IMPLICIT SCHEMES AND LU DECOMPOSITIONS*

A. Jameson
Courant Institute of Mathematical Sciences
New York University

E. Turkel
Courant Institute of Mathematical Sciences
New York University

Abstract

Implicit methods for hyperbolic equations are analyzed using LU decompositions. It is shown that the inversion of the resulting tridiagonal matrices is usually stable even when diagonal dominance is lost. Furthermore, these decompositions can be used to construct stable algorithms in multi-dimensions. When marching to a steady state, the solution is independent of the time. Alternating direction methods which solve for $u^{n+1} - u^n$ are unconditionally unstable in three-space dimensions and so the new method is more appropriate. Furthermore, only two factors are required even in three-space dimensions and the operation count per time step is low. Acceleration to a steady state is analyzed, and it is shown that the fully implicit method with large time steps approximates a Newton-Raphson iteration procedure.

*The research for the first author was supported by the Office of Naval Research N00014-77–C-0032, NR06l-243. The research for the second author was partially supported by the Department of Energy Grant DOE EY-76-C-02-3077 and partially supported by NASA Contract No. NAS1-14101 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665.

1 Introduction

The use of implicit methods to solve hyperbolic equations has been increasing in recent years (e.g, [1], [2], [7]). Although implicit methods are frequently unconditionally stable, the permissible time step may still be restricted by the need to maintain a desired level of accuracy. Two classes of problems may be distinguished for which implicit methods are likely to be advantageous. First, there are stiff problems which contain several time scales in which most of the energy is contained in the slow modes. Nevertheless, the time step of an explicit method would be limited by a stability criterion set by the speed of the fast mode. Secondly, there are problems in which only a steady-state solution is desired and the time-dependent equations are used merely as a device for the iterative solution of the steady-state equations.

Implicit methods have the disadvantage that they require the solution of a large number of coupled equations at each time step. Hence, the reduction in the number of time steps compared with an explicit method may be outweighed by the increase in the number of arithmetic operations required for each time step. With a typical alternating direction method one needs to invert block tridiagonal matrices. If these matrices can be inverted by Gaussian eliminations without pivoting, the inversion can be accomplished by the Thomas algorithm in $O(m^3N)$ operations where m is the block size and N is the number of unknowns (see [6]). For many standard algorithms, diagonal dominance is lost when the time step becomes large. It is then no longer clear that the Thomas algorithm is numerically stable.

Another difficulty with alternating direction methods is encountered in the three dimensional case. When marching to a steady state using large time steps, one wants to ensure that the numerical solution is independent of the size of the time steps. A simple way to do this is to solve for $\Delta u^n = u^{n+1} - u^n$ at each time step. The equations then have the form

$$Q^n \Delta u^n = \Delta t L u^n$$

(see for example [2]). In this case it is evident that in the steady state we have $Lu = 0$ independent of Δt. In the two-dimensional case alternating direction methods which solve for either u^{n+1} or Δu^n are equivalent. However, in the three-dimensional case the two approaches yield different schemes. The three-dimensional alternating direction algorithm is unconditionally stable in the linear case if one solves for u^{n+1}, but the steady state solution depends on Δt. On the

other hand if one solves for Δu^n to produce a steady solution independent of Δt, then the algorithm is unconditionally unstable for scalar problems. For the Euler equations the equation for the entropy is essentially a scalar equation. Hence, this method is not stable for inviscid fluid dynamics.

In this study we discuss a class of implicit methods in which pre-calculated LU decompositions are used to approximate the equations obtained by linearizing a Crank-Nicolson or fully implicit scheme. It is shown that this approach can be used to derive schemes which are unconditionally stable in any number of space dimensions and also yield a steady state solution which is independent of Δt. The operation count at each time step is also quite moderate because the LU decomposition produces equations which only require the inversion of m x m diagonal blocks for each factor. In three dimensions there are only two factors instead of the three factors of an alternating direction algorithm.

The matrices of an unfactored implicit algorithm are not diagonally dominant for large time steps. Thus, the usual sufficient conditions for using Gaussian elimination without pivoting are no longer satisfied. We show that the LU decomposition can often still be constructed in such a way that each factor is diagonally dominant. This ensures the numerical stability of the inversions required at each time step.

2 One-Dimensional Problems

Consider the one dimensional system

$$w_t + Aw_x = 0 \tag{2.1}$$

with A a constant matrix.

Then the Crank-Nicolson scheme is given by

$$\left(I + \frac{\Delta tA}{2}\delta\right) w^{n+1} = \left(I - \frac{\Delta tA}{2}\delta\right) w^n \tag{2.2}$$

or

$$\left(I + \frac{\Delta tA}{2}\delta\right) \left(w^{n+1} - w^n\right) = -\Delta tA\delta w^n$$

where δ is a central difference operator defined by

$$\delta w_j^n = \frac{w_{j+1}^n - w_{j-1}^n}{2\Delta x} \tag{2.3}$$

We also define forward and backward difference operators

$$D+w_j = \frac{w_{j+1} - w_j}{\Delta x} \qquad\qquad D-w_j = \frac{w_j - w_{j-1}}{\Delta x} \tag{2.4}$$

The solution of (2.2) requires the inversion of a block tridiagonal matrix. Instead, we can approximately factor (2.2) by

$$\left(I + \frac{\Delta t A}{4} D_+\right)\left(I + \frac{\Delta t A}{4} D_-\right)(w^{n+1} - w^n) = \Delta t A \delta w^n \tag{2.5}$$

Since $w^{n+1} - w^n$ is of order Δt the difference between the schemes (2.2) and (2.4) are terms of order $(\Delta t)^3$ and so the additional errors are of the order of the truncation error. For a bounded domain the operators $I + \frac{\Delta t A}{4} D_+$ and $I + \frac{\Delta t A}{4} D_-$ can be inverted directed by beginning at the left and right boundaries, respectively. Computational experience indicated that this method fails for large Δt. This is true even though (2.4) is unconditionally stable in terms of the usual initial value stability analysis. The reason for this is that if A has both positive and negative eigenvalues, the factors lose diagonal dominance. The inversion process then becomes numerically unstable.

