(t_0), \varepsilon_b| \geqslant \mathrm{Max}|\varepsilon(t)| \geqslant \mathrm{Min}|\varepsilon(t)|\\ \geqslant \mathrm{Min}|\varepsilon(t_0), \varepsilon_b|.\end{aligned} \tag{A1.9}$$

where $t > t_0$, Max and Min are the largest and smallest values of the set denoted respectively, and ε_b the boundary values of ε during the interval t_0 to t, respectively. We shall call this *the extremal condition.*

When a is not constant the behaviour of ε at an extremum can be clearly shown, for at a spatial maximum of ε, $\nabla\varepsilon = 0$ and $\nabla^2\varepsilon < 0$ and

$$\frac{\partial}{\partial t}(a\varepsilon) = \nabla \cdot (\underset{=}{\kappa} \cdot \nabla\varepsilon) = \kappa_{\parallel}\nabla_{\parallel}\varepsilon^2 + \kappa_{\perp}\nabla_{\perp}^2\varepsilon < 0 \tag{A1.10}$$

and the function $(a\varepsilon)$ decreases. Similarly at a minimum $(a\varepsilon)$ increases.

An operator $G(\varepsilon)$ is called extremal if the solutions of the equation

$$\varepsilon(t) = G\{\varepsilon(t_0), \varepsilon_b\} \tag{A1.11}$$

satisfy the extremal condition. It can be shown that an operator is extremal if and only if it is both differential and monotonic in the sense that if ε and δ are two sets of the variable such that if $\mathrm{Min}(\varepsilon) \geqslant \mathrm{Max}(\delta)$ then $\mathrm{Min}[G(\varepsilon)] \geqslant \mathrm{Max}[G(\delta)]$. When G is linear this monotonicity condition is equivalent to that of non-negativity in Eq. (A1.11).

APPENDIX 2: Miscellaneous Matrix Theorems

LEMMA 1. *A differential (or conservative) matrix, a, is singular*: for the determinant

$$a = \begin{vmatrix} a_{11} & a_{12} & \cdots \\ a_{21} & a_{22} & \\ \vdots & & \end{vmatrix} = \begin{vmatrix} 0 & 0 & \cdots \\ a_{21} & a_{22} & \\ \vdots & & \end{vmatrix} = 0 \tag{A2.1}$$

LEMMA 2. *A non-positive definite real matrix B has eigenvalues, λ, whose real parts satisfy* $\mathrm{Re}(\lambda) \leqslant 0$.

Let $A = \Phi + i\Psi$ be an eigenvector of B with eigenvalue $\lambda = \phi + i\psi$. Then

$$\begin{aligned}(A^*, BA) &= \lambda(A^*, A) = (\phi + i\psi)[(\Phi, \Phi) + (\Psi, \Psi)] \\ &= [(\Phi, B\Phi) + (\Psi, B\Psi)] + i[(\Phi, B\Psi) - (\Psi, B\Phi)],\end{aligned}$$

$$\mathrm{Re}(\lambda) = \phi = [(\Phi, B\Phi) + (\Psi, B\Psi)]/[(\Phi, \Phi) + (\Psi, \Psi)] \leqslant 0 \tag{A2.2}$$

since B is non-positive definite.

THEOREM 1. *A matrix (M − B), where M is a real, positive diagonal matrix (non-null) and B is a non-positive definite real matrix, is non-singular.*

Since the matrix $(M - B)$ is positive definite it follows from an identical proof to

Lemma 2 that the eigenvalues of $(M-B)$ have positive non-zero real parts. Hence the determinant of $(M-B)$ is non-zero and positive, for since $(M-B)$ is real the eigenvalues must be either real positive or occur as complex conjugate pairs.

LEMMA 3. *If a is a differential and conservative matrix, all the co-factors $|a|_{ij}$ in the determinant of a are equal.*

Consider the determinant

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & \\ a_{21} & & & \\ \vdots & & a_{ii}+1 & \cdots \\ & & \vdots & \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} & \cdots & 0 & \cdots & a_{ii} & \cdots \\ \vdots & & & & & & \\ a_{i1} & & \cdots & 1 & \cdots & a_{ii}+1 & \cdots \\ \vdots & & & & & & \end{vmatrix}$$
$$= |a|_{ij}, \tag{A2.3}$$

where the column j is arbitrary, since a is differential. Similarly since a is conservative, the value of the co-factor, is independent of i.

