NASA Contractor Report 178076

ICASE REPORT NO. 86-18

NASA-CR-178076
19860017080

ICASE

ADVANCES IN NUMERICAL AND APPLIED MATHEMATICS

Edited by
J. C. South, Jr.
and
M. Y. Hussaini

Contract Nos. NAS1-17070 and NAS1-18107
March 1986

INSTITUTE FOR COMPUTER APPLICATIONS IN SCIENCE AND ENGINEERING
NASA Langley Research Center, Hampton, Virginia 23665

Operated by the Universities Space Research Association.

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23665

ENTER:

DISPLAY 39/6/1

86N26552*# ISSUE 17 PAGE 2729 CATEGORY 34 RPT#: NASA-CR-178076
ICASE-86-18 NAS 1.26:178076 CNT#: NAS1-17070 NAS1-18107 86/03/00 593
PAGES UNCLASSIFIED DOCUMENT

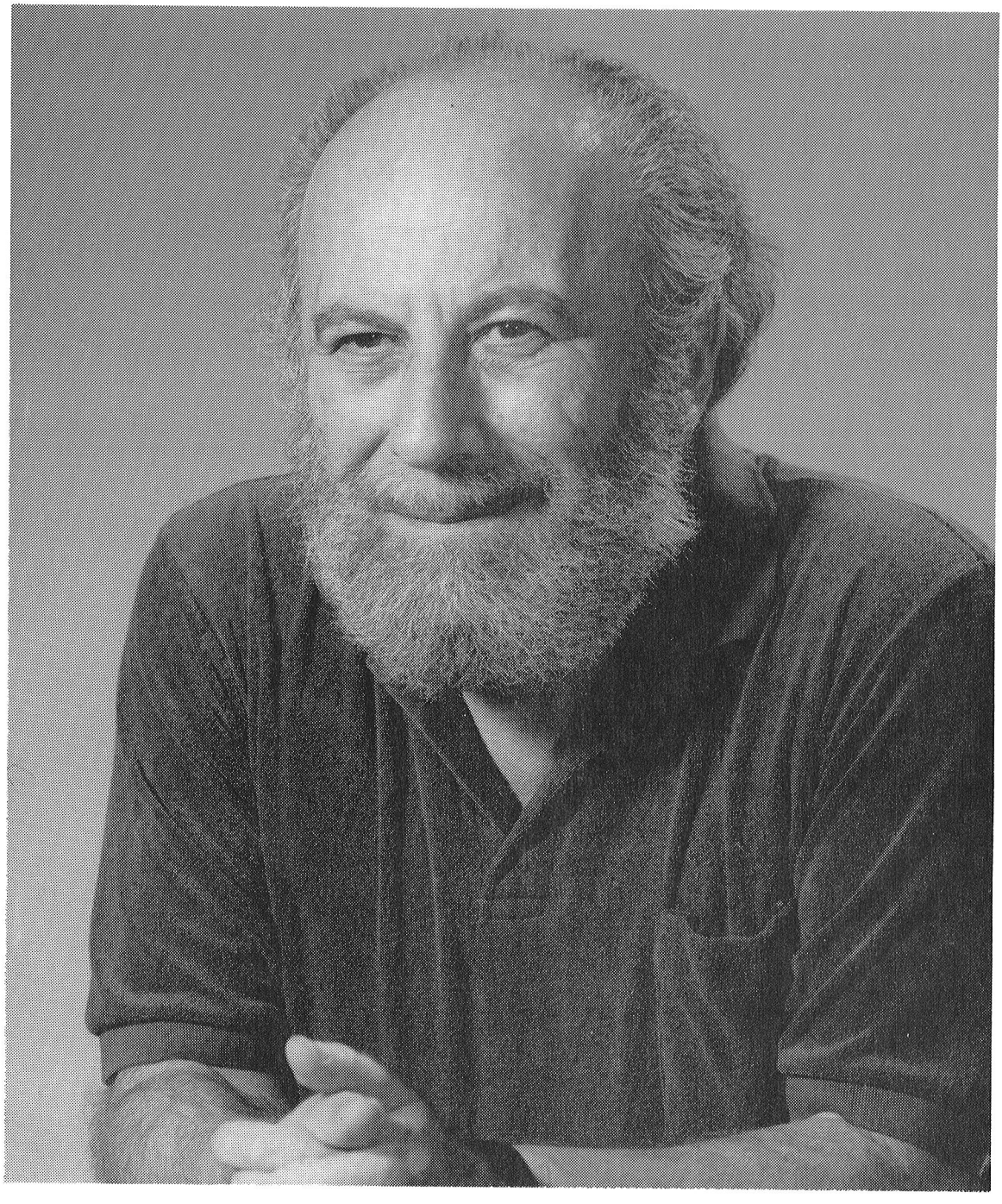
UTTL: Advances in numerical and applied mathematics TLSP: Final Report
AUTH: A/SOUTH, J. C., JR.; B/HUSSAINI, M. Y. PAT: A/ed.; B/ed.
CORP: National Aeronautics and Space Administration. Langley Research Center, Hampton, Va. AVAIL.NTIS
SAP: HC A25/MF A01
CIO: UNITED STATES Submitted for publication
MAJS: /*BOUNDARY VALUE PROBLEMS/*COMPUTATIONAL FLUID DYNAMICS/*NUMERICAL ANALYSIS/*SPECTRAL METHODS/*TRANSONIC FLOW/*VORTEX BREAKDOWN
MINS: / COMPUTATIONAL GRIDS/ FINITE ELEMENT METHOD/ GAS DYNAMICS/ INCOMPRESSIBLE FLOW/ NAVIER-STOKES EQUATION/ PIPES (TUBES)
ANN: This collection of papers covers some recent developments in numerical analysis and computational fluid dynamics. Some of these studies are of a fundamental nature. They address basic issues such as intermediate boundary conditions for approximate factorization schemes, existence and uniqueness of steady states for time dependent problems, and pitfalls of implicit time stepping. The other studies deal with modern numerical methods such as total variation diminishing schemes, higher order variants of vortex and particle methods, spectral multidomain techniques, and front tracking techniques. There is also a paper on adaptive grids. The fluid

ENTER:

DISPLAY 39/6/1

dynamics papers treat the classical problems of incompressible flows in helically coiled pipes, vortex breakdown, and transonic flows. For individual titles see N86-26553 through N86-26573.

ADVANCES IN NUMERICAL AND APPLIED MATHEMATICS



Dr. Milton E. Rose

DEDICATION

Dr. Milton E. Rose began his mathematical career in numerical analysis at the start of the computer era. He received the Ph.D. degree from New York University where he studied under Richard Courant. There, using a Univac I, he helped demonstrate the feasibility of studying floods in a large river system with dams and power stations (Ohio, Tennessee, and Mississippi rivers). His research has continued to emphasize the importance of developing efficient approximation methods for the numerical treatment of partial differential equations while keeping physical ideas at the forefront. His work has served as an inspiration for two generations of colleagues. In particular, his treatment of "Stefan problems," using enthalpy rather than temperature, has become the standard practice in the field.

Dr. Rose has continued his research while engaged in an active administrative career. He has served as Head of the Applied Mathematics Division, Brookhaven National Laboratory; Head of the Mathematical Sciences Section, National Science Foundation; Head of the Office of Computing Activities, National Science Foundation; Chairman of the Mathematics Department, Colorado State University; Chief of the Mathematics and Geosciences Branch, Energy Research and Development Administration; and has served as Director of the Institute for Computer Applications in Science and Engineering (ICASE) since September 1977. His administrative efforts produced remarkable improvements in the fields of computer applications that he managed for the U. S. government.

At ICASE, Dr. Rose has nurtured and brought to maturity an activity that has gained international recognition for its breadth and intellectual content.

On the occasion of his 60th birthday, a few of Dr. Rose's friends have produced this volume to show their appreciation for his wise and happy guidance and to challenge him to keep it up for the next 60.

Eugene Isaacson
March 1986

FOREWORD

This volume contains 21 research papers dedicated to Milton E. Rose on the occasion of his 60th birthday. The contributors are mathematicians and fluid dynamicists who have known and worked with Milt Rose during his tenure as Director of ICASE.

These research papers cover some recent developments in numerical analysis and computational fluid dynamics. Some of these studies are of a fundamental nature. They address basic issues such as intermediate boundary conditions for approximate factorization schemes, existence and uniqueness of steady states for time-dependent problems, pitfalls of implicit time stepping, etc. The other studies deal with modern numerical methods such as total-variation-diminishing schemes, higher order variants of vortex and particle methods, spectral multi-domain techniques, and front-tracking techniques. There is also a paper on adaptive grids. The fluid dynamics papers treat the classical problems of incompressible flows in curved pipes, vortex breakdown, and transonic flows.

The editors would like to take this opportunity to thank the authors for their excellent contributions and their promptness for meeting deadlines.

JCS and MYH
March 1986

TABLE OF CONTENTS

Section I

Section II

Section III

CONVERGENCE TO STEADY STATE OF SOLUTIONS OF BURGERS' EQUATION

Gunilla Kreiss
Royal Institute of Technology
Stockholm, Sweden

and

Heinz-Otto Kreiss
California Institute of Technology
Pasadena, California

Abstract

Consider the initial-boundary value problem for Burgers' equation. It is shown that its solutions converge, in time, to a unique steady state. The speed of the convergence depends on the boundary conditions and can be exponentially slow. Methods to speed up the rate of convergence are also discussed.

Research was partially supported by the Office of Naval Research under N00014-83-K-0422 and National Science Foundation Grant DMS-8312264. Additional support was provided by the National Aeronautics and Space Administration under NASA Contract No. NAS1-17070 while the authors were in residence at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA 23665-5225.

1. Introduction. In many gasdynamical problems one tries to calculate the steady state solution by solving the corresponding time dependant problem. One hopes that for $t \to \infty$ the solution converges to a unique steady state. Recently, M. D. Salas, S. Abarbanel and D. Gottlieb [1]considered the initial-boundary value problem

$$\begin{aligned} &u_t + \frac{1}{2}(u^2)_x = f(x), \quad t \geq 0, \quad 0 \leq x \leq \pi, \\ &u(x,0) = g(x). \end{aligned} \tag{1.1}$$

They used

$$f(x) = \sin x \cos x, \quad g(x) = b \sin x, \quad 0 < b,$$

and showed that the solution $u(x,t)$ of the above problem converges to a steady state $v(x)$, as $t \to \infty$, but that $v(x)$ depends on the initial data.

In this paper we consider the viscous problem

$$u_t + \frac{1}{2}(u^2)_x = \varepsilon u_{xx} + f(x), \quad t \geq 0, \quad 0 \leq x \leq 1, \quad \varepsilon > 0, \tag{1.2a}$$

with initial and boundary conditions

$$\begin{aligned} u(x,0) &= g(x), \\ u(0,t) = a, \quad u(1,t) &= b, \end{aligned} \tag{1.2b}$$

and the corresponding steady state problem

$$\begin{aligned} &\frac{1}{2}(y^2)_x = \varepsilon y_{xx} + f(x), \quad 0 \leq x \leq 1, \quad \varepsilon > 0, \\ &y(0) = a, \quad y(1) = b. \end{aligned} \tag{1.3}$$

For simplicity we restrict ourselves to two cases:

1) $a > 0 \geq b$, $a \geq -b$, $f(x) \equiv 0$,
2) $a = b = 0$, f is such that there exists an α with $0 < \alpha < 1$ such that $f(x) > 0$ for $0 < x < \alpha$, $f(x) < 0$ for $\alpha < x < 1$, $f(0) = f(1) = 0$, $f_x(0) \geq f_0 > 0$ and $f_x(1) \geq f_0$.

We will show that (1.3) has a unique solution and discuss the properties of $y(x)$. We shall also show that in all cases we consider, the limit of $y(x)$ as $\varepsilon \to 0$ exists. Thus, if

$$\lim_{t \to \infty} u(x,t) = y(x)$$

exists, we obtain a unique steady state solution of the inviscid equation (1.1) if we first let $t \to \infty$ and then $\varepsilon \to 0$. This is in contrast to the procedure in [1], where the two limit procedures are taken in the reverse order.

We shall prove that the eigenvalues of the eigenvalue problem

$$\lambda\varphi = -(y\varphi)_x + \varepsilon\varphi_{xx}, \quad \varphi(0) = \varphi(1) = 0, \tag{1.4}$$

are all negative. Therefore, the solution of (1.2) converges to the solution of (1.3) provided $u(x,0) = g(x)$ is sufficiently close to $y(x)$. In another paper we shall prove that $u(x,t)$ converges to $y(x)$ as $t \to \infty$ for arbitrary initial data. The speed of convergence is determined by the eigenvalues, λ_j, of (1.4). We shall show that the eigenvalue distribution depends on $f(x)$ and on a, b in the following way:

There is a constant $c > 0$ which does not depend on ε such that

$$\begin{aligned}
&(1) \quad \text{if} \quad a > -b,\ f \equiv 0 \quad \text{then} \quad 0 > -c/\varepsilon > \lambda_1 > \lambda_2 > \cdots \\
&(2) \quad \text{if} \quad a = -b,\ f \equiv 0 \quad \text{then} \quad -\lambda_1 = O(e^{-1/\varepsilon}) > 0,\ -c/\varepsilon > \lambda_2 > \lambda_3 > \cdots \\
&(3) \quad \text{if} \quad a = b = 0,\ \int_0^1 f(x)dx \neq 0, \quad \text{then} \quad -c > \lambda_1 > \lambda_2 > \cdots \\
&(4) \quad \text{if} \quad a = b = 0,\ \int_0^1 f(x)dx = 0, \quad \text{then} \quad -\lambda_1 = O(e^{-1/\varepsilon}) > 0,\ -c > \lambda_2 > \lambda_3 > \cdots
\end{aligned} \tag{1.5}$$

We expect a reasonable speed of convergence in the first and third case, while in the second and fourth case the speed should be extremely slow due to the eigenvalue $-\lambda_1 = O(e^{-1/\varepsilon})$. This is confirmed by numerical experiments. We see that at first $u(x,t)$ quite rapidly approches the same limit as the inviscid equation (1.1), which consists of solutions of the stationary equation

$$\frac{1}{2}(u^2)_x = f(x)$$

connected by a shock. Once the viscous shock has been formed, the solution of (1.2) becomes quasi-stationary and the shock creeps extremely slowly to the "right" position. We can explain the behavior, because by linearizing around the quasistationary solution we find that the eigenvalues of the corresponding eigenvalue problem have a similar distribution as earlier.

If $-\lambda_1 = O(e^{-1/\varepsilon})$ then the speed of convergence is so slow that the above method to calculate the steady state is impractical, see figures (1) and (3). However, we can use the same technique as Hafez, Parlette and Salas in [2] to speed up the convergence. See figures (2) and (4).

Unfortunatly, not only the speed of convergence but also the condition number of the stationary problem deteriorates. We have to calculate with $O(e^{1/\varepsilon})$ decimals to obtain correct

results. To avoid an excessive number of decimals we have used a quite large ε in our numerical calculations.

The situation becomes much better in a two dimensional case, which we discuss in the last section. Now there is a whole sequence of eigenvalues

$$-\mu_{1j} = O(j^2\varepsilon), \qquad j = 1, 2, \ldots,$$

close to zero. However, they are only algebraically and not exponentially close to zero. We indicate how to modify the procedure to accellerate the speed of convergence.

We believe that the viscous model (1.2) better explains what happens in actual calculations than the inviscid equation (1.1). Practically all numerical methods have some viscosity built in. Also, from a physical point of view, the solution we are interested in is the limit of solutions of a viscous equation.

Finally we want to point out that the appearance of small eigenvalues has also been observed by D. Brown, W. Kath, H. O. Kreiss and W. Henshaw, M. Naughton (private communication).

2. Uniqueness,existence and properties of the steady state solution. We start with uniqueness, which can be proven by standard techniques.

Lemma 2.1. If the steady equation (1.3) has a solution, then it is unique.

Proof. Let u, v be two solutions. Then $w = u - v$ is the solution of

$$\frac{1}{2}(pw)_x = \varepsilon w_{xx}, \quad p = u + v, \quad w(0) = w(1) = 0. \tag{2.1}$$

If $w \not\equiv 0$ then the zeros of w are isolated. Let $\overline{x}$ with $0 < \overline{x} \leq 1$ be the first zero to the right of $x = 0$. Without restriction we can assume that $w > 0$ for $0 < x < \overline{x}$, i.e. $w_x(0) \geq 0$ and $w_x(\overline{x}) \leq 0$. Integration of (2.1) gives us

$$-\varepsilon(|w_x(\overline{x})| + |w_x(0)|) = \varepsilon[w_x]_0^{\overline{x}} = \frac{1}{2}[pw]_0^{\overline{x}} = 0.$$

Thus $w_x(0) = w_x(\overline{x}) = 0$. We can consider (2.1) as an initial value problem with initial data $w(0) = w_x(0) = 0$ whose solution is $w(x) \equiv 0$, and the lemma is proved.

We shall now discuss the properties of the solution. Let us start with the case $f(x) \equiv 0$, $a > 0 \geq b$, $a > -b$. Integrating (1.3) gives us

$$\begin{aligned} \varepsilon y_x &= \frac{1}{2}y^2 - c, \quad 0 \leq x \leq 1, \\ y(0) &= a. \end{aligned} \tag{2.2}$$

The constant c has to be determined so that $y(1) = b$. We necessarily have $c = d^2/2 > a^2/2$, because with $c \le a^2/2$, $\quad y_x \ge 0$ for all x, and $y(1) = b$ cannot be satisfied. We can solve equation (2.2) explicitly. This is done by writing (2.2) in the form

$$2\varepsilon \int_a^{y(x)} \frac{d\tilde{y}}{\tilde{y}^2 - d^2} = \int_0^x d\tilde{x},$$

i.e.

$$\left(\frac{a+d}{a-d}\right)\left(\frac{y(x)-d}{y(x)+d}\right) = e^{dx/\varepsilon}$$

Therefore $y(1) = b$ implies $d = a + O(e^{-1/\varepsilon})$, and

$$y(x) = a\frac{1 - \tau e^{-a(1-x)/\varepsilon}}{1 + \tau e^{-a(1-x)/\varepsilon}}, \quad \text{with} \quad \tau = \frac{a-b}{a+b}. \tag{2.3}$$

Away from the boundary layer at x=1 we have $y(x) = a + O(e^{-a(1-x)/\varepsilon})$. Thus, for $\varepsilon \to 0$, $y(x)$ converges to a for $0 \le x < 1$.

If $a = -b$ we consider (2.2) on the interval $0 \le x \le \frac{1}{2}$, with boundary conditions $y(0) = a$, $\quad y(\frac{1}{2}) = 0$ and obtain a solution $y_1(x)$ of the form (2.3). The solution on the whole interval is given by

$$y(x) = \begin{cases} y_1(x), & \text{if } 0 \le x \le \frac{1}{2}, \\ -y_1(1-x), & \text{if } \frac{1}{2} \le x \le 1. \end{cases}$$

In figures (9) and (10) we have plotted $y(x)$ for two different sets of boundary values.

Consider case 2, where f only vanishes at $x = 0, \alpha, 1$ and $a = b = 0$. Without restrictions we can assume that

$$\int_0^1 f(x) \ge 0. \tag{2.4}$$

If this is not true, we transform the problem by introducing new variables,

$$\tilde{x} = 1 - x, \quad \tilde{f} = -f, \quad \tilde{y} = -y.$$

The new problem satisfies (2.4).

Lemma 2.2. Let $y(x)$ be the solution of (1.3), $F(x) = \int_0^x f(\xi)d\xi$ and $h(x) = \sqrt{2F(x)}$. Then

$$y_x(1) \le y_x(0) \le K_1, \quad K_1 = \max_{0 \le x \le 1} \{|h_x(x)|\} + |h_x(0)|.$$

Proof. Integration of (1.3) gives

$$\begin{aligned} \varepsilon(y_x - y_x(0)) &= \frac{1}{2}y^2 - F, \\ y(0) &= 0, \end{aligned} \tag{2.5}$$

where $y_x(0)$ is determined by $y(1) = 0$. If $u = y - h$, then u is the solution of

$$\begin{aligned} u_x &= y_x(0) - h_x + \varepsilon^{-1}uh + \frac{1}{2}\varepsilon^{-1}u^2, \\ u(0) &= 0. \end{aligned}$$

Assume that $y_x(0) > K_1$. It follows that $y_x(0) - h_x(x)$ is positive and thus u and u_x are positive for all $x > 0$. In particular $u(1) > 0$ and $y(1) = u(1) + h(1) > 0$, which contradicts $y(1) = 0$. Thus $y_x(0) \leq K_1$. Also

$$\varepsilon y_x(1) = \varepsilon y_x(0) - F(1) \leq \varepsilon y_x(0).$$

This proves the lemma.

Lemma 2.3. Let $y(x)$ be the solution of (1.3) and let ε be sufficiently small. If $F(1) > 0$ then $y(x) > 0$ for $0 < x < 1$ and $y(x)$ has exactly one maximum. If $F(1) = 0$ then there exists an $\overline{x}$ with $0 < \overline{x} < 1$ such that $y(x) > 0$ for $0 < x < \overline{x}$, and $y(x) < 0$ for $\overline{x} < x < 1$. Also $y(x)$ has exactly one minimum and one maximum. In both cases $|y(x)| \leq \max|F(x)|$.

Proof. At extrema $y_x = 0$ and

$$y_{xx} = -\varepsilon^{-1}f = \begin{cases} < 0 & \text{for } 0 < x < \alpha, \\ = 0 & \text{for } x = \alpha, \\ > 0 & \text{for } \alpha < x < 1. \end{cases} \tag{2.6}$$

Thus y cannot have a minimum to the left of a maximum. Since $y(0) = y(1) = 0$ there are only three possibilities, namely

$$y > 0 \quad \text{for} \quad 0 < x < 1, \quad \text{y has exactly one maximum,} \tag{2.7a}$$

$$y < 0 \quad \text{for} \quad 0 < x < 1, \quad \text{y has exactly one minimum,} \tag{2.7b}$$

$$\begin{aligned} &y > 0 \quad \text{for} \quad 0 < x < \overline{x}, \qquad 0 < \overline{x} < 1, \\ &y < 0 \quad \text{for} \quad \overline{x} < x < 1, \quad \text{y has exactly one maximum and one minimum.} \end{aligned} \tag{2.7c}$$

We shall prove that if $F(1) > 0$ then (b) and (c) are not possible, and that if $F(1) = 0$ then (a) and (b) are not possible.

Let $F(1) > 0$. Suppose (2.7b) holds. Then

$$y_x(0) \leq 0, \quad y_x(1) \geq 0.$$

By (2.5)

$$0 \leq \varepsilon(y_x(1) - y_x(0)) = -F(1) < 0. \tag{2.8}$$

This is a contradiction, so (2.7b) cannot hold. Now supposse (2.7c) is valid. Then $y_x(0) \geq 0$ and by (2.8)

$$y_x(0) \geq \varepsilon^{-1} F(1).$$

If ε is small enough this is impossible by lemma 2.2.

Let $F(1) = 0$. Assume that (2.7a) or (2.7b) are valid. By (2.8) $y_x(0) = y_x(1)$, which is only possible if $y_x(0) = y_x(1) = 0$. Differentiating (1.3) gives us

$$\varepsilon y_{xxx} = y y_{xx} + (y_x)^2 - f_x. \tag{2.9}$$

Thus

$$\begin{aligned} y(0) = y_x(0) = y_{xx}(0) = 0, \quad y_{xxx}(0) < 0, \\ y(1) = y_x(1) = y_{xx}(1) = 0, \quad y_{xxx}(1) < 0. \end{aligned}$$

This implies that y must change sign at least once, which contradicts the assumption, and therefore (2.7c) must hold.

It remains to show that $|y(x)|$ is bounded by $\max|F(x)|$. Since $y(0) = y(1) = 0$, the maximum absolute value of y is found at a local extrema, where $y_x = 0$. Thus, from (2.5) it follows that

$$|y(x)| \leq \max_{0 \leq x \leq 1} |F(x) - \varepsilon y_x(0)| \leq \max_{0 \leq x \leq 1} |F(x)|.$$

This finishes the proof.

We can use the usual singular perturbation methods to discuss the behavior of the solution in detail, see for ex. [3].

Theorem 2.1. Let $F(1) > 0$, assume that (1.3) has a solution and that ε is sufficiently small. Then $y(x)$ has a boundary layer at $x = 1$. For $1 - O(\varepsilon|\log(\varepsilon)|) \leq x \leq 1$, $y(x)$ is close to $w(x)$ which is the solution of

$$\varepsilon w_x = \frac{1}{2} w^2 - F(1), \quad -\infty < x \leq 1, \quad w(1) = 0. \tag{2.9}$$

In any interval $0 < x_0 \leq x \leq 1 - O(\varepsilon|\log(\varepsilon)|)$

$$y(x) = h(x) + \tilde{\varepsilon} u_1(x, \varepsilon), \quad h(x) = \sqrt{2F(x)} =: xg(x), \tag{2.10}$$

where u_1 and its derivatives are bounded independantly of ε. For $0 \le x \le x_0 < \alpha$ we have

$$y(x) = h(x) + \varepsilon u(\tilde{x}), \quad \tilde{x} = x/\sqrt{\varepsilon}, \tag{2.11}$$

where u and the derivatives $d^\nu u/d\tilde{x}^\nu$ are bounded independantly of ε. Thus, for $\varepsilon \to 0$, $y(x)$ converges to $h(x)$ for $0 \le x < 1$.

Proof. We indicate only the proof of (2.11). In the proof we shall use I_1, I_2 and I to denote the intervals $0 \le \tilde{x} \le 1$, $1 \le \tilde{x} \le x_0/\sqrt{\varepsilon}$ and $0 \le \tilde{x} \le x_0/\sqrt{\varepsilon}$, respectively. We shall also use

$$\|f\|_I := \max_{\tilde{x} \in I} |f(\tilde{x})|,$$

where I is an interval.

We introduce a new variable in (1.3),

$$y(x) = h(x) + \varepsilon u(x/\sqrt{\varepsilon}).$$

This gives us

$$u_{\tilde{x}\tilde{x}} - (\tilde{x}g(x) + \sqrt{\varepsilon}u)u_{\tilde{x}} - h_x u = -h_{xx}, \quad 0 \le \tilde{x} \le x_0/\sqrt{\varepsilon}, \quad u(0) = 0, \quad u(x_0/\sqrt{\varepsilon}) = u_0, \tag{2.12}$$

where $u_0 = u_1(x_0, \varepsilon)$ is bounded independantly of ε. From $x_0 < \alpha$ and the assumtion $f_x(0) \ge f_0 > 0$ it follows that $h_x(x) \ge h_0 > 0$ for $0 \le x \le x_0$. Therefore we can use the maximum principle. The maximum of u is found either on the boundary or at a local extrema, where $u_{\tilde{x}} = 0$. At local extrema

$$|u| \le |\frac{h_{xx}}{h_x}| \le \frac{1}{h_0}\|h_{xx}(x)\|_I =: \alpha.$$

Thus

$$\|u\|_I \le \max(u_0, \alpha). \tag{2.13}$$

Next we want to estimate $\|u_{\tilde{x}}\|_I$. First we consider the interval $I_1 = [0, 1]$. By (2.12) and (2.13) there are constants C_1 and C_2 such that

$$\|u_{\tilde{x}\tilde{x}}\|_{I_1} \le C_1\|u_{\tilde{x}}\|_{I_1} + C_2.$$

It is well known, see Landau [4], that one can estimate $\|u_{\tilde{x}}\|_{I_1}$ in terms of $\|u\|_{I_1}$, and $\|u_{\tilde{x}\tilde{x}}\|_{I_1}$, i.e. for every constant δ there is a constant $C(\delta)$ such that

$$\|u_{\tilde{x}}\|_{I_1} \le \delta\|u_{\tilde{x}\tilde{x}}\|_{I_1} + C(\delta)\|u\|_{I_1}.$$

Thus for $\delta = \frac{1}{2}(C_1)^{-1}$ we obtain a bound for $\|u_{\tilde{x}\tilde{x}}\|_{I_1}$, which gives us a bound for $\|u_{\tilde{x}}\|_{I_1}$. Especially, $|u_{\tilde{x}}(1)|$ is bounded.

In the remaining interval $I_2 = [1, x_0/\sqrt{\varepsilon}]$, we have

$$F \ge F(\sqrt{\varepsilon}) = \varepsilon f_x(0)(1 + O(\sqrt{\varepsilon})).$$

Thus

$$\tilde{x}g + \sqrt{\varepsilon}u = \sqrt{2F}/\sqrt{\varepsilon} + \sqrt{\varepsilon} \ge \sqrt{f_x(0)} + O(\sqrt{\varepsilon}),$$

i.e. for sufficiently small $\sqrt{\varepsilon}$

$$\tilde{x}g + \sqrt{\varepsilon}u \geq \frac{1}{2}\sqrt{f_x(0)}\,.$$

At local extrema of $u_{\tilde{x}}$, $u_{\tilde{x}\tilde{x}} = 0$ and we have, by (2.12),

$$|u_{\tilde{x}}| \leq \frac{2}{\sqrt{f_x(0)}}|h_{xx} - h_x u| \leq \frac{2}{\sqrt{f_x(0)}}(\|h_{xx}\|_{I_2} + \|h_x\|_{I_2}\|u\|_{I_2}) =: \beta.$$

Thus

$$\|u_{\tilde{x}}\|_{I_2} \leq \max(|u_{\tilde{x}}(1)|, |u_{\tilde{x}}(\frac{x_0}{\sqrt{\varepsilon}})|, \beta),$$

and $u_{\tilde{x}}$ is bounded independantly of ε in the whole interval. By differentiating (2.12) bounds for higher derivatives of u can be obtained.

It is also clear that as $\varepsilon \to 0$, $y(x)$ converges to $h(x)$. This finishes the proof.

If $F(1) = 0$ then the solution switches at $\overline{x}$ from $\sqrt{2F} + O(\varepsilon)$ to $-\sqrt{2F} + O(\varepsilon)$. In each subinterval $0 \leq x < \overline{x}$ and $\overline{x} \leq x \leq 1$ the local behavior of the solution is of the same type as in the first case. As $\varepsilon \to 0$, $y(x)$ converges to $h(x)$ for $0 \leq x < \overline{x}$ and to $-h(x)$ for $\overline{x} < x \leq 1$. In general, the position of $\overline{x}$ can only be obtained by detailed calculation. However, if $f(x)$ is antisymmetric around $x = \frac{1}{2}$ then $\overline{x} = \frac{1}{2}$. This is the only case we consider.

We shall now discuss the existence of a solution. For this we need two lemmata.

Lemma 2.4. For sufficiently large ε the steady state equation (1.2) has a solution.

Proof. By integrating (1.3) twice, we can write the equation in the form

$$y(x) = \frac{1}{2}\eta \int_0^x y^2(\xi)d\xi - \eta \int_0^x F(\xi)d\xi + \eta x c_0, \quad \eta = 1/\varepsilon,$$

$$\frac{1}{2}\int_0^1 y^2(\xi)d\xi - \int_0^1 F(\xi)d\xi + c_0 = 0,$$

or after the change of variable $y = \eta\tilde{y}$

$$\tilde{y}(x) = \frac{1}{2}\eta^2 \int_0^x \tilde{y}^2(\xi)d\xi - \int_0^x F(\xi)d\xi + x c_0,$$

$$\frac{1}{2}\eta^2 \int_0^1 \tilde{y}^2(\xi)d\xi - \int_0^1 F(\xi)d\xi + c_0 = 0.$$

For $\eta = 0$ the above equations have a unique solution. Therefore the same is true for all sufficiently small η. This proves the lemma.

Lemma 2.5. Let $p(x)$ be a smooth function. Consider the eigenvalue problem

$$\lambda\varphi = -(p\varphi)_x + \varepsilon\varphi_{xx}, \quad \varphi(0) = \varphi(1) = 0. \tag{2.14}$$

The eigenvalues are real and negative.

Proof. We introduce a new variable $\psi(x)$ by

$$\varphi(x) = e^{\frac{1}{2}\varepsilon^{-1}\int_1^x p(\xi)d\xi}\psi(x),$$

and obtain

$$\begin{gathered}\lambda\psi = \varepsilon\psi_{xx} - c\psi =: L\psi, \quad c(x) = \frac{1}{2}p_x(x) + \frac{1}{4\varepsilon}(p(x))^2, \\ \psi(0) = \psi(1) = 0.\end{gathered} \tag{2.15}$$

(2.15) is selfadjoint and therefore the eigenvalues are real. Let $\varphi \not\equiv 0$, λ be a solution of (2.14), and let $\tilde{x}$ be the first zero of φ to the right of $x = 0$. We can assume that $\varphi > 0$ for $0 < x < \tilde{x}$. Thus $\varphi_x(0) \geq 0$ and $\varphi_x(\tilde{x}) \leq 0$, and integration of (2.14) gives us

$$\lambda\int_0^{\tilde{x}} \varphi(x)dx = \varepsilon[\varphi_x]_0^{\tilde{x}} \leq 0.$$

It follows that $\lambda \leq 0$. If $\lambda = 0$, the only possible solution of (2.14) would be $\varphi(x) \equiv 0$. Thus $\lambda < 0$, which proves the lemma.

Now we can prove

Theorem 2.2. The equation (1.3) has a unique solution for all $\varepsilon > 0$.

Proof. We have already shown that (1.3) has a solution for sufficiently large ε. We will now employ continuation in ε to prove existance for all $\varepsilon > 0$. Assume we have shown existance for $\varepsilon > \bar{\varepsilon}$. We want to show that there is a solution for $\varepsilon = \bar{\varepsilon}$. By lemma 2.3 the solutions of (1.3) are uniformly bounded for $\bar{\varepsilon} < \varepsilon \leq \bar{\varepsilon} + 1$. Therefore the same is true for the first three derivatives. Thus we can select a sequence of solutions

$$y(x, \varepsilon_\nu), \quad \nu = 1, 2, \ldots, \quad \lim_{\nu\to\infty} \varepsilon_\nu = \bar{\varepsilon},$$

such that

$$\lim_{\nu\to\infty} \frac{d^j}{dx^j} y(x, \varepsilon_\nu) = \frac{d^j}{dx^j} y(x, \bar{\varepsilon}), \quad j = 0, 1, 2 \quad ,$$

and $y(x,\bar{\varepsilon})$ is the desired solution.Linearizing the equation around $y(x,\bar{\varepsilon})$ gives us

$$(y(x,\bar{\varepsilon})\delta y)_x = \varepsilon(\delta y)_{xx} + (\varepsilon - \bar{\varepsilon})y(x,\bar{\varepsilon}), \quad \delta y(0) = \delta y(1) = 0.$$

By the previous lemma $\lambda = 0$ is not an eigenvalue of the above equation and therefore we can solve (1.3) for all sufficiently small $\varepsilon - \bar{\varepsilon}$. This proves the theorem.

3. Speed of convergence. In this section we want to discuss the speed of convergence to steady state. We assume that the initial data $g(x)$ of (1.3) are sufficiently close to the solution of the steady problem, so that we only have to discuss the behavior of the solutions of the linearized equation

$$\begin{aligned} w_t + (yw)_x &= \varepsilon w_{xx}, \quad 0 \le x \le 1, \quad t \ge 0, \\ w(x,0) &= \tilde{g}(x), \\ w(0,t) &= w(1,t) = 0. \end{aligned} \tag{3.1}$$

To determine the speed of convergence we study the distribution of eigenvalues of

$$\lambda\varphi + (y\varphi)_x = \varepsilon\varphi_{xx}, \quad \varphi(0) = \varphi(1) = 0. \tag{3.2}$$

Theorem 3.1 . The eigenvalues of (3.2) are real and negative and their distribution is given by (1.5).

Proof . Lemma 2.5 tells us that the eigenvalues are real and negative. First we consider the case $f \equiv 0, \quad a > -b$. We write (3.2) in the selfadjoint form (2.15) with $p = y$. Let $\lambda = \lambda_1$ be the largest eigenvalue. The corresponding eigenfunction ψ_1 does not change sign, and we can assume that $\psi_1 > 0$ for $0 < x < 1$ and that $\max|\psi_1(x)| = 1$. We assume that $\lambda_1 > -a^2/8\varepsilon$. Then there is a constant K such that $c(x) + \lambda_1 > 0$ for $0 \ge x \ge 1 - K\varepsilon$. Thus ψ_1 is monotone in the interval $0 \le x \le 1 - K\varepsilon$, and therefore ψ_1 must have its maximum in the remaining interval, $1 - K\varepsilon \le x < 1$. By assumption $\max\psi_1(x) = 1$ and therefore there must be a constant $\delta > 0$ such that $\psi_{1x}(1) \le -\delta/\varepsilon$. Now consider the corresponding eigenfunction

$$\varphi_1(x) = e^{\frac{1}{2}\varepsilon^{-1}\int_1^x y(\xi)d\xi}\psi_1(x), \quad \varphi_{1x}(1) = \psi_{1x}(1), \quad 0 \le \varphi_1(x) \le \psi_1(x).$$

Integrating (3.2) gives us

$$-\delta \ge \varepsilon(\varphi_{1x}(1) - \varphi_{1x}(0)) = \lambda_1 \int_0^1 \varphi_1 dx \ge$$

$$\ge \lambda_1 \int_0^1 e^{\frac{1}{2}\varepsilon^{-1}\int_1^x y d\xi} dx = \lambda_1 \varepsilon d.$$

Thus

$$\lambda_1 < -\min(\frac{a^2}{8\varepsilon}, \frac{\delta}{d\varepsilon}),$$

and the theorem is proven for this case.

When $f \not\equiv 0$, $a = b = 0$, and $\int_0^1 f(x)dx > 0$ the corresponding estimate follows in the same way, since by theorem 2.1 there are constants $C_0 > 0$ and K such that

$$c(x) = \frac{1}{2}(h_x(x) + O(\sqrt{\varepsilon}) + \frac{1}{4}\varepsilon^{-1}(h(x) + O(\sqrt{\varepsilon}))^2) \geq C_0 > 0 \quad \text{for} \quad 0 \leq x \leq 1 - K\varepsilon.$$

We now consider the antisymmetric case when $a = -b$, $f \equiv 0$ or $a = b = 0$ and $f(x)$ is antisymmetric around $x = \frac{1}{2}$. We want to show that

$$-\lambda_1 = O(\varepsilon^{-1}e^{-1/\varepsilon}).$$

We shall use the fact that for our selfadjoint eigenvalue problem (2.15) the eigenvalue with the smallest absolute value, λ_1, satisfies

$$|\lambda_1| \leq \frac{\|L\phi\|_2}{\|\phi\|_2},$$

for any smooth function $\phi \not\equiv 0$ satisfying the boundary conditions. We chose

$$\phi(x) = e^{\frac{1}{2}\varepsilon^{-1}\int_{1/2}^{x} y(\xi)d\xi} - e^{-\frac{1}{2}\varepsilon^{-1}\int_{0}^{1/2} y(\xi)d\xi}$$

as trial function. $y(x)$ is antisymmetric around $x = \frac{1}{2}$, and $\phi(0) = \phi(1) = 0$. Also

$$L\phi = (\frac{y^2}{4\varepsilon} + \frac{y_x}{2})e^{-\frac{1}{2}\varepsilon^{-1}\int_0^{1/2} y(\xi)d\xi}$$

Both ϕ^2 and $(\frac{1}{4}\varepsilon^{-1}y^2 + \frac{1}{2}y_x)^2$ are symmetric around $x = \frac{1}{2}$. Therefore

$$\|L\phi\|^2 = 2\int_0^{1/2} (\frac{y^2}{4\varepsilon} + \frac{y_x}{2})^2 e^{-\varepsilon^{-1}\int_0^{1/2} y d\xi} dx,$$

$$\|\phi\|^2 = 2\int_0^{1/2} e^{-\varepsilon^{-1}\int_0^{1/2} y d\xi}(e^{\frac{1}{2}\varepsilon^{-1}\int_0^{x} y d\xi} - 1)^2 dx,$$

and by (2.5) and theorem 2.1

$$\lambda_1^2 \le \frac{\|L\phi\|^2}{\|\phi\|^2} = \frac{\int_0^{1/2} (\frac{y^2}{4\varepsilon} + \frac{y_x}{2})^2 dx}{\int_0^{1/2} (e^{\frac{1}{2}\varepsilon^{-1}\int_0^x y d\xi} - 1)^2 dx} \le C^2 \varepsilon^{-2} e^{-2D/\varepsilon},$$

where $C > 0$, $D > 0$ are constants which do not depend on ε.

We shall now estimate the size of the second eigenvalue for the case with an interior boundary layer at $x = \frac{1}{2}$. By assumption $y(x)$ is antisymmetric around $x = \frac{1}{2}$. Consider the eigenvalue problem (3.2) on half the interval, $0 \le x \le \frac{1}{2}$, and denote its solutions by

$$\tilde{\varphi}_i(x), \quad \tilde{\lambda}_i, \quad i = 1, 2 \ldots .$$

We know that $\tilde{\varphi}_i$ has $i - 1$ sign changes, and we have already shown how the $\tilde{\lambda}_i$'s are bounded away from zero. The function

$$\varphi_{2i}(x) = \begin{cases} \tilde{\varphi}_i(x) & \text{for } 0 \le x \le \frac{1}{2} \\ -\tilde{\varphi}_i(x - \frac{1}{2}) & \text{for } \frac{1}{2} < x \le 1 \end{cases}, \quad i = 1, 2 \ldots ,$$

will satisfy (3.2) on the full interval, $0 \le x \le 1$ with $\lambda = \lambda_{2i} = \tilde{\lambda}_i$. Also φ_{2i} changes sign $2(i-1)+1$ times. Thus φ_{2i} is the $2i^{th}$ eigenfunction and λ_{2i} is the $2i^{th}$ eigenvalue. Therefore λ_2 is bounded away from zero. This finishes the proof.

4. Numerical results. We shall discuss difference approximations for the time dependant problem (1.2) and the eigenvalue problem (3.2). We introduce gridpoints

$$(x_i = ih, t_j = jk), \quad i = 0, 1, \ldots \quad j = 0, 1, \ldots, N, \quad h = \frac{1}{N},$$

where N is a natural number and $k > 0$ is the time step. We also introduce gridfunctions

$$v_i^j = v(x_i, t_j).$$

We approximate (1.2) by the usual implicit method

$$(I - \varepsilon k D_+ D_-) v_i^{j+1} + \frac{1}{2} k D_0 (v_i^{j+1})^2 = v_i^j + k f_i, \quad i = 1, 2, \ldots, N-1 \qquad (4.1)$$

with initial and boundary conditions

$$u_i^0 = g_i, \quad i = 1, 2, \ldots, N-1,$$

$$u_0^j = a, \quad u_N^j = b, \quad j = 1, 2, \ldots$$

Here

$$h^2 D_+ D_- u_i = u_{i+1} - 2u_i + u_{i-1} \quad \text{and} \quad 2hD_0(u_i)^2 = (u_{i+1})^2 - (u_{i-1})^2$$

denote the usual centered difference operators. At every time step one has to solve a nonlinear system to determine u_i^{j+1} . This is done by the iteration

$$(I - \varepsilon k D_+ D_-) u_i^{(l+1)} = -\frac{1}{2} k D_0 (u_i^{(l)})^2 + u_i^j + k f_i, \quad l = 0, 1 \ldots \quad , \tag{4.2}$$

where $u^{(0)}$ is choosen by a predictor process.

In all our experiments the solution of (4.1) converges to a steady state solution. However, the speed of convergence depends on the location of the shock. If the shock is located at the boundary, corresponding to the first and third case of (1.5), then the convergence to steady state is quite rapid. See figure (5). If on the other hand the shock is located in the interior, corresponding to the other cases of (1.5), the convergence is, in general, very slow. When the shock is formed at an early stage it is in general in the "wrong" place, depending on the initial data. From then on, the the shock moves slowly to the correct position. See figures (1),(3). This process can be considered quasi-stationary, which makes it possible to use the same convergence acceleration as in [2].

Formally we can write our iteration (4.1) as

$$H(u^{n+1}) = u^{n+1} - u^n := r^n. \tag{4.3}$$

We can linearize the realation and obtain

$$(I - L) r^{n+1} = r^n. \tag{4.4}$$

In our case

$$L r_i = \varepsilon k D_+ D_- r_i - k D_0 (u_i^{n+1} r_i). \tag{4.5}$$

This is a discretization of the right hand side of the eigenvalue problem (2.14), with $p = u^n$. If the process is quasi-stationary we can consider L to be independant of n. Then we have

$$r^{n+j} = (I - L)^{-j} r^n$$

and

$$u^{n+p} = u^n + \sum_{j=0}^{p-1} (I - L)^{-j} r^n.$$

If the eigenvalues λ_i, of L are negative the eigenvalues κ_i, of $(I - L)^{-1}$ satisfy $|\kappa_i| < 1$ and

$$\lim_{p \to \infty} u^{n+p} = u^n + (I - (I - L)^{-1})^{-1} r^n = u^n + (I - L^{-1}) r^n \tag{4.6}$$

Instead of taking a large number of time steps we can take one large step, which we call an extrapolation step. We put

$$u = u^n + \beta e, \tag{4.7}$$

where e is the solution of the equation

$$Le = (L - I)r^n, \tag{4.8}$$

and β is a stabilizing parameter. We choose β in such a way that $H(u^n + \beta e)$ has no component in the direction of e, i.e.

$$\langle H(u^n + \beta e), e\rangle = 0,$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product. There are other possible choices, for example choose β such that

$$\|H(u^n + \beta e)\| = \min_{\beta} \|H(u^n + \beta e)\|.$$

Of course (4.7) is not the steady solution we are seeking. We use the new u to restart the time iteration, and make a new extrapolation step once a new quasi-stationary state is reached. In our experiments we use an a priori fixed number of time steps between the extrapolation steps. Better strategies are under development.

We have calculated the first eigenvalues and eigenvectors of the discrete linearized operator (4.5), provided u_i^{n+1} is the discrete steady state solution. The calculations show that the eigenvalues are negative and their distribution is of the same type as for the corresponding continous case. See table (1). In figures (6),(7) the first few eigenvectors are plotted. Note that in the case of an interior shock the first eigenvector is exponentially small away from the shock region. Also, we have no doubt, and it is confirmed by the calculations, that the position of the shock does not change the nature of the eigenvalue distribution. In fact, in the proof of theorem 3.1, y can be replaced by any function of the same structure.

In our case, when the shock is located in the interior, $(I - L)^{-1}$ has only one eigenvalue, κ_1, close to zero. All other eigenvalues are small. Therefore, when we have reached the quasi-stationary state, r^n is in the direction of the eigenvector corresponding to κ_1. See figure (8). Therefore we do not need to solve (4.8), and instead of (4.7) we use

$$u = u^n + \beta r^n. \tag{4.9}$$

In figures (2),(4) we have plotted u at different time stages to show how the convergence is accelerated.

5. A twodimensional case. Consider the following problem

$$u_t + (\frac{1}{2}u^2)_x = \varepsilon(u_{xx} + u_{yy}), \quad 0 \le x \le 1, \quad 0 \le y \le 1, \quad t \ge 0,$$

$$\begin{gathered} u(0,y,t) = a, \quad u(1,y,t) = -a, \quad a > 0, \\ u(x,0,t) = u(x,1,t) = w(x), \\ u(x,y,0) = g(x,y), \end{gathered} \tag{5.1}$$

where $W(x)$ is the solution of the one dimensional problem (1.3) with $b = -a$, and $f(x) \equiv 0$. See (2.3). A steady solution of (5.1) is $u(x,y) = w(x)$.

The speed of convergence can be studied by analyzing the corresponding eigenvalue problem

$$\mu\varphi + (w\varphi)_x = \varepsilon(\varphi_{xx} + \varphi_{yy}), \quad \varphi = 0 \text{ on the boundary.} \tag{5.2}$$

We can solve (5.2) by separation of variables. Let $\varphi(x,y) = X(x)Y(y)$. Then

$$(wX)' - \varepsilon X'' = \lambda X, \quad X(0) = X(1) = 0, \tag{5.3a}$$

$$Y'' = -qY, \quad Y(0) = Y(1) = 0, \tag{5.3b}$$

with $\mu = \lambda - \varepsilon q$. We recognize (5.3a) as (3.2). Therefore $-\lambda_1 = O(e^{-1/\varepsilon})$ and $-\lambda_j > O(1/\varepsilon)$, $j = 2,3,\ldots$. We can solve (5.3b). The solution is

$$Y_j(y) = \sin(j\pi y), \quad q_j = (j\pi)^2, \quad j = 1,2\ldots.$$

There is a whole sequence of eigenvalues, μ_{1j}, of order $O(\varepsilon)$. The eigenfunctions corresponding to this sequence, φ_{1j}, will be exponentially small away from the shock. All other eigenvalues will be of order $O(1/\varepsilon)$.

We expect that the time iteration will again lead to a quasi-stationary state, and that the residual will be composed of eigenfunctions corresponding to the eigenvalues of order $O(\varepsilon)$. Therefore e in (4.8) will be of the same form, and we can replace all components of e away from the shock by zero, thus obtaining a linear system of equations of order N instead of N^2. More details will be given in another paper.

REFERENCES

[1] M. D. Salas, S. Abarbanel, D. Gottlieb, *Multiple steady states for characteristic initial value problems*, Icase report No 84-57 , NASA CR-172486, November 1984.

[2] M. Hafez, E. Parlette, M. Salas, *Convergence acceleration of iterative solutions for transonic flow computations*, AIAA 85-1641.

[3] J. D. Cole, J. Kevorkian, *Perturbation methods in Applied Mathematics*, Springer 1981.

[4] E. Landau, *Einige Ungleichungen für zweimal differenzierbare Funktionen*, Proc. London Math. Soc. 13(1913) 43-49.

Table 1.

Eigenvalues of the eigenvalueproblem (3.2), y is the solution of (1.3). Three different cases were treated. The discretization is done according to (4.5), with $N = 100$ gridpoints. The eigenvalues were found using inverse iteration. Eigenvectors corresponding to case (1) are plotted in figure (6a,b).

	λ_1	λ_2	λ_3
$f(x) = \sin(2\pi x)/2$ $a = b = 0$ $\varepsilon = 0.04$	$-8.64 \cdot 10^{-3}$	-4.34	-5.32
$f(x) = \sin(2\pi x)/2$ $a = b = 0$ $\varepsilon = 0.02$	$-4.62 \cdot 10^{-6}$	-5.617	-5.622
$f(x) \equiv 0$ $a = 1,\ b = -1$ $\varepsilon = 0.02$	$-1.24 \cdot 10^{-9}$	-12.8	-13.5

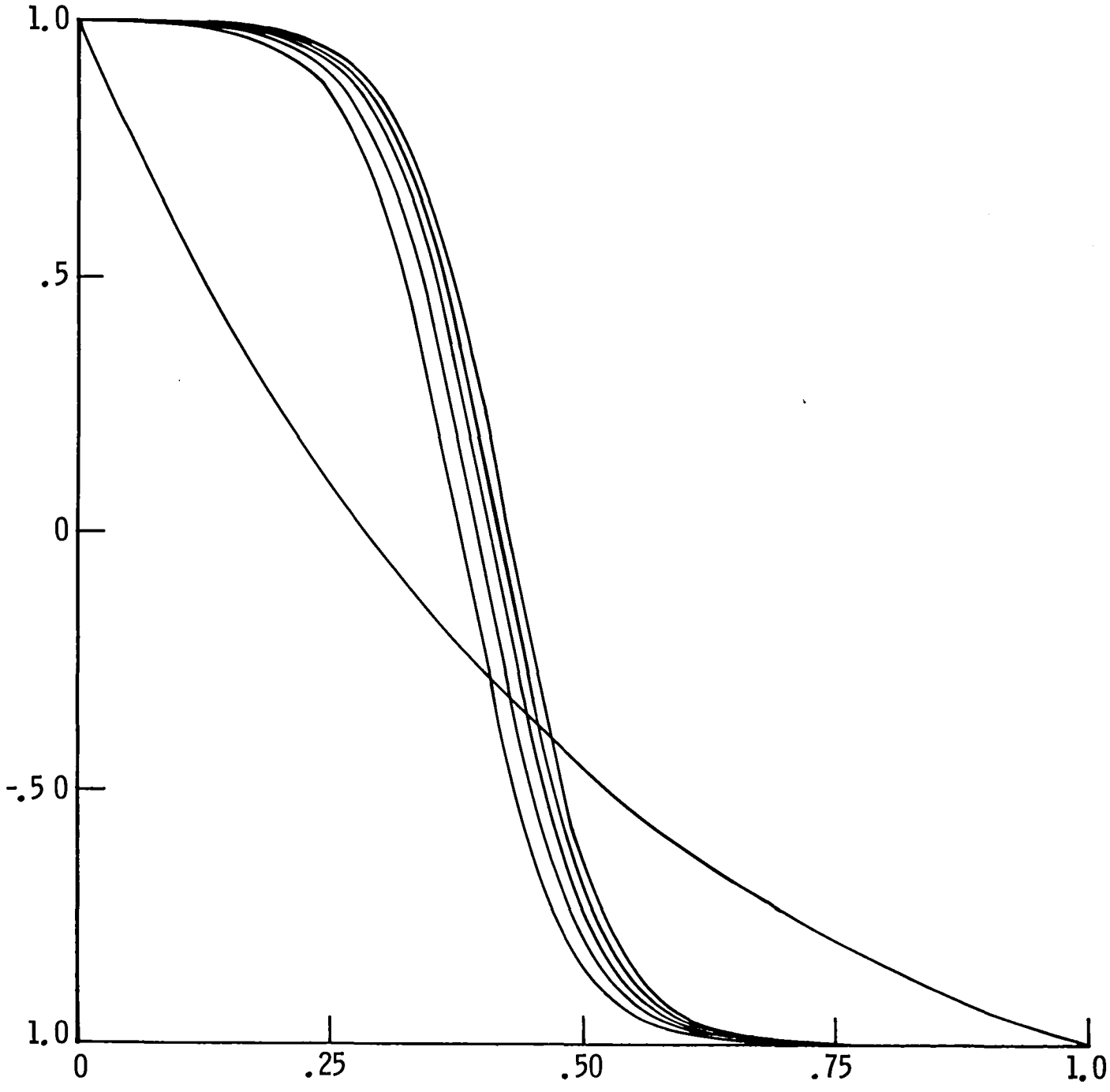


Figure 1. Convergence in time without convergence acceleration. Numerical solutions at different time stages for the case $\varepsilon = 0.05$, $f \equiv 0$, $a = 1$, $b = -1$, $u(x,0) = 1 + 2(e^{-2x} - 1)/(1 - e^{-2})$. Between each curve there are 200 time steps = 40 time units. The calculation is made with time step k = 0.2 and N=50 grid points.

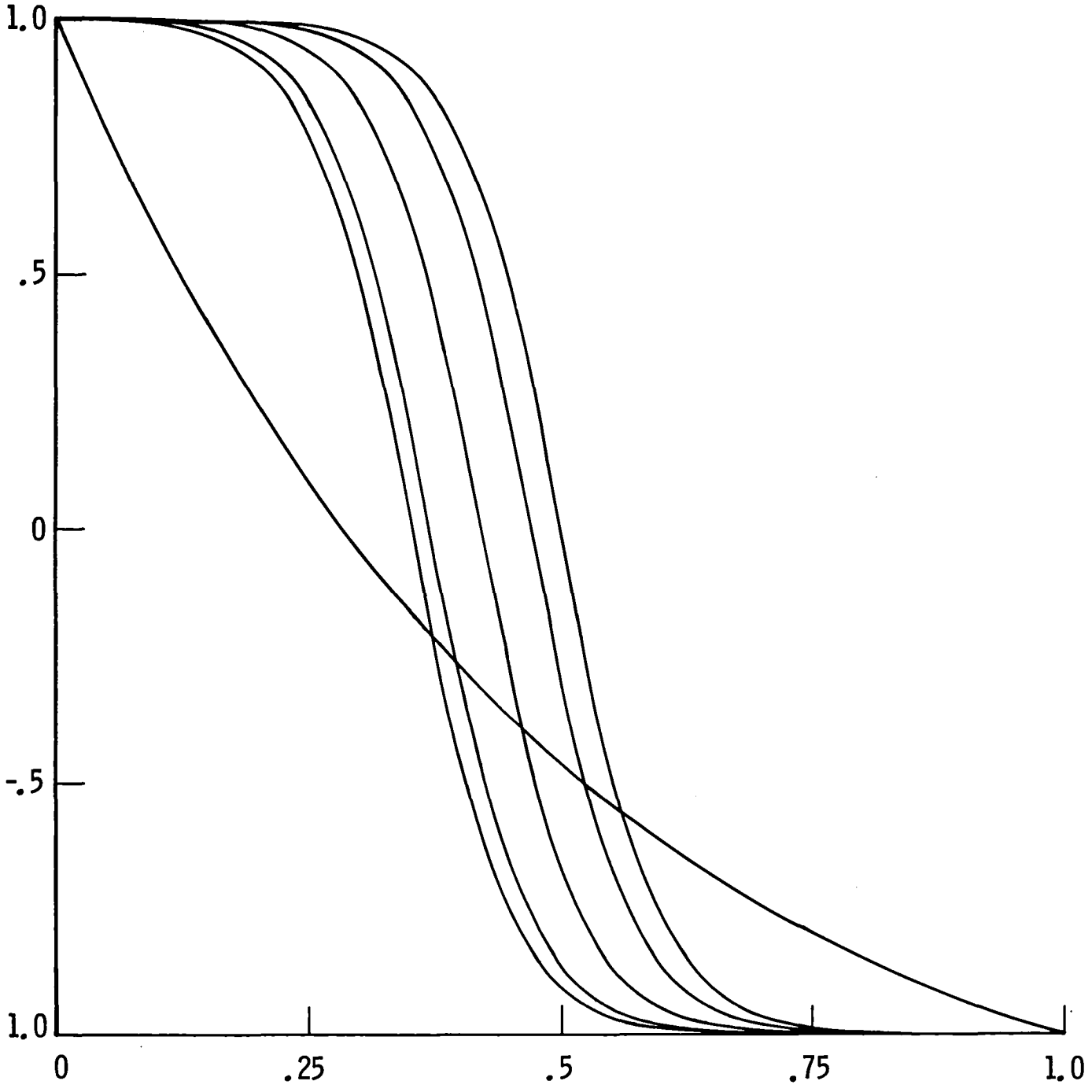


Figure 2. Convergence in time with convergence acceleration. Numerical solutions at different time stages for the same case as in figure 1. Between each curve there are 15 time steps and one extrapolation step. The same time step, k=0.2, and number of grid points , N=50, are used.

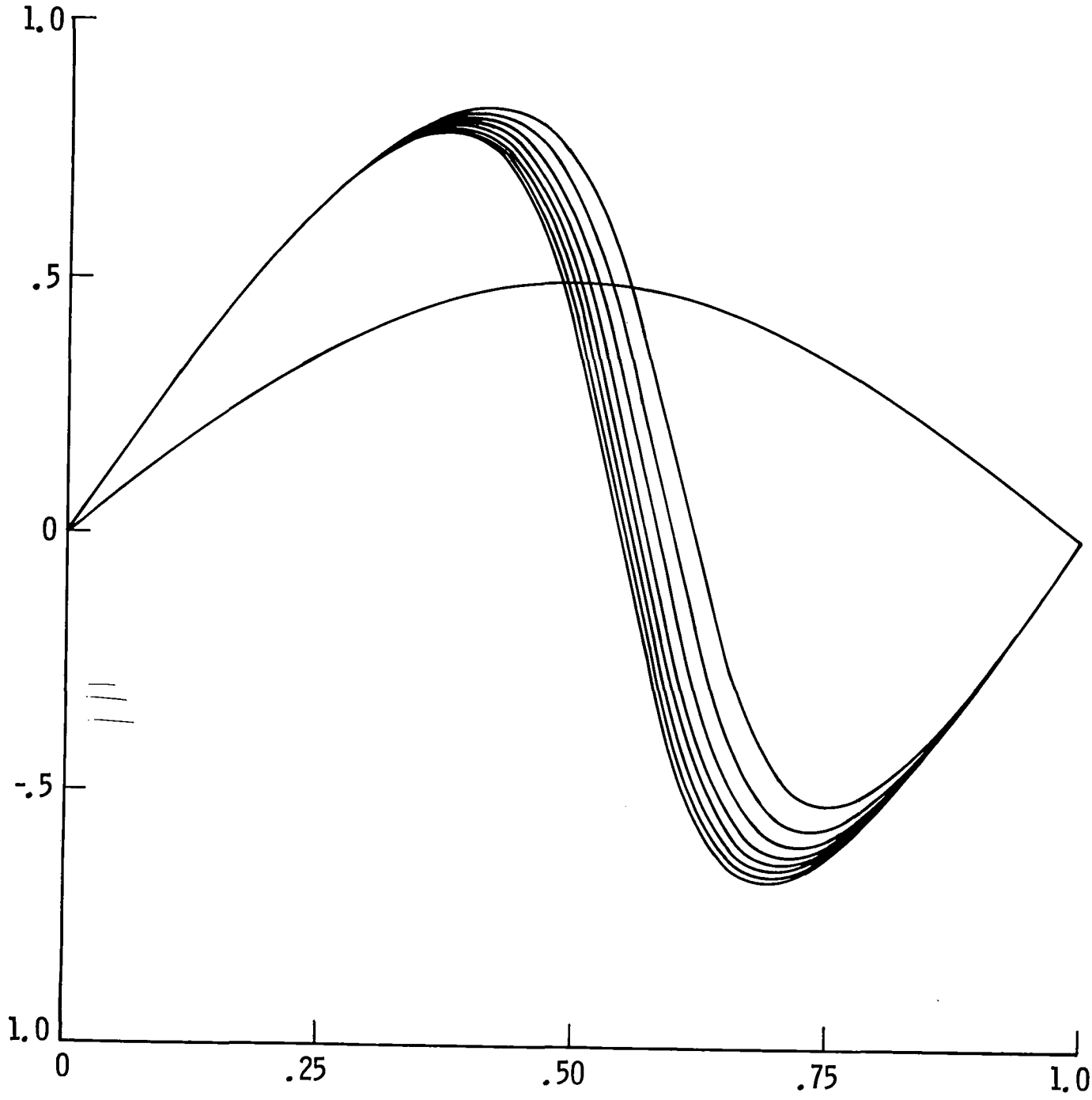


Figure 3. Convergence in time without convergence acceleration. Numerical solutions at different time stages for the case $\varepsilon = 0.04$, $f = \frac{\pi}{4}\sin(\pi x)\cos(\pi x)$, $a = b = 0$, $u(x,0) = \frac{1}{2}\sin(\pi x)$. Between each curve there are 100 time steps . The calculation is made with time step k = 0.1 and N=50 grid points.

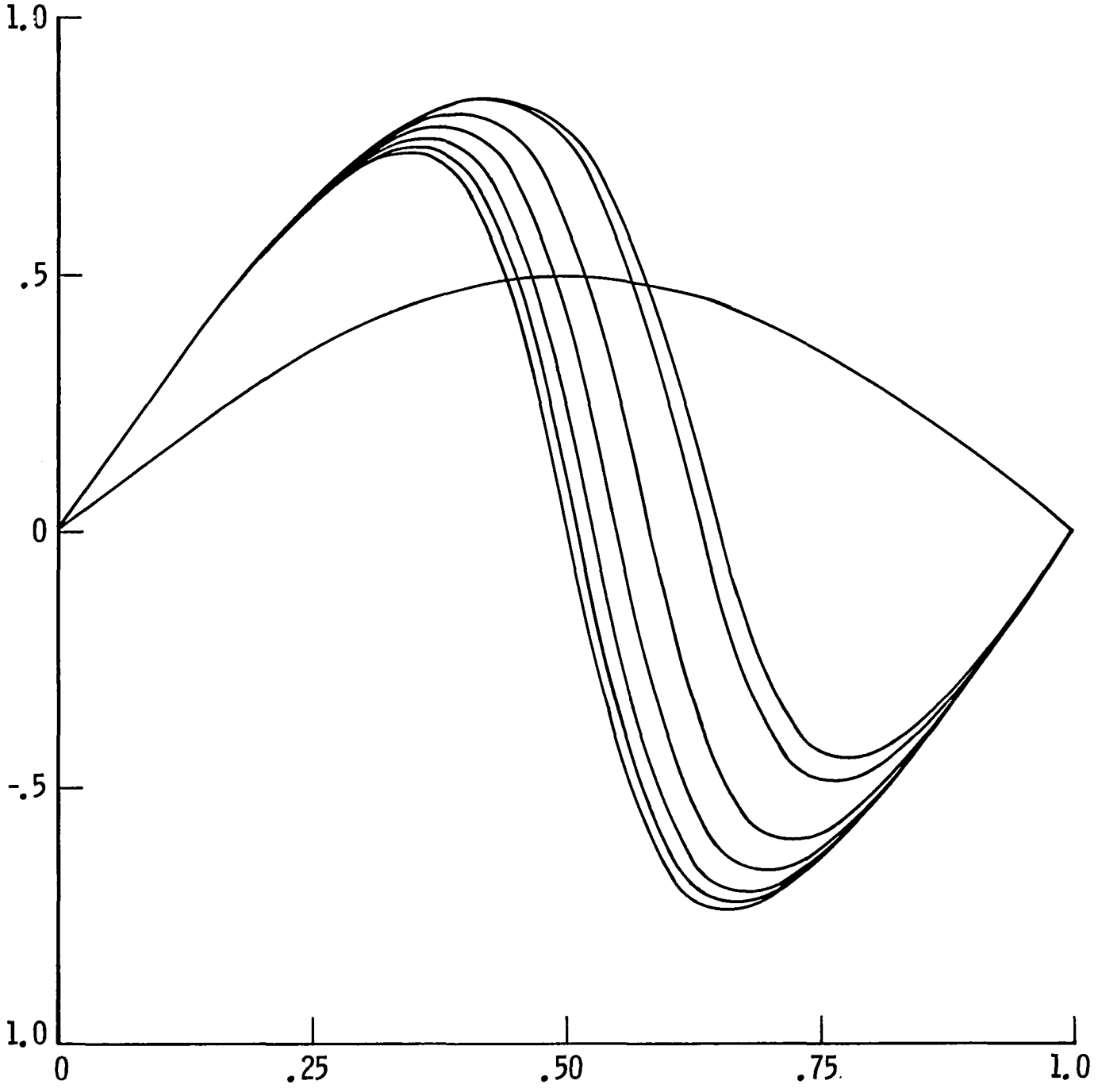


Figure 4. Convergence in time with convergence acceleration. Numerical solutions at different time stages for the same case as in figure 3. Between each curve there are 20 time steps and one extrapolation step. The same time step, k=0.1, and number of grid points , N=50, are used.

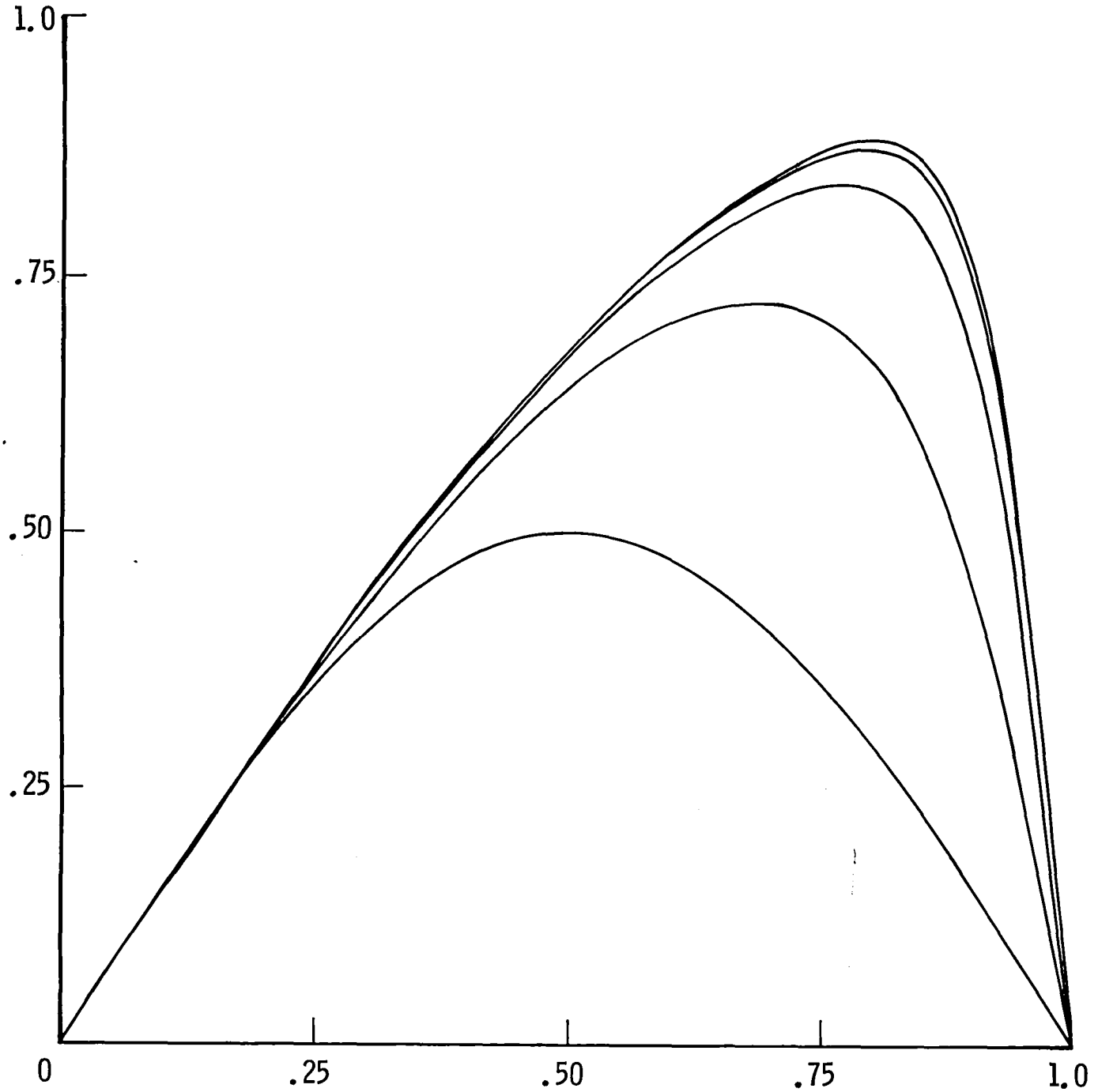


Figure 5. Convergence when the shock is located at the boundary. Here $\varepsilon = 0.04$, $f(x) = \frac{\pi}{4}\sin(\pi x)$, $u(x,0) = \frac{1}{2}\sin(\pi x)$, $N = 50, k = 0.1$. Between each curve there are 5 time steps.

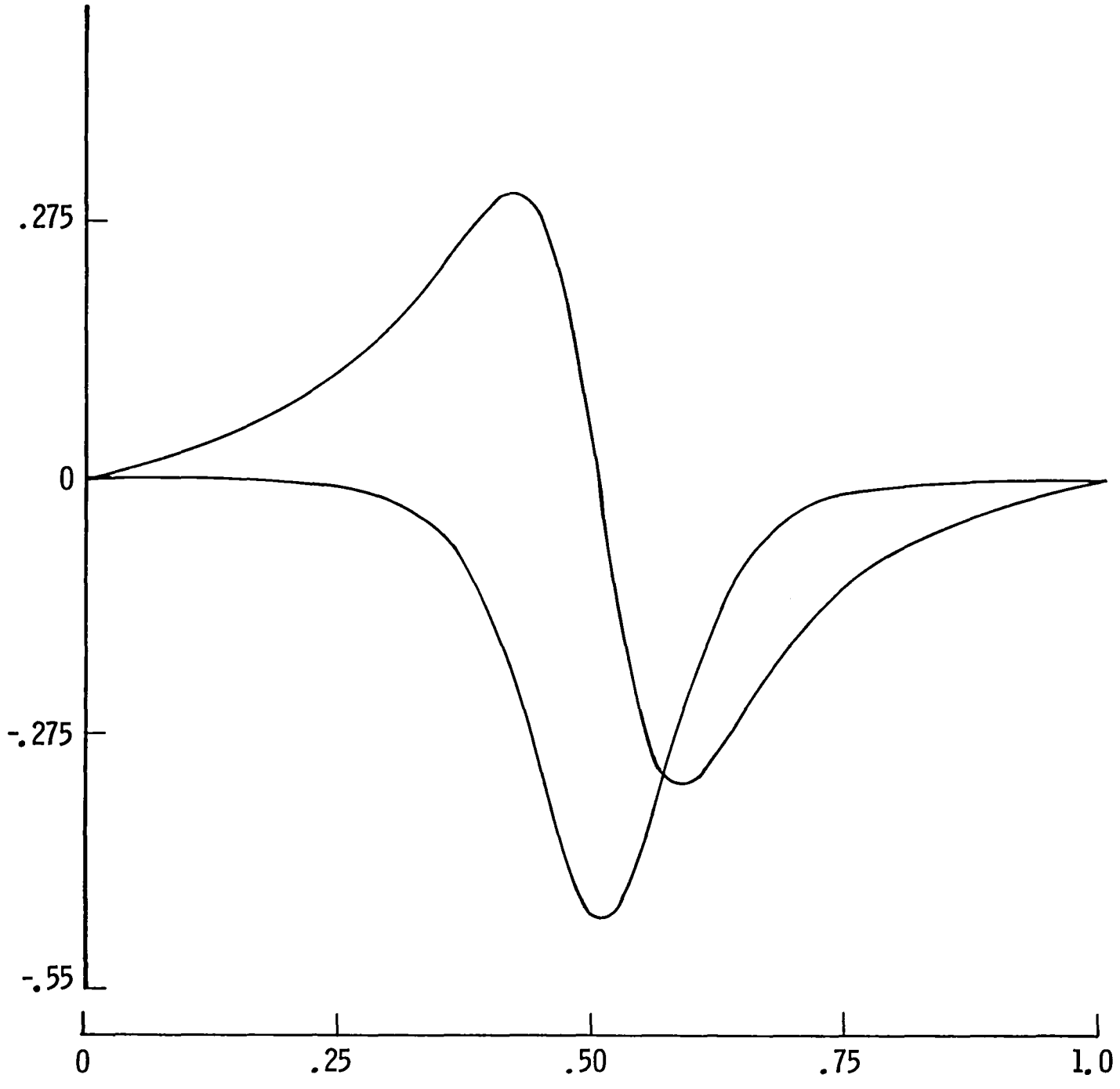


Figure 6a. Eigenvectors. The first two eigenfunctions of problem (3.2), when y, the solution of (1.3), has a shock in the interior. In this case $\varepsilon = 0.04$, $f(x) = \frac{\pi}{4}\sin(\pi x)\cos(\pi x)$, $a = b = 0$, $N = 100$.

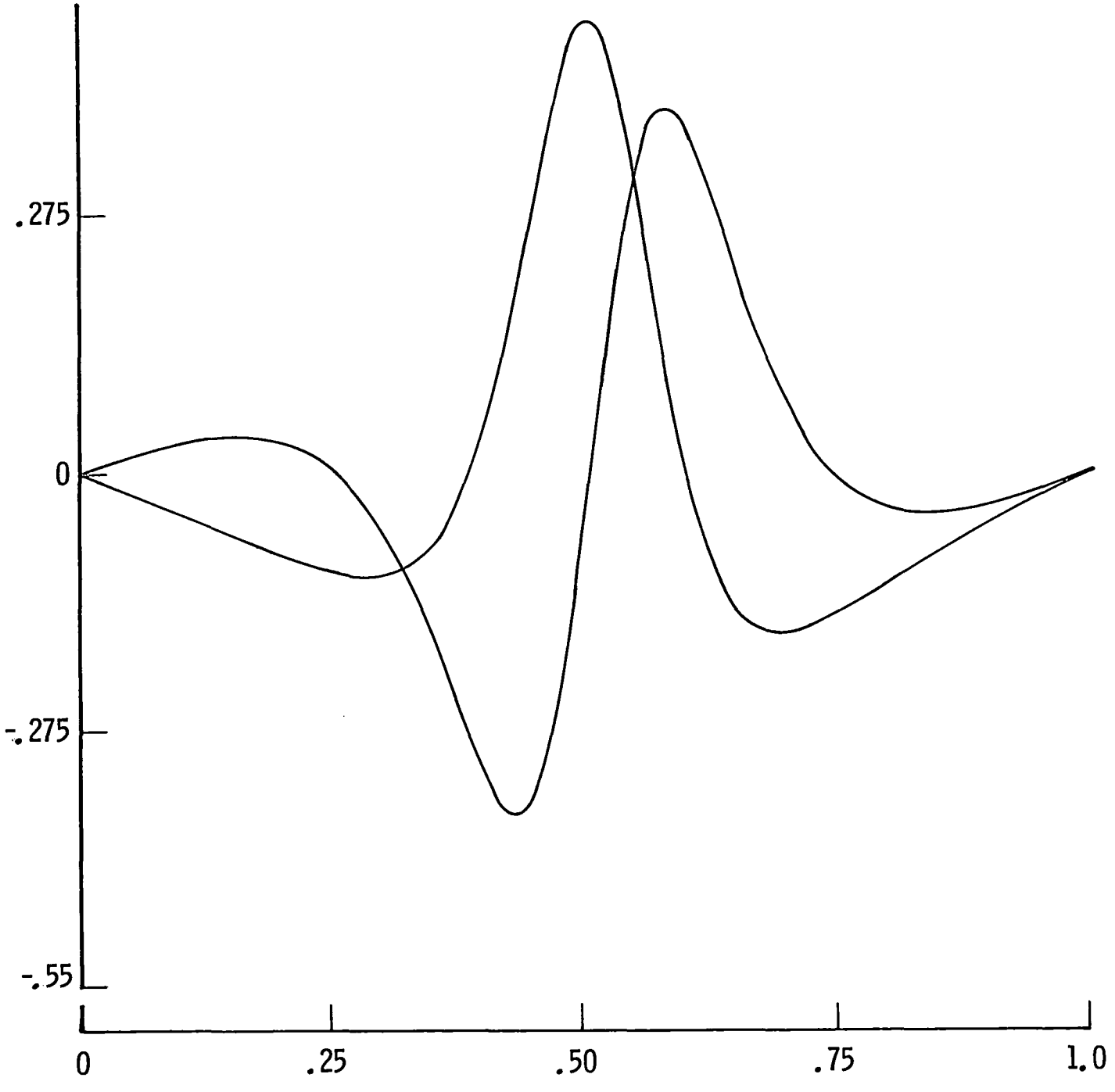


Figure 6b. Eigenvectors. The third and fourth eigenfunctions of problem (3.2), when y, the solution of (1.3), has a shock in the interior. In this case $\varepsilon = 0.04$, $f(x) = \frac{\pi}{4}\sin(\pi x)\cos(\pi x)$, $a = b = 0$, $N = 100$.

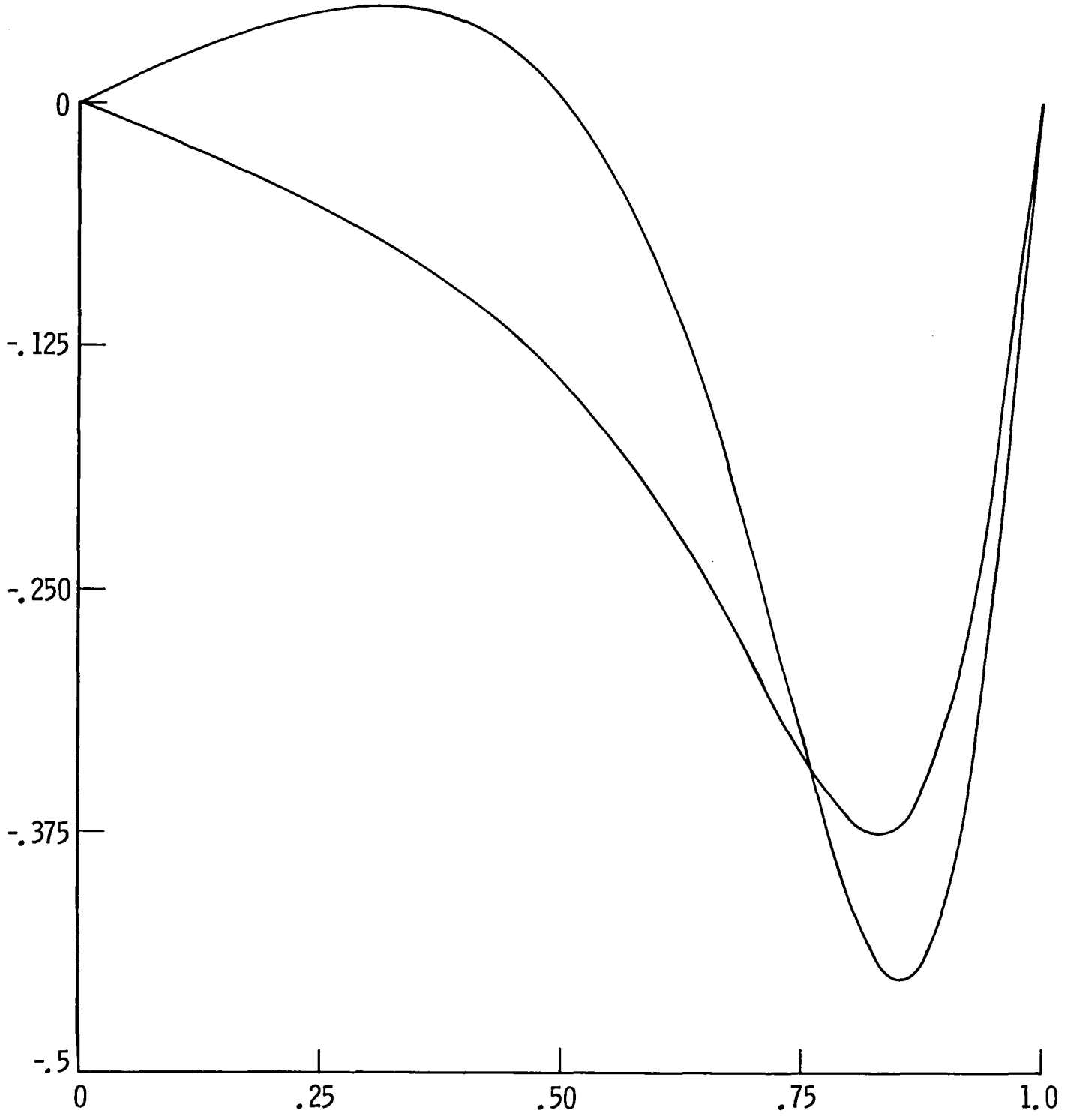


Figure 7. Eigenvectors. The first two eigenvectors, φ_1 and φ_2, of problem (3.2), when y, the solution of (1.3), has a shock $x = 1$. In this case $\varepsilon = 0.08$, $f(x) = \frac{\pi}{4}\sin(\pi x)$, $a = b = 0$, $N = 100$.

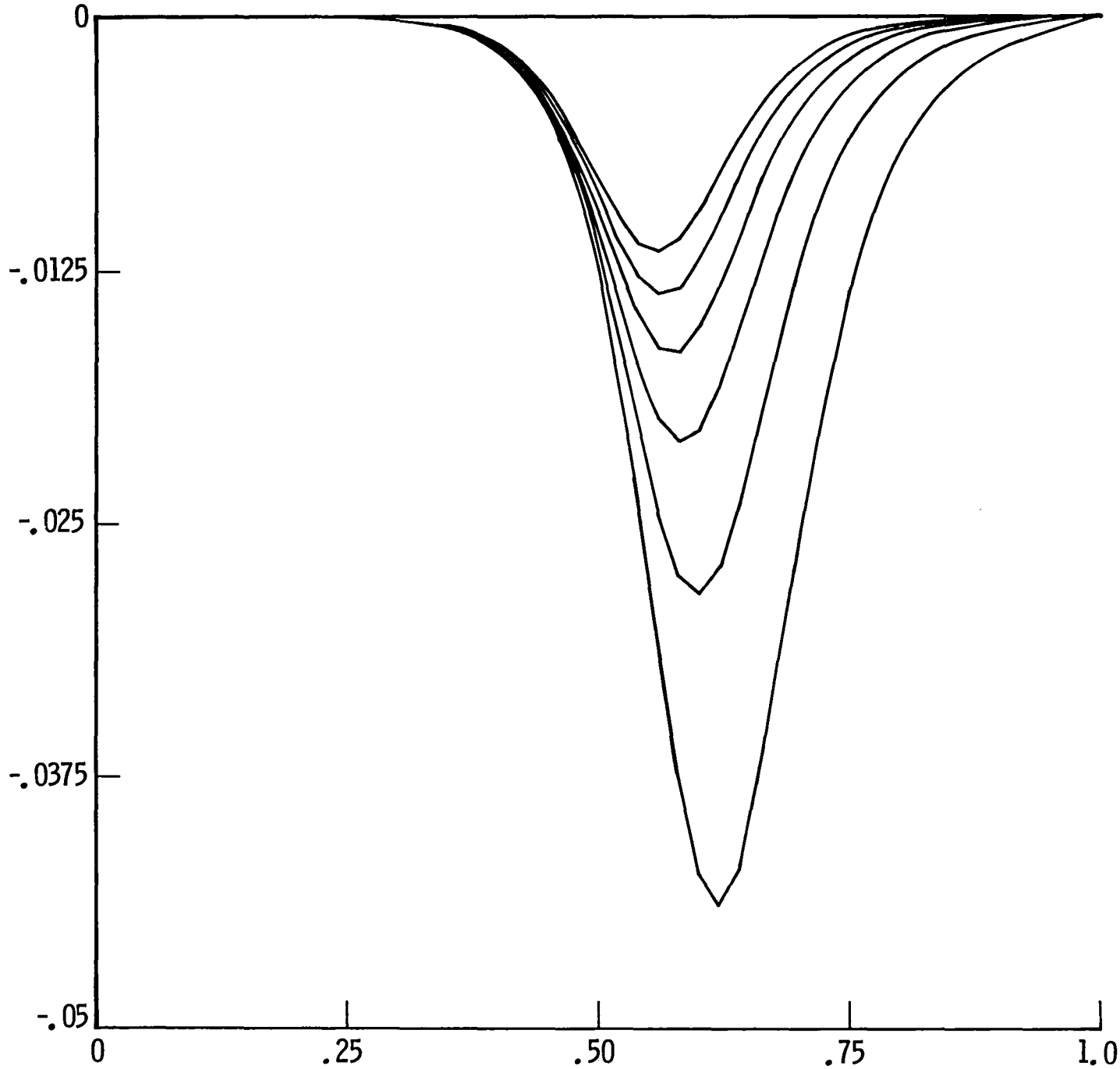


Figure 8. Differences between consecutive solutions at different time stages, when $\varepsilon = 0.04$, $f = \frac{\pi}{4}\sin(\pi x)\cos(\pi x)$, $a = b = 0$, $u(x,0) = \frac{1}{2}\sin(\pi x)$. Between each curve there are 100 time steps . The calculation is made with time step k = 0.1 and N=50 grid points.

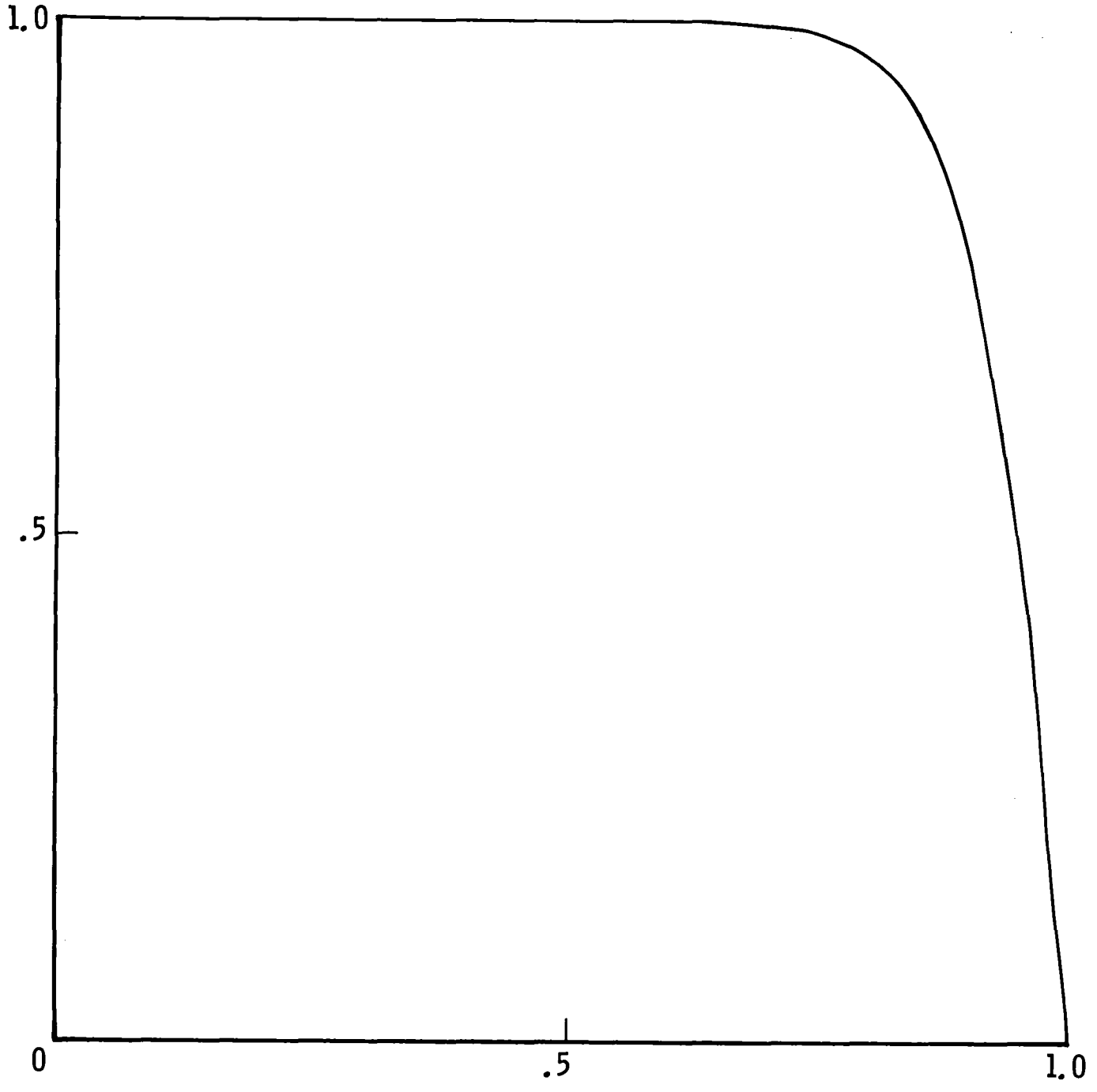


Figure 9. The solution of (1.2) when $f \equiv 0$, $a = 1$, $b = 0$ and $\varepsilon = 0.05$.

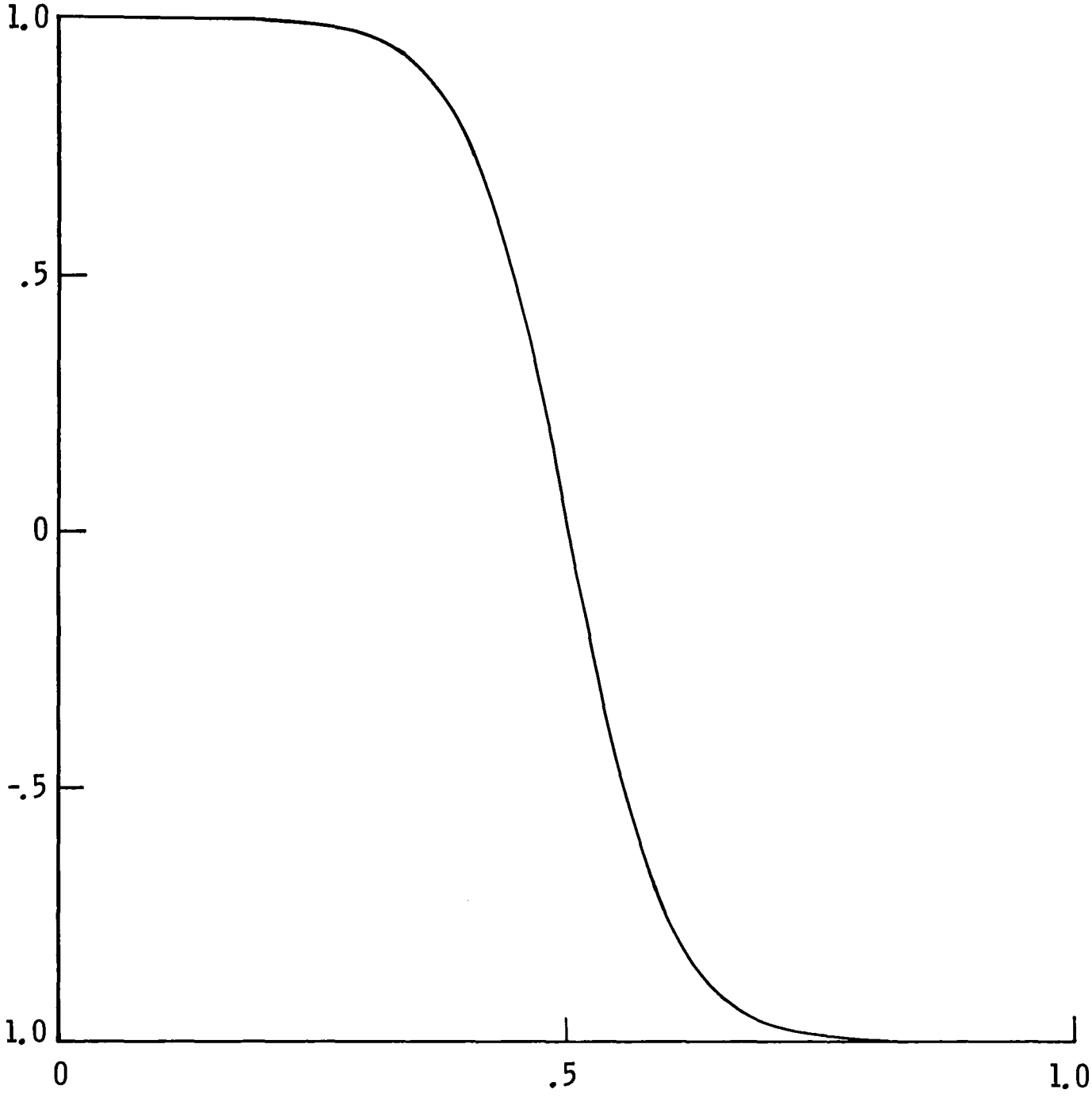


Figure 10. The solution of (1.2) when $f \equiv 0$, $a = 1$, $b = -1$ and $\varepsilon = 0.05$.

STABILITY ANALYSIS OF INTERMEDIATE BOUNDARY CONDITIONS IN APPROXIMATE FACTORIZATION SCHEMES

Jerry C. South, Jr.
NASA Langley Research Center

Mohamed M. Hafez
University of California, Davis

David Gottlieb
Brown University

Abstract

The paper discusses the role of the intermediate boundary condition in the AF2 scheme used by Holst for simulation of the transonic full potential equation. We show that the treatment suggested by Holst led to a restriction on the time step and suggest ways to overcome this restriction. The discussion is based on the theory developed by Gustafsson, Kreiss, and Sundström and also on the von Neumann method.

Research for the third author was supported in part by the National Aeronautics and Space Administration under NASA Contract Nos. NAS1-17070 and NAS1-18107 and under AFOSR 85-0303 while he was in residence at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA 23665-5225.

INTRODUCTION

Approximate factorization schemes are widely used to obtain efficient solutions to problems in Computational Fluid Dynamics. In many cases, they have provided a significant increase in efficiency over previously-used solution methods in particular problems. Some outstanding examples are the classical Alternating-Direction-Implicit method of Peaceman and Rachford [1], the Briley-McDonald Linearized Block Implicit scheme [2], and the Beam and Warming [3] Approximate Factorization (AF) scheme for the compressible Navier-Stokes equations. In the transonic potential-flow area, some AF schemes which have significantly improved solution efficiency are the work of Ballhaus and Steger [4], Ballhaus et al. [5], Holst [6], [7], and Jameson [8].

All of these schemes have the common feature that the solution procedure is broken down into a sequence of easily-implemented stages; i.e., easily-inverted matrix factors. Each of the stages usually requires boundary conditions for an "intermediate" variable (vector) which is not always a consistent approximation to the solution function desired. This feature can make satisfaction of implicit boundary conditions difficult, at best, and impossible, at worst. Dwoyer and Thames [9] demonstrated serious boundary-condition problems associated with the class of AF schemes called "Locally One-Dimensional," even in explicit schemes.

The present paper further highlights the importance of intermediate boundary conditions by focusing on a specific example--a boundary-induced stability restriction in Holst's AF2 scheme [6] for the transonic full-potential equation. An analysis of the effect of the intermediate boundary condition is given by use of the usual von Neumann method and also the methods of Gustaffson, Kreiss, and Sundstrom [10] and Osher [11].

ANALYSIS

Holst's scheme is a variation of the AF2 schemes presented in References 4 and 5. It will be referred to herein as "AF2Y," since in its implementation the y-operator is split, rather than splitting the x-operator as in References 4 and 5. For the purpose of analyzing the intermediate boundary-condition problem, it is illuminating to study the application of AF2Y to the two-dimensional (2-D) Laplace's equation in a rectangle. The present analysis is valid only for the subsonic flow condition, which is simpler by far than the transonic case. However, it is reasonable to assume that if boundary-induced instability is present in the subsonic case, it will also occur in the transonic case. In practice this was true.

The Discrete Problem

The following thin-airfoil problem is thus considered: We wish to solve the Laplace difference equation for the disturbance velocity potential

$$L\phi_{jk} = (a\delta_{xx} + b\delta_{yy})\phi_{jk} = 0 \tag{1}$$

where a and b are constant coefficients and δ_{xx} and δ_{yy} are central difference operators; e.g.,

$$\delta_{xx}\phi_{jk} = \phi_{j+1,k} - 2\phi_{jk} + \phi_{j-1,k}. \tag{2}$$

The boundary conditions are set on a rectangular region with Dirichlet conditions, $\phi = 0$, set on three sides (left, top, and right), representing vanishing disturbances, and a Neumann condition at the bottom boundary,

representing a thin-airfoil flow-tangency condition:

$$\phi_y = s(x) \qquad \text{at} \quad y = 0. \tag{3}$$

A discrete analog of Eq. (3) at $k = 1$ can be written as:

$$(\overrightarrow{\delta}_y + \overleftarrow{\delta}_y)\phi_{j,1} = 2\Delta y s(x) \tag{4}$$

where we use the following notation for one-sided, two-point differences:

$$\overrightarrow{\delta}_y \phi_{jk} = \phi_{j,k+1} - \phi_{jk} \tag{5}$$

$$\overleftarrow{\delta}_y \phi_{jk} = \phi_{jk} - \phi_{j,k-1}. \tag{6}$$

The difference operator (1) requires evaluation of $\delta_{yy}\phi_{j,1}$ at the boundary $k = 1$. Since this operator can be written as:

$$\delta_{yy} = \overrightarrow{\delta}_y - \overleftarrow{\delta}_y, \tag{7}$$

Equation (4) is used to eliminate $\overleftarrow{\delta}_y \phi_{j,1}$, which calls for a value of ϕ_{jk} below the boundary $k = 1$. Thus, the difference operator at $k = 1$ is:

$$L_B \phi_{j,1} = (a\delta_{xx} + 2b\overrightarrow{\delta}_y)\phi_{j,1} - 2b\Delta y s(x). \tag{8}$$

The AF2Y Scheme

The AF2Y scheme models a hyperbolic equation, $\sigma\phi_{yt} = \nabla^2\phi$, and is used as an iteration scheme:

$$(\alpha + b_1\overset{+}{\delta}_y)(-\alpha b_2\overset{+}{\delta}_y - a\delta_{xx})\Delta\phi_{jk} = \alpha\omega L\phi^n_{jk} \tag{9}$$

where n is the iteration counter,

$$b_1 b_2 = b \tag{10}$$

and $\Delta\phi$ is the correction

$$\Delta\phi_{jk} = \phi^{n+1}_{jk} - \phi^n_{jk}. \tag{11}$$

The scheme is implemented in two stages:

$$(\alpha + b_1\overset{+}{\delta}_y)f_{jk} = \alpha\omega L\phi^n_{jk} \tag{12}$$

$$(-\alpha b_2\overset{+}{\delta}_y - a\delta_{xx})\Delta\phi_{jk} = f_{jk}. \tag{13}$$

The intermediate variable f is defined by Eq. (13). The parameter α corresponds to a reciprocal "time" step, Δt^{-1}, and is usually cycled between small and large values to obtain rapid convergence. The parameter ω corresponds roughly to a relaxation factor which is usually close to 2.

The first stage (12) is bidiagonal, proceeding from the bottom boundary, $k = 1$, to the last interior row of mesh points, $k = K - 1$, for every j. The

second stage (13) is a tridiagonal solution which proceeds row-by-row, from $k = K - 1$ to $k = 1$, to obtain the correction $\Delta\phi_{jk}$. The second stage is initiated with the condition $\Delta\phi_{j,K} = 0$, corresponding to the vanishing disturbance, $\phi = 0$, at $k = K$.

The Intermediate Boundary Condition

The main problem in implementing the scheme is how to initiate the bidiagonal solution for f at $k = 1$. It seems reasonable, at first sight, to use a derivative condition on f at the boundary, as Holst [6] did; i.e.,

$$\overset{\leftrightarrow}{\delta}_y f_{j,1} = 0. \tag{14}$$

Comparison of Eqs. (14) and (12) implies that

$$f_{j,1} = \omega L_B \phi^n_{j,1}. \tag{15}$$

If this procedure is used with no further modification, it is unstable for small values of α (or large "time" steps) and fixed ω as described next.

Stability Analysis

A von Neumann (VN) analysis shows that the interior scheme (9) is stable for all modes under the restrictions

$$0 < \omega < 2 \tag{16}$$

$$\alpha > 0. \tag{17}$$

However, the boundary scheme, implied by Eqs. (15) and (13) taken together, is another matter.

A boundary condition more general than Eq. (14) for f can be considered. Let a "dummy-point" value for f be given as:

$$f_{j,0} = \gamma f_{j,1}. \tag{18}$$

Then the equation for $f_{j,1}$ is, from Eq. (12),

$$(\alpha + b_1)f_{j,1} = \alpha\omega L_B\phi^n_{j,1} + \gamma b_1 f_{j,1} \tag{19}$$

and Eq. (13) yields:

$$f_{j,1} = (-\alpha b_2\overset{\leftrightarrow}{\delta}_y - a\delta_{xx})\Delta\phi_{j,1} = \Omega L_B\phi^n_{j,1} \tag{20}$$

where

$$\Omega = \frac{\alpha\omega}{\alpha+b_1(1-\gamma)}. \tag{21}$$

To carry out a VN analysis, we substitute into Eq. (20) trial solutions

$$\phi^n_{j,k} = G^n\, e^{i(jp\Delta x+kq\Delta y)} \tag{22}$$

where $i = \sqrt{-1}$, p and q are wave numbers, and G is the amplification factor, to obtain:

$$(\alpha B + 2Ab_1 - i\alpha E)(G - 1) = -2\Omega b_1(A + B - iE) \tag{23}$$

where

$$\left.\begin{aligned} A &= a\,(1 - \cos\xi) \geq 0 \\ B &= b\,(1 - \cos\eta) \geq 0 \\ E &= b \sin\eta \\ \xi &= p\Delta x \\ \eta &= q\Delta y \end{aligned}\right\} \tag{24}$$

The stability condition, $|G|^2 < 1$, reduces to:

$$\Omega\{(A + B)[(2 - \Omega)b_1 A + (\alpha - \Omega b_1)B] + (\alpha - \Omega b_1)E^2\} > 0. \tag{25}$$

To maintain the inequality (25) the following stability restrictions are easily deduced:

$$0 < \Omega < 2 \tag{26}$$

$$\alpha > b_1\,\Omega. \tag{27}$$

For the case $\gamma = 1$, corresponding to the backward-Neumann condition on f (Eq. (14)), restrictions (26) and (27) reduce to Eq. (16) and

$$\alpha > b_1\,\omega \qquad (\gamma = 1). \tag{28}$$

The restriction (27) enforces a "time" step limitation on the scheme for fixed Ω, which will slow convergence; or a reduction in Ω, according to:

$$\Omega < \min\left(2, \frac{\alpha}{b_1}\right) \tag{29}$$

which in fact yields fast convergence and ensures stability.

It is noted that another useful type of boundary condition for f is given by

$$f_{j,0} = \frac{\alpha\beta}{b_1} L_B \phi^n_{j,1} \tag{30}$$

which gives the same form as Eq. (20) for $f_{j,1}$ with

$$\Omega = \frac{\alpha(\omega + \beta)}{\alpha + b_1} . \tag{31}$$

Both classes of schemes are implemented by initiating the bidiagonal march for f using Eq. (20), under restriction (29).

The restriction (29) was verified numerically in both a constant-coefficient, Cartesian-coordinate computer code for Laplace's equation and in the "TAIR" code [12] by using fixed values for α (i.e., no α-cycling) and Ω, and for various values of the coefficient b_1. In all cases, convergence was obtained when the restriction (29) was obeyed; and divergence occurred when it was violated.

The experiments with the TAIR code were especially interesting, since the coefficient b_1 varies along the airfoil surface. The test case chosen was the default "0"-type mesh for an NACA 0012 airfoil. It was found that the arithmetic mean of b_1 along the surface presented the crucial condition, rather than the maximum value, as might be expected.

The question arises as to why the TAIR code, which implements the AF2Y scheme with the boundary condition (14), operates so well since α is cycled between small values, which violate the restriction (28) and large values. The answer seems to be that α is increased within several rows adjacent to

the boundary to a value which (in the default mesh) meets the restriction (28), when smaller values of α are used in the remaining interior field. This "fix" was developed empirically by the authors of Reference 12; without this fix the code diverges. This procedure is not recommended in general, since it requires a discontinuous change in α. The assumption in the development of the factored scheme (9) is that α is constant throughout the mesh.

A seemingly attractive scheme, involving a discontinuity in α at the boundary, is as follows: Initiate the solution for f using Eq. (15) with $\omega = 1$, and change the second stage (13) at the boundary to:

$$\left(-2b\overset{\rightarrow}{\delta}_y - a\delta_{xx}\right) \Delta\phi_{j,1} = f_{j,1} = L_B^n \phi_{j,1}. \tag{32}$$

This procedure exactly annihilates the boundary residual (in the linear case) and represents a fully implicit satisfaction of the surface boundary condition. However, the factored operator at line $k = 2$ is no longer the interior-point operator, since the term $-\alpha b_2 \overset{\rightarrow}{\delta}_y$ in the inner factor is changed to $-2b\overset{\rightarrow}{\delta}_y$ discontinuously. It is possible to analyze such a scheme by the methods presented herein, but the line $k = 2$ must be considered as part of the boundary scheme. No details will be given here, but the analysis shows that setting $\omega < 4/3$ at $k = 2$ guarantees linear stability of the overall scheme. However, the amplification factor modulus $|G|$ exceeds unity only in a narrow frequency range of small η (Eq. (24)) when $\omega > 4/3$. Numerical experiments showed no sensitivity to the value of ω at $k = 2$. This scheme was always stable in tests with a constant-coefficient Cartesian-mesh code, even with $\omega = 1.8$ at $k = 2$. If the scheme was unstable for

highly stretched grids, setting $\omega < 4/3$ at $k = 2$ did not stabilize the scheme. In the variable-coefficient, nonlinear case, such a scheme is no faster than, and not as robust as, the scheme (20) with restriction (29).

Review of the Stability Theory

It is well known that in general the von Neumann analysis at a single line is neither sufficient nor necessary for checking stability. Trapp and Ramshaw [13] pointed out the usefulness of the VN analysis to study boundary schemes but recognized that no theoretical justification was known.

We wish to review briefly the stability theory for finite-difference approximations to initial boundary-value problems. A necessary condition for the stability of such a scheme is the Ryabenkii-Godunov condition. It states that the numerical scheme is unstable if there exists a solution of the type

$$\phi^n_{j,k} = G^n \psi_{j,k}, \qquad |G| > 1 \tag{33}$$

for the inner scheme and the boundary scheme. (It is also sufficient to check one boundary at a time.) Substituting (33) into (12) and (13) one finds that $\psi_{j,k}$ satisfies a constant coefficient second-order difference scheme whose solution is

$$\psi_{j,k} = \mu^k e^{i(j\rho\delta x)}. \tag{34}$$

Actually there are two possible μ's, but it is readily verified that only one of them satisfies $|\mu| < 1$ for $|G| > 1$; and, therefore, it is not a valid solution for the quarter-plane problem.

In Appendix A we show that there exists a solution of the form (34) to (12), (13), and (20) such that $|G| > 1$ and $|\mu| < 1$ if (29) is not satisfied. This proves that the scheme is unstable. By instability here we mean that unbounded solution occurs after a fixed number of time steps for any mesh--it precludes the possibility of reaching steady state.