To analyze this further we consider the general three-point approximation to (2.1) which is second order accurate in space. Let

$$\Delta w_j^n = w_j^{n+1} - w_j^n \tag{2.6}$$

Then, we have

$$\begin{aligned}\Delta w_j^n + \sigma\left(\Delta w_{j+1}^n - 2\Delta w_j^n + \Delta w_{j-1}^n\right) = \\ - \frac{\lambda A}{2}\left[\xi\left(w_{j+1}^{n+1} - w_{j-1}^{n+1}\right) + \left(1-\xi\right)\left(w_{j+1}^n - w_{j-1}^n\right)\right]\end{aligned} \tag{2.7}$$

Here, $\lambda = \frac{\Delta t}{\Delta x}$ and ξ denotes the weighting of the space differences at the new and old time levels. $\xi = \frac{1}{2}$ yields the Crank-Nicolson scheme while $\xi = 1$ yield the fully implicit method. σ is a free parameter; it is convenient to allow it to have the general form

$$\sigma = \sigma_1 + \sigma_2 \lambda^2 A^2 \xi^2 \tag{2.8}$$

(2.6) can be rewritten as

$$\Delta w_j^n + \frac{\lambda A \xi}{2}\Big(\Delta w_{j+1}^n - \Delta w_{j-1}^n\Big) + \sigma\Big(\Delta w_{j+1}^n - 2\Delta w_j^n + \Delta w_{j-1}^n\Big) = \frac{\lambda A}{2}\Big(w_{j+1}^n - w_{j-1}^n\Big) \tag{2.9}$$

or

$$Q\Big(w^{n+1} - w^n\Big) = -\lambda A \delta w \tag{2.10}$$

Q is a block traditional matrix. Omitting the effect of boundaries, Q can be replaced by LU where L and U have the form

$$\mathbf{L} = \begin{pmatrix} \ell_1 & 0 & \cdots & 0 \\ \ell_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & \ell_2 & \ell_1 \end{pmatrix}$$

$$\mathbf{U} = \begin{pmatrix} u_1 & u_2 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & u_2 \\ 0 & \cdots & 0 & u_1 \end{pmatrix} \tag{2.11}$$

where

$$\ell_1 = \alpha_1 + \beta_1 \lambda A \qquad \ell_2 = \gamma_1 - \beta_1 \lambda A \tag{2.12}$$

$$u_1 = \alpha_2 - \beta_2 \lambda A \qquad u_2 = \gamma_2 + \beta_2 \lambda A \tag{2.13}$$

and α_j, β_j may be matrix functions of A.

Given the matrix Q the LU decomposition is unique except for a diagonal matrix, i.e., given L, U the most general decomposition of Q is given by $Q = L'U'$ with $L' = LD$ and $U' = D^{-1}U$ for some nonsingular diagonal matrix D. The matrix D does not enter in any essential manner, and it will be chosen for convenience. In particular, we consider a scaling so that $\alpha_1 + \gamma_1 = I$.

For second order accuracy in space, one requires that

$$\begin{aligned} \alpha_1 &= \alpha_2 \\ \beta_1 = \beta_2 &= \frac{\xi}{2} \\ \gamma_1 &= \gamma_2 \end{aligned}$$

Hence, we have in (2.7)

$$\begin{aligned} \ell_1 &= \alpha + \frac{\xi\lambda A}{2} & \ell_2 &= \gamma - \frac{\xi\lambda A}{2} \\ u_1 &= \alpha - \frac{\xi\lambda A}{2} & u_2 &= \gamma + \frac{\xi\lambda A}{2} \end{aligned} \tag{2.14}$$

with $\alpha+\gamma = 1$. Thus, we have one free parameter, α, at our disposal. Multiplying L and U as given by (2.8) and comparing with (2.6) we find that

$$\alpha(1-\alpha) + \frac{\xi^2\lambda^2 A^2}{4} = \sigma = \sigma_1 + \sigma_2\lambda^2 A^2\xi^2 \tag{2.15}$$

and so

$$\alpha = \frac{1 + \left[1 - (4\sigma_2 - 1)\,\xi^2\lambda^2 A^2 - 4\sigma_1\right]^{\frac{1}{2}}}{2} \tag{2.16}$$

We stress that the inversion procedure is well conditioned if and only if the matrices L and U are diagonally dominant. The diagonal dominance of Q is only sufficient but not necessary. Hence, for these inversions to be well conditioned we require

$$\begin{aligned} \|\gamma - \frac{\xi\lambda A^{-1}}{2}\alpha + \frac{\xi\lambda A}{2}\| \leq 1 \\ \|\alpha - \frac{\xi\lambda A^{-1}}{2}\gamma + \frac{\xi\lambda A}{2}\| \leq 1 \end{aligned}$$

To demonstrate the importance of diagonal dominance for the L and U factors we consider the system

$$\mathbf{L}x = \mathbf{f}$$

with

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ b & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & b & 1 \end{pmatrix} \qquad \mathbf{f} = \begin{pmatrix} \varepsilon \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

The solution is

$$\begin{aligned} x_1 &= \varepsilon \\ x_j &= (-b)^{j-1}\varepsilon \end{aligned}$$

If b ¿ 1, the inversion process is not well posed even though the matrix is already in lower triangular form. The inverse of $\mathbf{L}$ is given by

$$\mathbf{L^{-1}} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -b & 1 & 0 & \cdots & 0 \\ b^2 & -b & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ (-b)^{n-1} & (-b)^{n-2} & & & 1 \end{pmatrix}$$

Hence, for $|b| > 1$ the condition number increase exponentially as n increases. Conversely, If $\mathbf{L}$ and $\mathbf{U}$ are diagonally dominant, then it is easy to show that the pivots in Gaussian elimination without pivoting cannot grow with increasing n.

