Numerical Analysis of Discrete Ordinate Methods

I. P. GRANT

Atlas Computer Laboratory, Chilton, Didcot, Berkshire, England

Abstract

It is shown that the DSN equations for spherical nonscattering time-dependent systems can be put in matrix form, and that one can then apply well-known theorems to obtain criteria for stability and monotonicity of the difference operator. The influence of the choice of auxiliary equations needed for full definition of the solution is analyzed in some detail.

Numerical illustrations of the theory are discussed.

1. INTRODUCTION

Although discrete ordinate methods have been employed for many years to solve problems in neutron transport theory and radiative transfer, a definitive mathematical analysis of the numerical procedures used remains to be evolved. As Lathrop and Carlson [1] have recently remarked, "this lack of mathematical guidance is regrettable, for when solution methods are proposed, errors cannot be analyzed and one is reduced to comparing results from what may be equally imprecise algorithms". They have attempted in their paper to alleviate their difficulties by employing physical arguments. Our object in this present paper, which covers similar ground, is to derive some mathematical understanding of the difficulties with the aid of well-known matrix theorems.

We have chosen to analyze the problems involved for spherically symmetric, nonscattering, time-dependent systems. Scattering introduces difficulties of another order, and we have preferred to consider them separately at this stage [2]. The transfer equation for our system may be written

$$L[I] = \frac{1}{c} \frac{\partial I}{\partial t} + \mu \frac{\partial I}{\partial r} + \frac{1 - \mu^2}{r} \frac{\partial I}{\partial \mu} = \sigma(r, t)[B(r, t) - I], \qquad (1.1)$$

where the specific intensity $I \equiv I(r, \mu, t)$ satisfies the initial and boundary conditions

$$I(r, \mu, t) = \Phi(r, \mu); \quad t = 0, -1 \le \mu \le 1, 0 \le r \le R$$

= $\Psi(\mu, t); \quad t > 0, -1 \le \mu < 0, r = R.$ (1.2)

GRANT

Both $\sigma(r, t)$ and B(r, t) are taken as prescribed functions over the region $0 \leq r \leq R$, $0 \leq t \leq T$. We shall assume that $\sigma(r, t)$ is, at worst, piecewise continuous and is suitably differentiable except at discontinuities. We shall assume B(r, t) to be continuous and to be suitably differentiable except possibly at points of discontinuity of $\sigma(r, t)$.

In the following section we set up discrete ordinate difference equations in much the same way as Lathrop and Carlson [1], and rewrite them in matrix form. Properties of the matrix and their bearing on the behavior of numerical solutions are set out in Section 3. In Section 4 we briefly examine discretization errors, and discuss how these are affected by different choices of the auxiliary equations required to fully define the numerical solution. Some numerical solutions of a test problem which can be solved analytically are examined in Section 5 in order to illustrate the theory.

2. The S_n Difference Operator

2.1. Notation

We begin by partitioning the range $0 \le t \le T$, $0 \le r \le R$, $-1 \le \mu \le 1$ into cells $\Omega \equiv (s + \frac{1}{2}, i + \frac{1}{2}, m + \frac{1}{2})$ centred on the point $(t_{s+1/2}, r_{i+1/2}, \mu_{m+1/2})$. We suppose that the partitioning is such that

$$0 \leq t_s < t_{s+1/2} < t_{s+1} \leq T; s = 0, 1, ..., S,$$

$$r_i < r_{i+1/2} < r_{i+1}; i = 1, 2, ..., N; 0 \leq r_1, r_{N+1} = R,$$

$$\mu_m < \mu_{m+1/2} < \mu_{m+1}; m = -M, -M + 1, ..., M - 1; \mu_{\pm M} = \pm 1.$$

We shall write for each cell

$$V_{i+1/2} = \int_{(i+1/2)} dV = \int_{r_i}^{r_{i+1}} 4\pi r^2 dr, \qquad (2.1)$$

$$W_{m+1/2} = \int_{(m+1/2)} d\mu = \int_{\mu_m}^{\mu_{m+1}} d\mu, -M \leqslant m \leqslant M-1, \qquad (2.2)$$

$$\Delta_{s+1/2} = \int_{(s+1/2)} dt = \int_{t_s}^{t_{s+1}} dt, \qquad (2.3)$$

together with

$$\beta_m = -\int_{-1}^{\mu_m} \mu \, d\mu = \frac{1}{2}(1-\mu_m^2), \qquad m = -M, -M+1, \dots, M, \quad (2.4)$$

and

$$\alpha_{m,i+1/2} = (A_{i+1} - A_i) \beta_m, \quad i = 1, ..., N; m = -M, ..., M,$$
 (2.5)

where $A_i = 4\pi r_i^2$. We shall regard the interval lengths $W_{m+1/2}$ and division points $\mu_{m+1/2}$ as the weights and abscissas of some quadrature formula over [-1, 1] (or [-1, 0] and [0, 1] separately). The precise nature of this rule is unimportant for our work. For a discussion of various possibilities we may refer, for example, to Carlson [3]. All that we shall do is to impose conditions

$$\sum_{-M}^{M-1} W_{m+1/2} = 2, \quad \sum_{-M}^{M-1} W_{m+1/2} \mu_{m+1/2} = 0, \quad \sum_{-M}^{M-1} W_{m+1/2} \mu_{m+1/2}^2 = \frac{2}{3}.$$
 (2.6)

We shall frequently assume that the partition of the μ range is symmetric, namely $\mu_{-m-1/2} = -\mu_{m+1/2}$, m = 0, ..., M - 1, and we shall then impose further conditions, namely,

$$\sum_{-M}^{-1} W_{m+1/2} = 1; \quad \sum_{-M}^{m-1} W_{k+1/2} \mu_{k+1/2} = -\beta_m, \quad m = -M + 1, ..., M \quad (2.7)$$

with $\beta_{\pm M} = 0$. In particular, $\beta_0 = -\frac{1}{2}$. Notice that we do not need to consider the points μ_m for integer *m* explicitly.