It should be noted here that VN analysis of the boundary scheme does not predict the existence of solutions of the form (33) with $|G| > 1$. In fact, Gottlieb and Turkel [15] gave an example of a boundary scheme (Scheme VI, p. 184 of Reference 15) coupled with a variant of MacCormack's scheme in the interior which is conditionally stable, yet the VN analysis of the boundary scheme shows unconditional instability. However, Goldberg and Tadmor showed that for a dissipative interior scheme (i.e., amplification--factor modulus bounded away from unity for all nonzero modes) VN stability of the boundary scheme excludes the possibility of an eigenvalue or a generalized eigenvalue. By an eigenvalue we mean a solution of the form (34) with $|G| > 1$ whereas a generalized eigenvalue is G such that $|G| = 1$. Thus, if the condition stated in (29) is satisfied no eigenvalue or generalized eigenvalue exists. In Appendix A we show it directly. The theory of Gustafsson, Kreiss, and Sundström [10] (see also, Osher [11]) states that for a system of first-order hyperbolic equations stability is assured if there is no eigenvalue or generalized eigenvalue. While their theory does not apply directly to the equation

$$\sigma\rho_{yt} = \nabla^2 \rho,$$

it can be modified to include this case.

As a concluding remark we should note that stability here implies convergence in the sense of Lax--the numerical solution converges to the analytic one as the mesh size tends to zero for fixed time t. This is clearly only a necessary requirement to reach steady state.

Two-Dimensional Numerical Results

A limited number of numerical tests for cases involving stretched grids and nonlinear transonic flow have convinced us that the discontinuous-α schemes (e.g., Eq. (32)) are not as reliable as the scheme using Eq. (20) with restriction (29). Some numerical results are presented in Tables 1 and 2. In the tables, the following identification is used for the various boundary schemes:

Scheme I: Original TAIR scheme; Eqs. (13) and (15) at boundary, with α increased at 3 lines adjacent to boundary to satisfy restriction (28) with 10% safety margin.

Scheme II: Exact annihilation of boundary residual; Eqs. (15) and (32), with $\omega = 1$ at boundary only.

Scheme III: Eq. (20) and restriction (29) with 10% safety margin.

Table 1 shows a series of numerical tests for incompressible flow over a circle, with varying degrees of mesh stretching near the boundary. The TAIR code was used with $\omega = 1.8$ at all points except as noted in schemes II and III, and with the default settings for the α-cycle ($\alpha_{min} = 0.07$, $\alpha_{max} = 1.5$). The mesh contained 101 points uniformly spaced around the circle and 21 points in the radial direction with stretched spacing. The first column lists the cell aspect ratio at the boundary, $\Delta x/\Delta y$ ($= b_1$), for each case. The next

three columns show the number of iterations required to decrease the starting residual by 10^{-4} for three schemes previously discussed. Divergence is indicated by an entry "D." It is seen that scheme III is significantly less sensitive to grid stretching in the normal direction than are the discontinuous-α schemes, I and II.

Table 1. Number of Iterations to Reduce Residual by 10^{-4} Incompressible Circle Flow, 101 by 21 Mesh

$\frac{\Delta x}{\Delta y}$	Scheme		
	I	II	III
0.5	44	43	34
1	72	36	51
10	68	43	47
20	99	53	48
100	212	D	34
1000	400	D	127

As previously mentioned, the empirically-developed default settings in the TAIR code provide for an increased value of α near the surface; the default value satisfies the restriction (28) only for the first case in Table 1, $\Delta x/\Delta y = 0.5$. For that case, convergence is obtained; the scheme diverges for the other listed cases for which the default setting violates restriction (28). In scheme I, the value of α near the surface met the restriction, and convergence was obtained for all the listed cases.

It should be noted again that the stability analysis presented herein is valid only for subsonic flow, when the AF2Y scheme is guaranteed to be hyperbolic in time. When the flow becomes locally supersonic, the linearized Eq. (1) will have $a < 0$, and a term which simulates ϕ_{xt} must be added for stability [16]. The effect of including such a term (e.g., in the second factor of Eq. (9)) has not been studied at present. With that cautionary remark, we present results for transonic cases in the next table.

Table 2 presents results for two transonic cases for an NACA 0012 airfoil: (1) Zero incidence with free-stream Mach number $M_\infty = 0.85$ and (2) 2^o incidence with $M_\infty = 0.75$. All cases were run with $\omega = 1.8$, but with different α- cycles. It can be seen that there is little difference in the convergence rate among the schemes, except that scheme II is noticeably slower than schemes I or III for case (2).

Table 2. Number of Iterations to Decrease Residual by 10^{-4} for Transonic Flow. NACA 0012, Default TAIR Mesh, 149 by 30

Flow Condition	Scheme		
	I	II	III
$M_\infty = 0.85$ Zero incidence	190	174	187
$M_\infty = 0.75$ 2^o incidence	190	360	226

Three-Dimensional Version of AF2Y

A three-dimensional (3-D) version of the AF2Y scheme is presented in Ref. 7. It is different from the 2-D version discussed up to now, in that the factors are reversed in order. That is, the scheme can be written in the present context as:

$$\left(\alpha - \frac{c}{b_2}\delta_{zz}\right)\left(b_2 - \frac{a}{\alpha}\delta_{xx}\right)\left(\alpha - b_1\vec{\delta}_y\right)\Delta\phi_{jk\ell} = \alpha\omega L\phi^n_{jk\ell} + \alpha b_2\left(\alpha - b_1\vec{\delta}_y\right)\Delta\phi_{j,k-1,\ell} \quad (35)$$

where

$$L\phi_{jk\ell} = \left(a\delta_{xx} + b\delta_{yy} + c\delta_{zz}\right)\phi_{jk\ell}. \quad (36)$$

Because the factors are reversed, we will refer to this scheme as AF2YR.

Here the third coordinate direction is z, which can be thought of as the spanwise coordinate for a wing. The x- and y-coordinates are still the streamwise and normal coordinates as in the 2-D problem. The boundary operator corresponding to Eq. (8) is:

$$L_B\phi_{j,1,\ell} = \left(a\delta_{xx} + 2b\vec{\delta}_y + c\delta_{zz}\right)\phi_{j,1,\ell} - 2b\Delta ys(x). \quad (37)$$

The scheme is implemented in three stages, as follows:

1. $$\left(\alpha - \frac{c}{b_2}\delta_{zz}\right) g_{j\ell} = \alpha\omega L\phi^n_{jk\ell} + \alpha b_2 f_{j,k-1,\ell} \quad (38)$$

2. $$\left(b_2 - \frac{a}{\alpha}\delta_{xx}\right) f_{jk\ell} = g_{j\ell} \quad (39)$$

3. $$\left(\alpha - b_1\vec{\delta}_y\right)\Delta\phi_{jk\ell} = f_{jk\ell}. \quad (40)$$

The solution for f proceeds in planes, outward from the wing surface, using the tridiagonal Eqs. (38) and (39). The third stage (40) proceeds inward, solving for the correction in a bidiagonal march.

Again, the main problem is how to initiate the first stage. In Reference 7, the boundary condition used for f is

$$f_{j,0,\ell} = 0. \tag{41}$$

We can again consider the more general boundary conditions studied previously,

$$f_{j,0,\ell} = \gamma f_{j,1,\ell} \tag{42}$$

or

$$f_{j,0,\ell} = \frac{\beta}{b_2} L_B \phi_{j,1,\ell} \tag{43}$$

corresponding to Eqs. (18) and (30), respectively. Actually, condition (42) can only be approximately modeled in the 3-D problem, with some splitting error in the first two factors. That is, we can approximate Eq. (42) by solving, at $k = 1$:

$$1. \qquad \left(\alpha - \frac{c}{b_2} \delta_{zz}\right) g_{j\ell} = \alpha\omega L_B \phi^n_{j,1,\ell} \tag{44}$$

$$2. \qquad \left[(1-\gamma)\, b_2 - \frac{a}{\alpha} \delta_{xx}\right] f_{j,1,\ell} = g_{j\ell}. \tag{45}$$

Equation (43) is easily implemented by replacing ω in Eq. (38) by $\omega + \beta$ and setting $f_{j,0,\ell} = 0$.

Stability Analysis of the 3-D AF2YR Scheme

A VN analysis of the 3-D interior scheme shows that VN stability is achieved under restrictions (16) and (17). VN analysis of the boundary scheme (42) shows that sufficient conditions for stability of the VN boundary scheme are:

$$0 < \omega < 1 - \gamma \tag{46}$$

and

$$\gamma \leq 1. \tag{47}$$

The same criteria are obtained in the 2-D counterpart of the AF2YR scheme with boundary condition (18). The corresponding criteria for boundary condition (43) are:

$$0 < \omega + \beta < 1. \tag{48}$$

At this time we have no numerical experiments to test the stability and convergence of the 3-D boundary conditions (42) or (43) and the criteria (46) or (48). However, some comments about the use of AF2YR versus AF2Y are in order.

In the AF2YR scheme, the use of boundary condition (42) or (43) makes the scheme parabolic at the surface; i.e., the time-like equation at the boundary is:

$$\sigma\phi_t = \nabla^2\phi \tag{49}$$

where

$$\sigma = b_2 (1-\gamma)/\omega \tag{50}$$

for Eq. (42), and where

$$\sigma = b_2/(\omega + \beta) \tag{51}$$

for Eq. (43). In the case of AF2Y, the boundary equation remains hyperbolic, like the interior scheme, with

$$-\sigma\phi_{yt} = \nabla^2\phi \tag{52}$$

where

$$\sigma = b_2/\Omega. \tag{53}$$

It is felt that for this reason AF2Y may lead to faster convergence. It would appear that there is no difficulty in implementing such a scheme in 3-D, as:

$$1.\quad (\alpha + b_1\overset{+}{\delta}_y)\, f_{jk\ell} = \alpha\omega L\phi_{jk\ell} \tag{54}$$

$$2.\quad (\alpha b_2 - c\delta_{zz})\, g_{j\ell} = f_{jk\ell} + \alpha b_2 \Delta\phi_{j,k+1,\ell} \tag{55}$$

$$3.\quad \left(1 - \frac{a}{\alpha b_2}\delta_{xx}\right) \Delta\phi_{jk\ell} = g_{j\ell}. \tag{56}$$

The factors in the second and third stages could also be interchanged. The first stage is initiated by using Eq. (20), and the same stability and restrictions (26) and (27) hold.

CONCLUDING REMARKS

We have studied the stability of the AF2Y scheme with several boundary conditions for the intermediate variable. The von Neumann method provides a useful tool for this study in view of the Goldberg-Tadmor theorem, and the results were verified in the two-dimensional case by the more complete GKSO theory.

In general, the boundary schemes place a limitation on α and ω which is more restrictive than the requirements for the interior scheme. Since small α is desirable to damp low-frequency errors, one strategy involves increasing α at or near the boundary to meet the boundary restriction while using smaller α in the interior mesh. Such "discontinuous-α" schemes require further analysis of the stability at the line next to the discontinuity since the scheme there is no longer the interior scheme. They diverge on certain stretched grids. A safer strategy is to decrease ω at the boundary to conform to the restrictions. This results in a more robust scheme; and it does not appear to suffer much, if any, loss in convergence rate.

In regard to the 3-D AF2Y scheme, the current implementation in the TWING code involves a reversal of the factors from the 2-D TAIR code. We refer to this scheme as AF2YR. Although the reversal of the factors makes no difference in the interior (for the linear constant-coefficient case), there is a significant difference at the boundary. The AF2YR boundary scheme is parabolic in time as opposed to hyperbolic for AF2Y. For this reason, there may be a preference for the AF2Y, as in the TAIR code.

APPENDIX A

Application of the GKSO Theory to the AF2Y Scheme

In the GKSO theory [10], [11], the interior and boundary schemes are considered as a coupled problem. Instead of substituting the Fourier solutions as in Eq. (22), the class of trial solutions is extended to

$$\phi_{jk}^{n} = G^{n} e^{i(jp\Delta x)} \mu^{k} \quad \text{(A1)}$$

where μ is a complex number not restricted to lie on the unit circle in the complex plane. Fourier modes are retained in the direction tangential to the boundary under study. The trial solutions are substituted into the interior and boundary schemes, Eqs. (9) and (20), to obtain, respectively:

$$\left[\alpha + b_1\left(1 - \frac{1}{\mu}\right)\right]\left[-\alpha b_2(\mu - 1) + 2A\right](G - 1) = \alpha\omega\left[-2A + b\left(\mu - 2 + \frac{1}{\mu}\right)\right] \quad \text{(A2)}$$

and

$$\left[-\alpha b_2(\mu - 1) + 2A\right](G - 1) = \Omega\left[-2A = 2b(\mu - 1)\right], \quad \text{(A3)}$$

where Eq. (8) for $L_B\phi_{j,1}^{n}$ is used for the right-hand side of Eq. (A3) and where we have used the notation of Eq. (24).

Equations (A2) and (A3) are two simultaneous equations for the unknowns G and μ. In the theory, we are concerned only with values of μ inside the unit circle, i.e., only those solutions which decay away from the

boundary. If the solution of Eqs. (A2) and (A3) yield $G > 1$ for $\mu < 1$, the scheme is unstable.

If Eq. (A3) is divided into Eq. (A2), G is eliminated; and there results an equation for μ:

$$\Omega[\alpha\mu + b_1(\mu - 1)][-2A + 2b\ (\mu - 1)] = \alpha\omega[-2\mu A + b(\mu - 1)^2]. \qquad (A4)$$

First, it will be shown that for

$$A = a(1 - \cos\ \xi) = 0\ ,$$

there is a value of μ inside the unit circle. When $A = 0$, Eq. (A4) reduces to two linear factors:

$$(\mu - 1)\{[2\Omega\ (a + b_1) - \alpha\omega]\mu - 2b_1\ \Omega + \alpha\omega\} = 0. \qquad (A5)$$

The root $\mu = 1$ is a solution of Eqs. (A2) and (A3) only when $\Omega = \omega$, corresponding to $\gamma = 1$. (See Eq. (21).) Then

$$G = 1 - \omega \qquad (A6)$$

and the restriction (16) must be satisfied. The other root is:

$$\mu = \frac{2b_1\Omega - \alpha\omega}{2\Omega\ (\alpha + b_1) - \alpha\omega} \qquad (A7)$$

which is less than 1.0 and is arbitrarily close to 1.0 as α approaches zero.

Using Eq. (A3), we can show that for any complex μ such that its real part is less than 1, $G < 1$ if and only if restrictions (26) and (27) are satisfied. Thus, let

$$\mu = \mu_R + i\mu_I \tag{A8}$$

where μ_R and μ_I are the real and imaginary parts. Substitution of Eq. (A8) into (A3) and multiplication by b_1 gives:

$$[\alpha b\,(1 - \mu_R) + 2Ab_1 - i\alpha b\mu_I](G - 1) = -\,2\Omega b_1[A + b\,(1 - \mu_R) - ib\mu_I]. \tag{A9}$$

The condition $G^2 < 1$ then yields:

$$\Omega\{A^2 b_1(2 - \Omega) + Ab(1 - \mu_R)[\alpha - \Omega b_1 + (2 - \Omega)b_1]$$

$$+ b^2(\alpha - \Omega b_1)[(1 - \mu_R)^2 + \mu_I^2]\} > 0. \tag{A10}$$

For $\mu_R < 1$, the restrictions (26) and (27) are sufficient to ensure the inequality (A10) for arbitrary positive values of A, b_1, and b, regardless of the magnitude of μ_I. When $A = 0$, a value $\mu < 1$ always occurs, as shown by Eq. (A7); and the scheme will be unstable unless restriction (27) is satisfied. Thus, restriction (27) is necessary; and when it is satisfied (for small α), restriction (26) will also be satisfied.

Acknowledgment

We thank Dr. Terry Holst of NASA Ames Research Center for supplying the original TAIR code and suggesting the circle test case and Dr. Eitan Tadmor of ICASE for discussions of the Goldberg-Tadmor theorem.

REFERENCES

[1] D. W. Peaceman and H. H. Rachford, Jr., "The Numerical Solution of Parabolic and Elliptic Differential Equations," J. Assoc. Comput. Mach., Vol. 8, 1955, pp. 359-365.

[2] W. R. Briley, "Solution of the Three-Dimensional Compressible Navier-Stokes Equations by an Implicit Technique," Proceedings of the 4th International Conference on Numerical Methods in Fluid Dynamics, 1974.

[3] R. M. Beam and R. F. Warming, "An Implicit Factored Scheme for the Compressible Navier-Stokes Equations." AIAA J., Vol. 16, April 1978, pp. 393-402.

[4] W. F. Ballhaus and J. L. Steger, "Implicit Approximate-Factorization Schemes for the Low-Frequency Transonic Equation," NASA TMX-73082, November 1975.

[5] W. F. Ballhaus, A. Jameson, and J. Albert, "Implicit Approximate-Factorization Schemes for the Efficient Solution of Steady Transonic Flow Problems," AIAA J., Vol. 16, June 1978, pp. 573-579.

[6] T. L. Holst, "An Implicit Algorithm for the Conservative, Transonic Full Potential Equation Using an Arbitrary Mesh," AIAA J., Vol. 17, October 1979, pp. 1038-1045.

[7] T. L. Holst, and S. D. Thomas, "Numerical Solution of Transonic Wing Flow Fields," AIAA Paper 82-0105, January 1982.

[8] A. Jameson, "Acceleration of Transonic Potential Flow Calculations on Arbitrary Meshes by the Multiple Grid Method," Proceedings of the AIAA 4th Computational Fluid Dynamics Conference, Williamsburg, Va., 1979, pp. 122-146.

[9] D. L. Dwoyer and F. C. Thames, "Accuracy and Stability of Time-Split Difference Schemes," Proceedings of the AIAA 5th Computational Fluid Dynamics Conference, Palo Alto, California, pp. 101-112.

[10] B. Gustafsson, H.-O. Kreiss, and A. Sundstrom, "Stability Theory of Difference Approximations for Mixed Initial Boundary Value Problems II," Math. Comput., Vol. 26, 1972, pp. 649-686.

[11] S. Osher, "Systems of Difference Equations with General Homogeneous Boundary Conditions," Trans. Amer. Math. Soc., Vol. 137, 1969, pp. 177-201.

[12] F. C.Dougherty, T. L. Holst, K. L. Gundy, and S. D. Thomas, "TAIR - A Transonic Airfoil Analysis Computer Code," NASA TMX-81296, May 1981.

[13] J. A. Trapp and J. D. Ramshaw, "A Simple Heuristic Method for Analyzing the Effect of Boundary Conditions on Numerical Stability," J. Comput. Phys., Vol. 20, 1976, pp. 238-242.

[14] M. Goldberg and E. Tadmor, "Scheme-Independent Stability Criteria for Difference Approximations of Hyperbolic Initial Boundary Value Problems," Math. Comput., Vol. 36, April 1981, pp. 603-626.

[15] D. Gottlieb and E. Turkel, "Boundary Conditions for Multistep Finite-Difference Methods for Time-Dependent Equations," J. Comput. Phys., Vol. 26, 1978, pp. 181-196.

[16] A. Jameson, "Iterative Solution of Transonic Flows over Airfoils and Wings, Including Flows at Mach 1," Comm. Pure Appl. Math., Vol. 27, 1974, pp. 283-309.

MULTIPLE STEADY STATES FOR CHARACTERISTIC INITIAL VALUE PROBLEMS

M. D. Salas
NASA Langley Research Center

S. Abarbanel
Tel-Aviv University, Tel-Aviv, Israel
and
Institute for Computer Applications in Science and Engineering

D. Gottlieb
Tel-Aviv University, Tel-Aviv, Israel
and
Brown University

Abstract

The time dependent, isentropic, quasi-one-dimensional equations of gas dynamics and other model equations are considered under the constraint of characteristic boundary conditions. Analysis of the time evolution shows how different initial data may lead to different steady states and how seemingly anomalous behavior of the solution may be resolved. Numerical experimentation using time consistent explicit algorithms verifies the conclusions of the analysis. The use of implicit schemes with very large time steps leads to erroneous results.

Research was supported in part by the National Aeronautics and Space Administration under NASA Contract Nos. NAS1-17070 and NAS1-18107 while the second and third authors were in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665-5225. The third author was also supported by AFOSR Grant 85-0303.

INTRODUCTION

Consider a steady, isentropic flow in a dual-throat nozzle with equal throat areas, and assume that the flow is choked; then it is well known [1] that the flow between the throats can be either completely subsonic or supersonic depending on the initial state of the flow and the path taken to reach the steady state. If we experiment numerically with the above problem using either the isentropic quasi-one-dimensional gas dynamics equation or some "simpler" model equation, then some of the results obtained are rather peculiar.

(1) If the initial data correspond to sufficiently high supersonic flow (or sufficiently low subsonic flow), then the steady state flow obtained between the two throats is indeed completely supersonic (subsonic).

(2) If the initial data are completely supersonic (or subsonic), but below a certain level (above a certain level), then the steady state flow contains a shock wave connecting the supersonic branch of the solution to the subsonic branch. For the model equations considered, the shock corresponds to an isentropic jump, and its location depends on the initial data.

(3) Results (1) and (2) above are observed when time accurate schemes are used. However, the implicit backwards Euler scheme with large time steps yields steady states that are not reachable through a time accurate path from any class of nontrivial initial conditions. These steady states include not only discontinuous solutions (as observed in [2]), but also unstable smooth solutions.

(4) The numerical treatment of boundary conditions is very important in obtaining the proper results. For example, with central space differencing one may have a stable algorithm that does not converge in time to a steady state if the sonic conditions are invoked in order to supply numerical boundary conditions.

The purpose of this paper is to present our findings, and to provide, where possible, a mathematical explanation of the observed behavior, thereby removing the apparent peculiarities. We will show that the nonuniqueness aspect of the steady state solution is a by-product of the fact that the boundary conditions for the evolution equations are prescribed along characteristic curves. This is true for the dual throat problems due to the sonic conditions imposed at the throats. The model problems were therefore chosen to show this behavior.

In Section 2 we study the model equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = u(1 - u).$$

The relevance of this model equation to the quasi-one-dimensional gas dynamical equations is somewhat peripheral. However, it is rich in the number of possible steady solutions that it admits, including unstable continuous and discontinuous solutions. In this section we discuss the proper way to formulate the characteristic boundary conditions for first order quasi-linear hyperbolic equations.

In Section 3 we consider the model equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = \sin x \cos x.$$

This model equation has solutions which qualitatively behave like those of the isentropic dual throat nozzle problem. The simplicity of the model, however, affords a detailed study of the possibilities for anomalous behavior. This model equation will also show us how to quantify such vague terms as sufficiently high (or low) supersonic (subsonic) initial conditions that were mentioned in (1) and (2) above. These results are summarized in Theorems 1 and 2.

In Section 4 a model scalar equation is developed which has all of the interesting physical aspects of the complete isentropic quasi-one-dimensional gas dynamic equations governing the dual throat nozzle problem. To develop this equation, our guideline was to retain the differential equation exhibiting the characteristic boundary condition and to model the other dependent variable by assuming constant total enthalpy during the time evolution. By comparing the theoretical results of the model equation to numerical calculations for the complete system of equations, this section shows that the proposed single equation is indeed a good model of the complete system. Here, by the "goodness" of the model we mean that all of the important features of the system are retained.

Recently Kreiss and Kreiss [4] have investigated the above model equations in the presence of a linear dissipative term of the form εu_{xx}. They show that in this case the solution is unique and discuss the convergence properties of their numerical scheme.

2. FIRST EXAMPLE

Here we consider the scalar hyperbolic partial differential equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = u(1 - u), \qquad 0 \leq x \leq 1, \quad t > 0,$$
$$u(x,0) = g(x). \tag{2.1}$$

For reasons mentioned in the introduction, and to be discussed in detail in Section 4, we are interested in cases that model physical situations in which the boundaries are characteristic. In practice, when (2.1) is solved numerically as a characteristic boundary value problem, the boundary conditions are imposed dynamically as follows:

$$u(0,t) = \begin{cases} 0 \quad \text{if} \quad u(\varepsilon_0,t) > 0 & (\varepsilon_0 = \Delta x) \\ \text{unspecified if} \quad u(\varepsilon_0,t) \leq 0 \end{cases} \tag{2.2a}$$

$$u(1,t) = \begin{cases} 0 \quad \text{if} \quad u(\varepsilon_1,t) < 0 & (\varepsilon_1 = 1 - \Delta x) \\ \text{unspecified if} \quad u(\varepsilon_1,t) \geq 0 \end{cases} \tag{2.2b}$$

There are two families of continuous steady states satisfying (2.1) and the analytical versions of (2.2):

$$u = 0 \tag{2.3}$$

$$u = 1 - e^{\eta - x} \qquad (0 \leq \eta \leq 1). \tag{2.4}$$

The stability theory of ordinary differential equations applied to the characteristic equation $du/dt = u(1 - u)$ easily shows that the steady state solution $u = 0$ is unstable.

There are also weak solutions connecting various branches (different η's) of (2.4). These discontinuous solutions are unstable as will be demonstrated now. Let

$$u_L = 1 - e^{\eta_1} e^{-x} \tag{2.5}$$

be a steady state corresponding to $\eta = \eta_1$,

$$u_R = 1 - e^{\eta_2} e^{-x} \tag{2.6}$$

be another branch.

Since we want to rule out "expansion shocks," i.e., discontinuities that do not obey the "entropy condition" $u_L > 0 > u_R$, we will consider only the case of $1 \geq \eta_2 > \eta_1 \geq 0$, although the analysis is unchanged if $\eta_2 < \eta_1$. For a steady state shock we require $u_L(x_S) + u_R(x_S) = 0$. This determines the shock location, x_S, to be

$$x_S = \ell n \frac{e^{\eta_1} + e^{\eta_2}}{2} . \tag{2.7}$$

We now ask, what will be the shock speed, $\dot{x}_S = \frac{1}{2}(u_L + u_R)$, if x_S is perturbed to $x_S + \varepsilon$? Upon substituting the perturbed shock position in (2.5) and (2.6), we get for the new shock speed

$$\frac{u_L + u_R}{2} = 1 - e^{-\varepsilon} \approx \varepsilon + 0(\varepsilon^2). \tag{2.8}$$

Thus, if $\varepsilon > 0$ ($\varepsilon < 0$) the shock will move to the right (left), showing that the solution with a shock is not stable.

We have thus shown that in the steady state we need consider only the smooth solutions in (2.4). We will now demonstrate that these solutions are reachable from initial data. The demonstration is first done for the case $\eta = 0$, $g(x) > 0$ for all $x > 0$, and $g(0) = 0$.

Consider the problem (2.1), and let

$$g(x) = b(1 - e^{-x}), \qquad b > 0. \tag{2.9}$$

The solution to this problem is readily verified to be

$$u(x,t) = b\,\frac{1 - e^{-x}}{e^{-t} + b(1 - e^{-t})}\,. \tag{2.10}$$

Clearly, as $t \to \infty$, $u(x,t) \to 1 - e^{-x}$, which is a proper steady state.

Suppose now $g(x)$ is not a multiple of the steady state but is a general initial condition still satisfying $g(0) = 0$, $g(x) > 0$. The characteristic equations are

$$\frac{dx}{dt} = u \tag{2.11}$$

$$\frac{du}{dt} = u(1 - u). \tag{2.12}$$

From (2.12) one gets

$$u = \frac{g(\xi)}{g(\xi) + (1 - g(\xi))e^{-t}} \tag{2.13}$$

where $\xi = \xi(x,t)$ is the origin of the characteristic passing through x and t. By inserting (2.13) in (2.11) and integrating again along the characteristic, we get the following implicit relation between ξ, x and t:

$$e^{x-\xi-t} = [g(\xi) + (1 - g(\xi))e^{-t}] \tag{2.14}$$

or, upon rearranging

$$g(\xi) = \frac{e^{x-\xi} - 1}{e^{t} - 1} . \tag{2.15}$$

The argument is now as follows: $x - \xi$ is finite $(0 \leq x-\xi < 1)$, and thus as $t \to \infty$, $g(\xi) \to 0$, but $g(\xi) \to 0$ only for $\xi \to 0$. Hence, for any finite x, as t increases, $g(\xi)$ takes the large time asymptotic form of

$$g(\xi) \sim \frac{e^{x} - 1}{e^{t} - 1} \qquad (t \gg 1). \tag{2.16}$$

Substituting (2.16) in (2.13) we get

$$u(x,t) \sim \frac{1 - e^{-x}}{1 - e^{-t}} \qquad (t \gg 1). \tag{2.17}$$

Thus, as $t \to \infty$, $u(x,t) \to 1 - e^{-x}$ regardless of the detailed form of the initial data.

For other types of initial data (e.g., $g(x) = 0$ for some $x = x_0$), the proof is the same with $\eta = x_0$ and the coordinate x transformed to $\bar{x} = x - x_0$.

If $g(x)$ has several simple zeros, then the interval $0 \leq x \leq 1$ is subdivided by the zeros. Their relative locations will determine the proper η.

In particular, if $g(x)$ is a periodic function, $g(x_j) = 0$, with $x_j = \frac{jx}{N}$, $j = 0,1\cdots,N$, then

$$\eta = x_N = 1$$

if

$$\text{(i)} \quad \operatorname{sgn} g'(0) = \operatorname{sgn} g'(1) > 0$$

or

$$\text{(ii)} \quad \operatorname{sgn} g'(0) = -\operatorname{sgn} g'(1) < 0 \tag{2.18a}$$

and

$$\eta = x_{N-1}$$

if

$$\text{(i)} \quad \operatorname{sgn} g'(0) = \operatorname{sgn} g'(1) < 0$$

or

$$\text{(ii)} \quad \operatorname{sgn} g'(0) = -\operatorname{sgn} g'(1) > 0 \tag{2.18b}$$

(where primes denote differentiation with respect to the independent variable). In summary, this example demonstrates the richness of possible steady state solutions.

(1) There is an unstable smooth solution, $u = 0$.

(2) There are unstable discontinuous solutions.

(3) There is a one-parameter family of smooth steady states,

$$u = 1 - e^{\eta - x}$$

with the value of the parameter depending only on the initial data, a direct consequence of the problem having characteristic boundary values.

It is interesting to note that if the right-hand side of equation (2.1) is taken to be $u(u-1)$, instead of $u(1-u)$, then there is only one possible stable steady state solution satisfying the boundary conditions (2.2), namely

$$u = 0.$$

Note that this was one of the unstable solutions of the previous case.

2.1 NUMERICAL RESULTS FOR THE FIRST EXAMPLE

2.1.1 Explicit Form

The conservative, upwind, first order scheme of Engquist-Osher (E-O), [3] is used to approximate the hyperbolic system of conservation laws represented by

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = h(x,u) \tag{2.19}$$

where h is a source term. Let u_i^n represent the discrete value of u at $t^n = n\Delta t$ and $x_i = i\Delta x$. The explicit E-O scheme for equation (2.19) is,

$$u_i^{n+1} = u_i^n - \frac{1}{2}\frac{\Delta t}{\Delta x}\left[\frac{1}{2}(1 - \delta_{i+1})(u_{i+1}^n)^2 + \delta_i(u_i^n)^2 - \frac{1}{2}(1 + \delta_{i-1})(u_{i-1}^n)^2\right]$$

$$+ h(i\Delta x, u_i^n)\Delta t \tag{2.20}$$

where the switch function δ_i is defined by

$$\delta_i = \begin{cases} 0 & u_i^n = 0 \\ \dfrac{u_i^n}{|u_i^n|} & u_i^n \neq 0. \end{cases} \tag{2.21}$$

As usual, Δt satisfies the Courant-Friedrichs-Lewy condition,

$$\Delta t \leq \frac{\Delta x}{\max|u_i^n|}, \tag{2.22}$$

and $\Delta x = L/100$, where L is the length of the interval of interest. For the explicit E-O scheme convergence was established according to the criterion

$$\max_i |u_i^{n+1} - u_i^n| < 1. \times 10^{-3}. \tag{2.23}$$

The relation given by (2.23) is equivalent to requiring the steady state operator of (2.20) to be less than 10^{-3}. Figure 1 compares the exact and computed steady states for equation (2.1) with initial conditions*

$$g(x) = -\sin 2\pi x. \tag{2.24}$$

Note that the steady state satisfies the condition (2.18bi) and that the initial conditions and steady state solution are such that no boundary

---

*Note that, because of the first order accuracy of the Engquist-Osher scheme, Figure 1 shows a slight discrepancy between the analytic and numerical solution. The same problem run with $x = 1/1000$ gives results that, on the scale of Figure 1, are indistinguishable from the analytic results. This comment holds for all other numerical experiments, where, in order to save computer time, we used 100 mesh points.

conditions are imposed at either end of the interval. The same steady state is also obtained with

$$g(x) = -x(x-1)(x - \tfrac{1}{2}). \tag{2.25}$$

Figure 2 compares the exact and computed steady states for initial conditions

$$g(x) = \sin 2\pi x. \tag{2.26}$$

The steady result is in agreement with the condition (2.18ai).

2.1.2 Implicit Form

The slow convergence to steady state characteristic of explicit schemes has stimulated research into various acceleration techniques. One of the most promising avenues for acceleration consists of recasting the discrete equation in implicit form. If we define the increment in time of u by

$$\Delta u_i = u_i^{n+1} - u_i^n, \tag{2.27}$$

then the E-O scheme in implicit form is

$$\tfrac{1}{2}(1 - \delta_{i+1})u_{i+1}^n \Delta u_{i+1} + \left(\frac{\Delta x}{\Delta t} + \delta_i u_i^n - \left(\frac{\partial h}{\partial u}\right)_i^n \Delta x\right)\Delta u_i - \tfrac{1}{2}(1 + \delta_{i-1})u_{i-1}^n \Delta u_{i-1}$$

$$= -\tfrac{1}{2}\left[\tfrac{1}{2}(1 - \delta_{i+1})(u_{i+1}^n)^2 + \delta_i(u_i^n)^2 - \tfrac{1}{2}(1 + \delta_{i-1})(u_{i-1}^n)^2\right] + h(i\Delta x, u_i^n)\Delta x, \tag{2.28}$$

where δ_i is defined as before by equation (2.21). To obtain equation (2.28), terms of order Δu_i^2 and higher are neglected. It is easy to see, by comparing equations (2.20) and (2.28), that the right-hand side of equation (2.28) is the steady state operator. For the implicit E-O scheme convergence was established by requiring that the steady state operator be less than 10^{-5} at all mesh points.

Figure 3 shows the steady state solution obtained using the implicit E-O scheme with

$$g(x) = \sin 2\pi x \tag{2.29}$$

and using infinite Courant number $\left(\frac{1}{\Delta t} = 0\right)$. The steady state obtained with the implicit form of the scheme corresponds to one of the unstable solutions of equation (2.1). The stable solution, for $g(x)$ corresponding to equation (2.29), was shown in Figure 2. The peculiar behavior of the implicit algorithm at large Courant numbers is further demonstrated in Figure 4 for

$$g(x) = - x(x - 1)(x - \tfrac{1}{2}) \tag{2.30}$$

and infinite Courant number. For this case, the steady state reached by (2.28) consists of a combination of stable and unstable steady, piecewise solutions of equation (2.1).

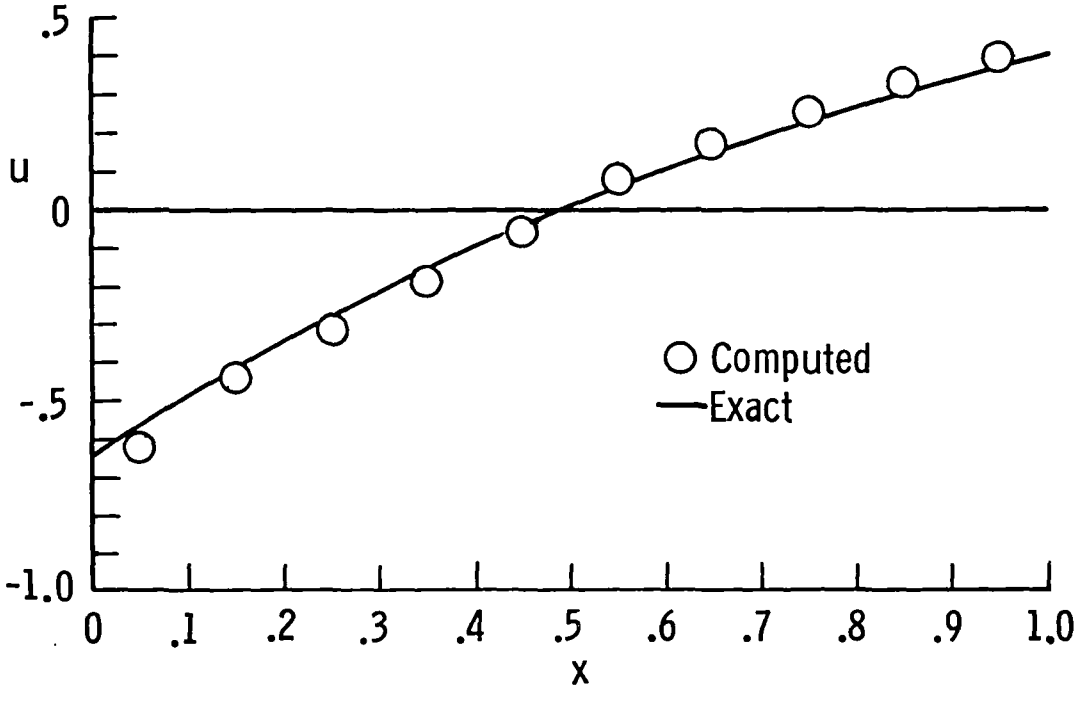


Figure 1. Exact and computed steady states for equation (2.1) with initial conditions (2.24) using a time accurate scheme.

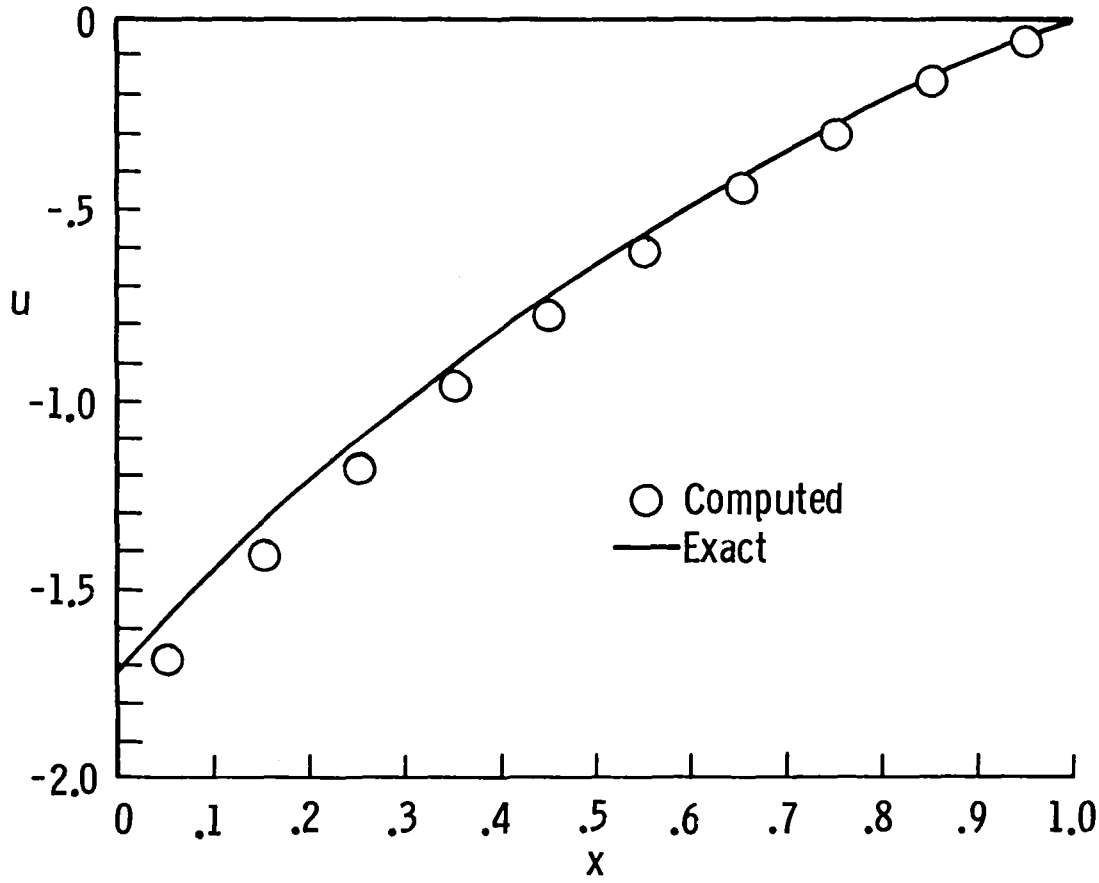


Figure 2. Exact and computed steady states for equation (2.1) with initial conditions (2.26) using a time accurate scheme.

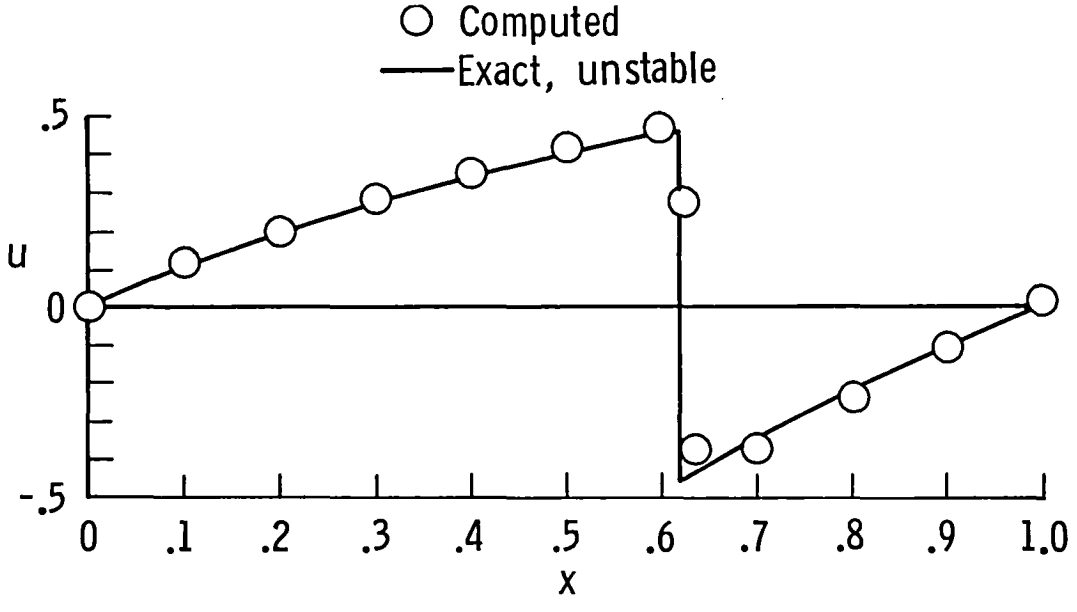


Figure 3. Exact and computed unstable steady states for equation (2.1) with initial conditions (2.29) using an implicit scheme with large Courant number.

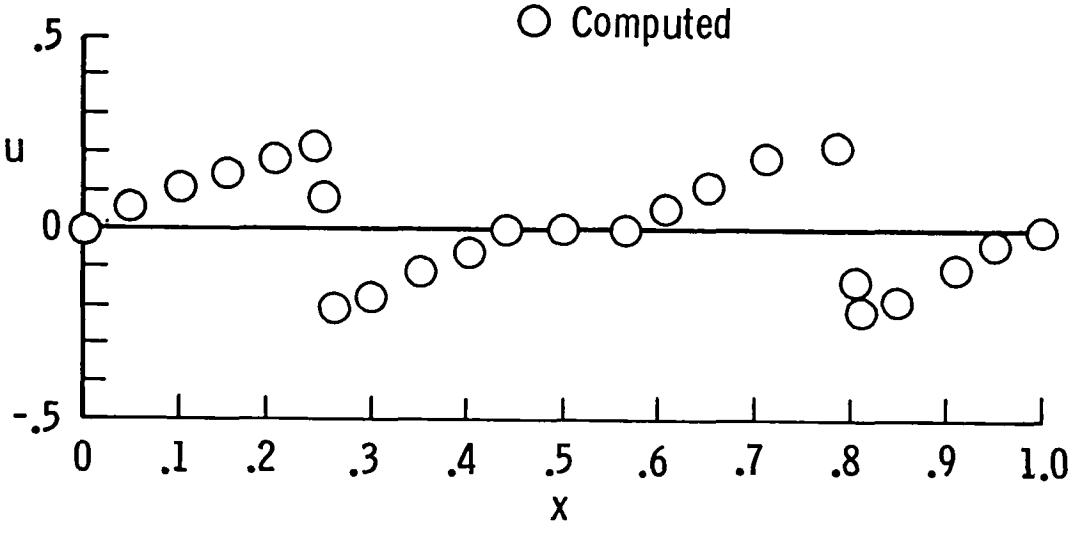


Figure 4. Computed steady state for equation (2.1) with initial conditions (2.30) using an implicit scheme with large Courant number.

3. SECOND EXAMPLE

We now shift our attention to another advection problem. The steady states of this problem are of a completely different nature than of those found in the previous example.

The partial differential equation under consideration is

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = \sin x \cos x, \qquad 0 \leq x \leq \pi, \quad t > 0 \tag{3.1}$$

$$u(x,0) = g(x), \qquad g(0) = g(\pi) = 0$$

with boundary conditions as given by (2.2).

Here we have two smooth steady state solutions,

$$\begin{aligned} u^+ &= \sin x \\ u^- &= -\sin x. \end{aligned} \tag{3.2a}$$

There is also an infinite number of possible discontinuous solutions of the form

$$\begin{aligned} u &= u^+ \qquad x < x_S \\ u &= u^- \qquad x > x_S \end{aligned}, \tag{3.2b}$$

where x_S, the "shock" location, is an arbitrary point in the interval $(0,\pi)$. Note that, in the steady state, the "shock" speed $u_S = (u^+ + u^-)/2$ is zero for any $0 < x_S < \pi$ and, therefore, (3.2b) is a legitimate steady state solution. In the above solutions we have already eliminated weak solutions that violate the "entropy condition," $u_L > 0 > u_R$.

We now ask two questions:

(i) From what class of initial conditions, if any, can either of the two smooth solutions, (3.2a), be reached and

(ii) Under what circumstances is a steady shock established, and can its location be predicted?

Consider first the two questions in the particularly simple case when

$$g(x) = \beta \sin x, \tag{3.3}$$

i.e., the initial data are proportional to a smooth steady state. For $\beta > 1$, Theorem 1 shows that the steady state is the smooth solution $u = u^+$. For

$\beta < -1$, a corollary of Theorem 1 leads to $u = u^-$.

Theorem 1: The solution of equation (3.1) with boundary conditions (2.2), initial conditions (3.3) and $\beta > 1$ satisfies

$$\lim_{t\to\infty} u(x,t) = \sin x.$$

Proof: The characteristic equations resulting from (3.1) are

$$\frac{dx}{dt} = u \tag{3.4}$$

$$\frac{du}{dt} = u\frac{du}{dx} = \frac{dF}{dx}, \qquad F = \frac{1}{2}\sin^2 x. \tag{3.5}$$

Again using $\xi = \xi(x,t)$ to designate the origin of a characteristic curve passing through (x,t), we integrate (3.5)

$$\frac{1}{2}u^2 - \frac{1}{2}g^2(\xi) = F(x) - F(\xi)$$

or

$$u = \pm\left[2F(x) - 2F(\xi) + g^2(\xi)\right]^{1/2} . \tag{3.6}$$

As $t \to 0$, $\xi \to x$ and we have to choose the positive branch of (3.6) because $\beta > 1$. Thus, using $F = (1/2)\sin^2 x$,

$$u = \left[\sin^2 x + (\beta^2 - 1)\sin^2 \xi\right]^{1/2}. \tag{3.7}$$

We claim now that for t large enough there is a unique correspondence between a point (x,t) and $\xi(x,t)$. In fact, if a shock wave were to appear at a certain time $t > 0$, it will, because of (3.7), separate two positive states. The shock wave will have a positive speed and consequently will propagate out of the domain. Therefore, for t large enough, we may substitute (3.7) into (3.4),

$$t = \int_{\xi}^{x} \frac{dy}{\left[2F(y) - 2F(\xi) + g^2(\xi)\right]^{1/2}} \tag{3.8}$$

or

$$t = \int_{\xi}^{x} \frac{dy}{\left[\sin^2 y + (\beta^2 - 1)\sin^2 \xi\right]^{1/2}} \tag{3.9}$$

For every $x < \pi$, the integrand in (3.9) cannot become singular except at the lower limit $y = \xi$, $\xi \to 0$. Thus, $t \to \infty$ as $\xi \to 0$ and the only possible solution for very large time is, from (3.7),

$$u \underset{t\to\infty}{\longrightarrow} \left[2F(x) - 2F(\xi) + g^2(\xi)\right]^{1/2}_{\xi\to 0} = \left[2F(x) - 2F(0) + g^2(0)\right]^{1/2} = \sin x,$$

which completes the proof.

Corollary: Suppose that β in (3.3) satisfies $\beta < -1$, then

$$\lim_{t\to\infty} u(x,t) = -\sin x.$$

Note that in view of (3.8) the results of Theorem 1 hold for any initial conditions $g(x)$ such that $g(0) = 0$, $g(x) > \sin x$. The corollary is thus also valid for any $g(x) < -\sin x$.

Still continuing with the case of $g(x) = \beta \sin x$, we now consider

$$0 < \beta < 1. \tag{3.10}$$

Here the steady state will be of the form (3.2b). We will show, however, in Theorem 2 that the shock location depends on the initial condition.

Theorem 2: The solution of equation (3.1) with boundary conditions (2.2), initial conditions (3.3), and $0 < \beta < 1$ satisfies

$$\lim_{t\to\infty} u(x,t) = \begin{cases} u^+ = \sin x, & 0 < x < x_s \\ u^- = -\sin x, & x_s < x < \pi \end{cases} \tag{3.11}$$

where

$$x_S = \pi - \sin^{-1}\sqrt{1-\beta^2} > \frac{\pi}{2}. \tag{3.12}$$

Proof: From the characteristic equation (3.5), with $0 < \beta < 1$, we get

$$u(x,t) = \pm[\sin^2 x - (1-\beta^2)\sin^2(\xi(x,t))]^{1/2}. \tag{3.13}$$

In the interval $(\pi - x_S, x_S)$, x_S as defined in (3.12), $u(x,t)$ cannot change sign because the radical in (3.13) cannot vanish in said interval. Since as $t \to 0$, $u(x,t)$ is positive, we conclude that

$$u(x,t) = [\sin^2 x - (1-\beta^2)\sin^2(\xi(x,t))]^{1/2}, \quad \pi - x_S < x < x_S. \tag{3.14}$$

In this interval the first characteristic equation (3.4) becomes

$$t = \int_{\xi}^{x} \frac{dy}{[\sin^2 y - (1-\beta^2)\sin^2(\xi(x,t))]^{1/2}} \tag{3.15}$$

since $t > 0$ we must have $\xi < x$ when $\pi - x_S < x < x_S$. As $t \to \infty$, $\xi(x,t)$ must therefore vanish in the limit. It is thus established that

$$\lim_{t\to\infty} u(x,t) = \sin x, \qquad (\pi - x_S < x < x_S). \tag{3.16}$$

Next consider the interval $[0, \pi - x_S)$. Formally as $t \to \infty$, in this leftmost interval, $\xi(x,t)$ must converge either to zero or π. However, any characteristic passing through (x,t) in the interval $[0, \pi - x_S)$ cannot emanate from any $\xi > x_S$ because this would mean a negative slope, and hence

a negative u in the interval $(\pi-x_s, x_s)$; this contradicts (3.16). Having established that $\lim_{t\to\infty} \xi(x,t) = 0$, we notice that formally it is possible for a characteristic curve, originating in the interval $[0,\pi-x_s)$, to start with a positive slope (required as $t \to 0$) and change slope in the interval. This, however, will result in a solution containing a "shock" that violates the "entropy condition" $u_L > 0 > u_R$. We thus have our next intermediate result

$$\lim_{t\to\infty} u(x,t) = \sin x, \qquad (0 \leq x < x_s). \tag{3.17}$$

It now remains for us to show that in the interval $x_s < x \leq \pi$ the solution must be negative and hence equal to $-\sin x$.

We first integrate (3.1) to get

$$\frac{\partial}{\partial t}\int_0^\pi u\,dx = -\int_0^\pi \left(\frac{u^2}{2}\right) x\,dx.$$

Suppose that at the point $0 < x_1 < x_2 \cdots < x_n < \pi$, $u(x,t)$ is discontinuous, since we admit only "shock" discontinuity $u^2(x_i^+) > u^2(x_i^-)$. Thus,

$$\frac{\partial}{\partial t}\int_0^\pi u(x,t)dx = \frac{1}{2}\,u^2(0,t) - \sum_{i=1}^{n}\left(u^2(x_i^+) - u^2(x_i^-)\right) - u^2(\pi,t) \tag{3.18}$$

from (3.13), $u^2(0,t) = 0$ and therefore,

$$\int_0^\pi u(x,t)dx < \int_0^\pi u(x,0)dx = 2\beta. \tag{3.19}$$

Let x_α be the point in which $u(x,\infty)$ changes sign. From (3.16), we have

$$x_s \leq x_\alpha$$

and from (3.19) we have

$$\int_0^{x_\alpha} \sin x - \int_{x_\alpha}^{\pi} \sin x \leq 2\beta, \tag{3.20}$$

thus,

$$-2 \cos_\alpha \leq 2\beta$$

or

$$x_\alpha \leq x_s \tag{3.21}$$

and therefore

$$x_\alpha = x_s. \tag{3.22}$$

This completes the proof.

It should be noted that, in general, x_s gives a lower bound on the location of the discontinuity whereas the area rule (3.19) yields an upper bound on it.

Corollary: Under the conditions of Theorem 2 with

$$-1 < \beta < 0$$

the solution still retains the form of (3.11) except that now

$$x_s = \sin^{-1} \sqrt{1 - \beta^2} < \frac{\pi}{2}.$$

For arbitrary initial data the general behavior is that described in Theorems 1 and 2 and their corollaries, i.e., one can get either solution (3.2a) or (3.2b). If a "shock" is present in the steady state, the upper and lower bounds for its location are given, for $g(x) > 0$, as follows:

$$\pi - \sin^{-1}\sqrt{\sin^2 z - g^2(z)} \leq x_s \leq \pi - \sin^{-1}\sqrt{1 - \frac{1}{4}\left(\int_0^{\pi} g(\eta)d\eta\right)^2}\,, \qquad (3.23)$$

where z maximizes the expression $\sin^2 x - g^2(x)$. For negative initial data the bounds are

$$\sin^{-1}\sqrt{\sin^2 z - g^2(z)} \leq x_s \leq \sin^{-1}\sqrt{1 - \frac{1}{4}\left(\int_0^{\pi} g(\eta)d\eta\right)^2}\,. \qquad (3.24)$$

The upper bound reflects the "area rule" (see (3.18)). The lower bound is the first point where $u(x,t)$ can change sign. For $g(x) > 0$, the upper bound becomes sharp (i.e., equals x_s), if $u(\pi,t) = 0$ for all t.

3.1 NUMERICAL RESULTS FOR THE SECOND EXAMPLE

3.1.1 Explicit Form

Equation (3.1) is discretized using the explicit E-O scheme given by equation (2.20). Numerical calculations were performed for initial conditions given by

$$g(x) = \beta \sin x, \qquad (3.25)$$

where β is a free parameter such that $0 \leq \beta \leq 1$. The steady state shock position as a function of β is plotted in Figure 5. The numerical results are in excellent agreement with the theoretical prediction given by equation (3.12). For any $\beta > 1$, the steady state obtained was u^+ given by equation (3.2a).

If one uses an algorithm employing central space differencing (e.g., MacCormack's scheme), it is then necessary to supply a numerical boundary condition. If the steady state value is used for the boundary condition, then the numerical algorithm, though stable, fails to converge to steady state. The reason is clearly due to the fact that the numerical boundary condition does not allow for a flux through that boundary. As a consequence we have (see (3.19))

$$\int_0^\pi u(x,t)dx = 2\beta$$

for all t, while the true steady state, u^+, requires

$$\lim_{t\to\infty} \int_0^\pi u(x,t) = 2.$$

3.1.2 Implicit Form

Equation (3.1) is discretized using the implicit E-O scheme given by equation (2.28). Once again, numerical calculations were performed for initial conditions given by equation (3.25). Now an additional free parameter is

$$\varepsilon = \frac{100}{\pi}\frac{\Delta x}{\Delta t} \tag{3.26}$$

which is a measure of how big Δt is taken in the numerical calculations. The results of these series of calculations are given in Figure 6. As indicated in the figure, if "small" Δt's are taken ($\varepsilon \geq 1/2$), then the steady state shock location calculated agrees with the theoretical prediction of equation (3.12). However, as Δt increases, the steady state shock position is found to the right of its theoretical location. For sufficiently high values of Δt (small ε's), the smooth solution is obtained.

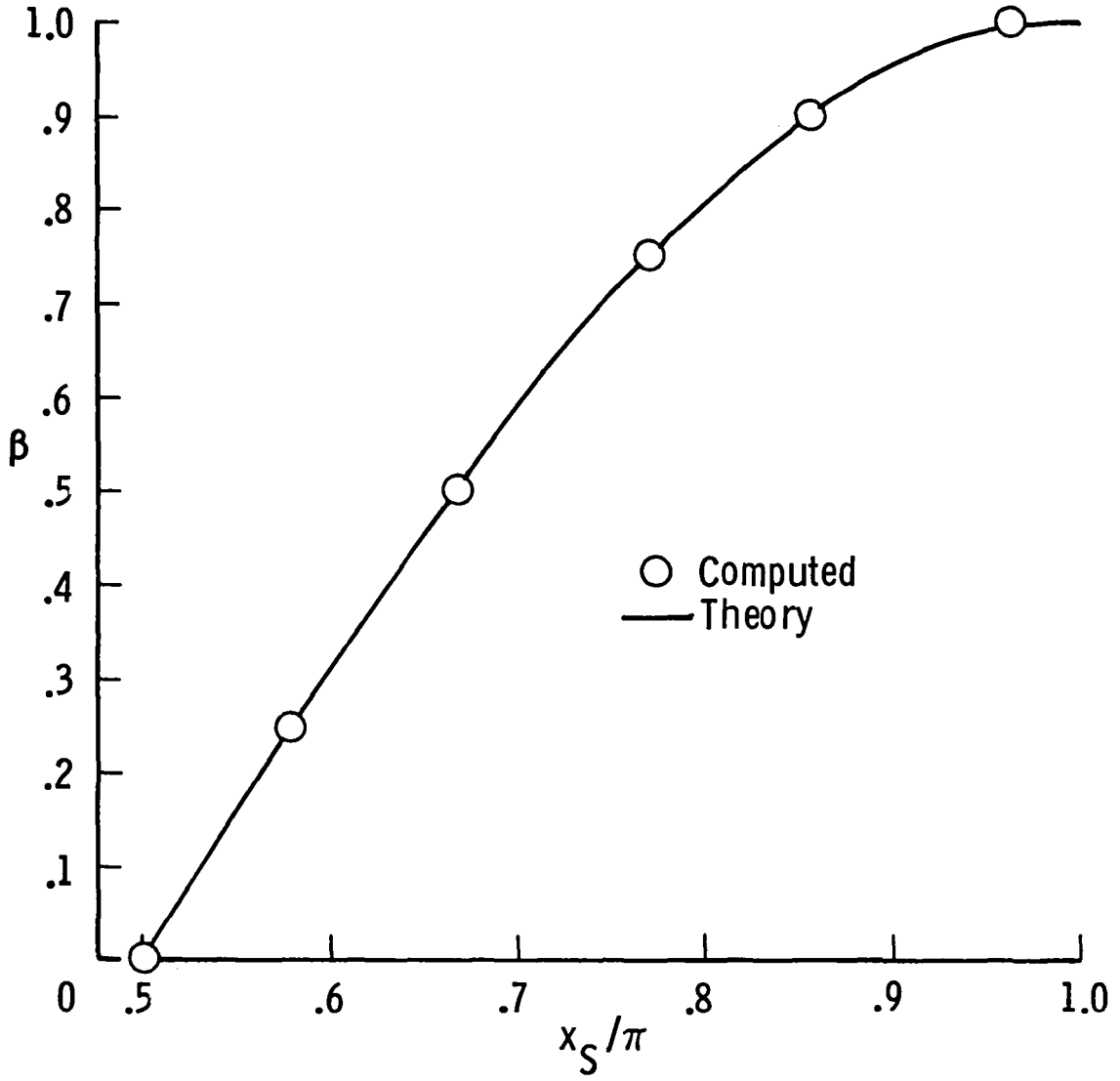


Figure 5. Computed and predicted steady state shock position for equation (3.1) with initial conditions (3.25) using a time accurate scheme.

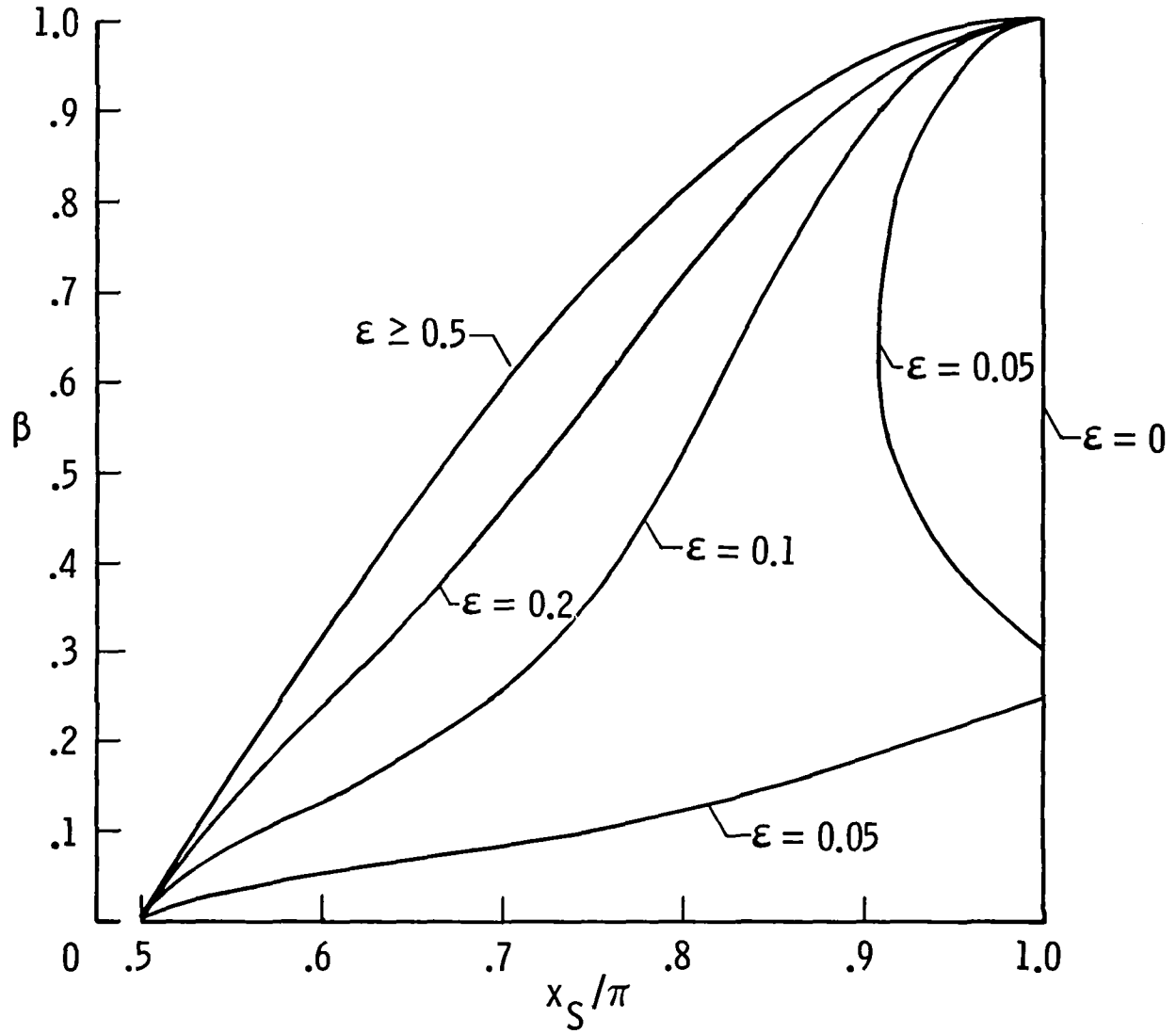


Figure 6. Computed and predicted steady state shock position for equation (3.1) with initial conditions (3.25) using an implicit scheme.

4. A MODEL FOR QUASI-ONE-DIMENSIONAL FLUID DYNAMICS

A characteristic boundary value problem, where boundary conditions are of the form (2.2), occurs in a double-throat Laval nozzle

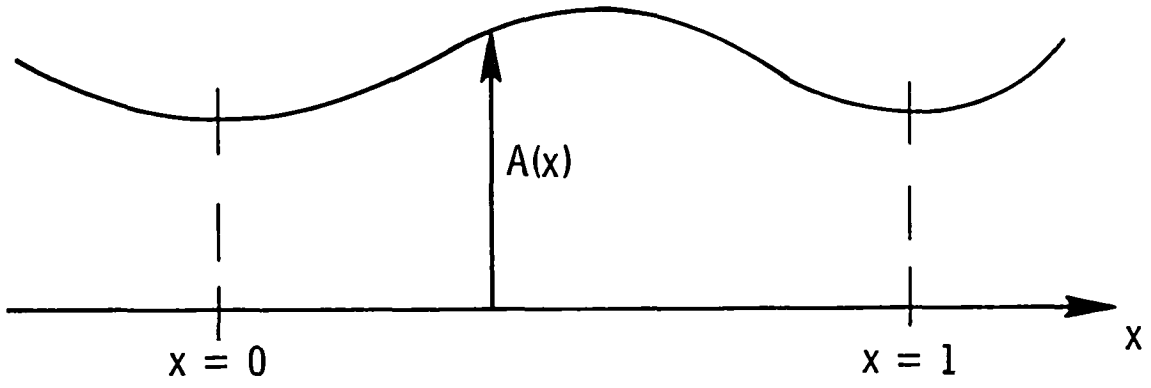


Figure 7. Sketch of double-throat nozzle

as shown in Figure 7. It is well known [1] that there are two possible smooth steady solutions, with sonic conditions at the throats. Between the throats, $0 < x < 1$, the flow can be either completely subsonic or supersonic, the exact Mach number distribution, in each case, being dependent on the nozzle area, $A(x)$, where $1 < A(x) \leq A_{max}$ in $(0,1)$, $A(0) = A(1) = 1$.

If one considers the isentropic case only, then the flow may be described by the quasi-one-dimensional partial differential equations for the Riemann variables,

$$\psi = u + \frac{2}{\gamma - 1} c, \qquad \phi = u - \frac{2}{\gamma - 1} c,$$

where u is the velocity, $c = (\gamma p/\rho)^{1/2}$ is the speed of sound, and γ is

the ratio of specific heats for ideal gases. The equations are

$$\frac{\partial \psi}{\partial t} + (u + c) \frac{\partial \psi}{\partial x} + ucF'(x) = 0, \tag{4.1}$$

$$\frac{\partial \phi}{\partial t} + (u - c) \frac{\partial \phi}{\partial x} - ucF'(x) = 0, \tag{4.2}$$

where $F'(x) = dF(x)/dx = d\big(\ell nA(x)\big)/dx$. This is a hyperbolic system whose time evolution is difficult to describe analytically. We therefore seek a model for this system so that with a single equation the most salient features are retained. We will present numerical evidence that analytical predictions resulting from this model equation agree very well with results found by numerical integration of the original system (4.1), (4.2).

The model is derived using a single assumption, namely that the total enthalpy is constant not only at steady state but also during the transient phase. The mathematical expression of this assumption is that

$$\phi^2 + \psi^2 + \frac{2(1 - a)}{a} \phi\psi = \frac{4c_*^2}{2a - 1} = \frac{16}{\gamma^2 - 1} c_0^2 \tag{4.3}$$

where

$$a = \frac{\gamma + 1}{4}, \tag{4.4}$$

c_0 is the stagnation sound speed, and c_* is the sonic sound speed, i.e., c_* is the sound speed at a sonic throat.

We now face the choice of solving (4.3) for either ψ in terms of ϕ, or vice versa. This dilemma is resolved by recognizing that our "physical" problem will impose characteristic boundary conditions on (4.2), and we would

like our model equation to retain this feature. Therefore, (4.2) is the relevant equation. Solving for ψ gives

$$\psi = -\frac{1-a}{a} + \frac{1}{a}\left[\frac{4a^2c_*^2}{2a-1} - (2a-1)\phi^2\right]^{1/2}, \tag{4.5}$$

where the positive branch was chosen in order to satisfy the steady state boundary condition at $x = 0$, i.e., at the first throat, where

$$\psi_* = \frac{2a}{2a-1}\,c_*; \qquad \phi_* = -\frac{2(1-a)}{2a-1}\,c_*. \tag{4.6}$$

Using (4.5) in (4.2), and defining

$$\hat{\phi} = \phi/\psi_* \tag{4.7}$$

the equation (4.2) takes the form

$$\frac{\partial\hat{\phi}}{\partial\tau} + \Lambda(\hat{\phi})\,\frac{\partial\hat{\phi}}{\partial x} = H(\hat{\phi})F'(x) \tag{4.8}$$

where

$$\Lambda(\hat{\phi}) = \hat{\phi} + \frac{1-a}{\sqrt{2a-1}}\,\sqrt{1-\hat{\phi}^2} \tag{4.9}$$

$$H(\hat{\phi}) = \left(\frac{2a-1}{4a}\right)\left[1 - 2\hat{\phi}^2 - \frac{2(1-a)}{\sqrt{2a-1}}\,\hat{\phi}\,\sqrt{1-\hat{\phi}^2}\right] \tag{4.10}$$

$$\tau = tc_*. \tag{4.11}$$

Notice that the time scale, τ, is determined by the sonic conditions.

For the sake of clarity let us first examine the simple case of $a = 1$ ($\gamma = 3$), which corresponds to the flow of products caused by detonating solid explosives. Equation (4.8) then becomes

$$\frac{\partial \hat{\phi}}{\partial \tau} + \hat{\phi}\,\frac{\partial \hat{\phi}}{\partial x} = \frac{1}{4}\,(1 - 2\hat{\phi}^2)F'(x), \qquad F(x) = \ell nA(x). \tag{4.12}$$

A smooth steady state solution of (4.12) with $\hat{\phi}(0) = 0$ is

$$\hat{\phi}^2(x) = \frac{1}{2}\left(1 - e^{-F(x)}\right), \tag{4.13}$$

since $A(0) = 1$, and so, as in (3.2a) we have two possible steady states. One is positive (supersonic) and the other is negative (subsonic):

$$\hat{\phi}^+ = \left(\frac{A(x) - 1}{2A(x)}\right)^{1/2} \tag{4.14}$$

$$\hat{\phi}^- = -\left(\frac{A(x) - 1}{2A(x)}\right)^{1/2}. \tag{4.15}$$

Bearing in mind the results of the previous sections, we will show that in the time evolution problem, $\hat{\phi}^+$ and $\hat{\phi}^-$ are reachable from different initial conditions. Clearly (4.14) and (4.15) can be connected by a steady shock - and again, because of the symmetry of $\hat{\phi}^+$ and $\hat{\phi}^-$, the steady shock location x_S could be anywhere in the interval $(0,1)$. We will show that here too bounds on x_S can be found and compare them with results of numerical integration of the original system (4.1), (4.2).

We will concentrate on the positive branch (4.14), showing that if the initial condition is given by

$$\hat{\phi}(x,0) = g(x) = \beta\hat{\phi}^{+} = \beta\left[\frac{A(x) - 1}{2A(x)}\right]^{1/2} \tag{4.16}$$

with

$$1 < \beta^2 < \frac{A_{max}}{A_{max} - 1}\ , \tag{4.17}$$

where A_{max} is the maximum area in the nozzle, then $\lim_{t\to\infty} \hat{\phi}(x,t) = \hat{\phi}^{+}(x)$. A solution of the second characteristic equation,

$$\frac{d\hat{\phi}}{d\tau} = \hat{\phi}\,\frac{d\hat{\phi}}{dx} = \frac{1}{4}\left(1 - 2\hat{\phi}^2\right)F'(x) \tag{4.18}$$

is given by

$$\left|1 - 2\hat{\phi}^2\right| = \left|1 - 2g^2\left(\xi(x,\tau)\right)\right| A\left(\xi(x,\tau)\right)/A(x), \tag{4.19}$$

where as before $\xi(x,\tau)$ is the origin of a characteristic curve passing through (x,τ). Since we have chosen (see (4.17)) $g^2(x)$ to be smaller than $1/2$, it follows from (4.19) that

$$\hat{\phi}(x,\xi) = \pm\left[\frac{A(x)\cdot\cdot A(\xi)\left[1 - 2g^2(\xi(x,\tau))\right]}{2A(x)}\right]^{1/2}\ , \tag{4.20}$$

where $\xi(x,\tau)$ is to be determined from the first characterisitc equation

$$\tau = \int_{\xi}^{x} \frac{dy}{\hat{\phi}(y,\xi)}\ . \tag{4.21}$$

From (4.16) we see that a positive (negative) β will initially select a positive (negative) branch of (4.20). By an argument similar to that used in Theorem 1, it remains for us to show that $\hat{\phi}$ thus initiated will not change

sign while evolving to steady state. This follows immediately from (4.20) if we use for $g(x)$ equation (4.16) with $\beta > 1$.

Next we consider the discontinuous steady state solution. The initial data are now taken so that $|g(x)| < \hat{\phi}^+$, see equation (4.14). A lower bound for x_S is found by inquiring about the zeros of (4.20) - the argument is the same as in the previous section. The radical in (4.2) is zero

$$A(x_S) = A(z)(1 - 2g^2(z)) \tag{4.22}$$

where, as before, z maximizes the expression $A(x)(1 - 2g^2(x))$. To find the upper bound we have to devise an "area rule" for equation (4.12). Because of the structure of the right-hand side of (4.12), it is no longer $\int_0^1 \hat{\phi}(x,\tau)dx$ which is conserved. To find the appropriate "area rule," we divide both sides of (4.12) by $1 - 2\hat{\phi}^2 > 0$. The resulting equation after integration by x over the interval may be written as

$$\frac{\partial}{\partial\tau}\int_0^1 \ell n\,\frac{1+\sqrt{2}\,\hat{\phi}}{1-\sqrt{2}\,\hat{\phi}}\,dx - \frac{1}{\sqrt{2}}\int_0^1 \frac{\partial}{\partial x}[\ell n(1-2\hat{\phi}^2)]dx = \frac{1}{\sqrt{2}}F(x)\Big|_0^1 = 0. \tag{4.23}$$

Under the usual area rule assumptions, $\hat{\phi}(0,\tau) = \hat{\phi}(1,\tau) = 0$, we have

$$\int_0^1 \ell n\,\frac{1+\sqrt{2}\,\hat{\phi}}{1-\sqrt{2}\,\hat{\phi}}\,dx = \text{const.} \tag{4.24}$$

Therefore, an upper bound for x_S is found from

$$\int_0^{x_S} \ell n\,\frac{1+\sqrt{2}\,\hat{\phi}^+}{1-\sqrt{2}\,\hat{\phi}^+}\,dx + \int_{x_S}^1 \ell n\,\frac{1+\sqrt{2}\,\hat{\phi}^-}{1-\sqrt{2}\,\hat{\phi}^-}\,dx = \int_0^1 \ell n\,\frac{1+\sqrt{2}\,g(x)}{1-\sqrt{2}\,g(x)}\,dx. \tag{4.25}$$

When $g(x) \leq \beta\hat{\phi}^{+}$, $(\beta < 1)$ we expect, as in the previous example, the upper and lower bounds on x_s to coincide. This was indeed verified in numerical experiments with a particular area distribution $A(x)$.

Recalling that (4.12) is a scalar model equation representing the systems (4.1), (4.2), we find it interesting to note that this 2×2 system also possesses an area rule, namely:

$$\frac{\partial}{\partial t}\int (\psi+\phi)dx = \frac{1}{2}\left[\left(\psi^2(1,t) + \phi^2(1,t)\right) - \left(\psi^2(0,t) + \phi^2(0,t)\right)\right]. \quad (4.26)$$

Under the assumption that $\phi(0,t) = \phi(1,t) = 0$; $\psi(0,t) = \psi(1,t)$, we have

$$\frac{\partial}{\partial t}\int (\psi + \phi)dx = 0. \quad (4.27)$$

We can now use this to test the "goodness" of our model by comparing the shock location predicted from (4.25) with that of the system, whose solution is found numerically. This comparison is carried out in the next section.

Having concluded the analysis of the $a = 1$ case, let us now return to the more general formulation (4.8). In particular, let us consider the case of $\gamma = 1.4$ $(a = .6)$, corresponding to air. We next show how (4.8) may be cast in a form similar to the "decoupled" one in (4.12). Multiply both sides of (4.8) by $r'(\phi)$ $(r' = dr/d\phi)$ to obtain

$$\frac{\partial r}{\partial \tau} + r\frac{\partial r}{\partial x} = \frac{H(\hat{\phi}(r))}{\hat{\phi}'(r)}F'(x) = K(r)(r_+ - r)(r - r_-)F'(x), \quad (4.28)$$

where

$$K(r) = \frac{\sqrt{1 - \frac{5}{9} r^2}\,\left(\sqrt{1 - \frac{5}{9} r^2} - \frac{2}{3} r\right)\left(r - \sqrt{\frac{3}{10}}\right)\left(r + \sqrt{\frac{3}{2}}\right)}{(1 - r^2)\left(r - \frac{3}{5}\sqrt{1 - \frac{5}{9} r^2}\right)\left(\frac{5}{3} r + 5\sqrt{1 - \frac{5}{9} r^2}\right)} \tag{4.29}$$

$$r_+ = \sqrt{\frac{3}{2}}\,, \qquad r_- = -\sqrt{\frac{3}{10}}\,. \tag{4.30}$$

The quantities r_- and r_+ are the values of r which, in the steady state, correspond to Mach numbers of zero and infinity, respectively. For general values of γ, $K(r)$, r_+, and r_- are replaced by $K(r,a)$, $r_+(a)$, and $r_-(a)$. $K(r,a)$ will have the same structure as in (4.29).

It is easy to verify that $K(r)$, given by (4.29), is a positive, slowly monotonically decreasing function in the relevant range $r_- \leq r \leq r_+$. In fact $K(r_-) \approx 2K(r_+) = .309$. In the case of $\gamma = 3$, i.e., equation (4.12), $r = \phi$ and we have $r_+ = -r_- = 1/\sqrt{2}$ and $K(r) =$ constant. It is thus clear that the topological behavior of (4.28) is the same as that of (4.12), and the arguments carry over. In particular the non-unique smooth steady states depend on the initial data in the same fashion with respect to β.

4.1 NUMERICAL RESULTS FOR QUASI ONE-DIMENSIONAL EQUATIONS

Here we study numerically equations (4.1) and (4.2) for $\gamma = 3$, namely:

$$\frac{\partial \psi}{\partial t} + \frac{\partial}{\partial x}\left(\frac{\psi^2}{2}\right) = -\frac{1}{4}\left(\psi^2 - \phi^2\right)F'(x) \tag{4.31}$$

$$\frac{\partial \phi}{\partial t} + \frac{\partial}{\partial x}\left(\frac{\phi^2}{2}\right) = \frac{1}{4}\left(\psi^2 - \phi^2\right)F'(x). \tag{4.32}$$

The area of the dual-throat nozzle is defined by

$$A(x) = \frac{(1 - d)^2 + \left(1 - d(2x - 1)^2\right)^2}{2(1 - d)\left(1 - d(2x - 1)^2\right)^2}, \qquad 0 \leq x \leq 1, \tag{4.33}$$

where d is a parameter related to the maximum area by

$$A_{max} = \frac{(1 - d)^2 + 1}{2(1 - d)} . \tag{4.34}$$

For the numerical experiments, we have used $d = 1/6$ which results in $A_{max} = 1.1$. The steady state Mach number distribution is

$$M(x) = A(x) \pm \sqrt{A^2(x) - 1} , \tag{4.35}$$

and the steady state solution to (4.31) and (4.32) as a function of the Mach number is

$$\psi = \sqrt{3}\ (1 + M)/(1 + M^2)^{1/2} \tag{4.36}$$

$$\phi = -\sqrt{3}\ (1 - M)/(1 + M^2)^{1/2} . \tag{4.37}$$

With the stagnation pressure and density used as reference values, the value of ψ_* is $\sqrt{6}$.

4.1.1 Explicit Form

Equations (4.31) and (4.32) are discretized using the explicit E-O scheme given by equation (2.20). Numerical calculations were performed with initial conditions corresponding to

$$\phi(x,0) = \beta \sqrt{6} \left[\frac{A(x) - 1}{2A(x)}\right]^{1/2}, \tag{4.38}$$

which is equivalent to (4.16), and with

$$\psi(x,0) = \sqrt{6} \left[\frac{A(x) + 1}{2A(x)}\right]^{1/2}, \tag{4.39}$$

or

$$\psi(x,0) = \sqrt{6} \left(1 - \beta^2 \left(\frac{A(x) - 1}{2A(x)}\right)\right)^{1/2} \tag{4.40}$$

The initial conditions given by (4.39) correspond to the exact, steady solution for ψ while those given by (4.40) correspond to conditions for ψ consistent with (4.38) and constant total enthalpy, (4.5). The steady state reached was the same in either case; therefore, the results reported here are for calculations with (4.40) only.

Figure 8 summarizes the numerical results. The figure compares the predicted steady state shock position as given by (4.25) for the model equation (4.12) and the computed position for the system (4.31) and (4.32). As is evident from the figure, the agreement is very good.

4.1.2 Implicit Form

Equations (4.31) and (4.32) are discretized using the implicit E-O scheme given by equation (2.28). Equations (4.38) and (4.40) are again used as initial conditions. The numerical results are summarized in Figure 9. As shown in the figure, the steady state shock position depends on the Courant number as measured by the parameter

$$\varepsilon = 100 \frac{\Delta x}{\Delta t} . \tag{4.41}$$

For values of $\varepsilon \geq 10$ the steady state shock position is the same as that predicted by the explicit form. For values of $\varepsilon < 10$ (large Δt), the steady state shock position bifurcates at certain values of β.

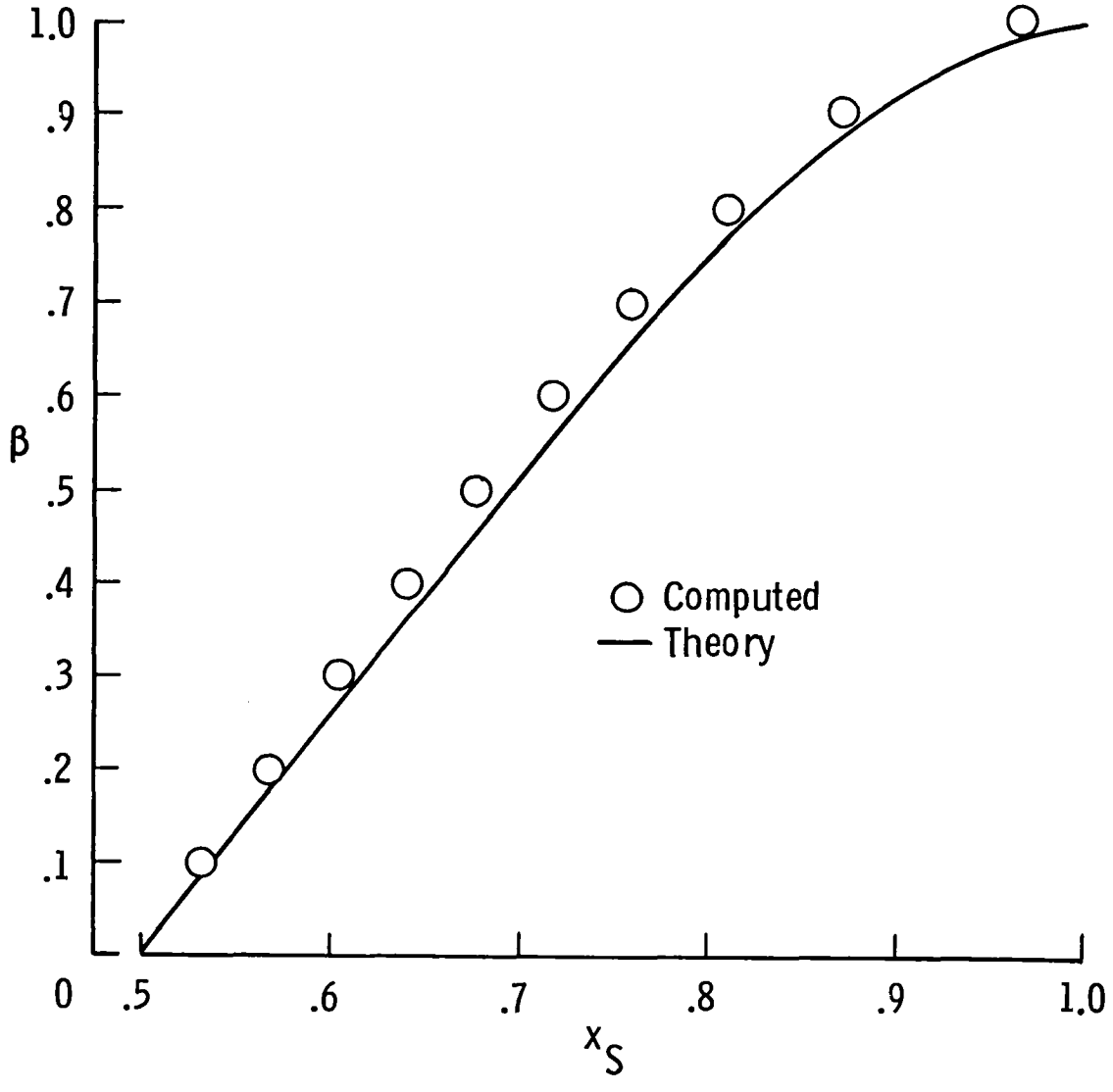


Figure 8. Predicted steady state shock position given by (4.25) for equation (4.12) and computed position for system (4.31) and (4.32) with initial conditions (4.38) and (4.40) using a time accurate scheme.

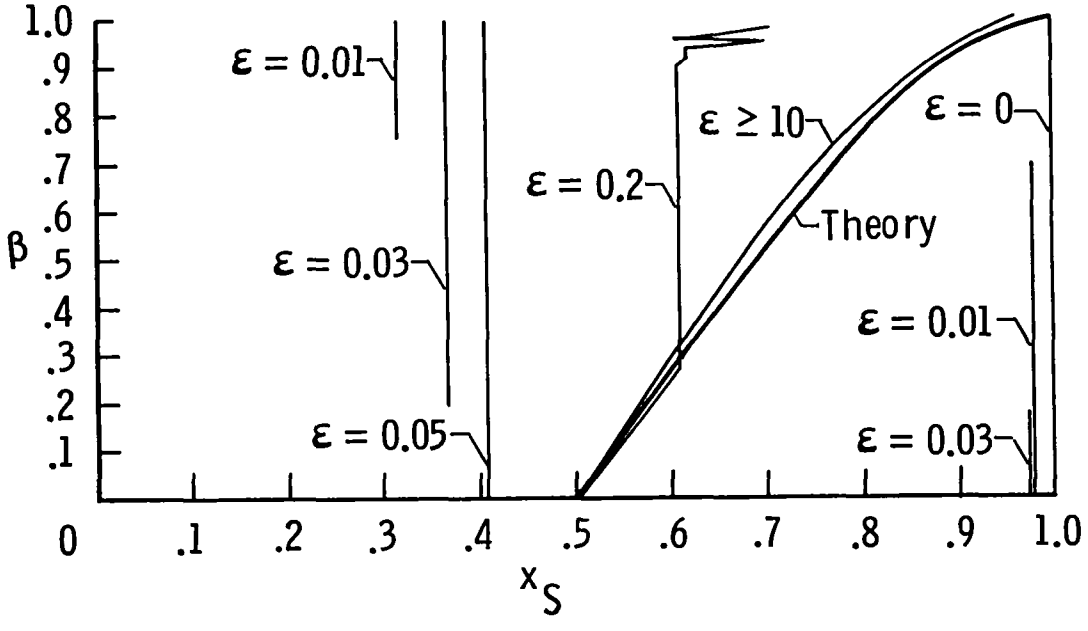


Figure 9. Predicted steady state shock position given by (4.25) for equation (4.12) and computed position for system (4.31) and (4.32) with initial conditions (4.38) and (4.40) using an implicit scheme.

CONCLUSIONS

In this paper we analyzed several model equations for characteristic initial boundary value problems and examined numerically these as well as the quasi-one-dimensional isentropic Euler equations of gas dynamics.

We showed that because of the characteristic nature of the boundary conditions the resulting steady states, whether smooth or discontinuous, depend on the initial data. Different initial conditions may yield different steady states. We also gave an example (see Section 2) of solution to the steady state equation which cannot evolve from the initial data. Thus from the point of view of the time-dependent equation, we find there are no non-unique steady states.

Another conclusion that one may draw is that in order to have complete confidence in the results, numerical schemes for characteristic initial boundary value problems should be time consistent and employ only suitable boundary conditions. Thus we have shown that implicit methods, even for finite Courant numbers, may yield solutions which are piecewise combinations of non-unique solutions of the steady state equations. In fact, such numerically implicit algorithms may converge to solutions which also include parts of unstable steady states.

REFERENCES

[1] L. Crocco, "One-Dimensional Treatment of Steady Gas Dynamics" in Fundamentals of Gas Dynamics, Vol. III of High Speed Aerodynamics and Jet Propulsion, Howard W. Emmons, ed., New Jersey, (1958), pp. 183-186.

[2] P. Embid, J. Goodman, and A. Majda, "Multiple Steady States for 1-D Transonic Flow," SIAM J. Sci. Stat. Comp., Vol. 5, No. 1 (1984), pp. 21-41.

[3] B. Engquist, and S. Osher, "Stable and Entropy Satisfying Approximations for Transonic Flow Calculations," Math. Comp., Vol. 34, (1980), pp. 45-75.

[4] G. Kreiss and H. O. Kreiss, "Convergence to Steady State of Solutions of Burgers' Equations," NASA Contractor Report No. 178017, ICASE Report No. 85-50, December 1985.

A MINIMUM ENTROPY PRINCIPLE IN THE GAS DYNAMICS EQUATIONS

Eitan Tadmor*
School of Mathematical Sciences, Tel-Aviv University
and
Institute for Computer Applications in Science and Engineering

ABSTRACT

Let $u(\bar{x},t)$ be a weak solution of the Euler equations, governing the inviscid polytropic gas dynamics; in addition, $u(\bar{x},t)$ is assumed to respect the usual entropy conditions connected with the conservative Euler equations. We show that such entropy solutions of the gas dynamics equations satisfy a minimum entropy principle, namely, that the spatial minimum of their specific entropy, Ess $\inf_{\bar{x}} S(u(\bar{x},t))$, is an increasing function of time. This principle equally applies to discrete approximations of the Euler equations such as the Godunov-type and Lax-Friedrichs schemes. Our derivation of this minimum principle makes use of the fact that there is a family of generalized entropy functions connected with the conservative Euler equations.

Research was supported in part by NASA Contract No. NAS1-17070 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665-5225. Additional support was provided in part by NSF Grant No. DMS85-03294 and ARO Grant No. DAAG29-85-K-0190 while in residence at the University of California, Los Angeles, CA 90024.

*Bat-Sheva Foundation Fellow

1. INTRODUCTION

Many phenomena in continuum mechanics are modeled by hyperbolic systems of conservation laws

$$\frac{\partial u}{\partial t} + \sum_{k=1}^{d} \frac{\partial f^{(k)}}{\partial x_k} = 0, \qquad (\bar{x} = (x_1,\cdots,x_d), t) \in \underline{R} \times [0,\infty), \tag{1.1}$$

where $\mathbf{f}^{(k)} \equiv \mathbf{f}^{(k)}(\mathbf{u}) = \left(f_1^{(k)},\cdots,f_N^{(k)}\right)^T$ are smooth nonlinear flux mappings of the N-vector of conservative variables $\mathbf{u} \equiv \mathbf{u}(\bar{x},t) = (u_1,\cdots,u_N)^T$. Friedrichs and Lax [3] have observed that the hyperbolic nature of such models is revealed by the property of most of those systems being endowed with a generalized

Entropy Function: A smooth convex mapping $U(\mathbf{u})$ augumented with entropy flux mappings $\vec{F} \equiv \vec{F}(\mathbf{u}) = (F^{(1)}(\mathbf{u}),\cdots,F^{(d)}(\mathbf{u}))$, such that the following compatibility relations hold

$$U_{\mathbf{u}}^T f_{\mathbf{u}}^{(k)} = F_{\mathbf{u}}^{(k)T}, \qquad k = 1,2,\cdots,d. \tag{1.2}$$

Multiplying (1.1) by $U_{\mathbf{u}}^T$ and employing (1.2), one arrives at an equivalent formulation of the compatibility relations (1.2), namely, that under the smooth regime we have on top of (1.1) the additional conservation of entropy

$$\frac{\partial U}{\partial t} + \sum_{k=1}^{d} \frac{\partial F^{(k)}}{\partial x_k} = 0. \tag{1.3}$$

Owing to the nonlinearity of the fluxes $\mathbf{f}^{(k)}(\mathbf{u})$, solutions of (1.1) may develop singularities at a finite time after which one must admit weak

solutions, i.e., those derived directly from the underlying integral conservative equations. Considering (1.1) as a strong limit of the regularized problem,

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{k=1}^{d} \frac{\partial \mathbf{f}^{(k)}}{\partial x_k} = \mu \sum_{k=1}^{d} \frac{\partial^2 \mathbf{u}}{\partial x_k^2}, \qquad \mu \downarrow 0, \tag{1.4}_{\mu}$$

then following Lax [9] and Krushkov [8], we postulate as an admissibility criterion for such limit solutions an entropy stability condition which manifests itself in terms of an

Entropy Inequality: We have, in the sense of distributions,

$$\frac{\partial U}{\partial t} + \sum_{k=1}^{d} \frac{\partial F^{(k)}}{\partial x_k} \leqq 0. \tag{1.5}$$

Weak solutions of (1.1), which in addition satisfy the inequality (1.5) for all entropy pairs $(U,\vec{F})$ connected with that system, are called entropy solutions.[1] Having a (weakly) nonpositive quantity on the L.H.S. of (1.5) is thus a consequence of viewing these entropy solutions as limits of vanishing dissipativity mechanisms. In particular, the inequality (1.5) implies that the total entropy in the domain decreases in time (we assume entropy outflux through the boundaries)

$$\frac{d}{dt} \int_{\bar{x}} U(\mathbf{u}(\bar{x},t))d\bar{x} \leq 0. \tag{1.6}$$

---

[1] Krushkov [8, p. 241] has termed such solutions simply as generalized solutions.

In this paper, we consider entropy solutions,

$$\mathbf{u} = (\rho, \mathbf{m}, E)^T \tag{1.7a}$$

of the Euler equations. These equations govern the inviscid polytropic gas dynamics, asserting the conservation of the density ρ, the momentum $\mathbf{m} = (m_1, m_2, m_3)^T$, and the energy E. Let $\mathbf{q} \equiv \frac{\mathbf{m}}{\rho}$ denote the velocity field of such motion. Then, expressed in terms of the pressure, p,

$$p = (\gamma-1)\cdot\left[E - \tfrac{1}{2}\cdot\rho|\mathbf{q}|^2\right], \qquad \gamma = \text{adiabatic exponent}, \tag{1.7b}$$

the corresponding fluxes in this case are given by[2]

$$\mathbf{f}^{(k)} = \left(m_k, q_k\cdot\mathbf{m} + p\cdot\mathbf{e}^{(k)}, q_k(E + p)\right)^T, \qquad k = 1,2,3. \tag{1.7c}$$

The main result of this paper asserts that entropy solutions of Euler equations satisfy the following

Minimum Principle: Let $\mathbf{u} \equiv \mathbf{u}(\bar{x},t)$ be an entropy solution of the gas dynamics equations (1.7) and let

$$S(\bar{x},t) \equiv S(\mathbf{u}(\bar{x},t)) = \ln(p\rho^{-\gamma}) \tag{1.8}$$

---

[2] With $\mathbf{e}^{(k)}$ denoting the unit Cartesian vectors $\mathbf{e}^{(k)} = \delta_{kj}$.

denote the specific entropy of such solution. Then the following estimate holds

$$\underset{|\bar{x}|\leq R}{\text{Ess inf}}\ S(\bar{x},t) \geq \underset{|\bar{x}|\leq R+t\cdot q_{max}}{\text{Ess inf}}\ S(\bar{x},t=0). \tag{1.9}$$

Here q_{max} stands for the maximal speed $|\mathbf{q}|$ in the domain.

The proof of this assertion is provided in Section 3 below. Prior to that we elaborate in Section 2 on the entropy inequality connected with the gas dynamics equations. In particular, Harten [5] has shown that there exists a whole family of entropy pairs associated with these equations, a fact which is essential in our derivation of the minimum principle.

As an immediate consequence of the minimum principle, we conclude that $\underset{\bar{x}}{\text{Ess inf}}\, S(\bar{x},t)$ is an increasing function of t for every entropy solution of (1.7). The following argument sheds additional light on this conclusion in the case of a piecewise-smooth flow. To this end, an arbitrary particle currently located at $(\bar{x},t)$ is traced backwards in time into its initial position at $t = 0$. Since the specific entropy of such particle remains constant along the particle path--except for its decrease when crossing backwards shock waves, it follows that its value $S(\bar{x},t)$ is greater or equal than that of the initial spatial minimum $\underset{\bar{x}}{\text{Ess inf}}\, S(\bar{x},t=0)$, as asserted. In contrast to the above 'Lagrangian' argument, the derivation of the minimum principle outlined below, is purely an 'Eulerian' one. It enables us to relax the regularity assumption on the flow, and--since we do not follow the characteristics, it equally applies to discrete approximations of the Euler equations.

In Section 4 we consider approximate solutions of the Euler equations, $w(\bar{x}_\nu,t)$, which respect the entropy decrease estimate (1.6),

$$\sum_\nu U(w(\bar{x}_\nu,t+\Delta t))\Delta\bar{x}_\nu \le \sum_\nu U(w(\bar{x}_\nu,t))\Delta\bar{x}_\nu. \tag{1.10}$$

We note that such approximate solutions are obtained by entropy stable schemes satisfying the cell entropy inequality

$$U(w(\bar{x}_\nu,t+\Delta t)) \le U(w(\bar{x}_\nu,t)) + \sum_{k=1}^{d} \frac{1}{\Delta\bar{x}_\nu}\left[F^{(k)}_{\nu+1/2} - F^{(k)}_{\nu-1/2}\right], \tag{1.11}$$

e.g., the Godunov-type and Lax-Friedrichs schemes [6]. We have

Minimum Principle: Let $w(\bar{x}_\nu,t)$ be an approximate solution of the gas dynamics equations (1.8) and let

$$S(\bar{x}_\nu,t) \equiv S(w(\bar{x}_\nu,t)) = \ln(p\rho^{-\gamma}) \tag{1.12}$$

denote the specific entropy of such solution. Assume that its total entropy decreases in time, (1.10). Then the following estimate holds

$$S(\bar{x},t+\Delta t) \ge \min_\nu\left[S(\bar{x}_\nu,t)\right]. \tag{1.13}$$

In the case of entropy stable schemes, (1.11), a more precise estimate is obtained which takes into account the support of the schemes' stencil.

The inequality (1.13) leads to an a'priori pointwise estimate on the approximate solution $w(\bar{x},t)$. Such pointwise estimates play an essential role

with regard to question of the convergence of entropy stable schemes. In particular, DiPerna [2, Section 7] has recently shown that in certain cases, such (two-sided) estimates are sufficient in order to guarantee the convergence of such schemes.

2. GENERALIZED ENTROPY FUNCTIONS OF THE EULER EQUATIONS

We consider the Euler equations for polytropic gas

$$\frac{\partial}{\partial t}\begin{bmatrix}\rho\\ \mathbf{m}\\ E\end{bmatrix} + \sum_{k=1}^{3}\frac{\partial}{\partial x_k}\begin{bmatrix}m_k\\ q_k\mathbf{m} + pe^{(k)}\\ q_k(E+\rho)\end{bmatrix} = 0. \tag{2.1}$$

It is well-known, e.g., [1], that for all smooth solutions of (2.1) the specific entropy[3]

$$S(\bar{x},t) = \ln(p\rho^{-\gamma}),$$

remains constant along streamlines, i.e.,

$$\frac{DS}{Dt} = \frac{\partial S}{\partial t} + \sum_{k=1}^{3} q_k \frac{\partial S}{\partial x_k} = 0. \tag{2.2a}$$

Let $h(S)$ be an arbitrary smooth function of S. Multiplying (2.2a) by $\rho h'(S)$ --prime denoting S-differentiation, we find

---

[3] After normalization, taking the specific heat constant to be $c_v = 1$.

$$\rho \frac{\partial h(S)}{\partial t} + \sum_{k=1}^{3} m_k \frac{\partial h(S)}{\partial x_k} = 0.$$

Adding this to the continuity equation which is premultiplied by $h(S)$,

$$\frac{\partial \rho}{\partial t} h(S) + \sum_{k=1}^{3} \frac{\partial m_k}{\partial x_k} h(S) = 0, \tag{2.2b}$$

we obtain after changing sign, a conservative entropy equation like (1.3) which reads [5]

$$\frac{\partial}{\partial t} [-\rho h(S)] + \sum_{k=1}^{3} \frac{\partial}{\partial x_k} [-m_k h(S)] = 0. \tag{2.3}$$

In order to comply with the further requirement of being a generalized entropy function, $U(\mathbf{u}) = -\rho h(S)$ has to be a <u>convex</u> function of the conservative variables $\mathbf{u} = (\rho, \mathbf{m}, E)^T$. A straightforward computation carried out by Harten [5, Section 2] in the two-dimensional case shows that the Hessian $U_{\mathbf{uu}}$ is positive definite if and only if

$$\rho[h'(S) - \gamma \cdot h''(S)] > 0.$$

Excluding negative densities we may summarize that <u>there exists a family of (generalized) entropy pairs</u> $(U, \vec{F})$ <u>associated with Euler equations</u> (2.1),

$$U(\mathbf{u}) = -\rho h(S), \quad F^{(k)}(\mathbf{u}) = -m_k h(S) \qquad k = 1,2,3, \tag{2.4a}$$

<u>generated by the smooth increasing functions</u> $h(S)$ <u>which satisfy</u>

$$h'(S) - \gamma \cdot h''(S) > 0. \tag{2.4b}$$

3. A MINIMUM ENTROPY PRINCIPLE

Let $u = (\rho, \mathbf{m}, E)^T$ be an entropy solution of the gas dynamics equations (2.1). Such a solution is characterized by the entropy inequality (1.7)

$$\frac{\partial U(\mathbf{u})}{\partial t} + \sum_{k=1}^{3} \frac{\partial F^{(k)}(\mathbf{u})}{\partial x_k} \leqq 0 \tag{3.1}$$

which holds for all entropy pairs $(U, \vec{F})$ connected with the equations.

To derive a minimum principle, we shall make use of an argument due to Lax [9, Section 3]. We begin with

Lemma 3.1: Let $\mathbf{u}$ be an entropy solution of the gas dynamics equations (2.1). Then for all nonpositive smooth increasing functions $h(S)$ satisfying (3.2b), we have

$$\int_{|\bar{x}| \leqq R} \rho(\bar{x},t) \cdot h(S(\bar{x},t)) d\bar{x} \geqq \int_{|\bar{x}| \leqq R + t \cdot q_{max}} \rho(\bar{x},0) \cdot h(S(\bar{x},0)) d\bar{x}. \tag{3.3}$$

Here q_{max} denotes the maximal speed $|\mathbf{q}|$ in the domain.

Proof: As in [10, Theorem 4.1] we integrate the entropy inequality (3.2a) over the truncated cone $C = \{|\bar{x}| \leqq R + (t - \tau) \cdot q_{max} | 0 \leqq \tau \leqq t\}$; if we let $(n_0, \bar{n})$ denote the unit outward normal, then by Green's theorem

$$\int_{\partial \mathbf{C}} \rho h(S) \cdot \left[n_0 + \sum_{k=1}^{3} q_k n_k \right] \partial \bar{x} \geqq 0. \tag{3.4}$$

The integrals over the top and bottom surfaces give us the difference between the left and right-hand sides in (3.3) and by (3.4) this difference is bounded from below by

$$-\int_{\text{mantle}} \rho h(S) \cdot \left[n_0 + \sum_{k=1}^{d} q_k n_k \right] \partial \bar{x}.$$

The result follows upon showing that the last quantity is nonnegative. Indeed, since by assumption $-\rho h(S) \geqq 0$, this is the same thing as

$$n_0 + \sum_{k=1}^{3} q_k n_k \geqq 0;$$

on the mantle we have

$$(n_0, \bar{n}) = \frac{1}{\sqrt{1 + q_{\max}^2}} \left(q_{\max}, \frac{\bar{x}}{|\bar{x}|} \right),$$

and hence

$$n_0 + \sum_{k=1}^{3} q_k n_k = \frac{1}{\sqrt{1 + q_{\max}^2}} \left(q_{\max} + \sum_{k=1}^{3} \frac{q_k x_k}{|\bar{x}|} \right) \geqq \frac{1}{\sqrt{1 + q_{\max}^2}} \left(q_{\max} - \sum_{k=1}^{3} \frac{|q_k|^2}{|q|} \right) \geqq 0$$

as asserted.

The discussion in Lemma 3.1 was restricted to smooth function $h(S)$; by passing to the limit, its conclusion (3.3) follows for any nonpositive nondecreasing function $h(S)$ satisfying (3.2b), whether smooth or not.

To derive the minimum entropy principle, we now make a special choice of such function, h(S), given by

$$h(S) = \mathrm{Min}[S - S_0, 0], \quad S_0 = \operatorname*{Ess\ inf}_{|\bar{x}| \leq R + t \cdot q_{max}} S(\bar{x}, 0). \tag{3.5}$$

The nonpositive function $h(S)$ is a nondecreasing concave one, hence admissible by (3.2b), and consequently (3.3) applies

$$\int_{|\bar{x}| \leq R} \rho(\bar{x},t) \cdot \mathrm{Min}[S(\bar{x},t) - S_0, 0] d\bar{x} \geq \int_{|\bar{x}| \leq R + t \cdot q_{max}} \rho(\bar{x},0) \cdot \mathrm{Min}[S(\bar{x},0) - S_0, 0] d\bar{x}. \tag{3.6}$$

Now, by the choice of S_0, the integral on the right of (3.6) vanishes since $\mathrm{Min}[S(\bar{x},0)-S_0,0]$ does. The inequality (3.6) then tells us that the integral on the left is also nonnegative. But since the integrand on the left is by definition nonpositive, this can be the case provided this integrand vanishes almost everywhere; that is, we have for almost all $\bar{x}$, $|\bar{x}| \leq R$

$$S(\bar{x},t) \geq S_0 = \operatorname*{Ess\ inf}_{|\bar{x}| \leq R + t \cdot q_{max}} S(\bar{x}, t=0)$$

and (1.9) follows.

The minimum entropy principle was deduced from the entropy inequality (3.2), which in turn was postulated based on the formal regularization introduced in (1.4). In general, other regularizations equally apply; in

particular, Euler equations are usually sought as the vanishing viscosity limit of the Navier-Stokes equations (here we take for simplicity the one-dimensional case)[4]

$$\frac{\partial}{\partial t}\begin{bmatrix}\rho\\ m\\ E\end{bmatrix}+\frac{\partial}{\partial x}\begin{bmatrix}m\\ qm+p\\ q(E+p)\end{bmatrix}=\mu\,\frac{\partial}{\partial x}\begin{bmatrix}0\\ \frac{\partial q}{\partial x}\\ q\,\frac{\partial q}{\partial x}\end{bmatrix},\qquad \mu\to 0. \tag{3.7}$$

Do the (generalized) entropy inequalities (3.2) remain valid on the basis of such limit? To answer this question we first note that if $U(\mathbf{u})$ is any entropy function, then thanks to its convexity the mapping $\mathbf{u}\to\mathbf{v}\equiv U_{\mathbf{u}}$ is one-to-one, and hence one can make the change of variables $\mathbf{u}=\mathbf{u}(\mathbf{v})$. Harten [5] has shown that such change of variables by each member of the family of entropy functions (2.4) puts the viscosity terms on the right of (3.7) into a negative semidefinite form. This makes apparent the dissipative effect of these viscosity terms. Indeed, if $T=c_v\cdot E-\frac{1}{2}\cdot|\mathbf{q}|^2$ denotes the absolute temperature, then direct manipulation of (3.7) yields, e.g., [1, Section 63], [12, Section 6.10],

$$\frac{\partial}{\partial t}\left[\rho h(S)\right]+\frac{\partial}{\partial x}\left[mh(S)\right]=\mu\cdot h(S)\,\frac{q_x^2}{T}, \tag{3.8}$$

from which we recover the entropy inequality (3.2a) for all smooth increasing functions $h(S)$. We note that the convexity condition was not assumed in this

---

[4] With μ combining the two viscosity coefficients in the general Navier-Stokes equations.

case. The merit of using the convexity condition, however, is that it enables us to deal with more general artificial viscosity terms, other than those appearing in the Navier-Stokes equations. Such artificial viscosity terms are frequently encountered in finite-difference approximations to the Euler equations; a specific example of this kind is studied in the next section.

