APPLICATION OF OSCILLATION MATRICES TO DIFFUSION-CONVECTION EQUATIONS

By Harvey S. Price, Richard S. Varga and Joseph E. Warren

Reprinted from Journal of Mathematics and Physics
Vol. 45, No. 3, September, 1966
Printed in U.S.A.

Reprinted from JOURNAL OF MATHEMATICS AND PHYSICS
Vol. 45, No. 3, September, 1966
Printed in U.S.A.

APPLICATION OF OSCILLATION MATRICES TO DIFFUSION-CONVECTION EQUATIONS

BY HARVEY S. PRICE, RICHARD S. VARGA AND JOSEPH E. WARREN

1. Introduction. Consider the transfer of heat or mass [7, 8] in a one-dimensional system which contains a homogeneous, incompressible flowing fluid. If the term which describes transport due to fluid motion (i.e., the convection term) is comparable in magnitude to the diffusion term, then the behavior of the system satisfies the following parabolic partial differential equation:

$$\frac{\partial c(x, t)}{\partial t} = \frac{\partial^2 c(x, t)}{\partial x^2} - \lambda \frac{\partial c(x, t)}{\partial x}, \qquad \lambda > 0, \tag{1}$$

where the diffusivity is taken to be unity and $c(x, t)$ represents the normalized concentration of heat or mass.

The following boundary conditions frequently apply:

$$\begin{aligned} &c(x, 0) = 0; 0 < x < l, \\ &c(0, t) = 1; t > 0, \qquad (\partial c/\partial x)(l, t) = 0; t > 0. \end{aligned} \tag{2}$$

With $l = \infty$ and the third condition of (2) replaced by $c(x, t) \to 0$ as $x \to \infty$, a routine use of the Laplace transform method shows that the concentration $c(x, t)$ at a fixed point in space is *monotonically increasing*:

$$c(x, t + \Delta t) > c(x, t), \qquad \Delta t > 0. \tag{3}$$

This condition (3) remains true when l is finite as well. Moreover, the concentration $c(x, t)$ must lie between zero and unity:

$$1 \geqq c(x, t) \geqq 0, \qquad 0 \leqq x \leqq l, t \geqq 0. \tag{4}$$

Our interest in this problem arose from the fact that standard finite difference approximations in space, such as (5), and time (such as the forward difference method of Section 5), yielded approximate concentrations which for fixed x exhibited damped *oscillations* in time about unity, thereby violating (3) and (4). We shall show that such damped oscillations can arise even with infinitely small time increments, i.e., with semi-discrete finite difference approximations, if the spatial mesh is sufficiently coarse. We first prove a necessary and sufficient condition (Theorem 1) for non-oscillation in the semi-discrete approximations to (1)–(2), and a sufficient condition (Theorem 2) for non-oscillation which applies to more general problems. The novelty of these results lies in the application of the theory of oscillatory matrices of Gantmakher and Krein [5, 6]. In the final section, a criterion for non-oscillation of time-discretizations is similarly given.

2. Semi-discrete central finite difference approximations. With a uniform space mesh $h = l/n$, the usual three-point (spatial) central difference approximation to (1) based on Taylor's series is [3, p. 141]:

$$\frac{dc_i(t)}{dt} = \frac{c_{i+1}(t) - 2c_i(t) + c_{i-1}(t)}{h^2} - \lambda\left[\frac{c_{i+1}(t) - c_{i-1}(t)}{2h}\right] + \tau_i, \quad 1 \leqq i \leqq n-1, \tag{5}$$

where $c_i(t) \equiv c(ih, t)$, and τ_i is an error term of order h^2 as $h \to 0$, which depends on higher spatial derivatives of $c(x, t)$. For the n-th mesh point, the boundary condition $c_x(l, t) = 0$ of (2) used in conjunction with the differential equation (1) similarly yields

$$\frac{dc_n(t)}{dt} = \frac{-2c_n(t) + 2c_{n-1}(t)}{h^2} + \tau_n, \tag{6}$$

where the error term τ_n is now of order h as $h \to 0$. Neglecting the error terms τ_i of (5) and (6) gives a system of ordinary differential equations which can be written in the form

$$\frac{d\mathbf{w}}{dt} = -A\mathbf{w}(t) + \mathbf{s}, \; t > 0, \tag{7}$$

where A is a real $n \times n$ matrix, and $\mathbf{w}(t)$ and $\mathbf{s}$ are column vectors with n components given explicitly by