Hence we require that

$$\left(\gamma - \frac{\xi\lambda A}{2}\right)^2 \geq \left(\alpha + \frac{\xi\lambda A}{2}\right)^2$$

and

$$\left(\alpha - \frac{\xi\lambda A}{2}\right)^2 \geq \left(\gamma + \frac{\xi\lambda A}{2}\right)^2$$

Since $\alpha + \gamma = 1$, the inversion algorithm is well conditioned if and only if

$$(\xi\lambda A)^2 \leq (\alpha - \gamma)^2 = (2\alpha - 1)^2 \tag{2.17}$$

We want the method to be unconditionally stable and so, (2.11) implies that α and γ must be functions of A or at least functions of the spectral radius of A.

For a well conditioned problem, (2.11) together with (2.10) requires that

$$\xi^2\lambda^2A^2 \leq (2\alpha - 1)^2 = 1 - (4\sigma_2 - 1)\xi^2\lambda^2A^2 - 4\sigma_1$$

or equivalently

$$4\sigma_2\xi^2\lambda^2A^2 \leq 1 - 4\sigma_1 \tag{2.18}$$

3 Analysis of Some Standard Schemes

We now consider some of the methods which can be derived from the general three-point scheme (2.6) and show that many of them lead to diagonally dominant **L** and **U** factors which yield a stable inversion process.

1. Standard second-order methods
   $\sigma_1 = \sigma_2 = 0$; so (3.3) is always satisfied. Hence, these methods are well conditioned for all ξ and all times steps.

2. 2 - 4 methods
   $\sigma_1 = \frac{1}{6} \quad \sigma_2 = 0$; (3.3) is always satisfied.

3. 4 - 4 methods
   $\xi = \frac{1}{2}, \sigma_1 = \frac{1}{6}, \sigma_2 = \frac{1}{3}$. In this case (3.3) implies that the inversion is well conditioned only if $\lambda A \leq 1$. This is confirmed by the numerical results of [5].

4. Scheme (2.4)
   $\sigma_1 = 0, \sigma_2 = \frac{1}{4}$ and so (3.3) implies that the method is well conditioned only if $\xi^2 \lambda^2 A^2 \leq 1$. This was conformed by computer runs.

5. Diagonally dominant schemes
   If we want schemes that are diagonally dominant, this can be achieved by choosing $\sigma_1 < 0, \sigma_2 < 0$ and $\sigma_1 \sigma_2 > \frac{1}{16}$. If $\sigma_1 < 0, \sigma_2 < 0$ then (3.3) is trivially satisfied. Hence, if the basic scheme is diagonally dominant, then the **L** and **U** factors are also diagonally dominant.

4 A Practical LU Decomposition

In section 2 we showed that an LU decomposition of form (2.8) is well conditioned if and only if

$$\xi^2 \lambda^2 A^2 = (2\alpha - 1)^2 \tag{4.1}$$

In section 3 we demonstrated that (4.1) is automatically satisfied for several well known schemes. In this case the LU decomposition is useful mainly for the purpose of analyzing the scheme because the resulting is a complicated matrix function of A. Furthermore, the introduction of boundaries complicated the LU factorization.

In order to generate new schemes which can be readily generalized to the multi-dimensional situation, we can reverse the approach by choosing the L and U factors as determining the scheme. We can then insure that the LU decomposition is quite simple and at the same time we can select the free parameter α so that (4.1) is always satisfied. Letting $|.|$ denote the absolute value of a matrix as determined by function theory, one choice for α is

$$\alpha = \frac{1}{2}(I + |\lambda A \xi|) \tag{4.2a}$$

For two-dimensional problems, $\xi = \frac{1}{2}$, this can be generalized by

$$\alpha = \frac{1}{2}\left(\frac{I}{2} + \left|A\frac{\Delta t}{\Delta x}\right| + \left|B\frac{\Delta t}{\Delta y}\right|\right) \tag{4.2b}$$

The absolute value of these matrices can be calculated by diagonalizing A and B independently. Although this approach is valid from a theoretical viewpoint, it is not computationally efficient. Instead, we can replace (4.2a) by

$$\alpha = \frac{1}{2}(1 + \rho\xi\lambda) \qquad\qquad \gamma = \frac{1}{2}(1 - \rho\varepsilon\lambda) \tag{4.3}$$

This choice of α satisfies (4.1) if ρ is equal to or greater than the spectral radius of A. This choice yields a scalar α which is computationally efficient. The extensions to several dimensions are discussed in section 6.

5 Boundary Treatment

There are two different approaches towards constructing boundary equations for those data that are not specified analytically, one approach is to put reasonable factors into the upper part of L and the lower corner of U. Having, by some other procedure, decided what equations one wants, one then uses the Sherman-Morrison formulas to correct the inverse for the given boundary treatment. This procedure can be expensive as another inverse is needed for each rank-one modification.

Instead, we shall include the boundary treatment within the LU decomposition. We shall concentrate on the left boundary, $x = 0$, which requires modification of the L matrix. Similar modifications affect the U matrix for the right boundary.

Assuming that the boundary treatment is of first order accuracy, one finds that L should be modified to have the form

$$\mathbf{L} = \begin{pmatrix} a - \frac{\xi\lambda A}{2} & c + \frac{\xi\lambda A}{2} & \\ \gamma - \frac{\xi\lambda A}{2} & \alpha + \frac{\xi\lambda A}{2} & 0 \\ 0 & \ddots & \ddots \end{pmatrix} \tag{5.1}$$

With $a+c=1$. We use linear extrapolation outside the domain for those variables not given analytically. This is equivalent to (5.1) with

$$\begin{aligned} a &= \alpha + 2\gamma \\ c &= -\gamma \end{aligned} \tag{5.2}$$

Using the theory of Gustafsson, Kreiss, and Sundstrom [4] one can show that the initial boundary value scheme is unconditionally stable for $\xi \geq \frac{1}{2}$. (5.1) requires the inversion of a 2 x 2 block matrix for the boundary values. The algorithmic aspects of the scheme are described in greater detail in section 7.