Finally we would like to remark on the previously mentioned Navier-Stokes equations. Our discussion above took into account only the viscosity contribution, neglecting heat conduction. Hughes, et al., [7] have shown that when the heat flux is also added, compare (3.7),

$$\frac{\partial}{\partial t}\begin{bmatrix}\rho\\ m\\ E\end{bmatrix} + \frac{\partial}{\partial x}\begin{bmatrix} m\\ qm + p\\ q(E + p)\end{bmatrix} = \mu \frac{\partial}{\partial x}\begin{bmatrix} 0\\ \frac{\partial q}{\partial x}\\ q\frac{\partial q}{\partial x}\end{bmatrix} + \kappa \frac{\partial}{\partial x}\begin{bmatrix}0\\ 0\\ \frac{\partial T}{\partial x}\end{bmatrix} \tag{3.9}$$

with κ denoting the heat conductivity constant, then only the 'physical' entropy, $U(\mathbf{u}) = -\rho S$ survives as the one which puts the additional heat flux into a symmetric negative-definite form. We would like to note in this connection the difference limit behavior of the Navier-Stokes flows depending on the viscosity and heat conductivity; Gilbarg [4] has shown that as $\kappa \to 0$ keeping μ fixed, we are led to a continuous thermally nonconducting shock layer, whereas for $\mu \to 0$ with κ fixed the convergence is to a (generally) discontinuous nonviscous shock layer. Consequently, the viscosity rather than the heat flux should play the major rule in an appropriate regularization model for the Euler equations.

4. DISCRETE APPROXIMATIONS OF THE EULER EQUATIONS

In this section we consider approximate solutions of the Euler equations, $w(x_\nu,t)$, whose total entropy decreases in time, compare (1.10)

$$\sum_\nu U(w(\bar{x}_\nu,t+\Delta t))\Delta\bar{x}_\nu \le \sum_\nu U(w(\bar{x}_\nu,t))\Delta\bar{x}_\nu. \tag{4.1}$$

Estimate (4.1) holds for <u>all</u> entropy functions $U = -\rho h(s)$ in (2.4). By passing to the limit, this applies to our previous choice of the function $h(s)$ in (3.5)

$$h(s) = \mathrm{Min}[S - S_0, 0], \tag{4.2a}$$

this time with a constant S_0 which is taken to be

$$S_0 = \min_\nu S(w(\bar{x}_\nu,t)). \tag{4.2b}$$

By our choice of S_0, we have $U(w(\bar{x}_\nu,t)) = 0$. The inequality (4.1) tells us that the left-hand side is therefore, nonnegative; consequently

$$S(\bar{x},t+\Delta t) - S_0 \ge h(S(x,t+\Delta t)) \ge 0$$

and (1.13) follows.

Approximate solutions which fulfill the required estimate (4.1) can be obtained by entropy stable schemes satisfying the cell entropy inequality (1.11)

$$U(w(\bar{x}_\nu,t+\Delta t)) \le U(w(\bar{x}_\nu,t)) + \sum_{k=1}^{d} \frac{1}{\Delta\bar{x}_\nu}\left[F^{(k)}_{\nu+1/2} - F^{(k)}_{\nu-1/2}\right]. \tag{4.3}$$

Examples of such entropy stable schemes include the Godunov-type and Lax-Friedrichs schemes, e.g., [6]. A more precise minimum principle follows in these cases, taking into account the support of the schemes' stencil. In particular, the (one-dimensional) Godunov scheme results from averaging of two neighboring Riemann problems [6], each of which satisfies (1.9). Consequently we have the

Minimum Principle (of the Godunov scheme): Let $\mathbf{w}(x_\nu,t)$ the Godunov approximate solution to the Euler equations (2.1). Assume that the appropriate CFL condition is met. Then the following estimate holds

$$S\big(\mathbf{w}(x_\nu,t+\Delta t)\big) \geq \min_{\nu-1\leq j\leq \nu+1} S\big(\mathbf{w}(x_j,t)\big). \tag{4.4}$$

Since the Lax-Friedrichs scheme coincides with a staggered Godunov's solver, the same conclusion, (4.4), holds. Another way to see this is outlined below; it makes no reference to Riemann's solution and can be generalized to the multidimensional problem.

To this end, we approximate the (for simplicity--one-dimensional) Euler equations with the Lax-Friedrichs scheme

$$\begin{aligned}\mathbf{w}(x_\nu,t+\Delta t) = \frac{1}{2}\Big[\mathbf{w}(x_{\nu+1},t) + \mathbf{w}(x_{\nu-1},t)\Big] \\ - \frac{\lambda}{2}\Big[f\big(\mathbf{w}(x_{\nu+1},t)\big) - f\big(\mathbf{w}(x_{\nu-1},t)\big)\Big], \quad \lambda \equiv \frac{\Delta t}{\Delta x}.\end{aligned} \tag{4.5}$$

We remark that the Lax-Friedrichs scheme can be derived from center differencing of the regularization model $(1.4)_{\Delta x}$. Lax has shown [9, Theorem

1.2] that if $\lambda \equiv \frac{\Delta t}{\Delta x}$ is sufficiently small, then solutions of this difference scheme satisfy the following cell entropy inequality

$$U(w(x_\nu, t + \Delta t)) \leqq \frac{U(w(x_{\nu+1}, t)) + U(w(x_{\nu-1}, t))}{2} - \frac{\lambda}{2}\left[F(w(x_{\nu+1}, t)) - F(w(x_{\nu-1}, t))\right] \tag{4.6}$$

for <u>all</u> entropy pairs $(U,F) = (-\rho h(S), -mh(S))$ in (2.4). by passing to the limit, this applies to our previous choice of the function $h(S)$ in (3.5)

$$h(S) = \text{Min}[S - S_0, 0], \tag{4.7a}$$

this time, with a contant S_0 which is taken to be

$$S_0 = \text{Min}[S(x_{\nu+1}, t),\ S(x_{\nu-1}, t)]. \tag{4.7b}$$

The inequality (4.6) now reads

$$\rho(x_\nu, t + \Delta t) \cdot h\big(S(x_\nu, t + \Delta t)\big) \geqq \left[\frac{1 + \lambda q(x_{\nu-1}, t)}{2} \rho(x_{\nu-1}, t) \cdot h\big(S(x_{\nu-1}, t)\big) + \frac{1 - \lambda q(x_{\nu+1}, t)}{2} \rho(x_{\nu+1}, t) \cdot h\big(S(x_{\nu+1}, t)\big)\right] \tag{4.8}$$

By our choice of the funtion $h(S)$ in (4.7), we have $h(S(x_{\nu\pm1}, t)) = 0$. The inequality (4.8) tells us that the left-hand side is therefore nonnegative; consequently

$$0 \leq h\big(S(x_\nu, t + \Delta t)\big) \leq S(x, t + \Delta t) - S_0$$

and the following minimum principle follows

$$S\big(w(x_\nu, t + \Delta t)\big) \geq \text{Min } S\big(w(x_{\nu \pm 1}, t)\big).$$

REFERENCES

[1] R. Courant and K. O. Friedrichs, Supersonic Flow and Shock Waves, Interscience, New York, 1948.

[2] R. J. DiPerna, "Convergence of approximate solutions to conservation laws," Arch. Rational Mech. Anal., Vol. 82 (1983), pp. 27-70.

[3] K. O. Friedrichs and P. D. Lax, "Systems of conservation laws with a convex extension," Proc. Nat. Acad. Sci. U.S.A., Vol. 68 (1971), pp. 1686-1688.

[4] D. Gilbarg, "The existence and limit behavior of the one-dimensional shock layer," Amer. J. Math., Vol. 73 (1951), pp. 256-274.

[5] A. Harten, "On the symmetric form of systems of conservation laws with entropy," J. Comput. Phys., Vol. 49 (1983), pp. 151-164.

[6] A. Harten, P. D. Lax, and B. Van Leer, "On upstream differencing and Godunov-type schemes for hyperbolic conservation laws," SIAM Rev., Vol. 25 (1983), pp. 35-61.

[7] T. J. R. Hughes, L. P. Franca, and M. Mallet, "Symmetric forms of the compressible Euler and Navier-Stokes equations and the second law of thermodynamics," Comput. Methods Appl. Mech. Engrg., to appear.

[8] S. N. Krushkov, "First-order quasilinear equations in several independent variables," Math. USSR-Sb., Vol. 10 (1970), pp. 217-243.

[9] P. D. Lax, "Shock waves and entropy" in Contributions to Nonlinear Functional Analysis (E. H. Zarantonello, ed.), pp. 603-634, 1971.

[10] E. Tadmor, "Skew-selfadjoint form for systems of conservations laws," J. Math. Anal. Appl., Vol. 703 (1984), pp. 428-442.

[11] E. Tadmor, "The numerical viscosity of entropy stable schemes for systems of conservation laws. I.," NASA Langley Research Center, ICASE Report 85-51, NASA CR-178021, 1985.

[12] G. Whitham, Linear and Nonlinear Waves, Wiley-Interscience, 1974.

A SPECTRAL MULTIDOMAIN METHOD

FOR THE SOLUTION OF HYPERBOLIC SYSTEMS

David A. Kopriva
Florida State University
and
Institute for Computer Applications in Science and Engineering

ABSTRACT

A multidomain Chebyshev spectral collocation method for solving hyperbolic partial differential equations has been developed. Though spectral methods are global methods, an attractive idea is to break a computational domain into several subdomains, and a way to handle the interfaces is described. The multidomain approach offers advantages over the use of a single Chebyshev grid. It allows complex geometries to be covered, and local refinement can be used to resolve important features. For steady-state problems it reduces the stiffness associated with the use of explicit time integration as a relaxation scheme. Furthermore, the proposed method remains spectrally accurate. Results showing performance of the method on one- dimensional linear models and one- and two-dimensional nonlinear gas-dynamics problems are presented.

Research was supported by the National Aeronautics and Space Administration under NASA Contract Nos. NAS1-17070 and NAS1-18107 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665-5225.

1. INTRODUCTION

In this paper we address the problem of efficiently computing Chebyshev spectral collocation approximations to quasilinear hyperbolic systems of the form

$$Q_t + A(Q)Q_x + B(Q)Q_y = 0 \qquad x,y \ \ D \subset R^2, \qquad t \geq 0 \tag{1}$$

with appropriate boundary and initial conditions. Here, Q is an m-vector and A and B are m×m matrices. This system is hyperbolic if for any constants k_1 and k_2 the matrix $T = k_1 A + k_2 B$ has only real eigenvalues and there exists a similarity transformation matrix, P, such that $PTP^{-1} = \Lambda$ is a real diagonal matrix.

In particular, we are interested in the solution of the Euler equations of gas dynamics which form a system of this type. The use of the nonconservation form is justified for problems in which shocks are fitted and in this situation spectral methods work well [1]. Problems of the type presented in Ref. [1] provide the motivation for what follows.

The typical Chebyshev spectral collocation procedure for the solution of the system (1) is described in several reviews such as those of Gottlieb, Hussaini, and Orszag [2], and Hussaini, Salas, and Zang [3]. First, the domain of interest is mapped onto the square $D' = [-1,1]\times[-1,1]$ and an $(N+M) \times (M+1)$ point mesh is generated with the collocation points defined by

$$\begin{aligned} x_i &= -\cos(i\pi/N) \qquad i = 0,1,\cdots,N \\ y_j &= -\cos(j\pi/M) \qquad j = 0,1,\cdots,M. \end{aligned} \tag{2}$$

Mesh point values of Q, designated by Q_{ij}, are associated with each of the collocation points (x_i,y_j). A global Chebyshev interpolant of order N in the x direction and order M in the y direction is then put through the mesh point values

$$Q_p(x,y) = \sum_{n,m=0}^{N,M} a_{nm}\, T_n(x)T_m(y). \tag{3}$$

Approximations to the derivatives at the collocation points are computed by differentiating the interpolant and evaluating the resulting polynomial at the collocation points. The computation of the derivatives can be accomplished in one of two ways (see Gottlieb, et al., [2]): The first is to take advantage of the fact that the sums for both the interpolant and its derivative reduce to cosine sums at the chosen collocation points. For example

$$\frac{dQ_p}{dx} = \sum_{n,m=0}^{N,M} a_{nm}T_n'(x)T_m(y) = \sum_{n,m=0}^{N,M} b_{nm}T_n(x)T_m(y) \tag{4}$$

where

$$b_{Nm} = 0,$$

$$b_{N-1,m} = 2Na_{nm} \tag{5}$$

and

$$c_n b_{nm} = b_{n+2,m} + 2(n+1)a_{n+1,m} \quad \text{for} \quad 0 \le n \le N-2.$$

The constant c_n is defined as $c_n = 2$ for $n = 0,N$ and $c_n = 1$ otherwise. The advantage of this form is that a fast cosine transform can compute the derivatives along each y line in $O(N \log N)$ operations.

The other approach to computing the derivatives is to write the differentiation operation as the product of a differentiation matrix and the a vector of the Q_{ij}'s. For example, along each y line the x derivative is

$$\left(\frac{d\mathbf{Q}_p}{dx}\right)_j = D(\mathbf{Q}_p)_j \tag{6}$$

where $(\mathbf{Q}_p)_j = [Q_{0,j}\ Q_{1,j}\ \cdots\ Q_{N,j}]^T$ and the elements of the matrix D are defined in Gottlieb et al., [2]. The amount of work with this procedure is of $O(N^2)$. What one loses in efficiency one gains as flexibility in the number of mesh points that can be used in each direction without adding storage.

No matter which way the spatial derivatives are computed, it is important to note that computing the Chebyshev derivative approximations requires only mesh point values. Derivatives at the end points require only points interior to the mesh so no extra procedure is required to compute derivatives at boundaries.

Once the spatial derivatives are approximated, what results is a system of ordinary differential equations in time for the variation of the solution at each collocation point (Method of Lines). Because the differentiation matrix is full, explicit methods are typically used to integrate the semi-discrete equations. In this paper, all time integrations will be performed with a fourth-order Runge-Kutta method.

The advantage of using this spectral method to solve (1) is that for solutions which are $C^{\infty}(D)$, the accuracy is better than any polynomial order (Canuto and Quarteroni, [4]). This is usually called "spectral accuracy" and asymptotic behavior can be observed if there are enough grid points to

adequately resolve the solution. It is thus possible to compute to a given spatial accuracy with fewer grid points than required by typical low-order finite difference approximations.

Balancing the high accuracy of the spectral method, however, are some major disadvantages of the typical Chebyshev collocation approach:

(1) It may not be easy or even possible to map $D \to D'$ globally.

(2) The collocation point distribution is global and predetermined. Local refinement of the mesh is not possible.

(3) The points are concentrated near the boundaries where they are typically not needed for hyperbolic problems.

(4) If explicit time integration is used the time step restriction in one dimension is proportional to $1/N^2$.

(5) For complete flexibility in the number of mesh points which can be used, the derivatives cost of $O(N^2)$ in each direction.

These problems can be reduced significantly by breaking up the region D into several subdomains D_k each of which has its own Chebyshev grid. With a stable and efficient method for computing the interfaces, the advantages of such an approach would be:

(1) Complicated geometries can be covered.

(2) Points can be distributed with some flexibility; local refinement is possible.

(3) In one dimension, with N points and K subdomains, the time step restriction increases to $\Delta t \propto K/N^2$.

(4) Derivative evaluation work with matrix multiplication decreases to $K(N/K)^2$ or $1/K$ that of a single grid.

The idea of breaking up the computational domain into subdomains each with a different grid is not new. For finite difference methods this is a currently popular approach (e.g., [5]). For spectral methods, however, previous applications have been limited to elliptic and parabolic problems. Orszag [6] first applied such a technique to solve elliptic problems. He enforced continuity of the function and its first derivative as the interface condition. Metivet and Morchoisne [7] and later, Morchoisne [8] computed multidomain solutions to the Navier-Stokes equations. Recently, Patera [9] and Korczak and Patera [10] have been using a spectral element method to solve the incompressible Navier-Stokes equations. Their method is very similar to the p finite-element methods developed by Babuska (see [10]) but uses Chebyshev interpolants. The treatment of the convective terms, however, does not lend itself to purely convective problems. For these problems, we describe the method below.

2. MULTIDOMAIN APPROACH

In this paper, we will break up the physical domain, D, into K subdomains D_k which do not overlap except for the common boundary points. Figure 1 shows a rectangular two-dimensional example of the situation with four subdomains. Each of the D_k are mapped onto a square $[-1,1]\times[-1,1]$. Spatial approximations at interior points of each subdomain are computed in the usual way. Across an interface, however, there are two values of the normal derivative. For example, at the y coordinate line interface between D_1 and D_2 in Figure 1, derivative approximations are available from the left and from the right. The problem is to choose properly information from

the right and the left to give a stable and consistent approximation to the differential equation at the interface.

Before discussing a multidomain method for the boundary value problem (1), we will first examine the one-dimensional case. In one dimension, we seek interface algorithms of the semidiscrete form

$$\frac{\partial Q^I}{\partial t} + A^L \frac{\partial Q^L}{\partial x} + A^R \frac{\partial Q^R}{\partial x} = 0 \tag{7}$$

where Q^I denotes the value of Q at an interface and the derivatives superscripted with L and R denote the two spectral approximations computed in the left and right, respectively. For consistency, we require that

$$A^L + A^R = A \tag{8}$$

and for efficiency we want A^L and A^R to be computed with little more work than is required for the computation of A itself.

To generate the coefficient matrices, consider first the linear <u>scalar</u> hyperbolic equation

$$u_t + \lambda u_x = 0 \qquad \lambda > 0. \tag{9}$$

Because the equation is hyperbolic, it is clear that the common interface point should depend only on information propagated from the left. Thus, the approximation <u>should</u> be

$$\frac{\partial u^I}{\partial t} + \lambda \frac{\partial u^L}{\partial x} = 0. \tag{10}$$

This is, of course, just upwind differencing at the interface and is equivalent to the way Gottlieb and Orszag [11] handled a tau approximation to equation (9). To simplify the computational logic to include cases where the coefficient, λ, is of either sign, the approximation (10) can be written as

$$\frac{\partial u^I}{\partial t} + \tfrac{1}{2}(\lambda + |\lambda|)\frac{\partial u^L}{\partial x} + \tfrac{1}{2}(\lambda - |\lambda|)\frac{\partial u^R}{\partial x} = 0. \tag{11}$$

If we now consider that this equation is a single component of a diagonalized system, where the diagonal matrix

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_n \end{bmatrix} = P^{-1} AP,$$

we can write the system as

$$\frac{\partial Q^I}{\partial t} + \tfrac{1}{2}(A + |A|)\frac{\partial Q^L}{\partial x} + \tfrac{1}{2}(A - |A|)\frac{\partial Q^R}{\partial x} = 0 \tag{12}$$

where $|A| = P|\Lambda|P^{-1}$. Formally, this is nothing more than the method of characteristics in one dimension.

We now propose to avoid the computation of the matrix absolute value by approximating it with a diagonal matrix

$$|A| \approx P\lambda^* IP^{-1} = \lambda^* I \tag{13}$$

where λ^* is chosen to lie between the largest and smallest elements of $|\Lambda|$. The boundary scheme is now of the form of Eq. (7) with

$$A^L = \tfrac{1}{2}(A + \lambda^* I) \qquad A^R = \tfrac{1}{2}(A - \lambda^* I). \tag{14}$$

This choice of coefficient matrices always has proper upwind dominance on all of the characteristic variables, but includes some downwind influence. To see this, re-diagonalize the system (7) and use u as the n^{th} component of the diagonalized system. Then the approximation to the method of characteristics causes the characteristic variables at the interface to be approximated by

$$\frac{\partial u^I}{\partial t} + \tfrac{1}{2}(\lambda_n + \lambda^*)\frac{\partial u^L}{\partial x} + \tfrac{1}{2}(\lambda_n - \lambda^*)\frac{\partial u^R}{\partial x} = 0. \tag{15}$$

In fact, this can be viewed as the purely upwind scheme with an error term: For the $\lambda_n > 0$ case,

$$\frac{\partial u^I}{\partial t} + \lambda_n \frac{\partial u^L}{\partial x} = (\lambda^* - \lambda_n)\left(\frac{\partial u^R}{\partial x} - \frac{\partial u^L}{\partial x}\right). \tag{16}$$

Thus, we have the spectrally accurate upwind approximation with an error term proportional to the difference of the right and left spectral derivatives. If the solution has the necessary smoothness, this difference should also decay spectrally and spectral accuracy of the approximation should be retained.

We will study the stability of the multidomain method with the interface approximation (14) numerically. An analytic study of stability is not possible at this time. Stability theory for Chebyshev approximations to hyperbolic initial-boundary value problems is not advanced enough to analyze an approximation which introduces some downwind influence at the interface.

We consider the two-domain approximation of the scalar equation (9) with the interface approximation (12) with $\lambda = 1$. The line segment [-2,2] is divided equally into two domains of [-2,0] on the left and [0,2] on the right. The semidiscrete approximation can be written as a system of ordinary equations with the two-domain coefficient matrix

$$\begin{bmatrix} D^L & 0 \\ 0 & D^R \end{bmatrix} \tag{17}$$

where D^L and D^R are the single domain differentiation matrices for the left and the right, modified to include the interface approximation. For this system to be time stable, that is, the solution does not grow unboundedly as $t \to \infty$, the eigenvalues of the coefficient matrix must have negative real parts.

Figure 2 shows how the eigenvalues change as λ^* varies when 6 points are used. The case of $\lambda^* = 0$ corresponds to simple averaging and is clearly not time stable. Choosing $\lambda^* > 0$ large enough moves the eigenvalues into the left half of the complex plane and the resulting approximation is time stable. The case of $\lambda^* = 1$ is the purely upwind case and the eigenvalues decouple into two single-domain patterns. If λ^* is chosen equal to, or larger than, the wave speed, λ_n, the approximation has the effect of adding a purely dissipative term to the equation and two purely real eigenvalues are created. If λ^* is very much larger than λ_n, however, the eigenvalues migrate to the right of the imaginary axis. The range of λ's for which the approximation is stable decreases as the disparity in the number of points

becomes larger; for very stiff systems, it may be necessary to use $|A|$ instead of λ^* at the interface.

It is interesting to note that the reverse situation, where there is more resolution on the upstream side of the interface, does not show this behavior and is stable for all $\lambda^* \geq 0$. For systems, this means that λ^* should be chosen to be only slightly larger than the smallest eigenvalue representing a characteristic moving from the coarse to the fine grid. For systems, this means that λ^* should be chosen to be only slightly larger than the smallest eigenvalue representing a characteristic moving from the coarse to the fine grid. We note, however, that the examples on which the scheme has been tested show that the approximation is robust over a wide range of choices of λ^*.

In two dimensions, the upwind weighted approximation is used in the direction perpendicular to the interface. Returning to Figure 1, along x coordinate lines, the y derivatives are continuous across the interfaces except at corners. At points not on the corners, then, we propose using

$$\frac{\partial Q^I}{\partial t} + A^L \frac{\partial Q^L}{\partial x} + A^R \frac{\partial Q^R}{\partial x} + B \frac{\partial Q^I}{\partial y} = 0 \tag{18}$$

where A^L and A^R are defined as above. Along x coordinate interfaces,

$$\frac{\partial Q^I}{\partial t} + A \frac{\partial Q^I}{\partial x} + B^T \frac{\partial Q^L}{\partial y} + B^B \frac{\partial Q^R}{\partial y} = 0 \tag{19}$$

where $B^T = 1/2\,(B + \mu^* I)$ and $B^B = 1/2\,(B - \mu^* I)$ and μ^* is an approximation to the eigenvalues of B. At corners, the weighted approximations are used in both directions.

3. NUMERICAL EXAMPLES

Numerical experiments on four model problems in one and two dimensions will be presented. The models include the scalar one-dimensional hyperbolic initial boundary value problem for a travelling Gaussian pulse, a linear system in one dimension, quasi-one-dimensional flow in a converging-diverging nozzle, and the transonic Ringleb problem. The Ringleb flow models the smooth nonlinear transonic flow in a curved duct and has an exact solution to which to compare.

A. Solution of a Linear Scalar Problem

The solution to the linear scalar problem

$$\frac{\partial u}{\partial t} + 2\frac{\partial u}{\partial x} = 0 \qquad x \in [-2,2], \quad t > 0 \qquad (20)$$

$$u(x,0) = \exp\left(-(x - x_0)^2/0.3\right) \qquad x \in [-2,2]$$

$$u(-2,t) = \exp\left(-(x - t - x_0)^2/0.3\right) \qquad t > 0$$

can be used to examine the effects of varying λ^* in the spatial approximation described in Eq. (15). The time integration for this and all following examples was a fourth-order Runge-Kutta technique. For this and the next model problem the time step was chosen so that the temporal errors were on the order of 10^{-10}. The main questions to be answered here are the effect of the $\lambda^* \neq 2$ on the accuracy of the solution and if reflections are a problem at the interface. Figure 3 shows the computed (circles) and exact (line) solutions for the pulse after it has propagated through the interface at $x = 0$ for two distributions of the mesh points and $\lambda^* = 6$.

The interface approximation Eq. (15) degrades the accuracy of the solution when compared to the purely characteristic interface, $\lambda^* = 2$, if equal resolution is not provided in each subdomain. In no case, however, is the global L_2 error larger than the global error for the characteristic interface. Furthermore, if λ^* remains fixed and the total number of points is increased, the error decay remains spectral. Figure 4 shows the pointwise errors of the solution to Eq. (20) for the situations represented in Figure 3 as λ^* is increased beyond the characteristic value of 2. The situation is worse when more resolution is used upstream of the interface because the approximation includes more and more downwind influence as λ^* is increased. In a practical computation, the effect of the boundary approximation would not be important if the solution were equally resolved in all subdomains.

Reflections at the interface are not visible in Figure 3 even though there is a factor of two difference in the number of collocation points. Gottlieb and Orszag [11] also noticed this for a tau approximation to the scalar wave equation. This is typical for the spectral approximations; examples with up to a factor of three and four in the ratio of the number of mesh points have not shown spurious reflections off of the interface.

B. A Linear System Example

The accuracy of the interface approximation will now be demonstrated with the 2×2 linear system

$$\begin{bmatrix} u \\ v \end{bmatrix}_t + \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}_x \qquad x \in [-2,2], \quad t > 0. \tag{21}$$

The coefficent matrix has eigenvalues +3 and -1 so the system has information which propagates in both directions and with different speeds across the interface at $x = 0$. The initial and boundary conditions were chosen so that the characteristic variables were the Gaussian pulses used in the scalar problem, Eq. (20). The coefficient λ^* for this case was chosen to be the maximum eigenvalue, $\lambda = +3$. Figure 5 shows the results for the two components of this system at a time when the characteristic pulses have crossed the interface. In Figure 5a there are twice as many points to the left of the interface as to the right and this is reversed for Figure 5b. The symbols represent the computed solutions and the solid lines represent the exact solutions.

A study of discrete L_2 errors for the system computations is shown in Tables I through III. Clearly, the error is spectral for all three situations. In fact, for an equal number of mesh points on either side of the interface, the error decay is exponential. For the problem of propagating pulses, where the features needing higher resolution are continually moving, it is not surprising that the best errors are obtained when there are an equal number of mesh points on both sides of the interface.

C. Quasi-One-dimensional Nozzle Flow

One potential point of concern in using the interface approximation given by Eq. (14) regards the stability of cases where one of the eigenvalues of the coefficient matrix is much larger than any other. Such a situation occurs at sonic points in an ideal gas flow where one of the characteristic speeds actually vanishes.

To test this situation the nonlinear problem of steady gas flow in a quasi-one-dimensional converging-diverging nozzle was solved with the multidomain method where an interface was placed at the sonic point. The quasilinear form of the Euler gas dynamics equations for time-dependent flow in a quasi-one-dimensional nozzle without shocks can be written as

$$\begin{bmatrix} P \\ u \end{bmatrix}_t + \begin{bmatrix} u & \gamma \\ a^2/\gamma & u \end{bmatrix} \begin{bmatrix} P \\ u \end{bmatrix}_x = \begin{bmatrix} \gamma u A_x(x)/A(x) \\ 0 \end{bmatrix} \tag{22}$$

where P is the logarithm of the pressure, u is the gas velocity, γ is the ratio of specific heats, and a is the sound speed. The coefficient matrix has eigenvalues of $u + a$ and $u - a$ so that one of them is zero at a sonic point. The steady flow is found as the large time limit of the unsteady flow described by (22).

The nozzle area is given by $A(x) = x/2 + 1/x$ so the throat occurs at $x = \sqrt{2}$. For the cases run, a subsonic inflow boundary was placed at $x = 0.2$ and characteristic boundary conditions were used. After the gas accelerates through the sonic value at the throat, it leaves the nozzle supersonically so no boundary conditions are applied at the outflow.

For the gas dynamics calculations in one dimension, $\lambda^* = \frac{1}{2}(|u+a| + |u-a|)$ was chosen since this corresponds to the diagonal elements of the absolute value of the coefficient matrix. Although the problem was solved for domain interfaces in both the subsonic and supersonic portions of the nozzle, only results for a single interface at the sonic point will be shown here. (The two-dimensional example below will include a variety of interface placements.)

Figure 6 shows the steady pressure in the nozzle computed with two domains and twice as many mesh points on the right as on the left. Our tests on a variety of grids have not shown any stability difficulties in computing steady flows when placing the interface at a sonic point.

D. Two-Dimensional Transonic Flow

A more complicated problem is the two-dimensional transonic Ringleb flow. This problem allows us to study the computational efficiency of the multidomain solution algorithm as outlined in the Introduction. Kopriva, et al., [12] used this problem for a comparison of the performance of the spectral method with a second-order finite-difference method. In this section we will compare the multidomain spectral method with the single domain spectral method.

The Ringleb flow is a simple example of a two-dimensional transonic flow for which there is an exact solution. (See, for example, Courant and Friedrichs [13].) The streamlines of the physical space solution appear at large distances as parabolas which are determined from a special hodograph solution of the potential equation for steady irrotational isentropic flow. By choosing two streamlines to represent solid walls, this problem models a steady transonic flow in a duct. Figure 7 shows the Mach contours of one such duct flow.

Again we will look for the large time solution of the unsteady gas dynamics equations, this time in two dimensions. The problem in the curved duct shown in Figure 7 is mapped onto a rectangle in the stream function-potential (ψ,ϕ) coordinate system derived from the exact solution. In this

coordinate system, the unsteady equations can be written as

$$Q_t = -R \tag{23}$$

where R is the steady state residual

$$R = AQ_\phi + BQ_\psi. \tag{24}$$

Since the solution is irrotational, the solution vector is chosen to be

$$Q = [P \; u \; v]^T \tag{25}$$

and the coefficient matrices are

$$A = \begin{bmatrix} U & \phi_x & \phi_y & 0 \\ a^2 \phi_x/\gamma & U & 0 & 0 \\ a^2 \phi_y/\gamma & 0 & U & 0 \\ 0 & 0 & 0 & U \end{bmatrix} \qquad B = \begin{bmatrix} V & \psi_x & \psi_y & 0 \\ a^2 \psi_x/\gamma & V & 0 & 0 \\ a^2 \psi_y/\gamma & 0 & V & 0 \\ 0 & 0 & 0 & V \end{bmatrix}$$

As before, P represents the logarithm of the pressure and (u,v) represent the velocity components in the Cartesian x and y directions, respectively. The matrix coefficients are computed from the mapping derived from the exact solution and the contravariant velocity components are $U = u\phi_x + v\phi_y$ and $V = u\psi_x + v\psi_y$.

The physical boundary conditions for this problem represent subsonic inflow at the entrance of the duct (at the lower left of Figure 7), supersonic outflow at the exit, and the sides are treated as impermeable boundaries (walls). So that the initial boundary value problem is well-posed the boundary conditions must be chosen carefully. See Kopriva, et al., [12] for details of the procedure which follows. For the subsonic inflow, we can specify only two quantities and have chosen the total enthalpy and the angle of the flow (so $V = 0$). The quantities P and U are computed from two conditions: The first is a compatibility equation derived from the pressure equation and the normal momentum equation. The second comes from differentiating the enthalpy equation in time. From U and the condition $V = 0$, the Cartesian velocities u, v can be computed. At the outflow, no boundary conditions are needed. Finally, at the walls the normal velocity, U, must vanish. The vector Q is computed by solving the tangential momentum equation for V and a compatibility equation which combines the normal momentum and pressure equations for P.

The system of equations (22) were discretized as described above, and fourth-order Runge-Kutta was used for the time integration. For a single domain, the Chebyshev spectral grid for the Ringleb problem with 16 streamwise and 8 normal mesh intervals is shown in Figure 8. It is clear that the spectral method strongly concentrates the grid points near the walls. The largest gradients, however, occur in the streamwise direction near the sonic line (as can be seen in Figs. 7 and 9) where the streamwise mesh distribution is coarsest. These two factors contribute to the fact that the time integration step is very small and that accuracy is degraded by the lack of resolution where it is needed.

A multidomain grid distribution for which performance will be compared to the single domain method is shown in Figure 10. Six domains now cover the duct and the same number of mesh intervals as for the single domain case are used. The divisions were chosen to demonstrate the kinds of situations which the multidomain method should be able to handle. Three divisions with $6 + 5 + 5$ mesh intervals are in the streamwise direction and two are in the normal direction. With this choice, two points occur where the corners of four domains come together. The first domain boundary in the streamwise direction was chosen to appear in a subsonic region of the duct. The second domain boundary in the streamwise direction was chosen to intersect the sonic line. By dividing the normal direction into two domains, the effective mesh spacing near the walls is doubled. Finally, note that by comparing Figure 10 to Figure 7 the sonic line also intersects the domain interface in the normal direction.

To allow comparison, Figure 11 shows the Mach number contours for both the single domain and the multidomain solutions. Note particularly that the sonic line remains smooth through the domain interfaces. Table IV summarizes the performance of the single domain spectral method compared with this particular choice of grid. First, note that even with this distribution of domains, the maximum error in the pressure for the multidomain computation has not been degraded from the single grid one. In fact, the error is five percent better.

The real advantage that the splitting has had for this case, however, is that the multidomain solution relaxes more quickly to steady state for a given number of intervals and accuracy. Figure 12 compares the rate at which the discrete L_2 norm of the residual of the pressure decays for the single and multidomain cases. The results are also summarized in Table IV. From the

trend of the graph, it should take over 2 1/2 times as many iterations for the single grid residual to decay to that of the single grid residual. This is a direct result of the fact that larger time steps can be used for the multi-domain case. The choice of λ^* also affects the convergence rate: larger values up to the stability limit give faster convergence to steady state.

The advantage of a k-domain derivative computation requiring 1/k the amount of work as a single domain computation does not show up in this example. In fact, as Table IV indicates, the average time per iteration (time step) requires the same amount of time at 0.5 sec. on the Langley Cyber 855. This is due to the fact that there is overhead in computing the interface approximation. Doubling the number of points in each direction with the same domain distribution decreases the time per iteration for the multidomain computation to 70% of the single domain cost. Though no attempt was made to compute the interface conditions efficiently, the number of points inside each domain will have to be large compared to the number of domains for the efficiency gained by being able to use fewer points in computing derivatives to become important.

The final advantage of a multidomain method which was listed in the Introduction is that flexibility in the choice of grid point distribution is now possible. A series of calculations were made with the duct being divided into two domain intervals in each direction. As with Figure 10, the direction across the duct was divided in half and the same number of mesh points was used. In the streamwise direction, however, only one domain boundary was inserted. This boundary was inserted in several places along the duct with different numbers of points on either side.

Results of some of the computations are summarized in Table V. The division is reported in terms of the fraction of the total variation of the velocity potential along the length of the duct. The first entry in the list places the division approximately near the bend of the duct where the gradients of the solution are the highest. It is clear that with a proper choice of grid it is possible to obtain better accuracy with the multidomain distribution of a given number of grid points than with a single grid. For the best case computed here, the error is about 2 1/2 times better for the multidomain calculation.

The problem of how to properly distribute points and subdomains in general is a major one and is beyond the scope of this paper. If they are poorly placed the error can be worse than the single domain error (see Table V). For now, it is not known how to obtain the optimal point and subdomain distribution. Rather, some knowledge of the behavior of the solution must be used as a guide.

CONCLUSIONS

We have described a simple approximation which allows a multidomain spectral solution of quasilinear hyperbolic equations. Numerical examples of linear equation models and ideal gas flow show that the method gives advantages in both accuracy and efficiency over using a single domain. Dividing up a computational domain into several subdomains gives the possibility of local refinement and allows some flexibility in the distribution of mesh points. It is possible to obtain better accuracy by doing so. Also, with multiple domains it is possible to take larger time

steps than with a single domain. This increases the efficiency for using time relaxation to acheive steady state solutions.

The use of a multidomain technique is also appropriate if discontinuities are fitted as boundaries. When shocks occur within a flow, subdomains would be arranged so that each shock lies on a subdomain boundary. In smooth parts of the solution, the technique described here would be used. Along shock interfaces, a shock fitting algorithm like that described in reference [1] can be used (Kopriva and Hussaini, to be published).

The theoretical issues which remain are many. Some theory for the range of values which λ^* can take for the method to be stable must be found. However, choosing λ^* to be the average of the largest and smallest eigenvalues of the coefficient matrix has always worked. Finally, like the problems associated with the p- version of the finite-element method, the choice of domain and point distribution for a given number of points is an open issue.

ACKNOWLEDGEMENTS

The author would like to thank Dr. S. F. Davis and Professor L. N. Trefethen for helpful comments and suggestions, and the Massachusetts Institute of Technology for computer equipment used in the course of the investigation.

REFERENCES

1. M. Y. Hussaini, D. A. Kopriva, M. D. Salas, and T. A. Zang, "Spectral methods for the Euler equations: part II - Chebyshev methods and shock fitting," AIAA J., 23 (1985), 234.

2. D. Gottlieb, M. Y. Hussaini, and S. Orszag, "Theory and application of spectral methods," in Spectral Methods for Partial Differential Equations, SIAM, Philadelphia, 1984.

3. M. Y. Hussaini, M. D. Salas, and T. A. Zang, "Spectral methods for inviscid, compressible flows," in Advances in Computational Transonics, (W. G. Habashi, Ed.), Pineridge Press, 1983.

4. C. Canuto and A. Quarteroni, "Error estimates for spectral and pseudospectral approximations of hyperbolic equations," SIAM J. Numer. Anal., 19 (1982), 629.

5. M. Berger and A. T. Jameson, "Automatic adaptive grid refinement for the Euler equations," AIAA J., 23 (1985), 561.

6. S. A. Orszag, "Spectral methods for problems in complex geometries," J. Comput. Phys., 37 (1980), 70.

7. B. Metivet and Y. Morchoisne, "Multi-domain spectral technique for viscous flow calculation," in Proceedings of the 4th GAMM conference on

Numerical Methods in Fluid Mechanics, (H. Viviand, Ed.), p. 207, Vieweg, 1982.

8. Y. Morchoisne, "Inhomogeneous flow calculations by spectral methods: mono-domain and multi-domain techniques," in Spectral Methods for Partial Differential Equations, (D. Gottlieb, M. Y. Hussaini, and R. G. Voigt, Eds.), p. 181, SIAM, Philadelphia, 1984.

9. A. T. Patera, "A spectral element method for fluid dynamics: laminar flow in a channel expansion," J. Comput. Phys., 54 (1984), 468.

10. K. Z. Korczak and A. T. Patera, "An isoparametric spectral method for solution of the Navier-Stokes equations in complex geometries," J. Comput. Phys., to appear.

11. D. Gottlieb and S. Orszag, "Numerical Analysis of Spectral Methods: Theory and Application," SIAM, Philadelphia, 1977.

12. D. A. Kopriva, T. A. Zang, M. D. Salas and M. Y. Hussaini, "Pseudospectral solution of two-dimensional gas-dynamic problems" in Proceedings of the 5th GAMM Conference on Numerical Methods in Fluid Mechanics, (M. Pandolfi and R. Piva, Eds.), Vieweg, 1984.

13. R. Courant and K. O. Friedrichs, Supersonic Flow and Shock Waves, New York, Springer-Verlag, 1976.

TABLE I. L_2 errors for the solutions to Eq. (20) with equal number of points on each side of the interface.

N	Error in u	Error in v
8	1.57×10^{-2}	1.49×10^{-2}
16	4.15×10^{-6}	4.86×10^{-6}
32	1.91×10^{-9}	1.91×10^{-9}

TABLE II. L_2 errors for the solutions to Eq. (20) with more points to the right of the interface.

N_L, N_R	Error in u	Error in v
8, 16	1.22×10^{-2}	1.05×10^{-2}
12, 24	2.45×10^{-4}	2.33×10^{-4}
16, 32	3.93×10^{-6}	3.93×10^{-6}

TABLE III. L_2 errors for the solutions to Eq. (20) with more points to the left of the interface.

N_L, N_R	Error in u	Error in v
16, 8	9.80×10^{-3}	1.04×10^{-2}
24, 12	3.48×10^{-4}	2.88×10^{-4}
32, 16	1.49×10^{-6}	2.30×10^{-6}

TABLE IV. Performance comparison for single and multidomain spectral computations.

Grids: Single Domain (SD) 17 × 9 points
Multidomain (MD) (7 + 6 + 6) × (5 + 5) points
(separated by domain)

Maximum Error	
SD	1.85×10^{-3}
MD	1.74×10^{-3}
Number of Steps to Reduce Residual Three Orders of Magnitude	
SD	> 1500
MD	780
Average Spectral Radius	
SD	0.9964
MD	0.9942
Average Time per Iteration	
SD	0.50 sec.
MD	0.50 sec.

TABLE V. Effect of streamwise mesh distribution on Ringleb calculation.

Grid	Division	Maximum Error
8 + 8	0.45 + 0.55	7.8×10^{-4}
8 + 8	0.50 + 0.50	9.3×10^{-4}
16(SD)	---	1.9×10^{-3}
10 + 6	0.34 + 0.66	1.2×10^{-2}

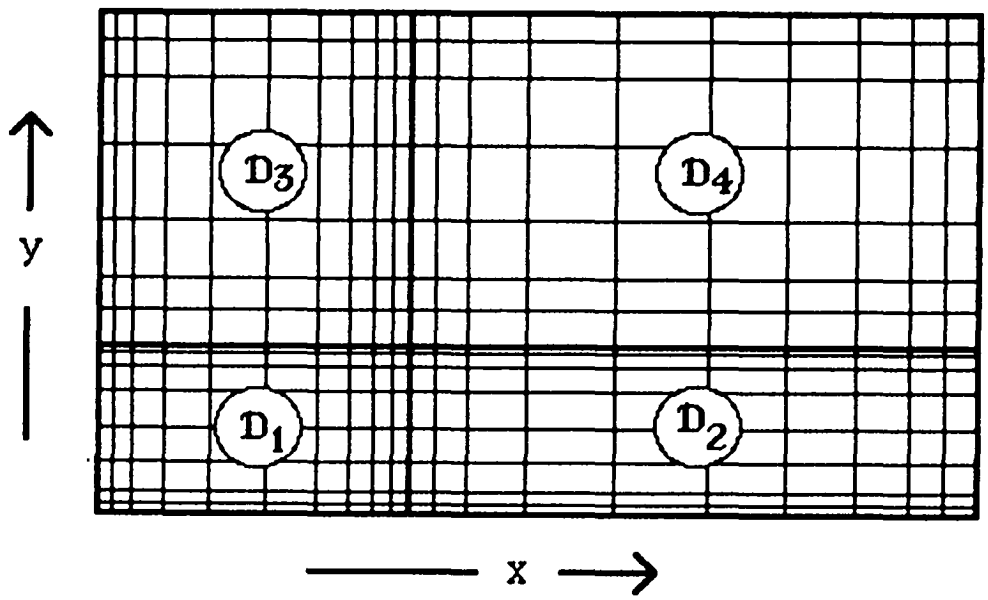


FIG. 1. Diagram of the two-dimensional subdomain structure used to divide a computational domain.

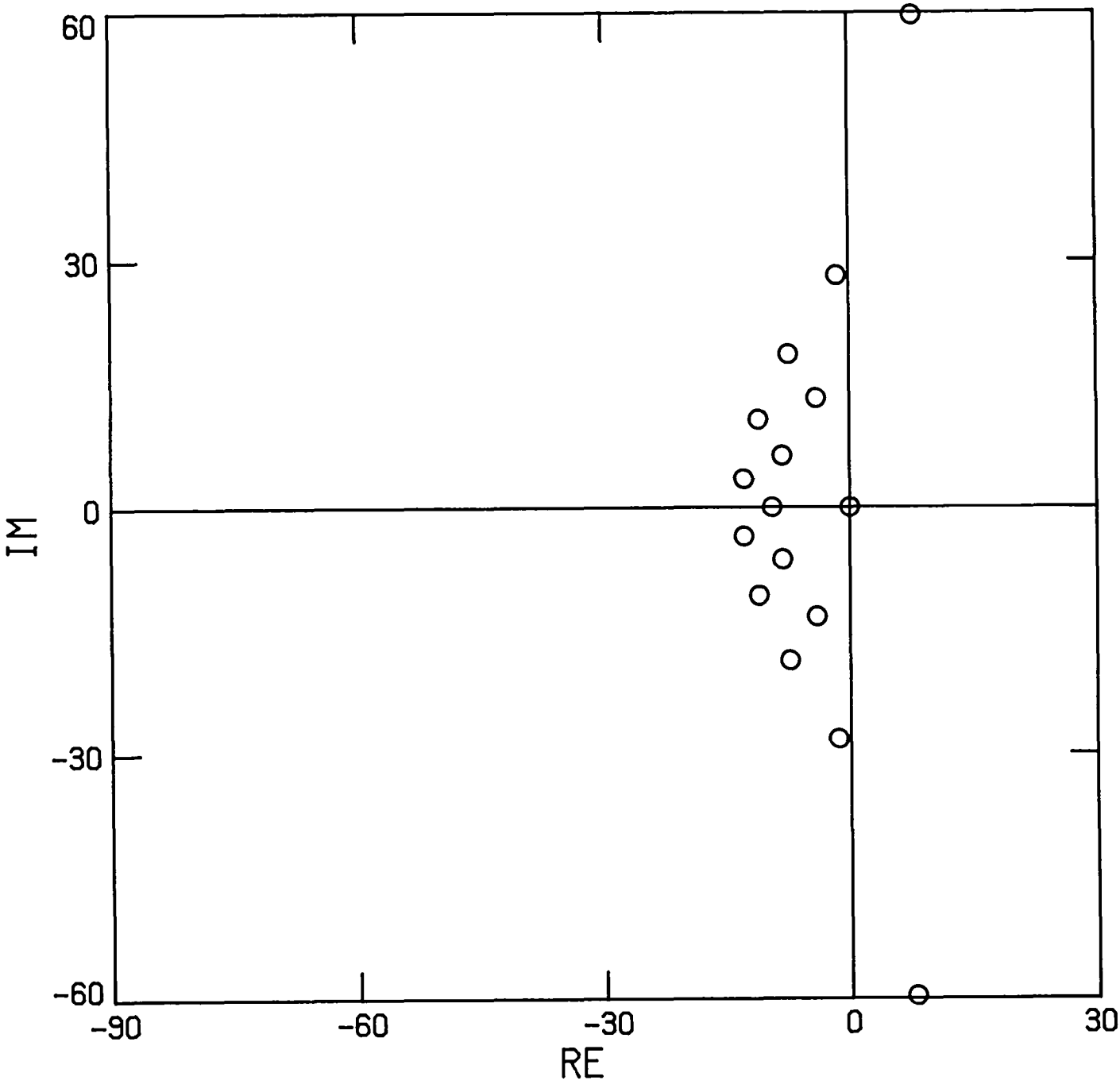


FIG. 2a. Effect on the eigenvalues of the two domain spatial approximation of the first derivative by varying λ^* in the boundary approximation: $\lambda^* = 0$.

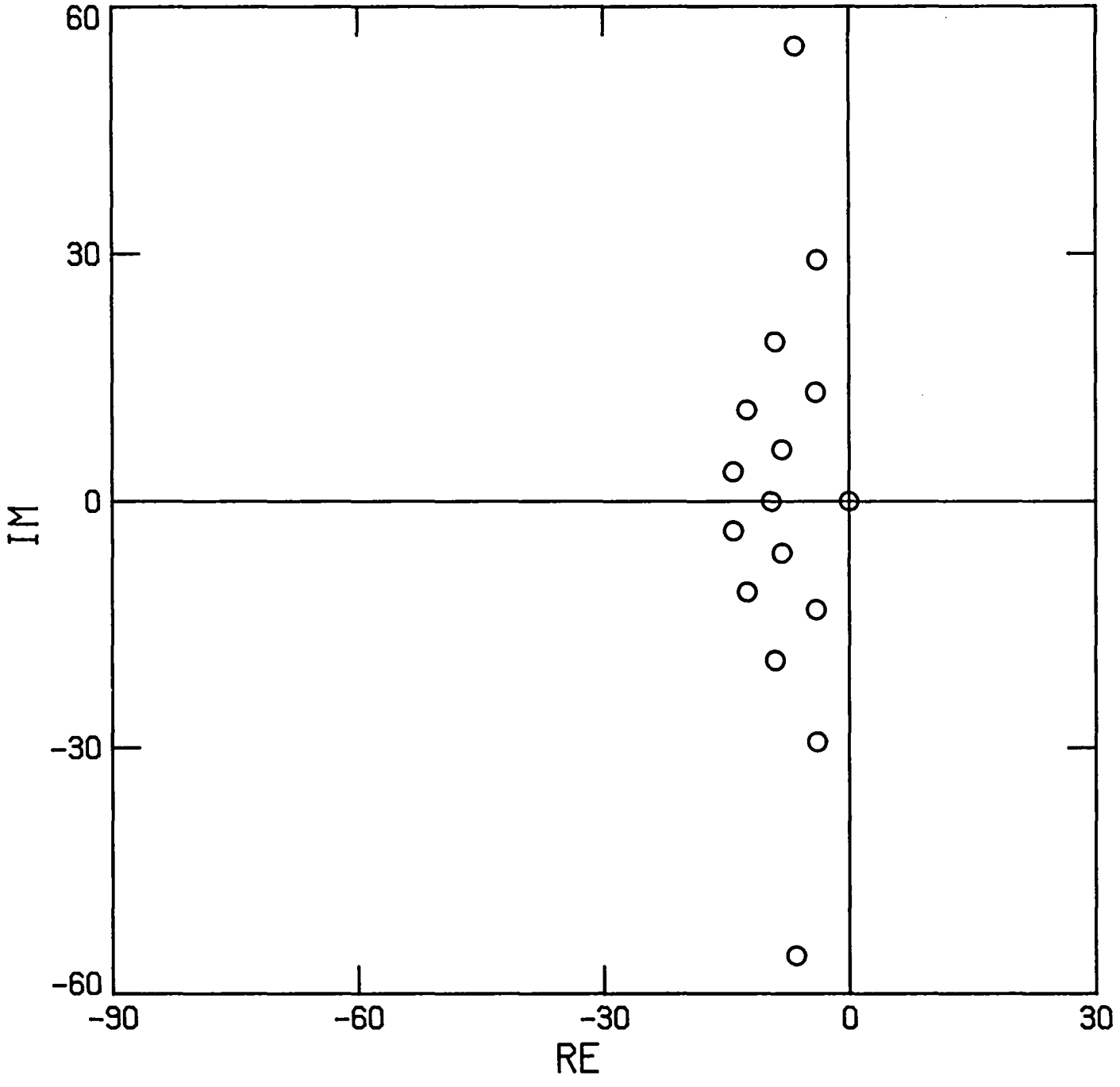


FIG. 2b. Effect on the eigenvalues of the two domain spatial approximation of the first derivative by varying λ^* in the boundary approximation: $\lambda^* = 0.5$.

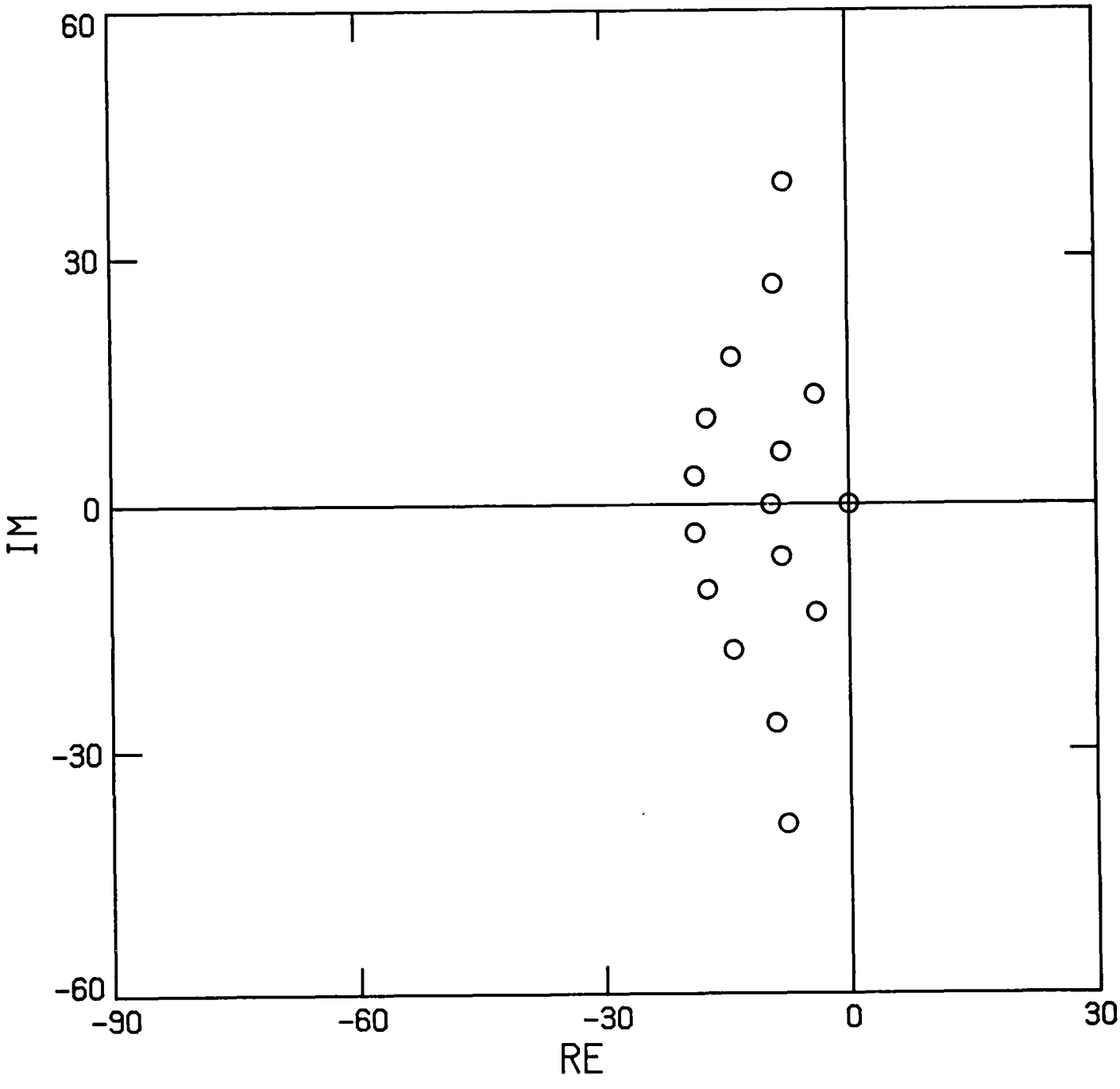


FIG. 2c. Effect on the eigenvalues of the two domain spatial approximation of the first derivative by varying λ^* in the boundary approximation: $\lambda^* = 1.0$ (purely upwind).

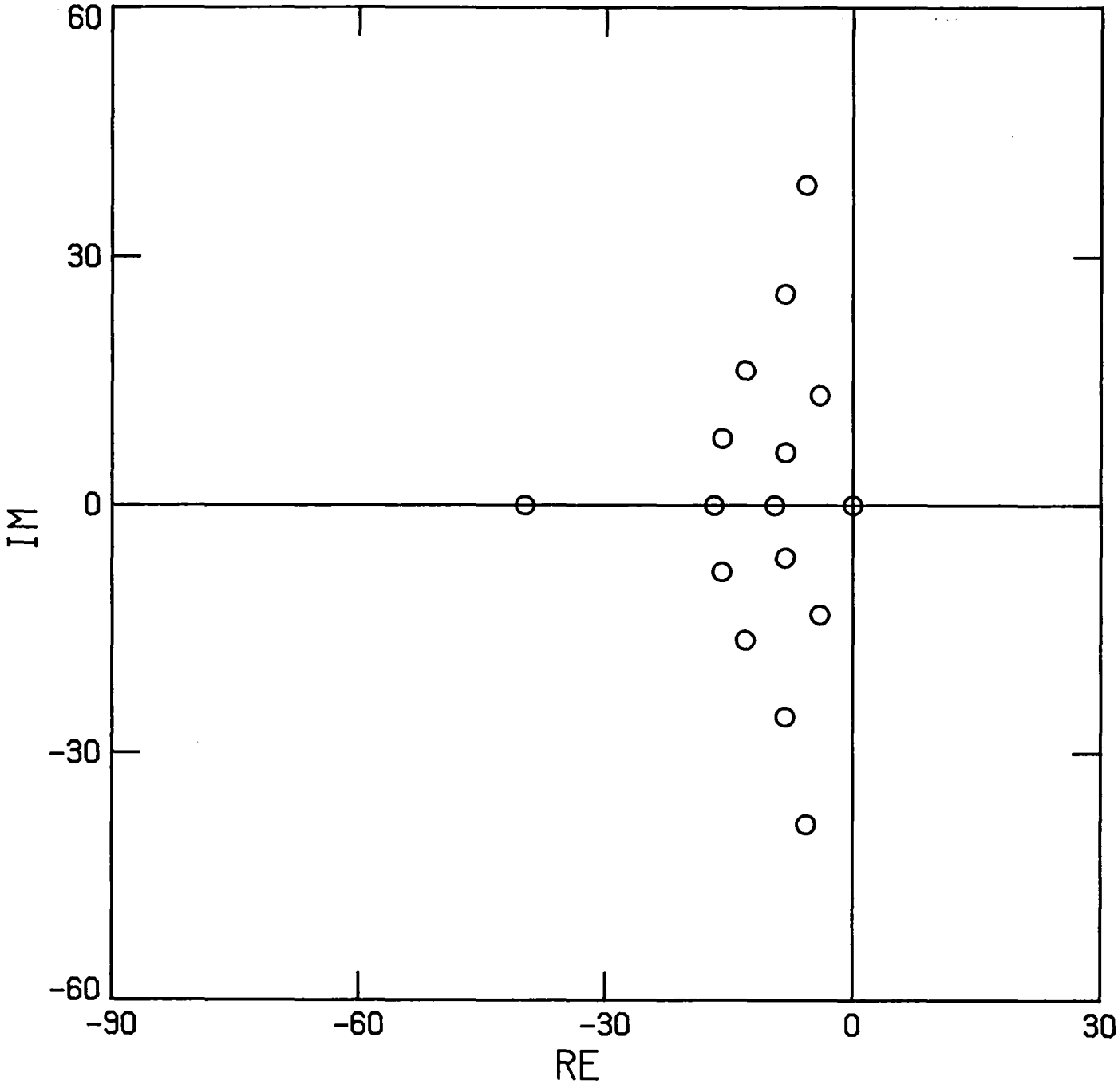


FIG. 2d. Effect on the eigenvalues of the two domain spatial approximation of the first derivative by varying λ^* in the boundary approximation: $\lambda^* = 1.1$.

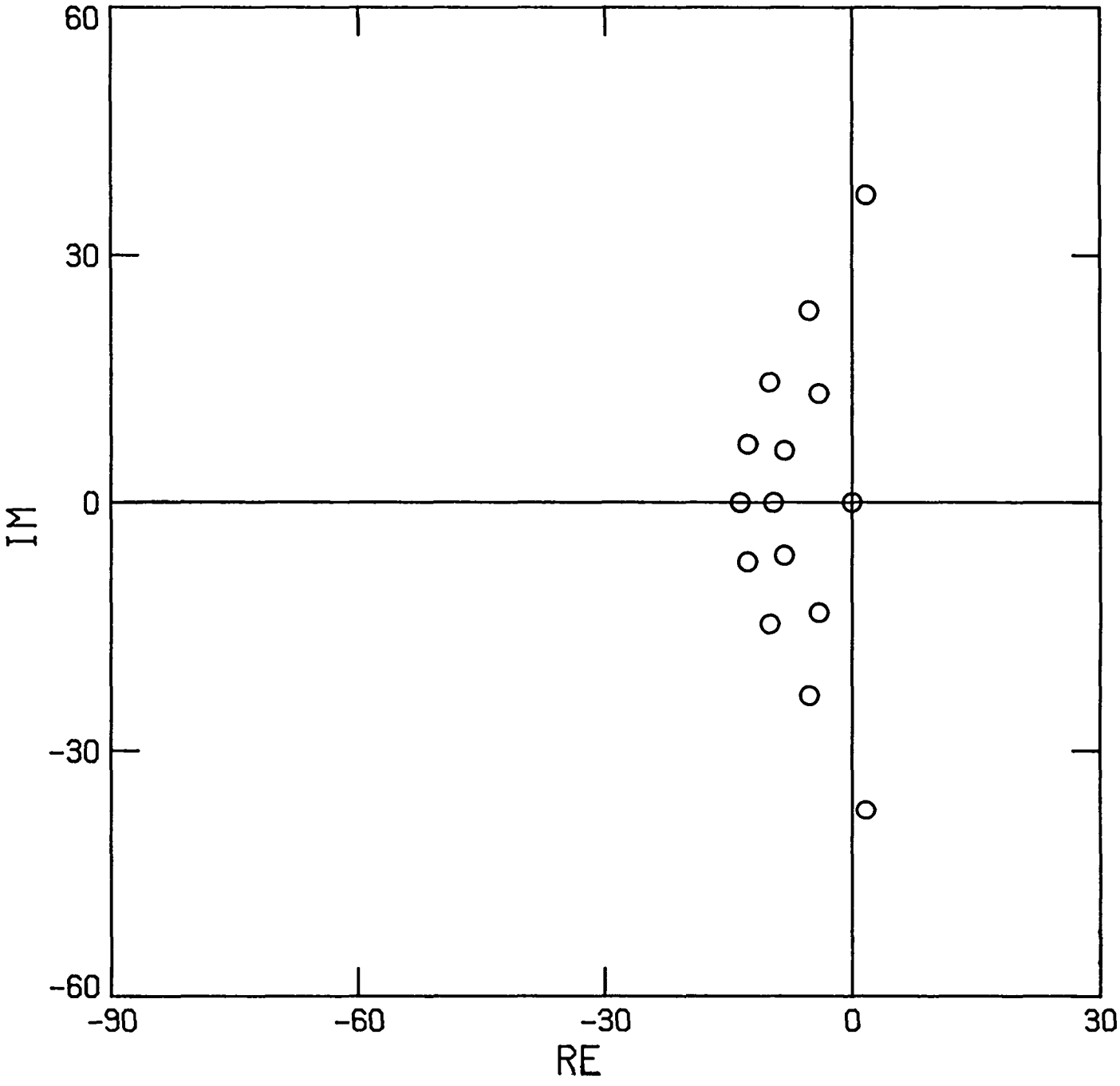


FIG. 2e. Effect on the eigenvalues of the two domain spatial approximation of the first derivative by varying λ^* in the boundary approximation: $\lambda^* = 5.0$.

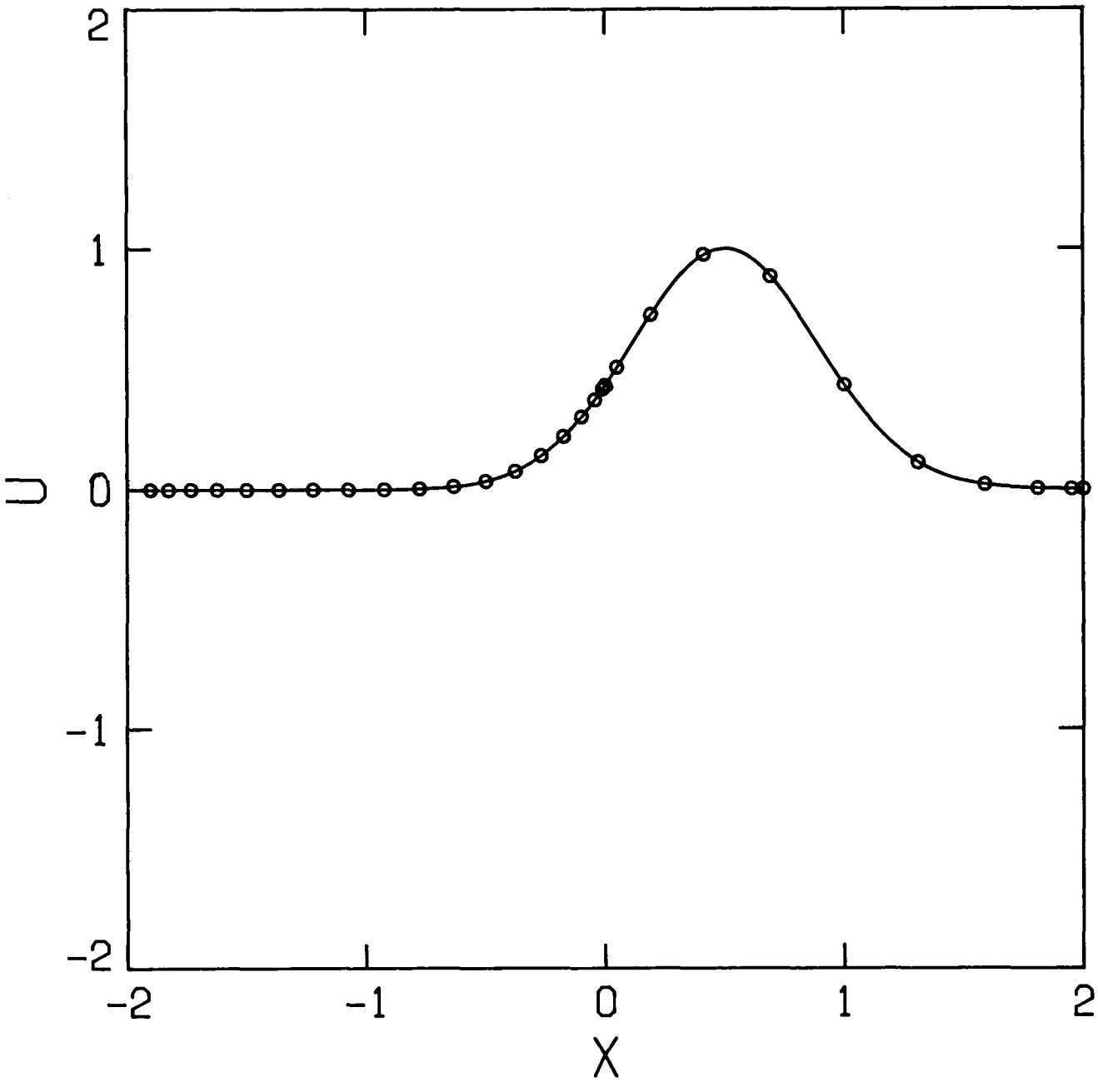


FIG. 3a. Solution of the scalar pulse problem Eq. (19) computed on two domains shown after the pulse has travelled from the left through the interface at x = 0. Computations are for 22 points left and 11 points right of the interface. The exact solution is the solid line; computed solutions are the circles.

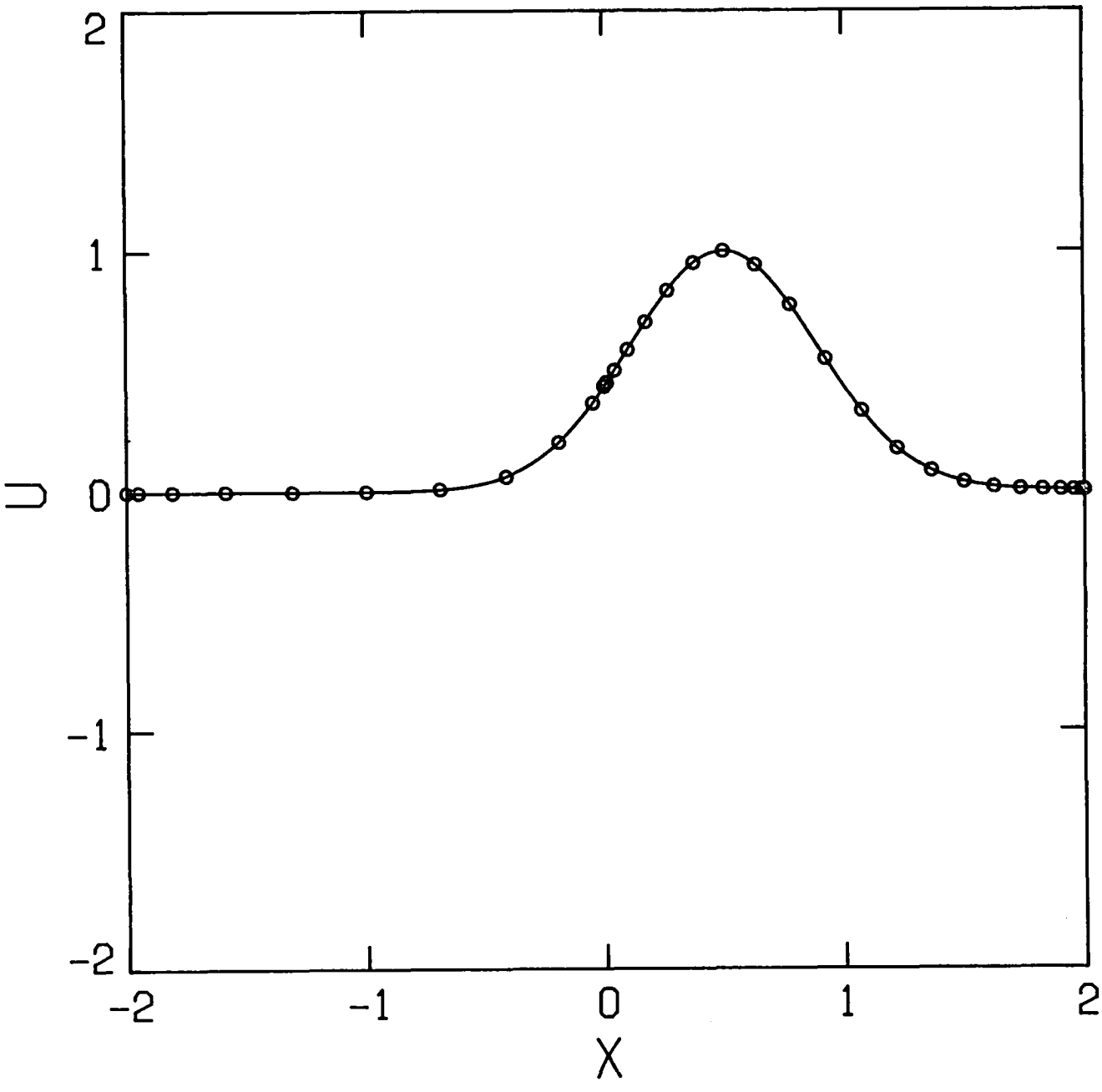


FIG. 3b. Solution of the scalar pulse problem Eq. (19) computed on two domains shown after the pulse has travelled from the left through the interface at $x = 0$. Computations are for 11 points left and 22 points right. The exact solution is the solid line; computed solutions are the circles.

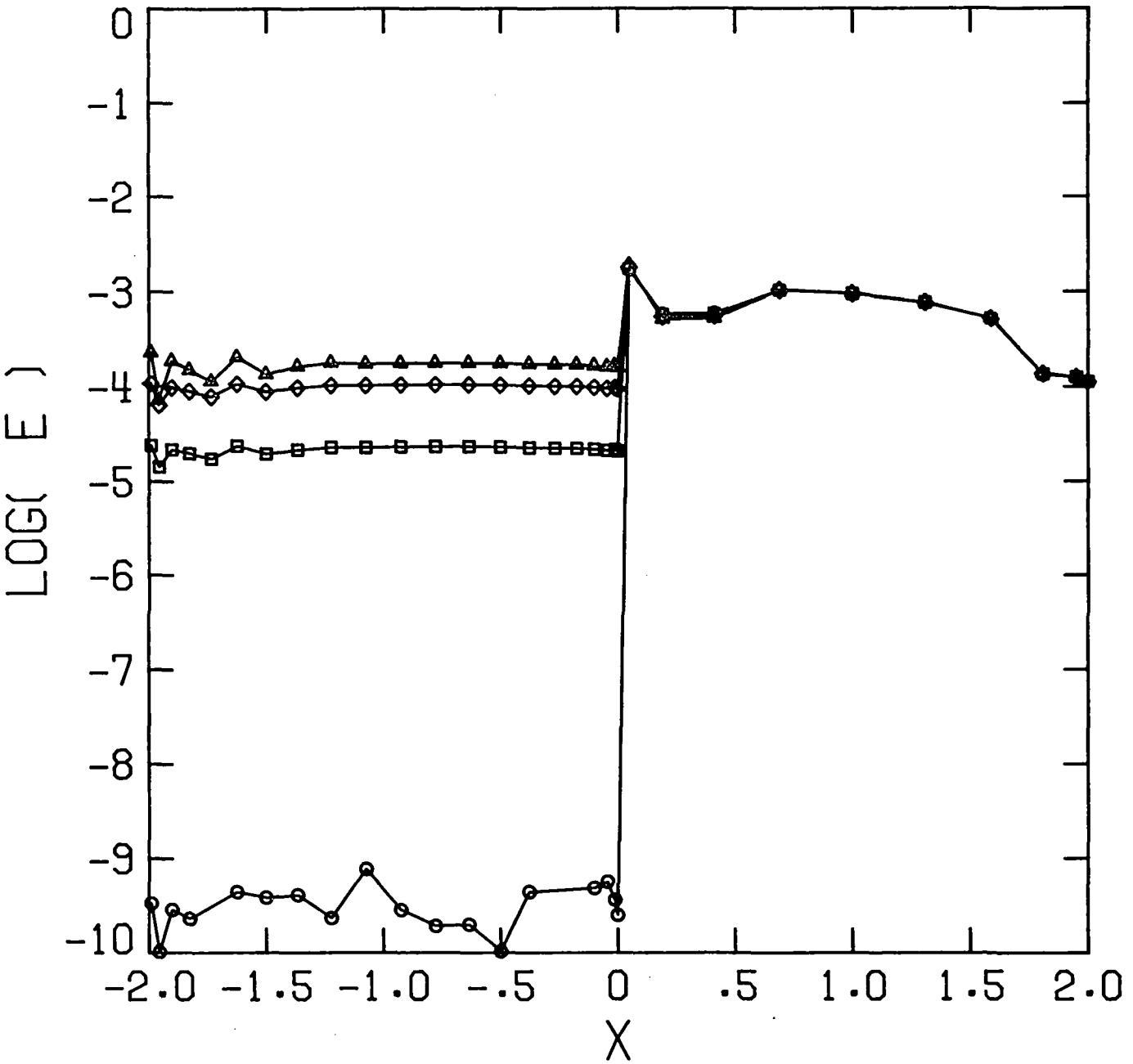


FIG. 4a. Pointwise errors as λ^* varies for the situation in Figure 3a.

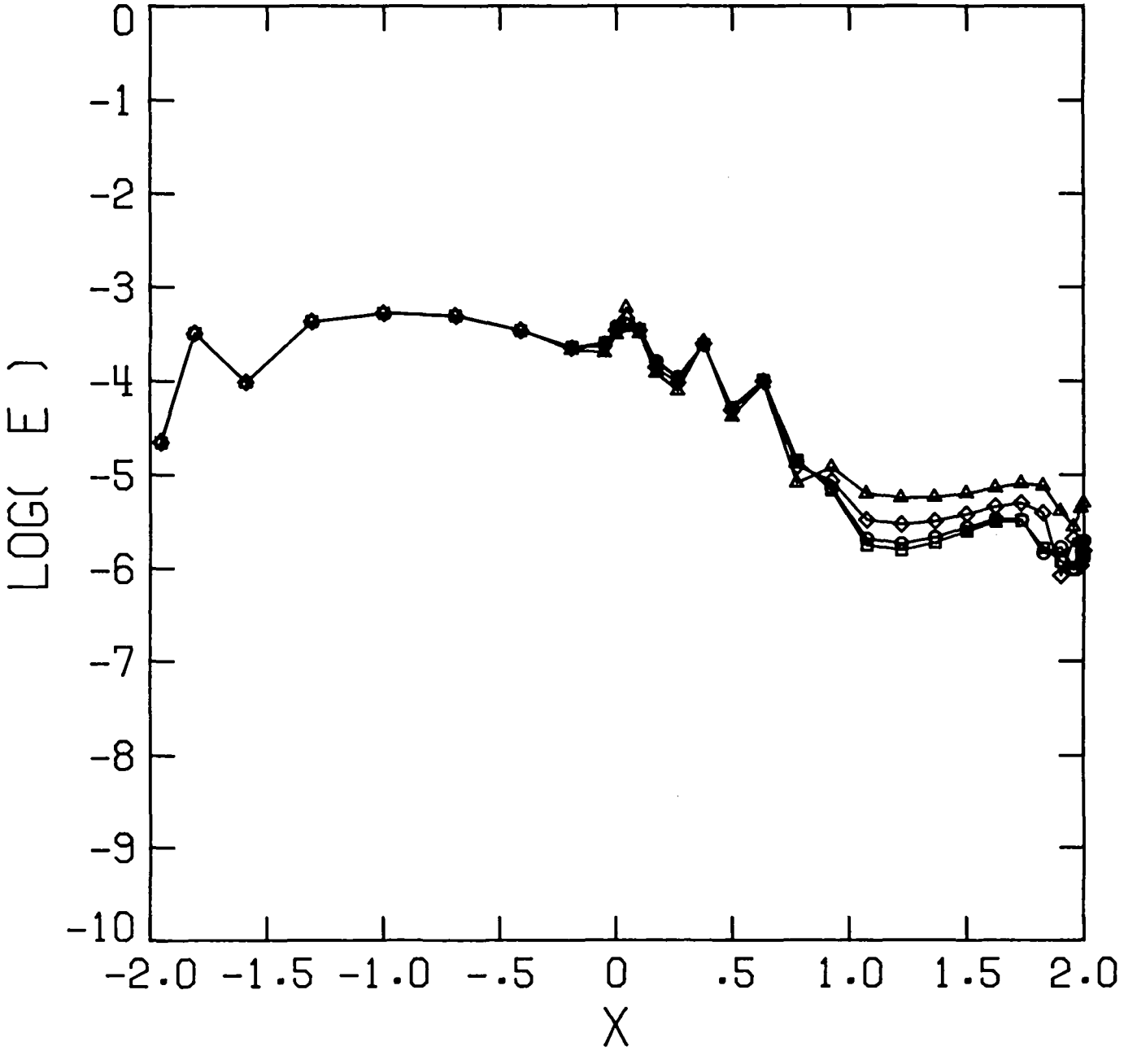


FIG. 4b. Pointwise errors as λ^* varies for the situation in Figure 3b.

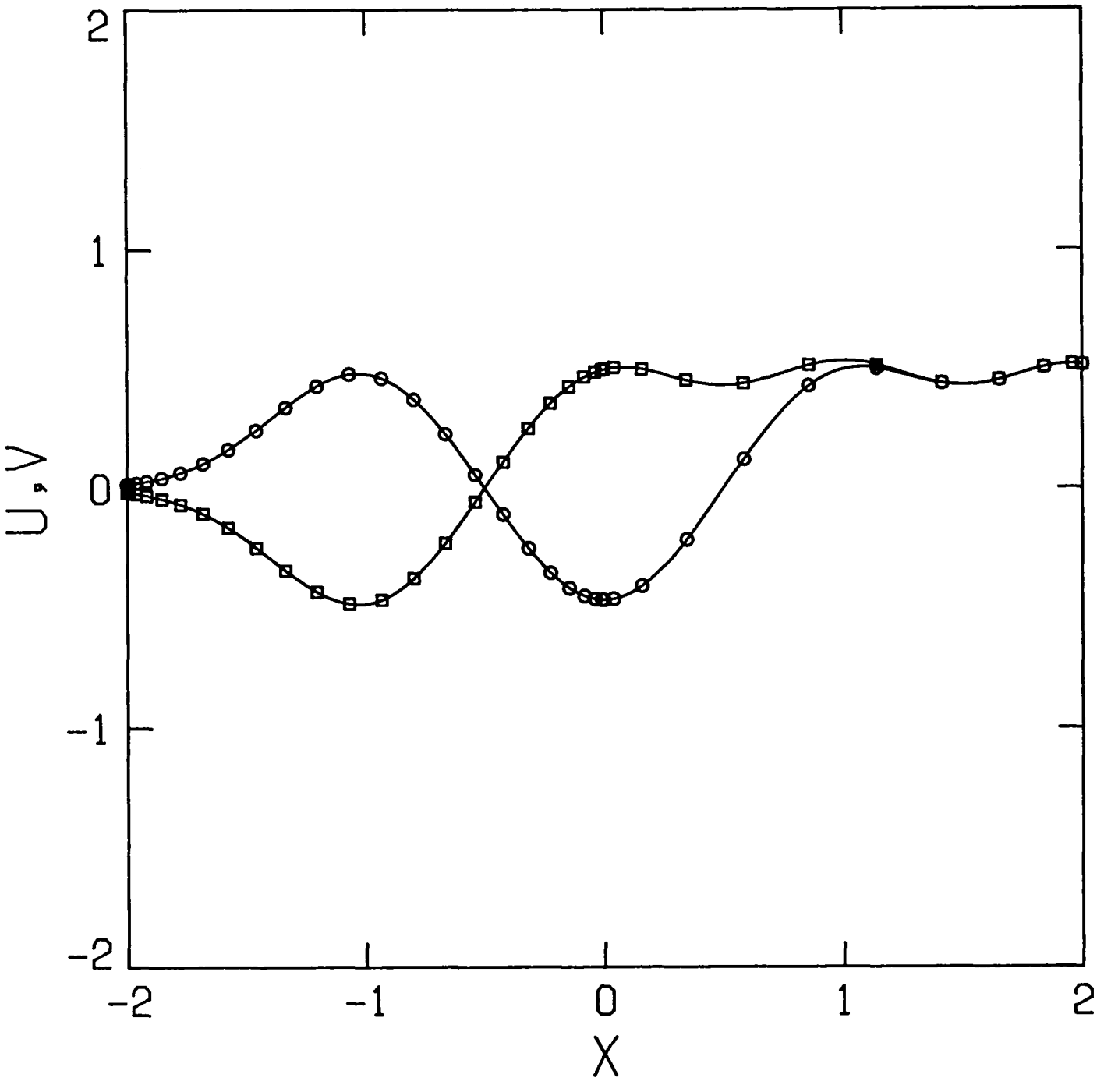


FIG. 5a. Graph of the two solutions u (circles) and v (squares) of the linear system Eq. (20) with 22 points on the left and 11 points on the right. The exact solutions are represented by the solid line.

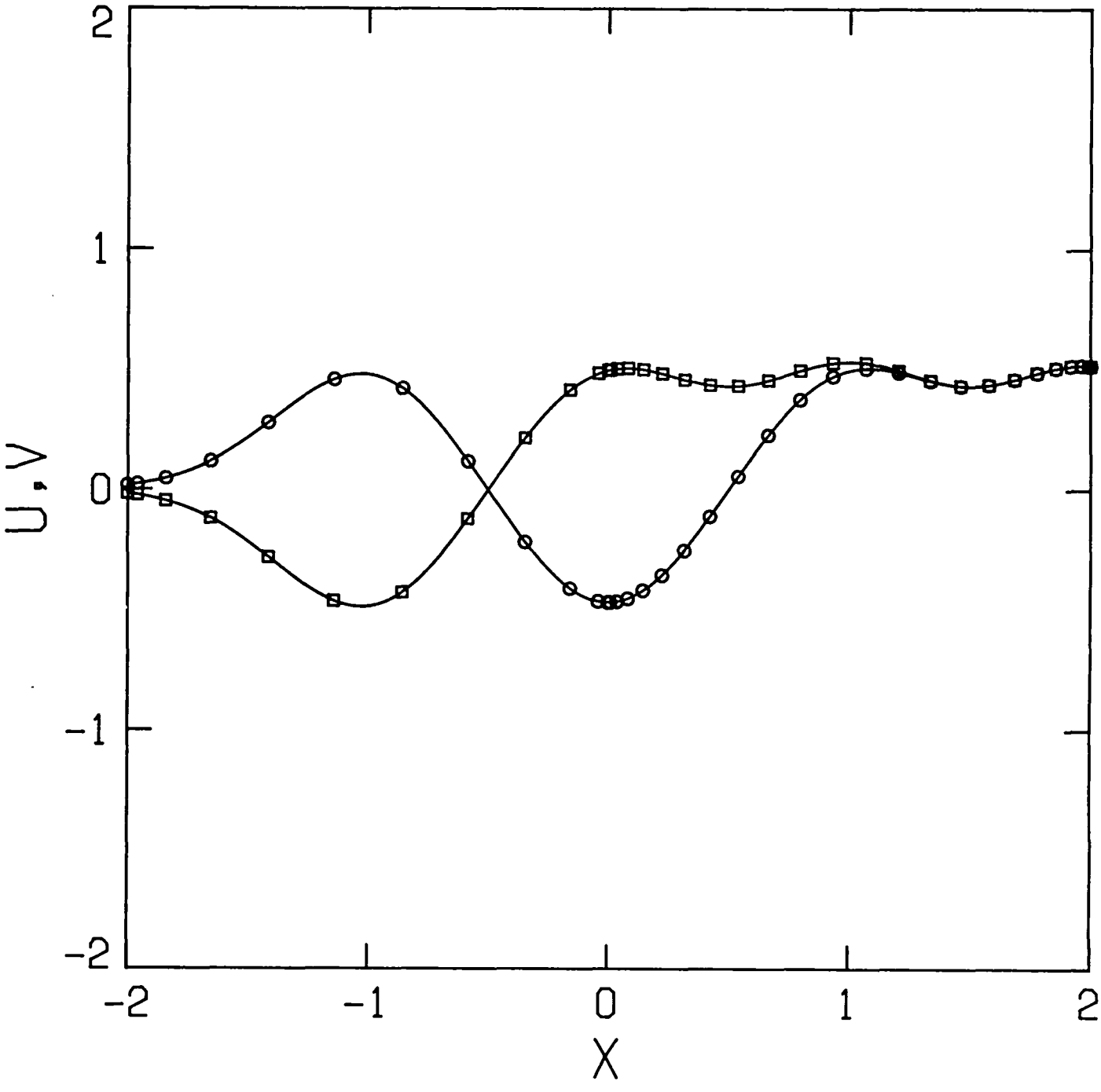


FIG. 5b. Graph of the two solutions u (circles) and v (squares) of the linear system Eq. (20) with 11 and 22 points on the left and the right. The exact solutions are represented by the solid line.

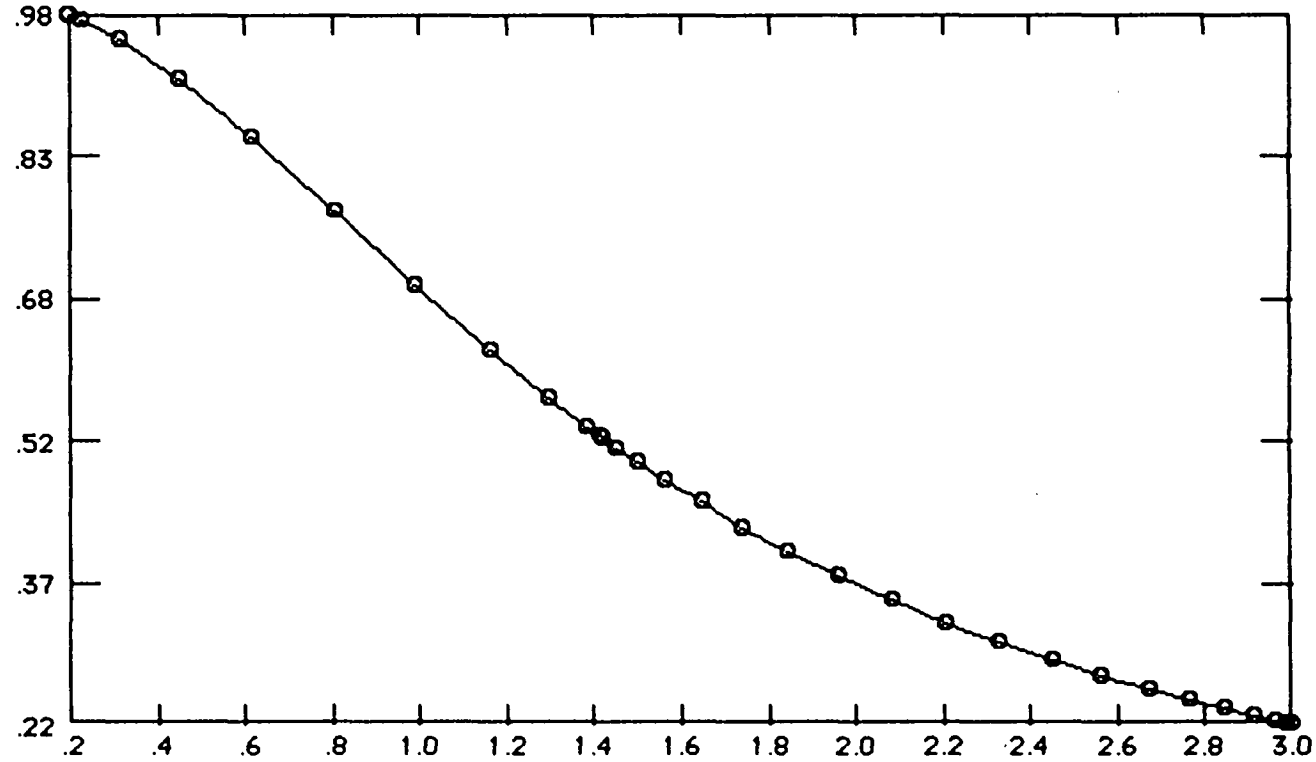


FIG. 6. Plot of the computed pressure in a converging-diverging nozzle where the interface is placed at the sonic point at $x = \sqrt{2}$. Twice as many points are used on the right as on the left of the interface.

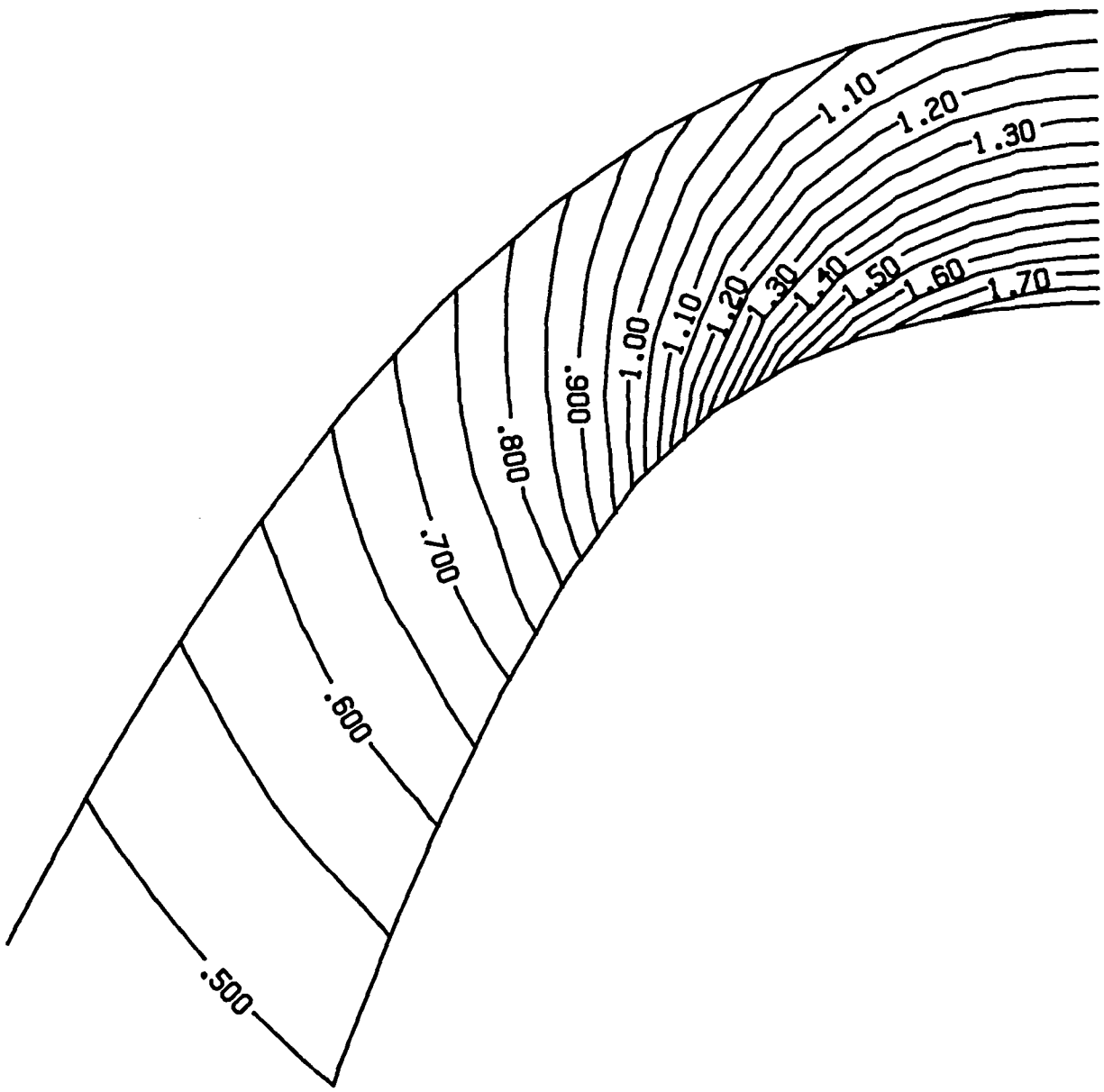


FIG. 7. Mach contours of the exact solution to the Ringleb problem which models transonic flow in a two-dimensional duct.

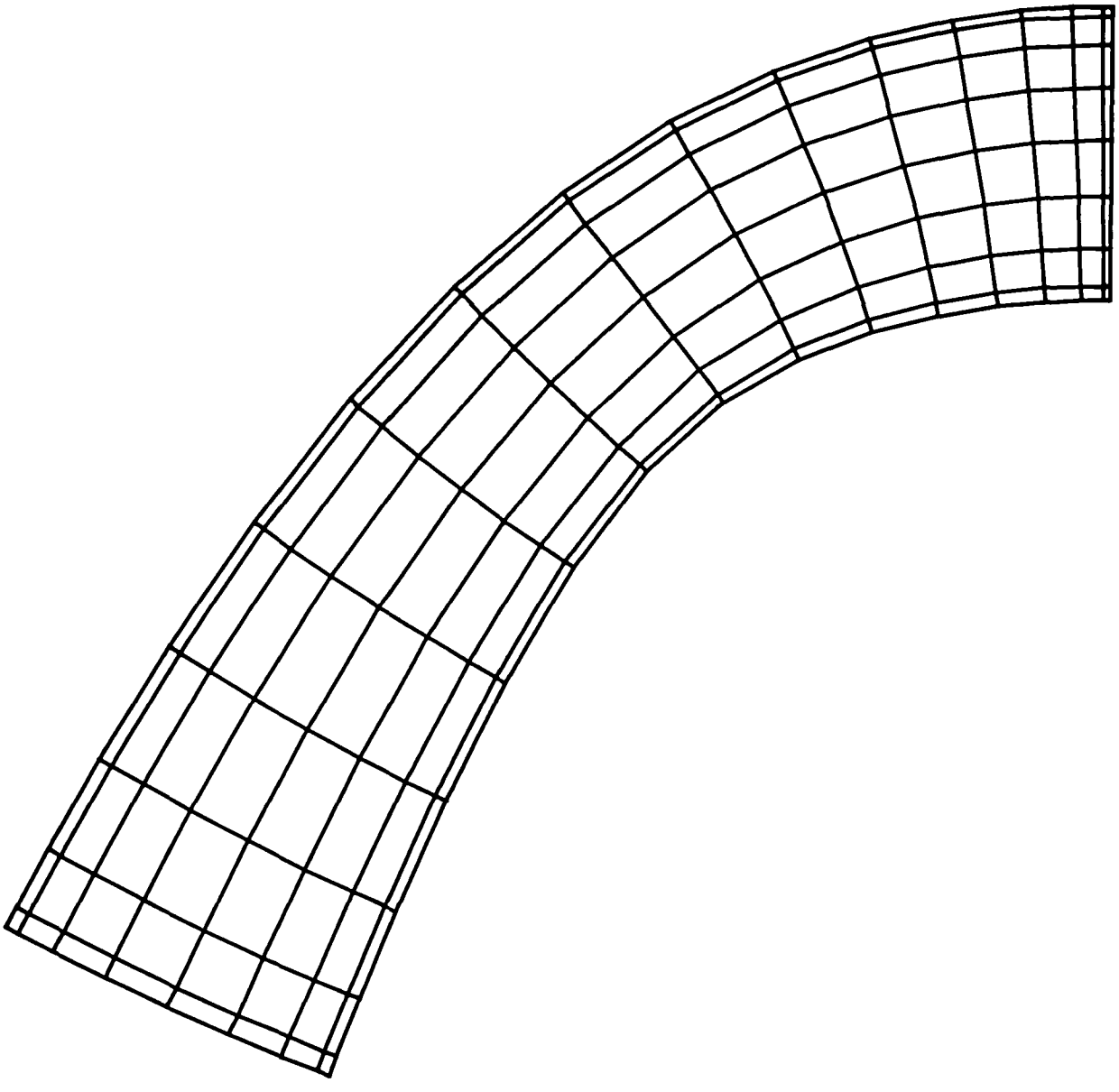


FIG. 8. Single domain Chebyshev grid for the Ringleb problem.

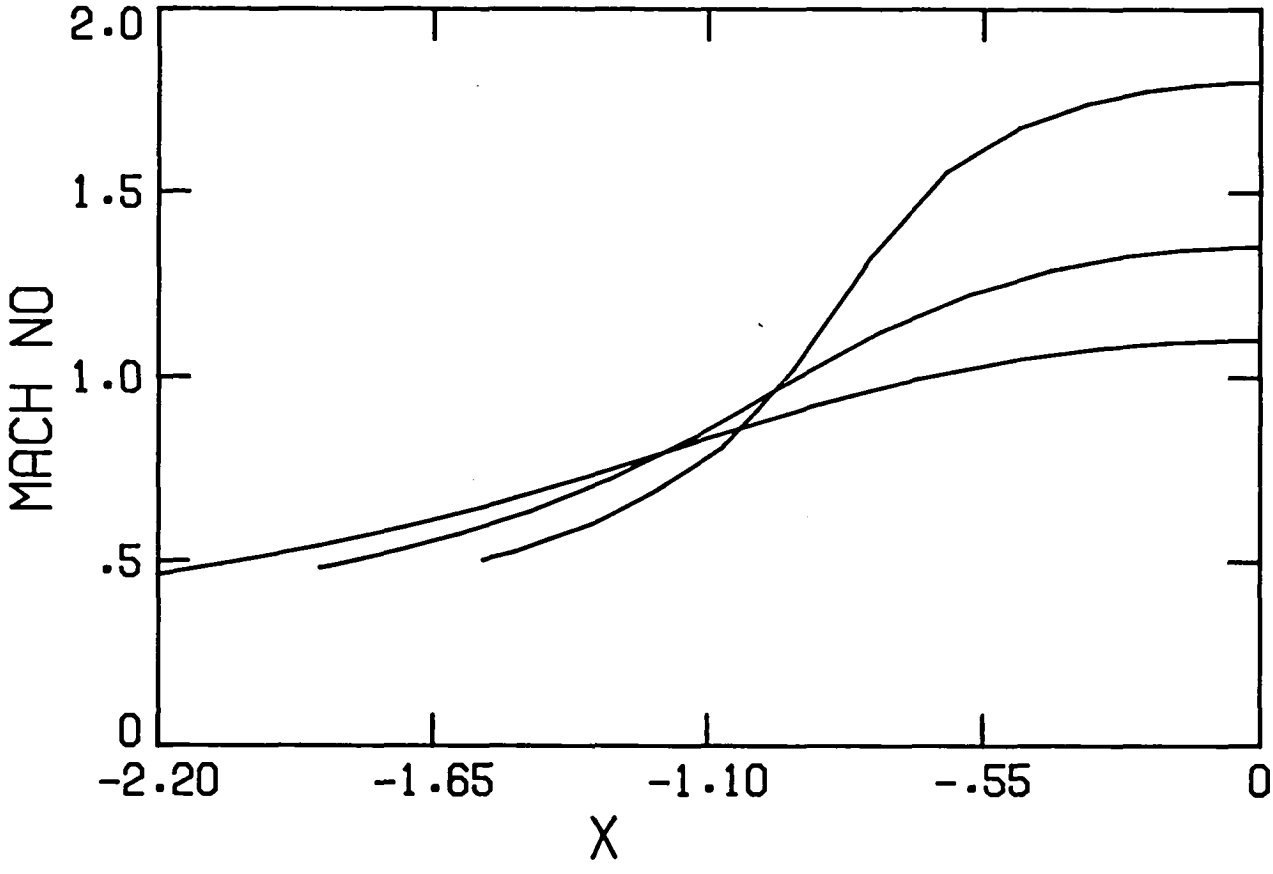


FIG. 9. Mach number variation along the lower wall, center streamline and upper wall for the Ringleb problem.

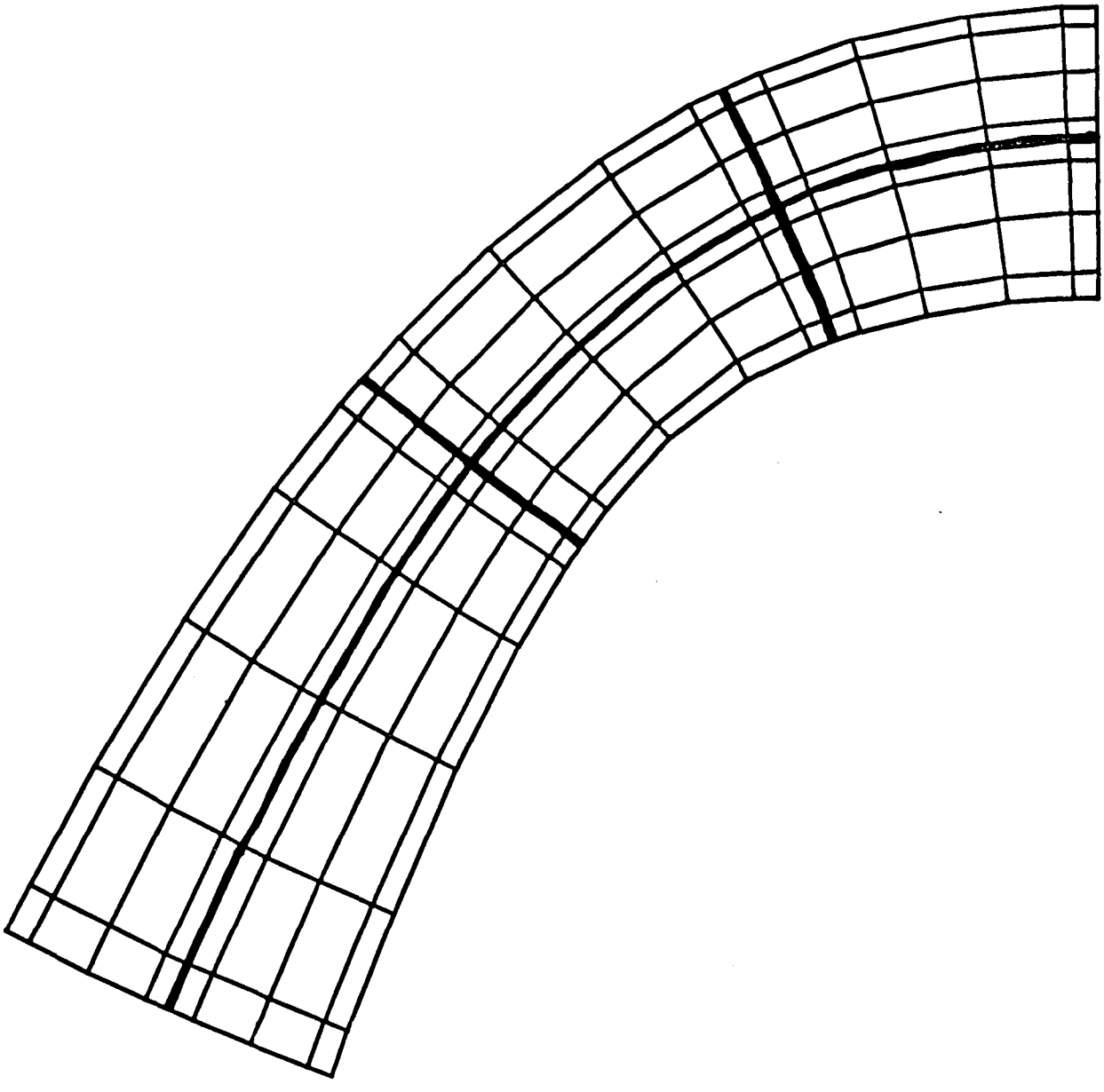


FIG. 10. Multidomain grid with six subdomains for the Ringleb problem.

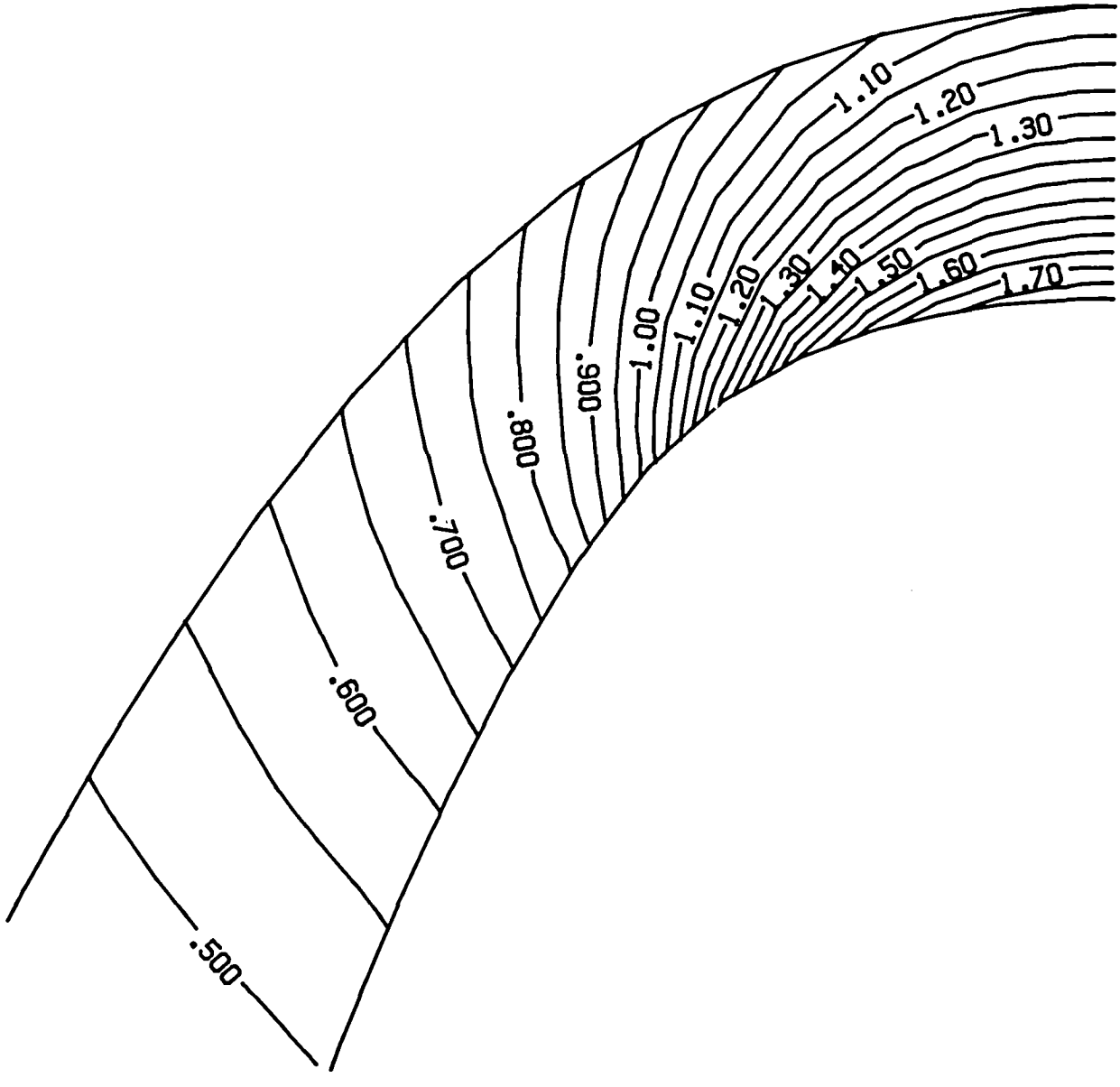


FIG. 11a. Mach number contours for single domain solution.