$$A = \frac{1}{h^2}\begin{bmatrix} +2 & -(1-\alpha) & & & \\ -(1+\alpha) & +2 & -(1-\alpha) & & \\ & \ddots & \ddots & \ddots & \\ & & -(1+\alpha) & +2 & -(1-\alpha) \\ & & & -2 & +2 \end{bmatrix};$$

$$\mathbf{w}(t) = \begin{bmatrix} w_1(t) \\ w_2(t) \\ \vdots \\ w_{n-1}(t) \\ w_n(t) \end{bmatrix}; \quad \mathbf{s} = \frac{1}{h^2}\begin{bmatrix} 1+\alpha \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{8}$$

and

$$\alpha \equiv \tfrac{1}{2}\lambda h. \tag{9}$$

We shall call $\mathbf{w}(t)$ the *semi-discrete approximation* of (1)–(2), in that the time variable t has not been discretized.

Our object is now to show that the matrix A has positive real and distinct eigenvalues for certain α. Let $D = \text{diag}(d_1, d_2, \cdots, d_n)$ be an $n \times n$ diagonal matrix having non-zero diagonal entries which alternate in sign, i.e., $d_i = (-1)^i |d_i|$, $1 \leqq i \leqq n$. Then, upon forming $D^{-1}AD$, it is easily seen, that for $0 \leqq \alpha < 1$, we can select the $|d_i|$ as a function of α so that

$$B \equiv D^{-1}AD = \frac{1}{h^2}\begin{bmatrix} 2 & \sqrt{1-\alpha^2} & & & \\ \sqrt{1-\alpha^2} & 2 & \sqrt{1-\alpha^2} & & \\ & \ddots & \ddots & \ddots & \\ & & \sqrt{1-\alpha^2} & 2 & \sqrt{2(1-\alpha)} \\ & & & \sqrt{2(1-\alpha)} & 2 \end{bmatrix}. \quad (10)$$

Several facts about the matrix $D^{-1}AD \equiv B$ can now be easily deduced. First, B is a real symmetric matrix. Next, A is an irreducibly diagonally dominant matrix with positive diagonal entries for $0 \leqq \alpha < 1$; as such, B is then positive definite [10, p. 23]. Moreover, this irreducible diagonal dominance implies that the successive principal minors of B are positive. Further, since the superdiagonal and subdiagonal of B have positive entries for $0 \leqq \alpha < 1$, it follows [5, p. 124] that B is an *oscillation matrix*, i.e., all minors of B, whether principal or not, are non-negative, and some positive power of B has all its minors positive. But, as an oscillation matrix has real *distinct* eigenvalues [5, p. 126], we deduce that the matrix A has real distinct positive eigenvalues for $0 \leqq \alpha < 1$.

We remark that the real distinct eigenvalue character of the matrix A, as proved above, is also essentially given in [4, Chapter X], and could have been established directly from a three-term recurrence relation between the upper left successive principal minors of $(\gamma I - A)$.

For the case $\alpha = 1$, the matrix A of (8) is then a lower bidiagonal matrix with diagonal entries all 2, so that the eigenvalues of A in this case are all 2. For $\alpha > 1$, there similarly exists a positive diagonal matrix such that $D^{-1}AD$ is a real *skew-symmetric* matrix. Thus, all of the eigenvalues γ_j of A are of the form

$$\gamma_j = (+2 + i\sigma_j)/h^2; \quad 1 \leqq j \leqq n, \sigma_j \text{ real}, \quad (11)$$

where it can be verified that

$$\max_j (\sigma_j) > 2\sqrt{2(\alpha - 1)} \cos [\pi/(n+1)]. \quad (12)$$

In other words, for any $\alpha > 1$, $-A$ is a *stable matrix* [1, p. 242; 2, p. 108], i.e., all the eigenvalues of $-A$ have negative real parts. On the other hand, there always exist eigenvalues of A with non-zero imaginary part. By way of contrast, for $0 \leqq \alpha \leqq 1$, all the eigenvalues of A are real. This proves the following:

THEOREM 1. Let A be the $n \times n$ matrix of (8). For any $h > 0$, $-A$ is a stable matrix. Moreover, all the eigenvalues of $-A$ are negative real numbers if and only if $0 < h \leqq 2/\lambda$. If $0 < h < 2/\lambda$, all the (negative real) eigenvalues of $-A$ are distinct.