6 Multidimensional LU Implicit Algorithms

In one dimension we constructed an approximate factorization which had the interpretation that both L and U were approximations to one sided differences. In two dimensions we can extend this technique.

Consider the equation

$$w_t + Aw_x + Bw_y = 0 \tag{6.1}$$

Let

$$\mathbf{L} = \begin{pmatrix} \ell_1 & & & & & \\ \ell_2 & \ddots & & & & \\ 0 & \ddots & \ddots & 0 & & \\ \ell_3 & & \ddots & \ddots & & \\ & \ddots & & \ddots & \ddots & \\ 0 & & \ell_3 & & \ell_2 & \ell_1 \end{pmatrix} \qquad \lambda = \frac{\Delta t}{\Delta x} = \frac{\Delta t}{\Delta y}$$

$$\mathbf{U} = \begin{pmatrix} u_1 & u_2 & 0 & u_3 & & 0 \\ & \ddots & \ddots & & \ddots & \\ & & \ddots & \ddots & & u_3 \\ & & & \ddots & \ddots & \\ & 0 & & & \ddots & u_2 \\ & & & & & u_1 \end{pmatrix}$$

where

It will be assumed that any shock waves contained in the flows to be computed are weak enough that the entropy and vorticity generated by the shock waves can be ignored without introducing serious errors. Consistent with this approximation we shall treat the exact potential flow equation in conservation form. Using Cartesian coordinates x, y, z we shall write this equation as

$$\frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) + \frac{\partial}{\partial z}(\rho w) = 0 \tag{6.2}$$

where ρ is the density and u, v, w are the velocity components. These are calculated as the gradient of the potential Φ.

$$u = \Phi_x, \quad v = \Phi_y, \quad w = \phi_z. \tag{6.3}$$

The flow is assumed to be uniform in the far field with a Mach number M_∞. At the body, the boundary condition is

$$u_n = 0 \tag{6.4}$$

where u_n is the normal velocity component. The density is computed from the isentropic formula.

$$\rho = \{1 + \frac{\gamma - 1}{2} M_\infty^2 (1 - q^2)\}^{\frac{1}{\gamma - 1}} \tag{6.5}$$

where ρ is the ratio of specific heats, and q is the speed,

$$q^2 = u^2 \ + v^2 \ + w^2 \tag{6.6}$$

With the normalization that $q = 1$ and $p = 1$ at infinity, the corresponding formulas for the pressure p and the local speed of

$$p = \frac{\rho^\gamma}{\gamma M_\infty^2}, \quad a^2 = \frac{\rho^{\gamma - 1}}{M_\infty^2}. \tag{6.7}$$

The shock jump conditions are

(a) continuity of Φ, implying a continuity of the tangential velocity component;

(b) continuity of ρu_n, where u_n is the normal velocity component.

Under the isentropic assumption the normal component of momentum is not conserved through the shock wave, leading to a body force which is an approximation to the wave drag. In any finite domain equations $(1) - (5)$ together with the shock jump relations (a) and (b) are equivalent to the Bateman variational principle that the integral

$$I = \int_\Omega p \, d\Omega \tag{6.8}$$

is stationary.

A difficulty with the formulation assuming potential flow is that corresponding to any solution of equation (1) there is a reverse flow solution, in which compression shock waves become expansion shock waves. In fact if central difference formulas are used throughout the domain, symmetric solutions, containing an expansion shock at the front and a compression shock at the rear, can be computed for a body with fore and aft symmetry such as an ellipse. This is a consequence of the absence of entropy from the formulation. In order to obtain a unique and physically relevant solution the shock jump relations (a) and (b) must be supplemented by the additional "entropy condition" that discontinuous expansions are to be excluded from the solution, corresponding to the fact that entropy cannot decrease in a real flow.

For this purpose the discrete approximation will be desymmetrized by the addition of artificial viscosity to produce an upwind bias in the supersonic zone. The added terms will be introduced in a manner such that the conservation form of equation (1) is preserved. Provided that the solution of the discrete equations converges in the limit as the cell width is reduced to zero, the correct shock jump relations consistent with the isentropic assumption will then be a natural consequence of the scheme.[10]

7 The Staggered Box Scheme

The basic idea of the numerical scheme is that cubes in the computational domain will be separately mapped to distorted cubes in the physical domain by inde-

pendent transformations from local coordinates X, Y, Z to Cartesian coordinates x, y, z as illustrated in Figure 1.

The mesh points are the vertices of the mapped cubes, and subscripts i, j, k will be used to denote the value of a quantity at

a mesh point. Subscripts $i + \frac{1}{2}, j + \frac{1}{2}, k + \frac{1}{2}$ will be used to denote points mapped from the centers of the cubes in the computational domain. In developing the difference formulas it will be convenient to introduce averaging the difference operators through the notation

$$\mu_X f = \frac{1}{2}(f_{i+\frac{1}{2},j,k} + f_{i-\frac{1}{2},j,k})$$

$$\delta_X f = f_{i+\frac{1}{2},j,k} - f_{i-\frac{1}{2},j,k}$$

with similar formulas for $\mu_y, \mu_z, \delta_y, \delta_z$. It will also be convenient to use notations such as

$$\begin{aligned}
\mu_{XX} f &= \mu_X(\mu_X f) \\
&= \frac{1}{4} f_{i+1,j,k} + \frac{1}{2} f_{i,j,k} + \frac{1}{4} f_{i-1,j,k} \\
\mu_{XY} f &= \mu_X(\mu_Y f) \\
\delta_{XX} f &= \delta_X(\delta_X f) \\
&= f_{i+1,j,k} + 2 f_{i,j,k} + f_{i-1,j,k} \\
\delta_{XY} f &= \delta_X(\delta_Y f)
\end{aligned}$$