Following convention, we suppress suffixes whose values are easily determined from the context in order to keep the equations reasonably tidy. We shall make use of integrals over segments of the boundary of the cell Ω as follows:

$$I_{s} = \int_{\Omega} \delta(t - t_{s}) I(r, \mu, t) \, dV \, d\mu \, dt / V_{i+1/2} W_{m+1/2}$$

$$I_{s+1} = \int_{\Omega} \delta(t - t_{s+1}) I(r, \mu, t) \, dV \, d\mu \, dt / V_{i+1/2} W_{m+1/2} \,.$$
(2.8)

and

These define mean specific intensities crossing the surfaces
$$t = t_s$$
, $t = t_{s+1}$, respectively. The notation $\delta(x)$ is used for Dirac's delta function. Similarly, we need

$$I_{i} = \int_{\Omega} \delta(r - r_{i}) \, \mu I(r, \, \mu, \, t) \, dV \, d\mu \, dt / W_{m+1/2} \mu_{m+1/2} \varDelta_{s+1/2} \,, \qquad (2.9)$$

$$I_m = \int_{\Omega} \delta(\mu - \mu_m) \frac{1}{r} I(r, \mu, t) \, dV \, d\mu \, dt / \frac{1}{2} (A_{i+1} - A_i) \, \Delta_{s+1/2} \,, \qquad (2.10)$$

together with the corresponding quantities I_{i+1} , I_{m+1} . We also need the cell average

$$\hat{I} = \int_{\Omega} I(\mathbf{r}, \, \mu, \, t) \, dV \, d\mu \, dt / V_{i+1/2} W_{m+1/2} \mathcal{A}_{s+1/2} \,. \tag{2.11}$$

2.2. The Difference Operator

Starting with (1.1) we write L[I] in divergence form

$$L[I] = \frac{1}{c} \frac{\partial I}{\partial t} + \frac{1}{r^2} \frac{\partial}{\partial r} (r^2 \mu I) + \frac{1}{r} \frac{\partial}{\partial \mu} [(1 - \mu^2) I].$$
(2.12)

We introduce the operator Λ_{Ω} by defining

$$\Lambda_{\Omega}[I] = \int_{\Omega} L[I] \, dV \, d\mu \, dt / V_{i+1/2} W_{m+1/2} \Delta_{s+1/2} \,. \tag{2.13}$$

In terms of our definitions (2.1)-(2.11) we find

$$A_{\Omega}[I] = \frac{I_{s+1} - I_s}{c\mathcal{A}_{s+1/2}} + \mu_{m+1/2} \frac{A_{i+1}I_{i+1} - A_iI_i}{V_{i+1/2}} + \frac{\alpha_{m+1}I_{m+1} - \alpha_mI_m}{V_{i+1/2}W_{m+1/2}}, \quad (2.14)$$

and with this notation, we may also write down the difference equivalent of (1.1), namely,

$$\begin{split} \Lambda_{\Omega}[I] &= \int_{\Omega} \sigma(r, t) [B(r, t) - I(r, \mu, t)] \, dV \, d\mu \, dt / V_{i+1/2} W_{m+1/2} \mathcal{L}_{s+1/2} \\ &= \sigma_{\Omega}[B_{\Omega} - I], \end{split}$$
(2.15)

say, where σ_{Ω} and B_{Ω} are average values in Ω .

In deriving (2.15) we have made no approximation which implies any particular form for the dependence of $I(r, \mu, t)$ on the independent variables. All that we have done is to replace a volume integration over Ω by an integration over its surface using Green's theorem. However, if we wish to calculate the intensity field cell by cell, we know not more than three of the six surface intensities, and Eq. (2.15) then gives only one relation for determining the remaining unknowns. We have to supply auxiliary equations in which we have assumed a specific functional form for $I(r, \mu, t)$ on its arguments. The equations that we shall employ are

$$X_{s+1/2}I_{s+1} + [1 - X_{s+1/2}]I_s = Y_{s+1/2}\overline{I},$$

$$X_{m+1/2}I_{m+1} + [1 - X_{m+1/2}]I_m = Y_{m+1/2}\overline{I},$$

$$X_{i+1/2}I_i + [1 - X_{i+1/2}]I_{i+1} = Y_{i+1/2}\overline{I}$$
(2.16)

and either

when $\mu_{m+1/2} < 0$, or

$$X_{i+1/2}I_{i+1} + [1 - X_{i+1/2}]I_i = Y_{i+1/2}I$$

when $\mu_{m+1/2} > 0$. It would be possible to set up relations of greater generality, but we consider that these are sufficient for our purposes. For each X and Y we shall assume $\frac{1}{2} \leq X \leq 1$, and $|1 - Y| \leq y < 1$, but we shall for the present impose no other restrictions.

2.3. Matrix Form of the Difference Equations

We have first of all to order the cells Ω . For this purpose, we need only consider a single time interval $[t_s, t_{s+1}]$, so that it is unnecessary to be explicit about the value of s in ordering. We shall see that we need an additional set of cells of zero weight in order to set up the boundary condition along the direction $\mu = -1$. For these cells we set m = -M - 1. We now order the symbols $\Omega \equiv (i + \frac{1}{2}, m + \frac{1}{2})$ as follows:

$$\begin{array}{cccc} (N+\frac{1}{2},-M-\frac{1}{2}), (N-\frac{1}{2},-M-\frac{1}{2}),..., (\frac{3}{2},-M-\frac{1}{2}), & \mu_{-M-1/2} = \mu_{-M} = -1, \\ (N+\frac{1}{2},-M+\frac{1}{2}), (N-\frac{1}{2},-M+\frac{1}{2}),..., (\frac{3}{2},-M+\frac{1}{2}) \\ & & \ddots \cdots \\ (N+\frac{1}{2},-\frac{1}{2}), (N-\frac{1}{2},-\frac{1}{2}),..., (\frac{3}{2},-\frac{1}{2}) \end{array} \right) \\ \mu_{m+1/2} < 0, \\ (2.17) \\ (\frac{3}{2},\frac{1}{2}), (\frac{5}{2},\frac{1}{2}),..., (N+\frac{1}{2},\frac{1}{2}) \\ & & \ddots \cdots \\ (\frac{3}{2},M-\frac{1}{2}), (\frac{3}{2},M-\frac{1}{2}),..., (N+\frac{1}{2},M-\frac{1}{2}) \end{array} \right) \\ \mu_{m+1/2} > 0.$$

We associate an index k with each cell, where $1 \le k \le (2M + 1) N = K$, and define an "output" intensity vector,

$$[u(k)]^{T} = \begin{cases} [I_{s+1}, I_{i}, I_{m+1}, I] & \mu_{m+1/2} < 0 \\ [I_{s+1}, I_{i+1}, I_{m}, \bar{I}] & \mu_{m+1/2} > 0. \end{cases}$$
(2.18)

The superscript T will be used to denote transposition. A subscript (s + 1/2) can be appended to u(k) in order to indicate the time interval involved.