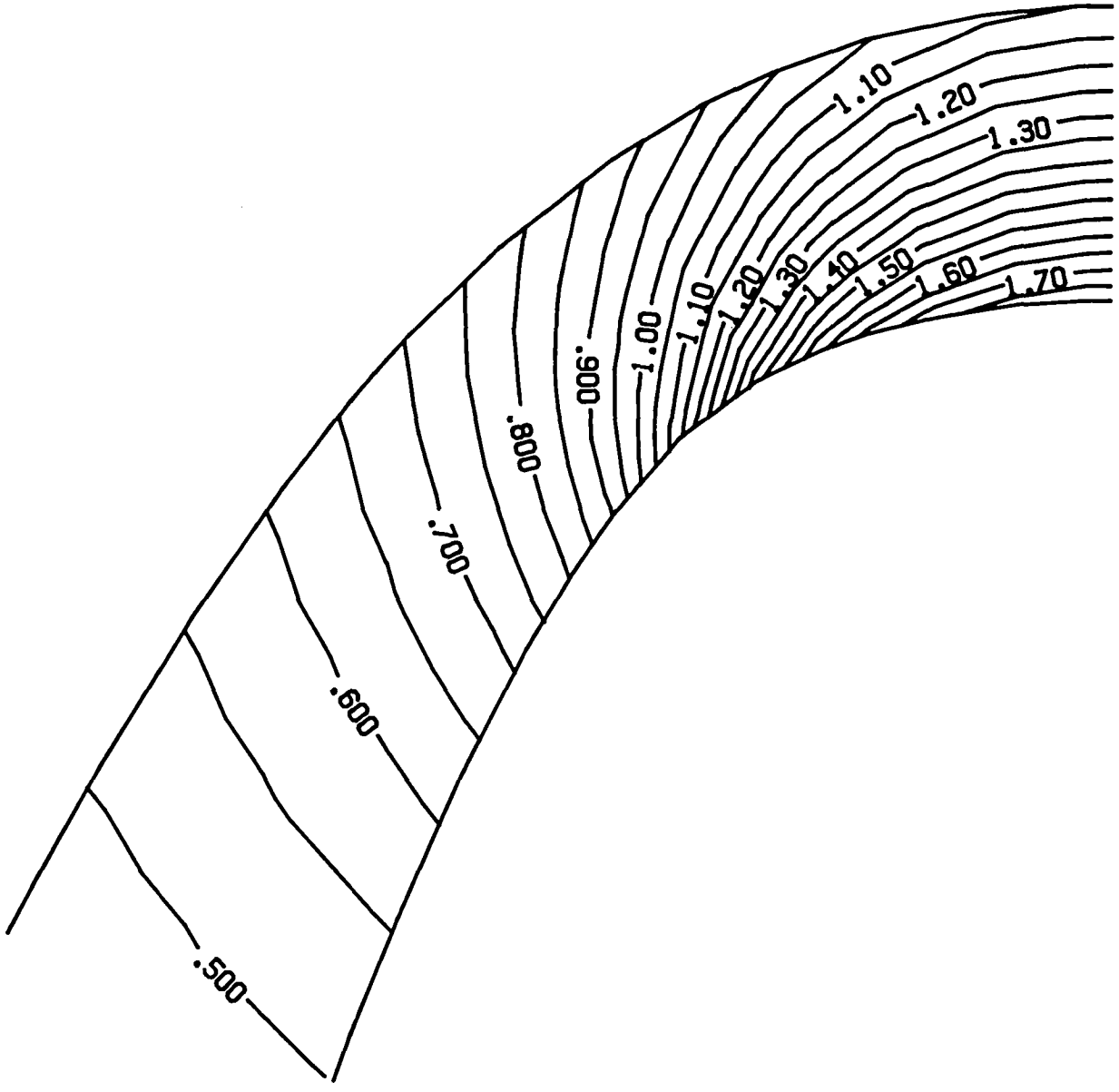


FIG. 11b. Mach number contours for six domain solution.

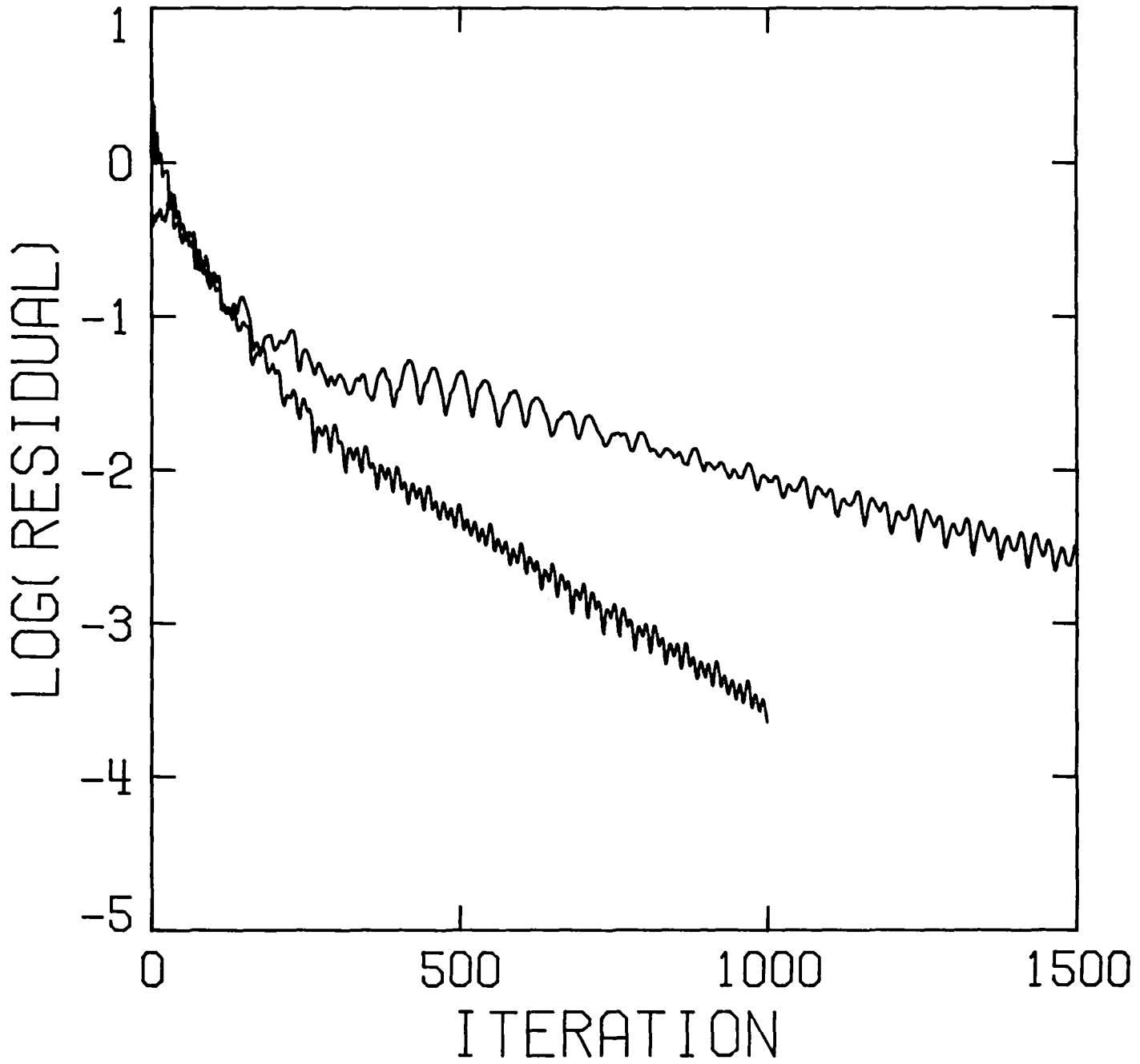


FIG. 12. Comparison of residual decay for single domain and multidomain solutions to the Ringleb problem.

ON SUBSTRUCTURING ALGORITHMS AND SOLUTION TECHNIQUES FOR THE NUMERICAL APPROXIMATION OF PARTIAL DIFFERENTIAL EQUATIONS

M. D. Gunzburger
Carnegie-Mellon University

R. A. Nicolaides
Carnegie-Mellon University

ABSTRACT

Substructuring methods are in common use in structural mechanics problems where typically the associated linear systems of algebraic equations are positive definite. Here these methods are extended to problems which lead to nonpositive definite, nonsymmetric matrices. The extension is based on an algorithm which carries out the block Gauss elimination procedure without the need for interchanges even when a pivot matrix is singular. Examples are provided wherein the method is used in connection with finite element solutions of the stationary Stokes equations and the Helmholtz equation, and dual methods for second-order elliptic equations.

Support for the first author was provided by the Air Force Office of Scientific Research under Grant No. AFOSR-83-0101 and by the Army Research Office under Contract No. DAAG-29-83-4-0084. The second author was supported under AFOSR Grant No. AFOSR-83-0231. Additional support was provided by the National Aeronautics and Space Administration under NASA Contract No. NAS1-18107.

1. THE SUBSTRUCTURING ALGORITHM IN THE POSITIVE DEFINITE CASE

The use of substructuring techniques in the numerical solution of problems governed by positive definite partial differential equations is in widespread use. The most notable case is found in structural mechanics, especially in connection with the equations of linear elasticity. For the sake of simplicity, here we describe the technique for the Dirichlet problem for the Poisson equation. Specifically, suppose we want to solve

$$\begin{aligned} -\Delta u &= f \quad \text{in} \quad \Omega \\ u &= 0 \quad \text{on} \quad \partial\Omega \end{aligned} \tag{1}$$

where Ω is, say, an open bounded region in $\mathbb{R}^2$ with boundary $\partial\Omega$. We subdivide the region Ω into open subregions Ω_i, $i = 1,\cdots,m$, such that $\overline{\Omega} = \bigcup_{i=1}^{m} \overline{\Omega}_i$ and $\Omega_i \cap \Omega_j = 0$ for $i \neq j$. We denote by Γ_{ij}, $1 \leq i < j \leq m$ the interfaces between regions Ω_i and Ω_j, i.e., $\Gamma_{ij} = \overline{\Omega}_i \cap \overline{\Omega}_j$. Of course, for particular choices of i and j in a given subdivision, Γ_{ij} may be empty. A sketch of a particular example with $m = 5$ is given in Figure 1.

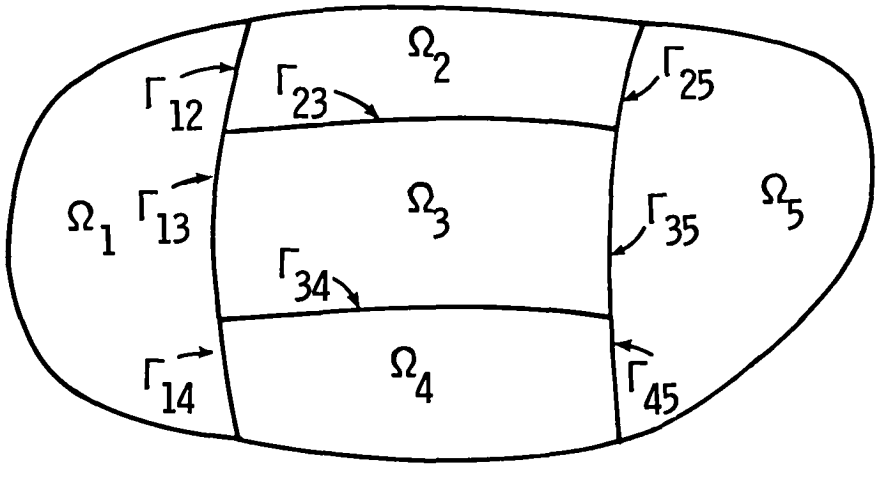


Figure 1. A subdivision of a region into five subregions.

We also subdivide Ω into a finite difference or finite element grid which in practice is much finer than the above subdivision of Ω into m subregions. We choose the two subdivisions so that the interfaces Γ_{ij} coincide with edges of the finite difference or finite element cells. The discretization of (1) proceeds in the usual manner. The essence of the substructuring algorithm is found in the particular choice for the ordering of the unknowns and equations, i.e., columns and rows, in the linear system resulting from the discretization of (1). Specifically, all unknowns and equations associated with the interior of a substructure Ω_i are numbered sequentially, one substructure at a time, and unknowns and equations associated with the interfaces Γ_{ij} are grouped together and numbered last. For example, in a typical finite difference discretization of (1), one associates equations and unknowns with nodes in the grid. In this case, we would group together all the unknowns in subregion Ω_1 together and number them first, then proceed to Ω_2, etc., and finally to Ω_m. Then we would number all the unknowns along the interfaces Γ_{ij}, $1 \leq i, j \leq m$. The equations would be numbered in the same way.[1] Likewise, in a finite element discretization of (1), some unknowns (trial functions) and equations (test functions) are associated with nodes or edges and these are

[1]The subdivision and numbering method described here applies to difference methods with stencils involving only nearest neighbors. The method may be extended in an obvious manner, e.g., by defining the interfaces to be more than one grid point in thickness, to methods having stencils with a greater degree of connectivity.

numbered in the same manner as in the finite difference case above.[2] In addition there may be test and trial functions more naturally associated with the finite elements themselves, and the equations and unknowns associated with these functions are grouped together with the other ones associated with the interior of the corresponding subregion Ω_i.

The net result of the above numbering schemes is that the linear system resulting from the discretization of (1) has the form

$$\begin{pmatrix} A_1 & & & & B_1 \\ & A_2 & & & B_2 \\ & & \ddots & & \vdots \\ & & & A_m & B_m \\ C_1 & C_2 & \cdots & C_m & A_0 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_m \\ U_0 \end{pmatrix} = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \\ F_0 \end{pmatrix}. \qquad (2)$$

In (2), the matrices A_i, $i = 1,\cdots,m$, in the finite element case, result in the case of both the test and trial functions being associated with the interior of the subregion Ω_i, $i = 1,\cdots,m$, respectively, while the matrix A_0 results from test and trial functions associated with the interfaces Γ_{ij}, $1 \leq i < j \leq m$. The matrices C_i, and B_i, represent trial, respectively test,

---

[2]Again, the method described here applies to the case where the test and trial functions vanish outside the elements which contain the associated node or edge. However, by defining the interfaces to be one or more elements thick, the method may be easily extended to other cases, e.g., cubic B-spline test and trial functions.

functions associated with the interior of Ω_i and test, respectively trial, functions associated with the interfaces. The vectors U_i, $i = 1,\cdots,m$, respectively denote the unknowns associated with the interior of Ω_i, $i = 1,\cdots,m$, while U_0 denotes the unknowns associated with the interfaces. All of these associations can also be made in the finite difference case.

It is well-known that the coefficient matrix of the linear system (2), resulting from a discretization of (1), is symmetric and positive definite. Indeed, $A_i = A_i^T$, $B_i = C_i^T$ for $i = 1,\cdots,m$ and $A_0 = A_0^T$. It is also easy to see that the matrices A_i, $i = 1,\cdots,m$, are themselves positive definite. In fact, these matrices are exactly the ones which would result from the analogous discretization of the problems

$$\begin{aligned} \Delta u &= f \quad \text{in} \quad \Omega_i \\ u &= 0 \quad \text{on} \quad \partial\Omega_i \end{aligned} \tag{3}$$

for $i = 1,\cdots,m$, where $\partial\Omega_i$ denotes the boundary of Ω_i. Note that this boundary may consist of both interfaces and a portion of the boundary $\partial\Omega$ of Ω, as is the case for Ω_1, Ω_2, Ω_4, and Ω_5 in Figure 1, or may consist wholly of interfaces as is the case for Ω_3 in that figure. Discretization of (3) results in a linear system with a coefficient matrix A_i, and thus A_i is clearly symmetric and positive definite. We note that even in the case of the Neumann problem, i.e., the boundary condition in (1) is replaced by $\partial u/\partial n = 0$ on $\partial\Omega$, the matrices A_i in (2) would still be, at least in the finite element case, symmetric and positive definite.[3] This is so because the problem (3) associated with the matrix A_i is now given by

$$\Delta u = f \quad \text{in} \quad \Omega_i$$

$$\frac{\partial u}{\partial n} = 0 \quad \text{on} \quad \partial\Omega_i \cap \partial\Omega \tag{4}$$

$$u = 0 \quad \text{on} \quad \partial\Omega_i \cap \Gamma_{ij}, \; j = 1,\cdots,m,$$

where we have set $\Gamma_{ij} = \Gamma_{ji}$. Since $\partial\Omega_i \cap \Gamma_{ij}$ is never empty, the matrix A_i associated with (4) is symmetric and positive definite.[4]

With the matrices A_i, $i = 1,\cdots,m$, being positive definite, one may proceed to solve (2) by a block elimination procedure. Symbolically, we may express the first m stages of this procedure by the relations

$$U_i = A_i^{-1}(F_i - B_i U_0), \qquad i = 1,\cdots,m, \tag{5}$$

which uniquely express U_i in terms of data and the interface unknowns U_0. The last stage of the process requires the solution of the linear system

$$DU_0 = G \tag{6}$$

where

$$D = A_0 - \sum_{i=1}^{m} C_i A_i^{-1} B_i \quad \text{and} \quad G = F_0 - \sum_{i=1}^{m} C_i A_i^{-1} F_i. \tag{7}$$

---

[3] If on $\partial\Omega_i \cap \partial\Omega$ something other than Dirichlet data is specified, then the matrix A_i also contains rows and columns associated with test and trial functions associated with nodes or edges on that portion of the boundary.

[4] Of course, the fact that A_i, $i = 0,\cdots,m$, are positive definite may be deduced directly from the fact the coefficient matrix of (2) is positive definite, i.e., the former is a necessary condition for the latter.

Of course, in (5) and (7) the inverses are not explicitly computed, but rather appropriate linear systems are solved. The solvability of the system (6) follows whenever the system (2) is solvable. In fact, if the system (2) is positive definite, so is the matrix D [1]. Once (6) is solved for U_0, (5) yields U_i, $i = 1,\cdots,m$.

Although we have described the substructuring algorithm in the context of the Poisson equation, the method can be applied in a similar manner to any positive definite problem. As noted above, the method has encountered great success in structural mechanics problems. However, in other fields where the governing equations are not positive definite or symmetric one may still order the equations and unknowns to produce linear systems such as (2), but these may not always be solved by a standard block elimination procedure. In the next two sections we describe a procedure to solve (2) even in the case of the matrices A_i being singular and show how the method may be implemented through an elimination procedure. In Section 4 we describe examples which lead to singular matrices A_i. Finally, in Section 5 we give some concluding remarks.

Incidentally, in almost all situations the use of a properly implemented substructuring algorithm will result in savings in computational costs when compared to a banded elimination procedure. For example, consider a discretization of Poisson's equation on a unit square. Suppose we have M subregions in each direction so that $m = M^2$ and suppose that each subregion is further subdivided by introducing an $n \times n$ grid. Thus, there are a total of Mm points in each direction. Banded elimination requires $O(M^4 n^4)$ operations, while the above substructuring algorithm can be implemented in, at most, $O(Mn^4 + M^4 n^3)$ operations. We note that this particular problem is not

particularly well-suited for substructuring methods. Also, the relative advantage of substructuring is greater when one considers three-dimensional problems or systems of partial differential equations.

We also note that substructuring ideas in connection with preconditioning techniques have been discussed in [2].

2. THE SOLUTION ALGORITHM IN THE GENERAL CASE

We begin by describing a method for solving (2) in the case where the matrices A_i are singular. The algorithm described here is a special case of a more general algorithm which applies to arbitrary matrices with arbitrary subdivisions into blocks, e.g., the matrix has no special structure and the matrices A_i may not only be singular, but may even be rectangular. The more general algorithm is described in [3]. We will describe the algorithm as applied to (2) and we will make use of pseudo-inverses in order to simplify the initial presentation. However, we emphasize that the algorithm may be implemented without the need for the explicit calculation of any pseudo-inverses; such an implementation is discussed in the next section. This is similar to the observation that the algorithm contained in (5)-(7) may be implemented without explicitly computing any inverses, e.g., by solving linear systems.

The system (2) is equivalent to

$$A_i U_i + B_i U_0 = F_i. \qquad i = 1,\cdots,m, \tag{8}$$

$$\sum_{i=1}^{m} C_i U_i + A_0 U_0 = F_0. \tag{9}$$

Now, U_i may be orthogonally decomposed in the form

$$U_i = Y_i + Z_i, \qquad i = 1,\cdots,m, \tag{10}$$

where

$$A_i Z_i = 0, \qquad i = 1,\cdots,m, \tag{11}$$

and Y_i is orthogonal to all vectors satisfying (11). In particular,

$$Y_i^T Z_i = 0, \qquad i = 1,\cdots,m. \tag{12}$$

Substitution of (10)-(11) into (8) yields that

$$A_i Y_i = F_i - B_i U_0, \qquad i = 1,\cdots,m. \tag{13}$$

Since Y_i is orthogonal to the null space of A_i, (13) yields that

$$Y_i = A_i^+(F_i - B_i U_0), \qquad i = 1,\cdots,m, \tag{14}$$

where A_i^+ denotes the pseudo-inverse of A_i. This relation states that Y_i is uniquely determined from the data and U_0. Note that (8) yields no information concerning Z_i as is to be expected since $A_i Z_i = 0$. Substituting (10) and (14) into (9) yields that

$$DU_0 = G - \sum_{i=1}^{m} C_i Z_i \tag{15}$$

where

$$D = A_0 - \sum_{i=1}^{m} C_i A_i^+ B_i \quad \text{and} \quad G = F_0 - \sum_{i=1}^{m} C_i A_i^+ F_i. \tag{16}$$

We may also decompose U_0 in the form

$$U_0 = Y_0 + Z_0 \tag{17}$$

where

$$DZ_0 = 0 \tag{18}$$

and Y_0 is orthogonal to all vectors satisfying (18). In particular,

$$Y_0^T Z_0 = 0. \tag{19}$$

Substitution of (17)-(18) into (15) yields that

$$DY_0 = G - \sum_{i=1}^{m} C_i Z_i \tag{20}$$

and, since Y_0 is orthogonal to the null space of D, (20) yields that

$$Y_0 = D^+(G - \sum_{i=1}^{m} C_i Z_i). \tag{21}$$

Again, it is not surprising that (15) yields no information concerning Z_0. Substitution of (17) and (21) into (14) then yields that

$$Y_i = A_i^+[F_i - B_i D^+(G - \sum_{j=1}^{m} C_j Z_j)] - A_i^+ B_i Z_0 \tag{22}$$

for $i = 1,\cdots,m$.

At this point we have shown that Y_i, $i = 0,\cdots,m$, may be uniquely expressed in terms of Z_i, $i = 0,\cdots,m$, by (21) and (22). It remains to show how to find the latter. The first step is to multiply (13) by $(I - A_i A_i^+)$. Since $A_i A_i^+ A_i = A_i$, we have that

$$(I - A_i A_i^+)(F_i - B_i U_0) = 0, \qquad i = 1,\cdots,m,$$

or substituting (17) and (21),

$$(I - A_i A_i^+)[F_i - B_i Z_0 - B_i D^+(G - \sum_{j=1}^{m} C_j Z_j)] = 0, \qquad i = 1,\cdots,m. \tag{23}$$

Now suppose we are able to determine bases for the null spaces of A_i, $i = 1,\cdots,m$, and D. We collect each of these basis sets into matrices N_i, $i = 0,\cdots,m$, i.e., N_i, $i = 0,\cdots,m$, have linearly independent columns,

$$DN_0 = 0 \quad \text{and} \quad A_i N_i = 0, \qquad i = 1,\cdots,m, \tag{24}$$

and the columns of N_0, respectively N_i, span the null space of D, respectively A_i, $i = 1,\cdots,m$. The number of columns in N_i is, of course, the dimension of the corresponding null spaces. Now, we may write that

$$Z_i = N_i \Lambda_i, \qquad i = 0,\cdots,m, \tag{25}$$

for some vectors Λ_i. Substituting (25) into (23) then yields that

$$\sum_{j=1}^{m} R_{ij} \Lambda_j = H_i, \qquad i = 1,\cdots,m, \tag{26}$$

where

$$R_{i0} = (I - A_i A_i^+)B_i N_0, \qquad H_i = (I - A_i A_i^+)[F_i - B_i D^+ G]$$

and (27)

$$R_{ij} = (I - A_i A_i^+)B_i D^+ C_j N_j, \qquad j = 1,\cdots,m.$$

Now letting

$$R = \begin{pmatrix} R_{10} & R_{11} \cdots R_{1m} \\ R_{20} & R_{21} \cdots R_{2m} \\ \vdots & \vdots \quad\quad \vdots \\ R_{m0} & R_{m1} \quad\quad R_{mm} \end{pmatrix}, \quad H = \begin{pmatrix} H_1 \\ H_2 \\ \vdots \\ H_m \end{pmatrix} \text{ and } \Lambda = \begin{pmatrix} \Lambda_0 \\ \Lambda_1 \\ \vdots \\ \Lambda_m \end{pmatrix}, \tag{28}$$

(26) may be expressed in the form

$$R\Lambda = H. \tag{29}$$

In general, R is a rectangular matrix. The number of rows in R is equal to the sum of the number of rows of the matrices A_i, $i = 1,\cdots,m$, and the number of columns of R is equal to the sum of the dimensions of the null spaces of A_i, $i = 1,\cdots,m$; and D. It can be shown [3] that the system (29) is a consistent system, and we may find its solution, for example, by forming

$$(R^T R)\Lambda = R^T H. \tag{30}$$

Suppose we can solve (30) for Λ. Then (28) yields Λ_i, $i = 0,\cdots,m$, (25) then yields Z_i, $i = 0,\cdots,m$, (21) and (22) yields Y_i, $i = 0,\cdots,m$, and finally (10) and (17) yield the solution U_i, $i = 0,\cdots,m$, of (2).

The algorithm described here is related in the following manner to the block elimination algorithm in Section 1. Suppose that the matrix of (2) and all the A_i's and D are nonsingular. Then, the algorithm of this section reduces to the standard block Gauss elimination procedure. Indeed, in this case, $A_i^+ = A_i^{-1}$, $D^+ = D^{-1}$ and $Z_i = 0$ so that $U_i = Y_i$ and the latter are determined uniquely by (14) and (21). Note the correspondence, in this case, between (14)-(15) and (5)-(6).

In the more general case, i.e., some or all of the A_i's and D being singular, it can be shown [3] that the rank deficiency of (30) is exactly that of the original coefficient matrix in (2). Therefore, if the latter is nonsingular, then so is R^TR and then Λ in (30) is uniquely determined. Since the Z_i's and Y_i's are uniquely determined from Λ, the algorithm produces the unique solution of (2). If the matrix of (2) is singular, so is R^TR and (30) does not have a unique solution. However, (30) may be solved anyway, either in terms of arbitrary parameters or by adding constraints. The number of parameters or constraints is equal to the dimension of the null space of R^TR which in turn is the same as the dimension of the null space of the coefficient matrix in (2). In any case, once a particular Λ is determined, then Z_i and Y_i are also determined.

In particular applications to the solution of partial differential equations, the dimension of the system (30) is small compared to that of the system (2). Indeed, typically $\dim(R^TR) = O(m)$, the number of subregions. For example, the dimension of the null spaces of the matrices A_i and D may be one or zero, in which case $\dim(R^TR) \leq m + 1$.

3. AN ELIMINATION IMPLEMENTATION

We begin by restating the algorithm of the previous section. Given the matrices $A_0,\cdots,A_m$, $B_1,\cdots,B_m$, $C_1,\cdots,C_m$ and the vectors $F_0,\cdots,F_m$, we find vectors $U_0,\cdots,U_m$ satisfying (2) by the following procedure.

1. Compute $A_i^+ F_i$ and $A_i^+ B_i$ for $i = 1,\cdots,m$.
2. Compute N_i, $i = 1,\cdots,m$, whose columns constitute a basis for the null space of A_i, $i = 1,\cdots,m$, respectively.
3. Compute $C_i(A_i^+ B_i)$, $C_i(A_i^+ F_i)$ and $C_i N_i$ for $i = 1,\cdots,m$.
4. Compute $D = A_0 - \sum_{i=1}^{m} C_i(A_i^+ B_i)$ and $G = F_0 - \sum_{i=1}^{m} C_i(A_i^+ F_i)$.
5. Compute $D^+ G$.
6. Compute N_0 whose columns constitute a basis for the null space of D.
7. Compute $D^+ C_i N_i$ for $i = 1,\cdots,m$.
8. Compute the matrices

$$R_{i0} = B_i N_0 - A_i (A_i^+ B_i)N_0 \quad \text{for } i = 1,\cdots,m,$$

$$R_{ij} = B_i(D^+ C_j N_j) - A_i(A_i^+ B_i)(D^+ C_j N_j) \quad \text{for } i,j = 1,\cdots,m$$

and the vectors

$$H_i = F_i - B_i(D^+ G) - A_i(A_i^+ F) + A_i(A_i^+ B_i)(D^+ G) \quad \text{for } i = 1,\cdots,m.$$

9. Assemble the results of step 6 into the matrix R and vector H according to (28) and then compute $R^T R$ and $R^T H$.
10. Solve the linear system $R^T R\Lambda = R^T H$ for Λ and then compute Λ_i, $i = 0,\cdots,m$, according to the partition of (28).
11. Compute $Z_i = N_i \Lambda_i$ for $i = 0,\cdots,m$.

12. Compute $Y_0 = (D^+ G) - \sum_{i=1}^{m} C_i N_i \Lambda_i = (D^+ G) - \sum_{i=1}^{m} C_i Z_i$.

13. Compute $U_0 = Y_0 + Z_0$.

14. Compute $Y_i = (A_i^+ F_i) - (A_i^+ B_i)U_0$ for $i = 1,\cdots,m$.

15. Compute $U_i = Y_i + Z_i$ for $i = 1,\cdots,m$.

Other than steps 1, 2, 5, 6, and 10, the above algorithm requries only matrix and matrix-vector multiplications. In this section we show how to carry out the other operations required by the algorithm through an elimination procedure. In particular, we will not need to explicitly calculate any pseudo-inverses of matrices.

We first describe how to carry out steps 1 and 2. Consider the linear system.

$$A_i Q = S = (B_i, F_i, 0) \tag{31}$$

where the right-hand side matrix S consists of the matrix B_i, the vector F_i, and some additional columns of zeroes. The number of these additional columns should be greater or equal to the dimension of the null space of A_i. This dimension will actually be determined during the elimination procedure.[5] We now proceed to solve (31) by Gauss elimination with partial pivoting. If the matrix A_i is singular, then one or more times during the elimination procedure we will not be able to locate a nonzero pivot element. In fact, the number of times this occurs is exactly the dimension of the null space of A_i. However, at such an occurrence, the corresponding column is

[5]See Section 5 concerning the effects that roundoff errors may have on the determination of this dimension.

already in the eliminated form so that we may skip over to the next column and continue the elimination process. At the end of the process, (31) has been reduced to the form

$$\overline{A}_i \, Q = J = (\overline{B}_i, \overline{F}_i, 0) \tag{32}$$

where $\overline{A}_i$ is upper triangular and in row echelon form. When A_i is singular, $\overline{A}_i$ will have zeros at the pivot location for exactly those columns for which no nonvanishing pivot element was found.

We now proceed to backsolve (32). No difficulty is encountered until a row is reached for which the pivot entry of $\overline{A}_i$ is zero. For the columns of Q corresponding to B_i and F_i, we may arbitrarily set (to something other than zero) the entry in the row corresponding to the zero pivot of $\overline{A}_i$. Then the backsolve procedure may continue until we reach another zero pivot entry, at which time we again arbitrarily specify an entry in the columns of Q corresponding to the columns B_i and F_i of S. While all this is going on we are also solving (32) for the columns corresponding to the zero columns of S. For these columns, whenever a zero pivot entry is encountered in $\overline{A}_i$, one of the elements in the corresponding row is set to one while the rest are set to zero. Each time a zero pivot entry is encountered, a different column is chosen for which one sets the arbitrary element to one. At the end of this backsolve procedure, (32) yields that

$$Q = (\hat{L}, \hat{K}, N_i).$$

Here the columns of N_i form a basis for the null space of A_i and $\hat{L}$ and $\hat{K}$ are particular solutions of the systems.

$$A_i L = B_i \quad \text{and} \quad A_i K = F_i. \tag{33}$$

The final step is to orthogonalize the columns of $\hat{L}$ and $\hat{K}$ with respect to the columns of N_i to yield

$$\tilde{Q} = (\tilde{L},\tilde{K},N_i).$$

Since $A_iN_i = 0$, $\tilde{L}$ and $\tilde{K}$ are still solutions of (33). Moreover, the columns of $\tilde{L}$ and $\tilde{K}$ are orthogonal to the null space of A_i and, therefore, are minimum norm solutions. By the uniqueness of the minimum norm solution, we have that

$$\tilde{L} = A_i^+ B_i \quad \text{and} \quad \tilde{K} = A_i^+ F_i.$$

Thus the above elimination procedure has accomplished the tasks of steps 1 and 2 of the algorithm.

The tasks of steps 5, 6, and 7 can be accomplished in an analogous manner. Also, if the matrix R^TR is nonsingular, then it may be easily solved by an ordinary Gauss elimination procedure. If it is singular then a solution in terms of arbitrary parameters may be determined in a manner similar to the above procedure for the system (31). We note that any sparsity or structure inherent in the matrices A_i may be exploited in the above procedure. However, in general, the matrix D will be dense. We will return to this point in the concluding section.

4. EXAMPLES

The Stationary Stokes Equation

Consider the stationary Stokes equations for the slow flow of a viscous fluid in a bounded region in $\mathbb{R}^2$. These are given by

$$\begin{aligned} \Delta \underline{u} - \text{grad } p &= \underline{f} \quad \text{in} \quad \Omega \\ \text{div } \underline{u} &= 0 \quad \text{in} \quad \Omega \\ \underline{u} &= 0 \quad \text{on} \quad \partial\Omega. \end{aligned} \tag{34}$$

Here $\underline{u}$ denotes the velocity, p the pressure, $\underline{f}$ the given body force and the viscosity coefficient has been absorbed into p and $\underline{f}$. Clearly, the pressure cannot be determined uniquely since we may add an arbitrary constant to the pressure and still satisfy (34).

A finite element approximation of the solution $(\underline{u},p)$ of (34) may be defined as follows. Given finite-dimensional spaces V^h and S^h for the discrete velocity and pressure fields, we seek $\underline{u}^h \in V^h$ and $p^h \in S^h$ such that

$$\begin{aligned} &\int_\Omega (\text{grad } \underline{u}^h : \text{grad } \underline{v}^h - p^h \text{ div } \underline{v}^h) d\Omega = -\int_\Omega \underline{f} \cdot \underline{v}^h \, d\Omega \quad \text{for all } \underline{v}^h \in V^h \\ &\int_\Omega q^h \text{ div } \underline{u}^h \, d\Omega = 0 \quad \text{for all } q^h \in S^h. \end{aligned} \tag{35}$$

Here we assume that the elements of V^h satisfy the boundary condition in (34). By choosing bases for the spaces V^h and S^h, (35) can be expressed as a linear algebraic system for the coefficients in the basis function expansions of $\underline{u}^h$ and p^h.

Now it is well-known that arbitrary choices of spaces V^h and S^h may not yield stable or accurate solutions. However, there are now known many element pairs for which (35) yields optimally accurate solutions [4], [5], [6]. One such pair is described as follows. Suppose S_h denotes a triangulation of the region Ω. We denote by V_h a finer triangulation derived from S_h by subdividing each triangle in S_h into four congruent triangles by joining the midsides. See Figure 2. We define S^h to consist of piecewise constant functions over the triangulation S_h

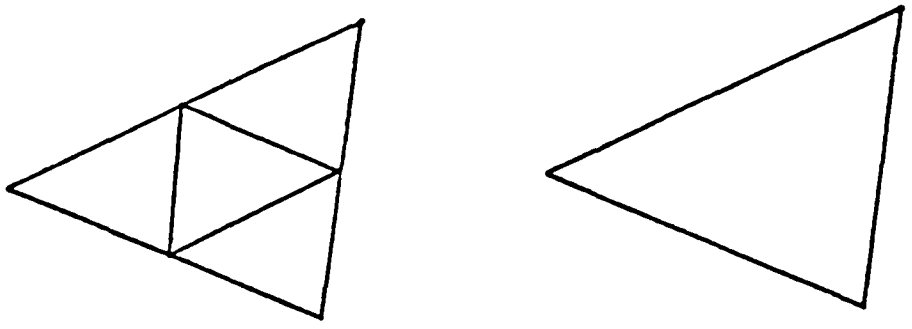


Figure 2. A triangle in S_h and the corresponding triangles in V_h.

and V^h to consist of piecewise linear functions over the triangulation V_h which are continuous over Ω and vanish on $\partial\Omega$. This combination is known to be stable and be optimally accurate [6].[6] The basis functions for V^h are easily associated with the vertices of the triangulation V_h while the basis functions for S^h are associated with the triangles in the triangulation S_h.

[6]See below for the necessary restriction on the pressure which yields this result.

Now let us consider a substructuring technique for the solution of (35). We assume that the interfaces Γ_{ij} between subregions are made up of edges of the triangulation S_h so that these interfaces do not cut across pressure triangles. One may easily arrange a numbering scheme for the unknowns and equations which yields a linear system of the form (2). For example, U_i consists of all velocity unknowns associated with vertices of V_h located in the interior of the subregion Ω_i and all pressure unknowns associated with the triangles of S_h which are also in Ω_i. Note that U_0 contains only velocity unknowns, namely those associated with vertices V_h which lie on the interfaces Γ_{ij} but not on $\partial\Omega$.

We have not constrained the pressure space and therefore the system (2) corresponding to this discretization of (34) is singular. In fact, its rank defficiency is one, and the null vector corresponds to the pressure function which is constant over Ω. On the other hand, the velocity approximation is uniquely determined by (2) [6]. Furthermore, it is easy to see that the submatrices $A_1,\cdots,A_m$ are singular. In fact, these matrices are exactly those which arise from the analogous discretization of the problem.

$$\Delta \underline{u} - \text{grad } p = \underline{f} \quad \text{in} \quad \Omega_i$$

$$\text{div } \underline{u} = f \quad \text{in} \quad \Omega_i$$

$$u = 0 \quad \text{on} \quad \partial\Omega_i.$$

Thus each of the matrices A_i has a single local pressure null vector, i.e., the dimension of N_i is one and N_i corresponds to the pressure function which is constant over Ω_i. On the other hand, since the velocity field can

be uniquely determined from (2) and since U_0 consists of only velocity unknowns, the matrix D in the linear system (15) is nonsingular, i.e., $N_0 = 0$. Thus, in this case, the system (30) has dimension m and has a one-dimensional null space, the latter following from the fact that the system (2) itself has a one-dimensional null space.

If we choose the pressure space S^h to consist of piecewise linear functions over the triangulation S_h which are continuous over Ω, while retaining the same velocity space, the situation changes drastically. For example, now the basis functions for S^h are more easily associated with the vertices of S_h. Now U_i contains pressure unknowns corresponding to vertices in S_h which are in the interior of Ω_i or lie on $\partial\Omega_i \cap \partial\Omega$. More important, U_0 now contains pressure unknowns associated with vertices of S_h which lie on Γ_{ij} but not on $\partial\Omega$. In this case the matrices A_i are nonsingular and the matrix D is singular with a one-dimensional null space.

The Helmholtz Equation

Now consider the problem

$$\begin{aligned} \Delta u + \lambda u &= f \quad \text{in} \quad \Omega \\ u &= 0 \quad \text{on} \quad \partial\Omega \end{aligned} \tag{36}$$

where λ is not near an eigenvalue of the operator $-\Delta$. Standard finite element or finite difference discretizations of (36) yield linear algebraic systems with coefficient matrices which are symmetric and indefinite, but which certainly may, by using a partial pivoting strategy, be stably inverted. Now consider the following specific situation. Let Ω be the

square $(0,\pi) \times (0,\pi)$ and let $\lambda = 13/4$. Since the eigenvalues of $-\Delta$ for this region are given by $(n^2 + m^2)$, $m,n = 1,2,\cdots$, we see that $\lambda = 13/4$ is not an eigenvalue and therefore the problem (36) leads to nonsingular coefficient matrices. Now, suppose we consider solving (36) by using the substructuring algorithm with the two subregions $\Omega_1 = (0,2\pi/3) \times (0,\pi)$ and $\Omega_2 = (2\pi/3,\pi)\ \Omega\ (0,\pi)$. Then the matrices A_i in (2) correspond to the coefficient matrix for the analogous discretization of the problem

$$\begin{aligned} \Delta u + \lambda u &= f \quad \text{in} \quad \Omega_i \\ u &= 0 \quad \text{on} \quad \partial\Omega_i. \end{aligned} \tag{37}$$

But the eigenvalues of $-\Delta$ for the region Ω_1 are given by $(n^2 + 9m^2/4)$, $m,n = 1,2,3,\cdots$, so that $\lambda = (13/4)$ is an eigenvalue of $-\Delta$ for the region Ω_1 and therefore the matrix A_1 is singular even though the system (2) is not.

Admittedly, this example is somewhat pathological in the sense that for random choices of regions, subregions, and parameters λ, the probability is zero that the matrices A_i in (2) will be singular. However, for particular choices of λ, Ω and Ω_i, one or more of the matrices A_i may be singular; after all, the above example is not really all that far-fetched. Of course, if any of the A_i's are singluar, the situation may be remedied by choosing a different subdivision of the region Ω; this in turn implies a complete reassembly of the coefficient matrix in (2). On the other hand, the algorithm of Sections 2 and 3 may be used whether or not any of the matrices A_i are singluar.

There is a small but nonvanishing probability that for some of the problems (37) λ, although not an eigenvalue of $-\Delta$ for the region Ω_i, is close to such an eigenvalue. If λ is close enough to such an eigenvalue, the matrix A_i, in finite precision arithmetic, may be mistakenly determined to be singular by the algorithm of Section 3. However, this will be the case only when the difference between λ and an eigenvalue is much smaller than the discretization error, i.e., of the order of the unit roundoff error of the machine, and no serious effect on the accuracy of the solution should result.

Dual Methods for Second-Order Elliptic Equations

For a third example, we consider dual methods for second-order elliptic partial differential equations. An example of these are methods based on the complementary energy principle in linear elasticity. For simplicity, we here consider the problem

$$\begin{aligned} u &= \nabla\phi \quad \text{in} \quad \Omega \\ \operatorname{div} \underline{u} &= f \quad \text{in} \quad \Omega \\ \underline{u}\cdot\underline{n} &= 0 \quad \text{on} \quad \Gamma_1 \end{aligned} \tag{38}$$

and

$$\phi = g \quad \text{on} \quad \Gamma_2$$

where again $\Gamma_1 \cap \Gamma_2 = \partial\Omega$ denotes the boundary of the bounded region $\Omega \subset \mathbb{R}^2$ and n denotes the unit outer normal to $\partial\Omega$. A finite element approximation of (38) may be obtained by choosing finite-dimensional spaces V^h and S^h and then seeking $\underline{u}^h \in V^h$ and $\phi^h \in S^h$ such that

$$\int_{\Omega} (\underline{u}^h \cdot \underline{v}^h + \phi^h \operatorname{div} \underline{v}^h) d\Omega = \int_{\Gamma_2} g \underline{v}^h \cdot \underline{n} d\Omega \quad \forall \underline{v}^h \in V^h$$

$$\int_{\Omega} \psi^h \operatorname{div} \underline{u}^h \, d\Omega = \int_{\Omega} f \psi^h \quad \forall \psi^h \in S^h.$$

We assume that the elements of V^h satisfy the boundary condition on Γ_1 in (38). The boundary condition on ϕ is natural in this formulation, which is one of its advantages.

In [7], the following choice of V^h and S^h was shown to yield stable and optimally accurate approximations, at least for polygonal domains. First, we subdivide Ω into quadrilaterals, and then subdivide each quadrilateral into four triangles by drawing the diagonals. For V^h we take all continuous piecewise linear vector fields with respect to the resulting triangulation and then define $S^h = \operatorname{div} V^h$. The resulting space S^h can be shown to be a subspace of all piecewise constants over the triangulation. See [7] for details.

In the implementation of the substructuring algorithm, we assume that the interfaces Γ_{ij} coincide with some of the edges of the quadrilaterals which initially defined our finite element triangulation of Ω, i.e., the interfaces do not cut through any of these quadrilaterals. The test and trial functions from V^h are associated with nodes while those from S^h are associated with the interior of the quadrilaterals. The matrices A_i in (2) now correspond to the discretization of the problem

$$\underline{u} = \nabla\phi \quad \text{and} \quad \operatorname{div} \underline{u} = f \quad \text{in} \quad \Omega_i$$

$$\underline{u} \cdot \underline{n} = 0 \quad \text{on} \quad \Gamma_1 \cap \partial\Omega_i, \; \phi = g \quad \text{on} \quad \Gamma_2 \cap \partial\Omega_i \tag{39}$$

and

$$\underline{u} = 0 \quad \text{on} \quad \Gamma_{ij} \cap \partial\Omega_i .$$

Because of the last boundary condition, the problem (39) is over constrained insofar as the variable $\underline{u}$ us concerned. Nevertheless, if $\Gamma_2 \cap \partial\Omega_i = 0$, i.e., a given subregion does not have part of its boundary coincide with that part of $\partial\Omega$ on which data for ϕ are given, then the problem (39) can only determine ϕ to an additive constant. This, for example, would be the case for subregion Ω_3 in Figure 1, i.e., an interior subregion. For such situations, i.e., $\Gamma_2 \cap \partial\Omega_i = 0$, the matrix A_i in (2) will again be singular, with a one-dimensional null space. Since (38) always uniquely determines $\underline{u}$, the matrix D of (16) will be nonsingular. The rank deficiency of the system (30) will be one or zero, depending on whether or not Γ_2 has vanishing measure, i.e., whether or not the problem (38) uniquely determines ϕ.

5. CONCLUDING REMARKS

Determination fo Zero Pivot Elements

A crucial step in the elimination algorithm presented in Section 3 is the determination of when all the elements in a column to be eliminated are already zero. This is necessary for the determination of the null spaces of the matrices A_i and D. In practice one would declare an element to vanish whenever its magnitude is less than some prescribed tolerance which should be proportional to the unit roundoff error of the machine. This naturally leaves open the possiblity of a very small but nonzero element being mistaken for a vanishing element. This situation can be avoided, at least when one is

solving partial differential equations, by first using high enough precision arithmetic, e.g., 60 or 64 bit floating point arithmetic, and by making sure that the algorithms used are stable. The former is easily arranged, while the latter points out the importance of rigorous mathematics. Indeed, if an algorithm is stable, as are the ones discussed in Section 4, and the machine precision is high enough, one should not encounter nonzero elements which are comparable in magnitude to the unit roundoff error unless the matrix in hand is singular or very nearly singular.

An alternative to the use of elimination type procedures is, of course, to employ methods based on orthogonal transformations. At the price of greater computational expense, such methods are less susceptible to ill effects due to roundoff error.

Parallelism

One of the attractions of substructuring algorithms is the obvious inherent parallelism both in the assembly and solution stages. The sets of matrices and vectors (A_i,B_i,C_i,F_i), $i = 1,\cdots,m$, can each be assembled independently. Furthermore, at least in the finite element case, we may write the matrix A_0 and the vector F_0 in the form

$$A_0 = \sum_{i=1}^{m} A_{0i}, \qquad F_0 = \sum_{i=1}^{m} F_{0i} \tag{40}$$

where the matrix A_{0i} and the vector F_{0i} represent the contribution to the matrix A_0 and vector F_0 coming from region Ω_i. Each of the sets (A_{0i}, F_{0i}), $i = 1,\cdots,m$, may be assembled in parallel. Thus, in the assembly stage, the sets $(A_i,B_i,C_i,F_i,A_{0i},F_{0i})$, $i = 1,\cdots,m$, may be assembled in parallel.

For example, each of the above sets may be assembled on separate processors, with no need for interprocessor communications. At the end of the assembly process, the concatenations of (40) must be performed. This step is not parallelizable, but represents a minor portion of the assembly process.

There is also a large degree of parallelism in the solution algorithm described at the beginning of Section 3. Steps 1, 2, and 3 are completely parallelizable, again with no interprocessor communications necessary. Furthermore, if the appropriate information can be transferred to the processors, steps 7, 8, 11, 14, and 15 and a portion of step 12 can also be computed in parallel. The only relatively major steps which are not parallelizable are steps 5 and 6.

The issue of parallelism in connection with substructuring algorithms has been studied in [8] in the context of a specific three-dimensional positive definite problem. That paper contains a discussion of operation counts which, for the most part, is also relevant in the present context.

Three-Dimensional Problems

As pointed out above, the major nonparallel steps in the computation are embodied in steps 5 and 6 in the algorithm of Section 3. Even on a serial machine these steps may be costly since, in general, they involve dense matrices. In two-dimensional problems, by keeping the number of subregions relatively small compared to the total number of elements in the triangulation, the size of these dense calculations can be kept small, i.e., the size of D can be of the order of the square root of the size of the A_i's. The latter usually are sparse, e.g., banded. A similar arrangement in three-dimensional problems would, in general, lead to a matrix D whose size

is of the order of the two-thirds power of the size of the A_i's, which may be unacceptably large. Furthermore, in steps 1 and 2 of the algorithm, the number of right-hand sides would be approximately equal to the number of columns of D and the size of the A_i's may be too large, when relatively few subregions are used. Therefore, for three-dimensional problems one must be especially careful to implement the algorithm in an efficient manner as possible.

These potential difficulties can be mitigated in a variety of ways. For example, many of the right-hand sides in the computations of step 1 of the algorithm are zero because any column of B_i which corresponds to an interface unknown which is not associated with $\partial\Omega_i$ would vanish. The corresponding row of C_i is also zero. Thus, one can avoid computations involving linear systems with zero right-hand sides and multiplications by zero vectors. The savings possible, in storage and computing time, by accounting for these features are relatively higher for three-dimensional problems.

Although, in general, the number of interface variables may be large for three-dimensional problems, in practice it is often the case that specific features of the domain Ω lead to a small number of such unknowns. For instance, in a wing-fuselage configuration, it is natural to consider the wing and fuselage to be different subregions and the interface between these two substructures is relatively small in extent. Indeed, it was exactly in this type of application that the terminology "substructuring" arose.

Finally we consider the most serious problem, namely that of the size of the matrix D. However, even here a judicious implementation can effect great savings. As a simple illustration consider the subregion structure of Figure 3 where we have now labeled the interface boundaries by Γ_i, $i = 1,\cdots,m-1$.

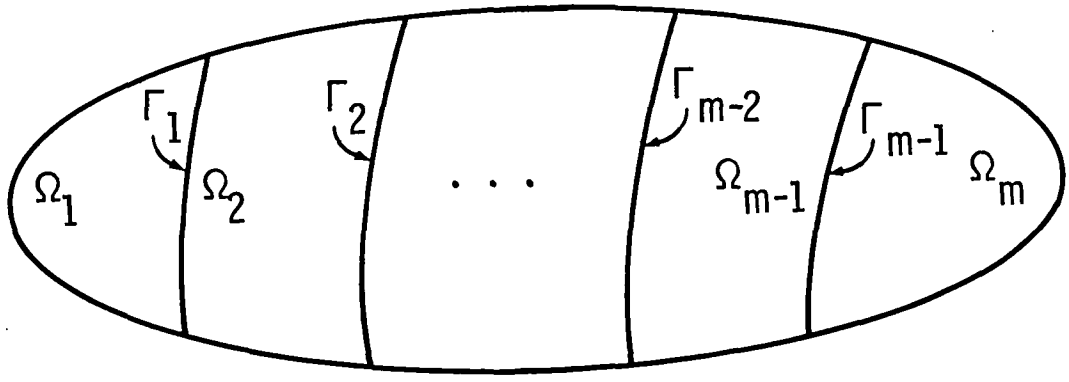


Figure 3. An example subdivision of the region Ω.

It is natural to order the interface unknowns U_0 one interface at a time, e.g., first those on Γ_1, then those on Γ_2, etc. It is not hard to see that the matrix D for this example is block tridiagonal, i.e., the unknowns corresponding to the interface Γ_i are connected only to the unknowns on the interfaces Γ_{i-1}, Γ_i, and Γ_{i+1}. By taking advantage of features such as this, the cost of step 5 and 6 of the algorithm can be greatly reduced, especially in three-dimensional settings. We note that these ideas are similar to those connected with one-way direction algorithms for positive definite problems [9].

REFERENCES

[1] F. Gantmacher, The Theory of Matrices, Chelsea, New York, 1960.

[2] G. Golub and D. Mayers, "Use of preconditioning over irregular regions," in Computer Methods in Applied Science and Engineering, VI, (R. Glowinski and J. Lions, Eds.), 1983, pp. 3-14.

[3] M. Gunzburger and R. Nicolaides, "Elimination with noninvertible pivots," Linear Algebra Appl., 64, 1985, pp. 183-189.

[4] V. Girault and P.-A. Raviart, Finite Element Approximation of the Navier-Stokes Equations, Springer, Berlin, 1979.

[5] J. Boland and R. Nicolaides, "Stability of finite elements under divergence constraints," SIAM J. Numer. Anal., 20, 1983, pp. 722-731.

[6] J. Boland and R. Nicolaides, "Stable and semistable low order finite elements for viscous flows," SIAM J. Numer. Anal., 22, 1985, pp. 474-492.

[7] G. Fix, M. Gunzburger, and R. Nicolaides, "On mixed finite element methods for first-order elliptic systems," Numer. Math., 37, 1981, pp. 29-48.

[8] L. Adams and R. Voigt, "A methodology for exploiting parallelism in the finite element process," in Proceedings of the NATO Workshop on High Speed Computations, (J. Kowolik, Ed.), Springer-Verlag, Berlin, 1984, pp. 373-392.

[9] A. George and J. Liu, Computer Solution of Large Sparse Positive Definite Systems, Prentice Hall, Englewood Cliffs, New Jersey, 1981.

Multiple Laminar Flows Through Curved Pipes*

Zhong-hua Yang† and H.B. Keller

Applied Mathematics, Caltech, Pasadena, CA 91125

Abstract

The Dean problem of steady viscous flow through a coiled circular pipe is studied numerically for a large range of Dean number and for several coiling ratios. We find that the solution family, as parameterized by Dean number, has numerous folds or limit points. Four folds and hence five branches of solutions are found. We speculate that infinitely many solutions can exist in this family for some fixed value(s) of D . More resolution and higher accuracy would be required to justify our conjecture and to find the rule of formation of new solution branches.

*This work was supported by the U.S. Department of Energy Office of Basic Energy Sciences (contract DE-AS03-76SF 00767), and by the Army Research Office (contract DAAG-29-81-K-0107). The calculations were done on the Caltech Applied Math IBM-4341 supplied and supported by the IBM Corporation.

†Permanent address: Shanghai University of Science and Technology, Jiading, Shanghai, China.

1. Introduction

Following the early work of Dean (1927, 1928) there have been several numerical studies of the steady, laminar, viscous flow of an incompressible fluid through a slightly curved pipe of circular cross section. In particular, Dennis (1980) with Collins (1975) and with Ng (1982) have computed such flows when the coiling ratio a/L is small. Here a is the pipe radius and L is the radius of curvature of the axis of the pipe. Also Van Dyke has applied the Stokes series and Dombes-Sykes technique (1978) to this problem. In all of this work the crucial parameter is the Dean number, D , defined as

$$D \equiv G\, a^3 \left(\frac{2a}{L}\right)^{1/2} / \mu\nu \tag{1.1}$$

where G is the constant pressure gradient driving the flow, μ is the viscosity and ν is the coefficient of kinematic viscosity. For small D and $a/L << 1$ all of the results agree.

In particular for a straight pipe, $a/L = 0$, the flow is the classical Poiseuille flow. However a slight curvature of the pipe axis induces a centrifugal force on the fluid which then forms a secondary flow, sending fluid outward along the symmetry axis and returning along the upper and lower curved surfaces. Thus a pair of symmetric vortices is superposed on the Poiseuille flow. These qualitative features are observed in all of the previously cited references for D small and $a/L << 1$. What happens as D and a/L increase? Few of the previous studies consider $a/L = O(1)$. Further, Van Dyke's expansions disagree with the finite difference calculations for larger values of D. And in Dennis & Ng (1982) dual solutions are found for the range $957.5 < D < 5000$; that is a four vortex solution is computed in addition to the standard two vortex flow described above.

In this paper we attempt to clarify the situation by determining the structure of the families of solutions that exist as D varies. In addition we show how this

structure changes as a/L increases (to 0.3). For this purpose we must retain the full Navier-Stokes equations and do not make the $a/L << 1$ simplifications. However no dramatic effects are found as a/L increases. Regarding the structure with respect to D we are not completely successful. Our results suggest, in analogy with the von Kármán swirling flows (Lentini & Keller 1980), that there may be *infinitely many steady flows* for some value (or interval) of D. However, we have found only five branches of such flows and believe that more numerical accuracy is required to completely settle the question. Indeed our first, cruder calculations revealed only three branches of solutions. Unfortunately the variation in flow patterns from one branch to the next are not as regular as those in the von Kármán swirling flows, so that we cannot have the same confidence in our current conjecture. Also, we do not see analytical regularities in the five flows we have detected.

After our study was completed we learned of related calculations in curved tubes by Winters and Brindley (1984) and by Winters (1984). However that work is mainly concerned with tubes of rectangular cross section, with a brief mention of the circular case in Winters and Brindley (1984). Bifurcations are obtained for the rectangular case but they do not examine the results we study here.

In section 2 we formulate the problem retaining the exact equations (valid to all orders in $\epsilon = a/L$). Expansions in Fourier series are introduced in section 3 to get a system of nonlinear two-point boundary value problems for the Fourier coefficients. Numerical methods are introduced in section 4. These employ centered differences and Newton's method with continuation or path following techniques introduced by H.B. Keller (1977). The results are presented and discussed in section 5.

2. General Formulation

We employ the notation used in Collins & Dennis (1975) and Dennis & Ng (1982) as indicated in Figure 1. The circular cross section of the tube in the (x, y)-plane has radius a with center at L on the x-axis. The tube is coiled about a

circle of radius L in the (x,z)-plane. With no pitch in the coil the tube thus forms a torus. Our equations are exact for this case. Dimensionless velocity components of the fluid are (u,v,w) at a point P with dimensionless polar coordinates (r,α). Here u is the radial and v is the angular component of velocity in the pipe cross section, w is the axial velocity normal to the cross section and $r \equiv r'/a$ where r' is the dimensional radius.

We seek flows independent of θ, the angular deviation from the (x,y)-plane. A stream function $\phi(r,\alpha)$ is introduced in terms of which the transverse velocity components are given by:

$$u(r,\alpha) = \frac{1}{r(1+\epsilon\, r\, \cos\, \alpha)}\, \frac{\partial \phi}{\partial \alpha}\,,$$
$$v(r,\alpha) = \frac{-1}{(1+\epsilon\, r\, \cos\, \alpha)}\, \frac{\partial \phi}{\partial r}\,. \qquad (2.1)$$

Here $\epsilon \equiv a/L$ is the "coiling ratio" and the continuity equation is thus satisfied. Using these velocity components in the Navier-Stokes equations we introduce the modified Laplacian

$$\tilde{\nabla}^2 \equiv \frac{1+\epsilon\, r\, \cos\, \alpha}{r}\, \left[\frac{\partial}{\partial r}\left(\frac{r}{1+\epsilon\, r\, \cos\, \alpha}\, \frac{\partial}{\partial r}\right) + \frac{\partial}{\partial \alpha}\, \left(\frac{\epsilon\, \sin\, \alpha}{1+\epsilon\, r\, \cos\, \alpha}\, \frac{\partial}{\partial \alpha}\right)\right] \qquad (2.2)$$

and the vorticity

$$\Omega = -\tilde{\nabla}^2 \phi\,, \qquad (2.3)$$

to get for the w-momentum equation

$$\tilde{\nabla}^2 w + \frac{1}{r(1+\,\epsilon\, r\, \cos\, \alpha)}\, \left(\frac{\partial \phi}{\partial r}\, \frac{\partial w}{\partial \alpha} - \frac{\partial \phi}{\partial \alpha}\, \frac{\partial w}{\partial r}\right) = -D\,, \qquad (2.4)$$

and on elimination of the pressure from the other momentum equations:

$$\tilde{\nabla}^2\Omega + \frac{1}{r(1+\epsilon\, r\, \cos\,\alpha)}\left(\frac{\partial\phi}{\partial r}\,\frac{\partial\Omega}{\partial\alpha} - \frac{\partial\phi}{\partial\alpha}\,\frac{\partial\Omega}{\partial r}\right)$$
$$+ \frac{2\epsilon\,\Omega}{(1+\epsilon\, r\, \cos\,\alpha)^2}\left(\sin\alpha\,\frac{\partial\phi}{\partial r} + \frac{\cos\alpha}{r}\,\frac{\partial\phi}{\partial\alpha}\right)$$
$$= \frac{w}{(1+\epsilon\, r\, \cos\alpha)^2}\left(\sin\alpha\,\frac{\partial w}{\partial r} + \frac{\cos\alpha}{r}\,\frac{\partial w}{\partial\alpha}\right) . \qquad (2.5)$$

The equations used in Dennis (1980) are obtained by setting $\epsilon = 0$ in (2.1)-(2.5) (*i.e.* they use the small coiling ratio approximation but we do not).

Boundary conditions on the wall of the tube, $r = 1$, yield:

$$w(1,\alpha) = \phi(1,\alpha) = \frac{\partial\phi}{\partial\alpha}\,(1,\alpha) = 0\ , \quad 0 \le \alpha \le \pi\ . \qquad (2.6)$$

We study here only flows symmetric about the x-axis for which:

$$w(r,\alpha) = w(r,-\alpha)\ , \quad \phi(r,\alpha) = -\phi(r,-\alpha)\ , \quad \Omega(r,\alpha) = -\Omega(r,-\alpha)\ . \qquad (2.7)$$

Thus on the symmetry axis we have:

$$\frac{\partial w}{\partial\alpha}\,(r,0) = \frac{\partial w}{\partial\alpha}\,(r,\pi) = 0\ ,$$
$$\phi(r,0) = \phi(r,\pi) = 0\ ,$$
$$\Omega(r,0) = \Omega(r,\pi) = 0\ . \qquad (2.8)$$

3. Fourier Series Expansions

To solve the boundary value problem posed in (2.2)-(2.8) we seek Fourier expansions of the stream function, axial velocity and vorticity in the forms:

$$
\begin{aligned}
a) \quad & \phi(r,\alpha) = \sum_{k=1}^{\infty} f_k(r) \sin\ k\alpha\ ; \\
b) \quad & w(r,\alpha) = \sum_{k=0}^{\infty} w_k(r) \cos\ k\alpha\ ; \\
c) \quad & \Omega(r,\alpha) = \sum_{k=1}^{\infty} g_k(r) \sin\ k\alpha\ .
\end{aligned} \tag{3.1}
$$

With these forms the symmetry conditions (2.7) and the implied boundary conditions (2.8) are satisfied.

Using the expansions (3.1) in the differential equations (2.3)- (2.5) and applying the orthogonality properties and other identities for the trigonometric functions yields an infinite system of coupled nonlinear, second order ordinary differential equations for the coefficient functions $\{f_k(r),\ w_k(r),\ g_k(r)\}$. Specifically we get from (2.3), with the notation $f_0(r) \equiv g_0(r) \equiv 0$:

$$
\begin{aligned}
&\frac{\epsilon r}{2}\left[\frac{d^2}{dr^2} - \frac{(k-1)(k-2)}{r^2}\right] f_{k-1}(r) + \left[\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr} - \frac{k^2}{r^2}\right] f_k(r) \\
&\qquad + \frac{\epsilon r}{2}\left[\frac{d^2}{dr^2} - \frac{(k+1)(k+2)}{r^2}\right] f_{k+1}(r) \\
&\qquad = -\frac{\epsilon r}{2}\, g_{k-1}(r) - g_k(r) - \frac{\epsilon r}{2}\, g_{k+1}(r)\ , \quad k \geq 1\ .
\end{aligned} \tag{3.2}
$$

From (2.4) we get, with $w_{-1}(r) \equiv 0$:

$$
\begin{aligned}
&\frac{\epsilon r}{2}\left[\frac{d^2}{dr^2} - \frac{(k-1)(k-2)}{r^2}\right] w_{k-1}(r) + \left[\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr} - \frac{k^2}{r^2}\right] w_k(r) \\
&\qquad + \frac{\epsilon r}{2}\left[\frac{d^2}{dr^2} - \frac{(k+1)(k+2)}{r^2}\right] w_{k+1}(r) \\
&\qquad = R_k(r) - \delta_{k,1}\epsilon\ r\ D - \delta_{k,0} D\ , \quad k \geq 0\ .
\end{aligned} \tag{3.3}
$$

From (2.5) we get, with $g_{-1}(r) \equiv 0$:

$$\left(\frac{\epsilon r}{2}\right)^2 \left[\frac{d^2}{dr^2} - \frac{(k-2)(k-3)}{r^2}\right] g_{k-2}(r) + \frac{\epsilon r}{2}\left[2\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr} - \frac{(k-1)(2k-3)}{r^2}\right] g_{k-1}(r)$$
$$+ \left\{\left[\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr} - \frac{k^2}{r^2}\right] + \frac{\epsilon^2 r^2}{2}\left[\frac{d^2}{dr^2} - \frac{k^2}{r^2}\right]\right\} g_k(r)$$
$$+ \frac{\epsilon r}{2}\left[2\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr} - \frac{(k+1)(2k+3)}{r^2}\right] g_{k+1}(r) + \left(\frac{\epsilon r}{2}\right)^2 \left[\frac{d^2}{dr^2} - \frac{(k+2)(k+3)}{r^2}\right] g_{k+2}(r)$$
$$= \frac{\epsilon r}{2} S_{k-1}(r) + S_k(r) + \frac{\epsilon r}{2} S_{k+1}(r) + P_k(r) + \frac{\epsilon}{2} Q_k(r) , \quad k \geq 1 . \tag{3.4}$$

We have used the Kronecker symbol $\delta_{i,j}$ and introduced the quantities R_k, S_k, P_k and Q_k as:

$$a) \quad R_k(r) \equiv \frac{(1+\delta_{k,0})^{-1}}{2r} \sum_{n=0}^{\infty} \left\{ \left[|n-k| \, f_{|n-k|}(r) + (n+k) f_{n+k}(r) \right] w_n'(r) \right.$$
$$\left. + \, n\left[f_{n+k}' + \text{ sign } (n-k) f_{|n-k|}'(r) \right] w_n(r) \right\}$$

$$b) \quad S_k(r) \equiv \frac{1}{2r} \sum_{n=1}^{\infty} \left\{ \left[|n-k| \, f_{|n-k|}(r) - (n+k) f_{n+k}(r) \right] g_n'(r) \right.$$
$$\left. - \, n\left[f_{n+k}' - \text{ sign } (n-k) f_{|n-k|}'(r) \right] g_n(r) \right\}$$

$$c) \quad P_k(r) \equiv \frac{1}{4} \sum_{n=0}^{\infty} \left\{ \left[w_n'(r) - \frac{n}{r} w_n(r) \right] \left[(1+\delta_{k,n+1}) w_{|n+1-k|}(r) - w_{n+1+k}(r) \right] \right.$$
$$\left. - \left[w_n'(r) + \frac{n}{r} w_n(r) \right] \left[(1+\delta_{n-1,k}) w_{|n-1-k|}(r) - w_{n-1+k}(r) \right] \right\}$$

$$d) \quad Q_k(r) \equiv \sum_{n=1}^{\infty} \left\{ \left[f_n'(r) - \frac{n}{r} f_n(r) \right] \left[g_{n+1+k}(r) - \text{ sign } (n+1-k) g_{|n+1-k|}(r) \right] \right.$$
$$\left. - \left[f_n'(r) + \frac{n}{r} f_n(r) \right] \left[g_{n-1+k}(r) - \text{ sign } (n-1-k) g_{|n-1-k|}(r) \right] \right\} \tag{3.5}$$

At the origin, $r = 0$, of the polar coordinates (r, α) continuity requires that $\phi(0,\alpha)$, $w(0,\alpha)$ and $\Omega(0,\alpha)$ be independent of α. From (3.1) we thus get that:

$$f_k(0) = w_k(0) = g_k(0) = 0 , \quad k = 1, 2, \ldots . \tag{3.6}$$

Note that a condition on $w_0(0)$ is not obtained but $w_0(0) = w(0,\alpha)$. The conditions (2.6) at $r = 1$ applied to (3.1a,b) yield:

$$\begin{aligned} a) \quad & f_k(1) = 0 \,, \quad k = 1,2,\ldots \\ b) \quad & f_k'(1) = 0 \,, \quad k = 1,2,\ldots \\ c) \quad & w_k(1) = 0 \,, \quad k = 0,1,2,\ldots \end{aligned} \tag{3.7}$$

The formal consistency of "order" of the system and number of boundary conditions seems to be off by one since all of the equations are second order and we do not have two boundary conditions on $w_0(r)$. This is easily remedied by noting that the equation in (3.3) for $k = 0$ can be reduced to a first order equation. To do this we multiply by r and integrate over $[0,r]$. In evaluating at $r = 0$ we use (3.6) and the assumptions that:

$$\lim_{r\to 0} \left[r\, w_0'(r)\right] = \lim_{r\to 0} \left[r^2 w_1'(r)\right] = 0 \,.$$

The result is the first order equation:

$$\frac{d}{dr}\, w_0(r) + \frac{\epsilon r}{2} \left[\frac{d}{dr}\, w_1(r) - \frac{2}{r}\, w_1(r)\right] = \frac{1}{2r} \sum_{n=1}^{\infty} n\, f_n(r)\, w_n(r) - \frac{r}{2}\, D \,. \tag{3.8}$$

The analytical problem is thus reduced to solving (3.2) for $k \geq 1$, (3.3) for $k \geq 1$, (3.4) for $k \geq 1$ and (3.8) subject to the boundary conditions (3.6) and (3.7).

4. Numerical Procedures

To solve or rather to approximate the solution of the problem formulated in Section 3 we first truncate the Fourier expansions, we then use difference approximations on the resulting system of O.D.E.s and finally we solve the nonlinear difference equations by means of Newton's method and continuation procedures. We describe these techniques below.

A. Truncation of the Fourier Expansions

Under the assumption that the series in (3.1) converge sufficiently rapidly we replace them by the finite trigonometric expansions obtained by setting

$$f_k(r) \equiv w_k(r) \equiv g_k(r) \equiv 0 \ , \quad k > K \ . \tag{4.1a}$$

When we use (4.1) in the equations (3.2)-(3.8) we obtain a system of $3K$ second order and one first order ordinary differential equations for the $3K+1$ quantities:

$$f_k(r) \ , \quad g_k(r) \ , \quad 1 \le k \le K \ ; \quad w_k(r) \ , \quad 0 \le k \le K \ . \tag{4.1b}$$

there are $6K+1$ boundary conditions in (3.6) and (3.7) when we terminate those relations at $k = K$. We seek to solve this two-point boundary value problem numerically.

B. Difference Approximations.

We place a uniform grid of points $r_j = jh$, $0 \le j \le M+1$ with $r_{M+1} = 1$ on the interval $0 \le r \le 1$. At each point of this grid we introduce approximations to the coefficients in (4.1b) with the notation

$$f_k(r_j) \doteq f_{k,j} \ , \quad g_k(r_j) \doteq g_{k,j} \ , \quad w_k(r_j) \doteq w_{k,j}$$

We employ the difference operators, for any mesh function, say u_j :

$$D_+u_j \equiv \frac{u_{j+1} - u_j}{h} \ , \quad D_-u_j \equiv \frac{u_j - u_{j-1}}{h} \ , \quad D_0u_j \equiv \frac{u_{j+1} - u_{j-1}}{2h}$$

Then the discrete or difference approximations to (3.2), (3.3) and (3.4) are taken to be:

$$\frac{\epsilon r_j}{2}\left[D_+D_- - \frac{(k-1)(k-2)}{r_j^2}\right] f_{k-1,j} + \left[D_+D_- + \frac{1}{r_j}D_0 - \frac{k^2}{r_j^2}\right] f_{k,j}$$
$$+\frac{\epsilon r_j}{2}\left[D_+D_- - \frac{(k+1)(k+2)}{r_j^2}\right] f_{k+1,j} = -\frac{\epsilon r_j}{2}\, g_{k-1,j} - g_{k,j} - \frac{\epsilon r_j}{2}\, g_{k+1,j}; \tag{4.2}$$

$$\frac{\epsilon r_j}{2}\left[D_+D_- - \frac{(k-1)(k-2)}{r_j^2}\right] w_{k-1,j} + \left[D_+D_- + \frac{1}{r_j}\, D_0 - \frac{k^2}{r_j^2}\right] w_{k,j}$$
$$+\frac{\epsilon r_j}{2}\left[D_+D_- - \frac{(k+1)(k+2)}{r_j^2}\right] w_{k+1,j} = R_{k,j} - \delta_{k,1}\epsilon r_j\, D \quad ; \tag{4.3}$$

$$\left(\frac{\epsilon r_j}{2}\right)^2\left[D_+D_- - \frac{(k-2)(k-3)}{r_j^2}\right] g_{k-2,j} + \frac{\epsilon r_j}{2}\left[2D_+D_- + \frac{1}{r_j}\, D_0 - \frac{(k-1)(2k-3)}{r_j^2}\right] g_{k-1,j}$$
$$+\left\{\left[D_+D_- + \frac{1}{r_j}\, D_0 - \frac{k^2}{r_j^2}\right] + \frac{\epsilon^2 r_j^2}{2}\left[D_+D_- - \frac{k^2}{r_j^2}\right]\right\} g_{k,j}$$
$$+\frac{\epsilon r}{2}\left[2D_+D_- + \frac{1}{r_j}\, D_0 - \frac{(k+1)(2k+3)}{r_j^2}\right] g_{k+1,j} + \left(\frac{\epsilon r_j}{2}\right)^2\left[D_+D_- - \frac{(k+2)(k+3)}{r_j^2}\right] g_{k+2,j}$$
$$= \frac{\epsilon r_j}{2}\, S_{k-1,j} + S_{k+1,j} + \frac{\epsilon r_j}{2}\, S_{k+1,j} + P_{k,j} + \frac{\epsilon}{2}\, Q_{k,j}\,; \tag{4.4}$$

Each of these difference equations is imposed for

$$j = 1, 2, \ldots, M\,,$$

$$k = 1, 2, \ldots, K\,.$$

The quantities $R_{k,j}$, $S_{k,j}$, $P_{k,j}$, and $Q_{k,j}$ are the obvious finite difference approximations to the quantities in (3.5) centered at r_j. Since only first

derivatives occur in these expressions we employ $D_0 w_{n,j}$ to approximate $w_n'(r_j)$, etc. The remaining first order equation (3.8) is centered at the points $r_{j-\frac{1}{2}} \equiv (j-\frac{1}{2})h$ as follows:

$$D_- w_{0,j} + \frac{\epsilon r_{j-\frac{1}{2}}}{2}\left[D_- w_{1,j} - \frac{2}{r_{j-\frac{1}{2}}}\left(\frac{w_{1,j} + w_{1,j-1}}{2}\right)\right]$$
$$= \frac{1}{2r_{j-\frac{1}{2}}}\sum_{n=1}^{K} n\left(\frac{f_{n,j} + f_{n,j-1}}{2}\right)\left(\frac{w_{n,j} + w_{n,j-1}}{2}\right) - r_{j-\frac{1}{2}}\,\frac{D}{2}\,, \tag{4.5}$$

for

$$j = 1, 2, \ldots, M+1 \ .$$

The boundary conditions (3.6) and (3.7a,c) go over into the corresponding conditions:

$$\begin{aligned} &a) \quad f_{k,0} = w_{k,0} = g_{k,0} = 0 \ , \quad k = 1, 2, \ldots, K \ ; \\ &b) \quad f_{k,M+1} = w_{k,M+1} = 0 \ , \quad k = 1, 2, \ldots, K \ ; \quad w_{0,M+1} = 0 \ . \end{aligned} \tag{4.6}$$

The remaining conditions, in (3.7b), are imposed in order to retain second order accuracy as:

$$D_0 \, f_{k,M+1} \equiv \frac{f_{k,M+2} - f_{k,M}}{2h} = 0 \ , \quad k = 1, 2, \ldots, K \ .$$

Of course the meshpoint r_{M+2} is not in $[0,1]$ and so the values $f_{k,M+2}$ seem extraneous. However they are eliminated by imposing the difference equations in (4.2) at $j = M+1$. The result, after using (4.6b) and the above, is for $\epsilon = 0$:

$$g_{k,M+1} = -\frac{2}{h^2} \, f_{k,M} \ , \quad k = 1, 2, \ldots, K \ . \tag{4.7}$$

For $\epsilon > 0$ we must add the terms:

$$\frac{\epsilon}{2} \left[g_{k-1,M+1} + g_{k+1,M+1} + D_+ D_- \, (f_{k-1,M} + f_{k+1,M}) \right]$$

The numerical problem is to solve the nonlinear system of difference equations in (4.2), (4.3), (4.4), (4.5) and (4.7). These form $3KM + K + M + 1$ equations. There are precisely that many unknowns $\{ f_{k,j} \ , \ w_{k,j} \ , \ g_{k,j} \}$ when the quantities in (4.6) are eliminated. We go further and use (4.7) to eliminate the K quantities $\{ g_{k,M+1} \}$. Then we have only $(3K+1)M + 1$ equations and unknowns.

C. Newton's Method and Continuation.

To solve the difference equations we use Newton's method combined with continuation procedures to insure good initial estimates of the solution as the parameters are varied. To do this efficiently the unknowns must be ordered in a manner that simplifies the structure of the Jacobian matrix. To describe our ordering we first introduce the vectors $\underset{\sim}{f}_j$, $\underset{\sim}{g}_j$ and $\underset{\sim}{w}_j$ of dimensions K, K and $K+1$, respectively, by:

$$\begin{aligned}
\underset{\sim}{f}_j^T &\equiv (f_{1,j},\ f_{2,j},\ldots,f_{K,j}) , \quad 1 \le j \le M ; \\
\underset{\sim}{g}_j^T &\equiv (g_{1,j},\ g_{2,j},\ldots,g_{K,j}) , \quad 1 \le j \le M+1 ; \\
\underset{\sim}{w}_j^T &\equiv (w_{0,j},\ w_{1,j},\ldots,w_{K,j}) , \quad 1 \le j \le M . \qquad (4.8)
\end{aligned}$$

Recall that (4.7) gives: $\underset{\sim}{g}_{M+1} = -\frac{2}{h^2}\ \underset{\sim}{f}_M$ (for the case $\epsilon = 0$) and so $\underset{\sim}{g}_{M+1}$ can be eliminated. The remaining $(3K+1)M+1$ unknowns are represented in the vector $\underset{\approx}{X}$ defined by:

$$\underset{\approx}{X}^T \equiv (w_{0,0}; \underset{\sim}{f}_1^T,\ \underset{\sim}{g}_1^T,\ \underset{\sim}{w}_1^T; \ldots; \underset{\sim}{f}_M^T,\ \underset{\sim}{g}_M^T,\ \underset{\sim}{w}_M^T) . \qquad (4.9)$$

Now we order the equations in a corresponding manner. That is for a fixed j-value (*i.e.* meshpoint) we take (4.5) and all of (4.2), (4.3) and (4.4) for $1 \le k \le K$. The equations ordered in this manner for $j = 1, 2, \ldots, M$ and finally (4.5) for $j = M+1$ can be written as a vector equation in the form

$$\underset{\approx}{G}(\underset{\approx}{X}; D, \epsilon) = \underset{\sim}{0} . \qquad (4.10)$$

Here $\underset{\approx}{G}$ has $(3K+1)$ components, each being one of the difference equations. We have indicated the dependence of these equations on the parameters D and ϵ as they play a special role in the continuation procedures. For a fixed value of D and ϵ we

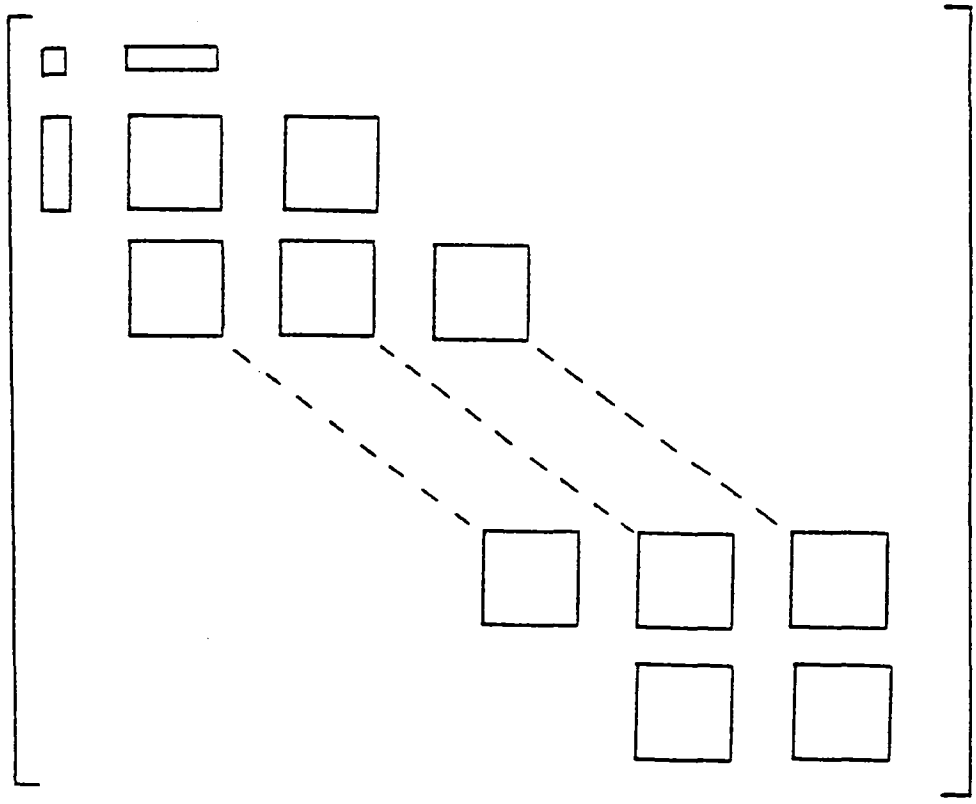
denote a solution of (4.10) by $\underset{\sim}{X} = \underset{\sim}{X}(D,\epsilon)$. When $D = 0$ and $\epsilon = 0$ an exact solution of the continuous problem is given by Poiseuille flow. Thus we easily get a solution of the discrete problem in this case. As D or ϵ deviates from zero we can use the Poiseuille flow as an initial estimate of the discrete solution in Newton's method applied to the system (4.10). This gives a sequence of iterates $\{\underset{\sim}{X}^{(\nu)}(D,\epsilon)\}$ defined by:

$$a) \quad \underset{\sim}{X}^{(0)}(D,\epsilon) \equiv \text{ initial estimate ,}$$

$$b) \quad \underset{\sim}{G}_{\underset{\sim}{X}}(\underset{\sim}{X}^{(\nu)};D,\epsilon)\ [\underset{\sim}{X}^{(\nu+1)} - \underset{\sim}{X}^{(\nu)}] = -\underset{\sim}{G}(\underset{\sim}{X}^{(\nu)};D,\epsilon),\ \nu = 0,1,2,\ldots. \quad (4.11)$$

Here $\underset{\sim}{G}_{\underset{\sim}{X}}$ is the Jacobian matrix which as a result of the above indicated ordering has the block-band structure indicated below. Each square block is a matrix of order $(3K+1) \times (3K+1)$. There are M

$\underset{\sim}{G}_{\underset{\sim}{X}} \equiv$



such rows of blocks. This array of blocks is bordered by one row and column as shown. All other elements in $\underset{\sim}{G}_{\underset{\sim}{X}}$ are zero. Most of the computing effort goes into

solving the linear algebraic systems in (4.11b). Thus to reduce the number of times this must be done we seek accurate initial estimates.

One way to obtain good initial estimates is to use two terms in a Taylor expansion of the solution with respect to changes in the parameter D, say. Thus we use:

$$\underset{\sim}{X}^{(0)}(D+\delta D,\epsilon)=\underset{\sim}{X}(D,\epsilon)+\delta D\underset{\sim}{X}_D(D,\epsilon) \tag{4.12a}$$

To obtain $\underset{\sim}{X}_D$ we note, from (4.10), that it satisfies:

$$\underset{\sim}{G}_{\underset{\sim}{X}}(\underset{\sim}{X}(D,\epsilon);D,\epsilon)\underset{\sim}{X}_D=-\underset{\sim}{G}_D(\underset{\sim}{X}(D,\epsilon);D,\epsilon) \tag{4.12b}$$

This system is similar to those in (4.11b). In fact when Newton's method has converged, the last time we solve (4.11b) we can also solve (4.12b) and thus $\underset{\sim}{X}_D(D,\epsilon)$ is determined with little extra work (*i.e.* only the backsolves and evaluation of $\underset{\sim}{G}_D$ need be done). Continuation with respect to ϵ can be done in an exactly similar manner.

The method described in (4.11), (4.12) is known as Euler-Newton continuation. It is extremely effective and usually converges quadratically. There are many refinements regarding step length procedures, efficient solution of the block-banded linear systems, approximation of Jacobians, etc., which we do not discuss here. Failure of the method to converge does occur, however, and it usually signals the presence of a bifurcation or fold point on the solution path (or family) being generated. Such points or solutions are called singular because the Jacobian matrix evaluated at these solutions is singular. Almost all such singular points are what we call simple folds or limit points. In particular a simple fold with respect to D is a singular solution, say $[\underset{\sim}{X}_0,D_0,\epsilon_0]$, which has the properties that:

$$
\begin{aligned}
&a) \quad \dim N(\underset{\sim}{G}^0_{\underset{\sim}{X}}) = 1 \ ; \ \ (i.e., \text{ all solutions of} \\
&\qquad \underset{\sim}{G}^0_{\underset{\sim}{X}} \underset{\sim}{\phi} = \underset{\sim}{0} \ \text{ are } \underset{\sim}{\phi} \equiv \alpha \, \underset{\sim}{\phi}_0 \ ; \ \ \alpha \in \Re \ , \ \text{some } \underset{\sim}{\phi}_0 \neq \underset{\sim}{0}) \\
&b) \quad \underset{\sim}{G}^0_D \notin \Re(\underset{\sim}{G}^0_{\underset{\sim}{X}}) \quad (i.e. \ \langle \underset{\sim}{G}^0_D, \underset{\sim}{\psi} \rangle \neq 0 \ \text{ for all solutions of} \\
&\qquad (\underset{\sim}{G}^0_{\underset{\sim}{X}})^T \underset{\sim}{\psi} = 0 \ ).
\end{aligned}
\tag{4.13}
$$

Here $\underset{\sim}{G}^0_{\underset{\sim}{X}} \equiv \underset{\sim}{G}_{\underset{\sim}{X}}(\underset{\sim}{X}_0; D_0, \epsilon_0)$ and $\underset{\sim}{G}^0_D \equiv \underset{\sim}{G}_D(\underset{\sim}{X}_0; D_0, \epsilon_0)$. All of the singular solutions we have found in this work have been such simple fold points. We have sought bifurcation points but have found none.

It is not difficult to circumvent the convergence problems near fold points. We do this by using pseudo-arclength continuation as introduced in Keller (1977). That is, we do not parametrize the solution path or family by D (as we assume has been done above) but rather introduce a new parameter s and a new scalar constraint and seek to solve the inflated or augmented system:

$$
\begin{aligned}
&a) \quad \underset{\sim}{G}(\underset{\sim}{X}(s), D(s), \epsilon) = 0 \\
&b) \quad N(\underset{\sim}{X}(s), D(s), s) \equiv \left\langle \dot{\underset{\sim}{X}}(s_0), \ \left[\underset{\sim}{X}(s) - \underset{\sim}{X}(s_0)\right] \right\rangle \\
&\qquad\qquad + \dot{D}(s_0) \left[D(s) - D(s_0)\right] \ + \ (s - s_0) = 0
\end{aligned}
\tag{4.14}
$$

Here $[\underset{\sim}{X}(s_0), \ D(s_0)]$ is a previously computed solution for ϵ fixed in the present discussion and for $s = s_0$. By $\dot{\underset{\sim}{X}} = \frac{d\underset{\sim}{X}}{ds}$ and $\dot{D} = \frac{dD}{ds}$ we denote the components of a tangent vector to the solution path $\{\underset{\sim}{X}(s), D(s)\}$. The constraint (4.14b) simply requires that the point $[\underset{\sim}{X}(s), D(s)]$ lie on the plane normal to this tangent at a distance $(s - s_0)$ from the point of tangency.

We use the scheme (4.14) when the previous Euler-Newton scheme begins to show signs of failure (*i.e.* too many iterations till convergence). We solve (4.14) by Newton's method. The Jacobian of this system is

$$\frac{\partial(\underset{\sim}{G}, N)}{\partial(\underset{\sim}{X}, D)} = \begin{pmatrix} \underset{\sim}{G}_{\underset{\sim}{X}} & \underset{\sim}{G}_D \\ N_{\underset{\sim}{X}} & N_D \end{pmatrix} \tag{4.15}$$

This Jacobian is nonsingular at regular solution points *and* at simple fold points. That is why our method has no difficulties in computing solution paths through folds. To solve for the Newton iterates we use the Bordering Algorithm described in Keller (1977) which is designed for systems with coefficients as in (4.15).

By differentiating in (4.14a) with respect to s we find that $[\dot{\underset{\sim}{X}}(s), \dot{D}(s)]$, the tangent to the solution path, satisfies:

$$\underset{\sim}{G}_{\underset{\sim}{X}} X(s) + \underset{\sim}{G}_D D(s) = 0 \tag{4.16a}$$

To solve this we first solve

$$\underset{\sim}{G}_{\underset{\sim}{X}} \underset{\sim}{\xi}(s) = -\underset{\sim}{G}_D \tag{4.16b}$$

and then set

$$\underset{\sim}{X}(s) = D(s)\ \underset{\sim}{\xi}(s) \tag{4.16c}$$

However since the scale of s has not been determined we choose it to represent (local) arclength along the solution path. Thus we require that

$$\langle \underset{\sim}{X}(s), \underset{\sim}{X}(s) \rangle + D^2(s) = 1$$

and using (4.16c) in the above we get

$$D(s) = \pm\left(\sqrt{1 - < \underset{\sim}{\xi}, \underset{\sim}{\xi} >}\right)^{-1} \tag{4.16d}$$

The sign here is chosen so that $\langle \dot{\underset{\sim}{X}}(s), \dot{\underset{\sim}{X}}(s_0) \rangle > 0$ which determines the orientation along the solution path.

We determine a new tangent only after having solved (4.14). Then we replace $[\dot{\underset{\sim}{X}}(s_0), \dot{D}(s_0)]$ by the new tangent $[\dot{\underset{\sim}{X}}(s), \dot{D}(s)]$ and proceed as before.

5. Results of Calculations

In addition to the stream function and axial flow velocity we have computed Re , the Reynolds number based on the mean axial velocity:

$$Re = 2\sqrt{2} \int_0^1 w_0(r) r dr \ ;$$

and the friction ratio (ratio of curved, γ_c , to straight, γ_s , wall friction):

$$\frac{\gamma_c}{\gamma_s} = 4\sqrt{2}\, \frac{Re}{D}$$

We have computed solution paths with D varying for the following sets of values of Fourier truncation, K , mesh spacing, h , and coiling ratio, ϵ:

$$I. \quad K = 10 \ , \quad h = \frac{1}{40} \ ; \quad \epsilon = 0 \ ;$$

$$II. \quad K = 10 \ , \quad h = \frac{1}{60} \ ; \quad \epsilon = 0 \ , \ \epsilon = 0.1 \ ;$$

$$III. \quad K = 20 \ , \quad h = \frac{1}{60} \ ; \quad \epsilon = 0 \ , \ \epsilon = 0.1 \ , \ \epsilon = 0.2$$

Starting from the trivial state with $u = v = w = 0$ for $\epsilon = 0$ and $D = 0$ we used continuation with D increasing as described in Section 4. In each of the three cases a simple fold was found and arclength continuation was used to accurately locate the fold and to traverse it. The solution branches were then continued with decreasing D and, in each case, another fold was found. Again these folds were located accurately and traversed to obtain a third branch in each of the three cases, now with D increasing. For cases I and II, extensions of these third branches continued well beyond where we could trust the numerical results. However for case III a third and fourth fold were found, leading to five branches of solutions. In Table 1 we list the critical value of the Dean number, Dm , at the m-th fold.

For cases II and III the fold solutions found for $\epsilon = 0$ were continued in ϵ up to 0.1 and for case III the continuation went up to $\epsilon = 0.2$. These results are also given in Table 1.

We call the family of solutions varying continuously with D in $D_{m-1} < D < D_m$ the "m-th branch" ($D_0 \equiv 0$). Our calculations seem to suggest that the analytic problem has infinitely many branches although we have computed only five of them. Graphs of γ_c/γ_s *vs* D are given for cases I and III in Figures 2 and 3, respectively. On the first branch, that emanating from $D = 0$, the solutions are of the classic form described by Dean — we call these "two-vortex" flows (see Figure 4). These two-vortex flows persist on the entire first branch and over most of the second branch down (in D values) to about $D \approx 5000$ where four-vortex solutions gradually appear. These four-vortex flows are formed in the calculations by the development, as D decreases on the second branch, of a small weak pair of vortices about the axis of symmetry near the outer edge of the tube. This vortex pair grows as D decreases and persists onto the third branch as D then increases (see Figure 5). The four-vortex flows remain on the entire third branch and onto the fourth branch down to $D \approx 14,000$ where six-vortex flows appear. We believe that, as this process continues, $2n$-vortex flows can form for all $n = 1, 2, \ldots$. Indeed on the fifth branch we have computed 8-vortex solutions at $D \approx 25,000$ (see Figure 9).

In Table 2 we compare our computed values of γ_c/γ_s on the first branch with various values reported in the literature (for two-vortex flows). The agreement is quite good. Dennis and Ng (1982) have also obtained four-vortex solutions over $957.5 < D < 5000$. We claim that these solutions are on the third branch. They were obtained accidentally in Dennis and Ng (1982) as a result of convergence difficulties with increasing D values near 5000. Then as D was decreased the solution "jumped" back onto the first branch. This is typical of the behavior to be

expected near folds if no special technique for traversing them is used. Thus the intermediate second branch was not obtained in Dennis and Ng (1982). In Table 3 we compare the values of the four-vortex solutions obtained in Dennis and Ng (1982) with our values on the third branch. The agreement leaves no doubt as to the identity of the two results. The somewhat larger discrepancies at $D = 5000$ is due, we believe, to inaccuracies in Dennis and Ng where convergence difficulties occurred. Graphs of the stream function and axial velocity contour lines on the third branch also agree well with those in Dennis and Ng.

Over the interval $D_4 < D < D_3$ we have obtained five solution branches. To give some idea of how the solutions change we show in Figures 4-8 plots of contour lines of the stream function and axial velocity for the solution with $D = 8000$ on each of the five branches. In addition we display in Figure 9 the results for $D = 25,000$ on the fifth branch. The contour lines in each figure are at levels that differ by one tenth the value between maximum and minimum values of the quantity plotted. The values of these maxima and minima are given with each figure. The small closed contours (or almost points) near the maxima or minima are at the levels of 0.995 or 1.005 , respectively, of the critical values.

Least squares fits of the γ_c/γ_s vs D curves with $\epsilon = 0$ have been made in the form

$$\frac{\gamma_c}{\gamma_s} \doteq a_m + b_m \, D^{1/3}$$

On branches $m = 1, 2$ and 5 we get the coefficient values:

$$a_1 = 0.3 \ , \quad a_2 = 0.25 \ , \quad a_5 = 0.15 \quad \text{and} \quad b_1 = b_2 = b_5 = 1/8 \ .$$

Other exponents have been used but the $1/3$ power seems to fit the data best. It is not clear, in light of the multiplicity of solutions and the unsettled nature of the solutions for large D , what the significance of "asymptotic solutions" for

$D \to \infty$ can be. Thus we do not address this problem here but merely present the above fits for whatever use they may be.

During the course of this work we have benefitted from conversations with Prof. A. Acrivos. We also wish to thank Prof. S.C.R. Dennis who first brought the matter of multiple solutions to our attention and suggested that we work on it.

References

Collins, W.M. & Dennis, S.C.R. 1975 The steady motion of a viscous fluid in a curved tube. *Q. J. Mech. Appl. Math.* 28, 133-156.

Dean, W.R. 1927 Note on the motion of fluid in a curved pipe. *Phil. Mag.* 4, 208-223.

Dean, W.R. 1928 The stream-line motion of fluid in a curved pipe. *Phil. Mag.* 5, 673-695.

Dennis, S.C.R. 1980 Calculation of the steady flow through a curved tube using a new finite-difference method. *J. Fluid Mech.* 99, 449-467.

Dennis, S.C.R. & Ng, M. 1982 Dual solutions for steady laminar flow through a curved tube. *Q. J. Mech. Appl. Math.* 35, 305-324.

Keller, H.B. 1977 Numerical solutions of bifurcation and nonlinear eigenvalue problems. In: *Applications of Bifurcation Theory* (ed. Rabinowitz), pp. 359-384. Academic Press.

Lentini, M. & Keller, H.B. 1980 The Kármán swirling flows. *SIAM J. Appl. Math.* 38, 52-64.

Van Dyke, M.D. 1978 Extended Stokes series: laminar flow through a loosely coiled pipe. *J. Fluid Mech.* 86, 129-145.

Winters, K.H. & Brindley, R.G.G. 1984 Multiple solutions for laminar flow in helically-coiled tubes. AERE-R 11373, U.K. Atomic Energy Authority, Harwell.

Winters, K.H. 1984 A bifurcation study of laminar flow in a curved tube of rectangular cross-section. TP-1104, U.K. ARE, Harwell.

Table and Figure Captions

Table 1. Critical Dean number, D_m , at the m-th fold in the solution branches.

Table 2. Comparison of γ_c/γ_s on the two-vortex solutions of various works with the present solutions on the first branch.

Table 3. Comparison of the four-vortex solutions of Dennis and Ng (1982) with the present solutions on the third branch.

Figure 1. The tube cross-sections showing coordinates, velocity components, axial flow distribution sketch and cross-flow streamlines sketch.

Figure 2. Friction ratio, γ_c/γ_s , *vs.* Dean number, D , for case I: $K = 10$, $h = 1/40$, $\epsilon = 0$.

Figure 3. Friction ratio, γ_c/γ_s , *vs.* Dean number, D , for case III: $K = 20$, $h = 1/60$, $\epsilon = 0$.

Figure 4. Axial velocity, w , and stream function, ϕ , contour lines: $D = 8000$, $K = 20$, $h = 1/60$, $\epsilon = 0$. First branch: Max $w = 0$, Min $w = 0$, Max $\phi = 23.986$, Min $\phi = 0$.

Figure 5. Same as in Fig. 4. Second Branch: Max $w = 625.956$, Min $w = 0$, Max $\phi = 23.497$, Min $\phi = 0$.

Figure 6. Same as in Fig. 4. Third Branch: Max $w = 594.777$, Min $w = 0$, Max $\phi = 22.962$, Min $\phi = -12.897$.

Figure 7. Same as in Fig. 4. Fourth Branch: Max w = 613.697, Min w = 0 , Max ϕ = 21.783, Min ϕ = −8.716.

Figure 8. Same as in Fig. 4. Fifth Branch: Max w = 622.831, Min w = 0 , Max ϕ = 20.679, Min ϕ = −4.676 .

Figure 9. Axial velocity, w , and stream function, ϕ , contour lines: $D = 25,000$, $K = 20$, $h = 1/60$, $\epsilon = 0$. Fifth branch: Max w = 1412.730, Min w = 0 , Max ϕ = 31.494, Min ϕ = −14.335 .

Table 1

	K	h	ε	D_1	D_2	D_3	D_4
I.	10	$\frac{1}{40}$	0	12,120	951	---	---
II.	10	$\frac{1}{60}$	0	12,752	950	---	---
III.	20	$\frac{1}{60}$	0	25,146	955	15,642	7,725
II.	10	$\frac{1}{60}$	0.1	19,963	1,130	---	---
III.	20	$\frac{1}{60}$	0.1	27,508	1,138	18,179	10,576
III.	20	$\frac{1}{60}$	0.2	30,071	1,358	20,440	14,907

Table 2

D	Collins & Dennis '75	Dennis & Ng '82	Dennis '80	This Work
1000	1.550	1.548	1.546	1.548
2000	1.852	1.847		1.848
3000		2.064	2.063	2.065
4000		2.237	2.237	2.238
5000	2.392	2.377	2.383	2.383

Table 3

	γ_c/γ_s		$w_o(0)$		Re	
D	Dennis & Ng'82	This Work	Dennis & Ng'82	This Work	Dennis & Ng'82	This Work
2000	1.8329	1.8338	1.0803	1.0795	192.9	192.8
3000	2.0463	2.0472	1.0514	1.0522	259.2	259.1
4000	2.2177	2.2172	1.0390	1.0389	318.8	318.9
5000	2.3662	2.3527	1.0332	1.0368	373.5	375.7

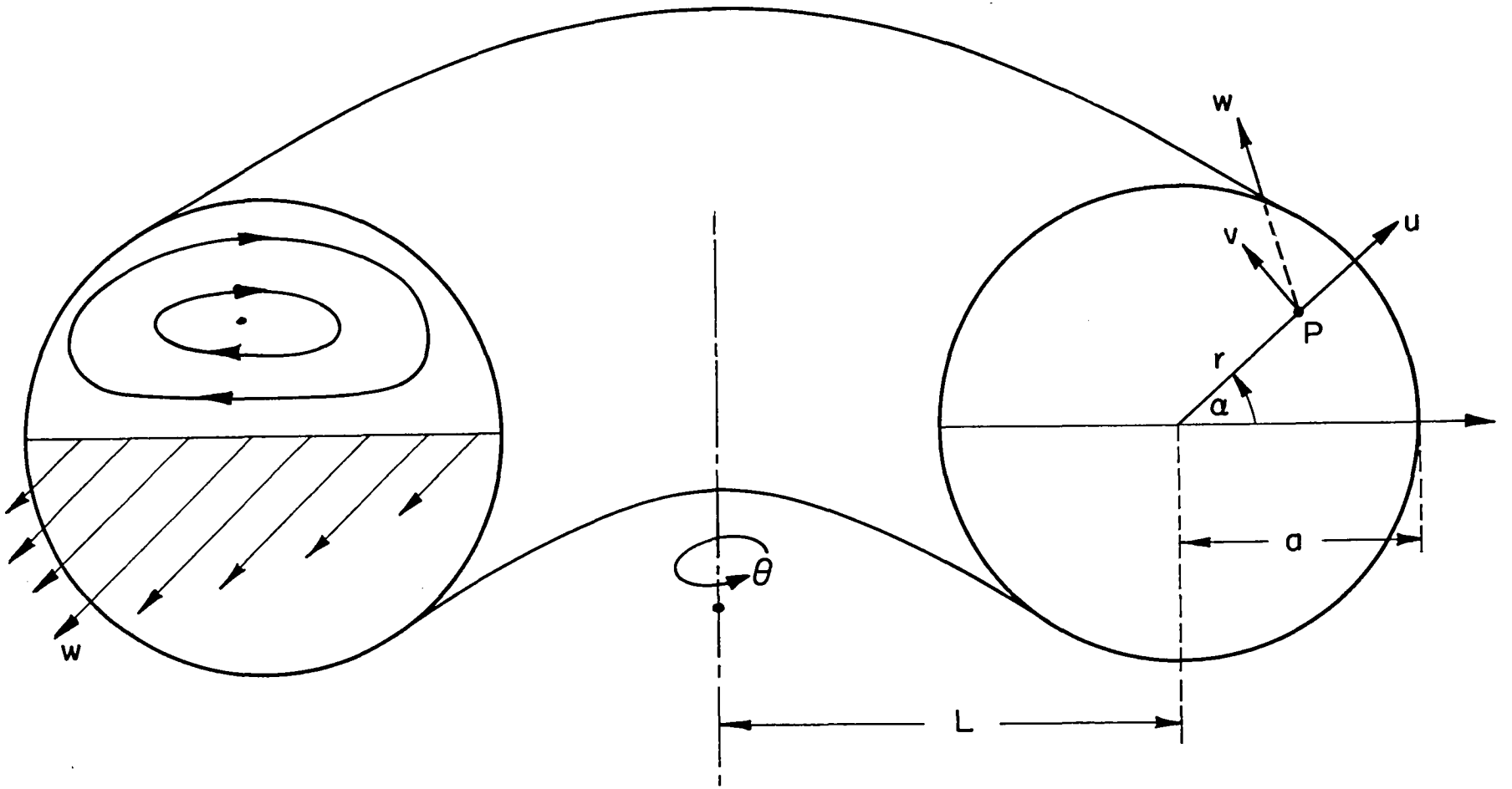


Figure 1.

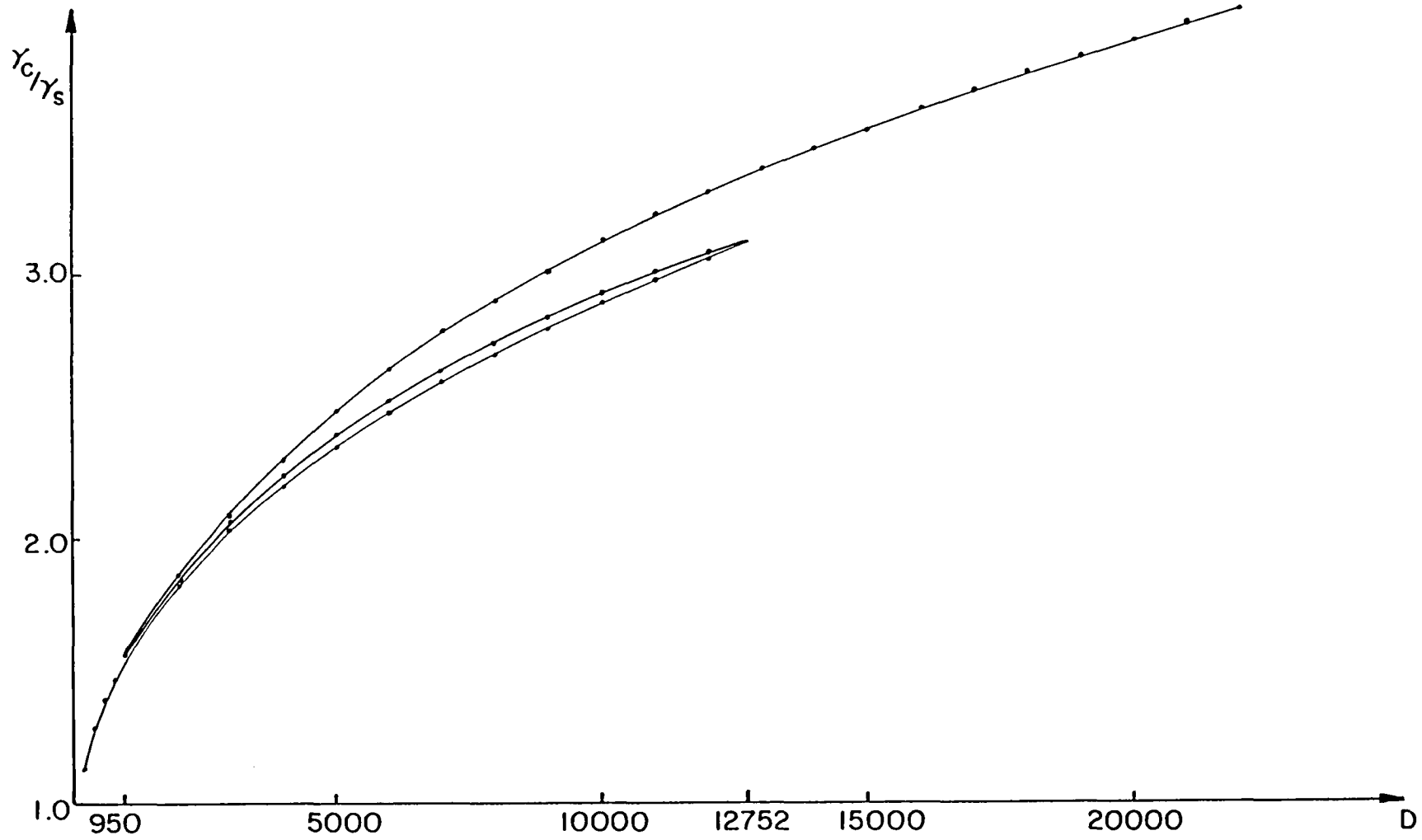


Figure 2.