Numbering the vertices of a particular cube from 1 to 8 as in Figure 1, the local mapping is constructed by a trilinear form in which the local coordinates lie in the range $-\frac{1}{2} \le X \le \frac{1}{2}, -\frac{1}{2} \le Y \le \frac{1}{2}, -\frac{1}{2} \le Z \le \frac{1}{2}$, so the vertices are at $X_i = \pm\frac{1}{2}, Y_i = \pm\frac{1}{2}, Z_i = \pm\frac{1}{2}$. Thus if the Cartesian coordinates of the i^{th} vertex of the mapped cube are x_i, y_i, z_i, the local mapping is defined by

$$x = 8 \sum_{i=1}^{8} x_i \, (\frac{1}{4} + X_i X) \, (\frac{1}{4} + Y_i Y) \, (\frac{1}{4} + Z_i Z) \tag{7.1}$$

with similar formulas for y, z. The potential Φ is assumed to have a similar form inside the cell:

$$\Phi = 8 \sum_{i=1}^{8} \Phi_i \, (\frac{1}{4} + X_i X) \, (\frac{1}{4} + Y_i Y) \, (\frac{1}{4} + Z_i Z) \tag{7.2}$$

These formulas preserve the continuity of x, y, z and at the boundary between any pair of cells, because the mappings in each cell reduce to the same bilinear form at the common face. At the center of a computational cell the derivatives of the transformation can be evaluated from equation (8) by formulas such as

$$\begin{aligned} x_X &= \frac{1}{4}\left(x_2 - x_1 + x_4 - x_3 + x_6 - x_5 + x_8 - x_7\right) \\ &= \mu_{YZ}\, \delta_X\, x \end{aligned}$$

Similarly it follows from equation (9) that

$$\Phi_X = \mu_{YZ}\, \delta_X\, \Phi, \qquad \Phi_{XY} = \mu_Z\, \delta_{XY}\, \Phi, \qquad \Phi_{XYZ} = \delta_{XYZ}\, \Phi$$

These formulas are simply an application of the box difference scheme.

Equation (1) will now be represented as a flux balance. For this purpose we introduce a secondary set of cells interlocking with the primary cells as illustrated in Figure 2.

In the computational domain the faces of the secondary cells span the midpoints of the primary cells. Since one secondary cell overlaps eight primary cells, in each of which there is a separate transformation, the secondary cells do not necessarily have smooth faces when they are mapped to the physical domain, but this is not important since their purpose is simply to serve as control volumes for the flux balance.

In order to derive the formula for the flux balance it is convenient to resort to tensor notation. Let the Cartesian and local coordinates be

$$x^1 = x, \quad x^2 = y, \quad x^3 = z$$

$$X^1 = X, \quad X^2 = Y, \quad X^3 = Z$$

The appearance of a repeated index in any formula will be understood to imply a summation over that index. Let H be the transformation matrix with elements $\frac{\partial x^i}{\partial X^j}$ and let h be the determinant of H. Let G be the matrix $H^T H$ with elements

$$g_{ij} = \frac{\partial x^k}{\partial X^i}\frac{\partial x^k}{\partial X^j} \tag{7.3}$$

Then G is the metric tensor. Also let g^{ij} be the elements of G^{-1}. Then the contravariant velocity components are

$$U = U^1, \quad V = U^2, \quad W = U^3$$

where

$$U^i = g^{ij} \frac{\partial \Phi}{\partial X^j} \tag{7.4}$$

It may be verified by applying the chain rule for partial derivatives that equation (1) can be written in the local coordinate system as

$$\frac{\partial}{\partial X^i}(\rho h U^i) = 0 \tag{7.5}$$

This corresponds to a well known formula for the divergence of a contravariant vector. In the computation of the density from equation (4) we now use the formula

$$q^2 = U^i \frac{\partial \Phi}{\partial X^i} \tag{7.6}$$

Also at a boundary $S(x, y, z)$ = constant, the condition that the normal velocity component is zero becomes

$$U^i \frac{\partial S}{\partial X^i} = 0$$

The mesh will be generated so that the boundary will coincide with faces of cells adjacent to the boundary. Thus the boundary condition will reduce to a simple form such as $V = 0$ on a cell face.

The formula for the local flux balance can now be written down by a second application of the box scheme on the secondary cells. Thus equation (12) is approximated by

$$\mu_{YZ}\ \delta_X\ (\rho h U) + \mu_{ZX}\ \delta_Y\ (\rho h V) + \mu_{XY}\ \delta_Z\ (\rho h W) = 0. \tag{7.7}$$

The physical interpretation of the quantities $\rho h U, \rho h V, \rho h W$ is that they are the fluxes across the faces of the secondary cell. Consequently this formula is equivalent to calculating the flux across the part of a face of a secondary cell lying in a particular primary cell by using values for ρ, h, U, V, W calculated at the center of that primary cell.

Adjacent to the body the flux balance is established on secondary cells bounded on one or more faces by the body surface as illustrated in Fig 3.

There is no flux across these faces and equation (14) is correspondingly modified.

Observe that equation (14) could also be derived from the Bateman variational principle. Suppose that the integral I defined by equation (7) is approximated by summing the volume of each primary cell multiplied by the pressure at its midpoint. Then on setting the derivative of I with respect to each nodal value $\Phi_{i,j,k}$ equal to zero to represent the fact that I is stationary, one recovers equation (14). In a finite element method using isoparametric trilinear elements the contribution of each cell would be calculated by an internal integration over the cell, allowing for the fact that according to the trilinear formulas p is not constant inside the cell.

The use of values of ρ, h, U, V, W calculated at the centers of the primary cells in equation (14), instead of values averaged over the relevant faces, simplifies the formulas at the expense of a "lumping error". Fortunately the contributions to the lumping error from adjacent primary cells offset each other. In fact, if we suppose the vertices of the cells to be generated by a global mapping smooth enough to allow Taylor series expansions of x, y, z as functions of X, Y, Z, then it can be seen from the interpretation of equation (14) as a box scheme that the local discretization error is of second order.