With this notation, Eq. (2.15) and (2.16) can be combined in the form

$$\begin{aligned} \mathbf{A}_{0}(k) \ u(k)_{s+1/2} &= \mathbf{A}_{1}(k) \ u(k)_{s-1/2} + \mathbf{A}_{2}(k) \ u(k-1)_{s+1/2} \\ &+ \mathbf{A}_{3}(k) \ u(k-l)_{s+1/2} + q(k)_{s+1/2} \end{aligned}$$

where the $A_i(k)$ are square 4×4 matrices, $[q(k)]^T = [0, 0, 0, B(k)]$, and l = N if $m \neq 0$, l = 2i - 1 if m = 0. The $A_i(k)$ are defined as follows:

$$\mathbf{A}_{0}(k) = \begin{bmatrix} X_{s+1/2} & \cdot & -Y_{s+1/2} \\ \cdot & X_{i+1/2} & -Y_{i+1/2} \\ a(k) & b_{+}(k) & c_{+}(k) & 1 \end{bmatrix}, \quad \mathbf{A}_{1}(k) = \begin{bmatrix} -(1 - X_{s+1/2}) & \cdot & \cdot \\ \cdot & \ddots & \cdot \\ a(k) & b_{+}(k) & c_{+}(k) & 1 \end{bmatrix},$$

$$\mathbf{A}_{2}(k) = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & -(1 - X_{i+1/2}) & \cdot \\ \cdot & b_{-}(k) & \cdot & \cdot \end{bmatrix}, \quad \mathbf{A}_{3}(k) = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & -(1 - X_{m+1/2}) & \cdot \\ \cdot & c_{-}(k) & \cdot \end{bmatrix}.$$

$$(2.20)$$

The X and Y coefficients have been defined above in Eq. (2.16),

$$a(k) = 1/c\sigma(k) \Delta_{s+1/2};$$

$$b_{+}(k) = |\mu_{m+1/2}| A_{i}/\sigma(k) V_{i+1/2}, \quad b_{-}(k) = |\mu_{m+1/2}| A_{i+1}/\sigma(k) V_{i+1/2}$$

(2.21)

when m < 0, or

$$b_{+}(k) = |\mu_{m+1/2}| A_{i+1}/\sigma(k) V_{i+1/2}, \quad b_{-}(k) = |\mu_{m+1/2}| A_{i}/\sigma(k) V_{i+1/2}$$

when $m \ge 0$; and

$$c_{+}(k) = \alpha_{m+1}/\sigma(k) V_{i+1/2} W_{m+1/2}, \quad c_{-}(k) = \alpha_{m}/\sigma(k) V_{i+1/2} W_{m+1/2},$$

in the general case. We note the relation $b_{+}(k) + c_{+}(k) = b_{-}(k) + c_{-}(k)$.

The restrictions on the X and Y coefficients ensure that $A_0(k)$ is invertible, so that (2.19) can be written in the form

$$-\mathbf{C}_{3}(k) u(k-l)_{s+1/2} - \mathbf{C}_{2}(k) u(k-1)_{s+1/2} + u(k)_{s+1/2} = \mathbf{C}_{1}(k) u(k)_{s-1/2} + S(k)_{s+1/2},$$
(2.22)

where, for k = 1, 2, ..., K,

$$C_{1}(k) = B_{1}(k) \begin{bmatrix} C_{s+1/2}(k) - D_{s+1/2}(k) & \cdot & \cdot \\ C_{i+1/2}(k) & \cdot & \cdot \\ C_{m+1/2}(k) & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot \end{bmatrix},$$

$$C_{2}(k) = B_{2}(k) \begin{bmatrix} \cdot & C_{s+1/2}(k) & \cdot & \cdot \\ \cdot & C_{i+1/2}(k) - D_{i+1/2}(k) & \cdot & \cdot \\ \cdot & C_{m+1/2}(k) & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot \end{bmatrix},$$

$$C_{3}(k) = B_{3}(k) \begin{bmatrix} \cdot & C_{s+1/2}(k) & \cdot \\ \cdot & C_{m+1/2}(k) & \cdot \\ \cdot & C_{m+1/2}(k) & \cdot \\ \cdot & C_{m+1/2}(k) & - D_{m+1/2}(k) & \cdot \\ \cdot & \cdot & 1 & \cdot \end{bmatrix},$$

with

$$B_{1}(k) = a(k) d(k)/X_{s+1/2}, B_{2}(k) = \bar{b}(k) d(k)/X_{i+1/2}, B_{3}(k) = \bar{c}(k) d(k)/X_{m+1/2},$$

$$C_{s+1/2}(k) = Y_{s+1/2}/X_{s+1/2}, C_{i+1/2}(k) = Y_{i+1/2}/X_{i+1/2}, C_{m+1/2}(k) = Y_{m+1/2}/X_{m+1/2},$$

$$D_{s+1/2}(k) = (1 - X_{s+1/2})/a(k) d(k), D_{i+1/2}(k) = (1 - X_{i+1/2})/\bar{b}(k) d(k),$$

$$D_{m+1/2}(k) = (1 - X_{m+1/2})/\bar{c}(k) d(k),$$

$$\bar{b}(k) = b_{+}(k)(1 - X_{i+1/2}) + b_{-}(k) X_{i+1/2},$$

$$\bar{c}(k) = c_{+}(k)(1 - X_{m+1/2}) + c_{-}(k) X_{m+1/2},$$

and

$$d(k) = \{1 + a(k) C_{s+1/2}(k) + b_{+}(k) C_{i+1/2}(k) + c_{+}(k) C_{m+1/2}(k)\}^{-1}.$$

