ACCELERATION OF CONVERGENCE BY SHIFTING THE SPECTRUM OF IMPLICIT FINITE DIFFERENCE OPERATORS ASSOCIATED WITH THE EQUATIONS OF GAS DYNAMICS

A. CHEER

Department of Mathematics, University of California, Davis, CA, U.S.A.

AND

M. SALEEM

Department of Mathematics, University of Missouri, St Louis, MO, U.S.A.

SUMMARY

Eigensystem analysis techniques are applied to finite difference formulations of the Navier–Stokes equations in one dimension. Spectra of the resulting implicit difference operators are computed. The largest eigenvalues are calculated by using a combination of the Frechet derivative of the operators and Arnoldi's method. The accuracy of Arnoldi's method is tested by comparing the rate of convergence of the iterative method with the dominant eigenvalue of the original iteration matrix.

On the basis of the pattern of eigenvalue distributions for various flow configurations, a *shifting* of the implicit operators in question is devised. The idea of shifting is based on the power method of linear algebra and is very simple to implement. This procedure has improved the rates of convergence of CFD codes (developed at NASA Ames Research Center) by 20%-50%. The sensitivity of the computed solution with respect to the *shift* is also studied. Finally, an adaptive shifting of the spectrum together with Wynn's acceleration algorithm are discussed. It turns out that the shifting process is a preconditioner for Wynn's method.

KEY WORDS Acceleration of convergence Steady state solution Finite difference Eigensystem analysis Shifting of the spectrum Euler equations Iterative method Frechet derivative Eigenvalue annihilation Rates of convergence

1. INTRODUCTION

It is well known that most problems of interest in the field of aerodynamics, especially in three dimensions, are presently unsolvable even when the existing analytical, computational and experimental tools are available. In spite of this difficulty, one favourite option is to solve the Navier–Stokes equations on a computer. A lot of work is being done in the development of efficient computer codes which simulate real physical flows around arbitrary aircraft wing geometries. The purpose of such algorithms is to obtain accurate solutions to the Navier–Stokes equations using the least possible computer time. This problem of 'accelerating' a given flow calculation has attracted many mathematicians and engineers. This has resulted in the development of numerous methods, some of which have become standard in the realm of computational

0271-2091/91/050443-20\$10.00 © 1991 by John Wiley & Sons, Ltd. Received 13 October 1989 Revised 29 March 1990 fluid dynamics. Multigrid methods, preconditioning of matrices and eigenvector annihilation are only a few techniques to be mentioned in this context. A new approach is discussed here.

This paper presents the problem of attaining accelerated convergence to the steady-state solution of the Navier-Stokes equations in one dimension. Discretization leads to an iterative scheme whose convergence is controlled by the eigenvalues of large, time-dependent matrices. These spectra are computed by using Arnoldi's method. The pattern of eigenvalues for different flow patterns suggests the possibility of various acceleration methods. The one presented here is based on *shifting the spectrum* of the operators in question, thus forcing the spectral radius to become smaller. This process speeds up the convergence of the iterative method. The results obtained are compared with the method of explicit eigenvalue annihilation. The shifting strategy is found to be superior. The amount of *shift* needed for a specific case depends on the ends of the spectrum, which is available from Arnoldi's method. Moreover, the rate of convergence is dependent on the shift factor. This relationship is obtained graphically by running different cases with varying flow parameters and is then compared with the theoretical estimates.

The homogeneous property of the flux Jacobians proved valuable in showing that the shifting procedure is justified. The analysis as carried out for the Euler equations is also applicable to viscous flows.

2. EQUATIONS OF FLOW AND THE ITERATIVE SCHEME

Finite difference formulation of the flow equations

The Navier-Stokes equation for flow in an nozzle is¹

$$\frac{\partial \mathbf{Q}}{\partial t} + \frac{1}{a} \frac{\partial (a\mathbf{F})}{\partial x} = \mathbf{S} + \frac{1}{Re} \frac{\partial \mathbf{F}_{\mathbf{v}}}{\partial x},\tag{1}$$

where Q is the vector of conserved variables, F is the flux vector, F_v is the viscous flux and S is the source term, which contains the information about area ratios of the nozzle. These vectors are given by

$$\mathbf{Q} = \begin{bmatrix} \rho \\ \rho u \\ e \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ u(e+p) \end{bmatrix}, \qquad \mathbf{S} = \begin{bmatrix} 0 \\ (p/a) \partial a/\partial x \\ 0 \end{bmatrix}, \qquad \mathbf{F}_{\mathbf{v}} = \begin{bmatrix} 0 \\ \tau_{\mathbf{xx}} \\ f \end{bmatrix}.$$

The equation of state is

$$p = (\gamma - 1)(e - \frac{1}{2}\rho u^2).$$

Viscous quantities are given by

$$\tau_{xx} = \frac{4}{3}\mu u_x, \qquad f = u\tau_{xx} - \frac{\mu}{\gamma - 1} (Pr)^{-1} \frac{\partial}{\partial x} c^2.$$

The description of the various symbols is as follows:

- ρ density
- u velocity
- e total fluid energy per unit volume
- p pressure
- y ratio of specific heats, equal to 1.4

- a cross-sectional area of the nozzle, equal to $1 4x(1-x)(1-a_{throat})$, where a_{throat} is the area at the throat; a_{throat} was fixed at 0.8
- Pr Prandtl number
- Re Reynolds number
- c speed of sound
- μ dynamic viscosity of the fluid.

Equation (1) is discretized in time by using a first-order implicit (backward Euler) approximation.² In this paper we consider the inviscid case, for which $F_v = 0$. This results in

$$\frac{\mathbf{Q}^{N+1}-\mathbf{Q}^{N}}{h} + \frac{1}{a}\frac{\partial}{\partial x}(a\mathbf{F}^{N+1}) = \mathbf{S}^{N+1},$$
(2)

where $h = \Delta t$ is the time step and $t = N\Delta t$.

The flux vector \mathbf{F} is a non-linear function of \mathbf{Q} , so that, by the chain rule of differentiation,

$$\frac{\partial \mathbf{F}}{\partial x} = \frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\mathbf{Q}} \frac{\partial \mathbf{Q}}{\partial x} = \mathbf{A} \frac{\partial \mathbf{Q}}{\partial x},$$

where A = dF/dQ is the flux Jacobian. In the case of the Euler equations for a nozzle with a uniform grid it can be shown that F = AQ, where the form of A can be found in Reference 3.