We can apply the previous ideas to the following more general problem

$$\frac{\partial c(x,t)}{\partial t} = \frac{\partial}{\partial x}\left\{K(x)\frac{\partial c(x,t)}{\partial x}\right\} - \lambda(x)\frac{\partial c(x,t)}{\partial x}; \quad 0 < x < 1, \quad t > 0, \quad (13)$$

where $K(x)$ and $\lambda(x)$ are given continuous positive functions in $0 \leqq x \leqq 1$, with the boundary conditions (2). Using a not necessarily uniform spatial mesh with $x_{i+1} = x_i + h_i$, $h_i > 0$, $0 \leqq i \leqq n - 1$, the spatial derivatives in (13) can be approximated [10, p. 178] by

$$\frac{\partial}{\partial x}\left\{K(x_i)\,\frac{\partial c(x_i, t)}{\partial x}\right\} = \frac{2K_{i+\frac{1}{2}}\,h_i^{-1}[c_{i+1}(t) - c_i(t)] - 2K_{i-\frac{1}{2}}\,h_{i-1}^{-1}[c_i(t) - c_{i-1}(t)]}{h_i + h_{i-1}} + \tau_i^{(1)}, \tag{14}$$

and

$$-\lambda(x_i)\,\frac{\partial c(x_i, t)}{\partial x} = -\lambda_i\left[\frac{c_{i+1}(t) - c_i(t)}{2h_i}\right] - \lambda_i\left[\frac{c_i(t) - c_{i-1}(t)}{2h_{i-1}}\right] + \tau_i^{(2)}, \tag{15}$$

where $K_{i+\frac{1}{2}} \equiv K[\frac{1}{2}(x_i + x_{i+1})]$. In general, the error terms $\tau_i^{(l)}$ in (14) and (15) are of the order $\bar{h}_i = \max(h_i, h_{i-1})$, but when $h_{i-1} = h_i$, these error terms are of order h_i^2. Thus, as before, by neglecting these error terms, we obtain

$$\frac{d\mathbf{w}(t)}{dt} = -A\mathbf{w}(t) + \mathbf{s}; \qquad t > 0, \tag{16}$$

where

$$A = \begin{bmatrix} +D_1 & -U_1 & & & \\ -L_2 & +D_2 & -U_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -U_{n-1} \\ & & & -L_n & +D_n \end{bmatrix} \tag{17}$$

and

$$\begin{cases} L_i = \dfrac{1}{h_{i-1}}\left\{\dfrac{2K_{i-\frac{1}{2}}}{(h_i + h_{i-1})} + \dfrac{\lambda_i}{2}\right\}; & 1 \leqq i \leqq n, \\ U_i = \dfrac{1}{h_i}\left\{\dfrac{2K_{i+\frac{1}{2}}}{(h_i + h_{i-1})} - \dfrac{\lambda_i}{2}\right\}; & 1 \leqq i \leqq n - 1, \\ D_i = L_i + U_i; \quad 1 < i < n - 1 & \text{and} \quad D_n = L_n. \end{cases} \tag{18}$$

If $\lambda_i(h_i + h_{i-1}) < 4K_{i-\frac{1}{2}}$ for all $1 \leqq i \leqq n - 1$, then the quantities L_i and U_i of (18) are all positive real numbers. Since $D_i = L_i + U_i$, then the matrix A of (17) is then irreducibly diagonally dominant with positive diagonal entries. The same device as previously applied in (10) shows that there is a nonsingular diagonal matrix C such that $C^{-1}AC$ is real and symmetric, and thus A has real eigenvalues. But as A is irreducibly diagonally dominant, then A has positive real eigenvalues. To deduce that these eigenvalues are distinct, we again use the fact that the superdiagonal and subdiagonal of A are positive, so that A is an oscillation matrix. This gives us

Theorem 2. If $\lambda_i(h_i + h_{i-1}) \leqq 4K_{i-\frac{1}{2}}$, $1 \leqq i \leqq n - 1$, then all the eigenvalues of the $n \times n$ matrix A of (17) are positive real numbers. If $\lambda_i(h_i + h_{i-1}) < 4K_{i-\frac{1}{2}}$, $1 \leqq i \leqq n - 1$, all the positive real eigenvalues of A are distinct.

If the matrix A has distinct real eigenvalues μ_i with associated eigenvectors $\mathbf{v}_i$, then the solution of either (7) or (16) can be expressed as

$$\mathbf{w}(t) = \sum_{i=1}^{n} d_i(1 - e^{-t\mu_i})\mathbf{v}_i, \qquad t \geqq 0, \tag{19}$$

where

$$A^{-1}\mathbf{s} = \sum_{i=1}^{n} d_i\mathbf{v}_i. \tag{20}$$

Thus, for all sufficiently small mesh spacings h_i, it is clear from Theorems 1 and 2 that the semi-discrete solutions of (19) are non-oscillatory. On the other hand, Theorem 1 shows that there exist non-real eigenvalues μ_i for $h > 2/\lambda$, so the solution $\mathbf{w}(t)$ of (19) exhibits *damped oscillations* for $h > 2/\lambda$, even for a semi-discrete approximation in which the time variable is left continuous.