It is convenient to incorporate flux incident on the boundaries into the source vector S(k), and to write

$$[S(k)]^{\mathsf{T}} = d(k) B(k)[C_{s+1/2}(k), C_{i+1/2}(k), C_{m+1/2}(k), 1] + [S_b(k)]^{\mathsf{T}}.$$

The vector $S_b(k)$ vanishes unless $k \equiv 1 \mod N$, and then we find

$$[S_b(k)]^{\mathrm{T}} = B_2(k)[C_{s+1/2}(k), C_{i+1/2}(k) - D_{i+1/2}(k), C_{m+1/2}(k), 1] u_b(k), k \equiv 1 \mod N$$
(2.24)

where

$$u_b(k) = I_{N+1,m+1/2}, \quad m < 0$$

= $I_{1,m+1/2}, \quad m \ge 0.$

The value of $I_{1,m+1/2}$ depends on what we suppose happens in the region $0 \le r \le r_1$. For example if $r_1 = 0$, a suitable condition is

$$I_{1,m+1/2} = I_{1,-m-1/2}, \quad m = 0, 1,..., M-1.$$

When $k \leq N(m = -M - 1)$, the first term in (2.22) disappears. This can be easily established by applying a limiting process in which $\mu_m \rightarrow \mu_{m-1} = -1$ to the basic equations (2.15) and (2.16). In addition to the relation $C_3(k) = 0$, we find that

$$C_{m+1/2}(k) = 1, \quad c_{+}(k) = -b_{+}(k) + b_{-}(k), \quad m = -M - 1, \ 1 \leq k \leq N.$$

(2.25)

This completes the specification of the difference equations to be solved.

The structure of our difference system becomes slightly more transparent if we partition the discrete space on which our solutions are defined into subspaces labeled by an index j = 1,..., 2M + 1. We associate the vector u^j with the *j*th subspace, where

$$[u^{j}]^{\mathrm{T}} = [u(N_{j} - N + 1)^{\mathrm{T}}, ..., u(N_{j})^{\mathrm{T}}].$$
 (2.26)

Equations (2.22) may now be written in the form

$$\mathbf{E}u_{s+1/2} = \mathbf{F}u_{s-1/2} + S_{s+1/2}, \qquad (2.27)$$

where we partition $u_{s+1/2}$ into vectors $u_{s+1/2}^{j}$, j = 1, ..., 2M + 1 as defined in (2.26).

581/2/4-4

The matrices E and F are respectively lower triangular and diagonal, and may be partitioned in a manner corresponding to the partition of u so that

$$\mathbf{F}_{j,j} = \text{diag}[\mathbf{C}_{1}(N_{j} - N + 1), \dots, \mathbf{C}_{1}(N_{j})], \qquad (2.28)$$
$$\mathbf{E}_{j,j} = \begin{bmatrix} I & & & \\ -\mathbf{C}_{2}(N_{j} - N + 2) & I & & \\ & -\mathbf{C}_{2}(N_{j} - N + 3) & I & \\ & & \ddots & \ddots & \\ & & & -\mathbf{C}_{2}(N_{j}) & I \end{bmatrix}, \qquad (2.29)$$

$$\mathbf{E}_{j,j-1} = \text{diag}[-\mathbf{C}_{3}(N_{j} - N + 1),..., -\mathbf{C}_{3}(N_{j})], j \neq M + 2$$

= $[\mathbf{E}_{j,j-1}(k, l)], j = M + 2,$ (2.30)

where

$$\mathbf{E}_{M+2,M+1}(k,l) = -\mathbf{C}_{3}(N_{j}-N+k)\,\delta(k,N-l+1)$$
(2.31)

is the matrix entry in the kth row and lth column. All other blocks of E and F are null. The special form of $E_{M+2,M+1}$ arises from the reverse ordering of radial points when μ changes sign.

The form of these equations deserves comment. Because E is lower triangular with nonnull diagonal elements, the solution of (2.27) is a trivial matter. All that is necessary is to solve the equations recursively in order of increasing j; for this reason, the procedure is known as the "Method of Directional Evaluation" [3]. The cell ordering (2.17) was chosen in order to simplify this progression. In practice it is only necessary to write Eq. (2.22) in the form

$$u(k)_{s+1/2} = C_1(k) u(k)_{s-1/2} + C_2(k) u(k-1)_{s+1/2} + C_3(k) u(k-l)_{s+1/2} + S(k)_{s+1/2}$$
(2.32)

to see that this can be done. However, the structure of Eqs. (2.27) is more convenient for studying properties of the difference operator, and we shall make use of this form in the next section.

One final remark: the method of directional evaluation can also be used for scattering problems, in which case iterative techniques are needed [3]. Essentially one adds an estimated scattering source to $S_{s+1/2}$ and solves the equations as before. The new solution can now be used to improve $S_{s+1/2}$, and the iteration is repeated until convergence is obtained. We shall not discuss this method further in the present paper.

3. MONOTONICITY AND STABILITY

In this section we study the properties of the solution of equations (2.22) or (2.27). Such a solution should be an approximation to a genuine solution of the original problem. We must therefore discuss the convergence of (2.27) and the consistency and stability of our initial-boundary-value problem. In addition, a genuine solution must, from physical arguments, be nonnegative, and so we shall need to consider the conditions under which we can be assured that every element of the discrete solution $u_{s+1/2}$ is non-negative. We shall write $u \ge 0$ to denote that all elements of the vector u are nonnegative. Similarly, we shall write $A \ge 0$ to show that all elements a_{ij} of the matrix A are nonnegative. Thus it is easy to see from the definitions of (2.20) and (2.22) that, for all k, $B_i(k) > 0$, i = 1, 2, 3; $C_{s+1/2}(k) > 0$, $C_{i+1/2}(k) > 0$, $C_{m+1/2}(k) > 0$; $D_{s+1/2}(k) \ge 0$, $D_{i+1/2}(k) \ge 0$, $D_{m+1/2}(k) \ge 0$.

For precision, we shall state our results as theorems.