In equation (2) the non-linear terms can be linearized about \mathbf{Q}^{N} by using a Taylor series. Thus

$$\mathbf{F}^{N+1} = \mathbf{F}^{N} + \mathbf{A}^{N} (\mathbf{Q}^{N+1} - \mathbf{Q}^{N}) + O(\|\Delta \mathbf{Q}^{N}\|^{2}),$$
(3a)

$$\mathbf{S}^{N+1} = \mathbf{S}^{N} + \mathbf{B}^{N}(\mathbf{Q}^{N+1} - \mathbf{Q}^{N}) + O(\|\Delta \mathbf{Q}^{N}\|^{2}),$$
(3b)

where $\mathbf{B}^N = d\mathbf{S}^N/d\mathbf{Q}^N$. Using equation (2) with equations (3) together with the definition $\Delta \mathbf{Q}^N = \mathbf{Q}^{N+1} - \mathbf{Q}^N$, it follows that

$$\frac{\Delta \mathbf{Q}^{N}}{h} + \frac{1}{a} \frac{\partial}{\partial x} [a(\mathbf{F}^{N} + \mathbf{A}^{N} \Delta \mathbf{Q}^{N})] = \mathbf{S}^{N} + \mathbf{B}^{N} \Delta \mathbf{Q}^{N}$$

or

$$\left(\mathbf{I} + \frac{h}{a\partial x}(a\mathbf{A}^{N}) - h\mathbf{B}^{N}\right)\Delta\mathbf{Q}^{N} = h\left(\mathbf{S}^{N} - \frac{1}{a\partial x}(a\mathbf{F}^{N})\right).$$
(4)

The spatial derivatives appearing in equation (4) are approximated by second-order central differences. This produces a matrix operator which is block tridiagonal. For the case of a nozzle with 100 grid points this gives rise to a 300×300 matrix.

Boundary conditions for the nozzle. There are three dependent variables $(\rho, u \text{ and } p)$ and therefore three boundary conditions must be specified at the entrance/exit. The type of boundary conditions can be determined by examining the characteristic paths¹ at the boundaries. Typically, the values of ρ and u are given, together with the inlet area and the exit pressure to which the gas must be expanded. Linear extrapolated characteristic boundary conditions obtained thus are shown in Figure 1 of Section 3. Dirichlet boundary conditions are taken to be the values obtained from the analytic solution, as in Reference 1.

Artificial dissipation added to the scheme. The implicit algorithm obtained in (4) encounters certain stability bounds, even though linear stability analysis for implicit methods in general shows unconditional stability. This is especially true in strongly non-linear cases such as flows

A. CHEER AND M. SALEEM

with shocks.⁴ In the nozzle code used here¹⁶ a time step of h=0.3 with linear extrapolated characteristic boundary conditions gave rise to instability in the solution, and the residual started to increase after 200 iterations.

Central differencing in space has an intrinsic property of introducing oscillations in the solution.⁶ The most common way of damping out spurious oscillations is to add to the complete algorithm some form of numerical dissipation with an error level that does not interfere with the accuracy of any physical viscous effects. This can be done by adding implicit or explicit dissipation. Moreover, this dissipation can be linear or non-linear. In the nozzle code a combination of second- and fourth-order dissipations is used. These are added explicitly for the RHS of equation (4) and implicitly for the block tridiagonal matrix. The exact model is explained in Reference 4. Second-order implicit dissipation stabilizes the algorithm and allows us to retain block tridiagonal inversions. The use of fourth-order dissipation matching the explicit terms produces larger stability bounds and enhances convergence. This will be reflected when the spectrum of the nozzle problem is obtained.

Iterative solution process

Once the boundary conditions, the fluxes, the dissipation terms, the RHS of the algorithm and the finite differences have been employed, the delta form (equation (4)) of the nozzle problem condenses to

$$\mathbf{W}^{N}(\Delta \mathbf{Q}^{N}) = \mathbf{R}^{N}.$$
(4a)

Here W is the matrix containing implicit smoothing, fluxes and boundary conditions, \mathbb{R}^N is the RHS with the source term, the viscous fluxes and any explicit dissipation terms, and \mathbb{Q}^N is the solution vector containing the conserved variables at each grid point, given by

$$\mathbf{Q}^{N} = [\rho_{1}, \rho u_{1}, e_{1}, \rho_{2}, \rho u_{2}, e_{2}, \dots, \rho_{J}, \rho u_{J}, e_{J}]^{T},$$

where J is the number of grid points used. Inversion leads to

$$\Delta \mathbf{Q}^N = (\mathbf{W}^N)^{-1} \mathbf{R}^N,$$

whence the solution \mathbf{Q}^{N+1} can be extracted as

$$\mathbf{Q}^{N+1} = \mathbf{Q}^N + (\mathbf{W}^N)^{-1} \mathbf{R}^N.$$
⁽⁵⁾

Now, since $(\mathbf{W}^N)^{-1}$ and \mathbf{R}^N both depend upon \mathbf{Q}^N , equation (5) can be written as

$$\mathbf{Q}^{N+1} = \mathbf{L}(\mathbf{Q}^N),\tag{6}$$

where, of course, $\mathbf{L}(\mathbf{Q}^N) = \mathbf{Q}^N + (\mathbf{W}^N)^{-1} \mathbf{R}^N$.

Equation (6) can be thought of as an operator equation. A study of its spectrum and convergence properties will be carried out in order to accelerate its convergence. When the nozzle algorithm is put in the form of equation (6), i.e. $\mathbf{Q}^{N+1} = \mathbf{L}(\mathbf{Q}^N)$, an application of \mathbf{L} on \mathbf{Q} means that the following steps have been performed:

- (a) setting up the initial conditions/data for the flow
- (b) setting up the boundary conditions
- (c) filling the block tridiagonal matrix and adding the implicit dissipation if required
- (d) computing the RHS including the source term, viscous terms and explicit dissipations terms
- (e) inverting the block tridiagonal matrix to extract the solution vector \mathbf{Q}^{N+1} .