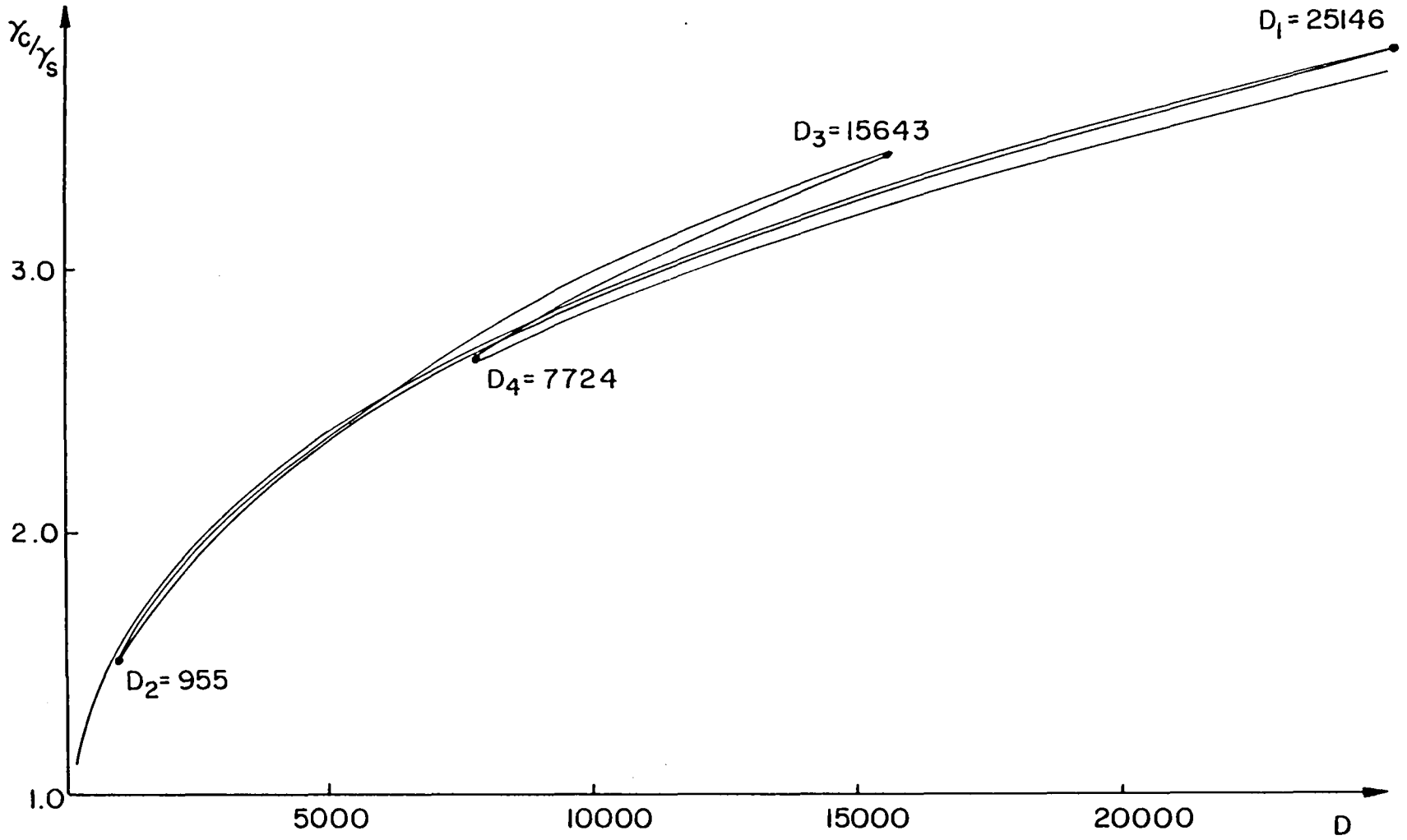
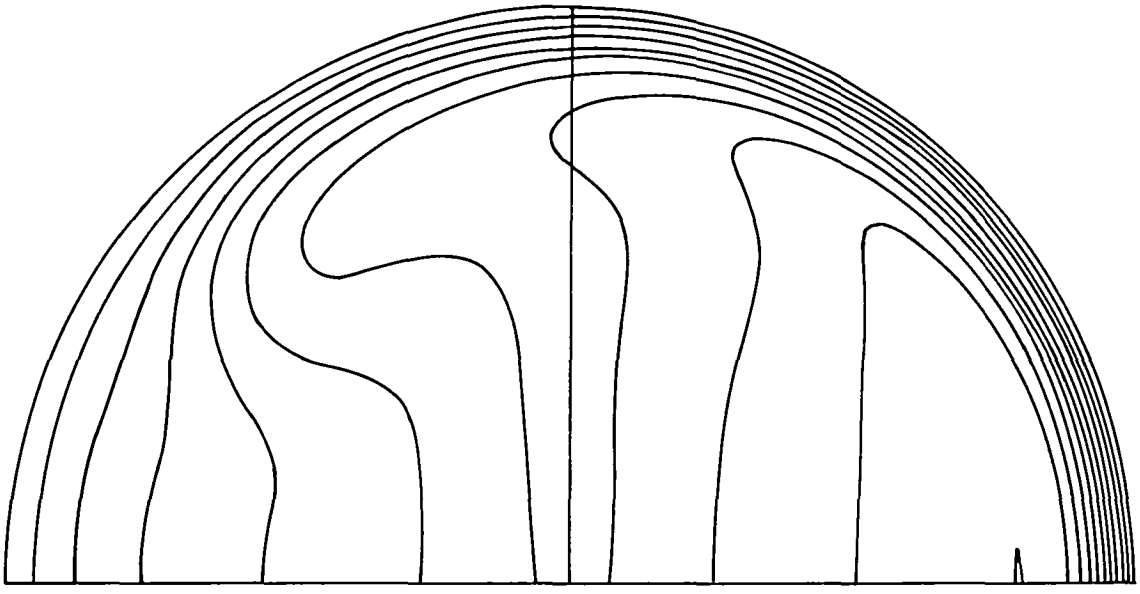
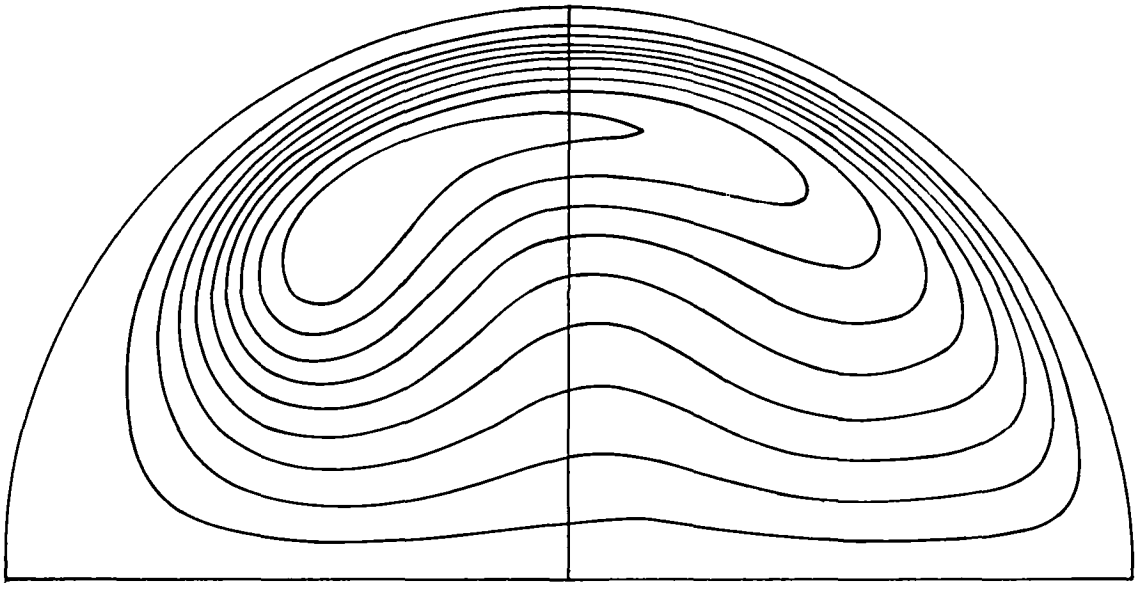

γ_c/γ_s
3.0
2.0
1.0
5000
10000
15000
20000
D
$D_1 = 25146$
$D_3 = 15643$
$D_4 = 7724$
$D_2 = 955$

Figure 3.

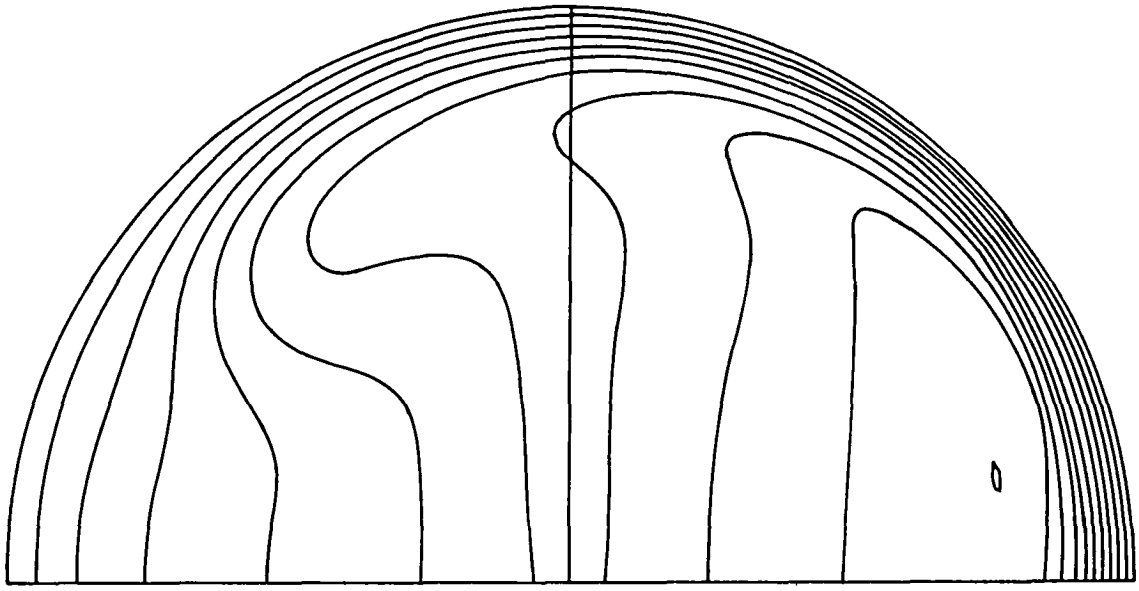


AXIAL VELOCITY CONTOURS

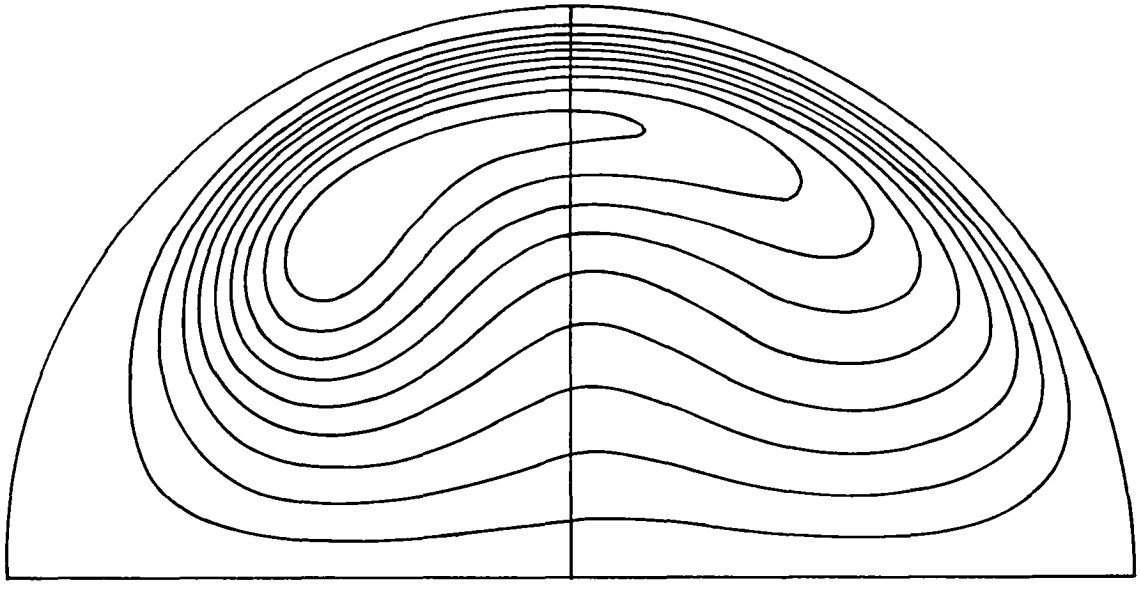


STREAM FUNCTION CONTOURS

Figure 4.

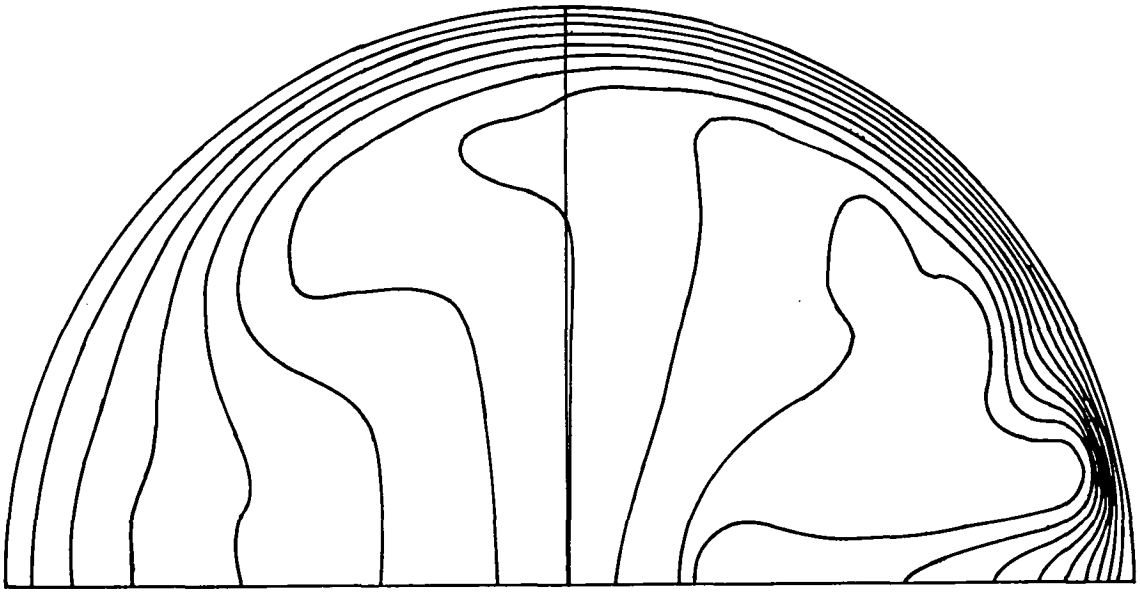


AXIAL VELOCITY CONTOURS

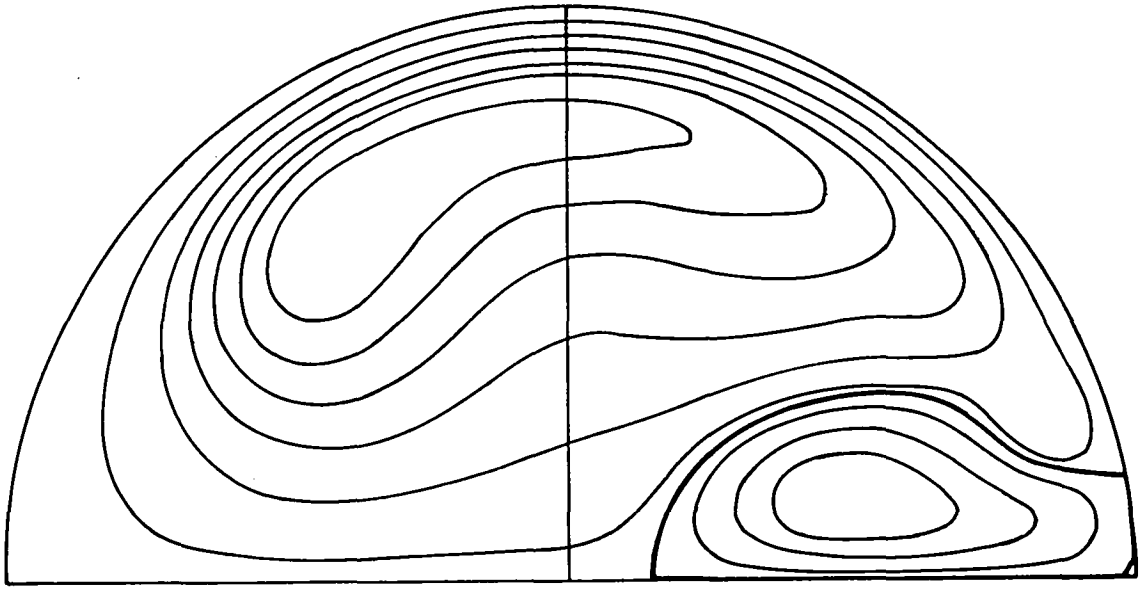


STREAM FUNCTION CONTOURS

Figure 5.

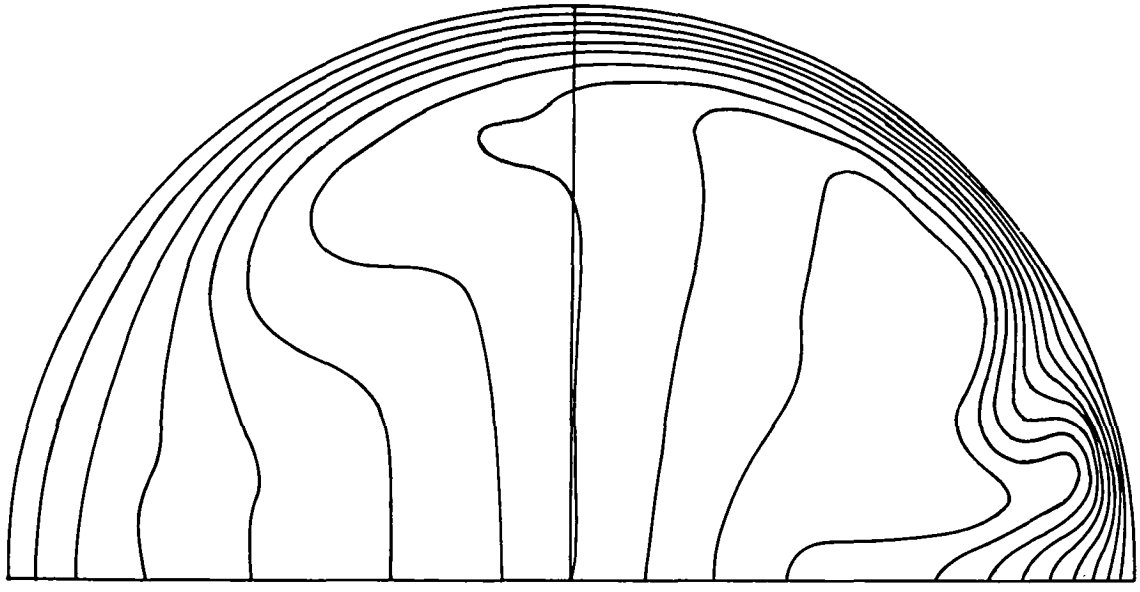


AXIAL VELOCITY CONTOURS

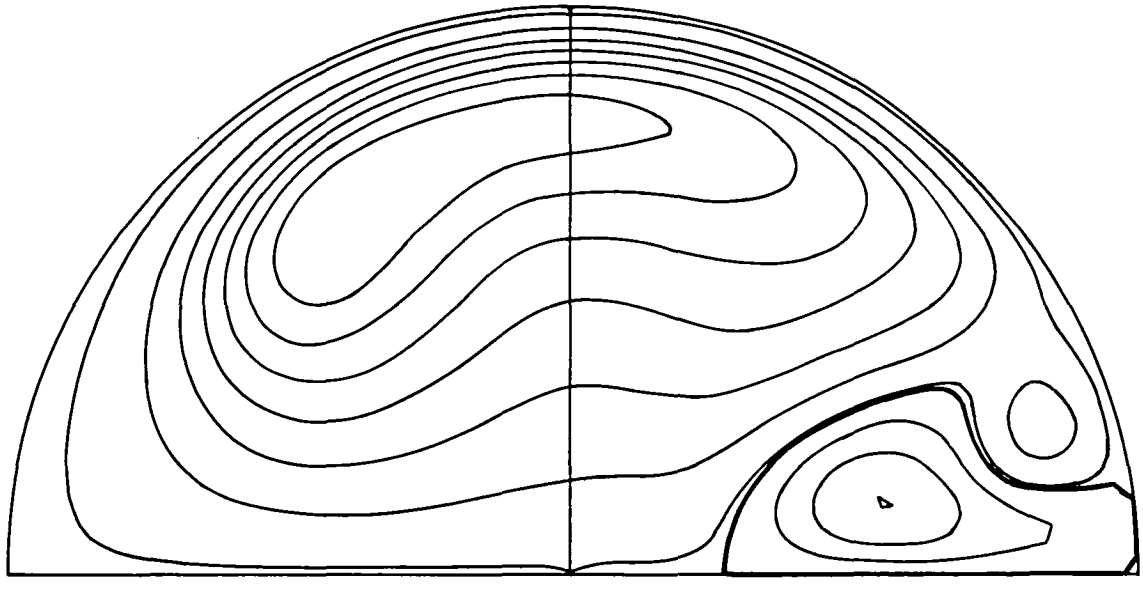


STREAM FUNCTION CONTOURS

Figure 6.

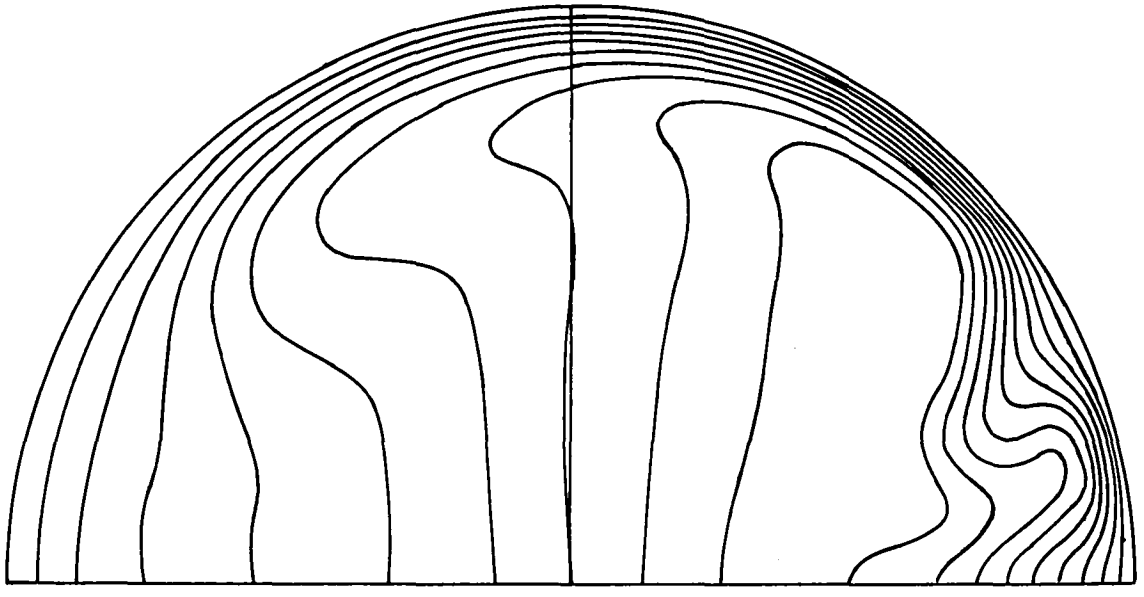


AXIAL VELOCITY CONTOURS

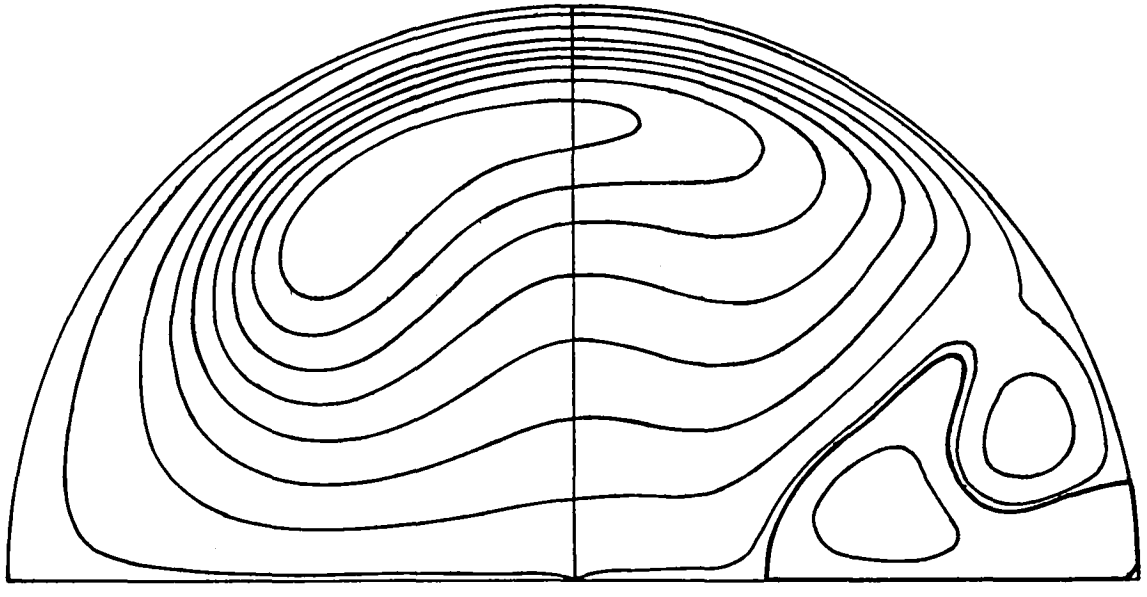


STREAM FUNCTION CONTOURS

Figure 7.

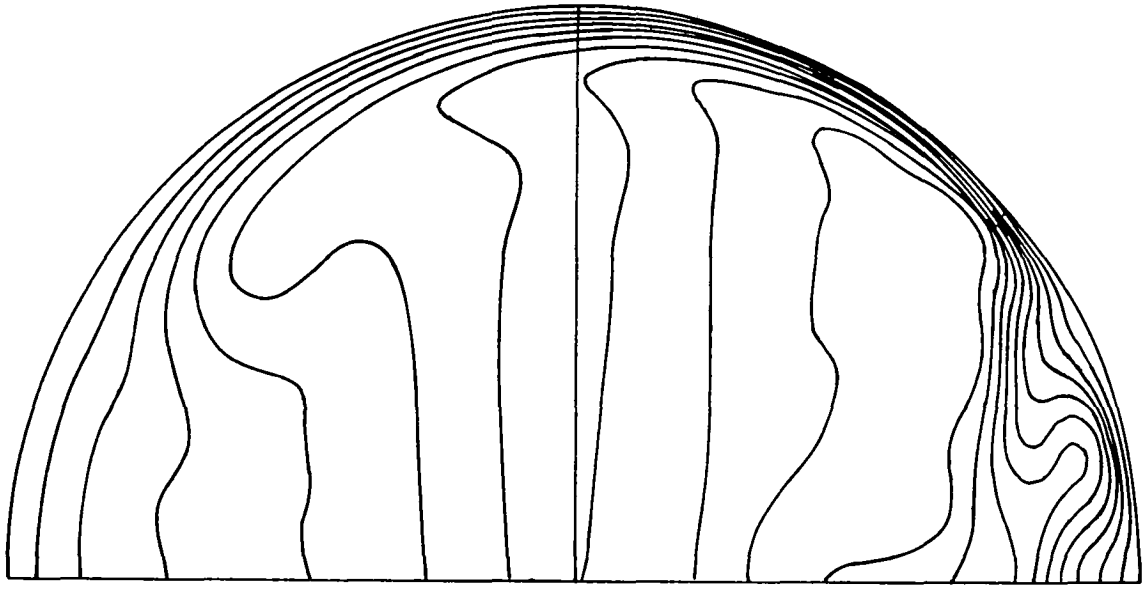


AXIAL VELOCITY CONTOURS

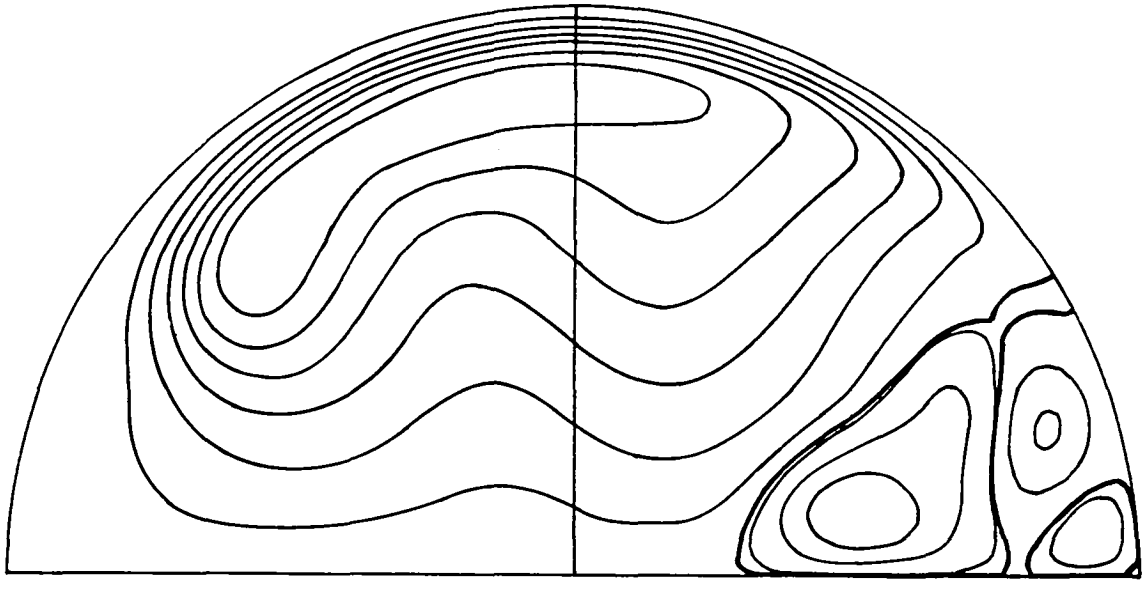


STREAM FUNCTION CONTOURS

Figure 8.



AXIAL VELOCITY CONTOURS



STREAM FUNCTION CONTOURS

Figure 9.

Calculations of the Stability of Some Axisymmetric Flows Proposed as a Model of Vortex Breakdown.

Nessan Mac Giolla Mhuiris

Institute for Computer Applications in Science and Engineering,
Mail Stop 132C, NASA Langley Research Center,
Hampton, Virginia 23665, USA.

ABSTRACT

The term "vortex breakdown" refers to the abrupt and drastic changes of structure that can sometimes occur in swirling flows. It has been conjectured that the "bubble" type of breakdown can be viewed as an axisymmetric wave travelling upstream in a primarily columnar vortex flow. In this scenario the wave's upstream progress is impeded only when it reaches a critical amplitude and it loses stability to some non-axisymmetric disturbance. We will investigate the stability of some axisymmetric wavy flows, which model vortex breakdown, to three dimensional disturbances viewing the amplitude of the wave as a bifurcation parameter. We will also look at the stability of a set of related, columnar vortex flows which are constructed by taking the two dimensional flow at a single axial location and extending it throughout the domain without variation. The method of our investigation will be to expand the perturbation velocity in a series of divergence free vectors which ensures that the continuity equation for the incompressible fluid is satisfied exactly by the computed velocity field. Projections of the stability equation onto the space of inviscid vector fields eliminates the pressure term from the equation and reduces the differential eigen problem to a generalized matrix eigen problem. Results are presented both for the one dimensional, columnar vortex flows and also for the wavy "bubble" flows.

1. Introduction: Vortex Breakdown.

The term "vortex breakdown" refers to the abrupt and drastic changes of structure that can sometimes occur in vortex flows. Observations by Peckham & Atkinson [1957] of breakdowns occurring in the leading edge vortex formed above a swept back lifting surface and a number of studies demonstrating the serious aerodynamic consequences of such events (the slopes of the lift, drag and moment curves are all altered by breakdown) stimulated early interest in the subject. Since that time the literature on vortex breakdown has burgeoned. The interested reader is referred to review articles by Hall [1972] and Leibovich [1978, 1984] for summaries both of the experimental observations that have been made and also of the theories that have been proposed to explain them.

Experimental observations are most easily made on vortex flows confined to tubes and the bulk of the available data is for such cases. In one apparatus, used by a number of researchers, water is passed radially inward through a set of guidevanes imparting swirl to the fluid which then enters axially into a test section (a frustrum of a cone of very small cone angle), by means of an annular channel formed between a bellmouth opening on the section and a centerbody whose tip is aligned with the cone axis. The boundary layer shed from the tip of the centerbody forms a well defined vortex core along the axis of the test section and dye injected through the tip allows for flow visualization.

With this type af apparatus two parameters are within the easy control of the experimentalist, namely the amount of swirl imparted to the inlet flow and the volume flow rate through the tube (effectively the Reynolds number of the flow). As the Reynolds number is increased for a sufficently large, fixed value of swirl the breakdown assumes one of two characteristic forms.

Both of these are characterized by a rapid deceleration of the axial velocity component, occurring in the axial distance on the order of one vortex core diameter, followed by the formation of a stagnation point and (in some frame of reference) a region of reversed flow along the axis. The two forms are easily distinguished in flow visualization studies as in one form, the spiral or S type breakdown, the tracer dye assumes a spiral shape rotating in the same sense as the inlet fluid, while in the other form, the bubble or B type breakdown, the dye assumes a form that looks much like a body of revolution placed in the fluid. Our interest will be in this latter form of breakdown which is sketched in Figure 1. Here we show "ideal" or averaged stream surfaces on which the fluid particles travel in helical paths. (Leibovich [1978]).

Faler & Leibovich [1977], Garg & Leibovich [1979] and the author [unpublished studies] have used the non-invasive techniques of laser doppler anemometry to measure the velocity fields both upstream and downstream of breakdown events. Outside a thin boundary layer along the tube wall the experimental data is well fitted by the analytic profiles,

$$V(r) = \frac{1}{r}Q\left(1 - e^{-\alpha r^2}\right) \tag{1.1}$$

$$W(r) = W_1 + W_2 e^{-\alpha r^2} \tag{1.2}$$

W and V being respectively the axial and azimuthal velocity components while W_1, W_2, Q and α are all constants (representative values are given by Garg & Leibovich [1979]). The profiles apply to the downstream flow only in the mean, as the flow there fluctuates with time.

Upstream of the recirculation zone the flow is axisymmetric and steady. After breakdown the vortex core expands to two or three times its upstream size and the constant W_2 in the mean axial velocity profile which had been positive upstream (jetlike flow) becomes negative (wakelike flow). Downstream, within a few vortex core diameters of the breakdown, a turbulent wake is

invariably established . This transition to turbulence "switched on" by the coherent breakdown structure provides a further incentive for its study.

Possibly motivated by the fact that the flows upstream of breakdown can be made to have a high degree of axial symmetry, most of the research to date assumes that axially symmetric processes are the important ones in vortex breakdown. As only axisymmetric disturbances can cause a change in the axial velocity component as measured on the axis and as a deceleration of this component is so pronounced in breakdown, it is clear that such disturbances play an important role. Nevertheless, all transitions occurring in vortex flows as documented by Faler [1976] are nonaxisymmetric and the flow within the bubble itself is unsteady with regular low frequency oscillations. Furthermore, the stagnation point that defines the start of the recirculation zone is not entirely fixed but wanders over a short range of the axis in a seemingly random fashion.

Leibovich [1984] proposed the following plausible scenario for the bubble type breakdown. A finite axisymmetric disturbance, triggered off downstream, moves upstream in a columnar flow that is nearly critical in the sense of Benjamin [1962]. (A supercritical flow, in this classification, allows for the upstream propagation of infinitesimal axisymmetric waves while a subcritical flow does not). Flows of the form (1.1,2) can indeed support axisymmetric dispersive waves and these can propagate upstream in some situations (Leibovich [1970], Randall & Leibovich [1973]). Moving in this direction, the cross sectional area of the tube decreases causing the wave to amplify and speed up. The conjecture is that, upon reaching some critical amplitude, these waves lose stability to a non-axisymmetric disturbance. The growth of the asymmetric mode at the expense of the axisymmetric wave, drains energy from it and this causes the wave to equilibrate at some axial location in the diverging tube, much as is seen in experiments.

Our aim is to study the stability of some inviscid, wavy axisymmetric flows to three dimensional disturbances with the amplitude of the waves viewed as a bifurcation parameter. We start with a columnar flow in cylinderical coordinates of the form $(0,\ V_0(r),\ W_0(r))$, (e.g. (1.1,2)). (For arbitrary C^1 functions, V_0 and W_0, all such flows satisfy Euler's equations). We then seek axisymmetric wavy perturbations to this flow which satisfy the equations of motion, at least approximately, for small amplitude. In terms of a stream function, ψ and a circulation, κ (Lamb [1932]) Leibovich [1972] found solutions to Eulers equations of the form,

$$\psi(r,x) = \psi_0(r) + \epsilon\phi(r)A(x) + O(\epsilon^2), \tag{1.3}$$

$$\kappa(r,x) = \kappa_0(r) + \epsilon\gamma(r)A(x), + O(\epsilon^2), \tag{1.4}$$

where x is a moving coordinate,

$$x = z - dt \tag{1.5}$$

and d is a constant that must found in the calculation. The velocity components are given by

$$u = -\frac{1}{r}\frac{\partial}{\partial z}\psi, \tag{1.6}$$

$$v = \frac{1}{r}\kappa, \tag{1.7}$$

$$w = \frac{1}{r}\frac{\partial}{\partial r}\psi. \tag{1.8}$$

The columnar base flow is represented by $\psi_0(r)$ and $\kappa_0(r)$.

The amplitude function, $A(z,t)$ is governed by a Korteweg de Vries equation which has both infinite and finite period solutions. The multiple scales analysis that was used to obtain these solutions is strictly valid only for long period waves which are also the most interesting solutions from a physical point of view. When doing the stability analysis we will confine our attention, for numerical reasons, to solutions of the finite period, $2H$ and these are given exactly in terms of

cnoidal functions (Whitham [1974]). The structure functions, $\phi(r)$, $\gamma(r)$ and the wave speed d are determined (numerically) from a second order, ordinary differential eigenvalue problem.

For certain base columnar profiles d is negative and consequently the axisymetric wave propagates upstream. Figure 2 is a plot of the streamlines (1.3) in a frame moving with the wave for such a case. The base columnar profile used here and throughout this paper is a purely swirling flow; $W_0(r) \equiv 0$ and $\alpha = 14$ in the notation of (1.1). The structure function ϕ has been normalized so that Max $\phi = 1$ and for this flow a recirculation zone (bubble) first appears in the streamline plot when the amplitude parameter, ϵ reaches a value of 0.0155. For the value of ϵ used here the plot is clearly reminiscent of the bubble type breakdown.

Our aim is to study the stability of the flows (1.3,4) to three dimensional disturbances viewing ϵ as a bifurcation parameter. The analysis will be carried out in a frame moving with the wave, i.e. using the coordinates (r, x, θ). As the base flow is dependent on both the radial and axial variables, r and x, the stability equations separate only in the azimuthal variable, θ. It will be in our interest also to study the stability of a related columnar flow that is constructed by taking the two dimensional flow (1.3,4) at a single axial station, $x = 0$, and extending it throughout the cylindrical domain without variation. For obvious reasons we will refer to this flow as the "mid-bubble" columnar flow and it is given explicitly as follows,

$$V_b(r) = V_0(r) + \frac{\epsilon}{r}\gamma(r), \tag{1.9}$$

$$W_b(r) = W_0(r) + \frac{\epsilon}{r}\phi'(r). \tag{1.10}$$

Plots of these profiles for various values of ϵ are given in Figures 3 and 4. Provided the wavy flow (1.3,4) varies only slowly along the axis (as it will do if the period, $2H$ of $A(x)$ is very large), we can look on the midbubble profiles as models for the full two dimensional flow. We conjecture

that the stability of these columnar flows (the equations for which separate in both x and θ) should also be indicative of the stability properties of the full two dimensional flow.

In the rest of this paper we describe the numerical scheme used to solve the stability equations, we discuss its implementation and verification on the computer and finally we give results obtained for the stability of the midbubble columnar and the axisymmetric wavy flows presented above.

2. Numerical Methods.

For incompressible fluids the physical law of mass conservation reduces to the constraint that the velocity vector of the fluid be divergence free. The pressure is not then a thermodynamic variable determined by an equation of state but rather can be thought of as a Lagrange multiplier adjusting itself instantaneously to ensure that this kinematical constraint on the velocity vector is met. There is no evolution equation for the pressure nor does it satisfy any predetermined boundary or initial conditions.

Numericists, seeking to solve the governing equations approximately, have found that their greatest difficulty lies in the treatment of the pressure variable. While many ingenious methods have been devised to overcome the difficulties, the treatment advocated in this work is in a mathematical sense the most natural and offers many computational advantages. Here, the pressure term is eliminated from the equations entirely and the divergence free condition is satisfied exactly by the numerically obtained approximation to the velocity vector. Moreover, as the components of the velocity are expanded in terms of series of polynomials that arise as the solution to a singular Sturm Liouville problem, whose excellent approximation properties are well

documented (e.g. Gottleib & Orszag [1977], Quarteroni [1983]) convergence of our approximation will be bound only by the smoothness of the solution and by the number of terms used in the component expansions. For infinitely differentiable velocity fields we should expect to achieve "exponential convergence" (Canuto et al.)

The essence of the method (originally due to Leonard & Wray [1982]) involves expanding the velocity in a series of divergence free vector fields each of which satisfy the same boundary conditions as the velocity. The infinite sums are truncated and substituted into the governing equations, which are the Navier-Stokes or Euler equations linearized about the appropriate base flow. Inner products are taken with vectors fields which satisfy inviscid boundary conditions. This eliminates the pressure term from the equations and reduces the differential eigenvalue problem to a matrix eigenvalue problem. The eigenvalues determine the stability of the base flow and the eigenfunctions are the set of coefficents in the expansions of the corresponding perturbation velocity fields.

To examine how this method works we recall that it is well known (Ladyshenskaya [1966]) that $L^2(D)$, the space of square integrable vector functions defined on a bounded domain D ($D \subset R^n$, n = 2,3) can be decomposed into those that are divergence free and whose normal components vanish on the boundary and those that can be expressed as the gradient of a differentiable function defined on D. For this paper we will consider vector fields, defined on the section of a cylinder T, which are periodic in both the axial and azimuthal variables, having as their axial period the tube length, 2H.

We will decompose $L^2(T)$ as follows.

$$L^2(T) = J(T) + J^{\perp}(T), \tag{2.1}$$

where,

$$J(T) = \left\{ \begin{array}{c} \\ \underline{u} \in L^2(T) \\ \\ \end{array} \begin{array}{l} (a)\ \nabla\cdot\underline{u} = 0 \text{ in } T, \\ (b)\ \underline{u}\cdot\underline{n} = 0 \text{ on } \partial T, \\ (c)\ \underline{u}|_{S_1} = \underline{u}|_{S_2}\,. \end{array} \right\} \tag{2.2}$$

S_1 and S_2 represent the ends of the cylinder. Given in this form $J(T)$ is clearly the space of

(a) incompressible, (b) inviscid, (c) periodic velocity fields.

The set of "viscous" velocity fields on T is a subset of $J(T)$ denoted $J^0(T)$.

$$J^0(T) = \left\{ \underline{u} \in J(T) \mid \underline{u} = 0 \text{ on } \partial T \right\}. \tag{2.3}$$

An alternative representation of $J(T)$ (Richtmyer [1978]) is given by,

$$J(T) = \left\{ \begin{array}{c} \\ \underline{u} \in L^2(T) \\ \\ \end{array} \begin{array}{l} (a)\ <\underline{u}, \nabla p> = 0 \text{ for all } p \in C^\infty(\overline{T}) \\ \\ (b)\ \underline{u}|_{S_1} = \underline{u}|_{S_2} \end{array} \right\} \tag{2.4}$$

where $\overline{T}$ is the closure of T and $<\cdot,\cdot>$ represents the usual inner product in $L^2(T)$,

$$<\underline{u},\underline{v}> = \int_T \underline{u}\cdot\underline{v}\, r dr d\theta dx\,. \tag{2.5}$$

The space $J(T)$ endowed with this inner product is a Hilbert space and a closed subspace of $L^2(T)$. The projection of $L^2(T)$ onto $J(T)$ will be denoted by Π. It is clear that vectors of the form ∇p are perpendicular to all $\underline{u}$ in $J(T)$ and in fact (Ladyshenskaya [1966]),

$$J^{\perp}(T) = \left\{ \underline{u} \in L^2(T) \mid \underline{u} = \nabla p \text{ for some } p \text{ in } C^1(\overline{T}) \right\}. \tag{2.6}$$

Π then has the following properties,

$$\Pi : L^2(T) \to J(T), \tag{2.7}$$

$$\Pi \underline{u} = \underline{u} \qquad \text{for all } \underline{u} \in J(T), \tag{2.8}$$

$$\Pi \nabla p = 0 \qquad \text{for all } p \in C^1(\overline{T}). \tag{2.9}$$

To determine the linear stability of a flow $\underline{U}$ to say, viscous disturbances which are periodic in x and θ we consider whether infinitesimal perturbations to $\underline{U}$ grow in time. Therefore we linearize the Navier-Stokes equations about $\underline{U}$ and seek solutions in $J^0(T)$ of the form,

$$\underline{u}(r,x,\theta)e^{-i\sigma t}. \tag{2.10}$$

The character of σ then determines the linear temporal stability of $\underline{U}$. If $\sigma = \alpha + i\beta$ where α, β are real then,

$$\begin{aligned} \beta > 0 \quad &=> \quad \underline{U} \text{ is unstable,} \\ \beta = 0 \quad &=> \quad \underline{U} \text{ is neutrally stable,.} \\ \beta < 0 \quad &=> \quad \underline{U} \text{ is stable} \end{aligned} \tag{2.11}$$

The equations that must be solved have the form,

$$i\sigma \underline{u} \;=\; E\underline{u} \;+\; \mathrm{Re}^{-1} S\underline{u}. \tag{2.12}$$

E and S are operators defined on $J(T)$ as follows,

$$S\underline{u} \;=\; -\Pi(\nabla \times \underline{\omega}) \tag{2.13}$$

and

$$E\underline{u} \;=\; \Pi(\underline{\omega} \times \underline{U} \;+\; \underline{\Omega} \times \underline{u}), \tag{2.14}$$

where $\underline{\Omega}$ and $\underline{\omega}$ are respectively the base and perturbation vorticities, ($\underline{\Omega} = \nabla \times \underline{U}$, $\underline{\omega} = \nabla \times \underline{u}$).

Some suitable nondimensionalization has introduced the Reynolds number,

$$\mathrm{Re} = \frac{U_0 R_0}{\nu}, \tag{2.15}$$

R_0 and U_0 being characteristic length and velocity scales associated with the base flow and ν is

the kinematic viscosity of the fluid. We can take R_0 to be the radius of the tube and U_0 to be the maximum value of the columnar base flow azimuthal velocity, $V_0(r)$.

The ∇p term in the Navier-Stokes equation has been eliminated by projection onto $J(T)$. Projection of the stability equation onto a finite dimensional subspace, $J_N(T)$ of $J(T)$ is achieved in practice by taking the inner product of the equation with basis vectors for $J_N(T)$. This process eliminates the operator Π from the equation, for given any vector $\underline{f}$ in $L^2(T)$ and any vector $\underline{\Delta}$ in $J(T)$ we have that,

$$<\Pi \underline{f}, \underline{\Delta}> \;=\; <\underline{f}, \underline{\Delta}>, \tag{2.16}$$

as projections are self adjoint and as the projection of any vector in $J(T)$ is itself.

It is worth emphasising that even when we are solving the viscous stability equations, we still project the governing equations onto the space of inviscid vector fields. The reason for this is that having found a velocity $\underline{u}$ such that the vector $\underline{f}$ defined as,

$$\underline{f} \;=\; i\sigma \underline{u} \;-\; \underline{\omega}\times\underline{U} \;-\; \underline{\Omega}\times\underline{u} \;+\; \nu \nabla\times\underline{\omega} \tag{2.17}$$

is orthogonal to all $\underline{\Delta}$ in $J(T)$ then $\underline{f} \in J^{\perp}(T)$ and so there exists a scalar function p (a pressure) with $\underline{f} = \nabla p$. If, however, $\underline{f}$ were in $J^{0^{\perp}}(T)$, which contains $J^{\perp}(T)$ then the existence of a pressure is not guaranteed and consequently $\underline{u}$ may not correspond to a physical solution.

Leonard and Wray [1982] demonstrated a divergence free vector function expansion for viscous velocity fields, defined on a cylindrical domain that are Fourier decomposable in both the axial and azimuthal variables. We will construct a somewhat different set of basis vectors here.

The velocity field, $\underline{u}$, satisfies the continuity equation and is Fourier decomposable in x and θ, which means in effect that only two of its three components, u, v, w are independent. This motivates the introduction of two vector families, $\underline{\chi}_n^{\pm}$ in an expansion of the form,

$$\underline{u} = \sum_{nkm} \left\{ a_{2nkm} \underline{\chi}_n^{+}(r) + a_{2n-1km} \underline{\chi}_n^{-}(r) \right\} e^{i(kz + m\theta)} \tag{2.18}$$

The components of the vectors $\underline{\chi}_n^{\pm}$ are found as follows. Expand two of the velocity components, say the first and the third, independently as,

$$u = \sum_{nkm} a_{2n-1km} f_n^{-}(r) \, e^{i(kz + m\theta)}, \tag{2.19}$$

$$w = \sum_{nkm} a_{2nkm} f_n^{+}(r) \, e^{i(kz + m\theta)}, \tag{2.20}$$

where $f_n^{\pm}(r)$ are complete sets of polynomials chosen to satisfy the boundary conditions that are imposed on u, w. The r and x components of $\underline{\chi}_n^{\pm}$ have now been picked and it remains for us to chose the θ components in a manner that ensures the vectors $\underline{\chi}_n^{\pm} \, e^{i(kz + m\theta)}$ are divergence free. Consider for example, $\underline{\chi}_n^{-}(r)$.

$$\nabla \cdot \left(\underline{\chi}_n^{-}(r) \, e^{i(kz + m\theta)} \right) = 0, \tag{2.21}$$

=>

$$(r f_n^{-}(r))' + im \chi_{n,\theta}^{-} = 0, \tag{2.22}$$

where the prime denotes a derivative with respect to r. This equation gives us the θ component of $\underline{\chi}_n^{-}(r)$. Rescaling, it is found that an expansion of the form (2.14) is possible for non-zero azimuthal wavenumbers where,

$$\underline{\chi}_n^{-}(r) = \left(im f_n^{-}(r), - (r f_n^{-}(r))', 0 \right), \tag{2.23}$$

$$\underline{\chi}_n^+(r) = \left(0, -rkf_n^+(r), mf_n^+(r)\right) \tag{2.24}$$

and such an expansion will guarantee that $\underline{u}$ is divergence free. This expansion is clearly incomplete for azimuthal wave number zero, $(m = 0)$, i.e. for axisymmetric flows. For that case the following expansion vectors can be used.

$$\underline{\chi}_n^-(r) = \left(ikf_n^-(r),\ 0, -\frac{1}{r}\left(rf_n^-(r)\right)'\right), \tag{2.25}$$

$$\underline{\chi}_n^+(r) = \left(0,\ f_n^+(r), 0\right) \tag{2.26}$$

The polynomials $f_n^{\pm}(r)$ must be chosen so that the vector $\underline{u}$ given by (2.14) satisfies appropriate (viscous or inviscid) boundary conditions on the walls of the domain, T and is single valued at the origin, $r = 0$. We will denote the polynomials used in the inviscid case by $a_n^{\pm}(r)$ reserving $f_n^{\pm}(r)$ for viscous expansions. We have then upon truncating (2.14) an approximation to $\underline{u}$ of the form,

$$\underline{u}_{NKM} = \sum_{n=1}^{N} \sum_{k=-K}^{K} \sum_{m=-M}^{M} a_{nkm}\, \underline{D}_{nkm}(r,x,\theta) \tag{2.27}$$

where,

$$\underline{D}_{nkm}(r,x,\theta) = \underline{\chi}_n^{\pm}(r;\, k,\, m)\, e^{i(kx + m\theta)}. \tag{2.28}$$

The projection vectors will have the same form as the expansion vectors, i.e. we will project with vectors, $\underline{\Delta}_{lpq}(r,x,\theta)$, where

$$\underline{\Delta}_{lpq}(r,x,\theta) = \underline{\xi}_l^{\pm}(r;\, k,\, m)\, e^{i(px + q\theta)} \tag{2.29}$$

for

$$l = 1,...,N; \quad p = -K,...,K; \quad q = -M,...,M$$

with the vectors $\underline{\xi}_l^{\pm}$ being given by equations (2.23 - 26) using the inviscid polynomials, $a_l^{\pm}(r)$ for

the components.

It is possible to choose the polynomials $a_l^{\pm}(r)$ and $f_n^{\pm}(r)$ in many different ways. Leonard & Wray [1982] in their consideration of certain turbulence simulations employed an unusual set of Jacobi polynomials to reduce the bandwidth of the final matrix system. These polynomials were also used by Spalart [1983] in his simulation of boundary-layer transition. Moser & Moin [1984] in their work on the infinite Taylor Couette system, used Tchebychev polynomials and incorporated the weight function, against which these polynomials are orthogonal, into the projection vectors. Here, we will construct the basis vectors from Legendre polynomials. All of the above sets are solutions to singular Sturm Liouville problems and consequently we can expect expansions in terms of any of these polynomials to exhibit excellent convergence properties.

The single valuedness criterion, which must be applied along the centre line of the tube for the vector $\underline{u}$, causes the polynomials $f_n^{\pm}(r)$ and $a_l^{\pm}(r)$ to depend on m, the azimuthal wavenumber (Joseph [1970]). One appropriate choice for $a_l^{\pm}(r)$ is,

$$\begin{aligned} a_l^{+}(r) &= rP_l(2r-1) && \text{for all } m, \\ a_l^{-}(r) &= (1-r)P_l(2r-1) && \text{if } |m| = 1, \\ a_l^{-}(r) &= r(1-r)P_l(2r-1) && \text{if } |m| \neq 1, \end{aligned} \tag{2.30}$$

where the radial variable has been scaled by the tube radius and $P_l(r)$ is the Legendre polynomial of order l (Abramowitz & Stegun [1970]). The corresponding choice for $f_n^{\pm}(r)$ is,

$$\begin{aligned} f_n^{+}(r) &= r(1-r)P_n(2r-1) && \text{for all } m, \\ f_n^{-}(r) &= (1-r)^2P_n(2r-1) && \text{if } |m| = 1, \\ f_n^{-}(r) &= r(1-r)^2P_n(2r-1) && \text{if } |m| \neq 1. \end{aligned} \tag{2.31}$$

As all of our stability problems separate in the azimuthal direction, this dependence on m presents no difficulty. We solve separate problems for each azimuthal wavenumber; so having chosen an m the expansion and projection sets are fixed throughout the calculation . Indeed, in theory there is no difficulty even if the problem at hand is truly three dimensional; however some care is required in implementing the method to ensure that the correct radial polynomial set is being used for each azimuthal component of the velocity.

3. Implementation and Verification of the Method.

Equation (2.12) is solved approximately by using the expansion $\underline{u}_{NKM}$ for $\underline{u}$ and taking inner products of the equation with the projection vectors, $\underline{\Delta}_{lpq}$ to get a generalized matrix eigen problem for the eigenvalues σ and the eigenvectors a (the coefficents in the expansion $\underline{u}_{NKM}$). This matrix problem can be written as,

$$\sigma \overline{A} a = \left(\overline{B} + \frac{i}{\mathrm{Re}} \overline{C} \right) a. \tag{3.1}$$

The matrix $\overline{A}$ is purely real and arises from the fact that the expansion and projection vectors are not orthonormal.

$$\overline{A}_{lnpkqm} = <\underline{\Delta}_{lpq}, \underline{D}_{nkm}>, = <\underline{\xi}_l, \underline{\chi}_n>\delta_{pk}\delta_{qm}, \equiv A_{\ln}\delta_{pk}\delta_{qm}, \tag{3.2}$$

The Kronecker delta symbol, δ_{ij} arises because the Fourier bases employed in the axial and azimuthal directions are orthogonal. The matrix $\overline{B}$ arising from the convection terms is also purely real.

$$\overline{B}_{lnpkqm} = <\underline{\Delta}_{lpq}, (\underline{\nabla}\times\underline{D}_{nkm})\times\underline{U} + \underline{\Omega}\times\underline{D}_{nkm}>. \tag{3.3}$$

Finally, the matrix $\overline{C}$ arising from the viscous terms is purely imaginary.

$$\overline{C}_{lnpkqm} = <\underline{\Delta}_{lpq}, \underline{\nabla}\times\underline{\nabla}\times\underline{D}_{nkm}>. \tag{3.4}$$

Using the orthogonality of the Fourier bases it can be written as,

$$\overline{C}_{lnpkqm} = C_{\ln}\delta_{pk}\delta_{qm}. \tag{3.5}$$

The form of the matrix $\overline{B}$ depends on the base flow $\underline{U}$. For columnar flows which are independent of x and θ it is possible to find a matrix B such that,

$$\overline{B}_{lnpkqm} = B_{\ln}\delta_{pk}\delta_{qm}. \tag{3.6}$$

The stability of these flows can be determined by solving the $O(N)$ generalized matrix eigen problem,

$$\sigma A a = \left(B + \frac{i}{\mathrm{Re}} C \right) a. \tag{3.7}$$

On the other hand, for the axisymmetric wavy base flow (1.3,4) we have

$$\overline{B}_{lnpkqm} = \hat{B}_{lnpk}\delta_{pk} \tag{3.8}$$

and the stability equation is the $O\Big((2K+1)\times N \Big)$ matrix equation,

$$\sigma \hat{A} a = \left(\hat{B} + \frac{i}{\mathrm{Re}} \hat{C} \right) a, \tag{3.9}$$

where $\hat{A}$ and $\hat{C}$ are the $O\Big((2K+1)\times N \Big)$ block diagonal matrices $A_{\ln}\delta_{pk}$ and $C_{\ln}\delta_{pk}$ respectively. The matrices depend parameterically on the wavenumbers of the projection and expansion vectors so we separate them into submatrices that can be evaluated independently of these and the other parameters (in particular ϵ) occurring in the base flow. The submatrices are evaluated once and then stored in the computer. The required integrations can be done at very little cost by utilizing the orthogonality properties of the expansion and projection polynomials. The full matrices are then be reassembled without the need for doing any further integrations.

One can always band the A and C matrices by appropriate choice of the polynomials $f_n^{\pm}(r)$ and $a_l^{\pm}(r)$. However the matrix B will generally be full, though for certain rather simple base flows such as the Hagen Pouiseille flow considered be Leonard & Wray [1982] it is also possible to band B. The matrix A was inverted to produce a regular eigenvalue problem in place of (3.1) and the QR algorithm was used to extract the eigenvalues. We also note that the matrix problem we get when considering the inviscid stability of base flows is a purely real one and consequently the eigenvalues occur, as they should do, in conjugate pairs.

A computer code has been written which implements the method we have been describing to solve the stability problems, both viscous and inviscid, for all columnar flows and for axisymmetric flows of the form (1.3,4). Both the direct and adjoint versions of the stability problems can be solved. The adjoint viscous stability problem is to find a $\underline{u}$ in $J^0(T)$ such that

$$i\lambda \underline{u} \;=\; E^* \underline{u} \;+\; \frac{1}{\mathrm{Re}} S\underline{u}. \tag{3.10}$$

The operator E^* is the adjoint operator to E and is given by,

$$E^* \underline{u} \;=\; -\Pi\Big(\underline{\Omega}\times\underline{u} \;+\; \underline{\nabla}\times(\underline{u}\times\underline{U})\Big) \tag{3.11}$$

The direct and adjoint spectra obtained by solving (2.12) and (3.10) should, of course, be conjugate to each other and how well a numerical scheme reproduces this theoretical result is a test of its accuracy.

We verified the code by calculating the stability of rotating Poiseuille flow,

$$\underline{U} \;=\; \Big(\;0,\quad V_1 r,\quad W_1(1-r^2)\;\Big) \tag{3.12}$$

Cotton et al. [1980] found that this flow with $V_1 = 0.2147$ and $W_1 = 1.0$ was neutrally stable to disturbances having azimuthal wavenumber, $m = 1$ and axial wavenumber, $k = -1$ for a

Reynolds number of 156. The following table lists the most unstable eigenvalue we found for this flow with the same wavenumber pair for the disturbance. The first column of the table gives N, the number of radial basis vectors that were used to obtain the eigenvalue given in the next two columns, N is also the order of the matrix problem that needs to be solved at each step.

Most unstable eigenvalue found for the rotating Poiseuille flow (3.12) with $m = 1$, $k = -1$, $V_1 = 0.2147$, Re = 156.0.

N	frequency	growth rate
4	-0.00029	.00334
6	-0.00279	.00101
10	-0.00284	.000001
14	-0.002847	.0000001
18	-0.002847898	.0000001379
22	-0.002847898	.0000001378

The convergence is exponential in N or some power of N and there is no evidence of significant roundoff error. The following table lists the corresponding eigenvalue found by solving the adjoint viscous stability problem for the same wavenumber pair and baseflow.

Eigenvalue found by doing the adjoint viscous stability problem for the flow (3.12), with $m = 1$, $k = -1$, $V_1 = 0.2147$, Re = 156.0.

N	frequency	growth rate
4	-0.00270	-.000094
6	-0.00280	-.000013
10	-0.00284	-.000004
14	-0.002847	-.0000002
18	-0.002847898	-.0000001379
22	-0.002847898	-.0000001379

Clearly the agreement between the adjoint and direct results is excellent. Inviscid stability results for flows of the form (3.12), obtained using our code also compare well with results in the

literature. These results instill confidence in the accuracy of the numerical method and in the code that implements it, at least for the case of columnar flows.

4. Stability Results for the Vortex Breakdown Model Flows.

In this section we will present the results obtained to date for the stability of the midbubble columnar, (1.9,10) and the wavy vortex (1.3,4) flows. Although these flows are inviscid we will consider their stability to both viscous and inviscid disturbances (i.e. we will solve the linearized Euler and the linearized Navier-Stokes equations for these flows). The justification for doing a viscous analysis is that the "real" flow is of course, viscous. Moreover, the inclusion of the higher order dissipative terms eliminates certain technical difficulties that arise due to the existence of critical layers in the neutrally stable eigenfunctions for columnar flows (Drazin & Reid [1981]).

We will begin by presenting the viscous results for the midbubble columnar flows. We found that thirty radial vector modes ($N = 30$) were adequate to resolve the most unstable eigenmode (i.e. the mode whose eigenvalue had the largest imaginary part) to three decimal places for these flows at low Reynolds numbers and that this number increased as the Reynolds number grew. Frequent checks were carried out on the accuracy of the computed eigenvalues both by increasing the order of the expansion and also by computing the adjoint spectrum for the same set of flow parameters. The difference between the most unstable eigenvalue as computed by the direct and adjoint versions of the code was always less than 1%.

Having fixed the number of radial expansion vectors in our system the viscous eigenvalues for the midbubble flows depend on four parameters,

$$\sigma = \sigma(m,k,\epsilon,Re). \tag{4.1}$$

The base columnar flow ($\epsilon = 0$) was found to be stable to all disturbances. It seems that even for very small values of ϵ (values for which there is no recirculation zone in the full two dimensional flow) the midbubble flows are unstable. This is documented in the following table which gives bracketing values for the critical Reynolds number for various values of ϵ.

Bracketing values for the critical Reynolds number. Various values of ϵ and $m = -1$.

ϵ	Stable for Re	Unstable for Re
0.000	Stable for all Re	
0.005	550	600
0.010	180	200
0.015	160	180
0.020	110	120
0.025	60	80
0.030	42	45

For large enough values of ϵ disturbances having both positive and negative azimuthal wavenumbers can destabilize the midbubble flows with the negative azimuthal modes giving rise in general to the largest values for the growth rates. In particular disturbances with azimuthal wavernumber, $m = -1$ were found to be the most dangerous. This is shown in the following table.

Bracketing values for the critical Reynolds number. Various values of m with $\epsilon = .03$.		
m	Stable for Re	Unstable for Re
-1	42	45
-2	90	100
-3	700	800
-4	1200	1400
-5	1400	1600

For fixed values of m and ϵ, a two parameter (k, Re) study was carried out. With $\epsilon = .03$ and $m = -1$ we obtain the stability diagram shown in Figure 5. The stability boundary appears to be a parabolic curve which is markedly asymmetric with respect to the $k = 0$ line. Within this curve the base flow is unstable to a range of axial wavenumbers; however, there is a "tongue" of stable wavenumbers that gradually thins out as the Reynolds number is increased. The $k = 1$ mode is the final one to be excited, this does not happen until Re = 4500 (approximately).

We now consider the stability of the midbubble flows to inviscid disturbances. The inviscisid stability of columnar flows is governed by an ordinary differential equation, the Howard-Gupta [1962] equation. A number of analytic results obtained from this equation exist in the literature. Leibovich & Stewartson [1982] showed that a sufficent condition for the instability of a columnar flow is that the function,

$$F(r) \equiv V(r)\Lambda'(r)\Big(\Lambda'(r)\Gamma(r) + W'(r)^2 \Big) \tag{4.2}$$

be negative somewhere in the domain of interest. (Λ is the angular velocity, $\frac{1}{r}V$ and Γ is the circulation rV.)

Th function $F(r)$ is easily evaluated for the midbubble profiles and this has been done for a range of values of ϵ. It was found that F first becomes negative only when a critical value of ϵ is reached, $\epsilon = 0.132$. Consequently the midbubble flows are guaranteed to be unstable by the Leibovich-Stewartson criterion for values of the parameter ϵ slightly below those needed to produce a recirculation region in the full two dimensional flow (recall this happens for $\epsilon = 0.155$).

A normal mode stability analysis was carried using the divergence free expansion method and the numerically obtained results confirm and somewhat extend the predictions of the Leibovich-Stewartson criterion. Once again we found that the disturbances giving rise to the largest growth rates had azimuthal wavenumbers, $|m| = 1$. Figure 6 shows how the maximum growth rate varies (almost linearly) with the amplitude parameter, ϵ. In the figure we can see that the normal mode analysis extends the previously obtained results in that midbubble flows with $\epsilon < 0.0132$ are also found to be unstable, even though $F(r) > 0$ for these flows. The maximum growth rates also increase with the size of the axial wavenumber, $|k|$, as shown in Figure 7.

We conclude that the midbubble flows are definitely unstable on both viscous and inviscid grounds for values of the parameter, ϵ below those needed to produce a reversed flow region in the full two dimensional wavy flow. Moreover the most destabilizing disturbances have azimuthal wavenumbers, $|m| = 1$ and short axial period, ($|k| > 1$). The unstable inviscid eigenfunctions tended to have regions of steep gradient near the origin and this made their resolution difficult.

While these midbubble stability results we have just reported tend to support the conjectures made about vortex breakdown, they are not encouraging for the numericist seeking to investigate the stability of the cnoidal wave flows (1.3,4). As we noted earlier, the stability equation

for these flows do not separate in x and consequently we have to solve $O(N(2K+1))$ matrix eigenvalue problems for each flow. It would seem to be necessary to include modes having a short axial period in our expansion, $\underline{u}_{NKM}$ which means that K will be large. The inclusion of these modes is also dictated by physical considerations. We should like the disturbance to include modes that scale with the dimensions of the bubble and in fact visualization studies indicate that the asymmetric unstable modes do have short axial periods. However, our experience with the midbubble flows show that these unstable modes are difficult to resolve radially, consequently the number of radial vector functions, N in our expansion must also be large.

With these constraints the normal mode analysis of the cnoidal wave flows becomes prohibitively expensive, (reacall that the number of operations needed to extract all the eigenvalues of a matrix is proportional to the cube of its order). To alleviate the cost problems most of the runs for these flows were done with the axial period of the cnoidal wave fixed at 1 ($H = 0.5$, units are tube radii). While such short period flows violate the assumptions needed to produce the solutions (1.3,4) it was hoped that these flows would be unstable to disturbances with smaller values for K. Indeed convergence studies indicate that adequate resolution in the axial direction was obtained with $K \approx 10$. Up to 40 radial modes were used in the study, leading to an $O(840)$ matrix eigenvalue problem when $K = 10$. (The calculations were performed on an Floating Point Systems 164 series vector processor at Cornell University, using a vectorized version of the QR algorithm, optimized for this machine and using some 16 megabytes of memory.) The adjoint problem was also solved on each run and only those eigenvalues which agreed well between the adjoint and direct calculations were considered.

Unfortunately the short period flows behave less like the midbubble flows and more like the underlying, stable columnar flows. This is indicated in Figure 8 which plots the least stable

growth rates found for the various values of ϵ as the Reynolds number is increased ($m = -1$ in this plot). All these short period cnoidal wave flows are stable, though marginally so. The disturbances with $|m| = 1$ are again the most unstable. Figure 9 shows the least stable growth rates found for some other azimuthal wavenumbers. The inviscid stability runs that were performed also failed to turn up any definite evidence of instability for these flows.

We have considered the viscous and inviscid stability of some axisymmetric flows which are said to model the bubble type of vortex breakdown. The investigation was carried out by expanding the perturbation velocity in terms of the new set of divergence free vectors presented in section 2. The stability results for the midbubble flows support the conjectured mechanism for breakdown and it was found that the most dangerous disturbances have a short axial period and azimuthal wavenumbers, $|m| = 1$. The two dimensional flows we considered all had rather short axial periods and these do not model the physical phenomena well. No conclusions as to the stability of these flows can be drawn because the normal mode analysis used here can only prove instability (by finding a growing disturbance). No evidence of instability was found but this may well be because we failed to include enough axial modes in our expansion for the disturbance. The study clearly points out that linear stability investigations for complex base flows are far from trivial from a computational point of view.

5. Acknowledgements

The author is happy to acknowledge the assistance of Professors Philip Holmes and Sidney Leibovich. Computations were carried out on equipment at Cornell University.

6. Bibliography

(1) Abramowitz, M. and Stegun, I., eds., 1970. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.* 10th ed. U.S. Gov. Printing Office.

(2) Benjamin, T.B., 1962. The theory of the vortex breakdown phenomenon. *J. Fluid Mech.* 14, 593.

(3) Canuto, C., Hussaini, M.Y., Quarteroni, A. and Zang, T.A. (To appear). *Spectral Methods with Applications to Fluid Dynamics.* Springer-Verlag, New York.

(4) Drazin, P. and Reid, W., 1981. *Hydrodynamic Stability.* Cambridge University Press.

(5) Faler, J.H. and Leibovich, S., 1977. Disrupted states of vortex flow and vortex breakdown. *Phys. Fluids* 20, 1385.

(6) Garg, A.K. and Leibovich, S., 1979. Spectral characteristics of vortex flow fields. *Phys. Fluids* 22, 2053.

(7) Hall, M.G., 1972. Vortex breakdown. *Ann Rev. Fluid Mech.* 4, 195.

(8) Howard, L.N. and Gupta, A.S., 1962. On the hydrodynamic and hydromagnetic stability of swirling flows. *J. Fluid Mech.* 14, 589.

(9) Gottlieb, D. and Orszag, S., 1977. *Numerical Analysis of Spectral Methods: Theory and Applications. CBMS-NSF Regional Conference Series on Applied Mathematics,* Vol. 26. SIAM, Philedelphia.

(10) Joseph, D.D., 1970. *Stability of Fluid Motions: Volume I. Springer Tracts in Natural Philosophy* Vol. 27, Springer-Verlag, New York.

(11) Ladyshenskaya, O. A., 1969. *The Mathematical Theory of Viscous Incompressible Flow.* Gordon and Breach, New York.

(12) Leibovich, S., 1970. Weakly nonlinear waves in rotating fluids, *J. Fluid Mech.* 42, 803.

(13) Leibovich, S., 1978. The structure of vortex breakdown. *Ann. Rev. Fluid Mech.* 10,221.

(14) Leibovich, S., 1984. Vortex stability and breakdown: Survey and extension. *AIAA J.* 22, 1192.

(15) Leibovich, S. and Stewartson, K., 1983. A sufficent condition for the instability of columnar vortices. *J. Fluid Mech.* 126, 335

(16) Leonard, A. and Wray, A., 1982. A new numerical method for simulation of three dimensional flow in a pipe. *Proc. International Conference on Numerical Methods in Fluid Dynamics,* 8th, Aachen. *Lecture Notes in Physics,* Vol. 170 (ed. E. Krause). Springer-Verlag, New York, pp. 335-342.

(17) Mac Giolla Mhuiris, N., 1986. *Numerical Calculations of the Stability of Some Axisymmetric Flows Proposed as a Model for Vortex Breakdown.* Ph.D. Dissertation, Cornell Univ.

(18) Moser, R.D. and Moin, P., 1984. *Direct Numerical Simulation of Curved Turbulent Channel Flow. NASA TM* 85974.

(19) Peckham, D. and Atkinson, S.A., 1957. Preliminary results of low speed wind tunnel tests on a Gothic wing of aspect ratio 1.0. *Aeronaut. Res. Counc. CP* 508.

(20) Quarteroni, A., 1983. Theoretical motivations underlying spectral methods. *Proc. Meeting INRIA - Novosibirsk,* Paris.

19

(21) Richtmyer, R.D., 1978. *Principles of Advanced Mathematical Physics: Volume I. Texts and Monographs in Physics,* Springer-Verlag, New York.

(22) Randall, J.D. and Leibovich, S., 1973. The critical state: a trapped wave model of vortex breakdown. *J. Fluid Mech.* 53, 495.

(23) Salwen, H., Cotton, F.W. and Grosch, C.E., 1980. Linear stability of Poiseuille flow in a circular pipe. *J. Fluid Mech.* 98, 273.

(24) Spalart, P.R., 1983. Numerical simulation of boundary layer transition. *Proc. International Conference in Fluid Dynamics,* 9th. *Lecture Notes in Physics,* Vol. 218 (ed. H. Akari). Springer-Verlag, New York, pp 531-535.

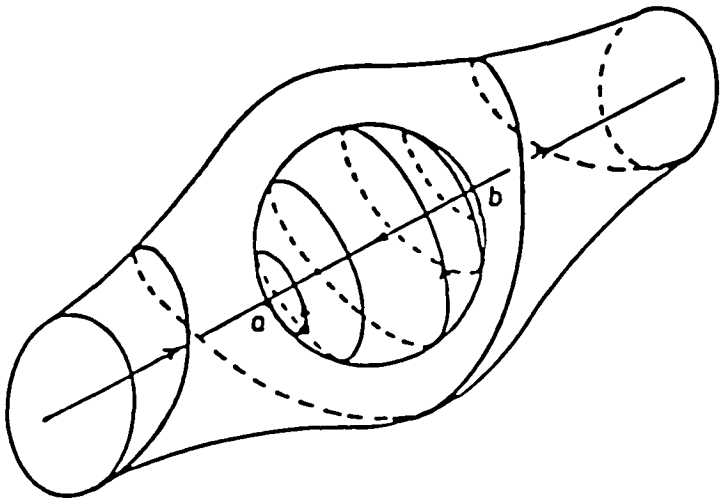


Figure 1. Axisymmetric bubble type vortex breakdown (after Leibovich [1978]).

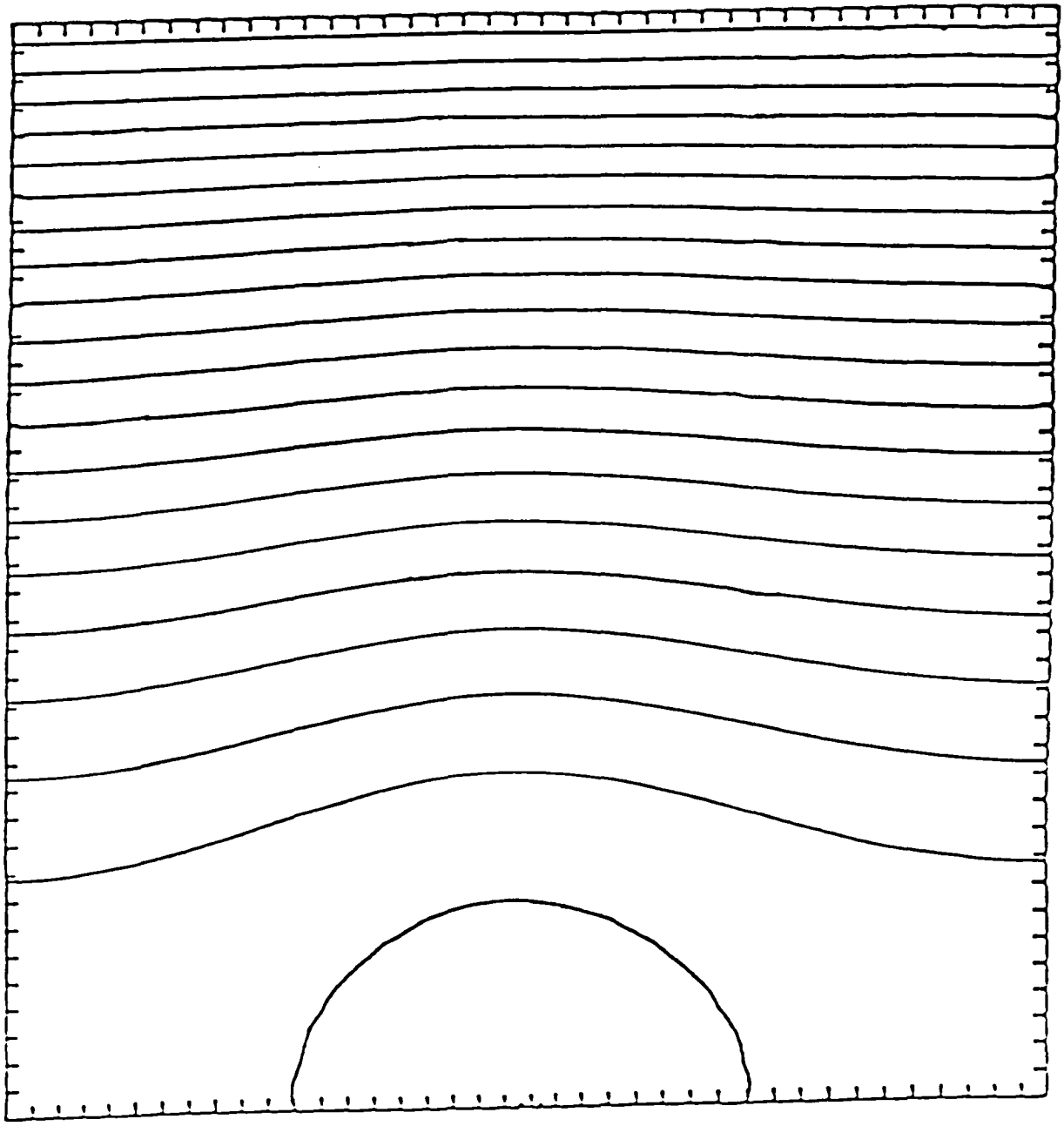


Figure 2. Streamlines (1.3), $\epsilon = .02$, $H = \pi$.

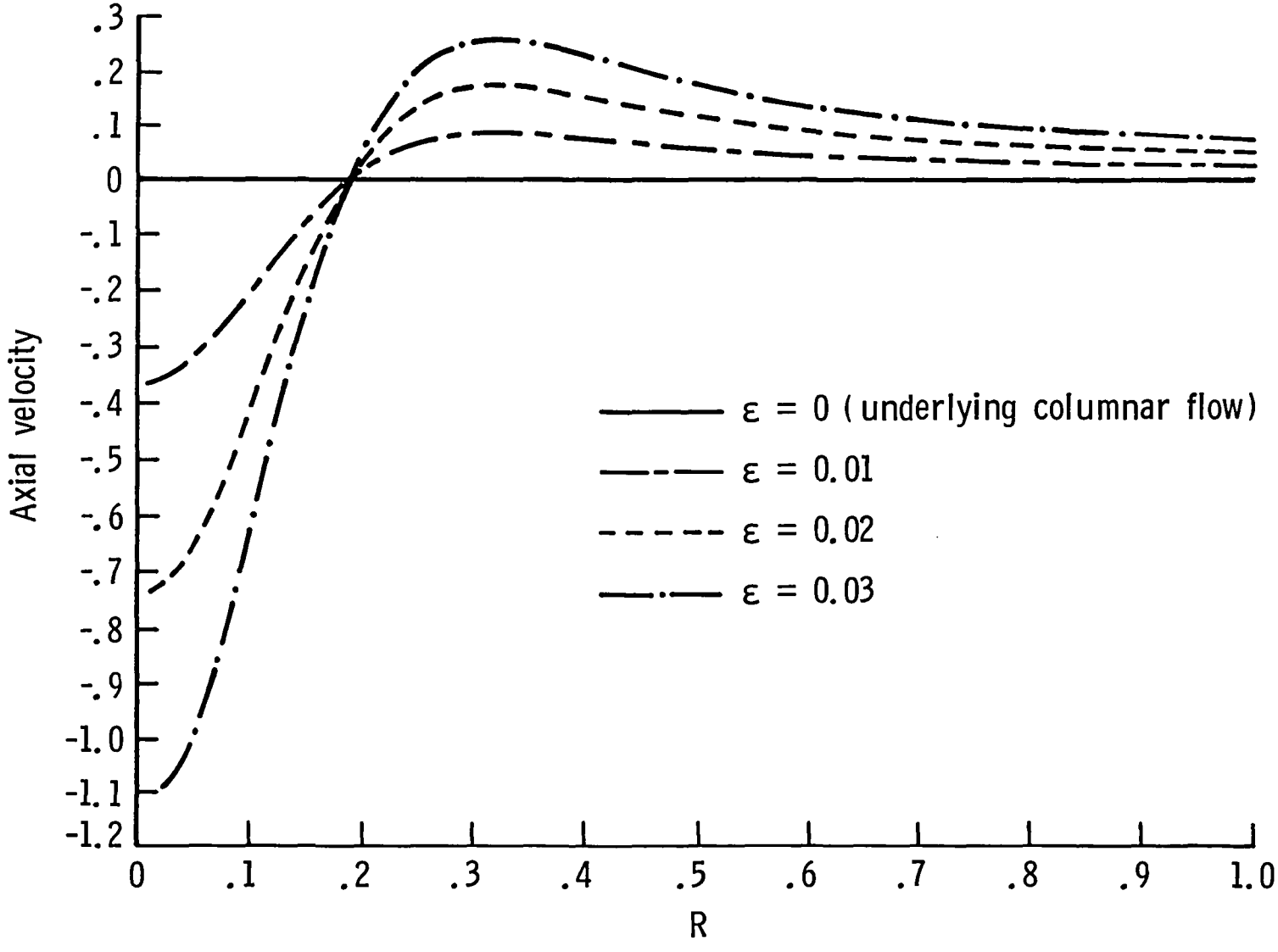


Figure 3. Axial velocity profiles for the midbubble flow.

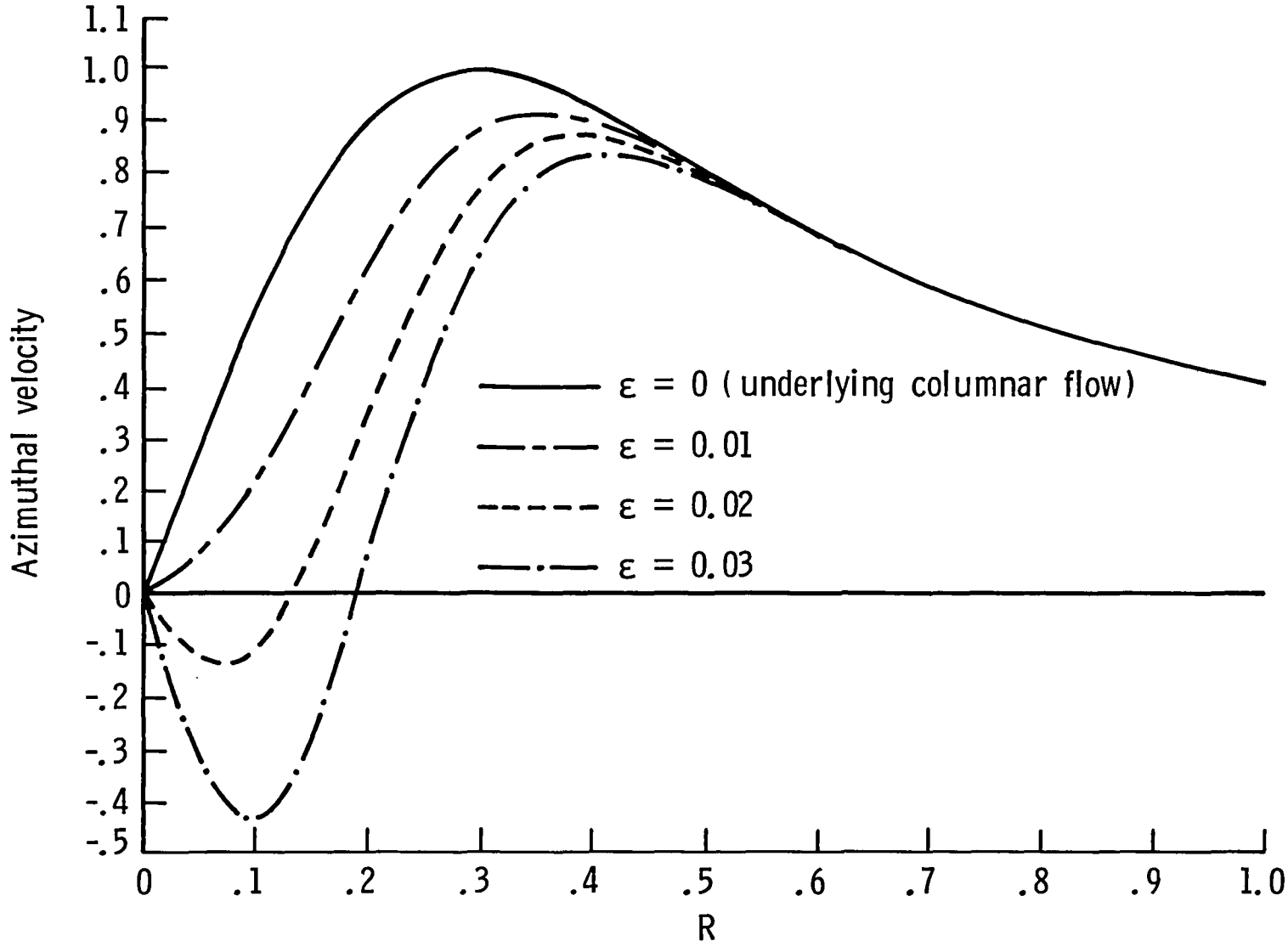


Figure 4. Azimuthal velocities for the midbuble flows.

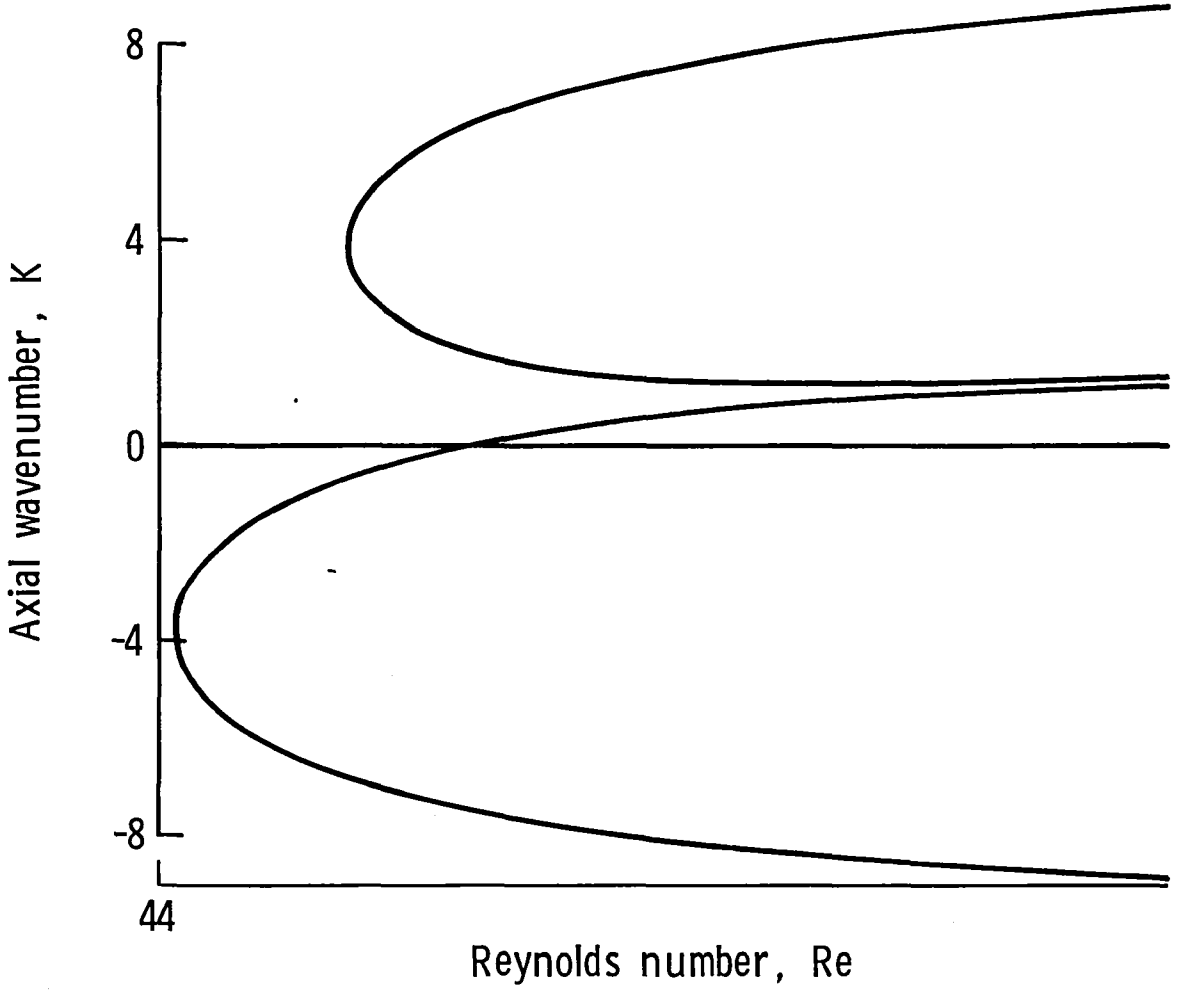


Figure 5. Stability diagram for the midbubble flow with $\epsilon = .03$ and $m = -1$.

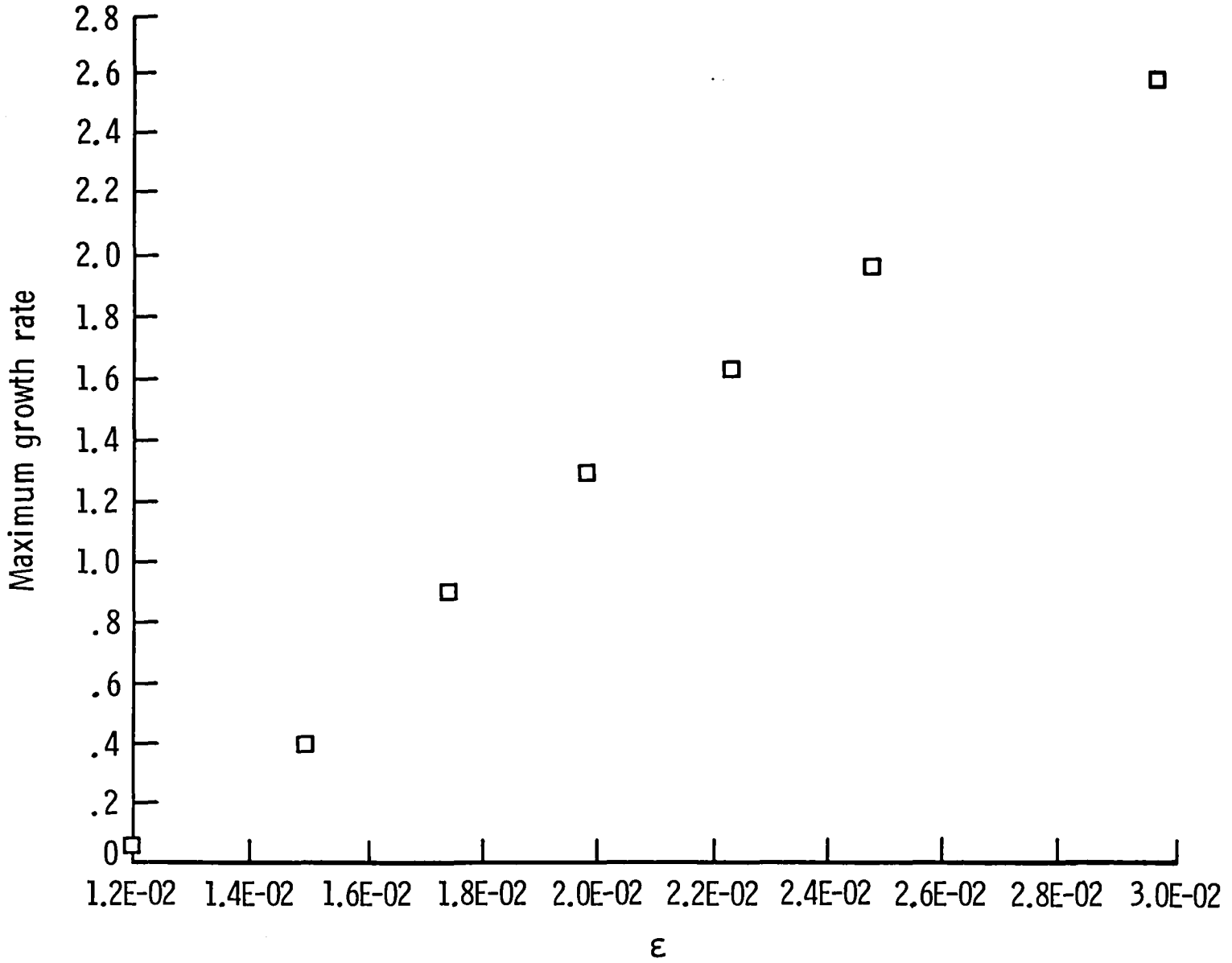


Figure 6. Maximum growth rates plotted against ϵ, $m = -1$ and $k = 6$.

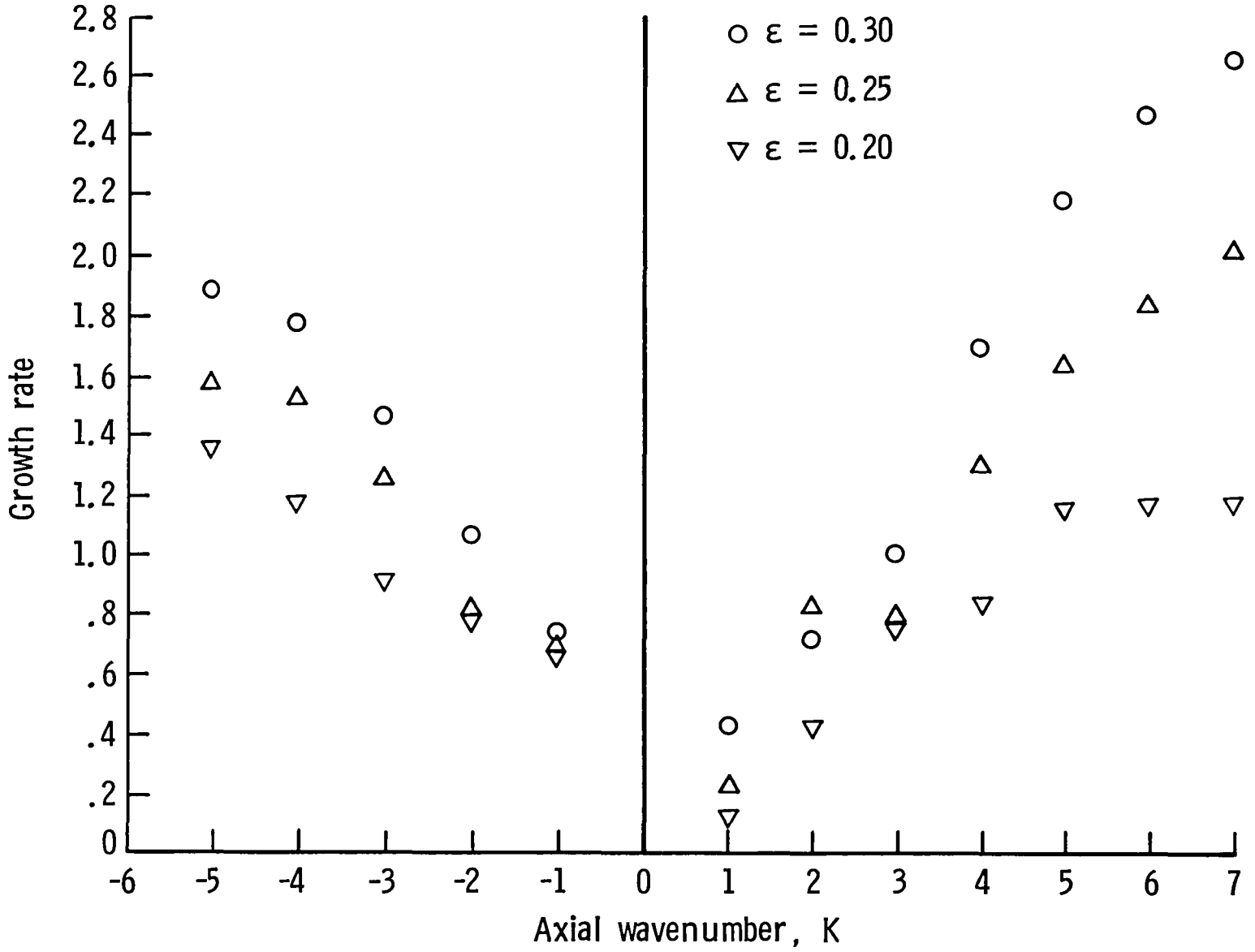


Figure 7. Maximum growth rates plotted against axial wavenumber, (inviscid analysis for the midbubble flow), $m = -1$.

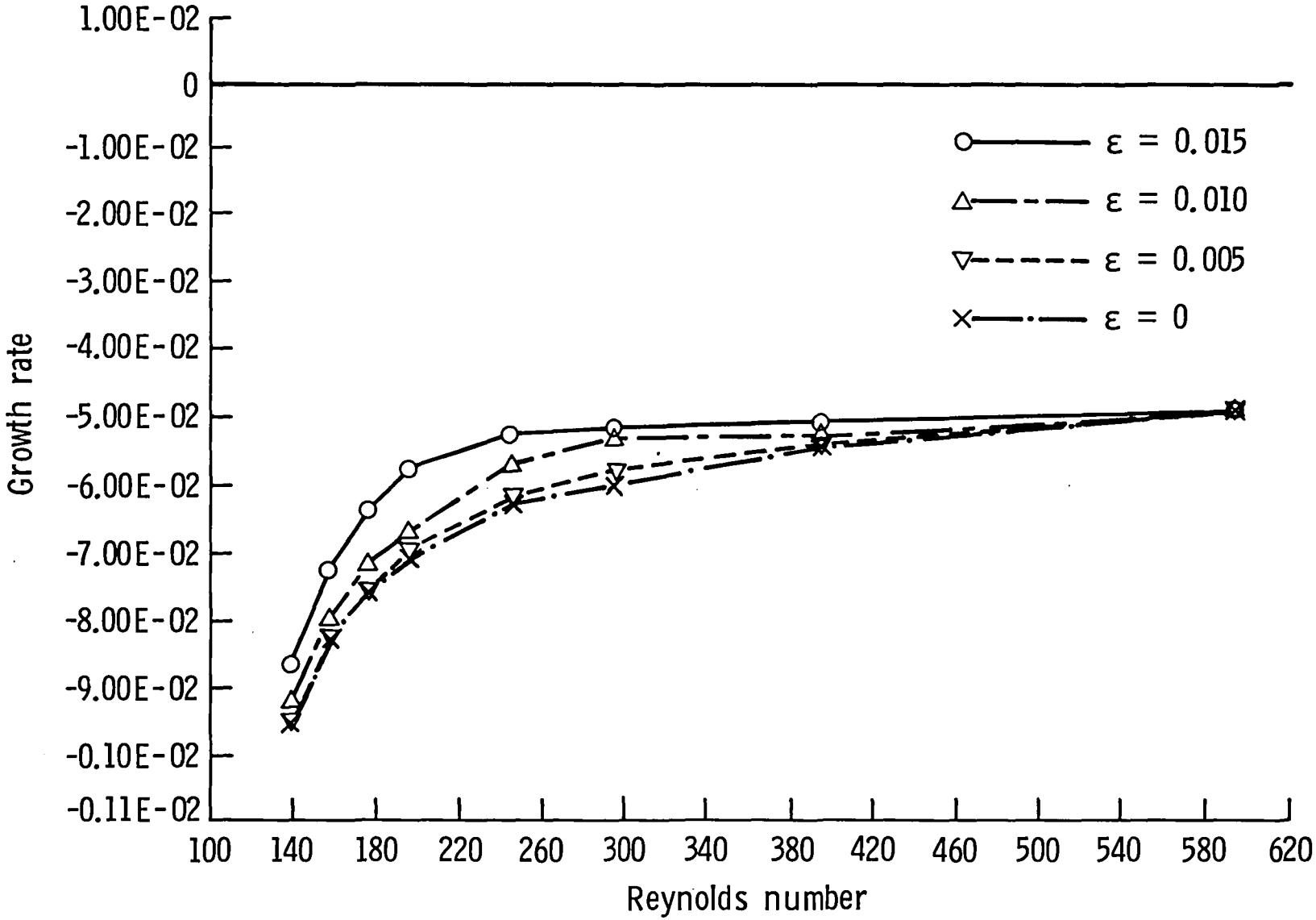


Figure 8. Least stable growth rates found for the full two dimensional flow (1.3,4), $H = .5$, $m = -1$.

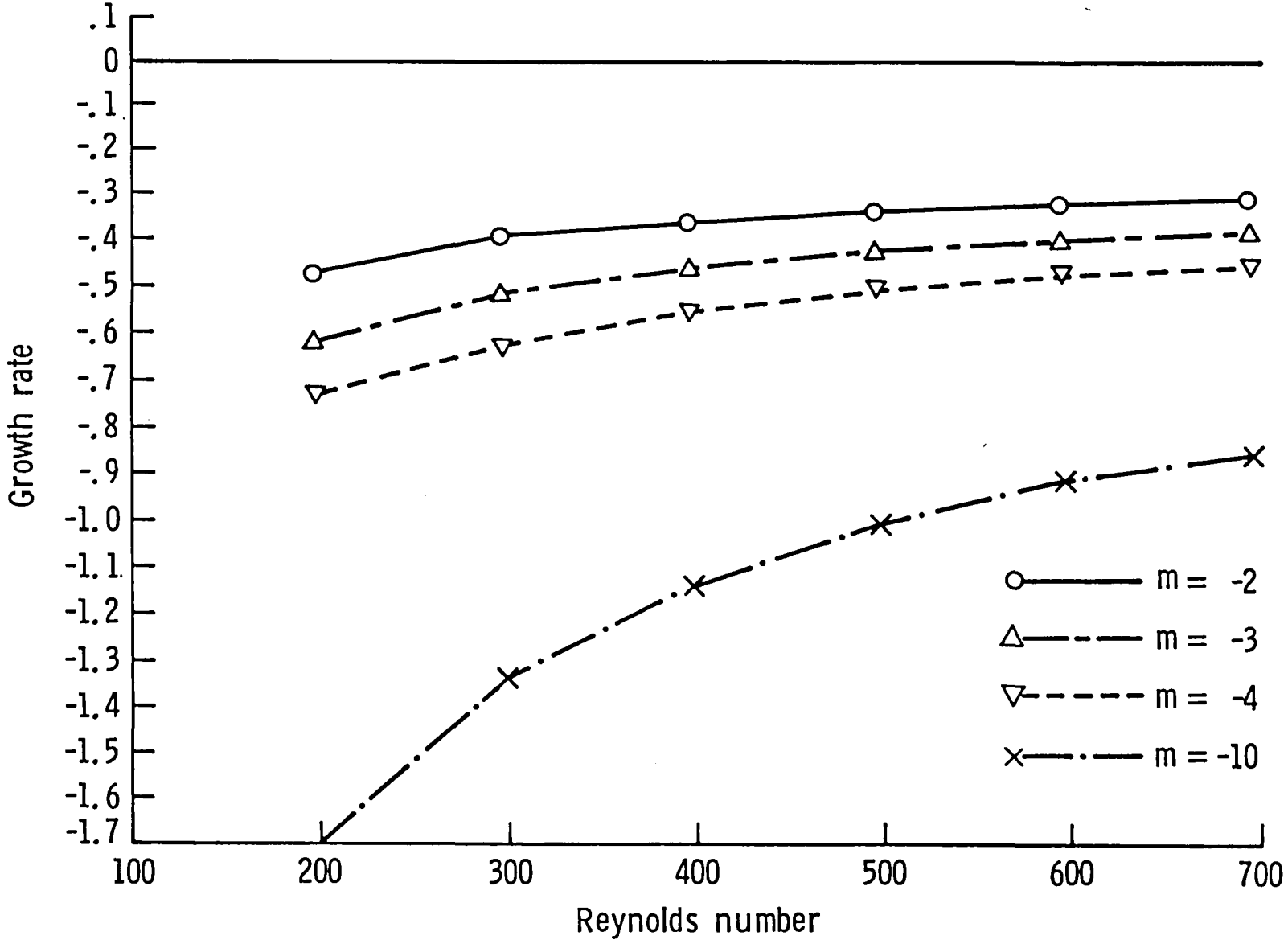


Figure 9. Least stable growth rates found for various values of m, $H = 0.5$, $m = -1$.

NUMERICAL STUDY OF VORTEX BREAKDOWN

M. Hafez
G. Kuruvila
Vigyan Research Associates, Inc.

and

M. D. Salas
NASA Langley Research Center

Abstract

The incompressible axisymmetric steady Navier-Stokes equations and the Euler equations are solved numerically to model the breakdown of a vortex. The solutions obtained for the Euler equations show a "vortex breakdown-like" structure, their behavior is very different from that of the Navier-Stokes solution which are obtained at low Reynolds number. The details of the numerical algorithms used are presented, and the results obtained are compared to those in the literature at the same Reynolds number.

Research was supported by the National Aeronautics and Space Administration for the first and second author under NASA Contract No. NAS1-17919.

1. INTRODUCTION

Under certain conditions, it has been observed that the vortex shed from the highly swept leading edges of a delta wing can change its structure abruptly. The change is characterized by either a spiral deformation of the vortex axis or the formation of a stagnation point along the vortex axis followed by a bubble of recirculating flow. Downstream of this structural change, the flow appears to be highly sensitive to perturbations and is usually turbulent. This sudden change in structure is known as vortex breakdown. The effect of vortex breakdown on the aerodynamics of a wing is very important, since it degrades the performance of the wing and can set a limit on the maximum attitude achievable by the wing. The phenomenon has been studied, both experimentally and theoretically, for the last 30 years, but no really satisfactory theory exists to explain it. The reader is referred to the two reviews of the subject given by Leibovich [1,2].

Our interest in vortex breakdown was aroused by the claim of Hitzel and Schmidt [3] that vortex breakdown could be predicted on the basis of the Euler equations. We consider that the numerical studies of flow over a delta wing by Hitzel and Schmidt are too superficial to warrant such a conclusion. The flow over a delta wing at high angles of attack is too complex and requires too many computational resources to allow an indepth study. How could the problem be formulated such that it would lend itself to an investigation of the relevance of the Euler equations vis-a-vis the Navier-Stokes equations? A drastic simplification of the problem is required. Fortunately, experimentalists have already achieved this by studying vortex breakdown within the confines of cylindrical tubes. In addition, two numerical investigations of the Navier-Stokes equations have been presented [4,5] for this problem; in one

case [4] steady solutions were obtained, while in the other [5] the solutions appeared unsteady. Perhaps an additional investigation could shed some light onto this problem. The purpose of this work is, therefore, to investigate the possibility of simulating vortex breakdown with the Euler equations, studying the relation of these solutions to those of the Navier-Stokes equations and comparing the latter to those in Refs. 4 and 5.