THEOREM 2. *If a is a differential and conservative matrix, then the matrix $(M-a)$ where M is a real, positive diagonal matrix, is monotone for sufficiently small M.*

The determinant

$$\begin{vmatrix} M_1 - a_{11} & -a_{12} & \cdots \\ -a_{21} & M_2 - a_{22} & \cdots \\ \vdots & \vdots & \end{vmatrix} = -\sum_i M_i |a|_{ii} + O(M_i M_j) \tag{A2.4}$$

since by Lemma 1, a is singular. The co-factor of the element ij is

$$|M-a|_{ij} = -|a|_{ij} + O(M_i). \tag{A2.5}$$

Hence making use of the results of Lemma 3 the inverse matrix is given by

$$([M-a]^{-1})_{ij} = \frac{1}{\sum_i M_i} + O(1). \tag{A2.6}$$

The solution thus obtained corresponds to the steady state of a differential, conservative process:

$$\sum_j (M-a)_{ij} \varepsilon_j = M_i \varepsilon_i^0, \tag{A2.7}$$

where

$$\varepsilon_j = \sum_i M_i \varepsilon_i^0 \Big/ \sum_i M_i. \tag{A2.8}$$

Lemma 4. *If a is a differential matrix, then the inverse non-singular matrix $[I-a]^{-1}$ satisfies*

$$\sum_j [(I-a)^{-1}]_{ij} = 1. \tag{A2.9}$$

For if a is differential then

$$\varepsilon_i = \sum_j [(I-a)]_{ij}^{-1}\,\varepsilon_j^0 \tag{A2.10}$$

and $\varepsilon_i = \varepsilon^0 = \varepsilon_j$ if all ε_j^0 are equal,and the result follows. Alternatively, the determinant

$$|I-a| = \begin{vmatrix} 1-a_{11} & -a_{12} & \cdots \\ -a_{21} & 1-a_{22} & \\ \vdots & & \end{vmatrix} = \begin{vmatrix} 1 & 1 & \cdots \\ -a_{21} & 1-a_{22} & \\ \vdots & & \end{vmatrix}$$

$$= \sum_j |I-a|_{ij} \text{ for arbitrary } i, \tag{A2.11}$$

where $|I-a|_{ij}$ is the co-factor of the element ij in the determinant $|I-a|$ of $I-a$, and the result (A2.9) follows directly. An equivalent result for a conservative matrix is readily proven.

Lemma 5. *If a is a differential (or conservative) matrix, and the inverse matrix $[(I-a)^{-1}]$ is monotone, then the transformed matrix*

$$a^1 = (I-a)^{-1}a = (I-a)^{-1} - I \tag{A2.12}$$

has a negative differential M-matrix form.

Since $(I-a)^{-1}$ is monotone $a_{ij}^1 = [(I-a)^{-1}]_{ij} \geqslant 0$ and $a_{ii}^1 = [(I-a)^{-1}]_{ii} - 1 = -\sum_{j \neq i} a_{ij}^1 < 0$.

Corollary. *$(I-a^1)$ is monotone.*

The following matrix identities are readily established:

$$[I-(a_1+a_2)] = (I-a_1)(I-a_1^1a_2^1)(I-a_2) \tag{A2.13}$$

and

$$[I-a_1a_2] = (I-a_1)[I-(a_1^1+a_2^1)](I-a_2), \tag{A2.14}$$

where the superscript 1 represents the transformation (A2.12). Hence

$$[I-(a_1+a_2)]^{-1} = (I-a_2)^{-1}(I-a_2^1)^{-1}[I-(a_1^{11}+a_2^{11})]^{-1}(I-a_1^1)^{-1}(I-a_1)^{-1} \tag{A2.15}$$

and the following theorem is a direct consequence of Lemma 5.

THEOREM 3. *If a_1 and a_2 are differential (or conservative) matrices and $(I-a_1)$ and $(I-a_2)$ are monotone then (a_1+a_2) is differential (or conservative) and $|I-(a_1+a_2)|$ is monotone.*

This theorem may be obviously generalised to a set of matrices $a_1, a_2 \cdots$ of the above type, for which $(I-a_i)$ is monotone.

CONDITION 1. A differential, conservative matrix, B, is non-positive definite if

$$\begin{aligned}(\varepsilon, B\varepsilon) &= \sum_{ij} \varepsilon_i B_{ij} \varepsilon_j \\ &= \sum_i \varepsilon_i \sum_{j \neq i} B_{ij}(\varepsilon_j - \varepsilon_i) \\ &= \sum_j \varepsilon_j \sum_{i \neq j} B_{ij}(\varepsilon_i - \varepsilon_j) \\ &= -\tfrac{1}{2} \sum_j \sum_{i \neq j} B_{ij}(\varepsilon_i - \varepsilon_j)^2 \\ &= -\tfrac{1}{4} \sum_j \sum_{i \neq j} (B_{ij} + B_{ji})(\varepsilon_i - \varepsilon_j)^2 \leqslant 0 \qquad \text{(A2.16)}\end{aligned}$$

for all values of ε_i. The condition $\sum_{i \neq j}(B_{ij}+B_{ji}) \geqslant 0$ or $B_{ii} \leqslant 0$ is clearly necessary and $(B_{ij}+B_{ji}) \geqslant 0$ sufficient.