The procedure (a)–(e) is, incidentally, one iteration of the nozzle code. Thus the phrases 'iteration of the code' and 'application of the operator L' will be used interchangeably. Furthermore, the term 'nozzle operator' will mean L, exceptions being notified in the context.

Spectrum of the nozzle operator

Once the nozzle algorithm is written in the iterative form

$$\mathbf{Q}^{N+1} = \mathbf{L}(\mathbf{Q}^N),\tag{7}$$

the eigenvalues of the operator L can be estimated as explained below. The approach presented here is fairly general, so that L can be any non-linear transformation and not necessarily the nozzle operator. Recall that the eigenvalues of a linear transformation T: $\mathscr{R}^M \to \mathscr{R}^M$ are actually the eigenvalues of its Jacobian, given by

$$\frac{\mathrm{d}\mathbf{T}}{\mathrm{d}x} = \frac{\partial(t_1, t_2, \ldots, t_m)}{\partial(x_1, x_2, \ldots, x_m)},$$

where

$$t_1 = t_1(x_1, x_2, \dots, x_m),$$

$$t_2 = t_2(x_1, x_2, \dots, x_m),$$

$$\vdots$$

$$t_m = t_m(x_1, x_2, \dots, x_m).$$

In this non-linear study it is necessary to introduce the corresponding Jacobian for L, which stems from the Taylor series of $L(Q^N)$ about some reference vector Q^* . One desirable choice for Q^* is the converged solution itself. Thus equation (7) becomes

$$\mathbf{Q}^{N+1} = \mathbf{L}(\mathbf{Q}^N) = \mathbf{L}(\mathbf{Q}^*) + \frac{d\mathbf{L}}{d\mathbf{Q}^*}(\mathbf{Q}^N - \mathbf{Q}^*) + O(||\mathbf{Q}^N - \mathbf{Q}^*||^2).$$

Notice that the terms of the order of $||Q^N - Q^*||^2$ can be neglected when Q^N approaches Q^* to a sufficient accuracy. This is particularly true near convergence. Thus

$$\mathbf{Q}^{N+1} = \mathbf{L}(\mathbf{Q}^*) + \frac{\mathrm{d}\mathbf{L}}{\mathrm{d}\mathbf{Q}^*}(\mathbf{Q}^N - \mathbf{Q}^*)$$
$$= \frac{\mathrm{d}\mathbf{L}}{\mathrm{d}\mathbf{Q}^*} \mathbf{Q}^N + \mathbf{L}(\mathbf{Q}^*) - \frac{\mathrm{d}\mathbf{L}}{\mathrm{d}\mathbf{Q}^*} \mathbf{Q}^*$$
$$= \mathbf{\mathring{A}} \mathbf{Q}^N + \mathbf{T}(\mathbf{Q}^*), \tag{8}$$

where $\dot{A} = dL/dQ^*$ and $T(Q^*) = L(Q^*) - (dL/dQ^*)Q^*$. The matrix \dot{A} defined here is not to be confused with the flux Jacobian A. From equation (8) it can be seen that

$$(\mathbf{Q}^{N+1}-\mathbf{Q}^N) = \mathbf{\dot{A}}(\mathbf{Q}^N-\mathbf{Q}^{N-1}), \qquad (9a)$$

$$(\mathbf{Q}^{N+1} - \mathbf{Q}^*) = \mathbf{\dot{A}}(\mathbf{Q}^N - \mathbf{Q}^*).$$
(9b)

If $\mathbf{d}_{n+1} = \mathbf{Q}^{N+1} - \mathbf{Q}^N$ denotes the successive differences and $\mathbf{e}_{n+1} = \mathbf{Q}^{N+1} - \mathbf{Q}^*$ denotes the error in the computed solution at time level n+1, then it follows from equations (9) that

$$\mathbf{d}_{n+1} = \mathbf{\mathring{A}} \mathbf{d}_n, \qquad \mathbf{e}_{n+1} = \mathbf{\mathring{A}} \mathbf{e}_n. \tag{10}$$

Equations (10) attach a vital importance to the matrix $\mathbf{\dot{A}}$ because of the fact that the successive differences and errors at any stage propagate according to the powers of $\mathbf{\dot{A}}$.

It can be easily shown that the residual of the solution will also satisfy an equation similar to equation (10). Consequently the spectral radius of the matrix Å controls the convergence of the nozzle problem. The eigensystem of the matrix Å will be studied in detail.

If a grid of 100 points is imposed on the nozzle, then the solution vector \mathbf{Q} has a dimension of 300. In this case the Jacobian $\mathbf{\mathring{A}} = d\mathbf{L}/d\mathbf{Q}^*$ is a 300 × 300 matrix with 90 000 entries. Computation and storage of its elements is not entirely impossible, but still, an efficient and economic method to obtain its eigenvalues must be adopted. The method used here is due to Arnoldi.⁷ Its algorithm to compute the dominant eigenvalues of $\mathbf{\mathring{A}}$ is described below.

For an arbitrary vector \mathbf{q}_1 , define

$$\hat{\mathbf{q}}_1 = \frac{\mathbf{q}_1}{\|\mathbf{q}_1\|}.$$

The algorithm then consists of the following steps.

For k = 1 to m, do

$$\mathbf{q}_{k+1} = \mathbf{\mathring{A}} \, \mathbf{\widehat{q}}_k - \sum_{j=1}^k c_{jk} \, \mathbf{\widehat{q}}_j,$$
$$c_{jk} = \langle \mathbf{\widehat{q}}_j, \, \mathbf{\mathring{A}} \, \mathbf{q}_k \rangle,$$

where $\langle x, y \rangle$ is the inner product of the vectors x and y,

$$\hat{\mathbf{q}}_{k+1} = \frac{\mathbf{q}_{k+1}}{\|\mathbf{q}_{k+1}\|};$$

next k.

The *m* eigenvalues of C, namely c_{ii} , are the approximations to 'some' *m* eigenvalues of Å.