THEOREM 1. Let **E** be the matrix of (2.27) and let v be a nonnegative vector. Then the equation $\mathbf{E}u = v$ possesses a unique solution which will be nonnegative if, for all cells, $C_{i+1/2}(k) \ge D_{i+1/2}(k)$, $C_{m+1/2}(k) \ge D_{m+1/2}(k)$, k = 1, 2, ..., K.

Proof. The simplest procedure is to exhibit the solution of Eu = v. We must, in the notation of (2.22) and (2.32), to construct u, beginning with u(1) = v(1), and then

$$u(k) = C_2(k) u(k-1) + C_3(k) u(k-1) + v(k)$$
(3.1)

for k = 2, 3, ..., K. Existence and uniqueness of solution follow immediately. For nonnegativity we see that the conditions $C_{i+1/2}(k) \ge D_{i+1/2}(k)$, $C_{m+1/2}(k) \ge D_{m+1/2}(k)$ imply that $C_2(k) \ge 0$, $C_3(k) \ge 0$ for all k. Since, by hypothesis, $v \ge 0$, we see that $u \ge 0$ since it is constructed from a sum of nonnegative terms.

COROLLARY. A sufficient condition for $\mathbf{E}^{-1} \ge 0$ is that $C_{i+1/2}(k) \ge D_{i+1/2}(k)$, $C_{m+1/2}(k) \ge D_{m+1/2}(k)$ for all cells, $\mathbf{k} = 1, 2, ..., K$.

This corollary is merely a restatement of the conclusions of Theorem 1. Notice that the condition is only sufficient to ensure $E^{-1} \ge 0$. Necessary conditions are much harder to find.

From (2.27) and Theorem 1 we see that $u_{s+1/2}$ is defined uniquely by

$$u_{s+1/2} = \mathbf{E}^{-1}(\mathbf{F} u_{s-1/2} + S_{s+1/2}) = \mathbf{E}^{-1} v_{s+1/2}$$
(3.2)

and that if the conditions of Theorem 1 are satisfied, $u_{s+1/2} \ge 0$ if $v_{s+1/2} = \mathbf{F} u_{s-1/2} + S_{s+1/2} \ge 0$. We have already seen that $S_{s+1/2} \ge 0$, so that if

 $u_{s-1/2} \ge 0$ it is sufficient that $\mathbf{F} \ge 0$, although once again this is not a necessary condition. This leads us to formulate the following.

THEOREM 2. Equations (2.27) have a unique solution $u_{s-1/2}$. A sufficient condition for this to be nonnegative is that $C_{i+1/2}(k) \ge D_{i+1/2}(k)$, $C_{m+1/2}(k) \ge D_{m+1/2}(k)$, $C_{s+1/2}(k) \ge D_{s+1/2}(k)$ for all cells, k = 1, 2, ..., K.

Next we must examine convergence to the solution of the continuous problem. We shall show in Section 4 that our difference operator is consistent in the following sense, namely that if M and $p = \frac{\Delta r}{c\Delta}$ be kept fixed, there exists a positive constant K such that as $c\Delta \to 0$,

$$|\Lambda_{\Omega}[I] - L_{\mathcal{M}}[I]| < K(c\Delta), \tag{3.3}$$

where $L_{M}[I]$ is the discrete ordinate differential expression

$$L_{M}[I] = \left[\frac{1}{c}\frac{\dot{o}}{\partial t} + \mu_{m+1/2}\frac{\partial}{\partial r}\right]I(r_{i+1/2}, \mu_{m+1/2}, t_{s+1/2}) \\ + \frac{2}{r_{i+1/2}}\left[\frac{\beta_{m+1}I(r_{i+1/2}, \mu_{m+1}, t_{s+1/2}) - \beta_{m}I(r_{i+1/2}, \mu_{m}, t_{s+1/2})}{W_{m+1/2}} + \mu_{m+1/2}I(r_{i+1/2}\mu_{m+1/2}t_{s+1/2})\right]$$
(3.4)

If the angular mesh is now refined suitably, we can choose a constant K', such that as $M \rightarrow \infty$

$$|L_{M}[I] - L[I]| < K'/M$$
 (3.5)

Now let $\mathbf{G} = \mathbf{E}^{-1}\mathbf{F}$. For a fixed step-length $c\Delta$, and fixed p, it is clear that $||u_{S+1/2}||$ remains bounded as $S \to \infty$, $(\Delta) \cdot S = T$ being kept fixed, if and only if $||\mathbf{G}^n||$ is uniformly bounded for all n > 0, with respect to some suitable choice of vector and matrix norms, that is to say if the approximation is stable. Since **G** is lower triangular, we may use the following result [4].

THEOREM 3. Let $\mathbf{G} \equiv \mathbf{G}(c\Delta, p)$ have eigenvalues γ_k , k = 1, 2, ..., K. A necessary and sufficient condition for \mathbf{G} to be stable is that $|\gamma_k| \leq \gamma < 1$ for all $k \neq k_0$ and $|\gamma_{k_0}| \leq 1 + O(c\Delta)$ for some single value of k_0 .

Thus we have only to examine the eigenvalues of G, which are easily written down by inspection. Some eigenvalues are zero, and the remainder are given by

$$\gamma_k = [B_1(k)(C_{s+1/2}(k) - D_{s+1/2}(k))], k = 1, 2, ..., K.$$
(3.6)

We are now in a position to apply our results to the sets of auxiliary equations in common use. The nomenclature for auxiliary equations is due to Carlson [3].

THEOREM 4. The Diamond Difference scheme defined by choosing auxiliary equations (2.16) with coefficients

$$Y_{s+1/2} = Y_{i+1/2} = Y_{m+1/2} = 1,$$
 $X_{s+1/2} = X_{i+1/2} = X_{m+1/2} = \frac{1}{2}$ (3.7)

is unconditionally stable. Nonnegativity of the solution cannot be guaranteed for finite mesh sizes.

Proof. From the definitions (2.22), we have

$$d(k) = l(k)/[l(k) + 2]$$
(3.8)

where

$$[l(k)]^{-1} = a(k) + b_{+}(k) + c_{+}(k) = a(k) + b_{-}(k) + c_{-}(k) > 0$$

Then

$$\gamma_k = 4a(k) d(k) - 1 = \frac{4a(k) l(k)}{l(k) + 2} - 1.$$

Since 0 < a(k) l(k) < 1 and l(k) > 0, we see that $-1 < \gamma_k < 1$, so that we have unconditional stability.