3. Non-central semi-discrete approximations. We now increase the generality of (1) by assuming that $\lambda = \lambda(x) \geqq 0$ for all x, $0 \leqq x \leqq 1$. With a uniform spatial mesh $h = l/n$, consider the following *non-central* finite difference approximations to (1)–(2), which serve to define the $n \times n$ matrix B:

$$\begin{aligned}
h^2[B\mathbf{c}(t)]_1 &\equiv (2 + \lambda_1 h)c_1(t) - c_2(t) = -h^2\frac{dc_1(t)}{dt} + (1 + \lambda_1 h) + h^2\tau_1; \\
h^2[B\mathbf{c}(t)]_2 &\equiv -(1 + 2\lambda_2 h)c_1(t) + \left(2 + \frac{3}{2}\lambda_2 h\right)c_2(t) - c_3(t) \\
&\qquad = -h^2\frac{dc_2(t)}{dt} - \frac{1}{2}\lambda_2 h + h^2\tau_2; \\
h^2[B\mathbf{c}(t)]_i &\equiv \frac{1}{2}\lambda_i h c_{i-2}(t) - (1 + 2\lambda_i h)c_{i-1}(t) + \left(2 + \frac{3}{2}\lambda_i h\right)c_i(t) \\
&\qquad - c_{i+1}(t) = -h^2\frac{dc_i(t)}{dt} + h^2\tau_i, \qquad 3 \leqq i \leqq n-1; \\
h^2[B\mathbf{c}(t)]_n &\equiv -2c_{n-1}(t) + 2c_n(t) = -h^2\frac{dc_n(t)}{dt} + h^2\tau_n.
\end{aligned} \tag{21}$$

Here, $\lambda_i \equiv \lambda(ih)$, and the error terms τ_i are of order h^2 as $h \to 0$ for $2 \leqq i \leqq n-1$, while τ_1 and τ_n are of order h as $h \to 0$. Neglecting the error terms τ_i in (21) yields

$$-B\mathbf{v}(t) = h^2\frac{d\mathbf{v}(t)}{dt} - \mathbf{g}, \qquad t > 0, \tag{22}$$

where $\mathbf{g}$ is the vector with components defined by

$$g_1 = (1 + \lambda_1 h); \quad g_2 = -\tfrac{1}{2}\lambda_2 h; \quad g_i = 0,\ 3 \leqq i \leqq n.$$

The matrix B so defined thus plays a role analogous to the matrix A in (7).

We now prove

Theorem 3. If $\lambda_i \geqq 0$ for $1 \leqq i \leqq n-1$, then the $n \times n$ matrix B defined by (21) has positive real distinct eigenvalues for *all* $h > 0$.

Proof. Defining $b^+_{i,j} = (-1)^{i+j}b_{i,j}$, $1 \leqq i, j \leqq n$, we first establish that all minors of the $n \times n$ matrix $B^+ = (b^+_{i,j})$ are nonnegative. It is easy to verify that an arbitrary minor of B^+ can be written as a product of elements of the matrix B^+ (i.e., the $b^+_{i,j}$'s) and minors[1]

$$B^+\begin{pmatrix} i_1, i_2, \cdots, i_p \\ k_1, k_2, \cdots, k_p \end{pmatrix}, \qquad \left(1 \leqq \begin{matrix} i_1 < i_2 < \cdots < i_p \\ k_1 < k_2 < \cdots < k_p \end{matrix} \leqq n\right),$$

whose elements satisfy

$$b^+_{i_\nu, k_{\nu+1}} > 0 \quad \text{and} \quad b^+_{i_{\nu+1}, k_\nu} > 0, \qquad 1 \leqq \nu \leqq p-1. \tag{23}$$

Furthermore, one can show [6, p. 345; 9, p. 79] that minors whose elements satisfy (23) can be written as products of elements of the matrix and minors of the form

$$B^+\begin{pmatrix} i, i+1, \cdots, i+p-1 \\ k, k+1, \cdots, k+p-1 \end{pmatrix}, \qquad 0 \leqq i-k \leqq 1. \tag{24}$$

We shall now show that the minors of (24) are nonnegative.