Products such as Åv which are needed for arbitrary v can be achieved by using the idea of the Frechet derivative of L as

$$\mathbf{\mathring{A}v} = \frac{\mathrm{d}\mathbf{L}}{\mathrm{d}\mathbf{O^*}}\mathbf{v} = \frac{\mathbf{L}(\mathbf{Q^*} + \varepsilon\mathbf{v}) - \mathbf{L}(\mathbf{Q^*} - \varepsilon\mathbf{v})}{2\varepsilon} + O(\varepsilon^2).$$

Another differencing which requires double the computer time, is the fourth-order Frechet derivative:

$$\mathbf{\mathring{A}v} = \frac{\mathrm{d}\mathbf{L}}{\mathrm{d}\mathbf{Q^*}} \mathbf{v} = \frac{1}{12\varepsilon} \left\{ -\mathbf{L}(\mathbf{Q^*} + 2\varepsilon\mathbf{v}) + 8\mathbf{L}(\mathbf{Q^*} + \varepsilon\mathbf{v}) - 8\mathbf{L}(\mathbf{Q^*} - \varepsilon\mathbf{v}) + \mathbf{L}(\mathbf{Q^*} - 2\varepsilon\mathbf{d}) \right\} + O(\varepsilon^4).$$

In order to retain a balance between accuracy and the cost of running the code, a second-order differencing was adopted. It was seen that a second-order differencing was optimal with ε chosen from the requirement

$$\varepsilon = 0.001 \times \frac{\|\mathbf{Q}^*\|}{\|\mathbf{v}\|},$$

where L_2 -norms were taken. This choice of epsilon is typical for Euler codes.⁸

The eigenvalues obtained by Arnoldi's method are good estimates to those of $\dot{A} = dL/dQ^*$ for the ends of the spectrum.⁹ This means that 'some' of the largest eigenvalues (in absolute value) computed by using Arnoldi's method will be good approximations to the actual ones.

In practice the eigenvalues obtained by Arnoldi's method are not necessarily the largest or the smallest ones of Å. If v, the starting vector, is an approximation to the dominant eigenvector (corresponding to the eigenvalue with largest modulus) of Å, then Arnoldi's method will tend to converge in that direction and return a set of eigenvalues including the dominant eigenvalue. If v happens to be 'some' linear combination of a few 'large' eigenvectors, then the spectrum of C (the upper Hessenberg matrix) will contain the corresponding 'large' eigenvalues. But the choice of v is

a fairly difficult one, since the eigenvectors of Å are not known in advance. Choices for starting vectors v are discussed in Reference 3. Typically, reasonably well-separated eigenvalues on the extremes of the spectrum (in this case the boundary of the unit circle) of Å will appear as eigenvalues of the reduced matrix C for relatively small values of *m*. Well-separated interior eigenvalues can converge as fast as or faster than the clustered extreme eigenvalues. The eigenvalues that converge slowest in a projection method like Arnoldi's are the ones that are tightly clustered in the middle of the spectrum.¹⁰

3. RESULTS FOR THE NOZZLE PROBLEM

In this section the results of computing the eigenvalues of the nozzle operator are presented. For this convergent-divergent nozzle of length 1.0, a fluid of density 1.0 enters at Mach 0.553. The variable cross-sectional area of the nozzle is given by $a(x) = 1 - 4x(1-x)(1-a_{throat})$, where $a_{throat} = 0.8$ is the area of the throat and $0 \le x \le 1$ (Figure 1). Thus the area at the ends is 1.0 whereas the area of the throat is 0.8. The shock location is fixed at x = 0.7. The various spectra depend on

- (a) grid size
- (b) boundary conditions—linear extrapolated characteristic BCs and Dirichlet BCs
- (c) time step
- (d) second and fourth-order dissipation terms
- (e) different phases of the convergence process.

Case studies

The eigenvalues of the complete Jacobian $\dot{A} = dL/dQ^*$ were computed for various cases and serve as the reference spectrum. In applying Arnoldi's method it is seen in general that the 'worst' eigenvalues are off in the second decimal place. The largest and the smallest ones (referred to as the 'ends' of the spectrum) are accurate to within 10^{-5} .

Grid

As j_{max} (the number of grid points) increases, the number of eigenvalues of large magnitude increases. For $j_{\text{max}} = 30$ there are 12 eigenvalues greater than 0.5. This number increased to 15 for



 $j_{max} = 100$. This indicates that for a fine grid, when using an acceleration method such as eigenvalue annihilation,¹¹ more eigenvalues must be annihilated in order to improve the rate of convergence substantially. With an increasing number of grid points it is evident from Figures 2–5 that the eigenvalues with large absolute values are well separated, making the nozzle code very suitable for an eigenvalue annihilation scheme. The eigenvalues of small magnitude are clustered near the origin. The eigenvalues with negative real parts (only one or two in each case) have been plotted such that they appear on the positive real axis.

Dissipation

In order to study the effects of second- and fourth-order dissipation on the spectrum, the code with linear extrapolated characteristic boundary conditions was run for 300 time steps until the residual fell below 4×10^{-7} . A solution at 300 iterations was thus saved. The dissipation terms were altered at this stage so the eigenvalues obtained correspond with this altered system of equations.

Without the second-order smoothing (EPS2) the spectrum did not change much. The eigenvalues became slightly enlarged but the largest ones were still distinct. In the inviscid case with dissipation there are 15 eigenvalues with modulus greater than 0.5, and $\lambda_{max} = 0.9693319$. With second-order dissipation turned off, the effect on λ_{max} is minimal and $\lambda_{max} = 0.9693434$. The larger end of the spectrum remains well separated with $|\lambda_1|_{EPS2=0} - |\lambda_2|_{EPS2=0} = 0.144434$. This is depicted in Figure 6.

The effect of removing the fourth-order dissipation (EPS4) was remarkable. The spectrum became 'inflated' and as such the smallest eigenvalues that were clustered together tended to become distinct as in Figure 7. The number of eigenvalues larger than 0.5 in magnitude increased

INVISCID NOZZLE, 100 GRID POINTS, 300 EIGENVALUES



Figure 2



INVISCID NOZZLE, 50 GRID POINTS, 150 EIGENVALUES LINEAR BCs, SMOOTHING EPS2 & EPS4, dt=0.2, 300 ITERATIONS ==eigenvalue -0.4475 shown with positive real part

INVISCID NOZZLE, 70 GRID POINTS, 210 EIGENVALUES LINEAR BCs, SMOOTHING EPS2 & EPS4, dt=0.2, 300 ITERATIONS •= eigenvalues - .643 & -0.2 shown with positive real part