APPENDIX 3: THE CORNER MATRIX

We define a corner matrix, associated with the corner $(K+\frac{1}{2}, L+\frac{1}{2})$ of the Lagrangian mesh, with non-zero components only between the cells 1, (K, L); 2, $(K+\frac{1}{2}, L)$; 3, $(K+1, L+1)$ and 4, $(K, L+1)$ as the form

$$b = \begin{pmatrix} M_1+(\alpha_1+\alpha_2-\beta_1) & -\alpha_1 & \beta_1 & -\alpha_2 \\ -\alpha_1 & M_2+(\alpha_1+\alpha_2+\beta_1) & -\alpha_2 & -\beta_1 \\ \beta_1 & -\alpha_2 & M_3+(\alpha_1+\alpha_2-\beta_1) & -\alpha_1 \\ -\alpha_2 & -\beta_1 & -\alpha_1 & M_4+(\alpha_1+\alpha_2+\beta_1) \end{pmatrix} \qquad \text{(A3.1)}$$

which we may write more compactly as

$$b_{ij} = -\alpha, \qquad b_{ik} = \beta, \qquad b_{il} = -\alpha^1, \qquad b_{ii} = M_i + (\alpha + \alpha^1 - \beta), \qquad \text{(A3.2)}$$

where the sequence (i, j, k, l) is taken cyclically around the cells $(1, 2, 3, 4)$, M_i is the cell "masses" $(=A_{ii} V_{ii})$ and

$$\begin{aligned} &\alpha = \alpha_1, \quad \alpha^1 = \alpha_2, \quad \beta = \beta_1 \qquad \text{if } i \text{ is odd,} \\ &\alpha = \alpha_2, \quad \alpha^1 = \alpha_1, \quad \beta = -\beta_1 \qquad \text{if } i \text{ is even.} \qquad \text{(A3.3)}\end{aligned}$$

The matrix equation

$$b\varepsilon = M\varepsilon^0 \tag{A3.4}$$

is clearly both differential and conservative.

It follows from Theorem 1 of Appendix 2 and the non-positive definite nature of $(M - b)$ that b is non-singular and its inverse is thus,

$$(b^{-1})_{ij} = |b|_{ij}/|b|, \tag{A3.5}$$

where $|b|_{ij}$ is the co-factor of the element ij in the determinant of b, $|b|$. The determinant

$$\begin{aligned}|b| = {} & M_1M_2M_3M_4 + M_2M_4(M_1 + M_3)(\alpha_1 + \alpha_2 - \beta_1)\\ & + M_1M_3(M_2 + M_4)(\alpha_1 + \alpha_2 + \beta_1)\\ & + M_1M_3[(\alpha_1 + \alpha_2 + \beta_1)^2 - \beta_1^2] + M_2M_4[(\alpha_1 + \alpha_2 - \beta_1)^2 - \beta_1^2]\\ & + (M_1M_2 + M_3M_4)[(\alpha_1 + \alpha_2)^2 - \alpha_1^2 - \beta_1^2]\\ & + (M_1M_4 + M_2M_3)[(\alpha_1 + \alpha_2)^2 - \alpha_2^2 - \beta_1^2]\\ & + 2(M_1 + M_2 + M_3 + M_4)(\alpha_1 + \alpha_2)(\alpha_1\alpha_2 - \beta_1^2).\end{aligned} \tag{A3.6}$$

In accordance with Lemma 1 of Appendix 2, $|b| = 0$ if $M_1 = M_2 = M_3 = M_4 = 0$, and in view of (53a), $|b| > 0$ if not in accord with Theorem 1.

The co-factors can be conveniently written in terms of the cyclic parameters (i, j, k, l) as in (A3.2).

$$\begin{aligned}|b|_{ii} = {} & M_jM_kM_l + M_jM_l(\alpha + \alpha^1 - \beta) + M_k(M_j + M_l)(\alpha + \alpha^1 + \beta)\\ & + M_j[(\alpha + \alpha^1)^2 - \alpha^2 - \beta^2] + M_l[(\alpha + \alpha^1)^2 - \alpha^{12} - \beta^2]\\ & + M_k[(\alpha + \alpha^1 + \beta)^2] + 2(\alpha + \alpha^1)(\alpha\alpha^1 - \beta^2).\end{aligned}$$

$$\begin{aligned}|b|_{ij} = {} & M_kM_l\alpha + (\alpha + \alpha^1)[M_l(\alpha - \beta) + M_k(\alpha + \beta)]\\ & + 2(\alpha + \alpha^1)(\alpha\alpha^1 - \beta^2).\end{aligned}$$

$$\begin{aligned}|b|_{ik} = {} & -M_jM_l\beta + (M_j + M_l)(\alpha\alpha^1 - (\alpha + \alpha^1)\beta - \beta^2)\\ & + 2(\alpha + \alpha^1)(\alpha\alpha^1 - \beta^2).\end{aligned}$$

$$\begin{aligned}|b|_{il} = {} & M_jM_k\alpha^1 + (\alpha + \alpha^1)[M_j(\alpha^1 - \beta) + M_k(\alpha^1 + \beta)]\\ & + 2(\alpha + \alpha^1)(\alpha\alpha^1 - \beta^2).\end{aligned} \tag{A3.7}$$