The introduction of lumped quantities in equation (14) is the source, however, of another difficulty. this is most easily seen by considering the case of incompressible flow in Cartesian coordinates. Setting $h = 1, \rho = 1$, equation (14) reduces in the two dimensional case to

$$\mu_{YY}\delta_{XX}\Phi + \mu_{XX}\delta_{YY}\Phi = 0$$

This is simply the rotated Laplacian as illustrated in Figure 4. The odd and even points are decoupled, leading to two independent solutions as sketched. In fact $\mu_{YY}\delta_{XX}\Phi$ and $\mu_{XX}\delta_{YY}\Phi$ are separately zero for $\Phi = 1$ at odd points, -1 at even points.

To overcome this difficulty, observe that it is due to the evaluation of the flux across the face labeled AB in Figure 5 using a value of calculated at the point A.

If we add a compensation flux $\varepsilon\ \Delta Y\ \Phi_{XY}$ across AB, the point at which Φ_X is effectively evaluated is shifted from A to B as is increased from 0 to 1/2. Taking the cell height ΔY as unity, consistent with the trilinear formula (8), the addition of similar compensation terms on all faces produces the scheme

$$\mu_{YY}\delta_{XX}\ \Phi + \mu_{XX}\delta_{YY}\ \Phi - \varepsilon\ \delta_{XXYY}\ \Phi = 0$$

Notice that setting $\varepsilon = \frac{1}{2}$ yields the standard five point scheme for Laplace's equation, while setting $\varepsilon = \frac{1}{3}$ yields the nine point fourth order accurate scheme.

In order to compensate for the lumping error in equation (14) in a similar manner, we first calculate influence coefficients giving the effective weight of $\delta_{XX}\Phi,\ \delta_{YY}\Phi,\ \delta_{ZZ}\Phi$ in equation (14) when the dependence of ρ on Φ_X, Φ_Y, Φ_Z is accounted for. These are

$$\begin{aligned} A_X &= \rho h \left(g^{11} - \frac{U^2}{a^2} \right) \\ A_Y &= \rho h \left(g^{22} - \frac{V^2}{a^2} \right) \\ A_Z &= \rho h \left(g^{33} - \frac{W^2}{a^2} \right) \end{aligned} \tag{7.8}$$

Now define

$$Q_{XY} = (A_X + A_Y)\ \mu_Z\ \delta_{XY}\ \Phi. \tag{7.9}$$

with similar formulas for Q_{YZ}, Q_{ZX}, and

$$Q_{XYZ} = (A_X + A_Y + A_Z)\ \delta_{XYZ}\ \Phi \tag{7.10}$$

Then the final compensated equation is

$$\begin{aligned} &\mu_{YZ}\ \delta_X\ (\rho h U) + \mu_{ZX}\ \delta_Y\ (\rho h V) + \mu_{XY}\ \delta_Z\ (\rho h W) \\ &-\varepsilon \left\{ \mu_Z\ \delta_{XY}\ Q_{XY} + \mu_X\ \delta_{YZ}\ Q_{YZ} + \mu_Y\ \delta_{ZX}\ Q_{ZX} - \frac{1}{2}\delta_{XYZ}\ Q_{XYZ} \right\} = 0 \end{aligned} \tag{7.11}$$

where $0 \leq \varepsilon \leq \frac{1}{2}$. This procedure has proved effective in suppressing high frequency oscillations in the solution.

This completes the definition of the discretization scheme for subsonic flow. It remains to add an artificial viscosity to desymmetrize the scheme in the supersonic zone. Instead of equation (12) we shall satisfy the modified flux balance equation

$$\frac{\partial}{\partial X}(\rho h U + P) \; + \frac{\partial}{\partial Y}(\rho h V + Q) \; + \frac{\partial}{\partial Z}(\rho h W + R) = 0$$

where the added fluxes P, Q and R are proportional to the cell width in the physical domain.Thus the correct conservation law will be recovered in the limit as the cell width decreases to zero. The added terms are designed to produce an upwind bias in the supersonic zone. As in the case of previous schemes for solving the potential flow equation in conservation form[5,6], they are modeled on the artificial viscosity of the nonconservative rotated difference scheme,[4] which has proved reliable in numerous calculations.

First we introduce the switching function

$$\mu = \max\left[0, \left(1 - \frac{a^2}{q^2}\right)\right]$$

Then P, Q, R are constructed to that

$$\begin{aligned} P \;\; approximates \;\; -\mu\,|U|\,\delta_X\,\rho \\ Q \;\; approximates \;\; -\mu\,|V|\,\delta_Y\,\rho \\ R \;\; approximates \;\; -\mu\,|W|\,\delta_Z\,\rho \end{aligned}$$

with an upwind shift in each case. Since $\mu = 0$ when $q < a$, the added terms vanish in the subsonic zone. In the numerical scheme equation (18) is actually modified by the addition of the terms

$$\delta_X P \; + \; \delta_Y Q \; + \; \delta_Z R$$

In order to form P, Q, R we first construct

$$\hat{P} = \mu h \frac{\rho}{a^2}\left(U^2\,\delta_{XX} + UV\,\mu_{XY}\,\delta_{XY} + WU\,\mu_{ZX}\,\delta_{ZX}\right)\Phi$$

$$\hat{Q} = \mu h \frac{\rho}{a^2}\left(UV\,\mu_{XY}\,\delta_{XY} + V^2\,\delta_{YY} + VW\,\mu_{YZ}\,\delta_{YZ}\right)\Phi$$

$$\hat{R} = \mu h \frac{\rho}{a^2}\left(WU\,\mu_{ZX}\,\delta_{ZX} + VW\,\mu_{YZ}\,\delta_{YZ} + W^2\,\delta_{ZZ}\right)\Phi$$

Then

$$P_{i+\frac{1}{2},j,k} = \begin{cases} \hat{P}_{i,j,k} & \text{if } U > 0 \\ \hat{P}_{i+1,j,k} & \text{if } U < 0 \end{cases}$$

with similar shifts for Q, R.