Figure 4

A. CHEER AND M. SALEEM



INVISCID NOZZLE, 100 GRID POINTS, 300 EIGENVALUES DIRICHLET BCs, dt=0.2, 300 ITERATIONS

INVISCID NOZZLE, 100 GRID POINTS, 300 EIGENVALUES LINEAR BCs, WITHOUT EPS2 SMOOTHING, dt=0.2, 300 iterations



Figure 6



INVISCID NOZZLE, 100 GRID POINTS, 300 EIGENVALUES LINEAR BCs, WITHOUT EPS4 SMMOTHING dt=0.2, 500 ITERATIONS

from 15 to 39. With this alteration the largest eigenvalues occur in a complex conjugate pair, $0.9502 \pm i0.0045$, with $|\lambda_1|_{EPS4=0} - |\lambda_3|_{EPS4=0} = 0.1154$. This indicates that in order to accelerate the convergence for this case, the effects of at least two eigenvalues (the complex conjugate pair) must be corrected for if an eigenvalue annihilation procedure is to be used. Moreover, as the number of eigenvalues of large magnitude increases, one must annihilate the contribution of more eigenvalues to gain the same rate of convergence.

An experiment with the dissipation turned off entirely was performed. With the solution Q^{300} saved at the 300th iteration, the eigenvalues of the nozzle operator were computed. Two eigenvalues with modulus greater than 1.0 were detected; see Figures 8 and 9. This was done with both types of boundary conditions. The total number of eigenvalues larger than 0.5 in magnitude increased to 42, out of which 37 were such that $|\lambda| \ge 0.6$. This experiment shows that the presence of a shock wave in the flow accounts for an inherent instability in the nozzle code, which is invisible because of the smoothing terms. Notice that this instability occurs in spite of the fully implicit construction of the code, even when a converged solution was used to restart the iterations.

4. ACCELERATION METHODS

This section deals with some existing techniques for the acceleration of convergence of iterative schemes. It gives a brief picture of *eigenvalue annihilation* applied to the nozzle code. The resulting improvement in the rate of convergence by this method is obtained, but the underlying difficulty of the exact computation of a few large eigenvalues is well known. Powerful methods such as Wynn's ε -algorithm are briefly mentioned. Finally, a new method based on *shifting the spectrum* of the implicit operators is derived. This method, in spite of its simple implementation, has shown good results with considerable savings in computer time.



INVISCID NOZZLE, 100 GRID POINTS, 300 EIGENVALUES LINEAR BCs, SMOOTHING EPS2=0.0, EPS4=0.0, dt=0.2, 300 ITERATIONS

INVISCID NOZZLE, 100 GRID POINTS, 300 EIGENVALUES DIRICHLET BCs, SMOOTHING EPS2=0.0,EPS4=0.0, dt=0.2, 300 ITERATIONS



Figure 9

Some eigenvalue annihilation techniques

In the one-dimensional problem of flow through a nozzle the largest few eigenvalues were discretely distributed. This phenomenon has been depicted in Figures 2–9. In such a case eigenvalue annihilation techniques based on Aitken's or Shanks' transformation¹² will work effectively. Explicit eigenvalue annihilation requires that a linear combination of two recent solutions be constructed as

$$\mathbf{Q} = \frac{\mathbf{Q}^{N+1} - \lambda \mathbf{Q}^{N}}{1 - \lambda}, \quad \text{where } \lambda = \lambda_{\max}.$$
(11)

This improved solution Q clearly converges according to the second largest eigenvalue and not as λ_{max} . In the nozzle code with Dirichlet boundary conditions, 50 grid points, $\Delta t = 0.25$, $\lambda_{max} = \lambda_1 = 0.927$ and $\lambda_2 = 0.79$, annihilation was tested. If λ_1 is annihilated, the residual starts falling as λ_2 , which is considerably faster. In this test the residual at 200 iterations is compared for three cases.

- (a) Without annihilation the residual is 0.381×10^{-8} .
- (b) With λ_1 annihilated the residual is less than 1.0×10^{-12} .
- (c) With λ_1 annihilated first and λ_2 annihilated after another 30 iterations the residual fell below 1.0×10^{-16} .

However, the rate of convergence of the method will eventually revert to λ_1 because of non-linear effects.

The idea of group annihilation is that the error associated with a number of eigenvalues is corrected for at the same time. This is particularly useful when the dominant eigenvalue occurs in a complex conjugate pair and both of them have to be annihilated at the same time. This idea is implemented in Reference 5. If more than two eigenvalues must be annihilated at some stage, existing techniques such as Wynn's ε -algorithm ¹³⁻¹⁵ can be used. It should be emphasized again that an exact estimate of eigenvalues to be annihilated is usually not available in advance. Wynn's ε -algorithm assumes an error distribution based on the largest p eigenvalues and annihilates all p of them. It requires storage of the solution vector at the previous p+2 iteration levels and some simple algebra to deduce an improved solution. This method has been applied to the nozzle problem¹⁶ to produce very good results. A method based on shifting the spectrum of Å is presented in Section 5.

5. SHIFTING THE SPECTRUM OF THE NOZZLE OPERATOR

The delta form of the nozzle algorithm as given in equation (4a) is

$$\mathbf{W}^{N} \Delta \mathbf{Q}^{N} = \mathbf{R}^{N}, \tag{12}$$

where the matrix \mathbf{W}^N contains the effect of flux Jacobians, time step, area ratios, geometry of the nozzle, linear and non-linear dissipations.