In accordance with Lemma 3 of Appendix 2 we note that all the co-factors take the same positive value $2(\alpha + \alpha^1)(\alpha\alpha^1 - \beta^2)$, and the inverse b^{-1} is therefore positive (Theorem 2) as $M_l \rightarrow 0$. Furthermore in accordance with Theorem 1 the diagonal co-factors are always positive and therefore $(b^{-1})_{ii} > 0$.

The departure from monotone character of the matrix b is shown by negative values of the off-diagonal elements of b^{-1}, and is a result of the anti-diffusion terms β_1 with positive sign. This causes problems with both the diagonal (across corner) elements $|b|_{ik}$ and, less expectedly, the direct (across face) ones $|b|_{ij}$ and $|b|_{il}$; the latter are, however, in general less restrictive. The physical explanation of these effects is straightforward:

(a) $|b|_{ik} < 0$. Suppose $\varepsilon_1^0 = \varepsilon_2^0 = \varepsilon_3^0 = 0$, $\varepsilon_3^0 \neq 0$ and suppose M_2, $M_4 \gg \alpha_1$, α_2; then the flux through the faces 2/3 and 3/4 does not markedly increase ε_2 or ε_4 from zero. The flux from 1 is thus dominated by the anti-diffusion flux to 3 decreasing ε_1 below zero.

(b) $|b|_{ij} < 0$. Suppose $\varepsilon_1^0 = \varepsilon_3^0 = \varepsilon_4^0 = 0$, $\varepsilon_2^0 \neq 0$ and suppose $M_4 \gg M_3$ and $\alpha_2 \gg \beta_1$, α_1. Rapid flux transfer increases ε_3 across the face 2/3, giving rise to the anti-diffusion flux 1 to 3 decreasing ε_1 below zero. Little compensating flow across face 1/2 occurs due to the small value of α_1, or across 1/4 due to the large mass of 4 (keeping ε_4 small).

It is evident from consideration of the co-factors (A3.7) that they only assume negative values if the values of M are large compared to the α, β coefficients, in particular if $M \gtrsim (\alpha_1 + \alpha_2)$. Examination of Eq. (49) shows that this parameter is the Taylor expansion parameter, i.e., the failure of the monotone character of b occurs for the same values of Δt for which the explicit treatment may be used. We may also note that the onset of non-monotonicity may be inhibited if the cell masses M_i are nearly equal.

Acknowledgments

I would like to thank A. J. Bennett for many helpful discussions on this work. This work was part of a programme supported by the Rutherford Laboratory.

References

1. R. D. Richtmyer and K. W. Morton, "Difference Methods for Initial Value Problems," Interscience, New York, 1967.
2. D. S. Kershaw, *J. Comput. Phys.* **39** (1981), 375.
3. A. Friedman, "Partial Differential Equations of the Parabolic Type," Prentice-Hall, Englewood Cliffs, N. J., 1964.
4. P. M. Morse and H. Feshbach, "Methods of Theoretical Physics," McGraw-Hill, New York, 1953.
5. D. S. Kershaw, *J. Comput. Phys.* **26** (1978), 43.
6. G. J. Pert, *J. Comput. Phys.* **39** (1981), 251.
7. S. I. Braginskii, "Reviews of Plasma Physics" (M. A. Leontovich, Ed.), Vol. 1, p. 265, Consultants Bureau, New York, 1965.
8. R. S. Varga, "Matrix Iterative Analysis," Prentice-hall, Englewood Cliffs, N. J., 1962.
9. L. Collatz, "The Numerical Treatment of Differential Equations," Springer-Verlag, Berlin, 1960.

10. G. J. PERT, *J. Comput. Phys.*, in press.
11. J. P. BORIS AND D. L. BOOK, *J. Comput. Phys.* **11** (1973), 38.
12. D. S. KERSHAW, *in* "Laser Program Annual Report," pp. 4–52, Lawrence Livermore Laboratory No. UCRL-50021-76, 1977.
13. C. W. GEAR, "Numerical Initial Value in Ordinary Differential Equations," Prentice-Hall, Englewood Cliffs, N. J., 1971.
14. K. G. DENBIGH, "The Thermodynamics of the Steady State," Methuen, London, 1965.