The motivation for these formulas is provided by the following analysis. When equation (12) is represented explicitly in quasilinear form, its leading terms are

$$\frac{\rho h}{a^2}\left\{(a^2 - q^2)\Phi_{ss} \; + \; a^2(\Delta\Phi - \Phi_{ss})\right\} = 0$$

where s is the local flow direction, and Δ is the Laplacian. In the transformed coordinate system

$$\frac{\partial \Phi}{\partial s} = \frac{U^i}{q}\frac{\partial \Phi}{\partial X^i}$$

so the leading terms of Φ_{ss} are $\frac{U^i U^j}{q^2}\frac{\partial^2 \Phi}{\partial X^i \partial X^j}$. According to the rotated difference scheme one should use upwind difference formulas to evaluate Φ_{ss} at supersonic points, as illustrated in Figure 6.

Now the upwind formula for Φ_{XX} can be regarded as an approximation to $\Phi_{XX} - \Delta X \Phi_{XXX}$. Similarly the upwind formula for Φ_{XY} yields an added term $\frac{1}{2}\ \Delta X\ \Phi_{XXY}\ +\ \frac{1}{2}\ \Delta Y\ \Phi_{XYY}$ and so on. The use of these formulas in the evaluation of $\frac{\rho h}{a^2}(a^2 - q^2)\Phi_{ss}$ thus produces an effective artificial viscosity

$$\begin{aligned} -\frac{\rho h}{a^2}\left(1 - \frac{a^2}{q^2}\right)\Big\{ & \Delta X\ U\left(U\ \Phi_{XXX} + V\ \Phi_{YXX} + W\ \Phi_{ZXX}\right) \\ & + \Delta Y\ V\left(U\ \Phi_{XYY} + V\ \Phi_{YYY} + W\ \Phi_{ZYY}\right) \\ & + \Delta Z\ W\left(U\ \Phi_{XZZ} + V\ \Phi_{YZZ} + W\ \Phi_{ZZZ}\right)\Big\} \end{aligned}$$

assuming that U, V, W are positive. Since $\frac{\partial \rho}{\partial (q^2)} = -\frac{\rho}{2a^2}$ it follows from equation (13) that

$$\rho_X = -\frac{\rho h}{a^2}\left(U\Phi_{XX} + V\Phi_{XY} + W\Phi_{XZ}\right)$$

Thus on setting $\Delta X = 1$, consistent with equation (8), leading terms of $-\left(\frac{\partial}{\partial x}\right)\left(\mu\, U\, \delta_X\, \rho\right)$ are

$$-\frac{\rho h}{a^2}\left(1 - \frac{a^2}{q^2}\right)\Delta X\Big(U\Phi_{XXX} + V\Phi_{YXX} + W\Phi_{ZXX}\Big)$$

which can be seen to be the desired quantity. Note that the construction of the artificial viscosity is based on the presumption of a smooth mesh in the supersonic zone.

Finally it remains to devise an iterative procedure for solving the nonlinear algebraic equations which result from the discretization. Following the same reasoning as was used for the iterative solution of the rotated difference scheme and earlier schemes in conservation form,[4–6] this is accomplished by embedding the steady state equation in an artificial time dependent equation. Thus we solve a discrete approximation to

$$\frac{\partial}{\partial X}(\rho hU + P) + \frac{\partial}{\partial Y}(\rho hV + Q) + \frac{\partial}{\partial Z}(\rho hW + R) = \alpha\Phi_{XT} + \beta\Phi_{YT} + \gamma\Phi_{ZT} + \delta\Phi_T$$

where the coefficients α, β, γ are chosen to make the flow direction timelike, as in the steady state, and controls the damping.

The complete numerical scheme thus calls for the following steps:

1. Calculate the contravariant velocity components and the density in each primary cell using the box scheme.
2. Calculate the flux balance on each secondary cell by a second application of the box scheme.
3. Add compensation terms to offset the effect of lumping errors.
4. Add artificial viscosity at points where the flow is locally supersonic to desymmetrize the scheme and enforce the entropy condition.
5. Add time dependent terms to embed the steady state equation in a convergent time dependent process which evolves to the solution.

8 Results

The finite volume scheme has been used in a number of calculations for swept wings and wing-cylinder combinations, and some results of these calculations are included in this section.[1] The scheme must be provided with the Cartesian coor-

[1]We would like to thank Frances Bauer for her valuable help in performing many of the numerical computations and obtaining the graphical output.

dinates of each mesh point. The meshes for our calculations have been generated by sequences of global mappings. This has the advantage of producing a smooth distribution of mesh points. In contrast with earlier methods in which the equation of motion was explicitly transformed, 4-6 these mappings are now used only to calculate the coordinates of the mesh points.

The following procedure has been used to generate the mesh for a swept wing. First we introduce parabolic coordinates in planes containing the wing section by the transformation

$$\bar{X} + i\bar{Y} = \left\{ \frac{\{x - x_0(z) + i(y - y_0(z))\}}{t(z)} \right\}^{\frac{1}{2}}$$

$$\bar{Z} = z$$

where z is the spanwise coordinate, $x_0(z)$ and $y_0(z)$ define a singular line located just inside the leading edge, and $t(z)$ is a scaling factor which can be adjusted so that the wing chord is covered by the same number of cells at every span station.

The effect of this transformation is to unwrap the wing to form a shallow bump

$$\bar{Y} = S(\bar{X}, \bar{Z})$$

as illustrated in Figure 7. Then we use a shearing transformation

$$X = \bar{X}, \quad Y = \bar{Y} - S(\bar{X}, \bar{Z}), \quad Z = \bar{Z}$$

to map the wing surface to the plane $Y = 0$. We now lay down a rectangular coordinate system in the X, Y, Z space, and finally generate the volume elements by the reverse sequence of transformations from X, Y, Z to x, y, z. The vortex sheet trailing behind the wing is assumed to coincide with the cut generated by the sheared parabolic coordinate system.