In the previous operator formulation of equation (12) it was shown that the solution can be obtained from the non-linear iterative process

$$\mathbf{Q}^{N+1} = \mathbf{L}(\mathbf{Q}^N),\tag{13}$$

where, in conjunction with equation (12), L was given by

$$\mathbf{L}(\mathbf{Q}^N) = \mathbf{Q}^N + (\mathbf{W}^N)^{-1} \mathbf{R}^N.$$

For the sake of simplicity in further analysis we assume that the cross-sectional area of the nozzle,

a(x), is constant. Thus in equation (12) W is given by (equation 4)

$$W = I + h\partial_x \frac{dF}{dQ} - h\frac{dS}{dQ}$$
$$= I + G,$$

where

$$\mathbf{G} = h \partial_x \frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\mathbf{Q}} - h \frac{\mathrm{d}\mathbf{S}}{\mathrm{d}\mathbf{Q}},$$

F is the flux vector, **S** is the source term for the nozzle, as defined in equation (1), and $\mathbf{R} = h(\mathbf{S} - \partial_x \mathbf{F})$. Noticing that **F** and **S** are both homogeneous of degree one with respect to the solution vector **Q**, the equations

$$\mathbf{F} = \frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\mathbf{Q}}\mathbf{Q} \quad \text{and} \quad \mathbf{S} = \frac{\mathrm{d}\mathbf{S}}{\mathrm{d}\mathbf{Q}}\mathbf{Q}$$

can be exploited to establish that

$$\mathbf{R} = \frac{\mathrm{d}\mathbf{R}}{\mathrm{d}\mathbf{Q}}\mathbf{Q}$$
$$= -\frac{\mathrm{d}}{\mathrm{d}\mathbf{Q}}(h\partial_x\mathbf{F} - h\mathbf{S})\mathbf{Q}$$
$$= \left(-h\partial_x\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\mathbf{Q}} + h\frac{\mathrm{d}\mathbf{S}}{\mathrm{d}\mathbf{Q}}\right)\mathbf{Q}$$
$$= -\mathbf{G}\mathbf{Q}.$$

Thus equation (12) can be written as

 $(\mathbf{I}+\mathbf{G})^N \Delta \mathbf{Q}^N = -\mathbf{G}^N \mathbf{Q}^N$

and

$$\mathbf{Q}^{N+1} = \mathbf{Q}^{N} - (\mathbf{I} + \mathbf{G}^{N})^{-1} \mathbf{G}^{N} \mathbf{Q}^{N}$$
$$= [\mathbf{I} - (\mathbf{I} + \mathbf{G}^{N})^{-1} \mathbf{G}^{N}] \mathbf{Q}^{N}$$
$$= (\mathbf{I} + \mathbf{G}^{N})^{-1} \mathbf{Q}^{N}, \qquad (14)$$

which is the unshifted algorithm, sometimes referred to as the generic algorithm. In deriving equation (14), the homogeneous property of the fluxes plays a very crucial role.

According to the power method, a shift sI should be introduced in the iteration matrix $I - (G^{N} + I)^{-1}G^{N}$. Since $(I + G^{N})^{-1}$ appearing therein is not known, such a shift is virtually impossible. The remedy is to introduce a shift in the matrix W, the implicit operator of equation (12). With this understanding, equation (12) becomes

$$(\mathbf{W}^N + s\mathbf{I})^N \Delta \mathbf{Q}^N = \mathbf{R}^N$$

and hence the corresponding equation (14) becomes

$$\mathbf{Q}^{N+1} = [\mathbf{I} - (\mathbf{G}^N + \mathbf{I} + s\mathbf{I})^{-1} \mathbf{G}^N] \mathbf{Q}^N$$
$$= \left(\mathbf{I} + \frac{1}{1+s} \mathbf{G}^N\right)^{-1} \mathbf{Q}^N.$$
(15)

Thus the eigenvalues of the iteration matrix in equation (14) are given by

$$\lambda = \frac{1}{1+g_i},$$

where g_i are the eigenvalues of the matrix \mathbf{G}^N . Similarly, the eigenvalues of the shifted iteration, equation (15), are given by

$$\Lambda = \frac{1}{1 + g_i / (1 + s)} = \frac{1 + s}{1 + s + g_i}$$

The shift s must be chosen as some linear combination of the eigenvalues of the iteration scheme. Moreover, it should be such that Λ , which controls the convergence of the shifted iteration, is less in magnitude than λ_{max} , which corresponds to the generic iteration scheme given in equation (14).

One simple value of s is $-(\lambda_{\max} + \lambda_{\min})/2$, where λ_{\max} and λ_{\min} are respectively the rightmost and leftmost eigenvalues of equation (14) and can be calculated in advance by using Arnoldi's method. Thus

$$\Lambda_{\max} = \frac{1 - \frac{1}{2}(\lambda_{\max} + \lambda_{\min})}{1 - \frac{1}{2}(\lambda_{\max} + \lambda_{\min}) + g}$$

where Λ_{\max} is attained for some $g = g_i$. Using the approximation $\lambda_{\max} = 1/(1+g)$, it follows that $g = (1/\lambda_{\max}) - 1$; thus

$$\Lambda_{\max} = \frac{1 - \frac{1}{2}(\lambda_{\max} + \lambda_{\min})}{1 - \frac{1}{2}(\lambda_{\max} + \lambda_{\min}) + (1/\lambda_{\max}) - 1}.$$
 (16)

This equation restricts the improvement in the rate of convergence. Λ_{max} will attain some definite value once the ends of the spectrum of the unshifted iteration are given. For example, if λ_{max} and λ_{min} are real and are symmetrically located about the origin, then $\lambda_{max} + \lambda_{min} = 0$. Consequently, from equation (16) $\Lambda_{max} = \lambda_{max}$, which means that there will be no improvement in the rate of convergence for such a specific case. This is expected already, since an attempt to shift an interval (-a, a) will only destroy its symmetry.

The amount of improvement by shifting is $\lambda_{max} - \Lambda_{max}$ and can be calculated easily from equation (16):

$$\lambda_{\max} - \Lambda_{\max} = \frac{(1 - \lambda_{\max})\lambda_{\max}(\lambda_{\max} + \lambda_{\min})}{2 - \lambda_{\max}(\lambda_{\max} + \lambda_{\min})}.$$

Once again it is easy to see that this difference is zero if $\lambda_{max} = -\lambda_{min}$. Other possible breakdowns are $\lambda_{max} = 1.0$ (which means an unstable iterative scheme) and $\lambda_{max} = 0$, for which case a positive shift s > 0 must be introduced. In all other cases it is clear that $\lambda_{max} - \Lambda_{max}$ will be some positive quantity as long as λ_{max} and λ_{min} are inside the unit circle. Thus there always exists a shift s which will force Λ_{max} to be less than or equal to λ_{max} . The optimal shift will therefore make the ends of the spectrum symmetric with respect to the origin.