To prove that the conditions of Theorems 1 and 2 must be violated, it is sufficient to establish a contradiction. Suppose that the conditions all hold. Written out, we see that they imply

$$a(k) \ge 1/4d(k), \quad \overline{b}(k) \ge 1/4d(k), \quad \overline{c}(k) \ge 1/4d(k).$$

Hence

$$\frac{1}{l(k)} = a(k) + \bar{b}(k) + \bar{c}(k) \ge \frac{3}{4 d(k)} = \frac{3(l(k) + 2)}{4l(k)}$$

by (3.8), which implies $l(k) \leq -2/3$. Since l(k) is positive, at least one of the inequalities of Theorem 2 must be violated, and nonnegativity cannot be guaranteed.

We shall see later that this imposes a practical restriction on the step lengths.

THEOREM 5. The Step Difference scheme defined by choosing Auxiliary equations (2.16) with coefficients

$$Y_{s+1/2} = Y_{i+1/2} = Y_{m+1/2} = X_{s+1/2} = X_{i+1/2} = X_{m+1/2} = 1$$
(3.9)

is unconditionally stable, and its solutions can be guaranteed nonnegative.

Proof. From the definitions (2.22), we have

$$d(k) = l(k)/[l(k) + 1]$$

with l(k) defined as above, so that

$$\gamma_k = a(k) \, d(k) = \frac{a(k) \, l(k)}{l(k) + 1} < \frac{1}{l(k) + 1} < 1$$

demonstrating stability. Nonnegativity is an immediate consequence of the observations that $D_{s+1/2}(k) = D_{i+1/2}(k) = D_{m+1/2}(k) = 0$.

4. DISCRETIZATION ERROR AND CONSISTENCY

We have already seen that the difference operator $\Lambda_{\Omega}[I]$ defined in (2.13) and (2.14) gives an exact relation representing the net radiation balance for a cell and it is straightforward to verify that the difference equations so defined are conservative in a global sense. However, we have been forced to introduce auxiliary equations to make the problem soluble, and these embody somewhat arbitrary approximations to describe the functional form of $I(r, \mu, t)$ within each cell. We have now to examine the discretization error arising from these approximations.

We first examine what happens if we take the cell dimensions $\Delta r = r_{i+1} - r_i$, $\Delta = t_{s+1} - t_s$ sufficiently small, keeping a fixed angular mesh. Suppose we associate the quantities \bar{I} , I_i ,..., with specific points inside and on the boundary of a cell so that

$$\begin{split} \bar{I} &= I(r_{i+1/2}, \mu_{m+1/2}, t_{s+1/2}), \\ I_i &= I(r_i, \mu_{m+1/2}, t_{s+1/2}), \end{split} \tag{4.1}$$

Suppose that

$$t_{s+1/2} = \frac{1}{2}(t_{s+1} + t_s)$$

and

$$\begin{aligned} r_{i+1/2} &= \theta_{i+1/2} r_{i+1} + (1 - \theta_{i+1/2}) r_i, \quad \mu_{m+1/2} < 0 \\ &= (1 - \theta_{i+1/2}) r_{i+1} + \theta_{i+1/2} r_i, \quad \mu_{m+1/2} > 0 \end{aligned}$$
(4.2)

If we expand I_i ,..., about the point $(r_{i+1/2}, \mu_{m+1/2}, t_{s+1/2})$ using Taylor's theorem, and insert in (2.14), we find

$$\begin{split} \Lambda_{\Omega}[I] &= \left\{ \frac{1}{c} \frac{\partial}{\partial t} + \mu_{m+1/2} \left[\frac{A_{i+1/2}(r_{i+1} - r_i)}{V_{i+1/2}} \right] \frac{\partial}{\partial r} \right\} \bar{I} \\ &+ \frac{A_{i+1} - A_i}{V_{i+1/2}} \left\{ \frac{\beta_{m+1}I_{m+1} - \beta_m I_m}{W_{m+1/2}} + \mu_{m+1/2} \bar{I} \right\} + O[(c\Delta)^2, (\Delta r)^2], \end{split}$$
(4.3)

where

$$\begin{aligned} A_{i+1/2} &= \theta_{i+1/2} A_{i+1} + (1 - \theta_{i+1/2}) A_i, \qquad \mu_{m+1/2} < 0 \\ &= (1 - \theta_{i+1/2}) A_{i+1} + \theta_{i+1/2} A_i, \qquad \mu_{m+1/2} > 0. \end{aligned}$$
(4.4)

In general,

$$\frac{A_{i+1/2}(r_{i+1}-r_i)}{V_{i+1/2}} = 1 + O(\Delta r), \qquad (4.5)$$

but the special choice

$$\theta_{i+1/2} = \frac{1}{2} + \frac{1}{6} (\operatorname{sgn} \mu_{m+1/2}) \frac{r_{i+1} - r_i}{r_{i+1} + r_i}$$
(4.6)

ensures that

$$\frac{A_{i+1/2}(r_{i+1}-r_i)}{V_{i+1/2}} = 1$$

exactly. Also,

$$\frac{A_{i+1} - A_i}{V_{i+1/2}} = \frac{2}{r_{i+1/2}} [1 + O(\Delta r)], \tag{4.7}$$

in general, with exact equality

$$\frac{A_{i+1} - A_i}{V_{i+1/2}} = \frac{2}{r_{i+1/2}}$$

when we make the special choice (4.6) again. Thus, provided \overline{I} has bounded first derivatives, we obtain inequality (3.3) in a straightforward manner. Inequality (3.5) follows by a similar argument in which we can make a choice

$$\theta_{m+1/2} = [\beta_{m+1/2} - \beta_{m+1}] / W_{m+1/2} \mu_{m+1/2}$$
(4.8)

to reduce the order of the truncation error to $O(W^2)$ if desired.