Case I. $i - k = 1$. In this case, if we choose $S^{(p)}$ to be a $p \times p$ diagonal matrix whose diagonal entries $s_{\alpha,\alpha}$ are given by

$$s_{\alpha,\alpha} = (3)^{\alpha-1}, \qquad 1 \leqq \alpha \leqq p,$$

then it is easy to verify from (21) that

$$(S^{(p)}) \cdot \begin{bmatrix} b^+_{i,i-1} & b^+_{i,i} & \cdots & b^+_{i,i+p-2} \\ b^+_{i+1,i-1} & b^+_{i+1,i} & \cdots & b^+_{i+1,i+p-2} \\ \cdots & \cdots & \cdots & \cdots \\ b^+_{i+p-1,i-1} & b^+_{i+p-1,i} & \cdots & b^+_{i+p-1,i+p-2} \end{bmatrix} \cdot (S^{(p)})^{-1}$$

is strictly diagonally dominant for all $2 \leqq i \leqq n-p+1$, and $1 \leqq p \leqq n-1$. Therefore, the minors

$$B^+\begin{pmatrix} i, i+1, \cdots, i+p-1 \\ i-1, i, \cdots, i+p-2 \end{pmatrix}, \qquad 2 \leqq i \leqq n-p+1; \qquad 1 \leqq p \leqq n-1,$$

are all positive.

Case II. $i - k = 0$. For this case, we shall consider minors of B rather than B^+. Since B and B^+ are similar, we have that

$$B\begin{pmatrix} i, \cdots, i+p-1 \\ i, \cdots, i+p-1 \end{pmatrix} = B^+\begin{pmatrix} i, \cdots, i+p-1 \\ i, \cdots, i+p-1 \end{pmatrix},$$

$$1 \leqq i \leqq n-p+1; \qquad 1 \leqq p \leqq n.$$

Defining

$$B(i,p) \equiv \begin{bmatrix} b_{i,i} & b_{i,i+1} & \cdots & b_{i,i+p-1} \\ b_{i+1,i} & b_{i+1,i+1} & \cdots & b_{i+1,i+p-1} \\ \cdots & \cdots & \cdots & \cdots \\ b_{i+p-1,i} & b_{i+p-1,i+1} & \cdots & b_{i+p-1,i+p-1} \end{bmatrix}, \tag{25}$$

[1] Here, we are using the notation of [5] to denote minors.

assume for the moment that $B(i, p)$ is monotone[2], and denote the matrix $B^{-1}(i, p)$ by $(r_{j,k}^{(i,p)})$. Hence, by definition

$$0 \leqq r_{p,1}^{(i,p)} = B^{+}\begin{pmatrix} i, i+1, \cdots, i+p-1 \\ i-1, i, \cdots, i+p-2 \end{pmatrix} \Big/ B\begin{pmatrix} i, i+1, \cdots, i+p-1 \\ i, i+1, \cdots, i+p-1 \end{pmatrix}. \tag{26}$$

It therefore follows from Case I that

$$B\begin{pmatrix} i, i+1, \cdots, i+p-1 \\ i, i+1, \cdots, i+p-1 \end{pmatrix} > 0, \qquad 1 \leqq i \leqq n-p+1; \qquad 1 \leqq p \leqq n,$$

if $B(i, p)$ is monotone for all $1 \leqq i \leqq n - p + 1$, $1 \leqq p \leqq n$.

We shall now show that the particular matrix $B(1, n) = (b_{i,j})$ defined by (21) is monotone; the proof for an arbitrary successive principal minor $B(i, p)$ follows along the same lines. Define the $n \times n$ matrix C by

$$C = \begin{bmatrix} 1 & \mathbf{0}^T \\ -\mathbf{d} & B(1, n-1) \end{bmatrix},$$

where $\mathbf{d}$ is the vector with $n - 1$ components d_i given by

$$d_1 = \textstyle\sum_{j=1}^{n-1} b_{1,j} = (1 + \lambda_1 h); \qquad d_i = 0, 2 \leqq i \leqq n.$$

Calling the first row of C the 0th row, we now define two M-matrices[3], M_1 and M_2, as follows:

$$\begin{cases} (M_1 u)_0 = 2; \\ (M_1 u)_i = -\tfrac{1}{2}\lambda_i h u_{i-1} + \tfrac{1}{4}(1 + 3\lambda_i h) u_i, & 1 \leqq i \leqq n-1; \\ (M_1 u)_n = 1; \end{cases}$$

$$\begin{cases} (M_2 u)_0 = \tfrac{1}{2}; \\ (M_2 u)_i = -u_{i-1} + 2u_i, & 1 \leqq i \leqq n. \end{cases}$$