It should be clear that the shift described here is not a linear translation as in the conventional linear algebraic approach. This means that each eigenvalue λ_i of the original process (12) is not shifted (translated) by $s = -(\lambda_{\max} + \lambda_{\min})/2$ to yield $\lambda_i - s$ as was originally intended. The shifting process adopted here defines a new and more favourable distribution of eigenvalues in which the largest eigenvalue is forced to become smaller in magnitude. For instance, if the generic iteration has $\lambda_{\max} = 0.5$ and $\lambda_{\min} = 0$, then a shift $s = -\frac{1}{2}(0.5) = \frac{1}{4}$ is chosen so that $\Lambda_{\max} = \frac{3}{7}$. Clearly, the difference in eigenvalues of the old and new iterations is $\lambda_{\max} - \Lambda_{\max} = \frac{1}{2} - \frac{3}{7} = \frac{1}{14}$, so there has been an improvement, but the amount of improvement is not $\frac{1}{4}$ as might have been intended.

For a nozzle with 50 grid points, a time step of 0.2 and linear characteristic boundary conditions, it has been calculated by Arnoldi's method that $\lambda_{max} = 0.969$ and $\lambda_{min} = -0.4475$, so the shift $s = -\frac{1}{2}(0.969 - 0.4475) = -0.26$. With this shift the new dominant eigenvalue Λ_{max} is 0.959. Through numerical experimentation a value s = -0.235 was found to be optimal. Explicit eigenvalue annihilation was also applied to the nozzle code. Theoretically a single annihilation step should improve the rate of convergence from $\lambda_{max} = 0.969$ to 0.838, which is the second-largest eigenvalue. In practice the rate of convergence reverts back to 0.969 during the course of iterations—a fact that has also been noticed by others.¹¹ The remedy is to repeat the annihilation process; but, in addition to the accuracy in computing the dominant eigenvalue, the procedure depends on when to apply the annihilation step and how often to repeat it. Using annihilation steps at 100, 150, 200, . . . iterations the nozzle code was run to 400 iterations. The CPU times on the VAX 11/780 for both the shift method and the annihilation method were found to be roughly the same. Table I compares the residuals at various iteration levels. Notice that the results are poor if λ_{max} was calculated to be 0.968 instead of 0.969.

Та	ble	

	Iterations					
	100	200	300	400	500	1000
Residual without shift Residual with shift Annihilating $\lambda = 0.969$ Annihilating $\lambda = 0.968$	$\begin{array}{c} 0.160 \times 10^{-3} \\ 0.301 \times 10^{-5} \\ 0.160 \times 10^{-3} \\ 0.160 \times 10^{-3} \end{array}$	$\begin{array}{c} 0.671 \times 10^{-5} \\ 0.107 \times 10^{-8} \\ 0.313 \times 10^{-8} \\ 0.934 \times 10^{-7} \end{array}$	$\begin{array}{c} 0.305 \times 10^{-6} \\ 0.377 \times 10^{-12} \\ 0.352 \times 10^{-12} \\ 0.197 \times 10^{-11} \end{array}$	$\begin{array}{c} 0.137 \times 10^{-7} \\ 0.92 \times 10^{-16} \\ 0.10 \times 10^{-16} \\ 0.21 \times 10^{-15} \end{array}$	0.618 × 10 ⁻⁹	0.108×10^{-15}





With this shifted version of the algorithm the spectrum of the nozzle operator was computed. At iteration number 200 the largest and smallest eigenvalues were found to be 0.926 and -0.922 respectively. Notice that the shifted spectrum has λ_{max} and λ_{min} symmetrically located about the origin as shown in Figure 10. Any further attempt to shift the spectrum will result in moving some parts of it closer to the boundary of the unit circle. On the basis of the data given above,

rate of convergence =
$$\left(\frac{\text{residual at } 200 \text{ iterations}}{\text{residual at } 100 \text{ iterations}}\right)^{1/100} = 0.9237.$$

Similar results were obtained by using Dirichlet BCs for a nozzle with 50 grid points with a time step of 0.25. The optimal value of the shifting parameter was -0.155, for which a comparison of residuals is given in Table II.

The shift reduced the largest eigenvalue from 0.940 to 0.877 and improved the rate of convergence considerably. The new spectrum is depicted in Figure 11. The symmetry of the largest

	Table II			
	Iterations			
	200	250	600	
Residual without shift Residual with shift	$\begin{array}{c} 0.381 \times 10^{-8} \\ 0.571 \times 10^{-13} \end{array}$	$0.861 \times 10^{-10} \\ 1.0 \times 10^{-16}$	1.0×10^{-16}	

INVISCID NOZZLE, 50 GRID POINTS, 150 EIGENVALUES DIRICHLET BCs, dt=0.25, 200 ITERATIONS, SHIFT=-0.155



Figure 11

eigenvalues is easy to observe. The effect of eigenvalue annihilation for this case has been considered in Section 4.

Finally, this procedure was carried out for a nozzle with 100 grid points and linear extrapolated characteristic BCs with a time step of 0.2. Table III summarizes the effect of shifting the iteration matrix by an amount s = -0.055.

For the unshifted case the largest and smallest eigenvalues were calculated to be 0.969 and -0.859 respectively. The spectrum of the shifted operator showed that the largest and smallest eigenvalues were 0.9601 and -0.9613 respectively. The fact that the largest eigenvalues are almost symmetrical w.r.t the origin suggests that any further shift is not possible. This optimal shift resulted in a saving of about 20% in the number of iterations. The corresponding spectrum is shown in Figure 12.

In this study the shift parameter was fixed for each individual flow through the nozzle. This parameter should, however, be large in the initial stages of iteration and settle down to its final value at convergence.

Table III						
	Iterations					
	100	200	300	400	500	
Residual without shift Residual with shift	$\begin{array}{c} 0.181 \times 10^{-3} \\ 0.435 \times 10^{-4} \end{array}$	0.856×10^{-6} 0.168×10^{-5}	0.380×10^{-6} 0.291×10^{-7}	0.169×10^{-7} 0.504×10^{-9}	$\begin{array}{c} 0.749 \times 10^{-9} \\ 0.850 \times 10^{-11} \end{array}$	

INVISCID NOZZLE, 100 GRID POINTS, 100 EIGENVALUES LINEAR BCs, dt=0.2, 500 ITERATIONS, SHIFT=-0.055



Figure 12



The eigenvalue λ_{max} which controls the rate of convergence of the iterative method can be calculated by Arnoldi's method or estimated by

rate of convergence =
$$\left(\frac{\text{residual at } N \text{ iterations}}{\text{residual at } M \text{ iterations}}\right)^{1/(N-M)}$$

For the algorithm presented in this paper the value of λ_{\min} which appears in the expression for s can be estimated by Arnoldi's method to be -0.5 for inviscid flow using a grid with 50 points and -0.85 using 100 grid points. With these λ_{\max} and λ_{\min} the shift parameter can be calculated at different iteration levels and the process applied adaptively.