2. MATHEMATICAL FLOW MODELS

An incompressible steady axisymmetric flow with swirl can be described in terms of a streamfunction, ψ; azimuthal vorticity component, ω; and a circulation, κ. In cylindrical coordinates (x,r,θ) the Navier-Stokes equations are:

$$r\left(\frac{\psi_r}{r}\right)_r + \psi_{xx} = r\omega$$

$$(u\omega)_r + (w\omega)_x + \left(\frac{\kappa^2}{r^3}\right)_x = \frac{1}{R_e}\left(\omega_{rr} + \left(\frac{\omega}{r}\right)_r + \omega_{xx}\right)$$

$$u\kappa_r + w\kappa_x = \frac{1}{R_e}\left(\kappa_{rr} - \frac{1}{r}\kappa_r + \kappa_{xx}\right) \qquad (2.1)$$

where $\kappa = rv$, $\omega = w_r - u_x$, and R_e is the Reynolds number defined in terms of the free-stream axial velocity, the vortex core radius, and the kinematic viscosity of the flow. The velocity components in the x,r,θ directions are denoted by w, u, and v, respectively. In terms of the streamfunction, w and u are given by:

$$w = \frac{\psi_r}{r}$$

$$u = -\frac{\psi_x}{r} \tag{2.2}$$

The inviscid equations are obtained from Eqs. (2.1) by letting $R_e \to \infty$. In the inviscid limit, it is clear that the circulation becomes constant along a streamline. It can also be shown that the total enthalpy, h, becomes constant along streamlines. Moreover, the vorticity component ω can be related to the gradients of the circulation and the total enthalpy by:

$$r\omega = r^2 \frac{dh}{d\psi} - \frac{1}{2} \frac{d(\kappa^2)}{d\psi} \tag{2.3}$$

where the total enthalpy is given by:

$$h = p + \frac{1}{2} (u^2 + v^2 + w^2) \tag{2.4}$$

and p is the pressure. Notice that in the absence of swirl ($v = 0$), ω/r becomes constant along a streamline. We also notice that the contribution to the vorticity (given by eqn. (2.3)) due to circulation term does not depend on the sign of κ but only on its magnitude. The functions $\kappa(\psi(x,r))$ and $h(\psi(x,r))$ are determined in terms of the specified inflow profile $\kappa(\psi(o,r))$ and $h(\psi(o,r))$ at the upstream boundary, provided that $\psi(x,r)$ is positive (i.e., outside a recirculation bubble). Inside the bubble, the κ and h distributions are not known. In fact, within the inviscid model a discontinuity is admissible across the streamline forming the bubble (the separating streamline). One way to avoid this problem is to invoke analytic

continuation of the functions $\kappa(\psi)$ and $h(\psi)$ for negative ψ. In the present work, the dependence of κ and h is known analytically for positive ψ from the assumed initial profiles. The same functional dependence is assumed for negative ψ. As a side point, it should be mentioned that since κ vanishes along a separating streamline ($\kappa = 0$ on the axis) and κ is analytically continued inside the bubble, it is reasonable to assume that κ changes sign inside the bubble; and, as a consequence, the swirl velocity in the bubble has the opposite sense to the swirl in the main flow. This is not the behavior observed with the viscous problem at Re = 100 and 200.

In solving the Navier-Stokes equations, the viscous terms play an important role in the neighborhood of the separating streamline by preventing the formation of discontinuous solutions. A similar role is played by the artificial viscosity terms in Euler calculations based on primitive variables (i.e., velocity components). However, it may be argued that although the artificial dissipation is critical in singling out a solution, the solution may be independent of its form and magnitude. In the least square formulation used in this work, there is no explicit or implicit artificial dissipation. In fact, the truncation errors for the central differences to be used are of a dispersive nature. It is, however, the assumption of analytic continuation of κ and h, for the inviscid problem, that rules out any discontinuities.

If we let the vortex core radius at the upstream boundary be $r = 1$ and the radius at the farfield boundary be $r = R$, the inflow profiles at $x = 0$ are given by:

$$
\begin{aligned}
u(r) &= 0 && 0 \leq r \leq R \\
v(r) &= Vr(2 - r^2) && r \leq 1 \\
v(r) &= V/r && r > 1 \\
w(r) &= 1 && 0 \leq r \leq R
\end{aligned}
\tag{2.5}
$$

where V is the maximum circumferential swirl velocity at the edge of the vortex core. These profiles are the same as those used in Ref. 4. From these profiles, it follows that the circulation at the upstream boundary is given by:

$$
\begin{aligned}
\kappa^2 &= 16V^2\psi^2\,(1 - \psi)^2 && \psi \leq 1/2 \\
\kappa^2 &= V^2 && \psi > 1/2
\end{aligned}
\tag{2.6}
$$

and that the vorticity component is given by:

$$
\begin{aligned}
r\omega &= 16V^2\,(1 + 2\psi^2 - 3\psi)\left(\frac{r^2}{2} - \psi\right) && \psi \leq 1/2 \\
r\omega &= 0 && \psi > 1/2
\end{aligned}
\tag{2.7}
$$

In terms of a perturbation streamfunction $\Psi = \frac{r^2}{2} - \psi$, the equation governing the inviscid flow is:

$$
\Psi_{xx} + r\,(\Psi_r/r)_r = -4V^2\alpha^2\Psi \tag{2.8}
$$

where

$$\alpha^2 = 4\ (1 + 2\psi^2 - 3\psi) \qquad \psi < 1/2$$

$$\alpha^2 = 0 \qquad \psi > 1/2 \tag{2.9}$$

We notice that since Ψ vanishes at the axis and if we require Ψ to vanish in the farfield, then the trivial solution $\Psi = 0$, corresponding to cylindrical stream surfaces, is a solution of the above equation. We are, however, interested in nontrivial solutions.

Using standard central difference approximation, equation (2.8) leads to a nonpositive definite matrix which is difficult to solve by standard relaxation schemes. To avoid this problem, a least-square variational formulation is obtained for the function

$$F(\Psi,u',w') = \iint_\Omega \left(\left(\frac{\Psi_r}{r} - w'\right)^2 r + \left(\frac{\Psi_x}{r} + u'\right)^2 r + \left(w'_r - u'_x + 4V^2 \frac{\alpha^2}{r}\psi\right)^2 \right) d\Omega \tag{2.10}$$

where Ω extends over the domain of interest and u' and w' are the perturbation velocities in terms of Ψ. The first term in parenthesis in the kernel of Eq. (2.10) corresponds to the definition of w', the second term corresponds to the definition of u', and the last term corresponds to Eq. (2.8). Each of these terms should vanish in the steady state. To form the kernel of Eq. (2.8), each of the above terms is multiplied by an arbitrary, but positive, weight function. The choice of r as the weight function for the first two terms and the cube of unit length for the last term was made to simplify the form of the resulting equations. From the function (2.10), the following Euler equations are easily obtained [6]:

$$\frac{\Psi_{xx}}{r} + \left(\frac{\Psi}{r}\right)_r - \frac{4V^2\alpha^2 g}{r} = (w'_r - u'_x)\left(1 + \frac{4V^2\alpha^2}{r}\right)$$

$$u'_{xx} - ru' = \Psi_x + w'_{rx} + g_x \qquad (2.11)$$

$$w'_{rr} - rw' = -\Psi_r + u'_{xr} - g_r$$

where

$$g = \frac{4V^2\alpha^2}{r}\Psi$$

3. NUMERICAL FORMULATION

3.1 Inviscid Problem

Equations (2.11) are solved using a staggered grid for Ψ, u', and w'. With Ψ defined at i,j nodes, u' between nodes of horizontal lines, and w' between nodes of vertical lines, the resulting discrete equations are:

$$\frac{(\Psi_{i+1,j} - 2\Psi_{i,j} + \Psi_{i-1,j})}{\Delta x^2} + \left[\frac{r_{i,j}}{\Delta r^2}\,\frac{\Psi_{i,j+1} - \Psi_{ij}}{r_{i,j+1/2}} - \frac{\Psi_{i,j} - \Psi_{i,j-1}}{r_{i,j-1/2}}\right] - \frac{4V^2\alpha^2}{r_{i,j}} g_{i,j}$$

$$= r_{i,j}\left(\frac{w'_{i,j+1/2} - w'_{i,j-1/2}}{\Delta r} - \frac{u'_{i+1/2,j} - u'_{i-1/2,j}}{\Delta x}\right)\left(1 + \frac{4V^2\alpha^2}{r_{i,j}}\right)$$

$$\frac{(u'_{i+3/2,j} - 2u'_{i+1/2,j} + u'_{i-1/2,j})}{\Delta x^2} - r_{i,j}u'_{i+1/2,j} = \frac{\Psi_{i+1,j} - \Psi_{i,j}}{\Delta x}$$

$$+ \frac{(w'_{i+1,j+1/2} - w'_{i+1,j-1/2} - w'_{i,j+1/2} + w'_{i,j-1/2})}{\Delta x \Delta r} + \frac{(g_{i+1,j} - g_{i,j})}{\Delta x}$$

$$\frac{(\hat{w}_{i,j+3/2} - 2\hat{w}_{i,j+1/2} + \hat{w}_{i,j-1/2})}{\Delta r^2} - r_{i,j+1/2}\hat{w}_{i,j+1/2} = -\frac{\Psi_{i,j+1} - \Psi_{i,j}}{\Delta r}$$

$$+ \frac{(\hat{u}_{i+1/2,j+1} - \hat{u}_{i-1/2,j+1} - \hat{u}_{i+1/2,j} + \hat{u}_{i-1/2,j})}{\Delta x \Delta r} - \frac{(g_{i,j+1} - g_{i,j})}{\Delta r} \qquad (3.1)$$

where

$$\begin{aligned} x_{i,j} &= (i-1)\,\Delta x \qquad i = 1,2\ldots I_{max} \\ r_{i,j} &= (j-1)\,\Delta r \qquad j = 1,2\ldots J_{max} \end{aligned} \qquad (3.2)$$

The first equation of (3.1) is solved for Ψ by the Zebra vertical line over-relaxation algorithm; the second equation is solved for u by direct inversion of horizontal lines; the last equation is solved for w by direct inversion of vertical lines. For the first equation of (3.1), the boundary condition consists of $\Psi = 0$ all around the domain. For the second equation, boundary conditions are required at $i = 1+1/2$ and $i = I_{max} - (1+1/2)$. Boundary conditions for $\hat{u}$ at $i = 1+1/2$ are obtained by solving

$$[\hat{u}_x - (\hat{w}_r + g)] - [\Psi_x + r\hat{u}] = 0 \qquad (3.3)$$

at $i = 2$; while at $i = I_{max} - (1+1/2)$, boundary conditions are obtained by solving

$$[\hat{u}_x - (\hat{w}_r + g)] + [\Psi_x + r\hat{u}] = 0 \qquad (3.4)$$

at $i = I_{max} - 1$. Each term in square brackets appearing in Eqs. (3.3) and (3.4) vanishes in the steady state. The change in sign between Eqs. (3.3) and (3.4) is introduced to add to the diagonal dominance of the discrete equations. Similarly, for w' the equation

$$[w'_r - (u'_x - g)] + [\Psi_r - rw'] = 0 \tag{3.5}$$

is solved at $j = 2$ to obtain w' at $j = 1+1/2$; while at $j = J_{max} - 1$, the equation

$$[w'_r - (u'_x - g)] - [\Psi_r - rw] = 0 \tag{3.6}$$

is solved to obtain w' at $j = J_{max} - (1+1/2)$.

3.2 Viscous Problem

The first equation of (2.1) is discretized using second-order-accurate central-difference formulas. To the second and third equations of (2.1) the time terms ω_t and κ_t are added to the left-hand side of each equation, respectively. The convection terms of these two equations are discretized using upwind first-order accurate formulas, while the diffusion terms are discretized using second-order-accurate formulas. The time terms are discretized using first-order backward derivatives. Unlike the inviscid problem, a nonstaggered grid is used for the viscous problem. The discrete equations are

$$\frac{\psi_{i+1,j} - 2\bar{\psi}_{i,j} + \bar{\psi}_{i-1,j}}{\Delta x^2} + \frac{r_{i,j}}{\Delta r^2}\left(\frac{\bar{\psi}_{i,j+1} - \bar{\psi}_{ij}}{r_{i,j+1/2}} - \frac{\bar{\psi}_{i,j} - \bar{\psi}_{i,j-1}}{r_{i,j-1/2}}\right) = r_{i,j}\omega_{i,j}$$

$$\frac{\bar{\omega}_{i,j} - \omega_{i,j}}{\Delta t} + \frac{u_{i,j} + |u_{i,j}|}{2}\frac{\bar{\omega}_{i,j} - \bar{\omega}_{i,j-1}}{\Delta r} + \frac{u_{i,j} - |u_{i,j}|}{2}\frac{\bar{\omega}_{i,j+1} - \bar{\omega}_{i,j}}{\Delta r}$$

$$+ \frac{w_{i,j} + |w_{i,j}|}{2}\frac{\bar{\omega}_{i,j} - \bar{\omega}_{i-1,j}}{\Delta x} + \frac{w_{i,j} - |w_{i,j}|}{2}\frac{\omega_{i+1,j} - \bar{\omega}_{i,j}}{\Delta x}$$

$$- \frac{u_{i,j}\bar{\omega}_{i,j}}{r_{i,j}} + \frac{\kappa^2_{i+1,j} - \kappa^2_{i-1,j}}{2\Delta x r^3_{i,j}} = \frac{1}{Re}\left[\frac{\bar{\omega}_{i,j+1} - 2\bar{\omega}_{i,j} + \bar{\omega}_{i,j-1}}{\Delta r^2}\right.$$

$$\left. + \frac{\bar{\omega}_{i,j+1} - \bar{\omega}_{i,j-1}}{2\Delta r\, r_{i,j}} - \frac{\bar{\omega}_{i,j}}{r^2_{i,j}} + \frac{\omega_{i+1,j} - 2\bar{\omega}_{i,j} + \bar{\omega}_{i-1,j}}{\Delta x^2}\right]$$

$$\frac{\bar{\kappa}_{i,j} - \kappa_{ij}}{\Delta t} + \frac{u_{i,j} + |u_{i,j}|}{2}\frac{\bar{\kappa}_{i,j} - \bar{\kappa}_{i,j-1}}{\Delta r} + \frac{u_{i,j} - |u_{i,j}|}{2}\frac{\bar{\kappa}_{i,j+1} - \bar{\kappa}_{i,j}}{\Delta r}$$

$$\frac{w_{i,j} + |w_{i,j}|}{2}\frac{\bar{\kappa}_{i,j} - \bar{\kappa}_{i-1,j}}{\Delta x} + \frac{w_{i,j} - |w_{i,j}|}{2}\frac{\kappa_{i+1,j} - \bar{\kappa}_{i,j}}{\Delta x}$$

$$= \frac{1}{Re}\left[\frac{\bar{\kappa}_{i,j+1} - 2\bar{\kappa}_{i,j} + \bar{\kappa}_{i,j-1}}{\Delta r^2} - \frac{\bar{\kappa}_{i,j+1} - \bar{\kappa}_{i,j-1}}{2\Delta r\, r_{i,j}} + \frac{\kappa_{i+1,j} - 2\bar{\kappa}_{i,j} + \bar{\kappa}_{i-1,j}}{\Delta x^2}\right] \tag{3.7}$$

Quantities with a bar are taken at the new time or iteration level. The time step is chosen equal to Δx. (Note that the identity $u_r + w_x = \frac{-u}{r}$ was used to simplify the convective terms of the second equation above.) The Eqs. (3.7) are solved by vertical line over-relaxation with the following boundary conditions:

At $x = 0$,

$$\psi = \frac{r^2}{2}$$

$$\kappa = 4V\psi(1 - \psi) \qquad 0 < r < 1$$

$$\kappa = V \qquad 1 < r < R$$

$$\omega = \psi_{xx}/r \ ;$$

at $r = 0$,

$$\psi = 0$$

$$\kappa = 0$$

$$\omega = 0 \ ;$$

at $r = R$,

$$\frac{\psi_r}{r} = 1$$

$$\kappa = V$$

$$\omega = 0 \ ;$$

at the outflow, $x = L$,

$$\psi_x = 0$$

$$\kappa_x = 0$$

$$\omega_x = 0$$

To improve the convergence rate of the viscous problem, the acceleration method described in Ref. 7 was used.

4. DISCUSSION OF RESULTS

In order to measure the deviation of a solution from the trivial solution $\psi = \frac{r^2}{2}$, we define the norm of the perturbation streamfunction as

$$||\Psi|| = LR\left(\sum_{i=1}^{I_{max}} \sum_{j=1}^{J_{max}} \Psi_{ij}^2\right)^{1/2} \Big/ (I_{max} J_{max}) \qquad (4.1)$$

Tests were performed to determine the required number of mesh points and the required locations of the farfield boundaries to achieve a certain level of accuracy. Two of these tests are illustrated in Figures 1 and 2 for the inviscid problem. Figure 1 shows the asymptotic behavior of the streamfunction norm as the number of mesh points in the axial direction is increased, holding all other parameters fixed. Figure 2 shows the effect of the location of the outflow boundary on the norm. From this study, it was concluded that for the inviscid problem a minimum spacing $\Delta r = \Delta x = 1/16$ was required and that $R \geq 2$, $L \geq 4$ was also required. The same requirements were found for the viscous problem for $100 \leq Re \leq 200$, except that the location of the outflow boundary had to be increased to $L \geq 10$. Figures 3 and 4 show that the same solution is obtained for the inviscid problem with $L = 5$ and $L = 10$. For all cases presented, residuals were driven to machine zero, $O(10^{-12})$.

A summary of the results is given in Figure 5. This figure shows the norm defined by Eq. (4.1) as a function of the square of the swirl parameter V. Two nontrivial branches were found for the inviscid problem. The first branch, indicated in the figure by the closed circles, corresponds to axisymmetric vortex breakdown-like solutions. Figs. (3), (4), and (6) illustrate the streamline topologies found in this branch. The same branch

was found by Ta'asan [8] using a multigrid algorithm to solve Eqs. (2.8). As shown in Figure 5, our results and those of Ta'asan are in good agreement. The problem with this branch is that as the swirl parameter is increased, the size of the bubble decreases. This behavior contradicts the experimental observations. The second inviscid branch, indicated by the open circles, intersects the first at approximately $V^2 = 0.575$. For values of V near the intersection of the two branches, the numerical algorithm developed a limit cycle where a single bubble splits in two. The two bubbles later coalesce and the cycle is repeated. The limit cycle prevented convergence to a steady state. Ta'asan only encountered the first branch and was able to continue this branch down to the axis, as shown in Figure 5. The streamlines corresponding to the second branch are illustrated in Figure 7. Obviously, this branch is not of the vortex breakdown type. It is believed that the second branch, although a solution to the least square problem, is not a solution of the original inviscid problem (Eqs. 2.1 with $Re \to \infty$). The evidence for this comes from inserting the least-square solutions into the original equations and evaluating the residuals. When this is done for the first branch, residuals of the order of Δx^2 are found. For the second branch, the residuals are of order Δx, and remain at the same level when the mesh is refined.

For the viscous problem, results are presented in Figure 5 for Reynolds numbers 100 and 200. The axial velocities obtained here and those obtained by Grabowski and Berger [4] are compared in Figure 8. The agreement is good when we consider that Grabowski and Berger used a much coarser but highly stretched mesh, slightly different outflow boundary conditions, and did not converge their solutions to the same level as in this work. It also appears that the

results obtained by Krause et al. [5] are anomalous, since they were unable to obtain steady-state solutions for the same cases studied here. Their failure to reach a steady state could be a result of the outflow boundary being at L = 5, too close to the inflow boundary. In our work, we found this location for the outflow boundary to lead to a large open bubble (see Figure 9), but the solution was steady nonetheless.

Figure 10 illustrates the changes in the bubble structure as the swirl parameter is increased with Re = 100 held fixed. The same is illustrated in Figure 11 for Re = 200. It is interesting to see the very rapid change in the norm that occurs at Re = 200 and $V^2 \approx 1.27$. (See Figure 5.) This behavior opens some questions about possible hysteresis and bifurcation at higher Reynolds number. However, our present approach is not capable of handling much higher Reynolds numbers well; and, therefore, these questions will be considered at a later time. Figure 12 shows the minimum value of the axial velocity component on the axis as a function of V for Reynolds numbers of 100 and 200. The point at which the recirculation bubble first appears corresponds to the first intersection of these curves with w = 0. The second intersection corresponds to the point at which the recirculation bubble lifts off the axis.

5. CONCLUDING REMARKS

Numerical solutions of the Euler equations were obtained and a vortex breakdown-like topology was observed. Those solutions were in good agreement with those obtained by Ta'asan [8]. For the Navier-Stokes equations, solutions were also obtained with vortex breakdown-like topology. These

latter solutions were in good agreement with the results reported in Ref. 4. The behavior of the inviscid solutions with increasing swirl was not consistent with the behavior of the Navier-Stokes solutions at low Reynolds number nor with experimental observations. (Experimental results showing bubble-type vortex breakdown are usually obtained at higher Reynolds numbers.) A future study will investigate the high Reynolds number limit of the Navier-Stokes equations and compare it to the Euler solutions obtained here.

References

[1] S. Leibovich, "Vortex Stability and Breakdown: Survey and Extension," AIAA J., Vol. 22, No. 9, 1984, pp. 1192-1206.

[2] S. Leibovich, "The Structure of Vortex Breakdown," Ann. Rev. Fluid Mech., Vol. 10, 1978, pp. 221-246.

[3] S. M. Hitzel and W. Schmidt, "Slender Wings with Leading-Edge Vortex Separation: A Challenge for Panel Methods and Euler Solvers," J. Aircraft, Vol. 21, No. 10, 1984, pp. 751-759.

[4] W. J. Grabowski and S. A. Berger, "Solutions of the Navier-Stokes Equations for Vortex Breakdown," J. Fluid Mech., Vol. 75, Part 3, 1976, pp. 525-544.

[5] E. Krause, E., X. G. Shi, and P. M. Hartwich, "Computation of Leading Edge Vortices," AIAA Paper No. 83-1907, Computational Fluid Dynamics Conference, Danvers, Massachusetts, 1983.

[6] R. Courant and F. John, Introduction to Calculus and Analysis, Vol. 2, Chapter 7, pp. 737-768.

[7] M. Hafez, E. Parlette, and M. D. Salas, "Convergence Acceleration of Iterative Solutions of Euler Equations for Transonic Flow Computations," AIAA Paper 85-1641, AIAA 18th Fluid Dynamics and Plasma-dynamics and Lasers Conference, July 16-18, 1985, Cincinnati, Ohio.

[8] S. Ta´asan, "A Multigrid Method for Vortex Breakdown Simulation," ICASE Report to appear.

[9] Xun-Gang Shi, "Numerische Simulation Des Aufplatzens von Werbeln," Ph.D. Thesis, September 1983, Technischen Hochschule Aachen, West Germany.

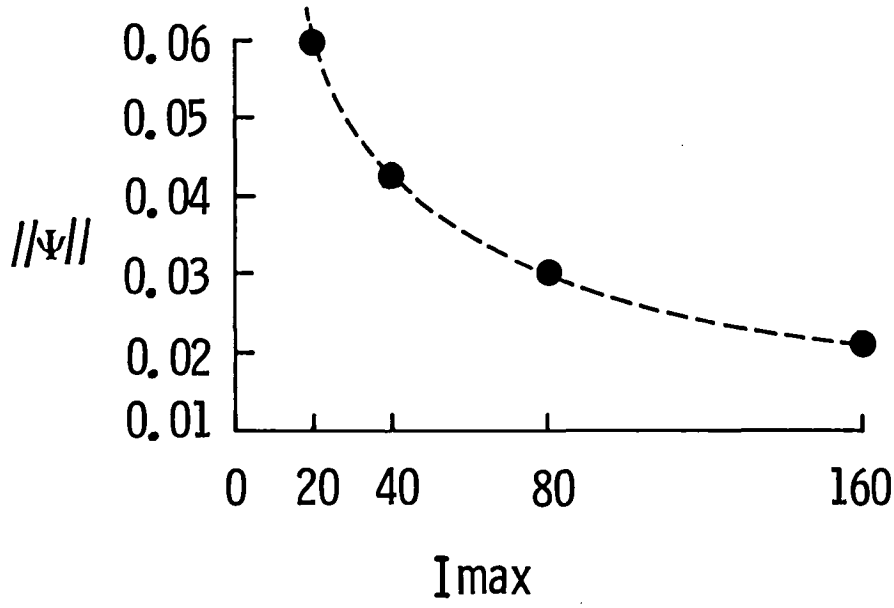


Figure 1. Convergence of the norm of the inviscid streamfunction Ψ with increasing resolution in the axial direction, holding L = 5, R = 2, V^2 = 0.4, and Δr = 1/16 fixed.

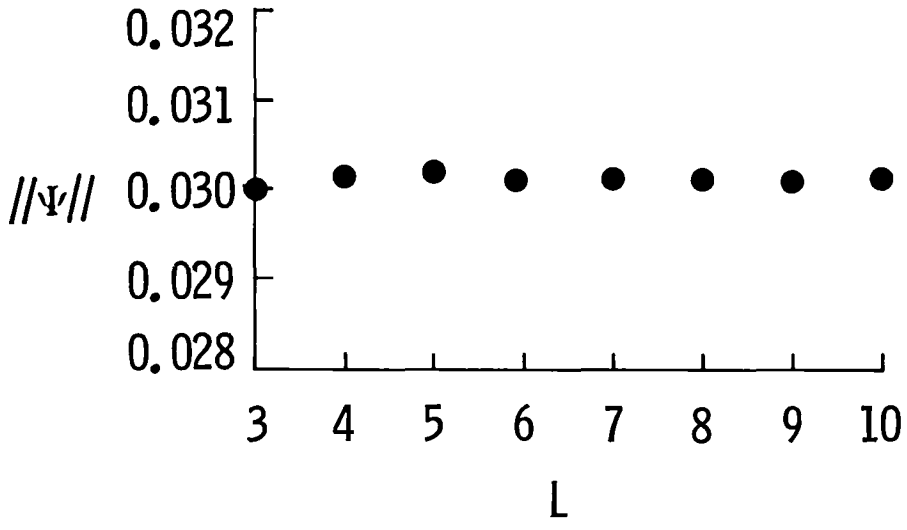


Figure 2. Effect of increasing the length of the domain on the norm of the inviscid streamfunction Ψ, holding R = 2, $\Delta r = \Delta x$ = 1/16.

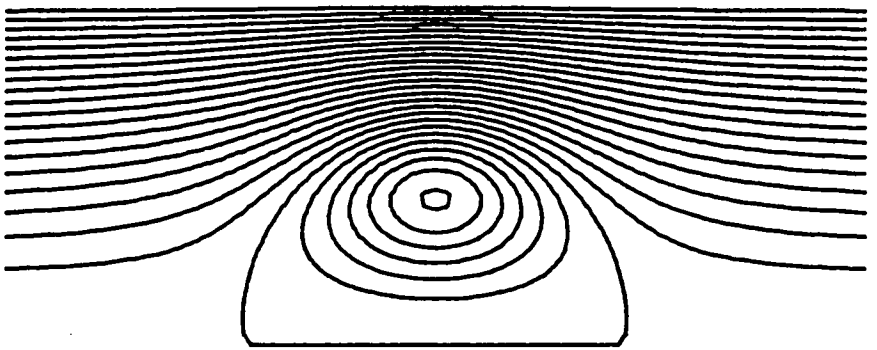


Figure 3. Computed streamline pattern for $V^2 = 0.2$, L = 5, R = 2, and $\Delta x = \Delta r = 1/16$.

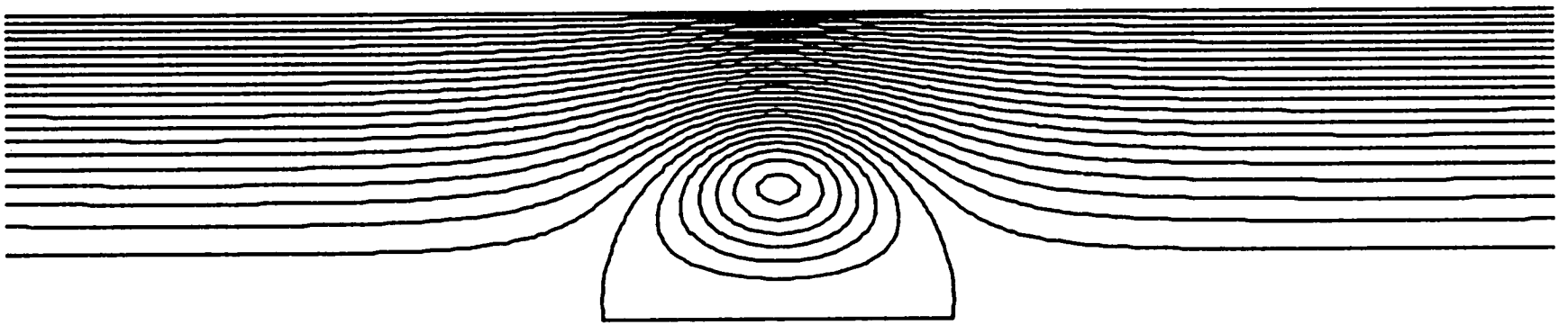


Figure 4. Computed streamline pattern for $V^2 = 0.2$, L = 10, R = 2, and $\Delta x = \Delta r = 1/16$.

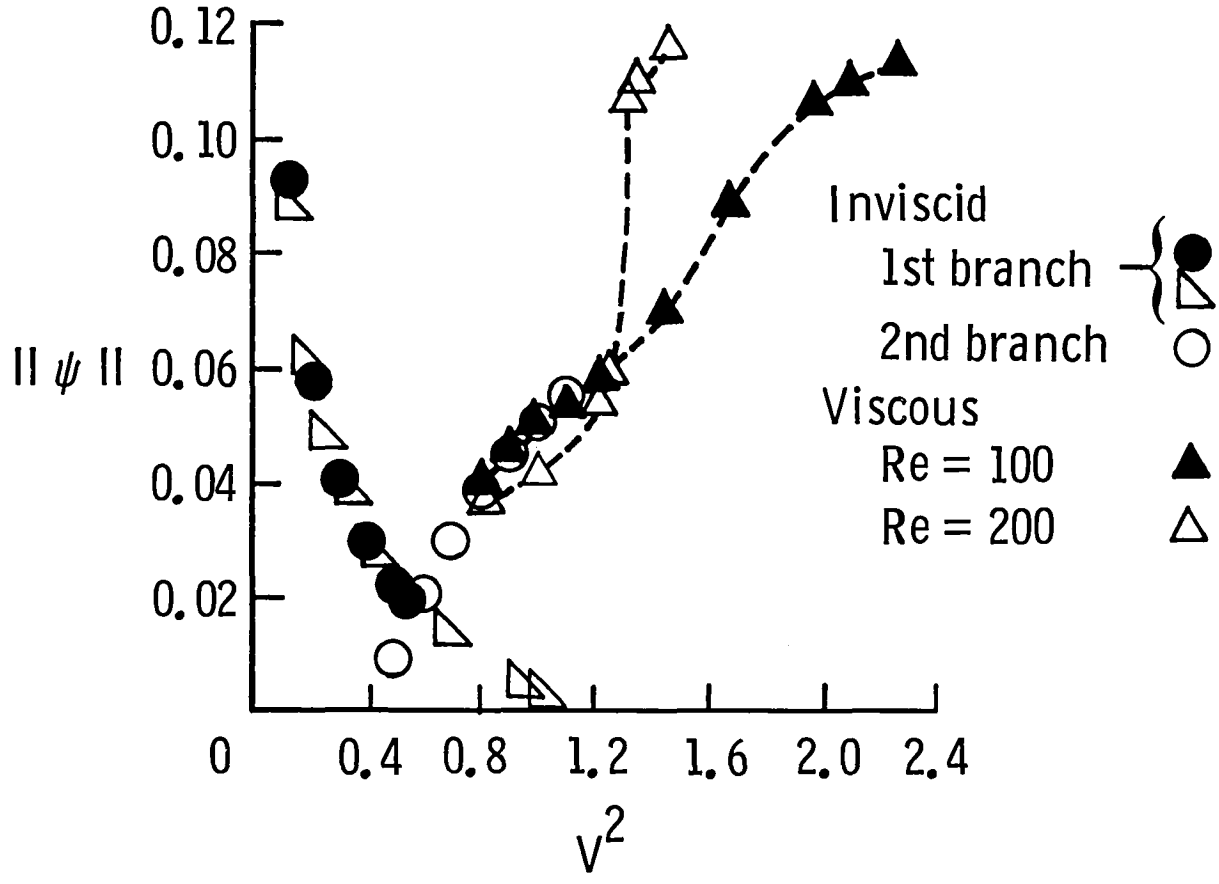


Figure 5. Norm of the streamfunction Ψ as a function of V^2.

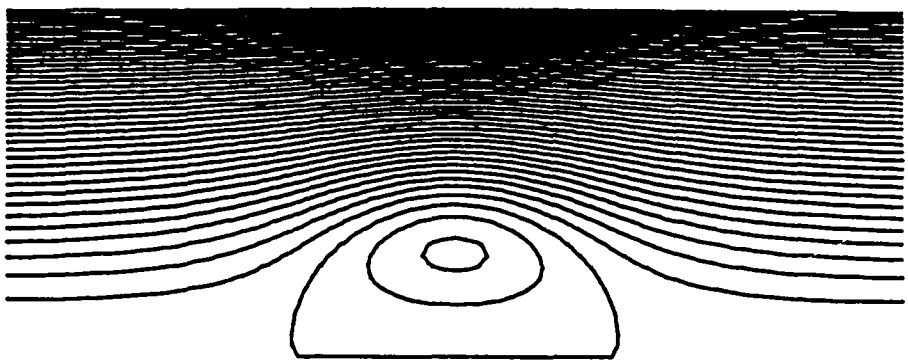


Figure 6. Computed streamline pattern for $V^2 = 0.5$, $L = 5$, $R = 2$, and $\Delta x = \Delta r = 1/16$.

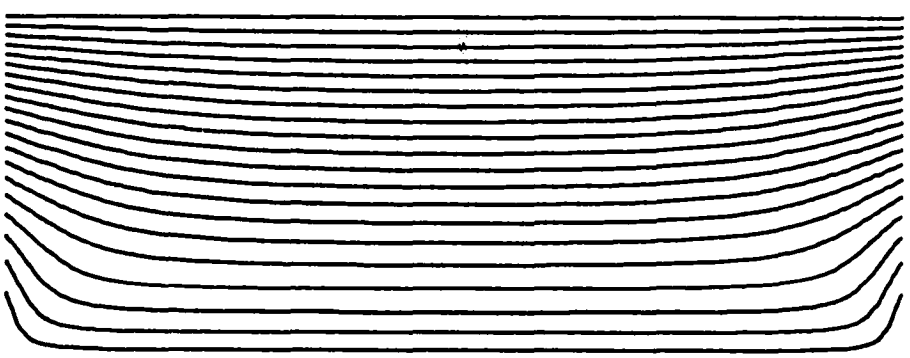


Figure 7. Computed streamline pattern for $V^2 = 0.9$, $L = 5$, $R = 2$, and $\Delta x = \Delta r = 1/16$.

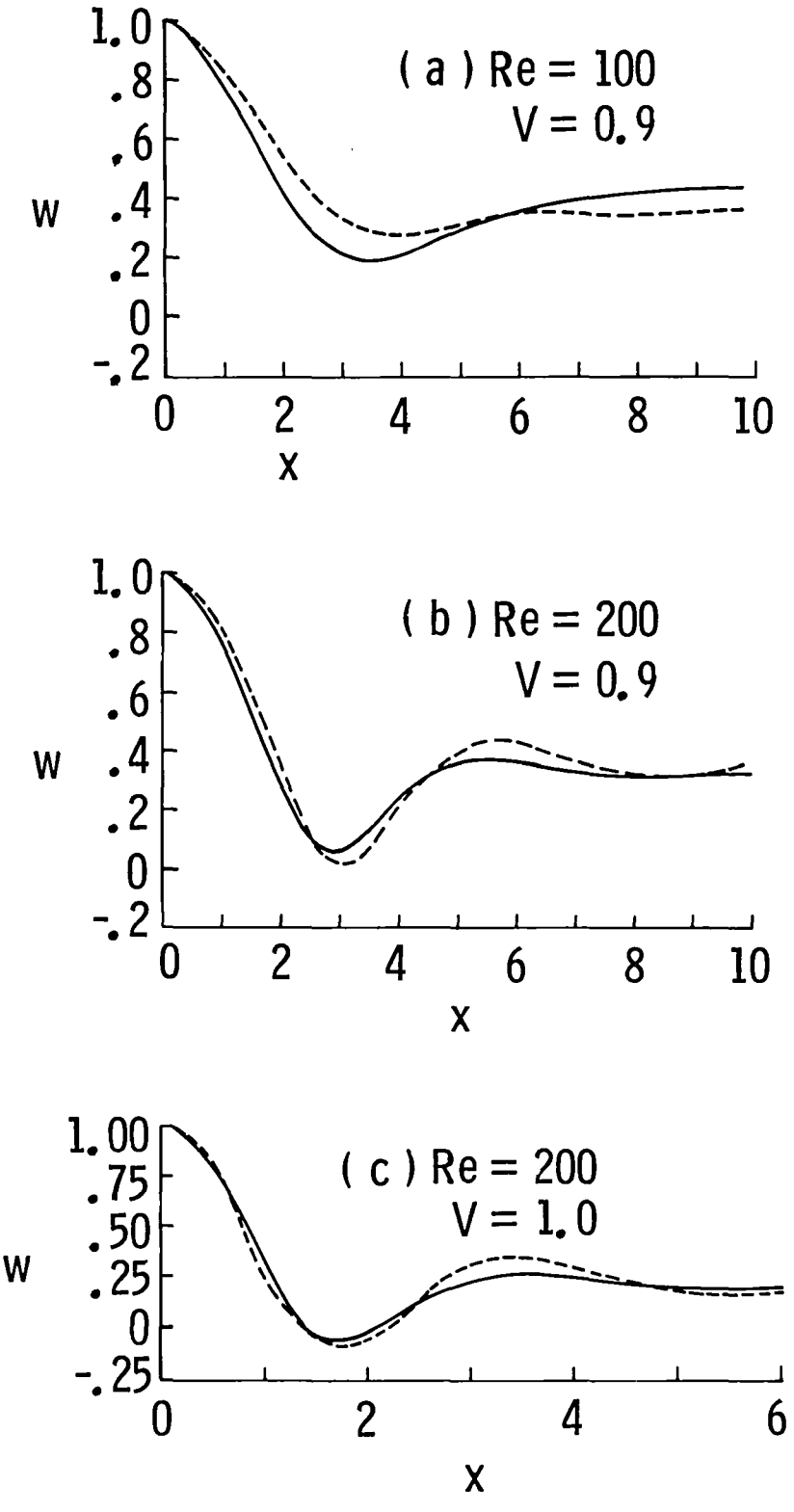


Figure 8. Comparison f velocity on vortex axis between present results (solid line) and those of Ref. 4 (dashed line).

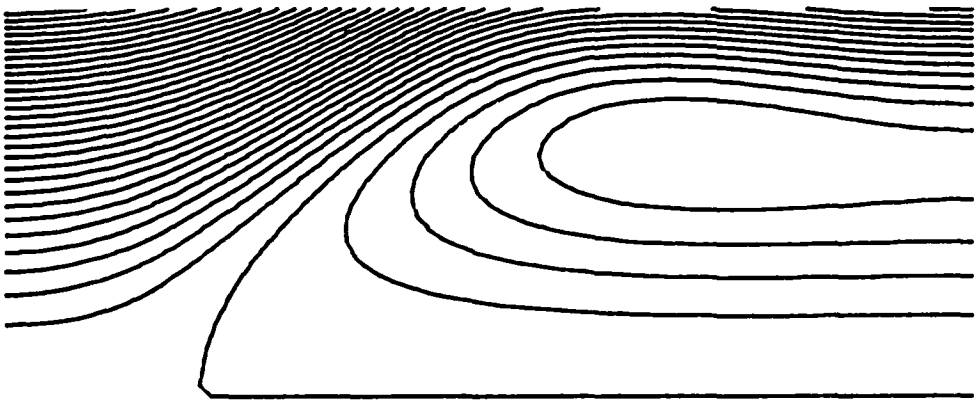
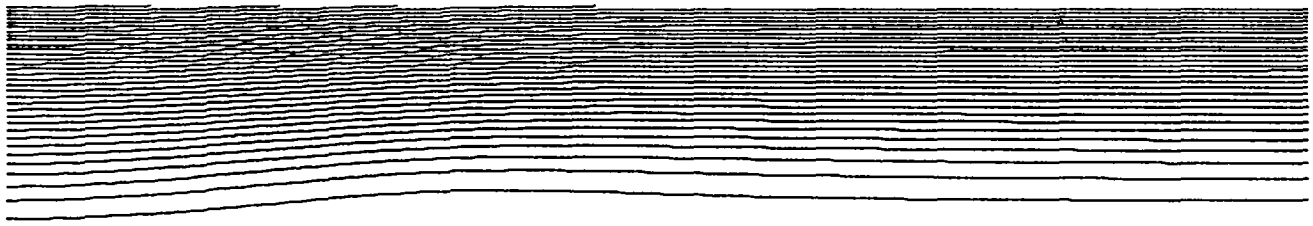
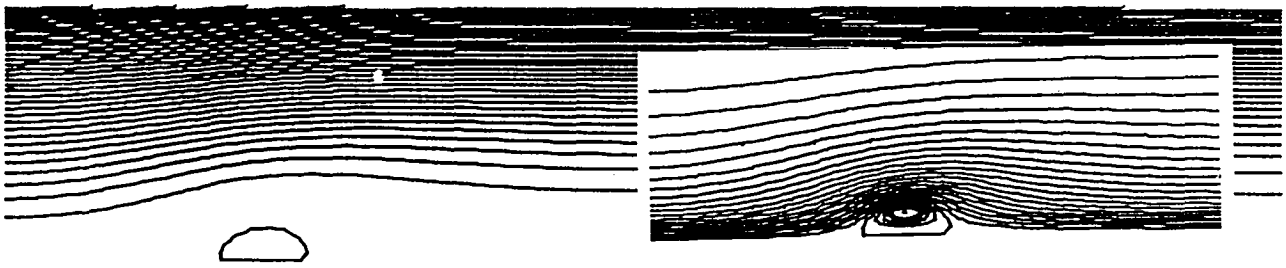


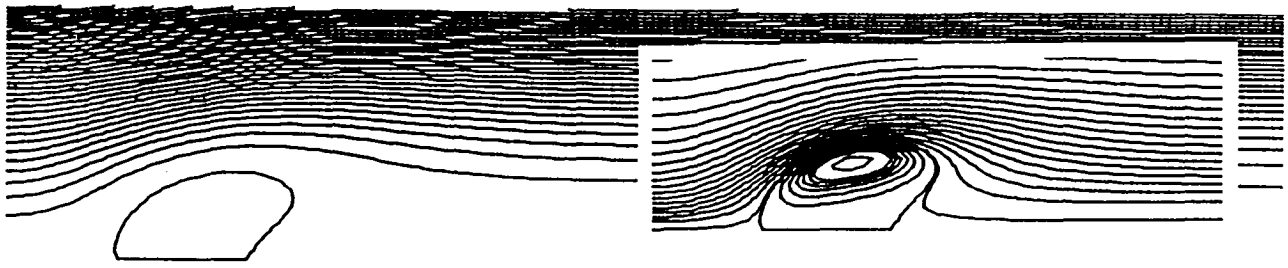
Figure 9. Computed streamline pattern for Re = 200, V = 0.8944, L = 5, R = 2, and $\Delta x = \Delta r = 1/16$. The shortness of the domain results in a large open bubble. This case corresponds to the same conditions of Figure 4.14, Ref. 9.



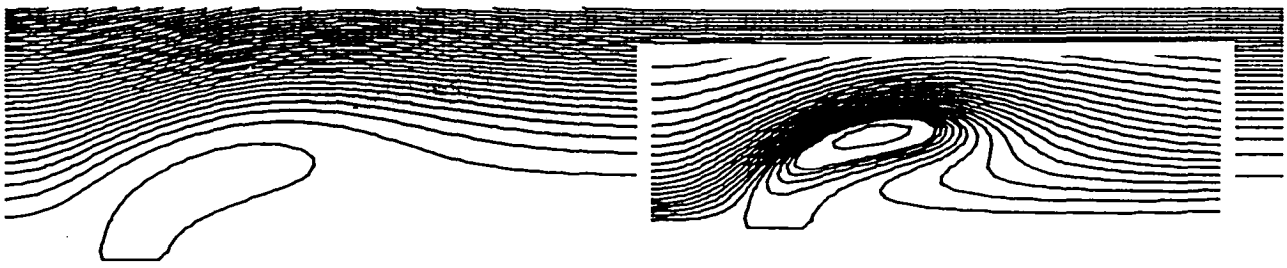
(a) V = 0.9



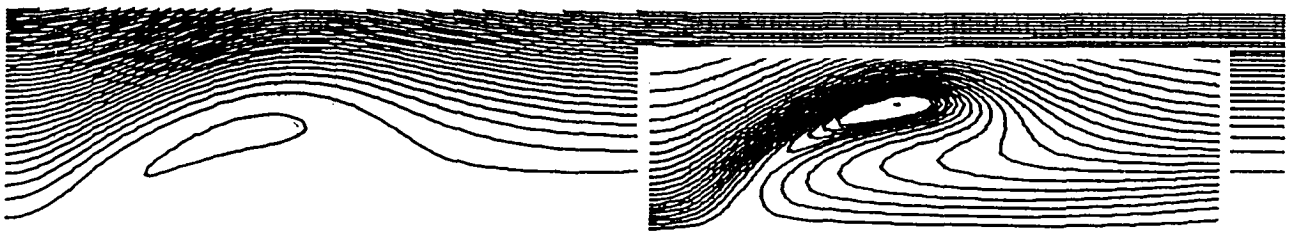
(b) V = 1.0



(c) V = 1.2

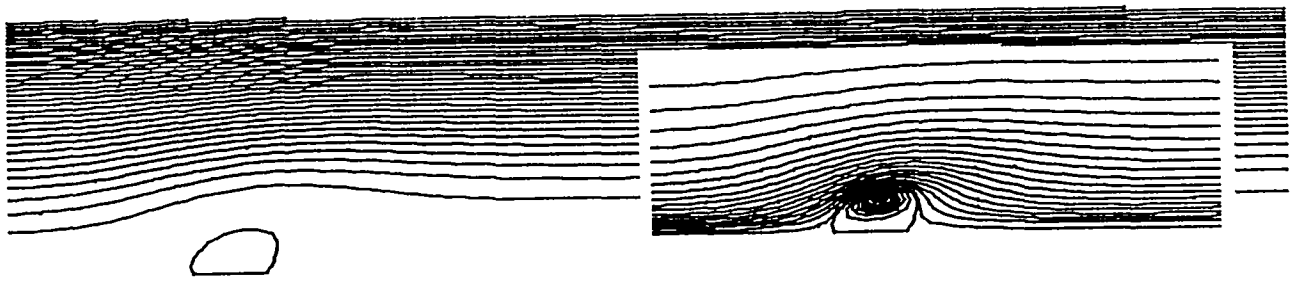


(d) V = 1.3

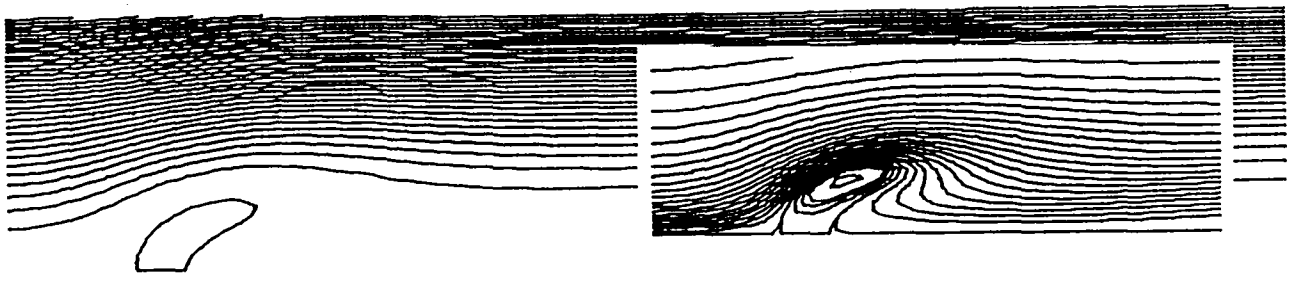


(e) V = 1.5

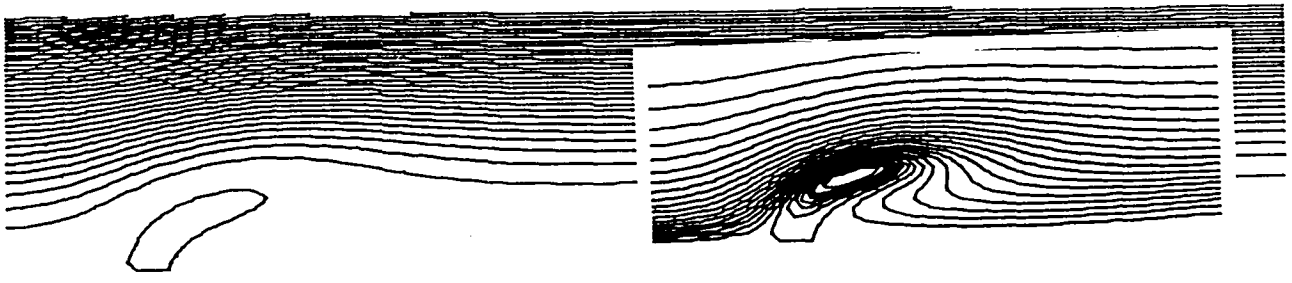
Figure 10. Computed streamline patterns for Re = 100, L = 10, R = 2, $\Delta x = \Delta r = 1/16$, and increasing values of V. Details of the bubble structure are shown on the insets.



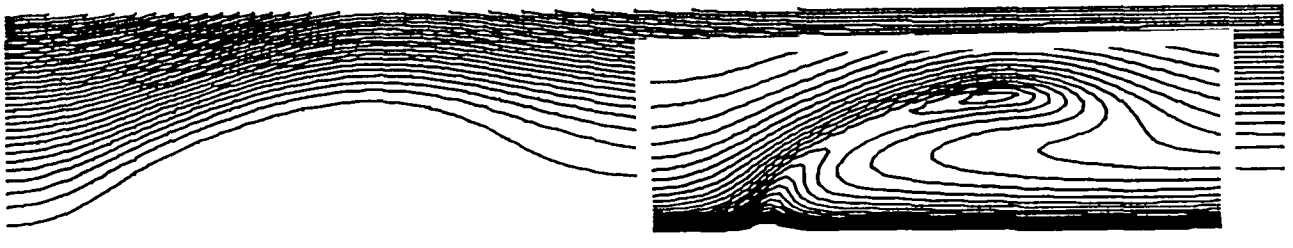
(a) V = 1.0



(b) V = 1.1



(c) V = 1.12



(d) V = 1.15

Figure 11. Computed streamline patterns for Re = 200, L = 10, R = 2, $\Delta x = \Delta r = 1/16$, and increasing values of V. Details of the bubble structure are shown on the insets.

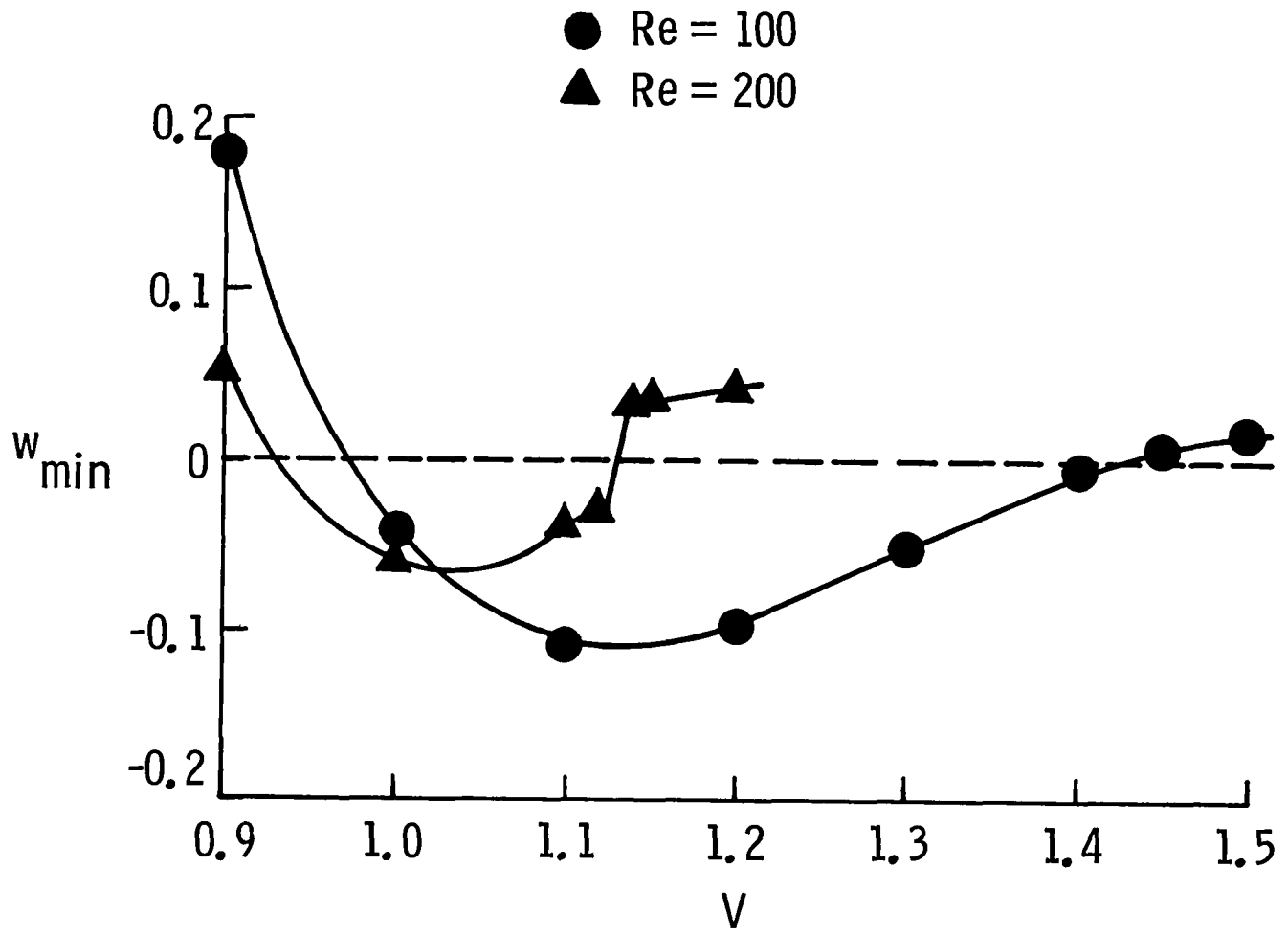


Figure 12. Minimum velocity on the axis as a function of V for Re = 100 and 200.

MULTIGRID METHOD FOR A VORTEX BREAKDOWN SIMULATION

Shlomo Ta'asan
Institute for Computer Applications in Science and Engineering

ABSTRACT

In this paper we study an inviscid model for a steady axisymmetric flow with swirl. The governing equation is a nonlinear elliptic equation which has more than one solution for a certain range of the swirl parameter. The physically interesting solutions have closed streamlines that look like vortex breakdown ("bubble"-like solutions). A multigrid method is used to find these solutions. Using an FMG algorithm (nested iteration), the problem is solved in just a few multigrid cycles.

Research was supported by the National Aeronautics and Space Administration under NASA Contracts No. NAS1-17070 and NAS1-18107 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665-5225.

1. INTRODUCTION

In this paper we study an inviscid model for steady axisymmetric flow with swirl, which has solutions with closed streamlines. These solutions have a structure similar to that observed experimentally as "bubble"-like solutions when vortex breakdown occurs [4].

Using a streamfunction-vorticity formulation to the axisymmetric incompressible Navier-Stokes equations, it was found [3] that one can reduce the problem to a single nonlinear elliptic equation for the streamfunction, in case of a special inflow flow and some regularity assumption on the vorticity. This nonlinear elliptic equation for the streamfunction has more than one solution. The trivial, represents a uniform flow and is of no physical interest. The other shows a "bubble"-like structure, the target of our numerical study.

In solving the problem numerically, the problem is reformulated in terms of a perturbed streamfunction, i.e., the deviation from the trivial solution. In terms of this perturbed streamfunction, the trivial solution is represented as an identically zero solution. Our goal then is to find non-zero solutions which have "bubble"-like form.

The approach we have taken in finding these solutions is to seek first for a bifurcation point from the trivial branch of solutions. By introducing a continuation parameter, we can then start marching on a branch of non-trivial solutions that bifurcate from that point. One choice of a continuation parameter is arc length [1]. Another choice, which is simpler but may not be good in general, is the norm of the perturbed streamfunction. The natural parameter in the problem, a swirl velocity parameter, is not good enough since it cannot "choose" the non-zero branch as can the former parameter. We

therefore choose the norm as a continuation parameter, making the swirl velocity parameter an unknown to be determined by the solution.

The multigrid approach used for solving the problem is similar to the one used in [5] for solving the Bratu problem. The relaxation in this method consists of three steps: (i) a local relaxation to smooth the error; (ii) a step to update the norm of the solution; and (iii) a step to update the swirl velocity parameter. An FMG algorithm (nested iteration) is used. That is, a solution for the prescribed norm is found first on the coarsest level, and then interpolated to finer levels, where on each level a few basic multigrid V-cycles are performed before proceeding to yet finer level.

The coarsest level, when solved to get an initial approximation for finer levels, uses a continuation method. Here the problem was solved first for a small norm, and then the norm is gradually increased until the prescribed norm is reached. Each time the norm is increased, the solution of the previous step was used as initial approximation. By solving for a bifurcation point from the trivial solution, a first approximation for the smallest norm problem was obtained.

Once a solution on the coarsest level is obtained for a prescribed norm, it is possible to solve finer grid problems without continuation.

The same problem we are discussing here was treated by a completely different method and is reported in [3]. There, a single grid method was used with a least squares formulation of the problem. The amount of work needed for that approach is considerably larger than the one reported here. Computed solutions by the two different formulations are in good agreement.

2. ON DERIVATION OF THE GOVERNING EQUATION

We summarize here the derivation of the equations used in the numerical process as given in [3]. In cylindrical coordinates (x,r,θ) the incompressible Navier-Stokes equations can be written in terms of a streamfunction ψ, vorticity ω, and circulation k as

$$r\left(\frac{\psi_r}{r}\right)_r + \psi_{xx} = r\omega \qquad (2.1a)$$

$$(u\omega)_r + (w\omega)_x + \left(\frac{k^2}{r^3}\right)_x = \frac{1}{Re}\left[\omega_{rr} + \frac{1}{r}\omega_r - \frac{\omega}{r^2} + \omega_{xx}\right] \qquad (2.1b)$$

$$uk_r + wk_x = \frac{1}{Re}\left[k_{rr} - \frac{1}{r}k_r + k_{xx}\right] \qquad (2.1c)$$

where $k = rv$, $\omega = w_r - u_x$ and Re is the Reynolds number. The velocity components in the x,r,θ directions are w, u, v, respectively, of which w and u are given in terms of the streamfunction by

$$w = \frac{\psi_r}{r} \qquad (2.2a)$$

$$u = -\frac{\psi_x}{r} \qquad (2.2b)$$

It is shown in [3] that in the inviscid case $(Re = \infty)$, one finds that the circulation k and the vorticity ω are functions of the streamfunction ψ only. Therefore, k and ω can be determined outside the "bubble" from the inflow boundary condition. In the model discussed it is assumed that the same functional dependence of k, ω on ψ is true also inside the bubble

(negative ψ). This imposes some regularity on the solution.

For the inflow conditions

$$v(0,r) = \begin{cases} V_0\, r(2 - r^2) & r \leq 1 \\ V_0/r & r > 1, \end{cases} \tag{2.3a}$$

$$w(0,r) = 1, \tag{2.3b}$$

it is possible to write k and ω in terms of the streamfunction as

$$k^2(0,r) = \begin{cases} 16\, V_0^2\, \psi^2(1 - \psi)^2 & \psi \leq 1/2 \\ V_0^2 & \psi > 1/2 \end{cases} \tag{2.4a}$$

$$\omega(0,r) = \begin{cases} 16\, V_0^2(1 + 2\psi^2 - 3\psi)(r^2/2 - \psi) & \psi \leq 1/2 \\ 0 & \psi > 1/2 \end{cases} \tag{2.4b}$$

and therefore, the equation obtained for ψ is

$$r(\psi_r/r)_r + \psi_{xx} = -\, 4V_0^2\, \tilde{\alpha}^2(\psi)(\psi - r^2/2) \tag{2.5a}$$

where

$$\tilde{\alpha}^2(\psi) = \begin{cases} 4(1 + 2\psi^2 - 3\psi) & \psi \leq 1/2 \\ 0 & \psi > 1/2\,. \end{cases} \tag{25b}$$

The reduction of the governing equations to a single nonlinear elliptic equation is possible if the relation $\psi = f(r)$ in the inflow boundary can be inverted to get $r = g(\psi)$. When $g(\psi)$ is introduced in the expression for

v at the inflow boundary, one has v as a function of ψ in that boundary and therefore $k(\psi)$, $\omega(\psi)$. Note that, in general, one cannot expect to analytically invert the relation $\psi = f(r)$, and so the reduction of the governing equations is possible only for very special inflows.

Numerical experiments were done in terms of $\phi = \psi - \frac{r^2}{2}$, which is a perturbation from the trivial solution $\psi = \frac{r^2}{2}$ that represents a uniform flow.

3. NUMERICAL ALGORITHM

3.1. Discretization

The equation for $\phi = \psi - r^2/2$ is given by

$$r\left(\frac{1}{r}\,\phi_r\right)_r + \phi_{xx} + 4\,V_0^2\,\alpha^2(\phi)\phi = 0, \qquad \Omega = (0,a) \times (0,b) \tag{3.1a}$$

$$\phi = 0, \qquad \text{on} \quad \partial\Omega \tag{3.1b}$$

where

$$\alpha^2(\phi) = \begin{cases} 4\,\phi - 1 + \frac{r^2}{2}\,(2\phi - 1 + r^2) & \phi + \frac{r^2}{2} < 1/2 \\ 0 & \text{otherwise} \end{cases} \tag{3.1c}$$

Equations (3.1) are discretized as

$$\frac{\phi^h_{i+1,j} - 2\phi^h_{ij} + \phi^h_{i-1,j}}{h^2} + \frac{r_j}{h^2}\left[\frac{2}{r_{j+1} + r_j}\,(\phi^h_{i,j+1} - \phi^h_{ij}) - \frac{2}{r_j + r_{j-1}}\,(\phi^h_{ij} - \phi^h_{i,j-1})\right] + V_0^2\,\alpha^2(\phi^h_{ij})\phi^h_{ij} = 0, \quad \text{in} \quad \Omega^h \tag{3.2a}$$

$$\phi^h_{ij} = 0, \quad \text{on} \quad \partial\Omega^h \tag{3.2b}$$

where $\Omega^h = \{nh,mh),\ 0 \leq nh \leq a,\ 0 \leq mh \leq b\}$.

3.2. General Strategy for Solving the Discretized Equations

Equation (3.2) has the trivial solution $\phi^h = 0$ for any V_0. This solution corresponds to a uniform flow and is not interesting physically. We seek solutions which represent vortex breakdown so that $\|\phi^h\|^2 \neq 0$, where

$$\|\phi^h\|^2 = h^2 \Sigma \phi_{ij}^2. \tag{3.3}$$

Iterating on equation (3.2) by any iterative method may lead us to the trivial solution. In order to rule out this possibility, we specify the norm of the discrete solution we want to find, while making free the swirl velocity parameter V_0.

To summarize, we solve equation (3.2) for (ϕ^h, V_0) under the constraint

$$\|\phi^h\|^2 = g_0, \tag{3.4}$$

where g_0 is given.

A relaxation scheme for (ϕ^h,V_0) in equation (3.2) together with the constraint (3.4) is described next.

3.3. Relaxation

Equations (3.2), (3.4) form a nonlinear system of equations for (ϕ^h,V_0). The relaxation used for this system has three steps: (i) a local process for

smoothing ϕ^h in equation (3.2); (ii) a global change to satisfy (3.4); and (iii) updating the swirl parameter V_0. That is, one relaxation consists of doing (i), (ii), and (iii) successively.

(i) local relaxation

Scan the point $(i,j) \in \Omega^h$ in lexicographic ordering; at each point (i,j) solve (3.2) approximately for ϕ^h_{ij} by applying one Newton iteration.

(ii) global step

Compute $\beta = \sqrt{g_0/\|\phi^h\|^2}$.

Then make the change

$$\phi^h_{ij} \leftarrow \beta\, \phi^h_{ij} , \qquad (i,j) \in \Omega^h.$$

(iii) updating V_0

Change V_0 such that the following equation holds

$$\langle L^h \phi^h + 4V_0^2\, \alpha^2(\phi^h)\phi^h,\ \phi^h\rangle = \langle f^h, \phi^h\rangle \tag{3.5}$$

where $L^h \phi^h$ is the discretization of $L\phi = r(\frac{1}{r}\phi_r)_r + \phi_{xx}$, $\langle \cdot,\cdot\rangle$ denotes the inner product, $\langle u,v\rangle = h^2 \sum_{ij} u_{ij} v_{ij}$, and f^h is the right-hand side of equation (3.2). (In a multigrid process f^h is nonzero on coarse grids.)

We now come to the description of the multigrid algorithm used to solve (3.2), (3.4) for (ϕ^h, V_0).

3.4.1. Basic Cycle:

Given a sequence of discretizations with mesh sizes $h_r > h_2 > \cdots > h_m$, where $h_k = 2h_{k+1}$. The h_k-grid equation is generally written as

$$L^k \phi^k = f^k \tag{3.6}$$

where L^k approximates L^{k+1} $(k < m)$ (e.g., they all are finite-difference approximations to the same differential operator). The algorithm for improving a given approximate solution $\tilde{\phi}^k$ to (3.6) is denoted by

$$\tilde{\phi}^k \leftarrow MG(k, \tilde{\phi}^k, f^k) \tag{3.7}$$

and is defined recursively as follows:

If $k = 1$, solve (3.6) by several relaxation sweeps; otherwise do steps (A) - (D):

(A) Perform ν_1 relaxation sweeps on (3.6), resulting in a new approximation $\bar{\phi}^k$.

(B) Starting with $\tilde{\phi}^{k-1} = I_k^{k-1} \bar{\phi}^k$, perform one cycle $\tilde{\phi}^{k-1} \leftarrow MG(k-1, \tilde{\phi}^{k-1}, L^{k-1} \tilde{\phi}^{k-1} + \bar{I}_k^{k-1}(f^k - L^k \bar{\phi}^k))$.

(C) Calculate $\bar{\bar{\phi}}^k = \bar{\phi}^k + I_{k-1}^k(\tilde{\phi}^{k-1} - I_k^{k-1} \bar{\phi}^k)$.

(D) Perform ν_2 additional relaxation sweeps on (3.6) starting with $\bar{\bar{\phi}}^k$ and yielding the final $\tilde{\phi}^k$ of (3.7).

In this algorithm I_k^{k-1}, $\bar{I}_k^{k-1}$ are fine-to-coarse grid transfer operators; I_{k-1}^k is an interpolation operator. We refer to the above cycle as $MG(\nu_1,\nu_2)$. In the notation of this section (3.6) includes both equations (3.2) and (3.4).

The basic cycle described above is for improving a given approximation on level k. The full multigrid (FMG) process involves solving the problem on the coarsest grid, interpolating it to finer grids, and making the cycle $MG(\nu_1,\nu_2)$ a few times after each refinement.

3.4.2. Full Multigrid Algorithm (FMG)

1. Solve (3.6) for $k = 1$, using a continuation method (see remark below).

2. Set $k = k + 1$ and
   $\tilde{\phi}^k = \Pi_{k-1}^k \tilde{\phi}^{k-1}$, where Π_{k-1}^k is a bicubic interpolation.

3. Perform $\gamma(k)$ times the cycle
   $\tilde{\phi}^k \leftarrow MG(k, \tilde{\phi}^k, f^k)$.

4. If $k < m$, go to step 2; otherwise stop.

A Remark on Step 1 of the FMG Algorithm (Continuation Method)

Since the problem involved is a nonlinear one, and we are using a Newton iteration, a good initial approximation may be needed to get fast convergence for $k = 1$ (the coarsest grid). This has been achieved by using a continuation process where we solve first for a small norm $\|\phi^h\|^2$, then gradually increasing it until the prescribed norm is obtained. Each time the norm is increased, the solution of the previous step is used as an initial

approximation. In order to get a good initial approximation for the smallest-norm problem, we have solved for the bifurcation point from the trivial branch of solutions.

3.5 Solving for the Bifurcation Point

At a bifurcation point (ϕ^*, V_0^*), the linearized problem of (3.1) must have a zero eigenvalue, and the corresponding eigenfunction gives rise to a second branch of solutions. Since $\phi = 0$ is a solution for any V_0, we may try to find a bifurcating branch from the trivial one $(0,V_0)$. The linearized equations around $(0,V_0)$ are given by

$$W_{xx} + r\left(\frac{1}{r} W_r\right)_r + 4V_0^2 \tilde{\alpha}^2(0)W = 0, \qquad \text{in } \Omega \tag{3.8a}$$

$$W = 0, \qquad \text{on } \partial\Omega. \tag{3.8b}$$

If there exists a bifurcating branch from the trivial one $(0,V_0)$, equation (3.8) has a solution (W^*,V_0^*) with $\|W^*\|_2 = 1$ where $\| \ \|_2$ denotes the L_2 norm.

We discretize (3.8) in a way similar to the discretization of (3.1). The constraint

$$\|W^h\|^2 = 1,$$

is added to ensure a non-zero solution to the problem. The process of solving the eigenvalue problem is identical to the process of solving (3.2), (3.4).

Once this linear eigenvalue problem is solved, we can use $\phi_0 = \pm\varepsilon W$ as an initial approximation for our original problem with a prescribed norm of ε. The sign is chosen such that ϕ_0 has negative values, to ensure that the total streamfunction $\psi = \frac{r^2}{2} + \phi$ will have closed streamlines with negative values (the bubble).

4. NUMERICAL RESULTS

Experiments were performed with equations (3.2), (3.3) using FMG algorithm of Section 3.4.2. In these experiments the domain was

$$\Omega^h = \{(nh, \ell h),\ 0 < nh < 5,\ 0 < \ell h < 2\}.$$

Three levels were used in the multigrid algorithm where the finest grid problem has mesh size 1/16. On the coarsest level 20 relaxations were performed while on finer grids $\nu_1 = \nu_2 = 3$, $\gamma(k) = 4$. In all numerical experiments $I_k^{k-1} = \bar{I}_k^{k-1}$ is injection, I_k^{k-1} is bilinear interpolation, and Π_{k-1}^k is bicubic interpolation.

Tables I-IX contain the L_2-norm of the residuals and the values of V_0^2 at the end of each cycle on the finest grid. Cycle #0 refers to the approximation obtained from the previous level as an initial guess. Figures 1-9 show the streamlines (contours of ψ) for the different cases. The value of V_0^*, the swirl parameter value for which bifurcation occurs is $V_0^* = 1.0069$ (computed on coarsest level).

The experiments clearly show that the multigrid method suggested is very efficient. In fact, as seen by the convergence history for V_0^2, it is enough

to take $\gamma(k) = 2$, instead of $\gamma(k) = 4$, i.e., by 2 FMG cycles the problem is already solved.

The results show that bigger bubbles are obtained for smaller swirl parameters, contradicting to what one would expect. This may be the result of the assumption made in the model, that the same functional dependence of k, ω on ψ holds inside as well as outside the bubble. A future study will investigate this point by solving the full systems (2.1), making no extra assumptions.

REFERENCES

[1] J. H. Bolstad, H. B. Keller, "A Multigrid Continuation Method for Elliptic Problems with Turning Points," to appear in SIAM J. Sci. Stat. Comput.

[2] A. Brandt, Multigrid Techniques: 1984 Guide with Applications to Fluid Dynamics. Monograph available as GMD-Studie No. 85, GMD-F1T, Postfach 1240, D-5205, St. Augustin 1, West Germany.

[3] M. M. Hafez and M. D. Salas: "Vortex Breakdown Simulation Based on a Nonlinear Inviscid Model," Proceedings of ICASE/NASA Workshop on Vortex Dominated Flows, (M. Y. Hussaini and M. D. Salas, eds.), Springer-Verlag, 1986.

[4] S. Leibovich: "Vortex Stability and Breakdown: Survey and Extension," AIAA J., Vol. 22, No. 9, 1984, pp. 1192-1206.

[5] K. Stüben and U. Trottenberg: "Multigrid Methods: Fundamental Algorithms, Model Problem Analysis and Applications," in Multigrid Methods, Lecture Notes in Mathematics, No. 960, (W. Hackbusch and U. Trottenberg, eds.), Springer-Verlag, 1982.

Table I. $\|\phi^h\|^2 = .005$

cycle #	$\|\text{Residuals}\|_2$	V_0^2
0	.362 (-1)	.95088
1	.986 (-3)	.96069
2	.843 (-4)	.96039
3	.148 (-4)	.96041
4	.745 (-5)	.96042

Table II. $\|\phi^h\|^2 = .05$

cycle #	$\|\text{Residuals}\|_2$	V_0^2
0	.948 (-1)	.68322
1	.232 (-2)	.68962
2	.251 (-3)	.68939
3	.113 (-3)	.68941
4	.918 (-4)	.68941

Table III. $\|\phi^h\|^2 = .11$

cycle #	$\|\text{Residuals}\|_2$	V_0^2
0	.122	.59214
1	.233 (-2)	.54739
2	.215 (-3)	.54732
3	.615 (-4)	.54733
4	.542 (-4)	.54733

Table IV. $\|\phi^h\|^2 = .15$

cycle #	$\|\text{Residuals}\|_2$	V_0^2
0	.135	.48347
1	.243 (-2)	.48803
2	.168 (-3)	.48798
3	.474 (-4)	.48798
4	.425 (-4)	.48798

Table V. $\|\phi^h\|^2 = .2$

cycle #	$\|\text{Residuals}\|_2$	V_0^2
0	.150	.42902
1	.242 (-2)	.43301
2	.193 (-3)	.43294
3	.366 (-4)	.43294
4	.266 (-4)	.43294

Table VI. $\|\phi^h\|^2 = .4$

cycle #	$\|\text{Residuals}\|_2$	V_0^2
0	.192	.30435
1	.271 (-2)	.30725
2	.239 (-3)	.30719
3	.177 (-3)	.30719
4	.176 (-3)	.30719

Table VII. $\|\phi^h\|^2 = .6$

cycle #	$\|\text{Residuals}\|_2$	V_0^2
0	.230	.24006
1	.303 (-2)	.24335
2	.218 (-3)	.24231
3	.188 (-3)	.24231
4	.175 (-3)	.24231

Table VIII. $\|\phi^h\|^2 = 1.0$

cycle #	$\|\text{Residuals}\|_2$	V_0^2
0	.295	.17139
1	.385 (-2)	.17303
2	.363 (-3)	.17302
3	.294 (-3)	.17302
4	.278 (-3)	.17302

Table IX. $\|\phi^h\|^2 = 2.0$

cycle #	$\|\text{Residuals}\|_2$	V_0^2
0	.428	.10176
1	.701 (-2)	.10276
2	.777 (-3)	.10275
3	.584 (-3)	.10275
4	.574 (-3)	.10275

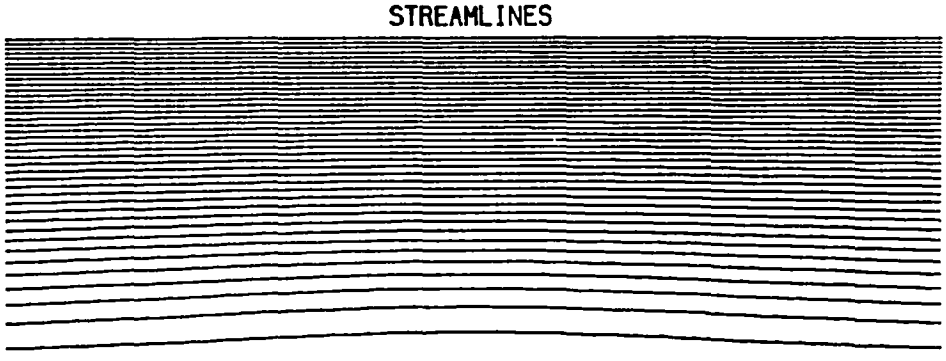


Figure 1. $\|\phi^h\|^2 = .005$, $V_0^2 = .96042$.

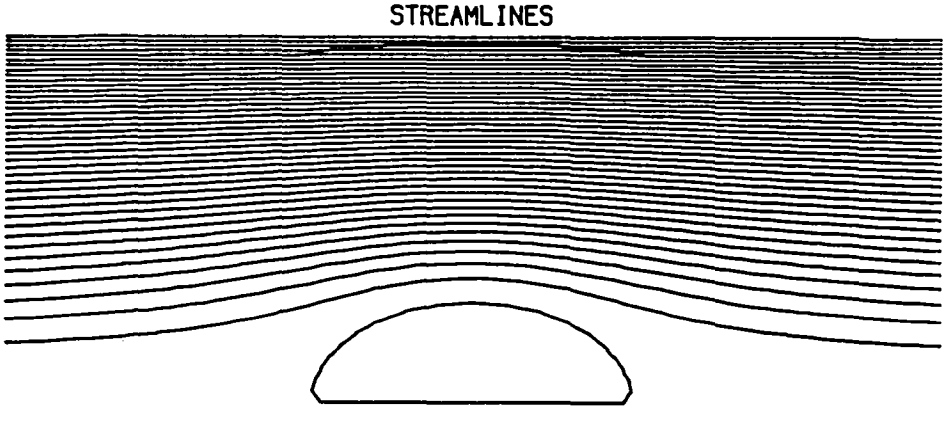


Figure 2. $\|\phi^h\|^2 = .05$, $V_0^2 = .68941$.

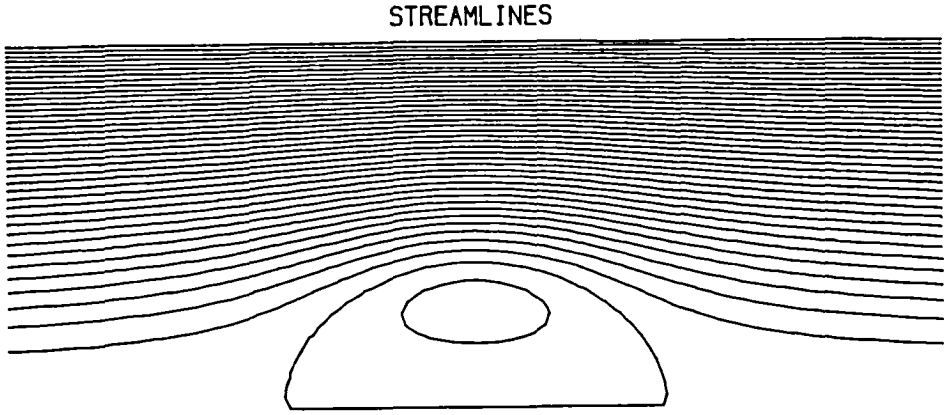


Figure 3. $\|\phi^h\|^2 = .11$, $V_0^2 = .54733$.

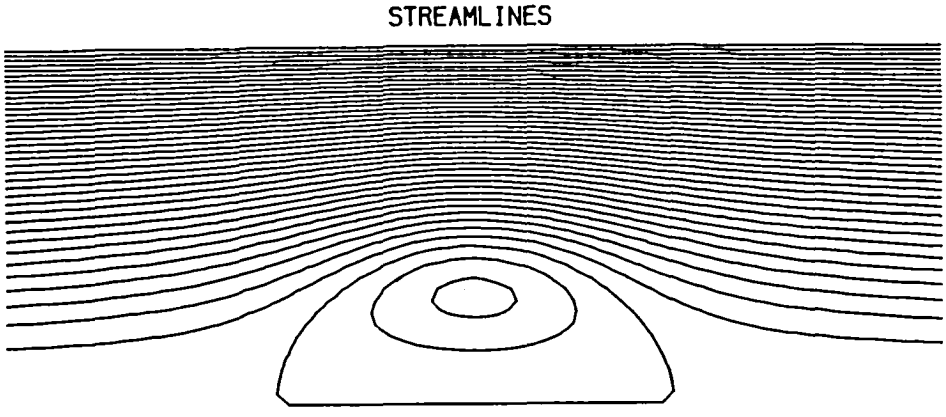


Figure 4. $\|\phi^h\|^2 = .15$, $V_0^2 = .48798$.

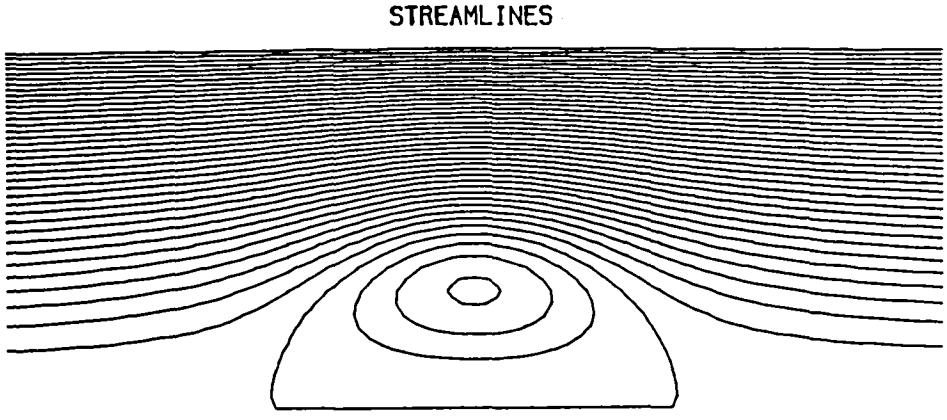


Figure 5. $\|\phi^h\|^2 = .2$, $V_0^2 = .43294$.

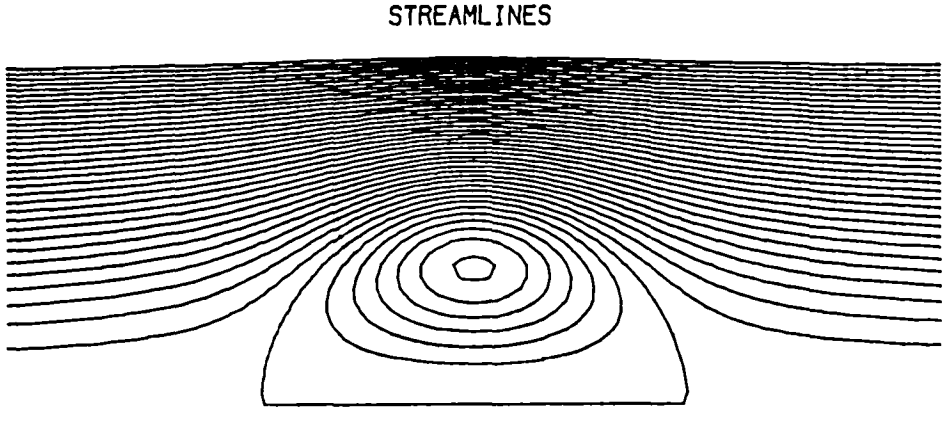


Figure 6. $\|\phi^h\|^2 = .4$, $V_0^2 = .30719$.

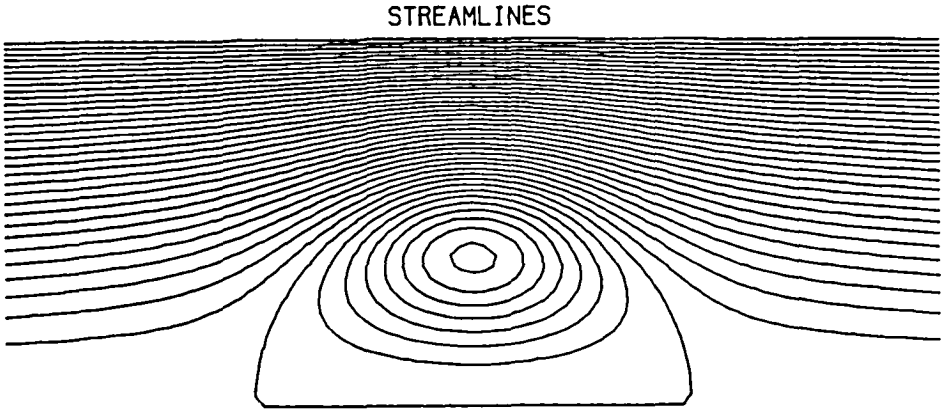


Figure 7. $\|\phi^h\|^2 = .6$, $V_0^2 = .24231$.

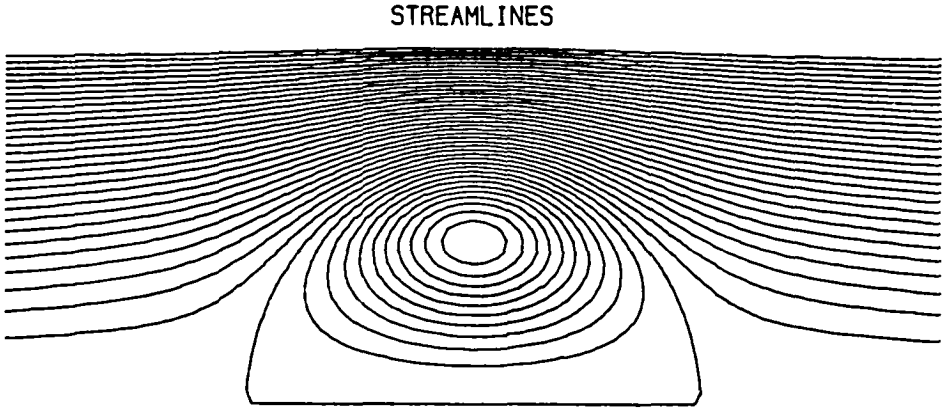


Figure 8. $\|\phi^h\|^2 = 1.0$, $V_0^2 = .17302$.

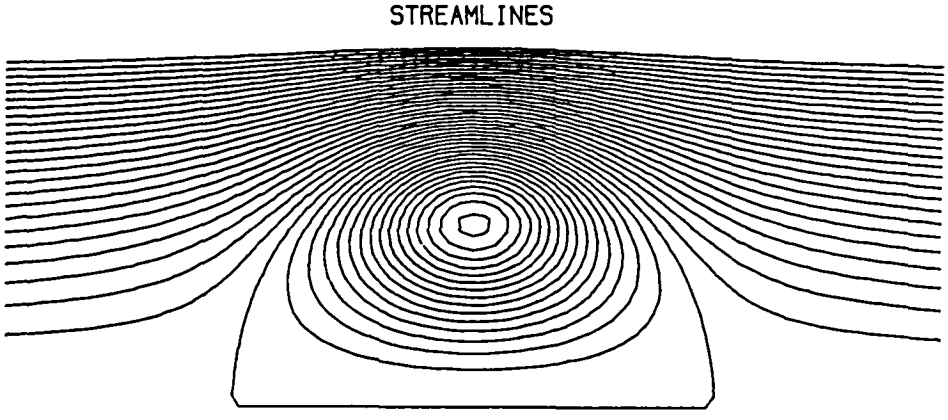


Figure 9. $\|\phi^h\|^2 = 2.0$, $V_0^2 = .10275$.

CONSTRUCTION OF HIGHER ORDER ACCURATE VORTEX AND PARTICLE METHODS

R. A. Nicolaides
Carnegie-Mellon University

ABSTRACT

The standard point vortex method has recently been shown to be of high order of accuracy for problems on the whole plane, when using a uniform initial subdivision for assigning the vorticity to the points. If obstacles are present in the flow, this high order deteriorates to first or second-order. This paper introduces new vortex methods which are of arbitrary accuracy (under regularity assumptions) regardless of the presence of bodies and the uniformity of the initial subdivision.

This work was supported by the Air Force Office of Scientific Research under Grant AFOSR-84-0137.

1. INTRODUCTION

There has been a growing interest recently in the theory and application of point vortex methods to the numerical solution of the incompressible Euler and Navier-Stokes equations. The impetus for the Euler case stems from the basic work of Dushane [6], Hald and Del Prete [7], and Hald [8], the Fourier analysis of Beale and Majda [1], [2], [3], and the Sobolev space approach of Raviart [12] and Cottet [4]. A recent paper by Cottet and Gallic [5] extends the latter approach to linear Burger's type equations with "viscosity" accounted for by splitting the convection and viscous parts and using a Green's function for the viscous computation. A method for introducing viscosity into particle methods for compressible flows is given by Monaghan and Gingold [9]. See also [10] and [11]. Apart from the first three of these references, the authors all obtain high order of accuracy error estimates, limited mainly by the regularity of the exact solution of the continuous equations. Unfortunately, the possibility of obtaining this accuracy is dependent on the existence of expansions similar in nature to the Euler-MacLaurin sum formula. If, for any reason, it is not possible to assert the existence of such expansions, the accuracy drops to first- or second-order, depending on the exact details of the algorithm and which errors are being estimated. If general boundaries (bodies) are present in the flow field, or if the initial subdivision of the flow field is not uniform, the necessary expansions will most likely cease to exist. Then questions arise as to how higher-order schemes may be constructed, and more important whether it is worthwhile to use them in view of the extra expense which is involved. The purpose of the paper is to give some possible answers to these questions.

In Section 2, the basic equations are given, and the simplest particle method is defined for comparison with some higher-order schemes. These schemes are introduced in Section 3. There, three methods for generating schemes of arbitrary accuracy are provided. An appendix contains some technical results about solving scalar hyperbolic equations with distributional data.

This paper is of an algorithmic nature and does not contain numerical results or precise error estimates. These will appear elsewhere.

2. MODEL PROBLEM

The incompressible Euler Equations in vorticity-velocity form are

$$\left.\begin{aligned} \omega_t + (u\omega)_x + (v\omega)_y &= 0 \qquad (2.1) \\ \operatorname{div}(u,v) = 0 \;:\; \operatorname{curl}(u,v) &= \omega \qquad (2.2) \end{aligned}\right\} \text{ in } \mathbb{R}^2$$

with initial condition

$$\omega(x,y,0) = \omega_0(x,y). \qquad (2.3)$$

The basic ideas for constructing higher-order schemes will be shown for (2.1) and (2.3), with (u,v) assumed given. For these linear problems it is not necessary to assume that (u,v) is solenoidal.

In this setting, we will now define the basic particle (or point vortex) method. Subdivide the plane into squares of side h, number the squares 1, 2, 3,... in some convenient way and define a distributional approximation

to $\omega_0(x,y)$ by

$$\omega_{0h}(x,y) = \sum_i h^2\, \omega(x_i,y_i)\, \delta(x-x_i,\ y-y_i) \tag{2.4}$$

where (x_i,y_i) denotes the center of the i^{th} mesh square, and $\delta(x-x_i,y-y_i)$ denotes the Dirac delta function with pole at (x_i,y_i). Now solve (2.1) and (2.2) with $\omega_0(x,y) \leftarrow \omega_{0h}(x,y)$. The well known solution to the latter problem is the distribution

$$\omega_h(x,y,t) \equiv \sum_i h^2\, \omega(x_i,y_i)\, \delta\big(x-\ X(x_i,y_i;t),\ y\ -\ Y(x_i,y_i;t)\big) \tag{2.5}$$

where $X(x_i,y_i,t)$ denotes the solution of the characteristic equation

$$dX/dt = u(X,Y,t) \qquad\qquad X(0) = x_i$$

and correspondingly for Y.

No use is made of the uniformity of the mesh in deriving (2.5). For a nonuniform mesh, h^2 in (2.5) is the area of the appropriate mesh square. In the error formulas below, h denotes the largest mesh length.

It is immediately clear from this definition that the particle approximation is <u>non-dissipative</u>, in the sense that no artificial viscosity is introduced because after the discretization of the initial condition is made (2.1) is solved exactly. In practice some ODE solver must be used to compute the trajectories, but in theory its error can be made arbitrarily small. This principle, of solving the exact equation with approximte data, seems to be common to particle methods generally and distinguishes them from finite

difference and finite element methods. The latter, at least, solves an approximate equation with exact data.

A rigorous error analysis of the method just defined can be found in [12]. This analysis is too complicated to reproduce here. Nevertheless, we need some simple guide to compare the accuracy of various schemes. It seems reasonable to look at the difference $\omega_0 - \omega_{0h}$ against a test function as a measure of "truncation error" since it is the only error made. Thus we define, for a given method of approximation and a given function ω_0 with compact support $\overline{\Omega}$ (where area $(\overline{\Omega}) = 1$ say)

$$\tau_h(\phi) = \iint (\omega_0 - \omega_{0h})\phi dxdy. \tag{2.6}$$

Here, the integration is performed over $\mathbb{R}^2$. The restriction that ω_0 has compact support is a matter of convenience rather than necessity and could be replaced by sufficiently rapid decay at large distances from the origin.

As an example, consider (2.4). Then we find

$$\tau_h(\phi) = \iint \omega_0\, \phi dxdy - \sum_i h^2(\omega_0\, \phi)(x_i,y_i). \tag{2.7}$$

This shows that a midpoint rule numerical integration is being used to approximate the integral, and under smoothness conditions it follows that as $h \to 0$

$$\tau_h(\phi) = O(h^2).$$

Clearly, higher-order integration formulas can be compared with each other on this basis. For a 2×2 product Gauss rule in each element, for example, we have $\tau_h = O(h^4)$.

Next, recall the important fact that in the nonlinear case it is necessary to compute the velocity field at each timestep by solving (2.2). Assume that this is to be done using the Green's function. Let W denote the number of arithmetical operations required to compute the velocity field at each particle position. If there are N particles, then $W \simeq CN^2/2$, for some constant C. Below, we will use W as a standard unit of work to compare various new algorithms. For the Gauss case therefore we have a work count of $16W$. From this we see that use of a higher-order rule does not necessarily assure a greater computational efficiency for typical values of h. In the next section, methods for obtaining high-order accuracy without such a large increase in the cost of the computation are defined.

3. HIGHER ORDER METHODS

The preceding remarks suggest that increasing the order of accuracy by adding more integration nodes may not be a good idea. It is natural to try to do the same thing by increasing the amount of information associated with each node. Specifically, in this section we shall associate with (x_i,y_i), m^{th} order distributions of the form

$$M_i(x,y) \equiv \sum_{|\alpha|\leq m} w_{i\alpha} D^{\alpha} \delta(x-x_i, y- y_i). \tag{3.1}$$

In (3.1), which generalizes the simple δ functions in (2.4), α denotes a multi-index, and $(x_i,y_i) \in \mathbb{R}^2$. Choice of the weights $w_{i\alpha}$ and the nodes (x_i,y_i) can be made in many ways. We shall give three methods in this section.

Method 1 (Direct Integration):

In this method, (x_i,y_i) are the corners of the elements, each of which has associated with it an expansion of the form (3.1). The weights in the expansion are chosen so that when ω_{0h} is substituted into (2.6), the second term gives a rule for integration of the function $(\omega_0 \phi)$, involving its values along with those of its derivatives through order m at the nodes. We shall consider the cases $m = 0$ and $m = 1$ in more detail.

Let $m = 0$. A rule for a square of side h with corners at P, Q, R, S which is exact for bilinear functions is

$$\iint f \, dxdy \simeq (h^2/4) \left(f(P) + f(Q) + f(R) + f(S)\right). \tag{3.2}$$

Using this as a composite rule implies the choice $w_{i00} = h^2 \omega(x_i,y_i)$ so that we define

$$M_i(x,y) \equiv h^2\omega(x_i,y_i) \, \delta(x-x_i, \, y-y_i). \tag{3.3}$$

Since this gives a rule which is locally exact for linear functions but not for all quadratics its accuracy is $0(h^2)$ in the sense of (2.6) while the work is 1W. This is essentially no different from the mid-point rule. In fact this rule is clearly analogous to the trapezoidal rule.

For a quadrilateral mesh, a bilinear mapping can be used to map the quadrilaterals onto a standard square in which (3.1) can be used. In some circumstances it may be desirable to use a triangular mesh instead of the quadrilateral one. An $0(h^2)$ rule for triangles analogous to (3.1) can then be used, avoiding the need to map the domains.

Now let $m = 1$. Analogous to (3.2) we have the formula

$$\begin{aligned}\iint f\,dxdy \approx\ & A\big(f(P) + f(Q) + f(R) + f(S)\big) \\ & + B\big(-f_x(P) + f_x(Q) + f_x(R) - f_x(S)\big) \\ & + C\big(-f_y(P) - f_y(Q) + f_y(R) + f_y(S)\big)\end{aligned} \tag{3.4}$$

where $A = h^2/4$, $B = C = h^3/24$, and P, Q, R, S denote the corners of the square $-h/2 \le x, y \le h/2$ labelled counterclockwise starting from the top right. Analogous to (3.3) there is the expression

$$M_i(x,y) \equiv \sum_{|\alpha|\le 1} w_{i\alpha}\, D^{\alpha}\, \delta(x-x_i,\ y-y_i). \tag{3.5}$$

In (3.5), the coefficients are computed from the composite rule based on (3.4). For the uniform square mesh we are using for illustration, the weights are

$$w_{i00} = A'\omega_0(x_i,y_i) + B'\omega_{0x}(x_i,y_i) + C'\omega_{0y}(x_i,y_i)$$

$$w_{i10} = -B'\omega_0(x_i,y_i)$$

$$w_{i01} = -C'\omega_0(x_i,y_i).$$

(3.4) is exact for cubic polynomials. It follows that this method is accurate in the sense of (2.6) to $O(h^4)$. To compute work units for this scheme, we observe that although there are only $\approx N$ particles there is some extra work associated with computation of derivatives of the velocity kernel. It turns

out that for this scheme the work units are $< 2\frac{1}{2} W$, a satisfactory figure. There is also some additional work required for computing the coefficients of the derivatives in (3.1). This amounts to having to integrate two more systems each of two odes, in addition to the characteristic odes (see appendix).

As in the previous case, rather than use a quadrilateral mesh it might sometimes be better to use a triangular one.

For a square mesh, the $m = 1$ scheme just discussed has an interesting property in the uniform case. This is the following: due to cancellations, the composite rule has weights of zero attached to the derivative unknowns at interior vertices. Hence the higher accuracy is achieved by corrections at the boundary. But this implies the use of a Euler-Maclaurin type expansion. Thus, if $\omega\phi$ has s continuous derivatives in $\mathbb{R}^2$ and compact support, by using nodal derivatives up to this order we can get accuracy $O(h^{s+1})$ merely by using the $m = 0$ scheme, since this is what the composite scheme reduces to on a uniform mesh in that case. This is another way to look at the results of [1] - [3].

Method 2 (Finite Element Approach):

The approach here uses a nodal finite element basis in the following way: let $\{\psi_{i\alpha}\}$ $|\alpha| \leq m$, $i = 1,2,\cdots,$ be the standard nodal basis functions associated with the i^{th} node (x_i,y_i) of a triangulation of the plane with maximum edge length h. These functions satisfy conditions of the form

$$D^{\beta} \psi_{i\alpha}(x_j,y_j) = \Delta_{ij}^{\alpha\beta},$$

where $\Delta_{ij}^{\alpha\beta}$ is a Kronecker delta. Then we define $w_{i\alpha}$ as

$$w_{i\alpha} = (-1)^{|\alpha|} \iint \psi_{i\alpha}(x,y)\ \omega_0(x,y)dxdy \tag{3.6}$$

where the integration is over the whole plane. We now have

$$\begin{aligned} \iint \omega_{0h}(x,y)\ \rho(x,y)dxdy &= \iint \sum_i \sum_{|\alpha|\leq m} w_{i\alpha}\ D^{\alpha}\delta(x-x_i,\ y-y_i) \\ &\qquad \times \phi(x,y)dxdy, \qquad \forall\ \phi \in C^m(\mathbb{R}^2) \\ &= \sum_i \sum_{|\alpha|\leq m} (-1)^{|\alpha|}\ w_{i\alpha}\ D^{\alpha}\ \phi(x_i,y_i) \\ &= \iint \omega_0(x,y)\ \phi^h(x,y)dxdy \end{aligned} \tag{3.7}$$

where ϕ^h is the finite element interpolant of ϕ on the given triangulation. Equation (2.6) then becomes

$$\tau_h(\phi) = \iint \omega_0(\phi - \phi^h)dxdy. \tag{3.8}$$

Since the error $|\phi - \phi^h|$ is formally $O(h^{r+1})$ where r is the degree of the highest order full polynomial space used, we can say here that τ_h is of this order.

This type of scheme differs from direct integration schemes in that no approximation of ω_0 is made. The test function only (often a convolution kernel in practice) is approximated and the result is integrated exactly. Because of this property, the rigorous error estimates for these methods

require minimal regularity on ω_0 unlike the direct integration methods where to achieve high accuracy requires ω_0 to have several smooth derivatives throughout $\mathbb{R}^2$. The $O(h^{r+1})$ estimate is in fact valid even if we know only $\omega_0 \in L^1(\mathbb{R}^2)$. If ω_0 has extra regularity it can be exploited to get higher accuracy by going to negative norm estimates of the finite element error. Smoothness of ϕ, however, is certainly required.

Two examples analogous to those considered above are the case of continuous linear elements on triangles, for which we can expect $O(h^2)$ accuracy with 1W work units, and full cubics - defined in terms of derivative unknowns at vertices, and function values at vertices and centroid for which the work will be somewhat larger than the values used so far (about $10\frac{1}{2}$ W units).

In general, the full range of finite element spaces is available for use.

Method 3 (Taylor/Moment Expansions):

Here we begin by subdividing the plane into arbitrary elements with mid-side nodes and arbitrary element shapes allowed in principle. Next, we define

$$\alpha_1!\ \alpha_2!\ w_{i\alpha} = (-1)^{|\alpha|} \iint (x-x_i)^{\alpha_1} (y-y_i)^{\alpha_2} \omega_0(x,y)dxdy \qquad (3.9)$$

in which (x_i,y_i) is an arbitrary point within the i^{th} element, and the integration is over the ith element. The $w_{i\alpha}$ are proportional to the moments of ω_0 restricted to the i^{th} element, about (x_i,y_i). It follows as above, that

$$\iint \omega_{0h}(x,y)\ \phi(x,y)dxdy = \iint \omega_0(x,y)\ \phi^{[m]}(x,y)dxdy \qquad (3.10)$$

where $\phi^{[m]}(x,y)$ is the piecewise polynomial function, in general discontinuous, equal in the i^{th} element to the Taylor expansion of $\phi(x,y)$ through m^{th} order terms, about the point (x_i,y_i). In this sense the local moment expansion defined by (3.1) and (3.9) "dualizes" into the local Taylor expansion about (x_i,y_i).

To get the accuracy of this scheme, we substitute into (2.6) to find that

$$\tau_h(\phi) = \iint \omega_0(\phi - \phi^{[m]})dxdy$$

so that denoting by h the largest linear dimension of the elements, we obtain accuracy $O(h^{m+1})$.

The moments method also needs only minimal regularity on ω_0 for full accuracy to be obtained. In practice, if $m = 1$ the point (x_i,y_i) should be chosen to be the center mass of ω_0 because then $w_{i\alpha} = 0$ for $|\alpha| = 1$, so we get second-order accuracy for the same work as with the lowest-order scheme. Using quadrilaterals for elements, with N vertices there are approximately N elements and so N particles. For $O(h^3)$ accuracy the interaction work count is $5W$, and for $O(h^4)$ is $8W$.

4. FURTHER REMARKS

There should be no difficulty in extending the ideas of Section 3 to three-dimensional particle methods of the kind suggested in [1] - [3] and [12].

Rigorous analysis using the Sobolev space setting has been carried out for both the finite element and moment expansion methods.

So far an insufficient amount of computation has been done to verify the error estimates and decide about the efficiency of the various methods.

ACKNOWLEDGEMENTS

Thanks to Chichia Chiu and Shenaz Choudhury for their help with this paper.

APPENDIX

A framework for finding distributional solutions of (2.1) with initial condition $\omega_{0h} = D^{\alpha}\,\delta(x-x_0,\ y-y_0)\ |\alpha| \leq m$ can be obtained starting from the following considerations. Let $X(x_0,y_0;t)$ and $Y(x_0,y_0;t)$ denote the characteristic curves of the equation (2.1); here, t parameterizes the curve and the generic point (x_0,y_0) denotes its origin at time $t = 0$. X and Y are computed by solving the ordinary differential equations

$$dX/dt = u(X,Y,t) \qquad dY/dt = v(X,Y,t)$$

$$X(0) = x_0 \qquad Y(0) = y_0.$$

At time t, let $J(x_0, y_0; t)$ denote the Jacobian of the flow map $\Phi : (x_0,y_0) \to (X,Y)$. The (nonlinear) case of most interest from the fluids viewpoint has $u_x + v_y = 0$, in which case $J(x_0,y_0;t) = 1$. We can obtain a formal analytical solution to (2.1) and (2.3) by writing the equation in terms of the material derivative as $d\omega/dt = 0$, integrating this equation over an arbitrary domain moving with the velocity field (u,v), say $\Omega(t)$, and then using the transport theorem to write

$$d/dt \iint_{\Omega(t)} \omega(X,Y,t)dXdY = 0,$$

from which it follows immediately that

$$\iint_{\Omega(t)} \omega(X,Y,t)dXdY = \iint_{\Omega(0)} \omega_0(x,y)dxdy.$$

Changing the variables on the right-hand side to X and Y respectively and recalling the arbitrariness of $\Omega(t)$ now gives

$$\omega(X,Y;t) = \omega_0\big(x(X,Y,t),\ y(X,Y,t)\big)J^{-1}\ (X,Y;t) \tag{A.1}$$

where $(x(X,Y,t),\ y(X,Y,t))$ is by inverting the equations $X = X(x,y:t)$, $Y = Y(x,y;t)$. The existence of a unique solution to these equations follows from ode theory provided u and v are smooth. Reversing the steps, it follows that (A.1) satisfies (2.1) given the required regularity of u, v, and ω_0.

Let $\phi \in C^m(\mathbb{R}^2)$; multiplying (A.1) by ϕ, integrating and changing the variables on the right to x and y we have

$$\iint \omega(X,Y,t)\ \phi(X,Y)dXdY = \iint \omega_0(x,y)\ \phi\big(X(x,y;t),\ Y(x,y,t)\big)dxdy, \tag{A.2}$$

or alternatively

$$\langle\omega,\ \phi\rangle = \langle\omega_0,\ \phi\circ(X,Y)\rangle \tag{A.3}$$

where $\circ$ denotes composition. If $X(\cdot,\cdot,\ t)$ and $Y(\cdot,\cdot,\ t)$, $Y(\cdot,\cdot,\ t) \in W^{m+1,\infty}(\mathbb{R}^2)$ (or $\in \mathcal{C}^{(m)}(\mathbb{R}^2)$), $\forall\ 0 \le t \le T$, then the right-side of (A.3) makes sense even if $\omega_0 + \omega_{0h} = D^\alpha\ \delta(x-x_0,\ y-y_0)|\alpha| \le m$. Thus a distribution ω is defined on $\mathcal{C}^{(m)}(\mathbb{R}^2)$ by (A.3). Therefore, we can pose the problem of finding ω_h satisfying

$$\langle\omega_h,\ \phi\rangle = \langle\omega_{0h},\ \phi\circ(X,Y)\rangle \qquad \forall\ \phi \in \mathcal{C}^{(m)}\ (\mathbb{R}^2). \tag{A.4}$$

A solution ω_h to (A.4) is given by

$$\omega_h(X,Y) = D^{\alpha}\, \delta(X - X(x,y;t),\ Y - Y(x,y;t))\Big|_{x=x_0,\ y=y_0}, \qquad (A.5)$$

the purely formal differentiations being performed w.r.t. x and y. Proof that (A.5) satisfies (A.4) is by direct computation.

If $|\alpha| = 0$ we recover the solution given in Section 2. Consider the case with $|\alpha| = 1$. Equation (A.5) gives

$$\begin{aligned} \omega_{h10} &= \delta_X(X-X_0,\ Y-Y_0)\ X_x(x_0,y_0,t) + \delta_Y(X-X_0,\ Y-Y_0)\ Y_x(X_0,y_0,t) \\ \omega_{h01} &= \delta_X(X-X_0,\ Y-Y_0)\ X_y(x_0,y_0,t) + \delta_Y(X-X_0,\ Y-Y_0)\ Y_y(x_0,y_0,t) \end{aligned} \qquad (A.6)$$

using the abbreviation X_0 for $X(x_0,y_0;t)$ and similarly Y_0. If the initial condition is

$$\omega_{h0} = a_{10}\ \delta_x(x-x_0,\ y-y_0) + a_{01}\ \delta_y(x-x_0,\ y-y_0),$$

then the solution to (A.4) of the required form as given by (A.6) is

$$\omega_h = a_{10}(t)\ \delta_X(X-X_0,\ Y-Y_0) + a_{01}(t)\ \delta_Y(X-X_0,\ Y-\ Y_0)$$

where

$$\begin{aligned} a_{10}(t) &= a_{10}\ \ X_x(x_0,y_0,t) + a_{01}\ X_y(x_0,y_0,t) \\ a_{01}(t) &= a_{10}\ Y_x(x_0,y_0,t) + a_{01}\ Y_y(x_0,y_0,t). \end{aligned} \qquad (A.7)$$

Letting M denote the matrix

$$\begin{bmatrix} X_x & X_y \\ Y_x & Y_y \end{bmatrix}$$

differentiation of the characteristic equations shows that

$$dM/dt = \nabla(u,v)M$$

and the initial condition for this system is $M(0) = 1$, the identity matrix. It will be necessary to solve this and analogous systems for the higher-order cases in order to compute the numerical approximations. Having solved it, $a_{10}(t)$ and $a_{01}(t)$ are given by (A.7).