It is then easily verified that

$$R \equiv M_1 M_2 - C \geqq 0.$$

If we define $\mathbf{e}$ to be the vector with all components unity and ξ to have components

$$\xi_0 = 1, \qquad \xi_i = 0, \qquad 1 \leqq i \leqq n,$$

then since $C\mathbf{e} \geqq \xi$,

$$0 \leqq M_2^{-1} M_1^{-1} R\mathbf{e} = \mathbf{e} - M_2^{-1} M_1^{-1} C\mathbf{e} \leqq \mathbf{e} - M_2^{-1} M_1^{-1} \xi. \tag{27}$$

It is easily verified that $M_1^{-1}\xi \geqq \frac{1}{2}\xi$, and that $M_2^{-1}\xi > 0$, so we have from (27)

$$0 \leqq M_2^{-1} M_1^{-1} R\mathbf{e} \leqq \mathbf{e} - \tfrac{1}{2} M_2^{-1}\xi < \mathbf{e},$$

[2] A real $n \times n$ matrix B is *monotone* if and only if B is nonsingular and $B^{-1} \geqq 0$, i.e., every element of the matrix B^{-1} is a nonnegative real number.

[3] A real $n \times n$ matrix $B = (b_{i,j})$ is an *M-matrix* if and only if B is monotone and $b_{i,j} \leqq 0$ for all $i \neq j$, $1 \leqq i, j \leqq n$.

which implies [10, p. 17] that the spectral radius $\rho(M_2^{-1} M_1^{-1}R)$ satisfies

$$\rho(M_2^{-1}M_1^{-1}R) < 1.$$

Therefore, we can express $(1 - M_2^{-1}M_1^{-1}R)^{-1}$ as the convergent matrix series

$$(1 - M_2^{-1}M_1^{-1}R)^{-1} = 1 + (M_2^{-1}M_1^{-1}R) + (M_2^{-1}M_1^{-1}R)^2 + \cdots \geqq 0.$$

Since $M_2^{-1}M_1^{-1}R$ is nonnegative, the above expression shows that $(1 - M_2^{-1}M_1^{-1}R)$ is an M-matrix, and as

$$C = M_1M_2(I - M_2^{-1}M_1^{-1}R)$$

is the product of three M-matrices, C is evidently a monotone matrix. From the definition of the matrix C, it follows that

$$C^{-1} = \begin{bmatrix} 1 & \mathbf{0}^T \\ B^{-1}\mathbf{d} & B^{-1} \end{bmatrix},$$

and as C is monotone, every entry of C^{-1} is necessarily nonnegative. Thus, B^{-1} is a nonnegative matrix, or equivalently, B is monotone.

In summary, collecting the results of Cases I and II, all the minors of B^+ are nonnegative. Since the superdiagonal and subdiagonal of B^+ have only positive entries, we again conclude [5, p. 126] that B^+ is an oscillation matrix, and as such B^+ has positive real distinct eigenvalues. Since B is diagonally similar to B^+ by definition, then B also has positive real distinct eigenvalues, completing the proof.

We conclude this section with some remarks. Although more tedious to describe in detail, arguments similar to those used in Theorem 3 further show that a matrix B can be derived so as to have positive distinct eigenvalues even if $\lambda(x)$ changes sign in $0 \leqq x \leqq l$, provided that enough mesh points are used between successive zeros of $\lambda(x)$. The derivation of the entries of this matrix B is altered in that whenever $\lambda_i < 0$ for some i, a *forward* spatial difference approximation for $c_x(ih, t)$ is used, rather than the *backward* difference approximation of (21). In other words, a forward or backward spatial difference approximation to $c_x(ih, t)$ is chosen, depending on the sign of $\lambda_i = \lambda(ih)$. As in Theorem 2, these results can be extended to the case (13) in which one has in addition variable diffusivity, i.e., $K(x)$ in (13) is a positive function of position.

4. Other semi-discrete approximations. The semi-discrete approximations of Sections 2–3 to (1)–(2) are obviously not the only ones which can be used. The following other approximations also come to mind.

a. *Lower-Order Non-Central Difference Approximations.* If we use the following lower order non-central difference approximations in (5) for $\lambda > 0$:

$$\lambda \frac{\partial c(x_i, t)}{\partial x} \doteq \lambda \left[\frac{c_i(t) - c_{i-1}(t)}{h}\right] \tag{28}$$

and retain the three-point central difference approximation to $\partial^2 c(x_i, t)/\partial x^2$, then each local error term $\bar{\tau}$ is now only of order h as $h \to 0$. On the other hand, the proof of Theorem 1 in Section 2 can be extended to show that the associated tridiagonal coefficient matrix $\tilde{A}$ will have positive real eigenvalues for *all* $h > 0$.