In practice we want to use finite-size cells, and it is desirable to understand what happens in this circumstance. It is helpful now to examine the analytic solution of (1.1) by the method of characteristics. This requires us to solve

$$dI/ds = \sigma[B - I] \tag{4.9}$$

along the ray-path defined by the equations

$$\frac{c \, dt}{ds} = 1, \qquad \frac{dr}{ds} = \mu, \qquad \frac{d\mu}{ds} = \frac{1 - \mu^2}{r}.$$
 (4.10)

Let $\tau(s)$ be the optical path length, defined by

$$\tau(s)=\int_0^s\sigma(s')\,ds',$$

where for brevity we have written $\sigma(s')$ for $\sigma(r(s'), t(s'))$ which is a prescribed function of its arguments. Then (4.9) has the solution

$$I(s) = I(0) e^{-\tau} + \int_0^{\tau} B(s') e^{-(\tau - \tau')} d\tau', \qquad (4.11)$$

where $\tau = \tau(s), \tau' = \tau(s')$.

For brevity, write $I_{\alpha} = I(\alpha s)$, $0 \leq \alpha \leq 1$, $\overline{I} = I_{\theta}$ for some fixed θ , $0 < \theta < 1$. We also write in the same notation

$$B_{\alpha} = \overline{B}[1 + \beta s(\alpha - \theta) + O(s^2)], 0 \leq \alpha \leq 1,$$

where max $|\beta s(\alpha - \theta)| < 1$, and similarly for $\sigma(s)$. Substituting in (4.11), we have

$$\tau = \bar{\sigma}s + O(s^2)$$

and

$$I_{\alpha} = I_0 e^{-\alpha \tau} + \overline{B}(1 - e^{-\alpha \tau}) \psi(\alpha, \theta), \qquad (4.12)$$

where

$$\psi(\alpha, \theta) = 1 + \beta s \left(\frac{\alpha}{1 - e^{-\alpha \tau}} - \theta - \frac{1}{\tau} \right) + O(s^2).$$

Now we find a relation between I_0 , $I_{\theta} = \overline{I}$, and I_1 by putting $\alpha = \theta$, 1 in turn in (4.12) and eliminating \overline{B} . We find

$$XI_1 + (1 - X)I_0 = Y\overline{I},$$
 (4.13)

where

$$\begin{split} X &= \frac{\psi(\theta,\,\theta)}{\psi(1,\,\theta)} \, Y \frac{1 - e^{-\theta \tau}}{1 - e^{-\tau}} \,, \\ Y &= \frac{\psi(1,\,\theta)}{e^{-\theta \tau} \psi(1,\,\theta) + (1 - e^{-\theta \tau}) \, \psi(\theta,\,\theta)} \,. \end{split}$$

In the particular case in which $\beta \to 0$, we see that $\psi(\alpha, \theta) \to 1$, and

$$X = \frac{1 - e^{-\theta_{\tau}}}{1 - e^{-\tau}}, \quad Y = 1 \qquad (\beta = 0). \tag{4.14}$$

Suppose we now apply (4.13) to a segment of a ray traversing a cell, and for simplicity put $\beta = 0$. Then from (4.14), for small τ

$$X(\tau) = \theta + \frac{1}{2}\theta(1-\theta)\tau + O(\tau^2), \tau \ll 1$$

so that

$$\theta I_1 + (1 - \theta) I_0 + O(\tau) = \tilde{I}, \ \tau \ll 1.$$
 (4.15)

At the other extreme, $\tau \gg 1$,

$$X(\tau) = 1 - e^{-\theta\tau} + O(e^{-\tau})$$

and so

$$I_1 + O(e^{-\theta \tau}) = \bar{I}, \ \tau \gg 1.$$
 (4.16)

More generally, if we retain terms of order βs (small) we find from (4.12)

$$I_{\alpha} = \bar{B}(1 - \beta/\bar{\sigma}) + O(s)$$

= $\bar{B} - \frac{1}{\bar{\sigma}} \frac{d\bar{B}}{ds}\Big|_{\alpha=\theta} + O(s),$ (4.17)

which we should expect to find as the limit for diffusion theory.

These results are suggestive for our choice of auxiliary equations (2.16). The standard choices, the Diamond and Step models were discussed in the last section and their properties are set out in Theorems 4 and 5. However, it is clear from our discussion above, that the choices $X_{i+1/2} = \theta_{i+1/2}$, etc., $Y_{i+1/2} = 1$ with $\theta_{i+1/2}$ defined by (4.6), will give a smaller truncation error in spherical geometry, particularly near the center of the sphere. But when the cell is optically thick, we can not expect this choice to give a satisfactory answer. For, if in the proof of Theorem 4 we consider what happens for large $\sigma(k)$, we see that

$$d(k) \sim 1 + O[\sigma(k)^{-1}]$$

and

$$\gamma_k \sim -1 + \frac{4}{c\sigma(k)\,\Delta}$$

so that the solution can be expected to oscillate in sign. This will remain true when $X > \frac{1}{2}$ except if we choose the Step model weights. However, if we make this choice, we increase the truncation error.

As a compromise, we may try using auxiliary equations of the type of (4.13) with weights given by (4.14) for a suitable choice of τ . The lines joining opposite faces of a cell are not, in general, characteristics. Nevertheless, such equations will interpolate between the two limits appropriate to small and large optical thickness, and it may be possible to do this in such a way that the solution will remain smooth and improve on the step model. We examine this possibility in a heuristic manner in connection with the problem of the next section,

GRANT

5. A TIME-DEPENDENT TEST PROBLEM

To illustrate our work we have chosen a simple problem whose solution can be written down immediately in closed form. This problem is the calculation of the radiation field in a uniform sphere of radius R with a source function B(t) which is independent of r. For the choice

$$\begin{array}{ll}
B(t) = B_0, & t < 0 \\
= B_1, & t > 0,
\end{array}$$
(5.1)

the solution may be written

$$I(r, \mu, t) = B_0(1 - e^{-\sigma s}) \qquad t \leq 0$$

= $B_1(1 - e^{-\sigma ct}) + B_0(e^{-\sigma ct} - e^{-\sigma s}) \qquad 0 \leq t \leq s/c \qquad (5.2)$
= $B_1(1 - e^{-\sigma s}) \qquad t \geq s/c,$

where

$$s = r\mu + [R^2 - r^2(1 - \mu^2)]^{1/2}.$$
 (5.3)

This provides a stringent test of the numerical method when $B_0 \gg B_1$, and we have chosen the values $B_0 = 100$, $B_1 = 1$. Length and time units are chosen so that $\sigma = c = 1$.