REFERENCES

[1] J. T. Beale and A. J. Majda, "Vortex Methods 1: Convergence in Three Dimensions," Math. Comp., Vol. 39, 1982, pp. 1-27.

[2] J. T. Beale and A. J. Majda, "Vortex Methods 2: Higher Order Accuracy in Two and Three Dimensions," Math. Comp., Vol. 39, 1982, pp. 29-52.

[3] J. T. Beale and A. J. Majda, "Higher Order Accurate Vortex Methods with Explicit Velocity Kernels," J. Comp. Phys., Vol. 58, 1985, pp. 188-208.

[4] G. H. Cottet, "Methodes Particulaires Pour L'equation D'Euler dans Le Plan," These de 3e cycle, Univ. P. et M. Curie, Paris, 1982.

[5] G. H. Cottet and S. Gallic, "A Particle Method to Solve Transport-diffusion Equations," Report 115, Centre de Math. Appl., Ecole Polytechnique, 1985.

[6] T. E. Dushane, "Convergence of a Vortex Method for Solving Euler's Equation," Math. Comp., Vol. 27, 1973, pp. 719-728.

[7] O. Hald and V. M. Del Prete, "Convergence of Vortex Methods for Solving Euler's Equations," Math. Comp., Vol. 32, 1978, pp. 791-809.

[8] O. Hald, "Convergence of Vortex Methods II," SIAM J. Numer. Anal., Vol. 16, 1979, pp. 726-755.

[9] J. J. Monaghan and R. A. Gingold, "Shock Simulation by the Particle Method SPH," J. Comp. Phys., Vol. 52, No. 2, November 1983, pp. 374-389.

[10] J. J. Monaghan and R. A. Gingold, "Kernel Estimates as a Basis for General Particle Methods in Hydrodynamics," J. Comp. Phys., Vol. 46, No. 3, June 1982, pp. 429-453.

[11] J. J. Monaghan, "Why Particle Methods Work," SIAM J. Sci. Stat. Comput., Vol. 3, No. 4, December 1982, pp. 422-433.

[12] P. A. Raviart, "An Analysis of Particle Methods," CIME course, Numerical Methods in Fluid Dynamics, Como (1983).

PSEUDO-TIME ALGORITHMS FOR THE NAVIER-STOKES EQUATIONS

R. C. Swanson
NASA Langley Research Center

E. Turkel
Tel-Aviv University, Israel
and
Institute for Computer Applications in Science and Engineering

ABSTRACT

A pseudo-time method is introduced to integrate the compressible Navier-Stokes equations to a steady state. This method is a generalization of a method used by Crocco and also by Allen and Cheng. We show that for a simple heat equation that this is just a renormalization of the time. For a convection-diffusion equation the renormalization is dependent only on the viscous terms. We implement the method for the Navier-Stokes equations using a Runge-Kutta type algorithm. This enables the time step to be chosen based on the inviscid model only. We also discuss the use of residual smoothing when viscous terms are present.

Research was supported in part by the National Aeronautics and Space Administration under NASA Contract Nos. NAS1-17070 and NAS1-18107 while the second author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665-5225.

I. INTRODUCTION

The solution of the compressible Navier-Stokes equations for flow about two- and three-dimensional complex aerodynamic configurations is still a time consuming problem on today's supercomputers. The resolution of the boundary layers requires the use of very fine meshes in the neighborhood of solid bodies. For a typical viscous flow the mesh can be several orders of magnitude finer (depending on the Reynolds number) than that required for an inviscid calculation. As an example, using a C-type mesh about an NACA 0012 airfoil, a typical mesh spacing near the body in the normal direction for an inviscid calculation is 1×10^{-2} chords. For a laminar viscous calculation with $Re = 5 \times 10^{3}$, this minimum cell height would be about 6×10^{-4} chords. For a turbulent calculation using an algebraic turbulence model and with $Re \simeq 3 \times 10^{6}$, the minimum cell height would be about 8×10^{-5} chords. In all cases a typical chordwise spacing at the midsection of the airfoil is about 5×10^{-2} chords.

Using an explicit method this fine mesh reduces the time step, due to stability requirements, that can be used. The time step restriction is caused by two factors. One contribution is due to the effect of the finer mesh on the inviscid portion of the calculation. When using an explicit method this reduction of the time step cannot be avoided without using a coarser mesh. It follows strictly from the need to include the entire domain of dependency in the numerical algorithm. Use of a local time step allows faster convergence to a steady state, but it does not remove the requirement to satisfy the convection stability condition in a local sense. A second difficulty is caused by the viscous terms. For an explicit method the time step is now dependent on the square of the mesh size rather than just the mesh size as

occurs for inviscid flow. Thus, even for a high Reynolds number flow the viscous time step will dominate when the mesh is sufficiently fine. In all these cases the use of an implicit scheme will alleviate the difficulties. In some ADI methods the Jacobian of the viscous terms is not used in the implicit portion of the code in order to improve the speed of the calculation [7]. We thus conclude that for both explicit and many implicit codes it is advantageous to account for the dependence of the time step on the viscous terms.

In this study we shall only discuss steady state problems which are solved by a pseudo time-dependent method. Hence, we can change all time derivatives as long as the steady state solution is not affected. One common device is to use a different time step in each zone. It is easier to calculate this local time step based on the inviscid equations. This provides an additional reason to eliminate the dependence of the time step on the viscous terms.

In this study we shall analyze a method used by Crocco [4] and also by Allen and Cheng [2]. They claim that the new scheme is unconditionally stable for a simple diffusion equation. We will show that in effect the scheme is a standard Euler forward-in-time central-in-space scheme. The time is artificially slowed down so as to satisfy the stability criterion. We then extend this scheme to the compressible Navier-Stokes equations using a Runge-Kutta scheme [9]. This modification enables us to choose our time step based on the inviscid equations. The modification automatically reduces the local time step in regions where the viscous time step is of importance. This enables us to use the inviscid time step in the far field while automatically accounting for viscous effects in the boundary layer. We will also look at residual smoothing for the heat equation.

II. SCALAR EQUATION

In this section we analyze and extend a scheme for the Navier-Stokes equations proposed by Crocco [4] and Allen-Cheng [2]. This scheme was also analyzed by Peyret and Viviand [6] and Roache [8], and we will extend their analysis.

We first consider the heat equation

$$w_t = \varepsilon w_{xx}. \tag{1}$$

The forward time centered space or Euler approximation to this scheme is given by

$$w_j^{n+1} = w_j^n + \frac{\varepsilon \Delta t}{(\Delta x)^2} \left(w_{j+1}^n - 2w_j^n + w_{j-1}^n\right). \tag{2}$$

This scheme is stable if

$$\nu = \frac{\varepsilon \Delta t}{(\Delta x)^2} \leq 1/2 \text{ , or } \Delta t \leq \frac{(\Delta x)^2}{2\varepsilon} \tag{3}$$

Crocco, and Allen/Cheng introduce the inconsistent scheme

$$w_j^{n+1} = w_j^n + \frac{\varepsilon \Delta t}{(\Delta x)^2} \left(w_{j+1}^n - 2w_j^{n+1} + w_{j-1}^n\right). \tag{4}$$

This scheme is unconditionally stable. If we are only interested in the steady state, then (4) yields the correct steady-state solution. We now rewrite (4) as

$$\frac{w_j^{n+1} - w_j^n}{\Delta t} = \frac{\varepsilon}{(\Delta x)^2} \left(w_{j+1}^n - 2w_j^n + w_{j-1}^n\right) - \frac{2\varepsilon}{(\Delta x)^2} \left(w_j^{n+1} - w_j^n\right)$$

or

$$\left(\frac{1}{\Delta\tau} + \frac{1}{\Delta t}\right)\left(w_j^{n+1} - w_j^n\right) = \frac{\varepsilon}{(\Delta x)^2}\left(w_{j+1}^n - 2w_j^n + w_{j-1}^n\right) \tag{5}$$

with

$$\Delta\tau = \frac{(\Delta x)^2}{2\varepsilon} . \tag{6}$$

Thus, for this model problem the Crocco scheme is identical with the Euler scheme (2) with an artificial time step Δt_e given by

$$\frac{1}{\Delta t_e} = \frac{1}{\Delta t} + \frac{2\varepsilon}{(\Delta x)^2} . \tag{7}$$

Thus, the unconditional stability is achieved by slowing down the time process. Note that as $\Delta t \to \infty$, $\Delta t_e \to (\Delta x)^2/2\varepsilon$, i.e., the stability limit for the Euler method. So choosing a large time step for (4) is equivalent to choosing Δt_e at the stability limit for (2), and we have merely scaled the time. This can also be derived from the modified equation given in [6]. If ε or Δx is not constant, this also introduces a local time step.

We next consider the convection-diffusion equation

$$w_t = aw_x + \varepsilon w_{xx} . \tag{8}$$

The Crocco scheme now becomes

$$\frac{w_j^{n+1} - w_j^n}{\Delta t} = \frac{a\left(w_{j+1}^n - w_{j-1}^n\right)}{2\Delta x} + \frac{\varepsilon}{(\Delta x)^2}\left(w_{j+1}^n - 2w_j^{n+1} + w_{j-1}^n\right) \tag{9}$$

or

$$\left(\frac{1}{\Delta\tau} + \frac{1}{\Delta t}\right)\left(w_j^{n+1} - w_j^n\right) = \frac{a\left(w_{j+1}^n - w_{j-1}^n\right)}{2\Delta x} + \frac{\varepsilon}{(\Delta x)^2}\left(w_{j+1}^n - 2w_j^n + w_{j-1}^n\right) \tag{10}$$

with $\Delta\tau$ given by (6). Thus, again this is equivalent to the Euler scheme with a time scaling that depends only on the viscous terms. Allen and Cheng utilized this scheme within a time-marching scheme proposed by Brailovskaya [3]. We generalize this by considering a general N-stage Runge-Kutta scheme.

Consider the two-dimensional equation

$$w_t = Hw + \varepsilon_1 w_{xx} + \varepsilon_2 w_{yy} \tag{11}$$

where Hw describes the hyperbolic or first-order terms. In [9] we describe a Runge-Kutta scheme where the viscous terms are frozen for all the stages. This is similar in philosophy to the Brailovskaya scheme. Using the Crocco formulation the $(K + 1)$-st stage becomes

$$\frac{w_{j,k}^{(K+1)} - w_{j,k}^{n}}{\alpha_{K+1}\,\Delta t} = H_D\, w_{j,k}^{(K)} + \frac{\varepsilon_1}{(\Delta x)^2}\left(w_{j+1,k}^{n} - 2w_{j,k}^{(K+1)} + w_{j-1,k}^{n}\right) + \frac{\varepsilon_2}{(\Delta y)^2}\left(w_{j,k+1}^{n} - 2w_{j,k}^{(K+1)} + w_{j,k-1}^{n}\right), \quad K=0,1,\cdots,N-1. \tag{12}$$

This reduces to a Runge-Kutta scheme

$$w_{j,k}^{(K+1)} = w_{j,k}^{n} + \alpha_{K+1}\,\Delta t_e\left[H_D\, w_{j,k}^{(K)} + P_D\, w_{j,k}^{n}\right] \tag{13}$$

where H_D, P_D are the approximations to the hyperbolic and parabolic parts respectively and

$$\frac{1}{\Delta t_e} = \frac{1}{\Delta t} + \frac{2\varepsilon_1}{(\Delta x)^2} + \frac{2\varepsilon_2}{(\Delta y)^2}\,. \tag{14}$$

We slightly generalize (14) by redefining Δt_e by

$$\frac{1}{\Delta t_e} = \frac{1}{\Delta t} + 2\kappa\left(\frac{\varepsilon_1}{(\Delta x)^2} + \frac{\varepsilon_2}{(\Delta y)^2}\right) \tag{15}$$

where κ is a constant that we can choose. The form of (15) no longer follows directly from the Crocco formulation. Instead κ will be chosen based on a stability analysis.

We choose Δt in (12) or (15) based on the hyperbolic (inviscid) stability condition. We then find Δt_e from (15) and advance to stage $(K + 1)$ using the Runge-Kutta scheme (13).

The constant κ in (15) can be chosen so that we recover the parabolic stability limitation when $H_D = 0$. The exact value of κ depends on the coefficients α_K in the Runge-Kutta formula. In order to see this more clearly we revert to the one-dimensional convection-diffusion equation (8). We replace all space derivatives by second-order central differences while the time derivative is kept continuous. We therefore have

$$w_t = \frac{a\left(w^n_{j+1} - w^n_{j-1}\right)}{2\Delta x} + \frac{\varepsilon}{(\Delta x)^2}\left(w^n_{j+1} - 2w^n_j + w^n_{j-1}\right). \tag{16}$$

We Fourier transform (16) to get

$$\hat{w}_t = \lambda \hat{w} \tag{17}$$

with

$$\lambda(\xi) = -\frac{2\varepsilon}{(\Delta x)^2}(1 - \cos\,\xi) + \frac{ia}{\Delta x}\sin\,\xi \qquad 0 \leq \xi \leq 2\pi. \tag{18}$$

A Runge-Kutta scheme for (16) or (17) is stable whenever $z(\xi) = \lambda(\xi)\Delta t_e$ lies within the stability domain that depends on $\alpha_1, \cdots, \alpha_N$ for all $0 \leq \xi \leq 2\pi$.

We consider the stability domain for the four-step scheme with $\alpha_1 = 1/4$, $\alpha_2 = 1/3$, $\alpha_3 = 1/2$, $\alpha_4 = 1$. This scheme has a stability condition along the imaginary axis of $\max_{\xi}|z| \leq 2\sqrt{2}$, i.e., for a hyperbolic problem ($\varepsilon = 0$) $\frac{a\Delta t_e}{\Delta x} \leq 2\sqrt{2}$. Along the negative real axis the stability condition is $|z| \lesssim 2.8$ and for a parabolic problem ($a = 0$) $\frac{2\varepsilon\Delta t_e}{(\Delta x)^2} \lesssim 2.8$. Hence for this case we would choose κ in (15) as $\kappa = 1.4$. We define the cell Reynolds number as

$$R_h = \frac{a\Delta x}{\varepsilon}. \tag{19}$$

The previous analysis shows that the Runge-Kutta scheme is stable for $R_h = 0$ and $R_h = \infty$. We do not have any proof that the scheme is stable for all R_h.

III. NAVIER-STOKES EQUATIONS

We now discuss the implementation of these ideas to the two-dimensional, compressible, Navier-Stokes equations. The extension to three dimensions is straightforward. We first consider the conservation form in Cartesian coordinates. We express the equations in the following form

$$\begin{aligned}
\rho_t &= H_1 \\
(\rho u)_t &= H_2 + (\lambda + 2\mu)\frac{\partial^2 u}{\partial x^2} + (\lambda + \mu)\frac{\partial^2 v}{\partial x \partial y} + \mu\frac{\partial^2 u}{\partial y^2} \\
(\rho v)_t &= H_3 + \mu\frac{\partial^2 v}{\partial x^2} + (\lambda + \mu)\frac{\partial^2 u}{\partial x \partial y} + (\lambda + 2\mu)\frac{\partial^2 v}{\partial y^2}
\end{aligned} \tag{20}$$

$$(\rho E)_t = H_4 + \frac{\gamma\mu}{Pr}\left(\frac{\partial^2 e}{\partial x^2} + \frac{\partial^2 e}{\partial y^2}\right)$$

$$+ (\lambda + 2\mu)u\frac{\partial^2 u}{\partial x^2} + \mu v\frac{\partial^2 v}{\partial x^2}$$

$$+ (\lambda + \mu)\left[v\frac{\partial^2 u}{\partial x \partial y} + u\frac{\partial^2 v}{\partial x \partial y}\right]$$

$$+ \mu u\frac{\partial^2 u}{\partial y^2} + (\lambda + 2\mu)v\frac{\partial^2 v}{\partial y^2}$$

where

$$e = E - \frac{(\rho u)^2 + (\rho v)^2}{2\rho},$$

and H_j denote first derivative terms (including the artificial viscosity and also the viscous dissipation function). The coefficients of viscosity (μ and λ), γ the specific heat ratio, and the Prandtl number Pr are all assumed (for the analysis) to be locally constant.

In deriving our results we shall ignore all cross derivatives (see, e.g., [1], [2]). Based on our previous analysis we add the following terms to the standard Runge-Kutta scheme.

$$\Delta\rho = K_1$$

$$\Delta(\rho u) = K_2 - 2\left[\frac{\lambda + 2\mu}{(\Delta x)^2} + \frac{\mu}{(\Delta y)^2}\right]\frac{\Delta(\rho u)}{\rho}\alpha\Delta t$$

$$\Delta(\rho v) = K_3 - 2\left[\frac{\mu}{(\Delta x)^2} + \frac{\lambda + 2\mu}{(\Delta y)^2}\right]\frac{\Delta(\rho v)}{\rho}\alpha\Delta t \tag{21}$$

$$\Delta(\rho E) = K_4 - \frac{2u}{\rho}\left[-\frac{\gamma\mu}{2Pr}\left(\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2}\right) + \frac{(\lambda + 2\mu)}{(\Delta x)^2} + \frac{\mu}{(\Delta y)^2}\right]\Delta(\rho u)\alpha\Delta t$$

$$- \frac{2v}{\rho}\left[- \frac{\gamma\mu}{2Pr}\left(\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2}\right) + \frac{\mu}{(\Delta x)^2} + \frac{(\lambda + 2\mu)}{(\Delta y)^2}\right]\Delta(\rho v)\cdot\alpha\Delta t$$

$$- \frac{2\gamma\mu}{\rho Pr}\left(\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2}\right)\Delta(\rho E)\cdot\alpha\Delta t$$

where $\Delta w = w^{n+1} - w^n$ and K_j denote the usual space derivative terms. For simplicity we have chosen $\kappa = 1$, and α denotes the constant in the Runge-Kutta scheme (28). Thus the density equation is unchanged. The second and third equations can be solved directly for $\Delta(\rho u)$, $\Delta(\rho v)$. Once $\Delta(\rho u)$, $\Delta(\rho v)$ are known the last equation can be solved for $\Delta(\rho E)$. As before these corrections imply an effective time step which automatically accounts for the viscous time step. In this case the effective time step differs for each equation.

We finally consider the Navier-Stokes equation in body fitted coordinates. This can be done either in a finite volume scheme or by using transformations. The result is the same in either case [9], and so we shall use a transformation for ease of presentation. Let $\xi = \xi(x,y)$, $\eta = \eta(x,y)$ be the body fitted coordinates. We choose the coordinate scaling so that $\Delta\xi = \Delta\eta = 1$. The Navier-Stokes equations (20) now become

$$\rho_t = \overline{H}_1$$

$$(\rho u)_t = \overline{H}_2 + \left[(\lambda + 2\mu)\xi_x^2 + \mu\xi_y^2\right]\frac{\partial^2 u}{\partial\xi^2} + \left[(\lambda + 2\mu)\eta_x^2 + \mu\eta_y^2\right)\frac{\partial^2 u}{\partial\eta^2}$$

$$+ (\lambda + \mu)\xi_x\,\xi_y\frac{\partial^2 v}{\partial\xi^2} + (\lambda + \mu)\eta_x\,\eta_y\frac{\partial^2 v}{\partial\eta^2} + \text{crossterms}$$

$$(\rho v)_t = \overline{H}_3 + \left[\mu\xi_x^2 + (\lambda + 2\mu)\xi_y^2\right]\frac{\partial^2 v}{\partial\xi^2} + \left[\mu\eta_x^2 + (\lambda + 2\mu)\eta_y^2\right]\frac{\partial^2 v}{\partial\eta^2}$$

$$+ (\lambda + \mu)\xi_x \xi_y \frac{\partial^2 u}{\partial \xi^2} + (\lambda + \mu)\eta_x \eta_y \frac{\partial^2 u}{\partial \eta^2} + \text{crossterms} \qquad (22)$$

$$(\rho E)_t = \overline{H}_4 + \frac{\gamma\mu}{Pr}\left[(\xi_x^2 + \xi_y^2)\frac{\partial^2 e}{\partial \xi^2} + (\eta_x^2 + \eta_y^2)\frac{\partial^2 e}{\partial \eta^2}\right]$$

$$+ \left[(\lambda + 2\mu)u\xi_x^2 + (\lambda + \mu)v\xi_x \xi_y + \mu u\xi_y^2\right]\frac{\partial^2 u}{\partial \xi^2}$$

$$+ \left[(\lambda + 2\mu)u\eta_x^2 + (\lambda + \mu)v\eta_x \eta_y + \mu u\eta_y^2\right]\frac{\partial^2 u}{\partial \eta^2}$$

$$+ \left[\mu v\xi_x^2 + (\lambda + \mu)u\xi_x \xi_y + (\lambda + 2\mu)v\xi_y^2\right]\frac{\partial^2 v}{\partial \xi^2}$$

$$+ \left[\mu v\eta_x^2 + (\lambda + \mu)u\eta_x \eta_y + (\lambda + 2\mu)v\eta_y^2\right]\frac{\partial^2 v}{\partial \eta^2} + \text{crossterms}$$

where $\overline{H}_j$ are first derivative terms and we have ignored all second cross derivative terms. As before this generates an appropriate correction term to the Runge-Kutta scheme. Equation (21) is now replaced by

$$\Delta\rho = \overline{K}_1$$

$$\Delta(\rho u) = \overline{K}_2 - 2\left[(\lambda + 2\mu)\xi_x^2 + \mu\xi_y^2 + (\lambda + 2\mu)\eta_x^2 + \mu\eta_y^2\right]\frac{\Delta(\rho u)}{\rho}\alpha\Delta t$$

$$- 2(\lambda + \mu)(\xi_x \xi_y + \eta_x \eta_y)\frac{\Delta(\rho v)}{\rho}\alpha\Delta t$$

$$\Delta(\rho v) = \overline{K}_3 - 2(\lambda + \mu)(\xi_x \xi_y + \eta_x \eta_y)\frac{\Delta(\rho u)}{\rho}\alpha\Delta t$$

$$- 2\left[\mu\xi_x^2 + (\lambda + 2\mu)\xi_y^2 + \mu\eta_x^2 + (\lambda + 2\mu)\eta_y^2\right]\frac{\Delta(\rho v)}{\rho}\alpha\Delta t \qquad (23)$$

$$\Delta(\rho E) = \overline{K}_4 - \frac{2\gamma\mu}{\rho Pr}\left(\xi_x^2 + \xi_y^2 + \eta_x^2 + \eta_y^2\right)\Delta(\rho E)\cdot\alpha\Delta t$$

$$- 2\Big[-\frac{u}{2}\frac{\gamma\mu}{Pr}\left(\xi_x^2 + \xi_y^2 + \eta_x^2 + \eta_y^2\right) + (\lambda + 2\mu)u\left(\xi_x^2 + \eta_x^2\right)$$

$$+ (\lambda + \mu)v(\xi_x\,\xi_y + \eta_x\,\eta_y) + \mu u\left(\xi_y^2 + \eta_y^2\right)\Big]\frac{\Delta(\rho u)}{\rho}\,\alpha\Delta t$$

$$- 2\Big[-\frac{v}{2}\frac{\gamma\mu}{Pr}\left(\xi_x^2 + \xi_y^2 + \eta_x^2 + \eta_y^2\right) + \mu v\left(\xi_x^2 + \eta_x^2\right)$$

$$+ (\lambda + \mu)u(\xi_x\,\xi_y + \eta_x\,\eta_y) + (\lambda + 2\mu)v\left(\xi_y^2 + \eta_y^2\right)\Big]\frac{\Delta(\rho v)}{\rho}\cdot\alpha\Delta t$$

where $\overline{K}_j$ represents the standard finite difference terms.

As before the density equation is unchanged by the viscous correction. Now, however, the two momentum equations are coupled together, unless the coordinate system is orthogonal. As we have two equations for $\Delta(\rho u)$ and $\Delta(\rho v)$, and we can easily solve these. To simplify the notation we define

$$z_1 = 1 + \frac{2\alpha\Delta t}{\rho}\left[(\lambda + 2\mu)\left(\xi_x^2 + \eta_x^2\right) + \mu\left(\xi_y^2 + \eta_y^2\right)\right]$$

$$z_2 = \frac{2\alpha\Delta t}{\rho}(\lambda + \mu)(\xi_x\,\xi_y + \eta_x\,\eta_y) \tag{24}$$

$$z_4 = 1 + \frac{2\alpha\Delta t}{\rho}\left[\mu\left(\xi_x^2 + \eta_x^2\right) + (\lambda + 2\mu)\left(\xi_y^2 + \eta_y^2\right)\right]$$

and

$$D = (1 + z_1)(1 + z_4) - z_2^2.$$

Then

$$\Delta\rho = \bar{K}_1$$

$$\Delta(\rho u) = \frac{\bar{K}_2\, z_4 - \bar{K}_3\, z_2}{D} \tag{25}$$

$$\Delta(\rho v) = \frac{\bar{K}_3\, z_1 - \bar{K}_2\, z_2}{D} .$$

As before given $\Delta(\rho u)$ and $\Delta(\rho v)$ we can solve for $\Delta(\rho E)$ directly from the energy equation in (23). We also note that if one uses the thin layer approximation (dropping all second ξ derivatives and cross derivatives in (22)) then these terms simplify slightly. In this case $\Delta\rho$, $\Delta\rho u$, $\Delta\rho v$ are still given by (25) with

$$z_1 = 1 + \frac{2\alpha\Delta t}{\rho}\left[(\lambda + 2\mu)\eta_x^2 + \mu\eta_y^2\right]$$

$$z_2 = \frac{2\alpha\Delta t}{\rho}(\lambda + \mu)\eta_x\,\eta_y$$

$$z_4 = 1 + \frac{2\alpha\Delta t}{\rho}\left[\mu\eta_x^2 + (\lambda + 2\mu)\eta_y^2\right] \tag{26}$$

$$\eta_x = -\frac{y_\xi}{J} \qquad \eta_y = +\frac{x_\xi}{J}$$

$$J = x_\xi\, y_\eta - x_\eta\, y_\xi$$

and

$$\left[1 + \frac{2\gamma\mu\alpha\Delta t}{\rho \mathrm{Pr}}(\eta_x^2 + \eta_y^2)\right]\Delta(\rho E) = \bar{K}_4$$

$$- 2\left[-\frac{u}{2}\left(\frac{\gamma\mu}{\mathrm{Pr}}\right)(\eta_x^2 + \eta_y^2) + (\lambda + 2\mu)u\eta_x^2 + (\lambda + \mu)v\eta_x\,\eta_y + \mu u\eta_y^2\right]\frac{\Delta(\rho u)}{\rho}\cdot\alpha\Delta t$$

$$- 2\left[-\frac{v}{2}\left(\frac{\gamma\mu}{\mathrm{Pr}}\right)(\eta_x^2 + \eta_y^2) + \mu v\eta_x^2 + (\lambda + \mu)u\eta_x\,\eta_y + (\lambda + 2\mu)u\eta_y^2\right]\frac{\Delta(\rho v)}{\rho}\cdot\alpha\Delta t.$$

IV. RESIDUAL SMOOTHING

As an alternative method of reducing the effect of the parabolic terms on the stability of the scheme we consider residual smoothing. With this technique one post-processes an explicit method with an implicit method. In practice one post-processes each equation separately and each direction separately so that only scalar tridiagonal matrices need be inverted. When using a multistage Runge-Kutta method, one can apply the residual smoothing after each stage, or at the end of the entire process, or any intermediate permutation.

In [10] it is shown that one can construct such a scheme for a hyperbolic equation so that the total method is unconditionally stable. It is further shown in [10] that it is not efficient to use a very large Δt even ignoring splitting errors. An optimal Δt is about two to three times larger than the explicit time step. We now consider the process for a parabolic problem in order to see the effect of viscous terms.

We, therefore, consider the heat equation

$$u_t = bu_{xx}. \tag{27}$$

We solve this equation by a k-stage Runge-Kutta scheme

$$\begin{aligned} u^{(1)} &= u^n + \alpha_1 \, \Delta t Q u \\ &\vdots \\ u^{(\ell+1)} &= u^n + \alpha_{\ell+1} \, \Delta t Q u^{(\ell)} \\ &\vdots \\ u^{n+1} &= u^{(k)} \end{aligned} \tag{28}$$

where $\alpha_1,\cdots,\alpha_k$ are given coefficients with $\alpha_k = 1$. Q is a difference approximation to u_{xx}. The amplification factor corresponding to (28) is

$$G = 1 + \beta_1 \Delta t\hat{Q} + \beta_2(\Delta t)^2 \hat{Q}^2 + \cdots + \beta_k(\Delta t)^k \hat{Q}^k \tag{29}$$

where $\beta_1 = 1$ and $\beta_\ell = \beta_{\ell-1}\ \alpha_{k-\ell+1}$, $\ell = 2,\cdots,k$. $\hat{Q}$ is the Fourier transform of Q. Hence, for second-order central differencing

$$\hat{Q} = -\frac{4b \sin^2(\theta/2)}{(\Delta x)^2}\ . \tag{30}$$

Residual smoothing consists of updating a stage (ℓ) by

$$(1 - \sigma D_2)\Delta u^{(\ell)} = u^{(\ell)} - u^n \tag{31}$$

where D_2 is again a second-order central difference approximation to u_{xx}, i.e., $D_2 \rightarrow (1,-2,1)$. We now consider two possibilities. In the first we apply (31) only after the final stage. Then the new amplification factor is

$$G_1(\theta) = 1 + \frac{\beta_1 \Delta t\hat{Q} + \beta_2(\Delta t)^2 \hat{Q}^2 + \cdots + \beta_k(\Delta t)^k \hat{Q}^k}{1 + 4\sigma \sin^2(\theta/2)}\ . \tag{32}$$

The second case we consider is applying (31) after every stage. The resultant amplification factor is

$$G_2(\theta) = 1 + \beta_1 \Delta t\hat{R} + \beta_2(\Delta t)^2 \hat{R}^2 + \cdots + \beta_k(\Delta t)^k \hat{R}^k \tag{33}$$

with

$$\hat{R} = \frac{\hat{Q}}{1 + 4\sigma \sin^2(\theta/2)}\ . \tag{34}$$

We now investigate the possibility that either of these schemes is unconditionally stable. To investigate this we need only consider Δt sufficiently large. We thus consider $\Delta t \to \infty$ with $\sigma \to \infty$. Then (32) becomes

$$G_1(\theta) \to 1 + \frac{(-1)^k \beta_k [\frac{4b\Delta t}{(\Delta x)^2} \sin^2(\frac{\theta}{2})]^k}{1 + 4\sigma \sin^2(\theta/2)} . \tag{35}$$

We thus see that for k even, $G_1(\theta) > 1$ and so (28) - (31) cannot be stable for Δt large. For $k = 1$ the scheme is identical with backward Euler for a scalar one-dimensional equation and, hence, unconditionally stable. For the second case we see that (33) has the same form as a standard Runge-Kutta method with $\hat{Q}$ replaced by $\hat{R}$, (34). Hence, it follows that the scheme is stable whenever $\Delta t\hat{R}$ is within the stability region of the scheme. As $\Delta t \to \infty$, so does σ and so there is a cancellation between the numerator and denominator; thus, $\Delta t\hat{R}$ remains bounded as Δt increases. We thus conclude that applying the residual smoothing after each stage can make the scheme unconditionally stable even for a Runge-Kutta method with an even number of stages.

We also see from the above argument that as Δt increases so must σ. In [9], [10] we show that for a hyperbolic equation

$$u_t + au_x = 0$$

that σ is proportional to $(a\Delta t/\Delta x)^2$. For the parabolic problem (27) it follows from (35) that σ should be proportional to $b\Delta t/(\Delta x)^2$. For the combined convection-diffusion equation σ will be related to the sum of two such contributions.

It follows from (33), (34) that if we apply residual smoothing after every stage then the stability polynomial has the same form as the original polynomial (29). The only difference is that $\hat{Q}$ is now replaced by $\hat{R}$. From (34) it follows that the ratio of $\hat{Q}$ to $\hat{R}$ is real. Hence, if $\hat{Q}$ is any complex number then $\hat{R}$ lies along the same ray in the complex plane but with a different amplitude. We therefore have shown that if the original scheme was unstable for a given direction then residual smoothing cannot stabilize the scheme. Furthermore, if the original scheme was conditionally stable then by choosing $\sigma = \sigma(\Delta t)$ sufficiently large we can make the scheme unconditionally stable. We have thus shown

Theorem: Let $\hat{Q}$ be the amplification factor for any approximation to the convection-diffusion equation and let (29) be the stability polynomial for a k stage Runge-Kutta scheme. We now apply residual smoothing, (31), after every stage of the scheme. If the original scheme was unconditionally unstable then the new scheme is still unconditionally unstable. If the original scheme was conditionally stable then the scheme with residual smoothing can be made unconditionally stable by choosing $\sigma(\Delta t)$ sufficiently large.

Hence, if the smoothing is applied at the end when solving a parabolic equation, then the scheme can be unconditionally stable only when using a multistage scheme with an odd number of stages. When the smoothing is done after each stage, the scheme can be stabilized for σ large. For a system with a hyperbolic portion and a small parabolic contribution, e.g., high Reynolds number Navier-Stokes, the residual smoothing is most effective with a

time step about twice that of the explicit convective portion. Hence, the question of unconditional stability is somewhat academic. In practice [8] the Runge-Kutta scheme for the Navier-Stokes equations is used with four stages and with the residual smoothing applied after each stage.

V. RESULTS

In this section we present some results for viscous flow obtained using the analysis of Sections II and III. We used a Runge-Kutta code to solve the Navier-Stokes equations for two flows about an airfoil section. The details of this code are discussed in [5], [9], [10], [11]. In these cases we considered only the thin-layer form of the Navier-Stokes equations.

For the first case we computed laminar flow over an NACA 0012 airfoil with a free-stream Mach number M_∞ of 0.5 and a Reynolds number Re_∞ of 5×10^3. The angle of attack (α) of the airfoil was zero degrees. Half-plane calculations were performed using a C-type grid consisting of 64 cells in the streamwise direction and 64 cells in the normal-like direction. The grid spacing at the airfoil surface was about 6×10^{-4} chords. The mesh spacing in the streamwise direction over the central part of the airfoil was $\Delta X = 0.05$ chords. Results for this case are shown in Figures 1a - 1c. As indicated in Figure 1b, the flow separates at $X = 0.817$ chords. The size of the recirculation zone is displayed in Figure 1c. The results are all independent of the time step procedure used to reach the steady state.

In Figure 1d convergence histories for this case for two calculations are shown. The residual displayed in this graph is the root mean square of the residual of the continuity equation. The calculations were started

impulsively by inserting the airfoil into a uniform flow and immediately enforcing the appropriate boundary conditions. Local time stepping and enthalpy damping (see [9]) were employed in each computation; no residual smoothing was used. For history A the Runge-Kutta scheme with the time step (Δt) limitation determined by convection was used; this required choosing a CFL = 1.0. For curve B a larger Courant-Friedrichs-Lewy (CFL) number was used by accounting for the diffusion limit on Δt with the pseudo-time algorithm. This allowed choosing CFL = 2.5 based on an inviscid criterion. There is additional work with the pseudo-time scheme. Nevertheless, the computational time required to reach a satisfactory level of convergence was reduced by a factor of 1.7.

In the second case we solved for turbulent flow over an NACA 0012 airfoil with $M_\infty = 0.5$, $Re_\infty = 2.89 \times 10^6$, and $\alpha = 0$ degrees. A 60×50 half-plane grid was used in the computations. The grid spacing at the surface was about 8.5×10^{-5} chords. The chordwise spacing at the midsection of the airfoil was about $\Delta X = 0.036$ chords. Numerical results for this case are presented in Figures 2a and 2b.

Figure 2c shows two convergence histories for this turbulent flow case. As in the laminar flow problem, the histories were obtained by computing without and with the effects on Δt due to diffusion. The pseudo-time algorithm was about 1.4 times faster in reaching steady state. This is close to the factor expected, since we were able to increase the CFL from 1.5 to 2.7, a factor of 1.8. We do not achieve this speedup of 1.8 since there is some reduction of the effective time step due to the diffusion terms.

VI. CONCLUSIONS

The use of the Crocco scheme for a scalar convection-diffusion equation introduces a scaling of the time step. This reduces the effective time step so that the viscous stability limit is automatically satisfied. As such the scheme cannot introduce any fundamental acceleration in reaching the steady state. The advantage of the scheme is that we do not need to explicitly account for the viscous time step restriction; it is done automatically. This can be done efficiently using Runge-Kutta type schemes. In addition, for variable coefficients or nonuniform meshes this introduces an effective local time step.

Using this scheme for a system of equations, e.g., Navier-Stokes, has the additional benefit that a different scaling is chosen for each equation. Thus each equation has its own appropriate (viscous) time step. This is equivalent to using a diagonal preconditioning [10] to accelerate the equations to a steady state. Computations demonstrate that we can gain a factor of between 1.5 and 2 with little programming effort.

We further show that if one uses residual smoothing to increase the time step then one must also account for the viscous terms. When the smoothing is applied after the completion of a Runge-Kutta cycle then unconditional stability is possible only if an odd number of stages is used. Applying the smoothing after each stage allows for unconditional stability for all multistage schemes provided σ is chosen sufficiently large.

REFERENCES

[1] S. Abarbanel and D. Gottlieb, "Optimal time splitting for two- and three-dimensional Navier-Stokes equations with mixed derivatives." J. Comput. Phys. **41**, 1-33 (1981).

[2] J. S. Allen and S. I. Cheng, "Numerical solutions of the compressible Navier-Stokes equations for the laminar near wake in supersonic flow." Phys. Fluids **13**, 37-52 (1970).

[3] Yu I. Brailovskaya, "A difference scheme for numerical solution of the two-dimensional nonstationary Navier-Stokes equations for a compressible gas." Soviet Phys. Dokl. **10**, 107-110 (1965).

[4] L. Crocco, "A suggestion for the numerical solution of the steady Navier-Stokes equations." AIAA J. **3**, 1824-1832 (1965).

[5] A. Jameson, W. Schmidt, and E. Turkel, "Numerical solutions of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes." AIAA paper 81-1259 (1981).

[6] R. Peyret and H. Viviand, "Computation of viscous compressible flows based on the Navier-Stokes equations." AGARD Report 212 (1975).

[7] T. H. Pulliam and J. L. Steger, "Recent Improvements in efficiency, accuracy, and convergence for implicit approximate factorization algorithms," AIAA paper 85-0360 (1985).

[8] P. J. Roache, Computational Fluid Dynamics, Hermosa Publishers, 1976.

[9] R. C. Swanson and E. Turkel, "A multistage time-stepping scheme for the Navier-Stokes equations." AIAA paper 85-0035 (1985).

[10] E. Turkel, "Acceleration to a steady state for the Euler equations." Numerical Methods for the Euler Equations of Fluid Dynamics, pp. 281-311, SIAM, Philadelphia, 1985.

[11] E. Turkel, "Algorithms for the Euler and Navier-Stokes equations on supercomputers." Progress and Supercomputing in Computational Fluid Dynamics (Edited by E. M. Murman and S. S. Abarbanel), 155-172, Birkhauser Publishing Co., Boston, 1985.

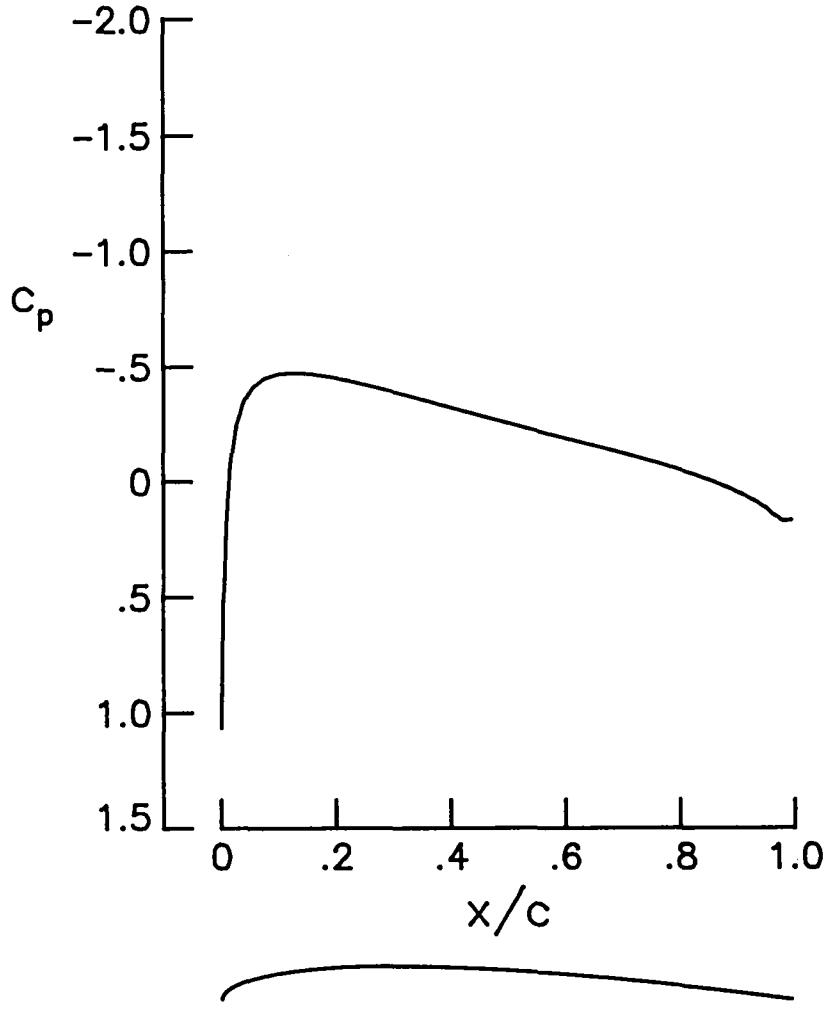


Figure 1a. Surface pressure distribution for laminar flow over an NACA 0012 airfoil ($M_\infty = 0.5$), $Re_\infty = 5 \times 10^3$, $\alpha = 0$ degrees).

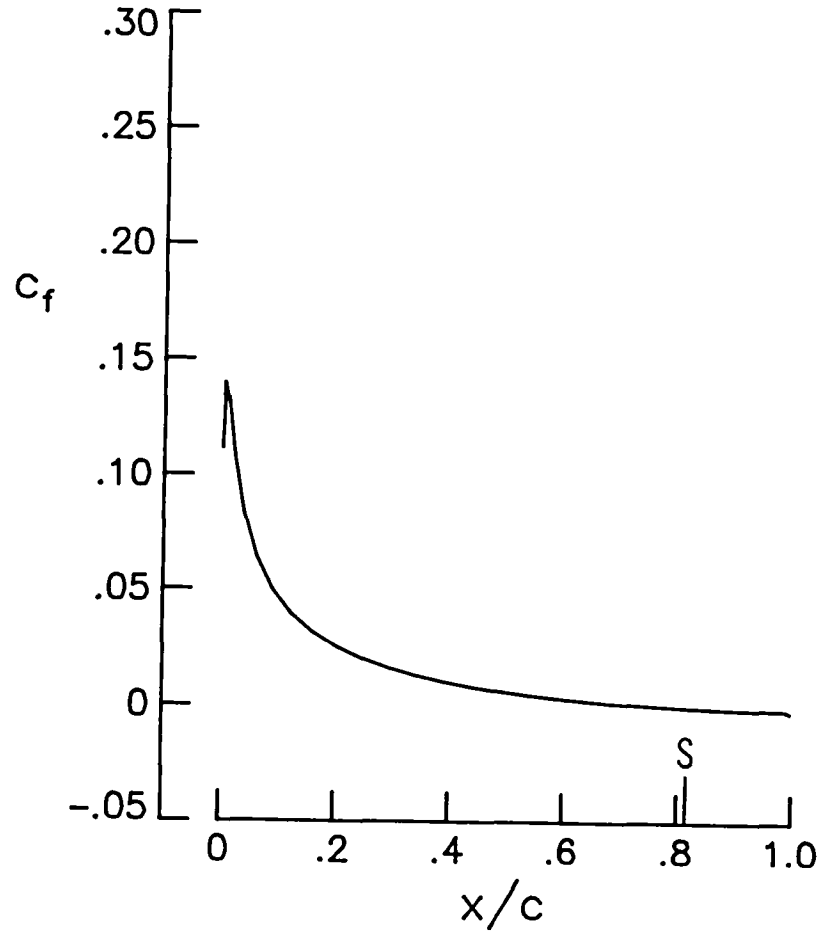


Figure 1b. Skin-friction (based on free-stream conditions) distribution for laminar flow over an NACA 0012 airfoil ($M_\infty = 0.5$, $Re_\infty = 5 \times 10^3$, $\alpha = 0$ degrees).

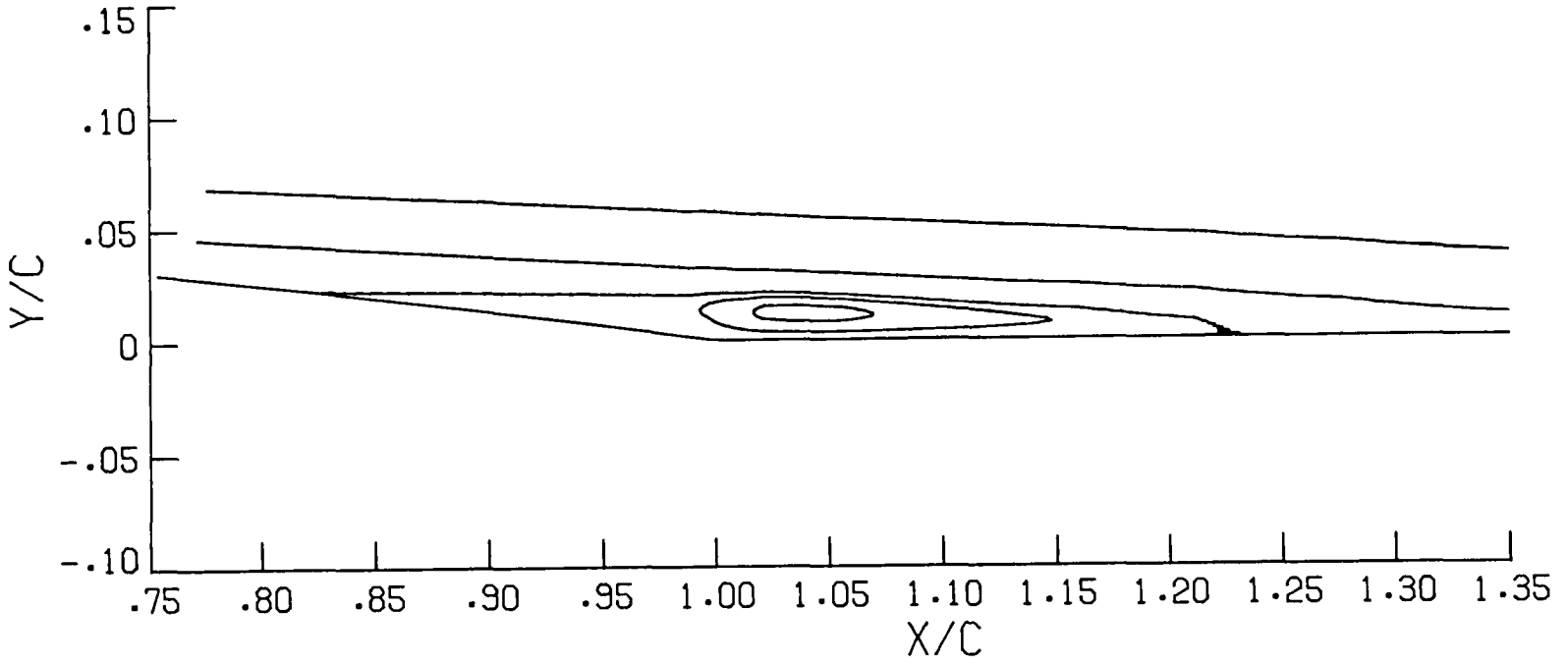


Figure 1c. Streamlines for upper surface at the trailing edge ($M_\infty = 0.5$, $Re_\infty = 5 \times 10^3$, $\alpha = 0$ degrees).

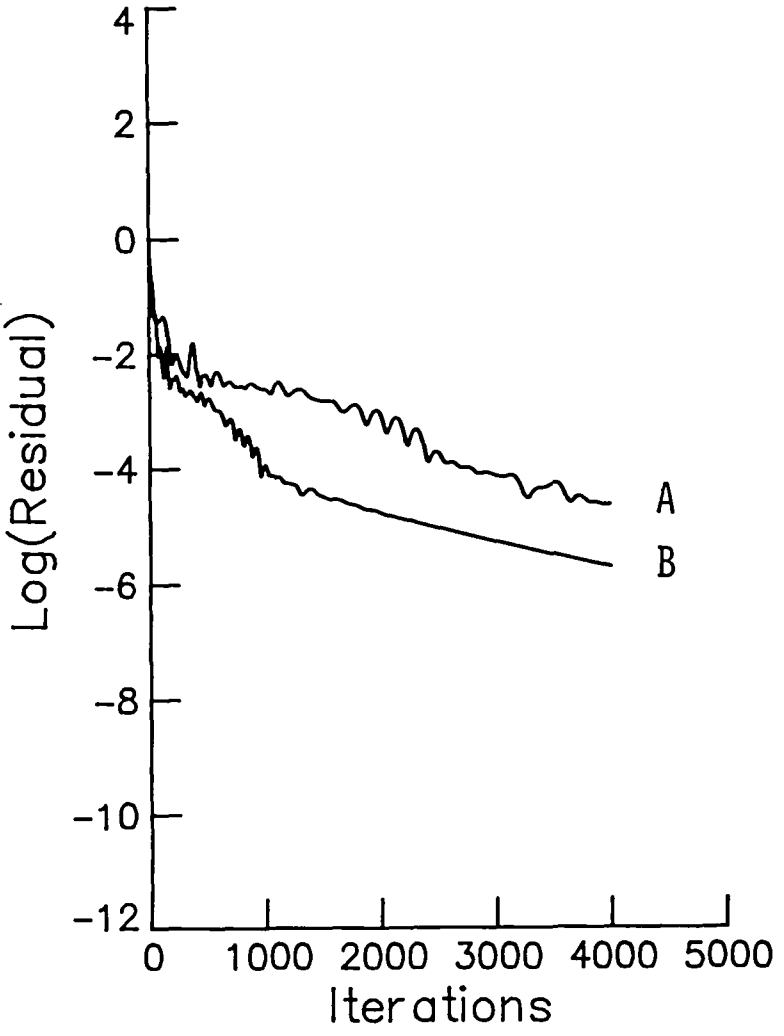


Figure 1d. Convergence histories for laminar airfoil flow calculations.

A -- Runge-Kutta scheme without pseudo-time algorithm (CFL number of 1.0).

B -- Runge-Kutta scheme with pseudo-time algorithm (CFL number of 2.5).

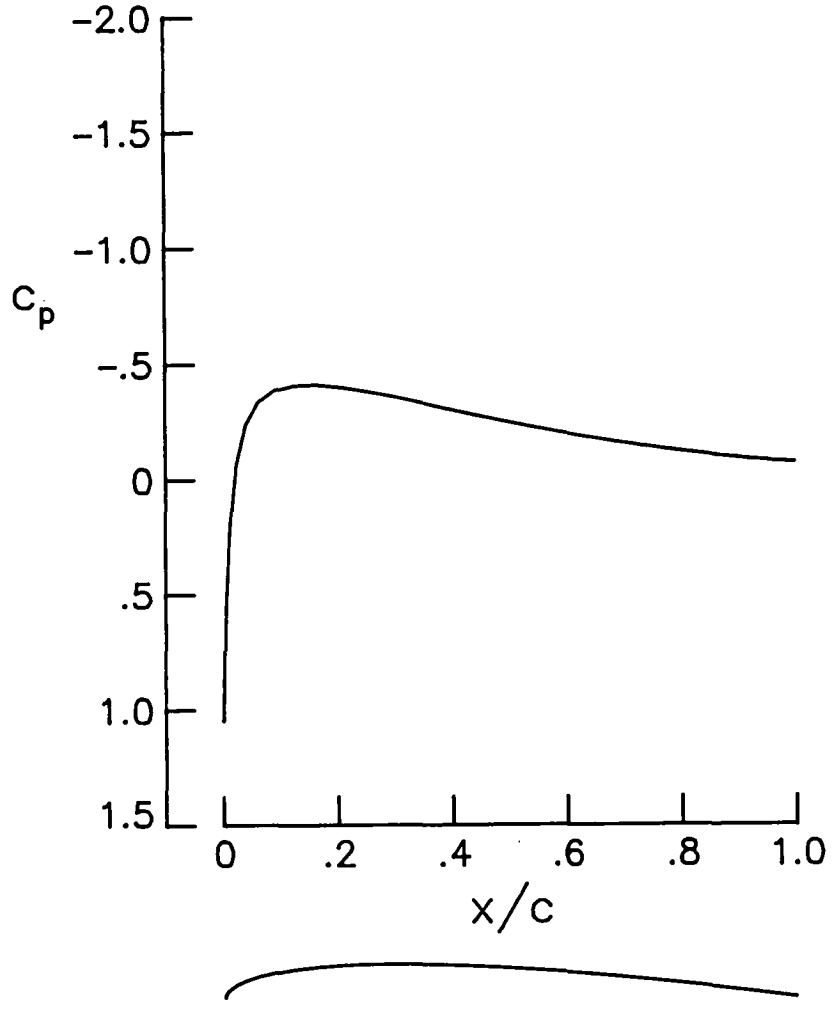


Figure 2a. Surface pressure distribution for turbulent flow over an NACA 0012 airfoil ($M_\infty = 0.5$, $Re_\infty = 2.89 \times 10^6$, $\alpha = 0$ degrees).

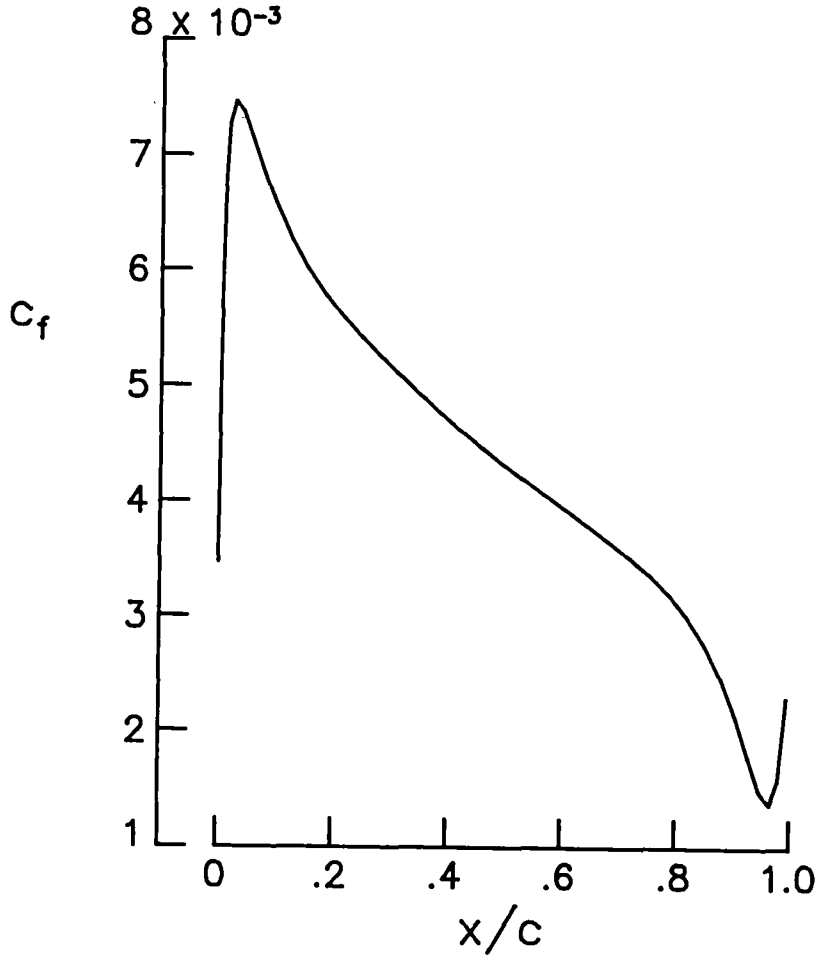


Figure 2b. Skin-friction (based on free-stream conditions) distribution for turbulent flow over an NACA 0012 airfoil ($M_\infty = 0.5$, $Re_\infty = 2.89 \times 10^6$, $\alpha = 0$ degrees).

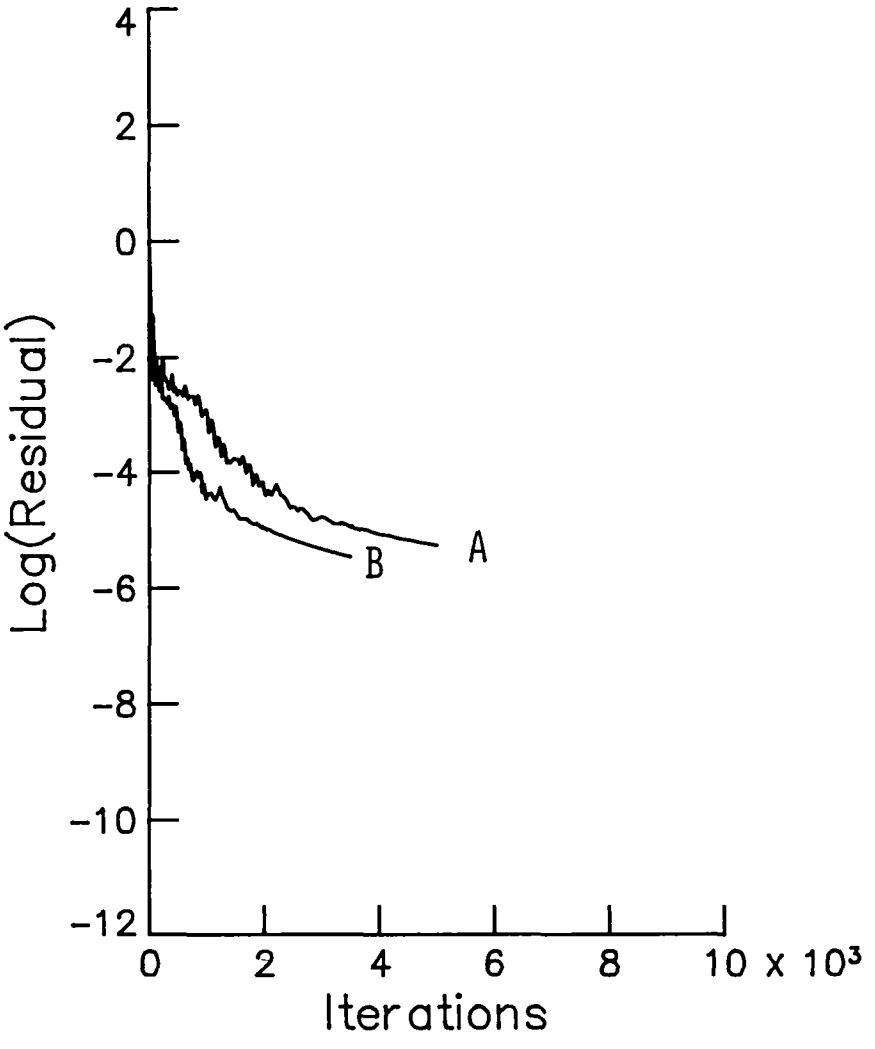


Figure 2c. Convergence histories for turbulent airfoil flow calculations.

A -- Runge-Kutta scheme without pseudo-time algorithm (CFL number of 1.5).

B -- Runge-Kutta scheme with pseudo-time algorithm (CFL number of 2.7).

CONDITIONS FOR THE CONSTRUCTION OF MULTI-POINT TOTAL VARIATION DIMINISHING DIFFERENCE SCHEMES

Antony Jameson*

and

Peter D. Lax**

ABSTRACT

Conditions are derived for the construction of total variation diminishing difference schemes with multi-point support. These conditions, which are proved for explicit, implicit, and semi-discrete schemes, correspond in a general sense to the introduction of upwind biasing.

*Princeton University, Mechanical and Aerospace Engineering Department, Princeton, New Jersey 08544.

**New York University, Courant Institute of Mathematical Sciences, New York, New York 10012.

Introduction

It is natural that the rapid evolution of increasingly powerful computers should inspire attempts to solve previously intractable problems by numerical calculation. One might imagine that within a fairly short time, advances in processing speed and memory capacity ought to reduce the simulation of physical systems governed by partial differential equations to a matter of routine. The numerical computation of solutions of nonlinear conservation laws has proved, in fact, to be perhaps unexpectedly difficult. Discontinuities are likely to appear in the solution, and schemes which are accurate in smooth regions tend to produce spurious oscillations in the neighborhood of the discontinuities. These oscillations can be eliminated by the use of strongly dissipative first order accurate schemes, but these schemes severely degrade the accuracy and usually produce excessively smeared discontinuities.

The scalar nonlinear conservation law in one space dimension

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0 \qquad (1)$$

provides a model which already contains the phenomena of shockwave formation and expansion fans. Thus it can be used to provide insight into the likely behavior of numerical approximations to more complex physical systems, while it is still simple enough to be fairly easily amenable to analysis. A rather complete mathematical theory of solutions to (1) is by now available [1-3].

Equation (1) describes wave propagation at a speed

$$a(u) = \frac{\partial f}{\partial u} \ .$$

The solution is constant along the characteristic lines

$$\frac{\partial x}{\partial t} = a(u)$$

provided that they do not intersect to form a shock wave. Tracing the solution backward along the characteristics, it can be seen that the total variation

$$TV(u) = \int_{-\infty}^{\infty} \left|\frac{\partial u}{\partial x}\right| dx$$

is constant prior to the formation of a shockwave, while it may decrease when the shockwave is formed. Corresponding to this property it may be observed that no new local extrema may be created and that the value of a local minimum is non-decreasing while the value of a local maximum is non-increasing. It follows that an initially monotone profile continues to be monotone.

It seems desirable that these properties should be preserved by a numerical approximation to (1). This will guarantee the exclusion of spurious oscillations in the numerical solution. Harten [4] has recently introduced the concept of total variation diminishing (TVD) difference schemes, which have the property that the discrete total variation

$$TV(v) = \sum_{k=-\infty}^{\infty} |v_k - v_{k-1}|$$

of the solution vector v cannot increase. Harten also devised procedures for constructing both explicit and implicit TVD schemes [4,5].

The purpose of this paper is to state and prove conditions for the construction of multi-point TVD schemes. Conditions are derived for explicit, implicit, and also semi-discrete operators to be TVD. The conditions are both necessary

and sufficient in the case of the explicit and semi-discrete schemes. The reasoning is a modification and extension of the reasoning used by Lax in an appendix to reference 5. The results were first presented in a lecture at ICASE in March 1984. The present paper is an amplification and revision of a Princeton University report issued under the same title in April 1984 [6]. In the intervening period Osher and Chakravarthy have given another proof that conditions (3.12) are sufficient for an explicit scheme to be TVD [7].

2. Conditions for Reduction of the ℓ_1 Norm

One dimensional difference operators act on doubly infinite sequences

$$u = \{u_k\} , \qquad -\infty < k < \infty . \tag{2.1}$$

The ℓ_1 norm of such a vector u is defined as

$$|u|_1 = \sum_{-\infty}^{\infty} |u_k| . \tag{2.2}$$

The space of all vectors u with finite ℓ_1 norm is denoted by ℓ_1.

A difference operator maps ℓ_1 into ℓ_1 and is of the form

$$A(u)_k = \sum_j a_j \, u_{k-j} . \tag{2.3}$$

The coefficients a_j depend on k, either explicitly or through dependence on u. In either case we write

$$a_j = a_j(k).$$

Theorem A: The operator A defined by (2.3) satisfies

$$|A(u)|_1 \leq |u|_1 \tag{2.4}$$

for all u in ℓ_1 if and only if

$$\sum_j |a_j(h+j)| \leq 1 \tag{2.5}$$

for all h.

An operator A satisfying (2.4) is a contraction.

Proof. The signum function is defined for every real u by

$$\text{signum } u = \begin{matrix} 1 \\ 0 \\ -1 \end{matrix} \begin{cases} \text{for } u > 0 \\ \text{for } u = 0 \\ \text{for } u < 0 \end{cases} . \tag{2.6}$$

Now set

$$s_k = \text{signum } A(u)_k; \tag{2.6*}$$

then, by definition (2.2) of the ℓ_1 norm and definition (2.6) of signum we have

$$\begin{aligned} |A(u)|_1 &= \sum_k |A(u)_k| = \sum_k s_k A(u)_k = \\ &= \sum_k s_k \sum_j a_j(k)\, u_{k-j} = \sum_{h,j} a_j(h+j)\, s_{h+j} u_h \\ &= \sum_h w_h u_h \leq \sum_h |w_h||u_h||, \end{aligned} \tag{2.7}$$

where

$$w_h = \sum_j a_j(h+j) s_{h+j} . \tag{2.7*}$$

Since s_k takes on the values ± 1 or 0, it follows from (2.7) that

$$|w_h| \leq \sum_j |a_j(h+j)| .$$

It follows therefore from assumption (2.5) that

$$|w_h| \leq 1$$

for all h. Setting this into (2.7) we deduce that (2.4) holds for all u in ℓ_1.

To show the necessity of (2.5) suppose on the contrary that it fails for some $h = h_0$. Set $u^{(0)}$ equal to

$$u^{(0)}_\ell = \begin{cases} 1 & \text{for } \ell = h_0 \\ 0 & \text{for } \ell \neq h_0 \end{cases} . \tag{2.8}$$

For this $u^{(0)}$ it follows from (2.3) that

$$A(u^{(0)})_k = a_{k-h_0}(k)$$

and so

$$|A(u^{(0)})|_1 = \sum_k |A(u^{(0)})_k| = \sum_k |a_{k-h_0}(k)| \tag{2.9}$$

$$= \sum_j |a_j(h_0+j)| > 1$$

since h_0 was so chosen that (2.5) is violated. On the other hand it is obvious from (2.8) that

$$|u^{(0)}|_1 = 1 \quad .$$

This combined with (2.9) shows that (2.4) fails for u(0).

For use in implicit schemes the following result is needed.

Theorem B: Define the operator B by

$$B(u)_k = \sum_j b_j(k) u_{k-j}. \tag{2.10}$$

B satisfies

$$|B(u)|_1 \geq |u|_1 \tag{2.11}$$

for all u in ℓ_1 if

$$b_0(h) - \sum_{j\neq 0} |b_j(h+j)| \geq 1. \tag{2.12}$$

An operator B satisfying (2.11) is called an expansion.

Proof: We define

$$s_k = \text{signum } u_k \tag{2.13}$$

Since $|s_k| \leq 1$,

$$|B(u)|_1 = \sum_k |B(u)_k| \geq \sum_k s_k B(u)_k \quad . \tag{2.14}$$

Analogously to (2.7), (2.7)* we have

$$\sum_k s_k B(u)_k = \sum_h w_h u_h \tag{2.15}$$

where

$$w_h = \sum_j b_j(h+j)\, s_{h+j} \quad . \tag{2.15}*$$

It follows readily from (2.12) that if $u_h \neq 0$,

$$|w_h| > 1 \quad .$$

Using (2.13) we get

$$\text{signum } w_h = \text{signum } u_h \quad .$$

These two imply that

$$\sum_h w_h u_h > |u|_1 \quad . \tag{2.16}$$

Combining (2.14), (2.15), and (2.16) we get (2.11).

We remark that (2.12) is far from being necessary for B to be expansive. For example, take the right shift operator T, with

$$b_j = \begin{cases} 1 & \text{for } j = 1 \\ 0 & \text{for } j \neq 1 \end{cases} \quad .$$

Clearly, T is an isometry:

$$(Tu)_1 = |u|_1,$$

but condition (2.12) is utterly violated.

Theorem A has a continuous analogue:

Theorem C: Let u(t) be a differentiable function of t real whose values lie in ℓ_1, and which satifies a differential equation of the form

$$\frac{du}{dt} = C(u), \tag{2.17}$$

where C is a difference operator, i.e., an operator of the form

$$C(u)_k = \sum_j c_j u_{k-j}. \tag{2.18}$$

The coefficients c_j may depend on k and t either directly or through a dependence on u. Then $|u(t)|_1$ is a nonincreasing function of t if and only if for all h and all t

$$c_0(h) + \sum_{j\neq 0} |c_j(h+j)| \leq 0, \tag{2.19}$$

Proof: Define $s_k(t)$ by

$$s_k(t) = \text{signum } u_k(t) \quad . \tag{2.20}$$

Then

$$|u(t)|_1 = \sum_k s_k(t)\, u_k(t) \quad . \tag{2.21}$$

Since each s_k is piecewise constant,

$$\frac{d}{dt}|u(t)|_1 = \sum_k s_k(t) \frac{du_k}{dt} \quad . \tag{2.22}$$

According to equation (2.17),

$$\frac{du_k}{dt} = \sum_j c_j(k)\, u_{k-j} \quad . \tag{2.23}$$

Setting this into the right in (2.22) we get, after relabeling the index of summation,

$$\frac{d}{dt}|u(t)|_1 = \sum_k s_k \sum_j c_j(k)\, u_{k-j} = \sum_h w_h\, u_h \tag{2.24}$$

where

$$w_h = \sum_j c_j(h+j)\, s_{h+j} \quad . \tag{2.25}$$

Suppose $u_h \neq 0$; then by (2.20), $s_h \neq 0$. Multiply (2.25) by s_h; using assumption (2.19) we get

$$s_h\, w_h = c_0(h) + \sum_{j\neq 0} c_j(h+j) s_h s_{h+j} \leq 0 \quad .$$

Since by definition, s_h and u_h have the same sign, it follows that for all h

$$u_h\, w_h \leq 0;$$

this relation clearly holds also when $u_h = 0$. Setting this into (2.24) we obtain

$$\frac{d}{dt}|u(t)|_1 \leq 0 \ ;$$

this proves that $|u(t)|_1$ decreases as t increases.

Next we indicate why condition (2.19) is necessary. Suppose it is violated at t_0, h_0. Let $u(t)$ be that solution of (2.17) whose value at t_0 equals

$$u_k(t_0) = \delta_{k,h_0} = \begin{cases} 1 & \text{for } k = h_0 \\ 0 & \text{for } k \neq h_0 \end{cases} .$$

Using (2.23) we get

$$u_k(t_0 + \varepsilon) = \delta_{k,h_0} + \varepsilon \sum_j c_j(k)\delta_{k-j,h_0} + O(\varepsilon^2) .$$

Summing with respect to k gives

$$\sum_k u_k(t_0 + \varepsilon) = 1 + \varepsilon \sum_j c_j(h_0 + j) + O(\varepsilon^2).$$

Since condition (2.19) is violated at t_0, h_0 we conclude that for ε small enough positive,

$$\sum_k u_k(t_0 + \varepsilon) > 1 \ .$$

Since

$$|u(t_0+ \varepsilon)|_1 > \sum_k u_k(t_0 + \varepsilon)$$

while

$$|u(t_0)|_1 = 1,$$

this shows that $|u(t)|_1$ is not a decreasing function of t, completing the proof of Theorem C.

3. Construction of Total Variation Diminishing Schemes

Theorems A, B, and C may be used to find conditions on the coefficients of a difference operator which guarantee that the total variation of a solution does not increase for

E) explicit schemes

I) implicit schemes

S) semi-discrete schemes.

The total variation of a vector u is

$$TV(u) = \sum_k |u_k - u_{k-1}| \ .$$

Using the right shift operator T:

$$T(u)_k = u_{k-1}$$

we can express TV(u) as

$$TV(u) = |(I-T)u|_1 \ . \tag{3.1}$$

We turn now to explicit (2J+1) point schemes

$$u^{n+1} = D(u^n) \tag{3.2}$$

where

$$D(u)_k = \sum_{-J}^{J} d_j(k) u_{k-j} \ . \tag{3.3}$$

We assume that the difference operator D preserves constants. In view of (3.3), this is the case if

$$\sum_j d_j(k) = 1 \tag{3.4}$$

for all k. Schemes (3.3) satisfying this condition can be written in the form

$$D(u)_k = u_k + \sum_{-J<j<J} e_j(k)(u_{k-j} - u_{k-j-1}) \tag{3.5}$$

or in operator notation

$$D = I + E(I-T), \tag{3.6}$$

where

$$E = \sum e_j T^j \quad . \tag{3.6}*$$

We want to find conditions which guarantee that D is TVD, i.e., satisfies for all u

$$TV(Du) \leq TV(u). \tag{3.7}$$

Using formula (3.1) this is the same as

$$|(I-T)Du|_1 \leq |(I-T)u|_1. \tag{3.7}*$$

Using formula (3.6) we can write

$$(I-T)D = (I+(I-T)E)(I-T) = A(I-T), \tag{3.8}$$

where

$$A = I + (I-T)E. \tag{3.8}*$$

We now set (3.8) into (3.7)*; denoting

$$(I-T)u = u^*$$

we obtain the equivalent inequality

$$|Au^*|_1 \leq |u^*|_1 \ . \tag{3.9}$$

This is certainly the case if A is an ℓ_1 contraction, for which we have derived in Section 2 the criterion (2.5):

$$\sum_j |a_j(h+j)| \leq 1 \tag{3.10}$$

where

$$(Au)_k = \sum_j a_j(k)u_{k-j} \quad .$$

It follows from (3.8)* that the coefficients a_j of A can be expressed in terms of the coefficients e_j of E as

$$a_0(k) = 1 + e_0(k) - e_{-1}(k-1) \tag{3.11}$$

and

$$a_j(k) = e_j(k) - e_{j-1}(k-1), \quad j \neq 0. \tag{3.11}*$$

It follows from these relations that

$$\sum_j a_j(h+j) = 1;$$

but then (3.10) can hold if and only if for all j and k

$$a_j(k) \geq 0.$$

Using (3.11), (3.11)* we can express this condition as follows:

$$\begin{aligned} &e_{-1}(k-1) \geq e_{-2}(k-2) \geq \ldots \geq e_{-J}(k-J) \geq 0, \\ &-e_0(k) \geq -e_1(k+1) \geq \ldots \geq -e_{J-1}(k+J-1) \geq 0, \\ &1 + e_0(k) - e_{-1}(k-1) \geq 0. \end{aligned} \tag{3.12}$$

Thus we have proved

Theorem E: The explicit scheme (3.3) is TVD if conditions (3.12) are satisfied for all k, where e_j are the coefficients appearing in formula (3.5) for D.

We turn next to implicit schemes:

$$F(u^{n+1}) = u^n \quad . \tag{3.13}$$

We take F to be a 2J+1 term difference operator that preserves constants. Such an F can be written in the form

$$F = I + G(I-T) \tag{3.14}$$

where

$$G(u)_k = \sum_{-J\leq j<J} g_j(k)u_{k-j} \qquad (3.14)^*$$

We want to find conditions under which scheme (3.13) is TVD, i.e., for all u

$$TV(Fu) \geq TV(u) \quad . \qquad (3.15)$$

Using formula (3.1) this is the same as

$$|(I-T)Fu|_1 \geq |(I-T)u|_1. \qquad (3.15)^*$$

Using formula (3.14) we can write

$$(I-T)F = (I+(I-T)G)(I-T) = B(I-T) \qquad (3.16)$$

where

$$B = I + (I-T)G. \qquad (3.16)^*$$

We set (3.16) into (3.15)*; denoting

$$(I-T)u = u^*$$

we obtain the equivalent inequality

$$|Bu^*|_1 \geq |u^*|_1. \qquad (3.17)$$

This is the case if B is an expansion. In theorem B we have derived criterion (2.12) that guarantees that an operator B is an expansion:

$$b_0(h) \geq \sum_{j\neq 0} |b_j(h+j)| + 1. \qquad (3.18)$$

It follows from (3.16)* that the coefficients b_j of B can be expressed in terms of the coefficients g_j of G as

$$b_0(k) = 1 + g_0(k) - g_{-1}(k-1)$$

and (3.19)

$$b_j(k) = g_j(k) - g_{j-1}(k-1), \quad j\neq 0$$

Adding up these relations we deduce that

$$b_0(k) = 1 - \sum_{j\neq 0} b_j(k+j);$$

but then (3.18) can hold if and only if for all k and for $j \neq 0$

$$b_j(k) \leq 0.$$

Using (3.19) these conditions can be restated as

$$g_0(k) \geq g_1(k+1) \geq \ldots \geq g_{J-1}(k+J-1) \geq 0 \tag{3.20}$$

and

$$-g_{-1}(k-1) \geq -g_{-2}(k-2) \geq \ldots \geq -g_{-J}(k-J) \geq 0. \tag{3.20}*$$

Thus we have proved

Theorem I: The implicit scheme (3.13) is TVD if conditions (3.20), (3.20)* are satisfied, where g_j are the coefficients of the operator G related by formula (3.14), (3.14)* to the operator F appearing in (3.13).

We remark that we can combine, as Harten does, theorems I and E to study implicit-explicit schemes of the form

$$F(u^{n+1}) = D(u^n). \tag{3.21}$$

Such a scheme is TVD if F satisfies the conditions of Theorem I and D the conditions of Theorem E.

Finally we turn to semi-discrete schemes:

$$\frac{du}{dt} = Hu, \tag{3.22}$$

with H some 2J+1 point difference operator. We assume that $u \equiv$ const is a solution of (3.22); this is the case if H annihilates all constant vectors. In this case H can be written in the form

$$H(u)_k = \sum_{-J \leq j \leq J} m_j(k)(u_{k-j} - u_{k-j-1}), \tag{3.23}$$

or in operator form

$$H = M(I-T). \tag{3.23}*$$

We want to find conditions on H which guarantee that TV(u) is a decreasing function of t for all solutions u of (3.22). By formula (3.1), this is the same as

$$|(I-T)u(t)|_1$$

being a decreasing function of t. So we multiply (3.22) by (I-T); using (3.23)* we get

$$\frac{d}{dt}(I-T)u = (I-T)\,M(I-T)u = C(I-T)u \tag{3.24}$$

where

$$C=(I-T)M. \tag{3.25}$$

Denoting

$$(I-T)u=u^*$$

(3.24) becomes

$$\frac{d}{dt}\,u^* = Cu^* \quad .$$

According to Theorem C, $|u^*|_1$ is a decreasing function of t if condition (2.19) of Section 2 is satisfied:

$$c_0(k) + \sum_{j\neq 0} |c_j(k+j)| \leq 0 \quad . \tag{3.26}$$

Using (3.25) we can express the coefficients c_j in terms of those of M as follows:

$$c_j(k) = m_j(k) - m_{j-1}(k-1) \quad . \tag{3.27}$$

Thus

$$\sum_j c_j(k+j) = 0;$$

It follows from this that (3.26) can hold if and only if

$$c_j(k+j) \geq 0, \; j \neq 0 \quad .$$

Using (3.27) we can restate this as

$$m_{-1}(k-1) \geq m_{-2}(k-2) \geq \ldots \geq m_{-J}(k-J) \geq 0 \qquad (3.28)$$

and

$$-m_0(k) \geq -m_1(k+1) \geq \ldots \geq m_{J-1}(k+J-1) \geq 0 \quad . \qquad (3.28)^*$$

Thus we have proved

Theorem S: The semi-discrete scheme (3.22) is TVD if conditions (3.28) and (3.28)* are satisfied, where the m_j are the coefficients of the operator M related by formula (3.23)* to the operator H.

4. Conclusion

The conservation law (1) describes a right running wave when a(u) is positive. Conditions (3.12) and (3.28) of Theorems (E) and (S) state that the explicit and semi-discrete schemes (E) and (S) are TVD if and only if the coefficients of the differences $u_{k-j} - u_{k-j-1}$ have the same sign as a(u) for $j > 0$, (points on the upwind side), and the opposite sign for $j < 0$, (points on the downwind side). If the differences are moved over to the right of equation (3.13), then condition (3.20) of Theorem (I) states that the implicit scheme (I) will be TVD if it satisfies a similar condition on the sign of its coefficients. In all three cases only the differences on the upwind side have the correct sign for consistency with (1), and can contribute to wave propagation in the correct direction. In this sense upwind biasing is a necessary feature of explicit TVD schemes, and it is also useful in the construction of implicit TVD schemes.

It is thus not surprising to find that most of the attempts to design schemes with the capability of capturing shockwaves and contact discontinuities, dating back to the early work of Courant, Isaacson and Rees [8], and Godunov [9], have introduced upwinding either directly or indirectly. Second order accurate upwind schemes have been devised by Van Leer [10], Harten [4], [5], Roe [11], Osher and Chakravarthy [12], and Sweby [13]. These all use flux limiters to attain the TVD property.

Another approach to the construction of TVD schemes stems from the observation that central difference formulas for odd and even derivatives have odd and even distributions of signs, and they can be superposed and combined with flux limiters to satisfy conditions (3.12) or (3.28). Upwind biasing is then produced indirectly by cancellation of terms of opposite sign. One possible

starting point for such a construction is a central difference scheme in which the numerical flux $1/2(f_{j+1} + f_j)$ is augmented by a third order dissipative flux. This scheme is the basis of a method which has been widely used to solve the Euler equations of compressible flow [14]. It can be converted into an attractively simple TVD scheme by the introduction of flux limiters in the dissipative terms [15]. The modified numerical flux retains a symmetric distribution of terms about the cell boundary $j + 1/2$. The resulting symmetric scheme is one of the variants of a class of symmetric TVD schemes recently proposed by Yee [16]. Her derivation follows an entirely different line of reasoning, building on the work of Davis [17], and Roe [18]. In comparison with upwind TVD schemes, symmetric TVD schemes offer a significant reduction of computational complexity, while exhibiting comparable shock capturing capabilities.

References

1. P.D. Lax, "Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves," SIAM Regional Series on Applied Mathematics, 11, 1973.

2. S.N. Kruzkov, "First Order Quasi-Linear Equations in Several Independent Variables," Math. USSR SB, 10, 1970, pp. 217-243.

3. O.A. Oleinik, "Discontinuous Solutions of Nonlinear Differential Equations," Uspekhi Mat. Nauk., 12, 1957, pp. 3-73, American Math. Soc. Transl., Series 2, 26, pp. 95-172.

4. A. Harten, "High Resolution Schemes for Hyperbolic Conservation Laws," New York University Report DOE/ER 03077-175, 1982, J. Comp. Phys., 49, 1983, pp. 357-393.

5. A. Harten, "On a Class of High Resolution Total Variation Stable Finite Difference Schemes," New York University Report DOE/ER/03077-176, 1982, SIAM J. Num. Anal., 21, 1984, pp. 1-21.

6. A. Jameson and P.D. Lax, "Conditions for the Construction of Multi-Point Total Variation Diminishing Difference Schemes," Princeton University Report MAE 1650, April 1984.

7. S. Osher and S. Chakravarthy, "Very High Order Accurate TVD Shemes," ICASE Report 84-44, Sept. 1984.

8. R. Courant, E. Isaacson, and M. Rees, "On the Solution of Nonlinear Hyperbolic Differential Equations," Comm. Pure Appl. Math., 5, 1952, pp. 243-255.

9. S.K. Godunov, "A Finite Difference Method for the Numerical Computation of Discontinuous Solutions of th Equations of Fluid Dynamics," Mat. Sbornik, 47, 1959, pp. 271-290, translated as JPRS 7225 by U.S. Dept. of Commerce, 1960.

10. B. Van Leer, "Towards the Ultimate Conservative Difference Scheme. II. Monotonicity and Conservation Combined in a Second Order Scheme", J. Comp. Phys., 14, 1974, pp. 361-370.

11. P.L. Roe, "Some Contributions to the Modelling of Discontinuous Flows," Proc. AMS/SIAM Seminar on Large Scale Computation in Fluid Mechanics, San Diego, 1983.

12. S. Osher, and S. Chakravarthy, "High Resolution Schemes and the Entropy Condition," ICASE Report NASA CR 172218, SIAM J. Num. Analysis, 21, 1984, pp. 955-984.

13. P.K. Sweby, "High Resolution Schemes Using Flux Limiters for Hyperbolic Conservation Laws," SIAM J. Num. Anal., 21, 1984, pp. 995-1011.

14. A. Jameson, "Solution of the Euler Equations by a Multigrid Method," Applied Math. and Computation, 13, 1983, pp. 327-356.

15. A. Jameson, "A Non-Oscillatory Shock Capturing Scheme Using Flux Limited Dissipation," Princeton University Report MAE 1653, April 1984, Lectures in Applied Mathematics, Vol. 22, Part 1, Large Scale Compuatations in Fluid Mechanics, edited by B.E. Engquist, S. Osher, and R.C.J. Sommerville, AMS, 1985, pp. 345-370.

16. H.C. Yee, "Generalized Formulation of a Class of Explicit and Implicit TVD Schemes," NASA TM86775, July 1985.

17. S.F. Davis, "TVD Finite Difference Schemes and Artificial Viscosity," ICASE Report 84-20, June 1984.

18. P.L. Roe, "Generalized Formulation of TVD Lax-Wendroff Schemes," ICASE Report 84-53, Oct. 1984.

SOME RESULTS ON

UNIFORMLY HIGH ORDER ACCURATE ESSENTIALLY

NON-OSCILLATORY SCHEMES

Ami Harten[1]
Department of Mathematics, UCLA and School of
Mathematical Sciences, Tel-Aviv University.

Stanley Osher[1], Bjorn Engquist[1]
Department of Mathematics, UCLA.

and

Sukumar R. Chakravarthy
Rockwell Science Center, Thousand Oaks, Ca.

ABSTRACT

We continue the construction and the analysis of essentially nonoscillatory shock capturing methods for the approximation of hyperbolic conservation laws. These schemes share many desirable properties with total variation diminishing schemes, but TVD schemes have at most first order accuracy in the sense of truncation error, at extrema of the solution. In this paper we construct an hierarchy of uniformly high order accurate approximations of any desired order of accuracy which are tailored to be essentially nonoscillatory. This means that, for piecewise smooth solutions, the variation of the numerical approximation is bounded by that of the true solution up to $O(h^{R-\frac{1}{2}})$, for $0 < R$ and h sufficiently small. The design involves an essentially non-oscillatory piecewise polynomial reconstruction of the solution from its cell averages, time evolution through an approximate solution of the resulting initial value problem, and averaging of this approximate solution over each cell. To solve this reconstruction problem we use a new interpolation technique that when applied to piecewise smooth data gives high-order accuracy whenever the function is smooth but avoids a Gibbs phenomenon at discontinuities.

[1]Research supported by NSF Grant No. DMS85-03294, ARO Grant No. DAAG29-85-K-0190, NASA Consortuim Agreement No. NCA2-IR390-403, and NASA Langley Grant No. NAG1-270.

I. INTRODUCTION

In this paper we consider numerical approximations to weak solutions of the hyperbolic initial value problem (IVP)

$$u_t + f(u)_x = 0 = u_t + a(u)u_x \tag{1.1a}$$

$$u(x,0) = u_0(x)\ . \tag{1.1b}$$

Here u and f are m vectors, and $a(u) = \partial f/\partial u$ is the Jacobian matrix, which is assumed to have only real eigenvalues and a complete set of linearly independent eigenvectors.

The initial data $u_0(x)$ are assumed to be piecewise-smooth functions that are either periodic or of compact support.

Let $v_j^n = v_h(x_j,t_n)$, $x_j = jh$, $t_n = na$, denote a numerical approximation in conservation form.

$$v_j^{n+1} = v_j^n - \lambda(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}) = (E_h \cdot v^n)_j\ . \tag{1.2a}$$

Here E_h is the numerical solution operator, $\lambda = \tau/h$, and $\hat{f}_{j+1/2}$, the numerical flux, is a function of $2k$ vector variables:

$$\hat{f}_{j+1/2} = \hat{f}(v_{j-k+1}^n \cdots v_{j+k}^n) \tag{1.2b}$$

which is consistent with (1.1a) in the sense that

$$\hat{f}(u,u,\ldots,u) = f(u)\ . \tag{1.2c}$$

We shall also consider a semi-discrete method of lines approximation to (1.1a) obtained by dividing (1.2a) by τ and letting $a \downarrow 0$

$$\frac{\partial v_j}{\partial t} = -\frac{1}{h}(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}) = \frac{((E_h - I)\cdot v)_j}{\Delta t} \tag{1.3}$$

with $\hat{f}_{j+1/2}$ again satisfying (1.2.b, c).

The numerical approximation in (1.2) is considered to be an approximation to the cell average of u:

$$v_j^n \approx \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x,t_n)\mathrm{d}x \,. \tag{1.4}$$

Accordingly we define its total variation in x to be:

$$\mathrm{TV}(v^n) = \mathrm{TV}(v_h(\cdot,t_n)) = \sum_j |v_{j+1}^n - v_j^n| \tag{1.5}$$

where $|\ \ |$ denotes any norm on R^m.

If the total variation of the numerical solution is uniformly bounded in h, for $0 \le t \le T$,

$$\mathrm{TV}(v_h(\cdot,t) \le C\,\mathrm{TV}(u_0)\,, \tag{1.6}$$

then any refinement sequence $h \to 0, \tau = O(h)$ has a subsequence $h_j \to 0$ such that

$$v_{h_j} \xrightarrow{L_1} u \tag{1.7}$$

where u is a weak solution of (1.1).

If all limit solutions (1.7) of the numerical solution (1.2) satisfy an entropy condition that implies uniqueness of the I.V.P. (1.1), then the numerical scheme is convergent (see, e.g. [5], [14]).

For the semi-discrete approximation, (1.3), we consider:

$$v_j(t) \approx \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x,t)\mathrm{d}x \,. \tag{1.8}$$

The analogous statements concerning TV and convergence are valid as well in this case, see, e.g. [12].

We shall now concentrate on the scalar case, $m = 1$. Extensions to systems will be discussed in sections III and V.

Recently total variation diminishing (TVD) schemes have been designed and analyzed [5], [6]. There the approximate solution is required to diminish the total variation (1.5) of the numerical solution in time:

$$\mathrm{TV}(v_h(\cdot,t_1)) \le \mathrm{TV}(v_n(\cdot,t_2) \ \text{ if } \ t_1 > t_2 \,. \tag{1.9}$$

These schemes trivially satisfy (1.6) with $C = 1$.

We were able to construct TVD schemes that in the sense of local truncation error are of high-order accuracy everywhere except at local extrema where they necessarily degenerate to first order accuracy (see [5], [6], [12], [14], [15], [17]). The perpetual damping of local extrema determines the cumulative global error of the "high-order TVD schemes" to be $O(h^{1+1/p})$ in the L_p norm. This improves by one order in steady state calculations, see [1].

In a sequence of papers of which this is the second, we show how to construct essentially non-oscillatory schemes (ENO) that are uniformly high-order accurate (in the sense of global error for smooth solutions of (1.1)) to any finite order.

In the first paper [7] we constructed a uniformly second-order accurate scheme which is non-oscillatory in the sense that the *number* of local extrema in the numerical solution is non-increasing. Unlike TVD schemes, which also have this property, members of this class are not required to damp the values of each local extremum in time, but are allowed occasionally to accentuate a local extremum.

In this paper the schemes (1.2) are constructed to be essentially non-oscillatory. Our goal is that, if the initial data $u_0(x)$ are *piecewise* smooth, then for h sufficiently small

$$\mathrm{TV}(v_h(\cdot,t+\Delta t)) \le \mathrm{TV}(v_h(\cdot,t)) + O(h^{N+1}) \tag{1.10}$$

where N is the order of accuracy of (1.2). This implies that, at each time step, the scheme is non-oscillatory modulo $O(h^{N+1})$.

The format of this paper is as follows. In section II we shall give the design principle and overview of the present method, including comparisons with TVD schemes. Section III consists of certain variants and extensions of the scheme including extensions to systems and to regions with boundaries. Section IV gives the interpolation algorithm, which is the crux of the method, along with the key result - Theorem (4.1). Several examples are also given. Section V gives further analysis of the interpolation method and an example showing that general non-oscillatory schemes need additional proper-

ties (which we believe to be true for the present methods) to guarantee convergence. We also analyze the truncation error of our methods in this section. Proofs of some technical results are given in an Appendix. We refer the reader to references [24] and [25] for numerical results using these methods.

II. Design Principle, Overview, and Examples.

In this section we describe how to construct ENO schemes of any desired accuracy.

Integrating the partial differential equation (1.1a) over the computational cell $(x_{j-1/2}, x_{j+1/2}) \times (t_n, t_{n+1})$, we get

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \lambda[\hat{f}_{j+1/2}(u) - \hat{f}_{j-1/2}(u)] , \tag{2.1a}$$

where

$$\hat{f}_{j+1/2}(u) = \frac{1}{\tau} \int_{t^n}^{t^{n+1}} f(u(x_{j+1/2},t))dt \tag{2.1b}$$

and

$$\bar{u}_j^n = \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x,t_n)dx . \tag{2.1c}$$

We shall also be interested in a semi-discrete approximation to (1.1), so we divide (2.1a) by τ and let $\tau \downarrow 0$:

$$\frac{\partial}{\partial t}\bar{u}_j = \frac{-[f(u(x_{j+1/2},t)) - f(u(x_{j-1/2},t))]}{h} , \tag{2.2a}$$

where again

$$\bar{u}_j = \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x,t)dt . \tag{2.2b}$$

We observe that although (2.1a) is a relation between cell-averages $\bar{u}_j^n$ and $\bar{u}_j^{n+1}$, the evaluation of the fluxes $\hat{f}_{j+1/2}(u)$ in (2.1b) requires knowledge of the solution itself, not its cell averages.

As in Godunov's scheme [4] and its second order extensions [20], [2], we derive our scheme as a

direct approximation to (2.1). We denote by v_j^n the numerical approximation to the cell averages $\bar{u}_j^n$ of the exact solution to (2.1) and set v_j^0 to be the cell averages of the initial data. Given $v^n = \{v_j^n\}$, we compute v^{n+1} as follows:

First we reconstruct $u(x,t_n)$ out of its approximate cell-averages $\{v_j^n\}$ to the appropriate accuracy and denote the result by $L(x;v^n)$. Next we solve the IVP:

$$v_t + f(v)_x = 0,\ v(x,0) = L(x;v^n) \tag{2.3}$$

and denote its solution by $v(x,t)$. Finally we obtain v_j^{n+1} by taking cell averages of $v(x,\tau)$:

$$v_j^{n+1} = \frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} v(x,\tau)dx\ . \tag{2.4}$$

The averaging procedure is TVD, as is the exact solution operator. We may conclude, therefore, that the design of ENO high order accurate schemes boils down to a problem on the level of approximation of functions: that of constructing an essentially non-oscillatory high-order accurate interpolant of a piecewise smooth function from its cell averages.

In section IV we shall construct an essentially non-oscillatory piecewise polynomial of order N, $Q^N(x;w)$ that interpolates a piecewise-smooth function $w(x)$ at the cell interface points:

$$Q^N(x_{j+1/2};w) = w(x_{j+1/2}) \tag{2.5a}$$

and satisfies, wherever $w(x)$ is smooth

$$\left(\frac{d}{dx}\right)^r Q^N(x \pm 0;u) = \left(\frac{d}{dx}\right)^r w(x) + O(h^{N+1-r}),\ r = 1,\ldots,N\ . \tag{2.5b}$$

The key result, contained in Theorem (4.1) in section IV below, is the following. For any piecewise smooth function $w(x)$, there exists an $h_0 > 0$ and a function $z(x)$, such that for $0 < h \le h_0$:

$$Q^N(x;w) = z(x) + O(h^{N+1}) \tag{2.6a}$$

$$\mathrm{TV}(z) \le \mathrm{TV}(w)\ . \tag{2.6b}$$

We shall use this polynomial together with two different approaches to design ENO schemes. These methods are:

RP: Reconstruction via the primitive function.

RD: Reconstruction via deconvolution.

We begin with RP. Let $W(x)$ be the primitive function of $u(x)$

$$W(x) = \int_a^x u(s)ds \ . \tag{2.7}$$

The lower limit shall play no role in what follows, so we choose it to be $a = x_{-1/2}$, for simplicity of exposition. Thus since we wish to reconstruct $u(x)$ out of its approximate cell averages v_j (dropping the t or n dependence) we have an approximation to $W(x_{j+1/2})$

$$W(x_{j+1/2}) = \sum_{k=0}^{j} v_k h \ . \tag{2.8}$$

In each cell $I_j:\{x/x_{j-1/2} \le x < x_{j+1/2}\}$, $Q^N(x;w)$ is a polynomial of degree N which interpolates $w(x_{j+1/2})$; i.e., for all j

$$Q^N(x_{j+1/2};w) = w(x_{j+1/2}) \ . \tag{2.9}$$

Thus $Q^N(x,w)$ is a continuuous piecewise polynomial, and both of $d/dx\ Q^N(x \pm 0;w)$ are globally well defined.

Our approximation to (1.1) can be obtained by solving (2.3) with

$$v(x,w) = d/dx\ Q^N(x;w^n) = L(x;v^n) \ ,$$

obtaining $v(x,t)$, $0 \le t \le \tau$ and then computing cell averages (2.4). This can be rewritten, using the divergence theorem, as:

$$v_j^{n+1} = v_j^n - \lambda(\hat{f}_{j+1/2}^n - \hat{f}_{j-1/2}^n) \ , \tag{2.10}$$

since

$$\frac{Q^N(x_{j+1/2};w^n) - Q^N(x_{j-1/2};w^n)}{h} = v_j^n$$

because of (2.5a) and (2.8).

Here $\hat{f}^n_{j+1/2}$ is computed by averaging the flux function $f(u)$ applied to $v(x_{j+1/2},t)$ as in (2.1b).

In the linear case:

$$u_t + au_x = 0 \; ; \tag{2.11}$$

this procedure is easy to carry out. The exact solution to IVP (2.3) is

$$v(x,t) = L(x - at;v^n) = \frac{d}{dx} Q^N(x - at;w^n) \; ; \tag{2.12}$$

thus the scheme becomes

$$v_j^{n+1} = (E_h \cdot v^n)_j = v_j^n - \lambda(\hat{f}^n_{j+1/2} - \hat{f}^n_{j-1/2}) = \tag{2.13a}$$

$$= v_j^n - \frac{1}{h}\Big([Q^N(x_{j+1/2};w^n] - Q^N(x_{j+1/2} - a\tau;w^n)]$$

$$-[Q^N(x_{j-1/2};w^n) - Q^N(x_{j-1/2} - a\tau;w^n)]\Big)$$

$$= \frac{1}{h}[Q^N(x_{j+1/2} - a\tau;w^n) - Q^N(x_{j-1/2} - a\tau;w^n)]$$

given the CFL restriction[1]

$$|a\lambda| \le 1 \; . \tag{2.13b}$$

The numerical flux functions $\hat{f}^n_{j+1/2}$ defined here involve values of $Q^N(x;v^n)$ for x between $x_{j-1/2}$ and $x_{j+1/2}$ if $a > 0$, or $x_{j+1/2}$ and $x_{j+3/2}$ if $a < 0$. Thus, unsurprisingly the resulting scheme has an upwind bias.

For general $f(u)$ the explicit solution to (2.3) can be difficult to obtain. However, for $N = 1$, the initial data are piecewise constant:

[1]The restriction can be easily removed in this constant coefficient case.

$$L(x;v^n) \equiv v_j^n, \ x_{j-1/2} \le x < x_{j+1/2} .$$

Thus the scheme becomes:

$$v_j^{n+1} = v_j^n - \lambda[f(v(x_{j+1/2}) - f(v(x_{j-1/2})] , \tag{2.14a}$$

where $v(x_{j+1/2}) \equiv v(x_{j+1/2},t)$, for $0 < t \le \tau$, if the CFL restriction

$$|\lambda f'(u)| < 1 , \tag{2.14b}$$

for all u such that: $\min(v_j^n, v_{j+1}^n) \le u \le \max(v_j^n, v_{j+1}^n)$, is satisfied.

This is precisely Godunov's scheme [4], which is the canonical three point, upwind, first order accurate method [9]. Thus our higher order methods are simply generalizations of Godunov's techniques to higher order ENO schemes. The first higher-order TVD (although the concept was not yet defined) Godunov type method was introduced by van Leer [20]. See [8], [2], and [20] for theoretical and practical results concerning such TVD methods. The difference here, of course, is that we allow our interpolant to be arbitrarily high-order accurate even at extrema, and we replace the restrictive TVD condition [6], [10], by the ENO property.

A key step in this method comes in solving to the Riemann problem (1.1a, b), with initial data consisting of two constant states

$$u(x,0) \equiv u_L, x \le 0$$

$$u(x,0) \equiv u_R, x > 0 .$$

The unique entropy condition satisfying similarity solution was obtained in [9]. The resulting scheme (2.14a) can be written:

$$v_j^{n+1} = v_j^n - \lambda(\hat{f}_{j+1/2}^G - \hat{f}_{j-1/2}^G) \tag{2.15a}$$

where

$$\hat{f}_{j+1/2}^G = f^G(v_j, v_{j+1}) = \begin{cases} \min\limits_{v_j \le u \le v_{j+1}} f(u) , & \text{if } v_j \le v_{j+1} \\ \max\limits_{v_j \ge u \ge v_{j+1}} f(u) , & \text{if } v_j > v_{j+1} \end{cases} . \tag{2.15b}$$

The corresponding semi-discrete approximation is just:

$$\frac{\partial}{\partial t} v_j = -\frac{1}{h}(\hat{f}^G_{j+1/2} - \hat{f}^G_{j-1/2}) . \tag{2.16}$$

Although the high-order explicit method described above can have a complicated flux function, its semi-discrete limit is much simpler. We merely take limits as in (2.2a) and arrive at

$$\frac{\partial}{\partial t} v_j = \frac{-1}{h}(\hat{f}^G\left(\frac{d}{dx} Q^N(x_{j+1/2} - 0;w^n), \frac{d}{dx} Q^N(x_{j+1/2} + 0;w^n)\right) \tag{2.17}$$

$$-\hat{f}^G\left(\frac{d}{dx} Q^N(x_{j-1/2} - 0;w^n), \frac{d}{dx} Q^N(x_{j-1/2} + 0;w^n)\right))$$

i.e., Godunov's method with more accurate constant initial states.

Next we use RD. This time we begin with $u(x)$ and denote by $\bar{u}(x)$ its mean over $(x - h/2, x + h/2)$, i.e.,

$$\bar{u}(x) = \frac{1}{h}\int_{x-h/2}^{x+h/2} u(y)dy = \int_{-1/2}^{1/2} u(x + sh)ds . \tag{2.18}$$

Denote by $\bar{u}_j = \bar{u}(x_j)$, the cell-averages of $u(x)$.

Again, given cell averages v_j which approximate $\bar{u}_j$ we wish to reconstruct $u(x)$ up to $O(h^{N+1})$ in an essentially non-oscillatory way. Here we again begin by constructing a piecewise polynomial interpolant of order N, which we again call $Q^N(x;v)$, that interpolates v at x, for each j:

$$Q^N(x_j;v) = v_j . \tag{2.19}$$

This time $Q^N(x;v)$ is a polynomial of degree N in the interval $x_j \le x < x_{j+1}$, with possible jump discontinuities in derivatives at the end points. Then we compute an essential non-oscillatory piecewise polynomial of degree $N - 1$ as follows:

$$I^{N-1}(x;v) = v_j + \sum_{r=1}^{N-1} \frac{1}{r!}(x - x_j)^r \tag{2.20}$$

$$m\left[\left(\frac{d}{dx}\right)^r Q^N(x_j - 0;v), \left(\frac{d}{dx}\right)^r Q^N(x_j + 0,v)\right] ,$$

defined for

$$x_{j-1/2} \le x \le x_{j+1/2} .$$

Here m is the min mod function:

$$m(x,y) = \begin{cases} s \min(|x|,|y|) & \text{if } \operatorname{sgn}x = \operatorname{sgn}y = s \\ 0 & \text{otherwise} \end{cases} \tag{2.21}$$

This gives us our approximation to v, which may have discontinuities at each $x_{j+1/2}$. We use this to obtain an approximant to $u(x)$ via a "deconvolution" procedure. We have approximate derivatives to $\bar{u}(x)$[2]:

$$\left(h\frac{d}{dx}\right)^r \bar{u}(x)|_{x=x_j} = h^v m\left(\left(\frac{d}{dx}\right)^r Q(x_j - 0;v),\right. \tag{2.22}$$

$$\left.\left(\frac{d}{dx}\right)^r Q(x_j + 0;v)\right) + O(h^{r+1}),\ r = 0,1,\ldots,N-1 .$$

At points of smoothness, we have

$$\left(h\frac{d}{dx}\right)^k \bar{u}(x) = \int_{-1/2}^{1/2} \left(h\frac{d}{dx}\right)^k u(x+sh)ds \tag{2.23}$$

$$= \sum_{r=0}^{N-k-1} \frac{1}{r!}\left(h\frac{d}{dx}\right)^{k+r} u(x) \int_{-1/2}^{1/2} s^r dr + O(h^N)$$

$$= \sum_{r=0}^{N-k-1} \left(h\frac{d}{dx}\right)^{k+r} u(x)\, \frac{1}{2^r(r+1)!}\, \frac{[1-(-1)^{r+1}]}{2} + O(h^N) ,$$

for $k = 0,1,\ldots,N-1$.

Thus we may write the Toëplitz upper triangular matrix equation:

[2]This will be shown for piecewise smooth $\bar{u}(x)$ in section IV.

$$\begin{bmatrix} \bar{u}(x_j) \\ h\frac{d}{dx}\bar{u}(x_j) \\ \cdot \\ \cdot \\ \left(h\frac{d}{dx}\right)^{N-1}\bar{u}(x_j) \end{bmatrix} = \tag{2.24}$$

$$\begin{bmatrix} 1 & 0 & \frac{1}{4\cdot 3!} & \cdots & \cdots \\ 0 & 1 & 0 & \frac{1}{4\cdot 3!} & \cdots \\ & & 1 & 0 & \\ & & & \cdot & \\ & & & & \cdot \\ 0 & \cdots & 0 & & 1 \end{bmatrix} \begin{bmatrix} u(x_j) \\ \left(h\frac{d}{dx}\right)u(x_j) \\ \cdot \\ \cdot \\ \left(h\frac{d}{dx}\right)^{N-1}u(x_j) \end{bmatrix}$$

This is easily inverted and gives us each of the terms $(hd/dx)^v\, u(x_j)$, up to $O(h^N)$.

We replace the left side of (2.24) by the approximations on the corresponding right side of (2.22) for each v. We invert this system in (2.24) and call the computed approximate values $(h^v d/dx)^v v(x_j)$.

For $x_{j-1/2} \le x < x_{j+1/2}$, we write our approximation as

$$L^{N-1}(x;v) = \sum_{v=0}^{N-1} \left[\left(h\frac{d}{dx}\right)^v v(x_j)\right] \frac{(x-x_j)^v}{h^v v!}\,. \tag{2.25}$$

We need the following:

LEMMA (2.1)

The cell average is preserved under this operation, i.e.:

$$\frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} L^{N-1}(x;u)dx = \bar{u} \tag{2.26}$$

Proof

A direct computation gives us:

$$\frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} L^{N-1}(x;u)dx = \sum_{\nu=0}^{N-1}\left[\left(h\frac{d}{dx}\right)^{\nu} u(x_j)\right]\frac{1}{2^{\nu}(\nu+1)!}\left[\frac{1-(-1)^{\nu+1}}{2}\right] = \bar{u}_j$$

from the first row of (2.24).

Now we contine our scheme construction as we did using RP. In the RD approach we approximate (1.1) by solving (2.3) with $v(x,0) = L^{N-1}(x;0) = L(x;v^n)$ and proceed as above. We again arrive at (2.10). In the linear case (2.11) the resulting numerical fluxes are defined via

$$\hat{f}^n_{j+1/2} = a\int_0^1 L^{N-1}(x_{j+1/2} - as\tau;v^n)ds\ , \tag{2.27}$$

given the CFL restriction (2.13b).

Also the semi-discrete algorithm for general f obtained via RP in (2.17) is replaced by its analogue with the numerical flux

$$\hat{f}^G(L^{N-1}(x_{j+1/2} - 0), L^{N-1}(x_{j+1/2} + 0))\ . \tag{2.28}$$

III. Variants of the Scheme.

The exact solution to the special initial value problem (2.3) can be difficult to compute. This is, of course, particularly true when the initial data is a piecewise polynomial of degree higher than zero, but is also usually true for general systems of equations for piecewise constant initial data, i.e., for Godunov's method. One can, however, obtain a convergent power series expansion for this solution see [22], [23].

Godunov's method is canonical in the class of (scalar) E schemes, defined in [9]. A consistent numerical flux yields a semi-discrete E scheme iff

$$[\mathrm{sgn}(u_{j+1} - u_j)]\hat{f}_{j+1/2} \le [\mathrm{sgn}(u_{j+1} - u_j)]\hat{f}^G_{j+1/2}\ , \tag{3.1}$$

or equivalently, iff its viscosity is greater than or equal to that of Godunov's method [18]. E schemes

are TVD and entropy condition satisfying; thus they always converge to the correct physical solution [19], [18]. Examples include the Engquist - Osher scheme and entropy corrections of Roe's scheme - see, e.g. [3], [16].