The mesh for the wing-cylinder calculations has been generated by a simple extension of this procedure, in which the cylinder is mapped to a vertical slit by a preliminary Joukowsky transformation, as sketched in Figure 8. With the fuselage

thus compressed into the symmetry plane, we then use the same sequence of mappings as for a swept wing on a wall. The use of a vertical slit rather than a horizontal slit, as was used by Newman and Klunker for small disturbance calculations,[12] allows the wing to be shifted vertically so that both low and high

wing configurations can be treated.

Figure 9 shows the result of a calculation for the ONERA M6 wing, for which experimental data is available.[13] The calculation was performed on a sequence of meshes. After the calculation on each of the first two meshes, the number of intervals was doubled in each coordinate direction, and the interpolated result was used as the starting point for the calculation on the next mesh. The fine mesh contained 160 intervals in the chordwise x direction, 16 intervals in the normal y direction, and 32 intervals in the spanwise z direction, for a totally of 81920 cells. 100 relaxation cycles were used on each mesh. Such a calculation requires about 90 minutes on a CDC 6600 or 20 minutes on a CDC 7600. Separate pressure distributions are shown for stations at 20, 45, 65 and 95 percent of the semi-span. The pressure coefficient at which the speed is locally sonic is marked by a horizontal line on the pressure axis, and the experimental data is overplotted on the numerical result, using circles for the upper surface and squares for the lower surface. The calculation did not include a boundary layer correction. It can be seen, however, that the triangular shock pattern is quite well captured, and that the calculated pressure distribution is a fair simulation of the experimental result. The result of this calculation is also in quite good agreement with the result of a previous calculation using the nonconservative rotated difference scheme.[14]

Figure 10 shows the result for the same wing mounted on a low and position on a cylinder. The configuration is scaled so that the radius of the cylinder is 0.25, while the wing tip station is 1.25. No experimental data is available in this case. The calculation shows an increase of lift, particularly near the wing root. This is to be expected, because the cylinder is set at the same angle of attack as the wing and will generate an upwash. The problem of computing the flow past a wing-fuselage combination is discussed at greater length in a companion paper,[15] in which an alternative mesh generating scheme is proposed.

9 Conclusion

The results displayed in Figures 9 and 10 serve to indicate the promise of the finite volume scheme. Its main advantage is the relative ease with which it can be adapted to treat a variety of complex configurations. Since the treatment of interior points is independent of the particular mappings used to generate the mesh, topologically similar configurations can be treated by the same flow computation

routine, provided that suitable mappings can be found to map them to the same computational domain.

This flexibility is achieved at the expense of an increase in the amount of time required for the computations, unless a very large memory capacity is available, because of the need to perform a numerical inversion o the transformation matrix defining the local mapping in each cell. If the inverse transformation coefficients are not saved they must be recalculated at every cycle. In this form the scheme requires about 50 percent more time than the rotated difference scheme to treat a swept wing on an equal number of mesh points. It is worth noting that the computing time could be substantially reduced by restricting the use of distorted cells to an inner region surrounding the body, with a transition to Cartesian coordinates in the outer region.

References

[1] Murman, E.M. and Cole, J.D., Calculation of plane steady transonic flows, AIAA Jour., Vol. 9, 1971, pp. 114-121.

[2] Murman, E.M., Analysis of embedded shock waves calculated by relaxation methods, Proceedings of AIAA Conf. on Computational Fluid Dynamics, Palm Springs, July 1973, pp. 27-40.

[3] Bailey, F.R. and Ballhaus, W.F., Relxation methods for transonic flows about wing-cylinder combinations and lifting swept wings, Proceedings of Third International Conference on Numerical Methods in Fluid Dynamics, Paris, July 1972, Lecture Notes in Physics, Vol. 19, Springer Verlag, 1973, pp.2-9.

[4] Jameson, Antony, Iterative solution of transonic flows over airfoils and wings, including flows at Mach 1, Comm. Pure Appl. Math., Vol. 27, 1974, pp. 283-309.

[5] Jameson, Antony, Numerical solution of nonlinear partial differential equations of mixed type, Numerical Solution of Partial Differential Equations III, SYNSPADE 1975, Academic Press, 1976, pp. 275-320.

[6] Jameson, Antony, Numerical computation of transonic flows with shock waves, Symposium Transsonicum II, Gottingen, September 1975, Springer Verlag, 1976, pp. 384-414.

[7] MacCormack, R.W., Rizzi, A.W. and Inouye, M., Steady supersonic flow fields with embedded subsonic regions, Proceedings of Conference on Computational Problems and Methods in Aero and Fluid Dynamics, Manchester, 1974.

[8] Rizzi, Arthur, Transonic solutions of the Euler equations by the finite volume method, Symposium Transsonicum II, Gottingen, September 1975, Springer Verlag, 1976, pp. 567-574.

[9] Bateman, H., Notes on a differential equation which occurs in the two dimensional motion of a compressible fluid and the associated variational problem, Proc. Roy. Soc. Series A, Vol. 125, 1929, pp. 598-618.

[10] Lax, Peter and Wndroff, Burton, Systems of conservatiokn laws, Comm. Pure Appl. Math., Vol. 13, 1960, pp. 217-237.

[11] Synge, J.L., and Schild, A., Tensor Calculus, Unviersity of Toronto Press, 1949, pp. 57-58.

[12] Newman, Perry A. and Klunker, E.B., Numerical modeling of tunnel wall and body shape effects on transonic flow over finite lifting wings, Aerodynamic Analyses Requiring Advanced Computers, Part 2, NASA, SP-347, 1975, pp. 1189-1212.

[13] Monnerie, B., and Charpin, F., Essais de buffeting d'une aile en fleche en trnassonique, 10^{e} Colloque d'Aerody-namique Appliquee, Lille, November 1973.

[14] Jameson, Antony and Caughey, D.A., Numerical calculation of the transonic flow past a swept wing, New York University ERDA Report COO 3077-140, May 1977.

[15] Caughey, D.A. and Jameson, Antony, Numerical calculation of transonic potential flow about wing-fuselage combiantions, AIAA Paper 77-677, June 1977.