Hence, as in Section 3, there is no restriction on h in terms of λ for non-oscillatory semi-discrete solutions. However, the local accuracy of these approximations does not compare favorably with the local accuracy of the semi-discrete difference approximations of (21) which also possess non-oscillatory solutions for all $h > 0$.

b. *Change of Variables.* If we define for constant $\lambda > 0$

$$\theta(x, t) = c(x, t) \exp\left(-\tfrac{1}{2}\lambda x\right); \qquad 0 \leqq x \leqq l, \qquad t \geqq 0, \tag{29}$$

then $\theta(x, t)$ satisfies the differential equation

$$\frac{\partial\theta(x, t)}{\partial t} = \frac{\partial^2\theta(x, t)}{\partial x^2} - \frac{\lambda^2}{4}\theta(x, t), \qquad 0 < x < l, \qquad t > 0, \tag{30}$$

where $\theta(x, 0) = 0$ for $0 < x < l$, $\theta(0, t) = 1$ for $t > 0$, and

$$\frac{\partial\theta(l, t)}{\partial x} = -\frac{\lambda}{2}\theta(l, t), \qquad t > 0. \tag{31}$$

Again, using standard three-point central difference approximations to the right side of (30), it is easy to show that the associated $n \times n$ coefficient matrix will have positive real eigenvalues for *all* $h > 0$. Moreover, the local accuracy of such an approximation is of order h^2 as $h \to 0$. Unfortunately, it can be shown that, upon transforming the θ's back to c's, that the related steady-state concentrations are all *greater than unity* for any $h > 0$. In other words, these non-oscillatory semi-discrete approximations have physically unacceptable steady-state values; the semi-discrete approximations of (21) on the other hand can be verified to possess the proper steady-state behavior.

5. Time discretizations. Having analyzed the oscillation problem for semi-discrete approximations, we now turn to time discretizations, which are of course necessary in practical computations. As we shall see, time discretizations *can* introduce oscillatory behavior even in cases where no such oscillation exist for the semi-discrete difference approximations.

The solution of (7) is given explicitly by

$$\mathbf{w}(t + \Delta t) = A^{-1}\mathbf{s} + \exp(-\Delta t A)\cdot\{\mathbf{w}(t) - A^{-1}\mathbf{s}\}, \qquad t \geqq 0, \qquad \Delta t \geqq 0, \tag{32}$$

where $\mathbf{w}(0) = \mathbf{0}$ from (2), and $A^{-1}\mathbf{s}$ will have all components unity if the steady-state solution of (7) agrees with that of the physical problem. Using a fixed time increment Δt, we approximate the exponential matrix $\exp(-\Delta t A)$ by

$$\exp(-\Delta t A) \doteqdot [Q(\Delta t A)]^{-1}P(\Delta t A), \tag{33}$$

where $Q(\Delta t A)$ and $P(\Delta t A)$ are real polynomials in $\Delta t A$, and $Q(\Delta t A)$ is non-singular. This approximation generates a sequence of vectors $\mathbf{z}(m\Delta t)$, defined by

$$\mathbf{z}[(m + 1)\Delta t] = A^{-1}\mathbf{s} + Q^{-1}(\Delta t A)\cdot P(\Delta t A)\{\mathbf{z}(m\Delta t) - A^{-1}\mathbf{s}\}, \quad m \geqq 0, \tag{34}$$

where $\mathbf{z}(0) = \mathbf{0}$ from (2), and $\mathbf{z}(m\Delta t)$ approximates the vector $\mathbf{w}(m\Delta t)$.

With the results of our previous theorems, we now assume that the semi-discrete difference matrix A possesses only positive real distinct eigenvalues μ_i, with associated eigenvectors $\mathbf{v}_i$. Using (20), we thus deduce that

$$\mathbf{z}(m\Delta t) = \sum_{i=1}^{n} d_i \left[1 - \left(\frac{P(\Delta t \mu_i)}{Q(\Delta t \mu_i)} \right)^m \right] \mathbf{v}_i, \qquad m \geqq 0. \tag{35}$$

Thus, for $d_i \neq 0$, the coefficient of $\mathbf{v}_i$ *oscillates* with m if and only if

$$\frac{P(\Delta t \mu_i)}{Q(\Delta t \mu_i)} < 0. \tag{36}$$

We now consider various standard approximations [10, p. 262] for $\exp(-\Delta tA)$, which arise from Padé approximations to e^z, and we further require that no oscillations occur.