Wynn's algorithm¹⁶ is very suitable in accelerating the convergence of the nozzle problem. Together with an adaptive shifting of the spectrum, Wynn's method produced even better results. Adaptive shifting can therefore be considered as a preconditioner of Wynn's method.

Sensitivity of the solution w.r.t. shift

The shift parameter is found to depend strongly on the eigenvalue distribution of the operators. This was confirmed by the fact that a different shift was required for each class of problems. Moreover, for different values of s the rates of convergence were calculated and plotted. For a nozzle with 50 grid points and linear extrapolated characteristic BCs, various cases were run by changing s. The rates of convergence at 200 time steps were obtained and plotted against s. Figure 13 shows that there is an optimal value of s beyond which the solution fails to converge as fast. This optimal value is s = -0.235. A similar behaviour was noticed when the flow was started with Dirichlet BCs.

The optimal value of s as suggested by the formula $s = -\frac{1}{2}(\lambda_{max} + \lambda_{min})$ is -0.25, whereas the value obtained from the plotted experimental data is -0.235. The corresponding spectrum with s = -0.235 is shown in Figure 10.

6. CONCLUSIONS

Shifting of the spectrum is a very simple but powerful idea. It stems from the usual power method of linear algebra. The amount of shift to be employed depends on the eigenspectrum of the problem in question. This is where the entire difficulty lies, but Arnoldi's method is an economical solution

to this problem. Since it extracts the ends of the spectrum⁹ very efficiently and accurately, it can be used to compute λ_{max} and λ_{min} . In the one-dimensional nozzle problem with 100 grid points Arnoldi's method gave a good estimate of λ_{max} and λ_{min} by using an additional 20 iterations of the code. This is small as compared to the amount of savings produced. Arnoldi's method actually pays for itself when used for shifting the spectrum of any linear or non-linear transformation. In using Arnoldi's method for computing spectra of non-linear operators, the corresponding eigenvectors could always be obtained at no extra cost. These eigenvectors, which belonged to a certain Krylov subspace, formed an orthogonal system under some assumptions.¹⁷ Arnoldi's method gave more information about the operators than was utilized in this paper.

The strategy of shifting the spectrum of implicit operators used in this paper was also applied to ARC2D, a two-dimensional problem. The results will be presented in a future paper.¹⁸

ACKNOWLEDGEMENTS

This work was supported in part by NASA grant #NCA2-172. The authors are grateful to the referee for his suggestions in improving this paper, especially in identifying the shifting method as a pseudo-time-step method, since in fact the shift is equivalent to changing the time step Δt to $\Delta t^* = \Delta t/(1+s)$. This can be seen from equations (14) and (15) in which $\Lambda_{max} = 1/[1+g_i/(1+s)]$ can be compared with $\lambda_{max} = 1/(1+g_i)$.

REFERENCES

- 1. Liepmann and Roshko, Elements of Gas Dynamics, Wiley, New York, 1957.
- R. Beam and R. F. Warming, 'An implicit finite-difference algorithm for hyperbolic systems in conservation law form', J. Comput. Phys., 22, 87-110 (1976).
- 3. M. Saleem, 'Spectrum analysis and convergence acceleration techniques applied to implicit finite difference approximations for the Euler and Navier-Stokes equations', *Ph.D. Thesis*, Department of Applied Mathematics, University of California, Davis, CA, 1988.
- 4. T. H. Pulliam, 'Implicit solution methods in computational fluid dynamics', Appl. Numer. Math., 2, 441-474 (1986).
- 5. M. Hafez and H. K. Cheng, 'Convergence acceleration and shock fitting for transonic aerodynamics computations', AIAA 75-51, 13th Aerospace Sciences Meeting, Pasadena, CA, 1975.
- 6. F. H. Harlow and A. Amsden, *Fluid Dynamics*, LASL Monograph, Los Alamos Scientific Laboratory of the University of California, Los Alamos, New Mexico, 87544, June 1971.
- 7. W. E. Arnoldi, 'The principle of minimized iterations in the solutions of matrix eigenvalue problems', Q. Appl. Math., 9, 17–29 (1951).
- L. E. Eriksson and A. Rizzi, 'A computer aided analysis of the convergence to steady state of discrete approximations to the Euler equations', J. Comput. Phys., 57, 90-128 (1985).
- 9. Y. Saad, 'Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices', *Linear Algebra Appl.*, 34, 269–295 (1980).
- 10. J. Cullum and R. A. Willoughby, Lanczos Algorithms for Large Symmetric Eigenvalue Computations, Vol. I, Birkhauser, Boston, 1985.
- 11. D. C. Jespersen and P. G. Buning, 'Accelerating an iterative process by explicit annihilation', SIAM J. Sci. Stat. Comput., 6, 639-651 (1985).
- 12. D. Shanks, 'Nonlinear transformations of divergent and slowly convergent sequences', J. Math. Phys., 34, 1-42 (1955).
- 13. P. Wynn, 'Acceleration techniques for iterated vector and matrix problems', Math. Comput., 6, 301-322 (1962).
- 14. P. Wynn, 'On the convergence and stability of the epsilon algorithm', SIAM J. Numer. Anal., 2, 91-121 (1966).
- C. Brezenski, 'Acceleration de la convergence en analysis numerique', Lecture Notes in Mathematics, 584, Springer, Berlin, 1977.
- A. Cheer, M. Saleem, T. H. Pulliam and M. Hafez, 'Analysis of the convergence history of flows through nozzles with shocks', AIAA-SIAM First Natl Fluid Dynamics Congr., AIAA-88-3795-CP, Cincinnati OH, 1988.
- 17. Y. Saad, 'Krylov subspace methods for solving large unsymmetric linear systems', Math. Comput., 37, 105-126 (1981).
- 18. M. Saleem and A. Cheer, 'Convergence acceleration and spectrum transformation of a two dimensional fluid flow problem', To be submitted.