We have computed numerical solutions with a variety of choices of mesh with R = 4. The choice of angular mesh is not critical for the determination of the mean intensity J(r, t) and net flux F(r, t), and most of our calculations were made with the DP_2 Gaussian Set (see [3], Table III). Greater interest is attached to the variation in the r and t meshes and the effect of different choices of auxiliary equation.

Our results bear out in general terms the theory of the preceeding sections. In a series of runs with fixed *r*-mesh

$$r = 0, 1, 2, 3, 3.5, 3.75, 3.875, 3.9375, 4.0$$

chosen with small increments near the outer boundary so as to give good definition in that part of the region in which the solution changes most rapidly, we took successively a time step $\Delta = 5$, 2.5, 1.25, .625, .3125. Negative intensities appear in the solution at least once for all cases with the Diamond auxiliary equations. Use of the interpolating auxiliary equations (4.13) and (4.14) with $\tau = c\Delta$, $\theta = \frac{1}{2}$, for the *t*-direction, and $\tau = \Delta r$, with θ given by (4.6) damped out the occurrence of negative intensities, but did not remove them completely. Of course, negative intensities cannot occur with the step auxiliary equations.

The emergent net flux, defined by

$$F(R, t) = 2\pi \sum W_{m+1/2} \mu_{m+1/2} I_{N+1,m+1/2}$$

is a convenient quantity to show the oscillation in time of the numerical solutions. The results for the Diamond scheme are displayed in Fig. 1, together with the



FIG. 1. Net emergent flux from a sphere as a function of time. Solutions of the diamond scheme are plotted for time steps $\Delta = 5$, 1.25, and .3125,

analytic solution. Dashed lines have been used to connect computed values plotted at the mid-points of time steps. Figure 2 shows the corresponding histogram for the Step scheme. Here, we have plotted the numerical solution as a step function. The slower convergence of this scheme as Δ decreases, arising from the increased truncation error compared with the Diamond scheme, is obvious.



FIG. 2. Net emergent flux from a sphere as a function of time. Solutions of the step scheme are plotted for time steps $\Delta = 5$, 1.25, .3125.

Figure 3 shows that the interpolating auxiliary equations represent some compromise between the Step and Diamond schemes, and provide a damping of the oscillations without too much increase in the truncation error.

The distribution of net flux at cell boundaries in the interior of the system discloses other features of the problem. Figure 4 shows what happens to the



FIG. 3. Net emergent flux from a sphere as a function of time. Solutions of the interpolating scheme are shown for $\Delta = 1.25$, 5. The solution for $\Delta = .3125$ is virtually indistinguishable from the analytic solution.

solution of the Diamond scheme for a time step $\Delta = 2.5$. After one step there is qualitative agreement with the analytic solution at t = 1.25. On the next time step (2.5 to 5.0) we see an unphysical oscillation appearing, with a negative flux in the outside meshes following the behavior shown in Fig. 1. The corresponding profiles for the step and interpolating schemes appear in Fig. 5. These are qualitatively more reasonable, but because of the coarseness of zoning both radial and timewise, the accuracy is not very high. The fact that the solution is as good as it is in these circumstances is not unsatisfying. We have performed some other calculations, but they add little more to our understanding of the problem, and we shall not discuss them here.

GRANT

6. DISCUSSION

We have shown that all the schemes discussed converge, in the limit of small mesh sizes, to the original differential equations. For finite mesh sizes, the usual choices of auxiliary equations lead to stable difference schemes, but the operators involved can not always be guaranteed monotone. When the solution is dominated by an imposed source this may be of no consequence, but trouble can arise particularly for decaying radiation fields as in the example studied in the last section. These troubles can be traced to cell dimensions which are too large in space, in time, or in both. The step equations are alone in providing a monotone difference operator, but have a larger truncation error than other schemes. The interpolating



FIG. 4. Distribution of net flux with radius for the diamond scheme with time-step $\Delta = 6.5$ averaged over steps.



FIG. 5. Distribution of net flux with radius for the step and interpolating schemes with time step $\Delta = 2.5$ averaged over a step.

scheme has a smaller truncation error, but approaches the step scheme in its ability to keep the solution nonnegative.

Experience suggests that this is at best a somewhat unsatisfactory way of dealing with these problems. The reason is that the adopted equations are not based on a coherent underlying theory. Subsequent work [2], so far only relating to the case of a time-independent emitting, scattering, and absorbing slab, has suggested that invariance arguments hold the key to a more satisfying solution of the difficulties we have discussed. This is because the S_n difference equations are conservative, and so can be rewritten in a form consistent with well-known principles of invariance [5], that is to say, with conservation in a global sense. The matrix coefficients can then be interpreted as approximate operators for reflection and transmission by finite cells. Since we have a calculus for rapidly computing reflection and transmission operators for multiple layers [6, 7], we can construct better approximations for the operators for finite cells from the S_n expressions for thinner cells. Explicit methods of solving the transformed equations without having recourse to iteration in scattering problems have been found to work well in practice [2].

ACKNOWLEDGMENT

I am indebted to Dr. G. E. Hunt for many discussions in the course of this work.

REFERENCES

- 1. K. LATHROP AND B. G. CARLSON, J. Computational Phys. 2, 173-197 (1967).
- 2. I. P. GRANT AND G. E. HUNT (to be published).
- 3. B. G. CARLSON, *in* "Methods of Computational Physics," Vol. 1 (B. Alder, S. Fernbach, and M. Rotenberg, Eds.), pp. 1–42. Academic Press, New York, (1963).
- 4. R. D. RICHTMYER, "Difference Methods in Initial Value Problems," p. 69. Interscience, New York (1957).
- 5. S. CHANDRASEKHAR, "Radiative Transfer", Ch. VII. Oxford University Press, London (1950). 6. R. REDHEFFER, J. Math. Phys. 41, 1-41 (1962).
- 7. R. W. PREISENDORFER, "Radiative Transfer on Discrete Spaces". Pergamon Press (1965).