a. *Forward-Explicit Approximation.* In this case, $P(\Delta tA) = I - \Delta tA$, $Q(\Delta tA) = I$. Thus, (36) is invalid if

$$0 < \Delta t \leqq \frac{1}{\max\limits_{1 \leqq i \leqq n} \mu_i}. \tag{37}$$

It should be mentioned that the criterion for stability [10, p. 268] of this explicit method similarly gives the restriction that

$$0 < \Delta t \leqq \frac{2}{\max\limits_{1 \leqq i \leqq n} \mu_i}.$$

b. *Backward-Implicit Approximation.* In this case, $P(\Delta tA) = I$, $Q(\Delta tA) = I + \Delta tA$. Since $P(\Delta t\mu_i)/Q(\Delta t\mu_i) = 1/(1 + \Delta t\mu_i)$, then the backward-implicit difference method is *non-oscillatory* for *any* $\Delta t > 0$.

c. *Crank-Nicolson Approximation.* In this case, $P(\Delta tA) = 2I - \Delta tA$, $Q(\Delta tA) = 2I + \Delta tA$. Thus, (36) is invalid if

$$0 < \Delta t \leqq \frac{2}{\max\limits_{1 \leqq i \leqq n} \mu_i}. \tag{38}$$

This criterion is actually much too restrictive since the small eigenvalues dominate the solution. Experimentally we found that the oscillations resulting from the time discretizations were eliminated, for all practical purposes, if

$$0 < \Delta t \leqq \frac{1}{\min\limits_{1 \leqq i \leqq n} \mu_i}. \tag{38'}$$

We remark that the forward-explicit and the backward-implicit approximations of $\exp(-\Delta tA)$ above, are *first-order* correct, i.e., these approximations have expansions for Δt small which agree only through linear terms in Δt with the corresponding expansion for $\exp(-\Delta tA)$. The Crank-Nicolson approximation is *second-order* correct, but is restricted by (38′). Finally, we merely state that the following Padé approximation $E_{2,1}(\Delta tA)$[10, p. 267], is *third-order* correct, and like the backward-difference approximation, is non-oscillatory for *any* $\Delta t > 0$.

Specifically, in this case

$$P(\Delta t A) \equiv I - \frac{2}{3}\Delta t A + \frac{1}{6}(\Delta t A)^2; \qquad Q(\Delta t A) = I + \frac{\Delta t}{3} A.$$

We should note here that as for the forward-explicit approximation, this approximation, $E_{2,1}$, has a criterion for stability given by

$$0 < \Delta t \leqq \frac{6}{\max_{1 \leqq i \leqq n} \mu_i}. \tag{39}$$

REFERENCES

[1] R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill Book Co., Inc., New York (1960), 321 pp.

[2] Birkhoff, Garrett and Saunders MacLane, *A Survey of Modern Algebra*, 3rd Edition, Macmillan, New York, (1965), 437 pp.

[3] L. Collatz, *The Numerical Treatment of Differential Equations*, 3rd Edition, Springer-Verlag, Berlin, (1960), 598 pp.

[4] Fort, Tomlinson, *Finite Differences and Difference Equations in the Real Domain*, Clarendon Press, Oxford (1948), 251 pp.

[5] F. R. Gantmakher, *Application of the Theory of Matrices* (translated and revised by J. L. Brenner) Interscience Publishers, New York (1959), 317 pp.

[6] F. R. Gantmakher and M. G. Krein, *Oscillation Matrices and Small Vibrations of Mechanical Systems*, Moscow (1950) (in Russian). (An English translation is available through the Office of Technical Service, Dept. of Commerce, Washington 25, D. C.), 408 pp.

[7] M. Jakob, *Heat Transfer*, John Wiley and Sons, New York (1949), 758 pp.

[8] D. W. Peaceman and H. H. Rachford, "Numerical Calculation of Multi-Dimensional Miscible Displacement," Soc. Petroleum Engineering Journal **24** (1962): 327–338.

[9] H. S. Price, "Monotone and Oscillation Matrices Applied to Finite Difference Approximations," Report No. 189, Gulf Research & Development Company, Pittsburgh, Pa. (April 22, 1965), 133 pp.

[10] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1962), 322 pp.

Gulf Research & Development Company, Pittsburgh, Pa.
Case Institute of Technology, Cleveland, Ohio
Kuwait Oil Company, Ahmadi, Kuwait

(Received August 24, 1965)