One property all E schemes share is the fact that they can be obtained by averaging a solution to a Riemann problem over each cell, where f is replaced by an approximation $\tilde{f}$, in equation (1.1) - see [18], [14]. Thus they retain the ENO property. We may let $\hat{f}^E_{j+1/2} = \hat{f}^E(v_j, v_{j+1})$ be any two point E flux and generalize our semi-discrete algorithm (2.17) to:

$$\frac{\partial}{\partial t} v_j = \frac{-1}{h}\left[\hat{f}^E\left(\frac{d}{dx} Q^N(x_{j+1/2} - 0; w^n), \frac{d}{dx} Q^N(x_{j+1/2} + 0; w^n)\right)\right. \tag{3.2}$$

$$\left. -\hat{f}^E\left(\frac{d}{dx} Q^N(x_{j-1/2} - 0; w^n), \frac{d}{dx} Q^N(x_{j-1/2} + 0; w^n)\right)\right].$$

We may generalize (2.28) analogously.

Next we replace the exact numerical flux:

$$\hat{f}^n_{j+1/2} = \frac{1}{\tau}\int_0^\tau f(v(x_{j+1/2}, t))dt \tag{3.3}$$

by an approximation based on a Taylor series as follows.

For $x_{j-1/2} < x < x_{j+1/2}$, we can compute the quantities

$$\left(\frac{\partial}{\partial t}\right)^r v(x,0)$$

for v the solution of (2.3), by using a Lax-Wendroff type of procedure.

For example:

$$\frac{\partial v}{\partial t}(x,0) = \frac{-\partial}{\partial x} f(v(x,0))$$

$$\frac{\partial^2 u(x,u)}{\partial x \partial t} = -f''(u(x,0))(u_x(x,0))^2 - f'(u(x,0))u_{xx}(x,0)$$

$$\frac{\partial^2 u(x,0)}{\partial t^2} = -f''(u(x,0))\frac{\partial u}{\partial t}(x,0)\frac{\partial u}{\partial x}(x,0) - f'(u(x,0))\ \frac{\partial^2 u}{\partial x \partial t}(x,0)\ .$$

Next we write an approximation to $v(x,t)$:

$$v^R(x,t) = v(x,0) + t\ \frac{\partial v}{\partial t}\ (x,v) + \cdots + \frac{t^R}{R!}\ \frac{\partial^R v(x,0)}{\partial t}\ . \tag{3.4}$$

Now we replace the integral in (3.3) by a quadrature rule

$$\int_0^1 f(v(x_{j+1/2}; os)ds \approx A_0 f(v(x_{j+1/2}, s_0)) + \cdots + A_k f(v(x_{j+1/2}, s_k)) \tag{3.5}$$

$$\text{for } 0 \le s_0 < s_1 \cdots < s_k \le 1\ .$$

Finally, we define each value of f above as:

$$f(v(x_{j+1/2}, s_r)) = f^G(v(x_{j+1/2} - 0, s_r),\ v(x_{j+1/2} + 0, s_r)) \tag{3.6a}$$

if we base our approach on Godunov's method. More generally we can replace Godunov's flux by its generalization.

$$f(v(x_{j+1/2}, s_r)) = f^E(v(x_{j+1/2} - 0, s_r),\ v(x_{j+1/2} + 0, s_r))\ . \tag{3.6b}$$

Thus we approximate (3.3) by a sum of piecewise constant Godunov methods, or approximate Godunov methods, evaluated at several time layers. The quadrature rule, and the value of R, determine the order of time accuracy of this method.

We note that this approximation need not preserve the essential non-oscillatory property. Nevertheless, due to the (nonlinear) nature of our ENO interpolant, the method works well numerically, as is seen from the results in [24] and [25].

Next we consider hyperbolic *systems* of conservation laws (1.1). In the linear case, $f(u) = Au$, where A is a constant matrix with a complete set of right and left eigenvectors $r^{(\nu)}$, $l^{(\nu)}$, corresponding to real eigenvalues $\lambda^{(\nu)}$, for $\nu = 1,\ldots,m$. We proceed formally as in (2.1), (2.2), (2.3), and it just becomes a matter of computing the vector valued function $L(x;v^n) = v(x,0)$ in (2.3).

We decompose an arbitrary m vector w as

$$w = \sum_{\nu=1}^{m} (l^{(\nu)} \cdot w) r^{(\nu)} = \sum_{\nu=1}^{m} w^{(\nu)}$$

using the usual l^2 inner product. These are used to construct $L(x;v^n)$ again via the RP or RD reconstruction approaches.

The RP approach proceeds by computing

$$w^{(\nu)}(x_{j+1/2}) = \sum_{k=0}^{j} v_k^{(\nu)} h \ .$$

Then we proceed, as in the scalar case, to compute each of $Q^N(x,w^{(\nu)})$, and finally by letting

$$L(x;v^n) = \sum_{r=1}^{m} \left[\frac{d}{dx} Q^N(x;w^{(\nu),n}) \right] r^{(\nu)} \ . \tag{3.7}$$

The RD approach begins by computing $Q^N(x;v^{(\nu)})$ which is a piecewise polynomial interpolant of order N that interpolates $v^{(\nu)}$ at each x_j. The rest of the reconstruction procedure is done as in the scalar case, and finally we replace (2.25) by

$$L^{N-1}(x;v) = \sum_{\nu=1}^{m} L^{N-1}(x;v^{(\nu)}) r^{(\nu)} \ .$$

For nonlinear systems we denote by $A_j = \partial f/\partial u\ (v_j)$, the Jacobian matrix evaluated at v_j, and define $\lambda_j^{(\nu)}$, $l_j^{(\nu)}$, and $r_j^{(\nu)}$ in the usual fashion. This time we decompose

$$v = \sum_{\nu=1}^{m} (l_{j_0}^{(\nu)} \cdot v)\ r_{j_0}^{\nu} = \sum_{\nu=1}^{m} v^{(\nu)J_0} \ . \tag{3.8}$$

For each v and each j_0, we shall construct an ENO scalar interpolant such that, in the cell $x_j \le x < x_{j+1}$. $Q^{NJ_0+0}(x,v^{(\nu)})$ is the unique Nth degree polynomial that interpolates $v^{(\nu J_0)}(x_j)$ for $j = j_0, j_0 + 1$ and $N - 1$ neighboring points as defined in section IV, and $Q^{NJ_0-0}(x,v^{(\nu)})$ which interpolates $v^{(\nu J_0)}(x_j)$ for $j = j_0, j_0 - 1$ and the appropriate $N - 1$ neighboring points.

We then construct the m-vector valued ENO piecewise polynomial of degree $N-1$ as follows:

$$I^{N-1}(x;v) = v_j + \sum_{\mu=1}^{N-1} \sum_{\nu=1}^{m} \frac{1}{\mu!}(x-x_j)^{\mu}\, m\left[\left(\frac{d}{dx}\right)^{\mu} Q^{NJ_0-0}(x_j - 0; v^{(\nu)}),\right.$$

$$\left.\left(\frac{d}{dx}\right)^{\mu} Q^{NJ_0+0}(x_j+0; v^{(\nu)})\right] r_j^{(\nu)}, \qquad (3.9)$$

for $x_{j-1/2} \le x < x_{j+1/2}$.

We may then deconvolve precisely as in the scalar case and arrive at a vector-valued version of (2.25). Moreover Lemma (2.1) is still valid.

The RP approach is done analogously.

Thus using either RP or RD we have enough information to compute the vector valued analogue of (3.2) - the semi-discrete algorithm. This time the canonical method is again Godunov's which uses the exact solution to the Riemann problem. Other, simpler approximate Riemann solvers may be used - e.g., Osher's [13], van Leer's for the Euler equations of compressible gas dynamics [21], or Roe's [15] with an entropy fix as suggested in [16], [17].

The explicit vector-valued construction follows the procedure of (3.5), (3.6), again using perhaps one of the approximate Riemann solvers to replace Godunov's method.

Various simplifications of these procedures are possible and will be discussed in future papers.

Next we discuss the influence of boundaries on our procedure.

We illustrate the idea by considering the linear equation (2.11), with $a \neq 0$, to be solved for $t, x > 0$, with initial data of compact support. If $a > 0$, then a physical boundary condition $u(0,t) = g(t)$ must be given. If $a < 0$, then no physical boundary conditions are needed.

The modifications needed are two-fold:

(1) At points sufficiently near the boundary our ENO interpolant will lack a choice of least oscillatory direction. We will choose only among interpolation points which lie inside the region. This

procedure has not led to stability problems in our numerical experiments. This can be explained by the adaptive nature of the stencil in the interior. However, in situations where discontinuities flow into or out of boundaries, oscillations may develop. These oscillations do not seem to pollute the solution globally according to our (now rather extensive) numerical experimentation. We regard this as essentially the same problem that we have when discontinuities intersect in the interior. We shall discuss these matters in future papers. Some relevant numerical experiments are presented in [24].

(2) Instead of an initial-value problem, at $x = 0$ we solve an initial-boundary value problem. This is easy in the scalar case - if $a > 0$, we just use the given boundary condition, and if $a < 0$, we need no boundary condition since the wave propagates to the left.

For general systems of equations we follow the same procedure, i.e., interpolating in the interior directions when forced to, and solving an initial-boundary Riemann problem - perhaps approximately. See [10] for more details about the latter.

One variant of the scheme which we do not recommend involves interpolation of the fluxes to obtain a high order method. This was done in [11] in a TVD context, and schemes of arbitrarily high order away from critical points of the function $f(u)$ in (1.1) were obtained. One might think that our ENO interpolant might be used on the fluxes using the decomposition of an E scheme into its "upwind" and "downwind" parts

$$d\hat{f}^{-}_{j+1/2} = \hat{f}^{E}_{j+1/2} - f(v_j)$$

$$d\hat{f}^{+}_{j+1/2} = f(v_{j+1}) - \hat{f}^{E}_{j+1/2}$$

as in [11]. The difficulty here occurs because of the lack of smoothness of $\hat{f}^{E}$ which generally occurs at sonic points. This degrades the accuracy to be at most third order in L^1 at sonic points, if, e.g., the Engquist - Osher flux is used, and second order for Godunov's or Roe's methods.

IV. Essentially Non-Oscillatory Interpolation and Some Examples

Consider a scalar mesh function $\{v_j\}_{j=-\infty}^{\infty}$.

We let $Q(x;v)$ be an interpolant:

$$Q(x_j;v) = v_j = v(x_j),\ j = \ldots -1,0,1,\ldots, \tag{4.1}$$

$$x_j = jh,\ h > 0\ .$$

We shall study a special piecewise polynomial interpolant of degree N, $Q^N(x;v)$, defined recursively as follows:

Definition (4.1)

$$Q^1(x;v) = v_j + (x - x_j)\,\frac{(v_{j+1} - v_j)}{h}\ ,\ x_j \le x < x_{j+1} \tag{4.2a}$$

$$= v[x_j] + [x - x_j]\,v[x_j, x_{j+1}]\ ,$$

where $v[x_{j-\mu},\ldots,x_{j+\nu}]$ denotes the usual coefficient in the Newton interpolant.

We also define:

$$K_{\min}^{(1)} = j,\ K_{\max}^{(1)} = j + 1\ . \tag{4.2b}$$

Suppose we have defined $Q^{N-1}(x,v)$ for $x_j \le x < x_{j+1}$, and that we also have $K_{\min}^{(N-1)}$, $K_{\max}^{(N-1)}$. Then we compute

$$a^N = v[x_{K_{\min}^{(N-1)}},\ldots,x_{K_{\max}^{(N-1)}+1}] \tag{4.2c}$$

$$b^N = v[x_{K_{\min}^{(N-1)}-1},\ldots,x_{K_{\max}^{(N-1)}}]$$

and proceed inductively.

If $|a^N| \ge |b^N|$, then

$$Q^N(x;v) = Q^{N-1}(x;v) + b^N \prod_{K=K_{\min}^{(N-1)}}^{K_{\max}^{(N-1)}} (x - x_K) \tag{4.2d}$$

with

$$K_{\min}^{(N)} = K_{\min}^{(N-1)} - 1 \,. \tag{4.2e}$$

Or if $|a^N| < |b^N|$, then

$$Q^N(x;v) = Q^{N-1}(x;v) + a^N \prod_{K=K_{\min}^{(N-1)}}^{K_{\max}^{(N-1)}} (x - x_K) \tag{4.2f}$$

with

$$K_{\max}^{(N)} = K_{\max}^{(N-1)} + 1 \,.$$

Thus, in each cell $x_j \leq x < x_{j+1}$, we have constructed a polynomial of degree N which interpolates $v(x)$ at $N + 1$ consecutive points which include x_j and x_{j+1}. It is designed so that all its derivatives are as small in absolute value as is possible, given the above constraints.

Remark (4.1)

This interpolant can introduce small oscillations of order h^{N+1} even for monotone and smooth data $v(x)$.

We use the following:

Example (4.1)

Let

$$v(x) = x^3 \tag{4.3a}$$

$$N = 2 \tag{4.3b}$$

$$x_j = (j - 1/2)h \,. \tag{4.3c}$$

The interpolant $Q^2(x;v)$ for x between x_1 and x_2 will interpolate $v(x)$ at x_1, x_2, x_3.

We rescale, letting

$$x' = \frac{x}{h} + \frac{1}{2} \tag{4.4a}$$

$$v(x) = \frac{4v(x)}{h^3} + \frac{1}{2}\,. \tag{4.4b}$$

We get

$$Q^2(x';v') = -5x' + 6(x')^2 \tag{4.4c}$$

so a new extrema occurs at $x' = 5/12$ i.e. at $x = -h/12$. The magnitude is $O(h^3)$ in the unscaled variables.

Our next result shows that this is the worst possible case for $N = 2$.

In fact for piecewise smooth data and h sufficiently small the largest possible spurious oscillations for Q^N will be $O(h^{N+1})$.

THEOREM (4.1)

For any piecewise smooth $v(x)$, possibly having jump discontinuities, there exist an $h_0 > 0$ and a function $z(x)$, such that, for all $h \le h_0$

$$Q^N(x;v) = z(x) + O(h^{N+1}) \tag{4.5a}$$

where

$$\mathrm{TV}(z) \le \mathrm{TV}(v) \tag{4.5b}$$

and we repeat:

$$Q^N(x_j;v) = v(x_j),\ j = 0, \pm 1, \pm 2, \ldots\,. \tag{4.5c}$$

PROOF

Consider the interval $x_j \le x < x_{j+1}$ and study two cases;

(i) v is smooth in $[x_j, x_{j+1}]$

(ii) v has a jump discontinuity in $[x_j, x_{j+1}]$.

Case (i): If v is smooth over the full interval of interpolation $[x_{K^{(N)}_{min}}, x_{K^{(N)}_{max}}]$, standard interpolation results imply $Q^N = v + O(h^{N+1})$, so we then take $z(x) = v(x)$. Otherwise, for h_0 small enough, there exists an interval containing $N + 1$ consecutive parts such that all divided differences $w[\ ,\ ,\ ]$ involving points in this interval are bounded independently of h. We call the point at the extreme right $x_{K^{(N)}}$. If $[x_{K^{(N)}}, x_{K^{(N)}+1}]$ contains a discontinuity in x, then

$$v[x_K, \ldots, x_{K^{(N)}+1}] = \frac{v(x_{K^{(N)}+1}) - v(x_{K^{(N)}})}{m!h^m} + O(h^{-m+1}) , \tag{4.6a}$$

where

$$m = K^{(N)} - K + 1 . \tag{4.6b}$$

This follows from the explicit form of $v[\ \ ]$.

Hence the definition of $Q^N(x;v)$ guarantees that, for h small enough, there will be no discontinuity of $v(x)$ in the interval of interpolation $[x_{K^{(N)}_{min}}, x_{k^{(N)}_{max}}]$. The result above is still valid: $Q^N = v + O(h^{N+1})$.

Case (ii): We may suppose h_0 is small enough so that $v(x)$ has only one discontinuity in $[x_{K^{(N)}_{min}}, x_{K^{(N)}_{max}}]$, and it is in $[x_j, x_{j+1})$. For a given interval of interpolation we may decompose:

$$v = w + H \tag{4.7}$$

where w is Lipschitz continuous and H is piecewise constant with a single jump which occurs in $[x_j, x_{j+1})$.

We have in $[x_j, x_{j+1})$:

$$Q^N(x;v) = Q^N(x;w) + Q^N(x;H) , \tag{4.8a}$$

where:

$$Q^N(x;w) = \sum_{\nu=K_{\min}^{(N)}}^{K_{\max}^{(N)}} v[x_\nu,\ldots,x_{K_{\max}^{(N)}}] \prod_{u=j+1}^{K_{\max}^{(N)}} (x - x_u) , \tag{4.8b}$$

and

$$|v[x_\nu,\ldots,x_{K_{\max}^{(N)}}]| \le C \; h^{\nu - K_{\max}+1} \tag{4.8c}$$

(where C always denotes any universal positive constant).

This implies that

$$\left|\frac{d}{dx} Q^N(x;w)\right| \le C . \tag{4.9}$$

By Rolle's Theorem the interpolant $Q^N(x;H)$ of the piecewise constant function must have an extremum in every interval $(x_\nu, x_{\nu+1})$ for $\nu \neq j$, $K_{\min}^{(N)} \le r \le K_{\max}^{(N)}$. This makes a total of $N - 1$ extrema. Since the interpolant is of degree N, it must be monotone in $[x_j, x_{j+1}]$.

Thus, for $h = 1$, we have

$$\max_{x_j \le x \le x_{j+1}} \left|\frac{d}{dx} Q^N(x;H)\right| \ge C > 0 . \tag{4.10a}$$

For general h, the scaling gives

$$\max_{x_j \le x \le x_{j+1}} \left|\frac{d}{dx} Q^N(x;H)\right| \ge \frac{C}{h} . \tag{4.10b}$$

Thus (4.8)-(4.10) imply that $Q^N(x;v)$ is monotone in $[x_j, x_{j+1}]$. We take

$$z(x) = Q^N(x;v) . \tag{4.11}$$

On the interval $[x_j, x_{j+1}]$

$$\mathrm{TV}(Q^N(x;v)) = |v_{j+1} - v_j| \le \mathrm{TV}(v) . \tag{4.12}$$

The theorem is proven.

We also have

Remark (4.2)

Let $v(x)$ be piecewise polynomial of degree $\leq N$. Then in any interval $[x_j, x_{j+1}]$ in which $v(x)$ is not discontinuous the interpolant is exact

$$Q^N(x;0) \equiv v(x) .$$

We now compute "second" and "third" order accurate approximations to the linear problem.

$$u_t = -u_x \tag{4.13}$$

Using RP for $N = 2$, we have, for $x_{j-1/2} \leq x < x_{j+1/2}$,

$$Q^2(x;w) = w_{j-1/2} + (x - x_{j-1/2})v_j + \tag{4.14a}$$

$$+ \frac{1}{2}\frac{(x - x_{j-1/2})(x - x_{j+1/2})}{h}\,\bar{m}(\Delta_- v_j, \Delta_+ v_j)$$

where

$$\bar{m}(x,y) = x \quad \text{if} \quad |x| \leq |y| \tag{4.14b}$$

$$\bar{m}(x,y) = y \quad \text{if} \quad |y| > |x|$$

and

$$\Delta_{\mp} v_j = \mp(v_{j\mp1} - v_j) . \tag{4.14c}$$

The algorithm becomes

$$v_j^{n+1} = v_j^n - \lambda\Delta_-\left[v_j^n + \left(\frac{1 - \lambda}{2}\right)\bar{m}(\Delta_- v_j^n, \Delta_+ v_j^n)\right] . \tag{4.15}$$

This is a TVD scheme [6] for $\lambda < 1$ which is second order accurate with a first order degeneracy at critical points.

For $N = 2$ with the RD approach, a simple calculation gives the algorithm for $|\lambda| < 1$:

$$v_j^{n+1} = v_j^n - \lambda\Delta_-[v_j^n + \left(\frac{1-\lambda}{2}\right)m[\Delta_- v_j^n + \frac{1}{2}\bar{m}(\Delta_-\Delta_+ v_j^n, \Delta_-\Delta_- v_j^n), \tag{4.16}$$

$$\Delta_+ v_j^n - \frac{1}{2}\bar{m}(\Delta_-\Delta_+ v_j^n, \Delta_-\Delta_+ v_j^n)]] \, .$$

(In [7] we obtained a similar algorithm, with both of the $\bar{m}$ replaced by m. We proved that the scheme in [7] was truly non-oscillatory.)

The scheme (4.6) is truly second order accurate, even at critical points and converges for $|\lambda| < 1$, at least according to extensive numerical tests.

Using $N = 3$ in the RP approach gives us for $x_{j-1/2} \le x < x_{j+1/2}$:

$$\text{If } |\Delta_- v_j| \le |\Delta_+ v_j| \, , \text{ then:} \tag{4.17a}$$

$$Q^3(x;w) = w_{j-1/2} + (x - x_{j-1/2})v_j + \frac{1}{2}\frac{(x - x_{j-1/2})(x - x_{j+1/2})}{h}\Delta_- v_j$$

$$+ \frac{1}{6h^2}(x - x_{j-3/2})(x - x_{j-1/2})(x - x_{j+1/2})\bar{m}(\Delta_-\Delta_- v_j, \Delta_-\Delta_+ v_j) \, .$$

$$\text{If } |\Delta_- v_j| > |\Delta_+ v_j| \, , \text{ then} \tag{4.18b}$$

$$Q^3(x;w) = w_{j-1/2} + (x - x_{j+1/2})v_j + \frac{1}{2}\frac{(x - x_{j-1/2})(x - x_{j+1/2})}{h}\Delta_+ v_j$$

$$+ \frac{1}{6h^2}(x - x_{j-3/2})(x - x_{j+1/2})(x - x_{j+3/2})\bar{m}(\Delta_-\Delta_+ v_j, \Delta_+\Delta_+ v_j) \, .$$

Then our numerical scheme becomes for $|\lambda| < 1$:

$$v_j^{n+1} = v_j^n - \lambda\Delta_-[v_j^n + \left(\frac{1-\lambda}{2}\right)\bar{m}(\Delta_- v_j^n, \Delta_+ v_j^n) \tag{4.19}$$

$$+ \begin{bmatrix} \frac{1}{6}(\lambda - 1)(\lambda - 2)\bar{m}(\Delta_-\Delta_- v_j^n, \Delta_-\Delta_+ v_j^n), & \text{if } |\Delta_- v_j^n| \le |\Delta_+ v_j^n| \\ \frac{1}{6}(\lambda - 1)(\lambda + 1)\bar{m}(\Delta_-\Delta_+ v_j^n, \Delta_+\Delta_+ v_j^n) \, , & \text{if } |\Delta_- v_j^n| > |\Delta_+ v_j^n| \end{bmatrix}$$

This scheme is third order accurate except perhaps at points where u_x or $u_{xx} = 0$, at which it may degenerate to second order accuracy.

For $N = 3$ using RD we have for $x_j \le x < x_{j+1}$:

$$Q^3(x;v) = v_j \tag{4.20}$$

$$+\frac{(x-x_j)}{h}\Delta_+ v_j + \frac{(x-x_j)(x-x_{j+1})}{2h^2}\bar{m}(\Delta_-\Delta_+ v_j, \Delta_-\Delta_+ v_j)$$

$$+\frac{1}{6h^3}\begin{cases}(x-x_{j-1})(x-x_j)(x-x_{j+1})\bar{m}(\Delta_-\Delta_-\Delta_+ v_j, \Delta_-\Delta_+\Delta_+ v_j), \text{if } |\Delta_-\Delta_+ v_j| \le |\Delta_+\Delta_+ v_j| \\ (x-x_j)(x-x_{j+1})(x-x_{j+2})\bar{m}(\Delta_-\Delta_+\Delta_+ v_j, \Delta_+\Delta_+\Delta_+ v_j), \text{if } |\Delta_-\Delta_+ v_j| > |\Delta_+\Delta_+ v_j|\end{cases}.$$

We can derive a globally third order accurate scheme by using (2.20), (2.24), (2.25), and (2.27). We omit the details here.

It should be stressed that our algorithms are to be obtained recursively using the computer. We have written down a few numerical fluxes here just to give the reader some idea of what they look like.

V. FURTHER THEORETICAL RESULTS AND EXAMPLES

While Theorem (4.1) is encouraging in that it shows us that the interpolant Q^N is indeed essentially nonoscillatory, more analysis needs to be done. The schemes designed in section II do not use this function in a simple enough fashion for us to prove the desired estimate (1.10), even if $v_h(x,t)$ is piecewise smooth.

As a step in this direction we consider the method based on RP applied to a piecewise continuous function. A canonical example involves the interpolant $Q^N(x;g)$, where $g(x)$ is the primitive of a Heaviside function normalized to be:

$$g(x) = \alpha - x, x \le \alpha \tag{5.1a}$$

$$g(x) = 0, x > \alpha \tag{5.1b}$$

for $0 \leq \alpha < 1$.

We let $h = 1$ and compute the least oscillatory piecewise polynomial $Q^N(x,g)$ which interpolates $g(x)$ at $x = j$ for each integer j. By Remark (4.2) we have

$$Q^N(x;g) = g(x) \text{ for } x \leq 0 \text{ and } x \geq 1 . \tag{5.2}$$

We need only compute $Q^N(x;g)$ for $0 < x < 1$. We wish the reconstructed function $d/dx\ Q^N(x,g)$ to be a non-oscillatory approximation to $g(x)$. This reduces to showing that on $0 \leq x \leq 1$

$$-1 \leq \frac{d}{dx} Q^N(x;g) \leq 0 \tag{5.3a}$$

$$\frac{d^2}{dx^2} Q^N(x;g) \geq 0 . \tag{5.3b}$$

The least oscillatory polynomial on the interval $0 \leq x \leq 1$ will be one of the $N + 1$ polynomials of degree N, $Q_K^N(x;g)$, which interpolates $g(x)$ at the $N + 1$ consecutive points

$$\{K - N, K - N + 1,\ldots,0, 1,\ldots,K\}$$

for $1 \leq K \leq N$.

In the proof of Theorem (4.1) we showed that *any* polynomial which interpolates the derivature of this function $g'(x)$ through these $N + 1$ points is monotone on the interval $0 \leq x \leq 1$ In contrast we have

Example (5.1)

$$Q_1^N = (\alpha - x) + \frac{(x + N - 1)(x + N - 2)\ldots x}{N!} [1 - \alpha] \tag{5.4a}$$

thus:

$$\frac{d}{dx} Q_1^N (x,g)|_{x=1} = -1 + (1 - \alpha)[1 + \cdots + \frac{1}{N}] > 0 .$$

for $N(\alpha)$ sufficiently large, when α is fixed: $1 > \alpha > 0$.

Thus, in order for the inequalities (5.3) to be valid, we need the special properties of the least oscillatory interpolant of $g(x)$. We have:

Lemma (5.1)

The least oscillatory polynomial of degree N is Q_k^N iff

$1 - K/N \leq \alpha < 1 - (K-1)/N$, $K = 1,2,\ldots,N$.

Finally we have:

Lemma (5.2)

The polynomial obtained in the statement of Lemma (5.1) satisfies the inequalities (5.3).

We shall present the proofs of these claims in the Appendix.

Next we consider the method based on RP applied to a smooth perturbation of a Heaviside function $g'(x)$. We find here two new problems.

(1) The error between $d/dx\, Q^N(x;g)$ and $g'(x)$ in the cell next to the interval containing the discontinuity need not be $O(h^N)$ - it can be as bad as $O(h)$ for $N > 1$.

(2) The variation in this cell can increase - i.e., $\mathrm{Var}[d/dx\, Q^N(x;g)]$ in this cell can exceed that of $g'(x)$ in this cell by $O(h^2)$ for $N > 2$.

On the plus side we note that these are somewhat pathological examples, that the error and growth in variation are indeed decaying with h, and that two cells away from the discontinuity all seems well in that the error and possible variation growth appear to be $O(h^N)$. Nevertheless we expect to investigate other ENO interpolation procedures as well as alternative reconstruction techniques, with an aim towards removing these (hopefully minor) problems.

Example (5.2)

Let

$$g(x) = \frac{(x + Bh)^2}{2} + a(x + Bh), x > -Bh \tag{5.4a}$$

$$g(x) = -x - Bh, x \leq -Bh \text{ for } 1 > B > 0 . \tag{5.4b}$$

Then the function we are approximating, $g'(x)$, satisfies

$$g'(x) = -1, x \leq -Bh \tag{5.5a}$$

$$g'(x) = x + Bh + a \quad x > -Bh . \tag{5.5b}$$

We shall obtain $Q^N(x;g)$ which interpolates $g(x)$ at grid points $x_j = jh$, $j = 0, \pm 1$. We are interested in Q^N for $0 \leq x \leq h$. We shall arrange a and B so that

$$\frac{d}{dx} Q^N(x;g) \text{ for } N = 2 \text{ and } 3$$

both have $O(h)$ pointwise error compared to that of $g'(x)$ on this interval.

We do this as follows:

For $0 \leq x \leq h$:

$$Q^1(x;g) = g(0) + \frac{x}{h}(g(h) - g(0)) .$$

Next we arrange a and B so that the three consecutive points $(-h,g(-h))$, $(0,(g(0))$, and $(h,g(h))$ are collinear:

$$g(h) - 2g(0) + g(-h) = 0 \tag{5.6a}$$

or

$$g(-h) = h^2 - \frac{(B - 1)^2h^2}{2} - ah(B - 1) = h(1 - B) \tag{5.6b}$$

or

$$\frac{h}{2}(B-1)^2 + (a+1)(B-1) - h = 0\,. \tag{5.6c}$$

We solve this obtaining

$$B(h,a) = 1 + \frac{h}{a+1} + O(h^2)\,, \tag{5.6d}$$

and since we want $0 \le B \le 1$, we take $a < -1$.

Thus we have for $0 \le x \le h$

$$Q^2(x;g) = Q^1(x,g)\,,$$

which interpolates $g(x)$ at $x = -h,0,h$ and

$$\frac{d}{dx}\,Q^2(x;g) = \frac{g(h) - g(0)}{h}$$

which clearly differs from $g'(x)$ by $O(h)$ at some points in this interval.

We also claim that $d/dx\ Q^3(x,g) - g'(x)$ is $O(h)$ in this interval as well. It is easy to see that Q^3 will be chosen to interpolate $g(x)$ at $x = -h,0,h,$ and $2h$. Thus in our interval of interest:

$$Q^3(x;g) = Q^1(x;g) + \frac{(x-h)x(x+h)}{6h^3}\left[\frac{(B-1)h^2}{2} + ah(B-1) + h(B-1)\right].$$

Thus

$$\frac{d^2}{dx^2}\,Q^3(x;g) = \frac{x}{h^2}(B-1)\left[\frac{(B-1)h}{2} + (1+a)\right]$$

and then

$$\operatorname*{var}_{0\le x\le h}\frac{d}{dx}\,Q^3(x;g) = \frac{1}{2}(B-1)\left[\frac{(B-1)h}{2} + (1+h)\right] = \frac{h}{2} + O(h^2)$$

and the error is still $O(h)$ since $\mathrm{var}_{0\le x\le h}\, g'(x) = h$.

Our next example well allow for an *increase* in variation in this cell, although it will still decay with h as $h \downarrow 0$. Let:

$$g(x) = \frac{(x + Bh)^3}{6} + b(x + Bh), x > -Bh \tag{5.7a}$$

$$g(x) = -x - Bh, x \leq -Bh \tag{5.7b}$$

for $1 > B > 0$.

Then the function we approximate is:

$$g'(x) = -1, x \leq -Bh \tag{5.8a}$$

$$g'(x) = \frac{(x + Bh)^2}{2} + b\,. \tag{5.8b}$$

This time we want the points $(-h,g(-h))$, $(0,g(0))$, $(h,g(h))$, and $(2h, g(2h))$ to all lie on the same parabola. This means that

$$O = h^3 + h^3 \frac{(B - 1)^3}{6} + (1 + b)h(B - 1) \tag{5.9a}$$

or

$$B = 1 - \frac{h^2}{b + 1} \tag{5.9b}$$

Thus we take $b > -1$.

On the usual interval $0 \leq x \leq h$ we have

$$Q^1(x;g) = g(0) + \frac{x}{h}(g(h) - g(0)) \tag{5.10a}$$

$$Q^2(x;g) = Q^1(x;g) + \frac{x(x - h)}{2h^2}[g(h) - 2g(0) + g(-h)] \tag{5.10b}$$

and by (5.9):

$$Q^3(x;g) = Q^2(x;g) \tag{5.10c}$$

The function $g'(x)$ is monotone on the interval $0 \leq x \leq h$ as long as b is not $O(h^2)$ so:

$$\operatorname*{var}_{0\le x\le h} g'(x) = \frac{1}{2} h^2(1 + 2B)$$

while

$$\operatorname*{var}_{0\le x\le h} \frac{d}{dx} Q^3(x,g) = \frac{|g(2h) - 2g(h) + g(0)|}{h} = h^2(1 + B) .$$

Thus an oscillation of order $h^2/2$ is induced in what should be a third order method.

We note that the discontinuity in $g'(x)$ is rigged so that it occurs at a distance $O(h^2)$ from a grid point. This is a bit pathological, but is certainly possible.

This oscillation is maintained even when we increase the order. For example, in the same interval it can be easily shown:

$$Q^4(x;g) = Q^2(x;g) + \frac{(x - 2h)(x - h)x(x + h)}{24h^4} [g(-h) - \tilde{g}(-h)]$$

where $\tilde{g}(x)$ is the continuation of the cubic polynomial $g(x)$ to x negative.

Thus

$$g(-h) - \tilde{g}(-h) = h(1 - B) - \left[\frac{h^3}{6} (B - 1)^3 + bh(B - 1)\right]$$

$$= h(1 + b)(1 - B) + \frac{h^3}{6} (1 - B)^3 = h^3 + O(h^9)$$

A simple calculation gives us

$$\operatorname*{var}_{0\le x\le h} \frac{d}{dx} Q^4(x;g) = h^2\left[\frac{5}{6} + B\right] + O(h^5) ,$$

and we again have a variation increase $O(h^2)$ in this interval.

Next we show that a scheme, which is non-oscillatory for relevant data in that new extrema do not develop on the initial data as h increases, can still be extremely unstable.

Example 5.4

We approximate

$$u_t = -u_x \tag{5.11}$$

by:

$$v_j^{n+1} = v_j^n - \lambda \Delta_- \overline{m}(v_{j+1}^n, v_j^n) \, . \tag{5.12}$$

We take as initial data

$$v_j^0 = 0, \, j \leq -1 \tag{5.13a}$$

$$v_0^0 = a \tag{5.13b}$$

$$v_1^0 = \epsilon - a \tag{5.13c}$$

$$v_j^0 = 0, \, j \geq 2 \tag{5.13d}$$

for $0 < \epsilon \ll a, \, 0 < \lambda < 1/2$.

An explicit computation gives us:

$$v_j^1 = v_j^0 \text{ if } j \leq -1, j \geq 2 \tag{5.14a}$$

$$v_0^1 = a(1 + \lambda) - \lambda\epsilon \tag{5.14b}$$

$$v_1^1 = (\epsilon - a)(1 + \lambda) \, . \tag{5.14c}$$

Thus the "shape" of the initial data is invariant in time and

$$v_0^n \to \infty$$
$$v_1^n \to -\infty \, .$$

Now we analyze the truncation error TE for our two methods. We begin with RP applied to the linear equation (2.11) and arrive at (2.13). In this case a precise expression for TE is:

$$\text{TE} = \frac{-1}{h^2} \Delta_- [Q^N(x_{j+1/2}; W) - W(x_{j+1/2}) - \tag{5.15a}$$

$$- Q^N(x_{j+1/2} - a\tau;W) + W(x_{j+1/2}-a\tau)]$$

We recall

$$W(x) = \int^x \bar{u}(s)ds \quad \text{with} \tag{5.15b}$$

$$W(x_{j+1/2}) = Q^N(x_{j+1/2};w) \tag{5.15c}$$

and

$$\left(\frac{d}{dx}\right)^\nu Q(x;W) - \left(\frac{d}{dx}\right)^\nu W(x) = O(h^{N+1-\nu}) \tag{5.15d}$$

in regions where $W(x)$ is sufficiently smooth.

It is clear that the TE is $O(h^N)$ as long as the coefficient multiplying the $h^{N+1-\nu}$ term is differentiable when for $\nu = 1$. This will be true in general if the stencil of points used for the interpolant in two consecutive intervals is invariant under translations. This is true in smooth regions if none of the derivatives of $u(x)$ up to order $N - 1$ vanish in a neighborhood of this interval.

We thus have

Theorem (5.1)

TE for the explicit and semi-discrete methods based on RP approximating a linear equation is of order

$$\text{TE} = O(h^N), \text{ if } \left(\frac{d}{dx}\right)^r u(x) \neq 0, r = 1,2,...,N - 1 \tag{5.16a}$$

$$\text{TE} = O(h^{N-1}) \text{ otherwise .} \tag{5.16b}$$

For the full nonlinear problems the algorithm (2.14) can easily be shown to satisfy estimate (5.16b) above.

The computational evidence is that (5.16a) is valid under conditions stated there for the non-

linear case. We believe this to be true, but do not prove it here.

Next we state:

Theorem (5.2)

TE for the explicit and semi-discrete methods based on RP for general nonlinear equations is at least

$$\mathrm{TE} = O(h^{N-1}) \,. \tag{5.17}$$

Finally we analyze TE based on RD. Recall we are given via interpolation the values:

$$a_\nu = h^\nu (d/dx)^\nu \, \bar{u}(x_j) + O(h^{N+1}), \; \nu = 0,1,\ldots,N-1$$

Next we compute

$$b_\nu = h^\nu \left(\frac{d}{dx}\right)^\nu u(x_j) + O(h^N)$$

using the matrix equality:

$$\begin{bmatrix} a_0 \\ \cdot \\ \cdot \\ \cdot \\ a_{N-1} \end{bmatrix} = \begin{bmatrix} 1 & \alpha_1 & \cdots & \cdots & \alpha_{N-1} \\ 0 & 1 & \alpha_1 & \cdots & \alpha_{N-2} \\ 0 & 0 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdot & & & & \\ \cdot & & & & \\ 0 & 0 & & 0 & 1 \end{bmatrix} \begin{bmatrix} u(x_j) \\ \left(h\frac{d}{dx}\right) u(x_j) \\ \cdot \\ \cdot \\ \cdot \\ \left(h\frac{d}{dx}\right)^{N-1} u(x_j) \end{bmatrix} + \tag{5.18a}$$

$$+ h^N \left(\frac{d^N}{dx}\right) u(x_j) \begin{bmatrix} \alpha_N \\ \alpha_{N-1} \\ \cdot \\ \cdot \\ \cdot \\ \alpha_1 \end{bmatrix} + O(h^{N+1})$$

where

$$\alpha_\nu = \frac{1 + (-1)^\nu}{2^\nu(\nu + 1)!} . \tag{5.18b}$$

Call C the upper triangular Toeplitz matrix on the right above. We approximately invert the system, obtaining

$$\begin{bmatrix} b_0 \\ \cdot \\ \cdot \\ \cdot \\ b_{N-1} \end{bmatrix} = C^{-1} \begin{bmatrix} a_0 \\ \cdot \\ \cdot \\ \cdot \\ a_{N-1} \end{bmatrix} = \begin{bmatrix} u(x_j) \\ \left(h \frac{d}{dx}\right) u(x_j) \\ \cdot \\ \cdot \\ \cdot \\ \left(h \frac{d}{dx}\right)^{N-1} u(x_j) \end{bmatrix} \tag{5.19}$$

$$+ C^{-1} h^N \left(\frac{d}{dx}\right)^N u(x_j) \begin{bmatrix} \alpha_N \\ \alpha_{N-1} \\ \cdot \\ \cdot \\ \cdot \\ \alpha_1 \end{bmatrix} + O(h^{N+1}) .$$

Next we compute the function $L^{N-1}(x;u)$ as in (2.25), for $x_{j-1/2} \le x < x_{j+1/2}$

$$L^{N-1}(x;u) = \sum_{\nu=0}^{N-1} b_\nu \frac{(x - x_j)^\nu}{h^\nu \nu!} \tag{5.18}$$

$$= u(x) - \frac{\left[\left(h \frac{d}{dx}\right)^N u(x_j)\right]}{N!\, h^N} (x - x_j)^N$$

$$+ \left[\left(h\frac{d}{dx}\right)^N u(x_j)\right]\left[1, \frac{x-x_j}{h}, \ldots, \frac{(x-x_j)^{N-1}}{h^{N-1}}\right] C^{-1} \begin{bmatrix} \alpha_N \\ \alpha_{N-1} \\ \cdot \\ \cdot \\ \alpha_1 \end{bmatrix}$$

$$+ O(h^{N+1})$$

To show that TE = $O(h^N)$ in this case, we need only prove that

$$D^{(x)}[L^{N-1}(x;u) - u(x)] = O(h^N) \tag{5.19}$$

for u smooth. This follows by the smoothness in both x_j and x up to $O(h^{N+1})$ of the remaining terms on the right side of (5.18).

Thus we have:

Theorem 5.2

TE for the explicit and semi-discrete methods based on RD approximation for general systems of equations is $O(h^N)$.

We also note:

Remark (5.1)

We have been unable, so far, to prove that these methods are indeed essentially non-oscillatory although our present results show that the interpolation upon which the whole framework is based does indeed have this property.

Remark (5.2)

If $u_0(x)$ has two neighboring discontinuities and h is not sufficiently small, our present

methods, for $N > 2$, can result in nontrivial spurious overshoots. We shall remedy this difficulty in subsequent papers.

Appendix

We shall provide the (lengthy) details of the proofs of Lemmas (5.1) and (5.2).

Proof of Lemma (5.1)

We shall use induction on N. The result is trivially true for $N = 1$. Suppose it is true up to N. We consider the interval

$$1 - \frac{K}{N+1} \leq \alpha < 1 - \left(\frac{K-1}{N+1}\right)$$

We divide this into two parts

$$I_R\colon\ 1 - \left(\frac{K-1}{N}\right) \leq \alpha < 1 - \left(\frac{K-1}{N+1}\right) < 1 - \left(\frac{K-2}{N}\right) \tag{A1.a}$$

$$I_L\colon\ \left(1 - \frac{K}{N+1}\right) \leq \alpha < 1 - \left(\frac{K-1}{N}\right) \tag{A1.b}$$

after verifying

$$1 - \left(\frac{K-1}{N+1}\right) < 1 - \left(\frac{K-2}{N}\right)$$

$$K - 2 < N$$

and

$$1 - \frac{K}{N+1} < 1 - \left(\frac{K-1}{N}\right)$$

$$\frac{K-1}{N} < \frac{K}{N+1}$$

$$N/K - N + K - 1 < \frac{N}{K}$$

$$K - 1 < N$$

Thus by the induction hypothesis: for $\alpha \in I_R$, $Q^N(x;g) = Q^N_{K-1}(x;g)$, (if $K = 1$, I_R is empty), and for $\alpha \in I_L$, $Q^N(x;g) = Q^N_K(x;g)$.

We wish to show that for $\alpha \in I_R \cup I_L$ that $Q^{N+1}(x;g) = Q^{N+1}_k(x;g)$. Using the iterative definition of $Q^{N+1}_K(x;g)$, we must compare the two Newton coefficients

$$-R = (\alpha - K) - \binom{N+1}{1}(\alpha - (K-1)) + \cdots + (-1)^{K-1}\binom{N+1}{K-1}(\alpha - 1) \tag{A2.a}$$

$$-S = (\alpha - (K-1)) - \binom{N+1}{1}(\alpha - (K-2)) + \tag{A2.b}$$

$$+ \cdots + (-1)^{K-2}\binom{N+1}{K-1}(\alpha - 1)\,.$$

We wish to show

$$|R| \le |S| \tag{A3}$$

for these values of α.

To prove this we need the following:

Fact (A1)

$$\binom{n}{0} - \binom{n}{1} + \cdots + (-1)^K\binom{n}{k} = (-1)^K\binom{n-1}{K}$$

for $0 \le K \le n - 1$

Fact (A2)

$$K\binom{n}{0} - (K-1)\binom{n}{1} + \cdots + (-1)^{K-1}\binom{n}{K-1} = (-1)^{K-1}\binom{n-2}{K-1}$$

for $1 \le K \le n-1$.

Proof of Fact A1

Again we do it by induction. It is true for $K = 0$. Suppose it is true for K. Add $(-1)^{K+1}\binom{n}{K+1}$ to both sides of the equality. On the right we have

$$(-1)^{K+1}\left(\frac{n!}{(n-K-1)!(K+1)!} - \frac{(n-1)!}{(n-K-1)!K!}\right)$$

$$= (-1)^{K+1}\left(\frac{(n-1)!}{(n-K-2)!(K+1)!}\right)\left(\frac{n}{n-K-1} - \frac{(K+1)}{(n-K-1)}\right)$$

$$= (-1)^{K+1}\binom{n-1}{K+1}$$

Proof of Fact A2

Using induction. We see that it is true for $K = 1$. Suppose it is true for K. Then

$$(K+1)\binom{n}{0} - K\binom{n}{1} + \cdots + (-1)^{K}\binom{n}{K} = K\binom{n}{0} -$$

$$- (K-1)\binom{n}{K} + \cdots + (-1)^{K-1}\binom{n}{K}$$

(by Fact (A1)),

$$= (-1)^{K-1}\binom{n-2}{K-1} + (-1)^{K}\binom{n}{K}$$

(by the induction hypothesis)

$$= (-1)^K \left(\frac{(n-1)!}{K!(n-1-K)!} - \frac{(n-2)!}{(K-1)!(n-1-K)!} \right)$$

$$= (-1)^K \frac{(n-2)!}{K!(n-2-K)!} \left(\frac{n-1}{(n-1-K)} - \frac{K}{(n-K-K)} \right)$$

$$= (-1)^K \binom{n-2}{K}.$$

Using these facts, we have:

$$R = (1-\alpha)(-1)^{K-1} \binom{N}{K-1} + (-1)^{K-2} \binom{N-1}{K-2} \tag{A4.a}$$

$$S = (1-\alpha)(-1)^{K-2} \binom{N}{K-2} + (-1)^{K-3} \binom{N-1}{K-3} \tag{A4.b}$$

We note

$$(-1)^{K-2} R = \binom{N-1}{K-2} - (1-\alpha) \binom{N}{K-1}$$

$$\geq \binom{N-1}{K-2} - \binom{N}{K-1} \left(\frac{K-1}{N} \right)$$

$$= \binom{N-1}{K-2} (1-1) = 0$$

so

$$|R| = \binom{N-1}{K-1} - (1-\alpha) \binom{N}{K-1}.$$

Also

$$(-1)^{K-2} S = (1-\alpha) \binom{N}{K-2} - \binom{N-1}{K-3}$$

$$\geq \frac{K-1}{N+1}\binom{N}{K-2} - \binom{N-1}{K-3}$$

$$= \binom{N-1}{K-3}\left(\frac{N(K-1)}{(N+1)(K-2)} - 1\right) = \binom{N-1}{K-3}\left(\frac{N-K+2}{(N+1)(K-2)}\right) \geq 0$$

$$|S| = (1-\alpha)\binom{N}{K-2} - \binom{N-1}{K-3}.$$

We check

$$|S| \geq |R|$$

$$(1-\alpha)\left[\frac{N!}{(K-2)!(N-K+2)!} + \frac{N!}{(K-1)!(N-K+1)!}\right]$$

$$\geq \frac{(N-1)!}{(K-2)!(N-K+1)!} + \frac{(N-1)!}{(K-3)!(N-K+2)!}$$

$$\frac{(K-1)}{N+1}[N(K-1) + N(N-K+2)] \geq$$

$$\geq (K-1)(N-K+2) + (K-1)(K-2)$$

$$N \geq N$$

For I_L the two Newton coefficients are: the same R and

$$\tilde{S} = (1-a)(-1)^K\binom{N}{K} + (-1)^{K-1}\binom{N-1}{K-1}$$

This time we have

$$(-1)^{K-1}R = (1-\alpha)\binom{N}{K-1} - \binom{N-1}{K-2} \geq \frac{K-1}{N}\binom{N}{K-1} - \binom{N-1}{K-2} = 0$$

and

$$(-1)^{K-1}\,\tilde{S} = \binom{N-1}{K-1} - (1-\alpha)\binom{N}{K} \geq \binom{N-1}{K-1} - \left(\frac{K}{N+1}\right)\binom{N}{K}$$

$$= \binom{N-1}{K-1}\left(1 - \frac{N}{N+1}\right) > 0\,.$$

Finally we check

$$|\tilde{S}| \geq |R|$$

or

$$\binom{N-1}{K-1} + \binom{N-1}{K-2} \geq (1-\alpha)\left[\binom{N}{K-1} + \binom{N}{K}\right]$$

$$\frac{(N-1)!}{(K-1)!(N-K)}\left[1 + \frac{K-1}{N-K+1}\right]$$

$$\geq \frac{KN}{N+1}\,\frac{(N-1)!}{(K-1)!(N-K)!}\left[\frac{1}{(1-K+1)} + \frac{1}{K}\right]$$

$$\frac{N}{N-K+1} \geq \frac{KN}{N+1}\left[\frac{N+1}{K(N-K+1)}\right]$$

Thus Lemma (5.1) is proven.

Proof of Lemma (5.2)

We start with a general geometric result.

Fact (A.3)

Given

$$\frac{d}{dx}\,Q_K^N(1) \leq 0 \qquad \text{(A5.a)}$$

$$\frac{d}{dx}\,Q_K^N(0) \geq -1 \qquad \text{(A5.b)}$$

$$\frac{d}{dx} Q_K^N(1) \geq \frac{d}{dx} Q_K^N(0) . \tag{A5.c}$$

Then

$$\frac{d^2}{dx^2} Q_K^N(x) \geq 0 \text{ for } 0 \leq x \leq 1 . \tag{A6}$$

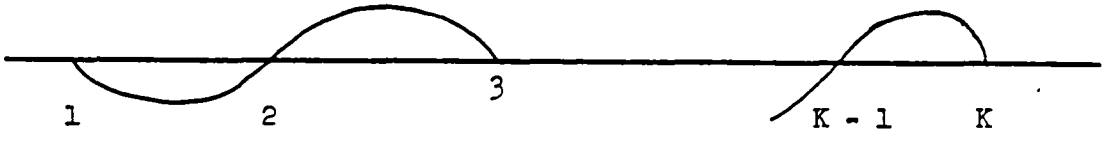
Proof of Fact A3

Rolle's Theorem tells us that $d/dx\ Q_K^N(x,g) = 0$ at least once in each interval $(1,2),\ldots,(K-1,K)$ and $d/dx\ Q_K^N(x,g) = -1$ at least once in each of $(K-N, K-N+1),\ldots,(-2,-1), (-1,0)$. If $K = 1$, this means that $d^2/dx^2\ Q_K^N(x,g) = 0$ at least $N-2$ times for $x < 0$. Thus $d/dx\ Q_K^N(x,g)$ is monotone for $0 \leq x < 1$. If $K = N$, then a similar argument shows that $d^2/dx^2\ Q_N^N(x;g) = 0$ at least $N-2$ times for $x > 1$ and the same monoticity result follows. Given (A5(c)), this takes care of these two cases.

For $1 < K < N$ we proceed as follows. If $d/dx\ Q_K^N = 0$ at least once in addition to these values mentioned above for $1 \leq x \leq K$, then it equals 0 at least K times for $x \geq 1$ and -1 at least $N-K$ time for $x \leq 0$. By our usual argument this means that it is monotone on (0,1), and we are finished. Similarly if $d/dx\ Q_K^N(x) = -1$ at an additional point for $K-N \leq x \leq 0$, the same conclusion follows.

If both of these possibilities are false, then the graph of $Q_K^N(x;g)$ looks like for $1 \leq x$:

(a) K odd



(b) K even

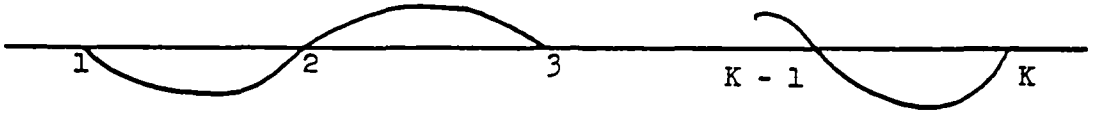
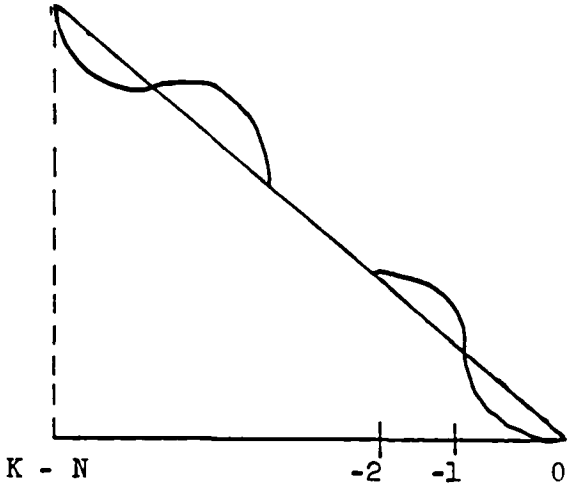


Fig. A1: Q_K^N for $x \geq 1$

and for $K - N \leq x \leq 0$, the graph looks like:

(a) $N - K$ odd



(b) $N - K$ even

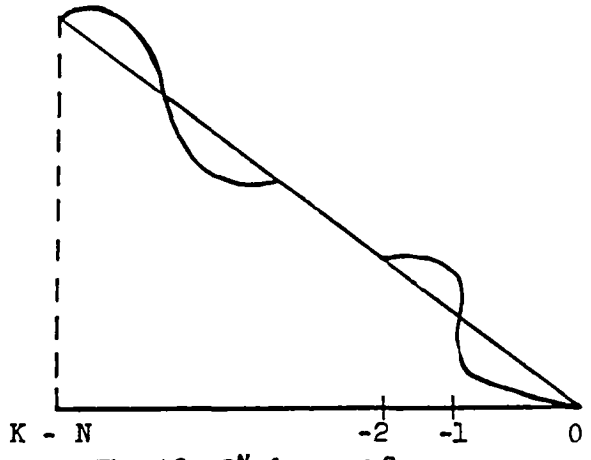


Fig. A2: Q_K^N for $x \leq 0$

If the leading coefficient of Q_K^N vanishes, then we have a polynomial of degree $N - 1$, and its derivative is monotone on $(0,1)$ as per our usual argument.

Otherwise we consider the following cases.

Case (1) K and $N - K$ even. Then if the leading coefficient is positive it follows from glancing at Fig (A2.b) that $d/dx\ Q_K^N = -1$ for some $x < K - N$ and we are finished. If the coefficient is negative then Fig. (A1.b) shows us that $d/dx\ Q_K^N = 0$ for some $x > K$ and we are again finished.

Case (2) K and $N-K$ odd. Then if the coefficient is positive Fig (A1.a) shows $d/dx\ Q_K^N = 0$ for some $x > K$. If the coefficient is negative then Fig (A2.a) shows $d/dx\ Q_K^N = -1$ for some $x < K - N$.

Case (3) K odd, $N-K$ even. If the coefficient is positive then Fig (A1.a) shows $d/dx\ Q_K^N$ vanishes for some $x > K$. If the coefficient is negative then Fig. (A2.b) gives us the desired result.

Case (4) K even, $N-K$ odd. If the coefficient is positive then Fig. (A1.a) gives us the desired result. If the coefficient is negative then Fig (A1b) does it.

To prove Lemma (5.2) we need only verify the inequalities (A5). We finally write down the formula for Q_K^N:

Lemma (A1)

$$Q_K^N = \alpha - x + \sum_{j=1}^{K} \frac{(x+N-K)\cdots(x-j+1)}{(N-K+j)!} \tag{A.7}$$

$$\left[(-1)^{j-1}(1-\alpha)\binom{N-K+j-1}{N-K}\right.$$

$$\left. + (-1)^{j-2}\binom{N-K+j-2}{N-K}\right]$$

where we define $\binom{A}{B} = 0$ if either $B < 0$ or $B > A$.

Proof:

Clearly $Q_K^N = \alpha - x$ for $x = K-N, K-N+1,\ldots,0$.

For $x = 1,2,\ldots,k$, we need

$$\alpha - x + \sum_{j=1}^{K} \binom{N-K+x}{N-K+j}\left[(1-\alpha)(-1)^{j-1}\binom{N-K+j-1}{j-1}\right. \tag{A.8}$$

$$\left. + (-1)^{j-2}\binom{N-K+j-2}{j-2}\right] = 0$$

This will follow if, for all integers $M \geq 0$:

Fact (A.4)

$$1 = \sum_{j=1}^{v} \binom{M+v}{M+j}(-1)^{j-1}\binom{M+j-1}{M}, \quad v \geq 1 \tag{A.9a}$$

Fact (A.5):

$$v = \sum_{j=1}^{v} \binom{M+v+1}{M+j+1}(-1)^{j-1}\binom{M+j-1}{M}, \quad v \geq 1 \tag{A.9b}$$

Proof of Facts (A.4) and (A.5)

We shall again use induction: For $M = 0$ we need

$$0 = -\left[\sum_{j=1}^{v} (-1)^{j-1}\binom{v}{j} - 1\right] = \sum_{j=0}^{v} (-1)^{j}\binom{v}{j} \tag{A.10}$$

This follows from Fact (A.1) for $v = n = K + 1$.

We also need

$$v = \sum_{j=1}^{v} \binom{v+1}{j+1}(-1)^{j-1} \tag{A.11}$$

$$= \sum_{j=-1}^{v} \binom{v+1}{j+1}(-1)^{j-1} - \binom{v+1}{0} + \binom{v+1}{1} = v$$

by (A.10).

Suppose both Facts are true up to M. For $M + 1$ we need

$$1 = \sum_{j=1}^{\nu} \binom{M+1+\nu}{M+1+j}(-1)^{j-1}\binom{M+j}{M+1} \tag{A.12}$$

$$= \sum_{j=1}^{\nu} \left(\frac{M+1+\nu}{M+1+j}\right)\left(\frac{M+j+1}{M+1}\right)\binom{M+\nu}{M+j}(-1)^{j-1}\binom{M+j-1}{M}$$

$$-\sum_{j=1}^{\nu} \frac{1}{M+1}\binom{M+1+\nu}{M+1+j}(-1)^{j-1}\binom{M+j-1}{M}$$

$$= \frac{M+1+\nu}{N+1} - \frac{\nu}{M+1} = 1 \quad \text{(by the induction hypothesis)}$$

Now we show (A.9b) is true for $M+1$ by induction on ν. It is clearly true for $\nu = 1$. If it is true for $\nu - 1$, we consider

$$1 = \sum_{j=1}^{\nu} \binom{M+\nu}{M+j}(-1)^{j-1}\binom{M+j-1}{M}$$

$$= \sum_{j=1}^{\nu} \binom{M+\nu+1}{M+j+1}(-1)^{j-1}\binom{m+j-1}{M} + (-1)^{\nu-1}\binom{M+\nu-1}{M}$$

$$+ \sum_{j=1}^{\nu} \binom{M+\nu}{M+j+1}\binom{M+j-1}{M}(-1)^{j}$$

$$= -(\nu-1) + \sum_{j=1}^{\nu} \binom{M+\nu+1}{M+j+1}(-1)^{j-1}\binom{M+j-1}{M}$$

or

$$\nu = \sum_{j=1}^{\nu} \binom{M+\nu+1}{M+j+1}(-1)^{j-1}\binom{M+j-1}{M}$$

Now we verify (A5.a)

$$\frac{d}{dx} Q_K^N(0) = -1 + \sum_{j=1}^{K} (-1)^{j-1} \frac{(N-K)!}{(N-K+j)!}$$

$$\left[(1-\alpha)(-1)^{j-1}\binom{N-K+j-1}{j-1}+(-1)^{j-2}\binom{N-K+j-2}{j-2}\right] \geq -1$$

or

$$(1-\alpha)\sum_{j=1}^{k}\left(\frac{1}{N-K+j}\right)-\sum_{j=1}^{K}\frac{(j-1)}{(N-K+j)(N-K+j-1)} \geq 0$$

or

$$-\alpha\sum_{j=N-K+1}^{N}\frac{1}{j}+(N-K)\sum_{j=N-K+1}^{N}\frac{1}{j(j-1)} \geq 0 \tag{A.13}$$

We use the identity: for $A < B$

$$\sum_{j=A}^{B}\frac{1}{j(j-1)}=\sum_{j=A}^{B}\left(\frac{1}{j-1}-\frac{1}{j}\right)=\frac{1}{A-1}-\frac{1}{B} \tag{A.14}$$

So (A.13) becomes

$$-\alpha\sum_{j=N-K+1}^{N}\frac{1}{j}+(N-K)\left(\frac{1}{N-K}-\frac{1}{N}\right) \geq 0$$

$$\frac{K}{N} \geq \left(\sum_{j=N-K+1}^{N}\frac{1}{j}\right)\alpha$$

$$\frac{K}{N} \geq \left(\sum_{j=N-K+1}^{N}\frac{1}{j}\right)\left(1-\left(\frac{K-1}{N}\right)\right)$$

$$0 \geq \sum_{j=N-K+1}^{N}\frac{1}{j}(N-K+1)-K\,.$$

If we replace the right side above by $K+1$, we get

$$\sum_{j=N-K}^{N}\frac{1}{j}(N-K)-K-1$$

$$=\sum_{j=N-K+1}^{N}\frac{1}{j}(N-K+1)-K-\sum_{j=N-K+1}^{N}\frac{1}{j}\,.$$

Thus the right side above is decreasing with K, and we need only verify the inequality for $K = 1$

$$0 \geq \frac{M}{M} - 1 = 0$$

Next we compute:

$$\frac{d}{dx} Q_K^N(1) = -1 + \left(\sum_{\nu=1}^{N-K+1} \frac{1}{\nu} \right)(1 - \alpha)$$

$$+ \sum_{j=2}^{k} \frac{(N-K+1)!(j-2)(-1)^{j-2}}{(N-K+j)!} \left[(1-\alpha)(-1)^{j-1} \binom{N-K+j-1}{j-1} \right.$$

$$\left. + (-1)^{j-2} \binom{N-K+j-2}{j-2} \right] \leq 0$$

Rearranging terms and simplifying gives us:

$$(-\alpha) \left[\sum_{j=1}^{N-K+1} \frac{1}{j} - \sum_{j=2}^{K} \left(\frac{N-K+1}{N-K+j} \right) \frac{1}{j-1} \right] \tag{A.15}$$

$$+ \sum_{j=2}^{K} \frac{(N-K+1)}{(N-K+j-1)(N-K+j)} \leq 1 .$$

. Now

$$\sum_{j=2}^{K} \frac{N-K+1}{(j-1)(N-K+j)} = \sum_{j=2}^{K} \frac{1}{j-1} - \sum_{j=2}^{N} \frac{1}{N-K+j}$$

The first term in (A.15) this becomes

$$(1-\alpha) \sum_{j=K}^{N} \frac{1}{j}$$

Using the identity (A.14), the second term becomes:

$$(N-K+1) \left[\frac{1}{N-K+1} - \frac{1}{N} \right] = 1 - 1 + \frac{K-1}{N} = \frac{K-1}{N}$$

So we have to check:

$$(1-\alpha)\sum_{j=K}^{N}\frac{1}{j}+\frac{K-1}{N}\leq 1$$

$$\frac{K}{N}\sum_{j=K}^{N}\frac{1}{j}+\frac{K-1}{N}\leq 1\,.$$

Again, if we replace K by $K+1$ on the left side above, it increases by $1/N\ \Sigma_{j=K+1}^{N}\ 1/j$. Thus we need only verify:

$$\frac{N}{N}\frac{1}{N}+\frac{N-1}{N}\leq 1$$

$$1\leq 1\,.$$

The last step is to verify that:

$$\frac{d}{dx}\,Q_K^N(0)\leq\frac{d}{dx}\,Q_K^N(1)$$

or

$$-\alpha\sum_{j=N-K+1}^{N}\frac{1}{j}+\frac{K}{N}\leq(1-\alpha)\sum_{j=k}^{N}\frac{1}{j}+\frac{K-1}{N}$$

or

$$\frac{1}{N}\leq(1-\alpha)\sum_{j=K}^{N}\frac{1}{j}+\alpha\sum_{j=N-K+1}^{N}\frac{1}{j}$$

$$0\leq\sum_{j=K}^{N-1}\frac{1}{j}+\alpha\left(\sum_{j=N-K+1}^{N-1}\frac{1}{j}-\sum_{j=K}^{N-1}\frac{1}{j}\right)$$

$$0\leq(1-\alpha)\sum_{j=K}^{N-1}\frac{1}{j}+\alpha\sum_{j=N-K+1}^{N-1}\frac{1}{j}\,.$$

Thus Lemma (5.2) is proven.

BIBLIOGRAPHY

[1] S.R. Chakravarthy and S. Osher, "A new class of High Accuracy Total Variation Diminishing Schemes for Hyperbolic Conservation Laws," *AIAA paper* #85-0363.

[2] P. Colella and P.R. Woodward, "The piecewise-parabolic method (PPM) for gas-dynamical simulations," *J. Comp. Phys.*, v. 54 (1984), 174-201.

[3] B. Engquist and S. Osher, "Stable and entropy condition satisfying approximations for transverse flow calculations," *Math Comp*, v. 34, (1980) pp. 45-75.

[4] S.K. Godunov, "A finite difference method for the numerical computation of discontinuous solutions of the equations to fluid dynamics," Mat. Sb., 47 (1959), pp. 271-290.

[5] A. Harten, "On a class of High Resolution Total-Variation-Stable Finite-Difference Schemes," *SINUM*, v. 21, pp. 1-23 (1984).

[6] A. Harten, "High resolution schemes for hyperbolic conservation laws," *J. Comp. Phys.*, 49 (1983), pp. 357-393.

[7] A. Harten and S. Osher, "Uniformly high-order accurate non-oscillatory schemes, I.," *MRC Technical Summary Report #2823*, May 1985, submitted to SINUM.

[8] S. Osher, "Convergence of Generalized MUSCL Schemes," *NASA Langley Contractor Report 172306*, (1984). SINUM v. 22, (1985), pp. 947-961.

[9] S. Osher, "Riemann solvers, the entropy condition, and difference approximations," *SINUM*, v. 21, (1984), pp. 217-235.

[10] S. Osher and S. Chakravarthy, "Upwind schemes and boundary conditions with applications to Euler equations in general geometries," *J. Comp, Phys* v. 50, (1983) pp. 447-481.

[11] S. Osher and S. Chakravarthy, "Very high order TVD schemes," *ICASE Report #84-44*, (1984), Hampton, VA.

[12] S. Osher and S.R. Chakravarthy, "High-resolution schemes and the entropy condition," *SINUM*, v. 21, (1984), pp. 955-984.

[13] S. Osher and F. Solomon, "Upwind schemes for hyperbolic systems of conservation laws," *Math. Comp.*, v. 38 (1982), pp. 339-377.

[14] S. Osher and E. Tadmor, "On the convergence of difference approximations to conservation laws," submitted to *Math-Comp*.

[15] P.L. Roe, "Approximate Riemann solvers, parameter vectors, and difference schemes," *J. Comp. Phys.*, v. 43 (1981), pp. 357-372.

[16] P.L. Roe, "Some contributions to the modeling of discontinuous flows," in *Lectures in Applied Mathematics*, v. 22, (1985) pp. 163-193.

[17] P.K. Sweby, "High resolution schemes using flux limiters for hyperbolic conservation laws," *SINUM*, v. 21, (1984), pp. 995-1011.

[18] E. Tadmor, "Numerical viscosity and the entropy condition for conservative difference schemes," *NASA Contractor Report 172141*, (1983), *NASA Langley, Math Comp.*, v. 43, (1984).

[19] B. Van Leer, "Towards the ultimate conservative difference scheme IV. A New approach to numerical convection," *J. Comp. Phys.*, 23 (1977), pp. 276-298.

[20] B. van Leer, "Towards the ultimate conservative difference scheme V. A second order sequel to Godunov's method," *J. Comp. Phys.*, v. 32, (1979) pp. 101-136.

[21] B. van Leer, "Flux-vector splitting for the Euler equations," *Proc. 8th International Conference on Numerical Methods in Fluid Dynamics*, Germany, June 28 - July 2, 1982.

[22] M. Ben-Artzi and J. Falcowitz, "An upwind second-order scheme for compressible duct flows," *Siam J. Sci. Comp.*, (1986) to appear.

[23] E. Harabetian, "A convergent series expansion for hyperbolic systems of conservation laws," *NASA Contractor Report 172557, ICASE Report #85-13*, 1985.

[24] S.R. Chakravarthy, A. Harten, and S. Osher, "Essentially non-oscillatory shock-capturing schemes of uniformly very high accuracy," *AIAA 86-0339*, (1986), Reno, NA.

[25] A. Harten, "On high-order accurate interpolation for non-oscillatory shock capturing schemes," *MRC Technical Summary Report #2829*, University of Wisconsin, (1985)

ON NUMERICAL DISPERSION BY UPWIND DIFFERENCING

Bram van Leer
Delft University of Technology
Delft, The Netherlands

ABSTRACT

Upwind-biased difference schemes for the linear one-dimensional convection equation are defined. It is demonstrated that the numerical dispersion caused by such schemes changes sign in the middle of the allowed CFL-number range. This makes it possible to annihilate dispersive errors in two steps.

1. INTRODUCTION

Upwind differencing is a way of differencing convection terms. For the scalar convection equation

$$u_t + au_x = 0, \tag{1}$$

discretized on a uniform grid $\{j\Delta x, n\Delta t\}$, the best-known upwind-difference approximation is the explicit first-order scheme of Courant, Isaacson and Rees (CIR) [1],

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\,\frac{u_j^n - u_{j-1}^n}{\Delta x} = 0, \quad a \geqslant 0, \tag{2.1}$$

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\,\frac{u_{j+1}^n - u_j^n}{\Delta x} = 0, \quad a < 0. \tag{2.2}$$

Introducing the Courant-Friedrichs-Lewy (CFL) number

$$\sigma = a\,\frac{\Delta t}{\Delta x}, \tag{3.1}$$

we may rewrite (2) as

$$u_j^{n+1} = (1 - |\sigma|)u_j^n + |\sigma|u_{j-s}^n, \tag{3.2}$$

where

$$s = \operatorname{sgn}\sigma. \tag{3.3}$$

The scheme is stable, even in the maximum norm, under the CFL condition

$$|\sigma| \leq 1. \tag{4}$$

The value of u_j^{n+1} given in (3.2) may be regarded as an approximation, by linear interpolation, to the value of the exact solution

$$u_j^{n+1} = u(x_j - \sigma\Delta x,\ t^n), \tag{5}$$

which, for non-integer σ, gets lost in the process of discretization. The interpolation at t^n involves only the two nodal points nearest to $x_j - \sigma\Delta x$. Thus, the numerical domain of dependence of u_j^{n+1} is upwind-biased.

The upwind bias becomes more obvious as larger values of the CFL number are allowed. If m is an integer such that

$$m \leq \sigma \leq m + 1, \tag{6.1}$$

a stable upwind scheme is [2]

$$u_j^{n+1} = (m + 1 - \sigma)u_{j-m}^n + (\sigma - m)u_{j-m-1}^n. \tag{6.2}$$

Upwind differencing is often compared to central differencing, where the numerical domain of dependence of u_j^{n+1} at t^n is centered on

x_j, the outcome usually being that upwind differencing is considered superior but more complicated (because of the search implied in (6.1)) and central differencing inferior but simpler (no search needed). Upwind differencing, it is said, stays closer to the physics contained in the convection equation. If this indeed is desirable, one should be able to measure the benefit. That, apparently, is not so easy: to date, very few quantitative theorems have been proven supporting the upwind claim to a higher accuracy.

One piece of evidence can be found in [3] where Fromm's [4] "zero-average phase-error" scheme (an upwind-biased scheme of second-order accuracy) is shown to yield the lowest L_2-error in convecting a step function, in comparison to all other second-order schemes based on the same data. This suggests the use of upwind schemes for shock-propagation problems, an area of application in which these schemes indeed are unrivalled [5].

Another quantitative statement was presented by me without proof in [6]; it concerns the lack of numerical dispersion by upwind schemes at which Fromm already hinted. This will be the subject of the remainder of the paper.

2. AN OPERATIONAL DEFINITION OF UPWINDING

To avoid cluttering up the formulas, I shall restrict the value of the CFL number to the interval [0,1].

Definition. A scheme for Eq. (1) of the general form

$$u_j^{n+1} = \sum_k c_k(\sigma) u_{j+k}^n \tag{7.1}$$

is called upwind-biased for the CFL-number range [0,1] if its coefficients satisfy the symmetry relation

$$c_k(1 - \sigma) = c_{-k-1}(\sigma). \tag{7.2}$$

Eq. (7.2) does not imply consistency of scheme (7.1) with Eq. (1); for this we need to impose two more conditions:

$$\sum_k c_k(\sigma) = 1, \tag{8.1}$$

$$\sum_k k c_k(\sigma) = -\sigma. \tag{8.2}$$

A detailed analysis is needed to find the condition on the coefficients that will ensure stability of the scheme for all values of σ in the range indicated.

It is possible to make scheme (7.1) yield the correct translated initial-value distribution for integer values of σ; this clearly is useful. The additional condition needed is

$$c_k(0) = 0, \qquad k \neq 0. \tag{9}$$

3. NUMERICAL DISPERSION BY UPWIND SCHEMES

When updating the solution with a scheme of the form (7.1), we generally introduce both dispersive and dissipative errors. That is, the Fourier components of the initial-value distribution are convected by the scheme at the wrong speed, while also being damped. Only for integer values of σ these errors can be avoided simultaneously. For non-integer values of σ all consistent stable schemes of the form (7.1) must be dissipative, since they are not invariant under time reversal. With upwind-biased schemes at least the dispersion may be avoided, as shown below.

Lemma. For any scheme that is upwind-biased for the CFL-number range [0,1], the result of one step with CFL number σ followed by a step with CFL number $1 - \sigma$ is free of dispersion.

Proof. Assume initial values according to

$$u_j^n = u_0^n e^{i\alpha j}; \tag{10}$$

any upwind-biased scheme with CFL number $\sigma \in [0,1]$ may then be written as

$$u_j^{n+1} = g(\sigma,\alpha)u_j^n \tag{11.1}$$

with amplification factor

$$g(\sigma,\alpha) = \sum_{k=-K}^{K-1} c_k(\sigma)e^{i\alpha k}, \quad K \geqslant 1. \tag{11.2}$$

The same scheme applied with a CFL number $1 - \sigma$ has an amplification factor

$$g(1-\sigma,\alpha) = \sum_{k=-K}^{K-1} c_k(1-\sigma)e^{i\alpha k}; \tag{12.1}$$

by virtue of (7.2) we have

$$g(1-\sigma,\alpha) = \sum_{k=-K}^{K-1} c_{-k-1}(\sigma)e^{i\alpha k}. \tag{12.2}$$

Introducing $\ell = -k-1$ leads to

$$\begin{aligned} g(1-\sigma,\alpha) &= \sum_{\ell=-K}^{K-1} c_\ell(\sigma)e^{-i\alpha(\ell+1)} \\ &= e^{-i\alpha} \sum_{\ell=-K}^{K-1} c_\ell(\sigma)e^{-i\alpha\ell} \\ &= e^{-i\alpha}g^*(\sigma,\alpha). \end{aligned} \tag{12.3}$$

The composite scheme, with a CFL number of 1, has an amplification factor

$$g(1-\sigma,\alpha)g(\sigma,\alpha) = e^{-i\alpha}g^*(\sigma,\alpha)g(\sigma,\alpha)$$

$$= e^{-i\alpha}|g(\sigma,\alpha)|^2, \tag{13.1}$$

to be compared to the amplification factor for the exact solution at a CFL number of 1:

$$u_j^{n+1} = e^{-i\alpha}u_j^n. \tag{13.2}$$

The two factors are identical in phase. □

The above lemma has an interesting consequence.

Corollary. An upwind-biased scheme for the CFL-number range [0,1] has no dispersion for a CFL number of $\frac{1}{2}$.

Proof. Apply the previous lemma to the case $\sigma = \frac{1}{2}$. Since $\sigma = 1 - \sigma = \frac{1}{2}$, the two steps with the upwind scheme have the same amplification factor

$$g\left(\frac{1}{2},\alpha\right) = e^{-i\alpha/2}\left|g\left(\frac{1}{2},\alpha\right)\right|, \tag{14}$$

with the correct phase $-\alpha/2$. □

A geometric interpretation of this corollary for the CIR scheme is given in Figure 1.

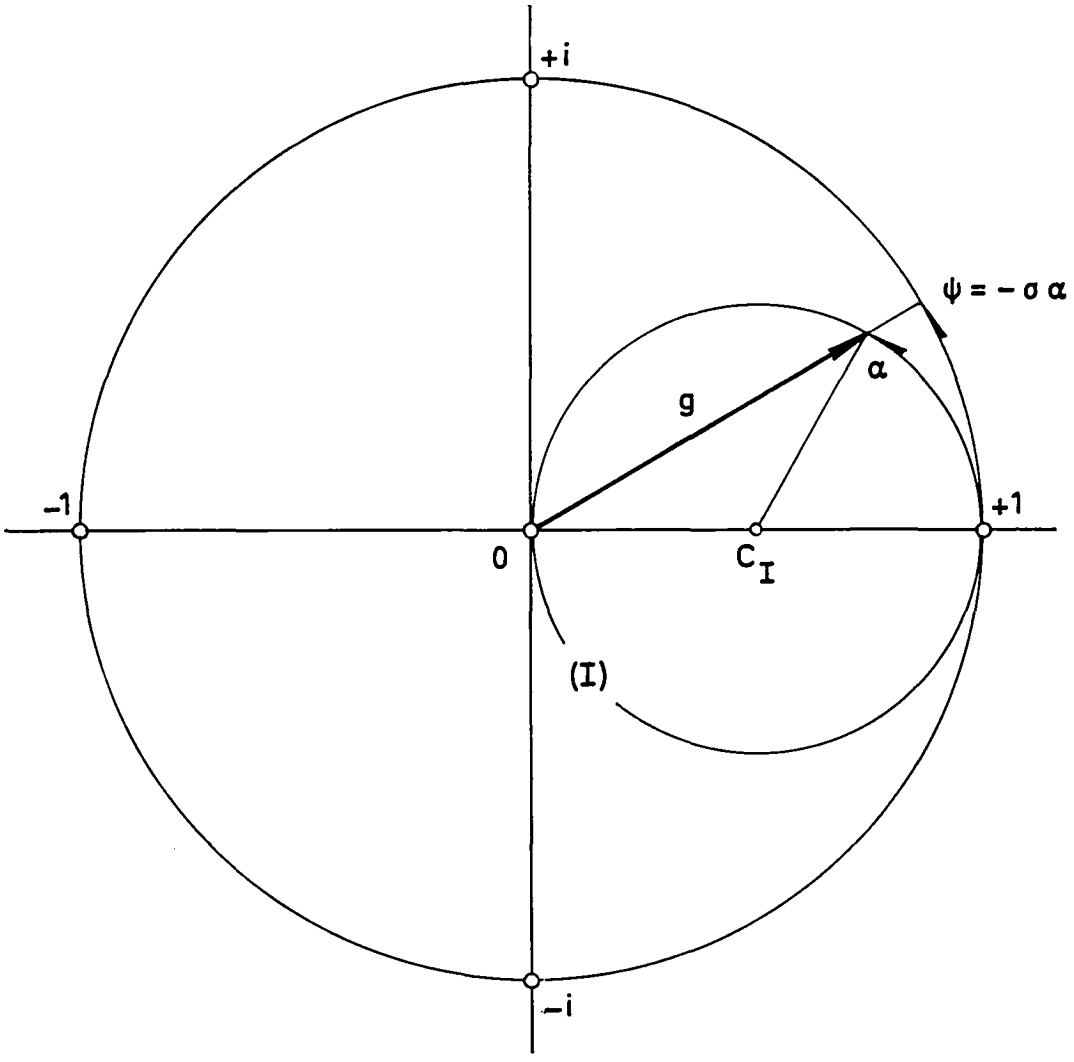


Figure 1. An illustration of the upwind property that arg $g(\sigma,\alpha) = -\sigma\alpha$ for $|\sigma| = \frac{1}{2}$, for the CFL scheme; the drawing is for $\sigma = -\frac{1}{2}$. The locus of $g(\sigma,\alpha)$ is the circle (I) with radius $|\sigma|$ and center C_I in $1 - |\sigma|$ on the real axis; arg $g(\sigma,\alpha)$ is called ψ.

It further follows that for any value of α the dispersive error changes sign when σ passes through 0 (illustrated for the CIR scheme by Figure 2), while the damping factor $|g(\sigma,\alpha)|$ goes through an extremum [6]. For all practical schemes this extremum is an absolute minimum. Thus, in an upwind-biased scheme, minimum dispersion and maximum dissipation go hand in hand. This, again, leads to the representation of moving discontinuities with comparatively little ringing.

Besides upwind-biased schemes for a CFL-number range of the type [m,m+1] there are upwind-biased schemes for the range [m-1,m+1]. These are obtained by shifting the center of a central-difference scheme upwind over m meshes. An example is the fully one-sided, second-order scheme for the CFL-number range [-2,0],

$$u_j^{n+1} = -\frac{\sigma}{2}(1-\sigma)u_{j-2}^n + \sigma(2-\sigma)u_{j-1}^n + \frac{1}{2}(1-\sigma)(2-\sigma)u_j^n. \quad (15)$$

The coefficients of this scheme satisfy the relation

$$c_k(2-\sigma) = c_{-k-2}(\sigma); \quad (16)$$

Accordingly, a step with CFL number σ should be followed by a step with CFL number $2-\sigma$ in order to achieve zero dispersion at a net CFL number of 2.

For central-difference schemes the corresponding relation is

$$c_k(-\sigma) = c_{-k}(\sigma); \quad (17)$$

hence, annihilation of phase errors cannot be combined with a net advancement in time.

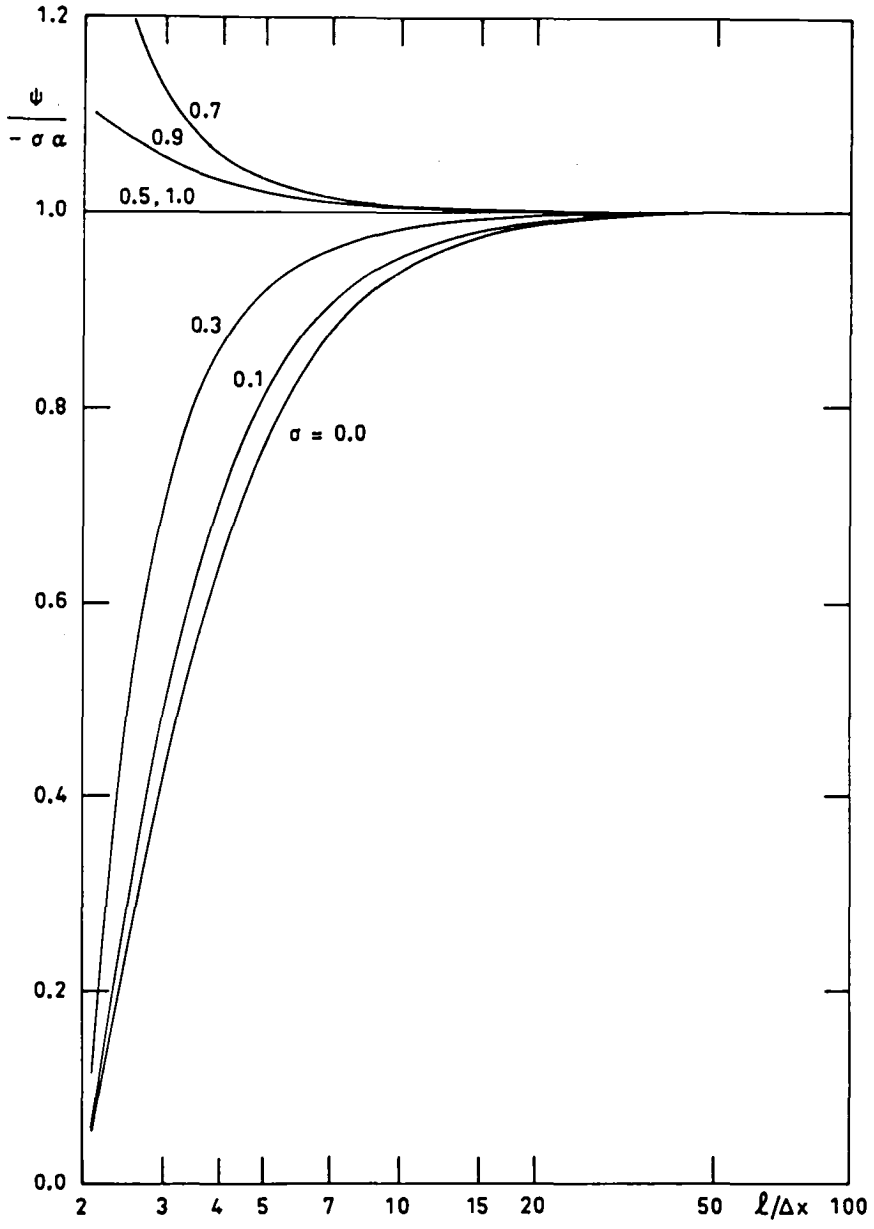


Figure 2. Velocity dispersion versus wavelength for the CFL scheme. The wavelength ℓ is related to α by $\alpha = 2\pi\Delta x/\ell$; the ratio of computed to exact convection speed is evaluated as $\psi/(-\sigma\alpha)$.

REFERENCES

[1] R. Courant, E. Isaacson, and M. Rees, Comm. Pure Appl. Math. 5 (1952), pp. 243-255.

[2] W. L. Miranker, Numer. Math. 17 (1971), pp. 124-142.

[3] P. Wesseling, J. Engrg. Math. 7 (1973), pp. 19-31.

[4] J. E. Fromm, J. Comput. Phys. 3 (1968), pp. 176-189.

[5] P. R. Woodward and P. Colella, J. Comput. Phys. 54 (1984), pp. 115-173.

[6] B. van Leer, J. Comput. Phys. 23 (1977), pp. 276-299.

AZTEC: A FRONT TRACKING CODE BASED ON GODUNOV'S METHOD

BLAIR K. SWARTZ
and
BURTON WENDROFF
Theoretical Division, Group T-7, MS B284
Los Alamos National Laboratory
Los Alamos, NM 87545

ABSTRACT

AZTEC (Adaptive Zoom Tracking - Experimental Code) is a code to solve the one-dimensional gas dynamic equations in a variable area duct with specific implementation for plane, cylindrical, and spherical geometries. The program uses a fixed, locally and adaptively refinable grid, together with a set of moving grid points which migrate through the fixed grid. The moving points represent shocks or contact discontinuities, and they can be created or destroyed, usually as the result of a collision. Mass, energy, and momentum (the last only in the constant area case) are exactly conserved, except after a collision; in that case the conservation error is reduced to invisible levels by spatially localized partial time stepping. The basic difference scheme for both the fixed and moving grid is Godunov's method, with the Riemann solver used to compute both cell boundary fluxes and the speeds of the moving points. Tracking of rarefaction waves on the moving grid is difficult with this method since the waves must be represented as piecewise constant. In one version of AZTEC the rarefaction waves are recorded on the fixed grid with the Lax-Wendroff difference scheme with a small additional viscosity, and most of the numerical experiments have been performed with this version. In another version the polytropic gas equation of state has been replaced by one in which the pressure is a continuous piecewise linear function of specific volume at constant entropy. With this assumption the solution of each Riemann problem is piecewise constant, and our method is exact until the wave structure becomes too complicated. Some preliminary numerical results are exhibited for this version.

* Sponsored by the U. S. Department of Energy under contract W-7405-ENG.36. The publisher recognizes the U. S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U. S. Government purposes.

1. AFTER SOD.

Sod's survey paper [1] was a milestone in the development of numerical methods for one dimensional gas dynamics, for it clearly exposed the shortcomings of some methods which were in vogue at the time. It seems appropriate to point out that two techniques which were not included in the survey are adaptive grid refinement and the method of characteristics. Proper application of the latter requires some form of front tracking, so that the programming of both methods is considerably more complicated than for shock capturing schemes.

While just a modest amount of localized grid refinement will improve a shock capturing method, there are pitfalls. We refer the reader to [2]. The method of characteristics has two interpretations. In the first, the characteristic curves become coordinate lines. Since there are three characteristics for the gas dynamic equations , two of them must be chosen. The natural choices are the u+c and u-c characteristics. In the case of isentropic flow, this means that differencing along the characteristics requires no interpolation. In the non isentropic case values on the third characteristic must be obtained by interpolation.

The second expression of the method of characteristics is a form of upstream differencing. The idea is roughly the following. Write the gas dynamic equations, or any hyperbolic system, in the form $w_t + Aw_x = 0$. Let l_j , $j=1, \cdots, n$, be the left eigenvectors of A , with eigenvalues λ_j . Then

$$l_j (w_t + \lambda_j w_x) = 0. \tag{1.1}$$

This is differenced explicitly, using backward spatial differences for positive λ_j and forward differences for negative λ_j . More precisely,

$$l_j^n (w_i^{n+1} - w_i^n + \mu_j (w_i^n - w_k^n)) = 0 \tag{1.2}$$

where $\mu_j = \lambda_j \Delta x / \Delta t$, and $k = i-1$ if $\mu_j > 0$, $k = i+1$ if $\mu_j < 0$. If there are discontinuities present, they must be tracked through the grid in both versions.

AZTEC combines grid refinement and tracking, using conservative differencing. The tracking is most easily done with Godunov's method, using moving grid points to locate the discontinuities. A condition for the stability of Godunov's method is that the fluxes on the cell boundaries remain constant during a time step. We found that the simplest way to do this in our context was to remove fixed grid points near the moving ones by locally coarsening the spatial grid. This is inaccurate if the moving point is in a region with spatial variation, but we counteract that with a local grid refinement which,

as described later, refines in both space and time. The Riemann solver, which provides the fluxes for the conservative difference equations also determines the speeds of the moving points, as suggested in [3].

In section 2 we give details of the grid refinement procedure. In section 3 we discuss the moving grid. In section 4 we exhibit the result of some computations. In section 5 we present some preliminary results for a piecewise linear equation of state.

2. GRID REFINEMENT.

The one dimensional gas dynamic equations for a variable area duct are

$$(a(x)\rho)_t + (a(x)\rho v)_x = 0 \tag{2.1}$$

$$(a(x)\rho v)_t + (a(x)(\rho v^2 + p))_x = p a_x$$

$$(a(x)\rho E)_t + (a(x)v(\rho E + p))_x = 0,$$

where ρ is the mass density, v is the velocity, $E = e + (1/2)v^2$, e is internal energy, and p is the pressure with equation of state $p = p(\rho, e)$. The quantity $a(x)$ is the area function.

Our program was originally written for slab geometry. It was pointed out to us by J.M.Hyman that an easy way to extend a fixed grid slab code to handle variable area is to introduce area-weighted variables. Thus, we let

$$w = (a(x)\rho, a(x)\rho v, a(x)\rho E)^T$$

so that the equations become

$$w_t + (af(w/a))_x = g. \tag{2.2}$$

where f is the flux vector given by

$$f = (\rho v, \rho v^2 + p, v(\rho E + p))^T$$

and

$$g = (0, pa_x, 0)^T$$

Suppose that we have a uniform grid of N cells indexed by i. The quantity w_i^n will be the average of w in the cell at time n. x_i is the coordinate of the cell center; $x_{i+1/2}$ is the coordinate of the interface between cell i and $i+1$. The area at a cell edge is

$$a_{i+1/2} = a(x_{i+1/2}).$$

but the area of the cell center is defined by

$$a_i = a(x_{i-1/2}, x_{i+1/2})$$

where

$$a(x,y) = \left| \int_x^y a(s)ds \; (y-x)^{-1} \right|$$

and

$$a(x,x) = a(x).$$

The basic conservative difference equation is

$$\Delta x \; w_i^{n+1} = \Delta x \; w_i^n - \Delta t \, [(af)_{i+1/2} - (af)_{i-1/2}] + g \, \Delta x \, \Delta t. \tag{2.3}$$

If the cell interface with index $i+1/2$ is *internal*, that is, if cells i and $i+1$ are both present on the grid, we allow two possible definitions of the numerical flux $f_{i+1/2}$.

The *Godunov* flux is obtained by solving the Riemann problem centered at $x = 0$ with left state given by w_i/a_i and right state given by w_{i+1}/a_{i+1}. The flux function evaluated at $x = 0, t > 0$ is then used for $f_{i+1/2}$.

The *Lax–Friedrichs* flux is defined as follows. Set

$$w_{i+1/2} = .5[w_{i+1} + w_i - \Delta t / \Delta x \, (f_{i+1} - f_i) a(x_{i+1/2})],$$

and then

$$f_{i+1/2} = f(w_{i+1/2}/a_{i+1/2}).$$

In the uniform area case if the fluxes at both cell boundaries are Lax-Friedrichs fluxes then w_i^{n+1} becomes the two-step Lax-Wendroff scheme.

The choice of fluxes is part of the experimentation with AZTEC. However, an invariable strategy that we have implemented is to always use Godunov fluxes on the finer grids (if they exist), at the external boundaries, and at cells in a neighborhood of a moving grid point. If the Lax-Friedrichs flux is used at all on the coarse grid, it is in an expansion region.

The grids are defined in terms of cells rather than points. The symbol j will always identify a grid level, $j = 1,2, \cdots J$. The maximum number of grid levels, J, is an input parameter. Level 1 represents the coarsest grid, with $N(1) = N$ cells each of length $\Delta x(1) = \Delta x$. Level 2 is a refinement of level 1 obtained by dividing each cell of

level 1 in half, so that $\Delta x(2) = .5\Delta x(1)$, and $N(2) = 2N(1)$. Thus, $\Delta x(j) = 2^{-(j-1)}\Delta x$, and $N(j) = 2^{j-1}N$.

Since the refinement is local and adaptive, not all cells on every level will be advanced at every time step. There are two kinds of cells, *live* and *dead*. At the start of a time step the level 1 cells are all live. For $j > 1$, a cell on level j will be live only if its parent cell on level $j-1$ is live and if certain tests of the state variables on level $j-1$ indicate that refinement (splitting) of the parent cell is required.

There will be two kinds of live cells, *sterile* and *fertile*. A sterile cell is one which is not to be split and which therefore must be advanced by the difference equation. A fertile cell is one which splits into two daughter cells on the next level and which is therefore *not* advanced by the differential equation. The advancement of a sterile cell requires the computation of fluxes at the cell boundaries, but computation of the flux at a fertile cell boundary will be needed only if that cell is not contiguous at that boundary to a fertile cell on the same level. Since AZTEC is designed for serial computation we have tried to avoid redundant calculation of fluxes. This arrangement, which is not quite as complicated as it sounds, is shown schematically in Figure 2.1.

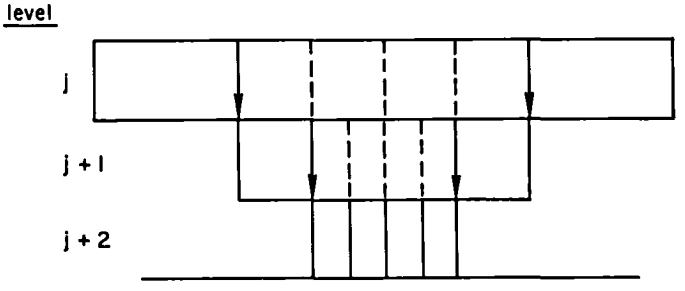


Fig. 2.1. Boundary fluxes at grid interfaces.

The boxes represent cells on the indicated grid level. The vertical sides of the boxes represent the cell boundaries. A dotted line means that the flux is not computed on that grid level. The flux is computed at a solid line. A solid line with an arrow means that the flux computed at that grid is used at the next grid level. Thus, at an interface between grids j and $j+1$, the coarse grid flux supplies the boundary condition for the finer grid. This generalization of [4] enables us to maintain conservation.

Since our difference scheme is explicit, the time step for level j must be half that for level $j-1$. An example of the evolution of the space-time grid is shown in Figure 2.2.

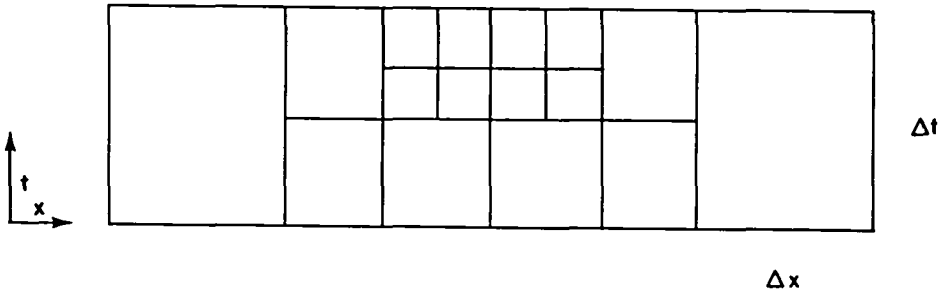


Fig. 2.2. Space-time refinement.

Note that refinement has occurred between the coarse time steps. Here is the algorithm for setting up and advancing the grids.

BEGIN ALGORITHM

j = 1

1 ic(j) = 0 **ic(j) is 0 for the first pass through level j, 1 for the second pass**

2 call CREATE(j) **Determine and label the fertile and sterile cells on level j, label and provide data for the live cells on level j+1. Set nc(j+1) = number of live cells on level j+1.**

call FLUX(j) **Compute and store fluxes on level j at those interfaces which are a boundary of at least one sterile cell.**

call ADVANCE(j) ** Compute w^{n+1} on level j and overwrite on w^n, for sterile cells only**

if j < J and nc(j+1) ≠ 0

then j = j+1 and go to 1

3 else if ic(j) = 0

then if j = 1

then step finished

else ic(j) = 1

go to 2

else j = j-1

call CONST(j) * *The total conserved quantity (mass, for example) in a fertile parent cell is defined to be the sum of the conserved quantities of the two daughter cells.**

go to 3

END ALGORITHM

The subroutine CREATE requires further discussion. First, it must determine which cells on level j are to be split. This is done by performing two tests for each cell. If l is the cell index, then one of the tests looks for moving grid points in the cells $l-2, l-1, l, l+1, l+2$ (see section 3). If there are any, then cell l splits into two. Of course, special provisions have to be made for cells close to the boundaries. The second test splits cell l if there is a compression in the same neighborhood as above: other criteria could be included. Now, suppose that in advancing from t to $t + \Delta t$ CREATE finds that a cell on level j must be split into two daughter cells on level $j+1$. There are two possibilities: the daughter cells were present at the previous time step and were advanced to time t by the algorithm, or they were not. In the former case no new data need be created for the daughters. In the latter case data is obtained by interpolation. If the parent cell on level j has index i, the interpolation is as follows. Let L and R be the indices of the left and right daughter cells, respectively, and let

$$w_L = 1.25w_i - .25(a_i/a_{i+1})w_{i+1}.$$

$$w_R = .75w_i + .25(a_i/a_{i+1})w_{i+1}$$

If both w_L/a_L and w_R/a_R lie between w_{i-1}/a_{i-1} and w_{i+1}/a_{i+1}, accept w_L and w_R as the interpolated values. If not, let

$$w_L = .75w_i + .25(a_i/a_{i-1})w_{i-1}$$

$$w_R = 1.25w_i - .25(a_i/a_{i-1})w_{i-1}.$$

Use these as the interpolated values unless the above monotonicity test fails, in which case set

$$w_L = (a_L/a_i)w_i.$$

$$w_R = (a_R/a_i)w_i.$$

The latter is also used if cell i is at the boundary of the physical domain.

3. THE MOVING GRID.

The moving grid points will move through the fixed grid and exchange data with the fixed cells. We have chosen to do this for the finest grid only: this is arranged by having one of the refinement tests look for moving points in the two cells on each side of the current cell. If the points are not allowed to move more than the length of one cell in one time step, they cannot leave the fine grid.

The moving gridpoints define boundaries of skewed space time cells in which the conservation laws are applied just as they were for the fixed grid. For the two points $x < y$ shown in Figure 3.1, the difference equation is

$$[y-x+(\sigma_y-\sigma_x)\Delta t\,]\hat{w}_{xy} = (y-x)w_{xy} - \Delta t\,[a(y,y+\sigma_y\Delta t)F_y - a(x,x+\sigma_x\Delta t)F_x] + \bar{g}\,\Delta t\ . \tag{3.1}$$

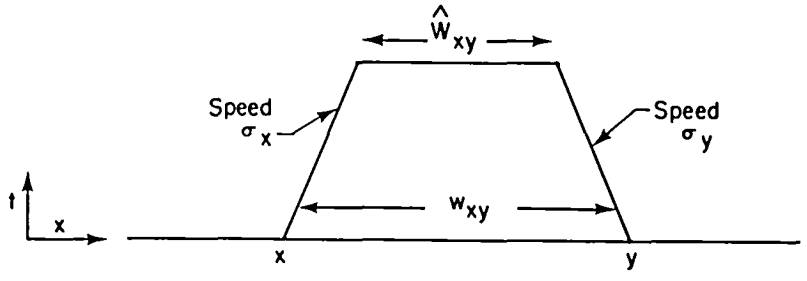


Fig. 3.1. Space-time cell defined by moving points.

The quantity w_{xy} is both the right value for the discontinuity at x and the left value for the one at y. The term $\bar{g}\,\Delta t$ only appears in the momentum equation. To avoid false accelerations it must have the following form. If, in the momentum equation,

$$F_y = \rho v_y^2 + p_y\,,$$

and similarly for F_x, then

$$\bar{g} = (1/2)(p_x + p_y)[a(y,y+\sigma_y\Delta t) - a(x,x+\sigma_x\Delta t)].$$

There are two things that must be provided in the basic difference equation above: the speeds σ and the fluxes F. These are obtained from the solution of Riemann problems. In order to do this we must first recover the hydrodynamic variables from the area weighted variables. Suppose, for the moment, that we have done this properly. Then when the grid point at position x is to be moved there will be associated with it a left state u_- and a right state u_+. We find the complete solution of the Riemann

problem for these two states. Then we decide which ray or rays are to be followed. For example, if the point x is a contact discontinuity and if the solution of the Riemann problem has a sufficiently strong contact discontinuity, then we take the new speed to be the speed of the contact. The new flux F is $f - \sigma u$ evaluated on this ray (this takes account of the fact that the ray is not necessarily vertical in the space-time plane). Note that $f - \sigma u$ is continuous across every ray in the solution of the Riemann problem. More generally, the point x might spawn several new moving points. If x is the result of a collision with another point or with a reflecting boundary, then we could follow all the shocks and contacts which emerge.

The complete logic of the procedure for deciding which rays to keep is too complicated to give in complete detail here, but we can give an outline of it. First, the Riemann solver produces a list of speeds and fluxes and identifiers for each sufficiently strong wave which is present in the solution. Thus, a shock corresponding to the characteristic $v+c$ is identified as a 3-shock, and a speed and flux are given for it. A rarefaction corresponding to the characteristic $v-c$ is identified as a 1-wave, and for it the speeds and fluxes on the leading and trailing edges are provided. Next, tactical decisions are made in a subroutine called TRACK, which has the job of creating and destroying moving points, advancing the moving points and checking for collisions, maintaining stability on the moving grid, and communicating with the most refined portions of the fixed grid.

Here is how TRACK works. First, the points are collected into blocks. Each block is such that the rightmost point of one block is separated by five or more full fixed cells from the leftmost point of the next block, as in Figure 3.2.

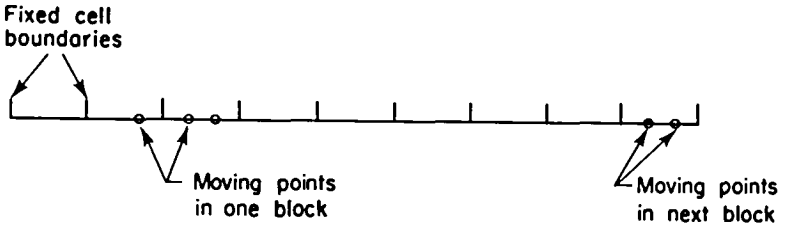


Fig. 3.2. Blocks of moving points.

Each block is processed independently of the others, so what follows refers to the points of one block. In order to improve resolution in the variable area case, if two

adjacent points are two or more full cells apart some fixed grid points between the pair are treated as moving. These are called separator waves. Next, each moving point is provided with a left and right average value of the area-weighted variable w so that if $w_-(x)$ and $w_+(x)$ are respectively the left and right states of x, and if x and y are adjacent points ($x < y$) then $w_+(x) = w_-(y)$. This is done using a combination of fixed cell data and moving point data obtained from eq. (3.1) for the previous time step, depending on the separation of the points. Of course, this is done conservatively. The hydrodynamic variable corresponding to $w_+(x)$ is $u_+(x) := w_+(x)/a(x,y)$. Now the Riemann solver is called for each point in the block. For the typical grid point all the rays returned by the solver are assumed to define new points which are inserted into the list of moving points. There are exceptions to this; for example, at a left reflecting boundary only rays with non-negative speeds are retained. The list is ordered by position if the positions are unequal and by speed otherwise, as in Figure 3.3.

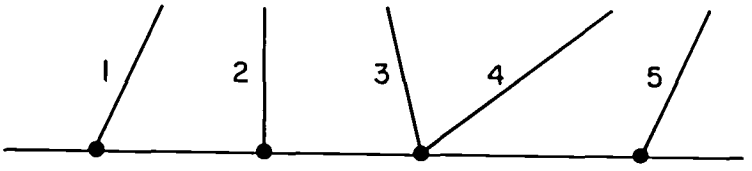


Fig. 3.3. Ordering of moving points.

At this stage we have many more points than we want or need, but most of them will be deleted at the end of the time step.

The reason for retaining so much information is that this gives us a procedure for maintaining stability during a collision or close approach of moving points. If a collision occurs at time $t + \delta t$, $0 < \delta t < \Delta t$, the current block of points is advanced to $t + \delta t$ using eq. (3.1) with Δt replaced by δt. Then we attempt to finish the time step by advancing from $t + \delta t$ to $t + \Delta t$, checking again for a collision, etc. The use of blocks causes this partial time-stepping to be spatially localized, unless the moving grid is evenly distributed in the fixed grid. The idea now is that any collision which occurs at this time is "exact," which means the following. If the points x and y in Figure 3.1

collide at time $t + \Delta t$, then $x + \sigma_x \Delta t = y + \sigma_y \Delta t$, so that the left side of eq.(3.1) is zero. On the other hand, even if the source term were not present, the right side of that equation cannot be expected to be zero. Indeed, consider the case shown in Figure 3.4.

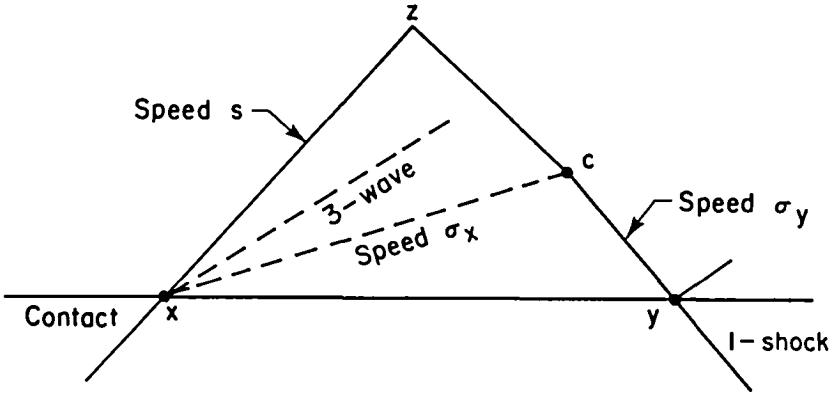


Fig. 3.4. Exact collision.

In this situation, a contact and a 1-shock have arrived at x and y, respectively. The Riemann solver has produced at x a contact with speed s and a 3-rarefaction wave, while at y the solution is a 1-shock and some other waves that play no role. Two bad things happen if we suppress the rarefaction wave. First, because the solution is not constant along the ray yz we can expect an instability to develop. Second, making the appropriate substitutions into the right side of eq.(3.1), we have

$$R_1 := (y - x)a(x,y)u_0 - \delta t\,[a(x,z)su_1 - a(y,z)\sigma_y u_0]$$

$$-\delta t\,[a(y,z)f(u_0) - a(x,z)f(u_1)].$$

Even if the area factors were constant, this would not be zero unless $u_1 = u_0$. On the other hand, if we include the leading edge of the rarefaction in the list of moving points, then the first collision occurs at c. Then the right side of eq.(3.1) becomes

$$R_2 := \bar{\delta}tf(u_0)[a(y,y+\sigma_y\bar{\delta}t) - a(x,x+\sigma_x\bar{\delta}t)].$$

In the constant area case, $R_2 = 0$, hence the nomenclature exact. In other words, an exact collision is one in which the state between the two intersecting rays is constant. For such a collision eq.(3.1) is identically correct if the area is constant. When the area is variable, the error in the mass and energy conservation is second order in the mesh size.

In Figure 3.5 we can see how the collision between the contact and the shock will actually occur.

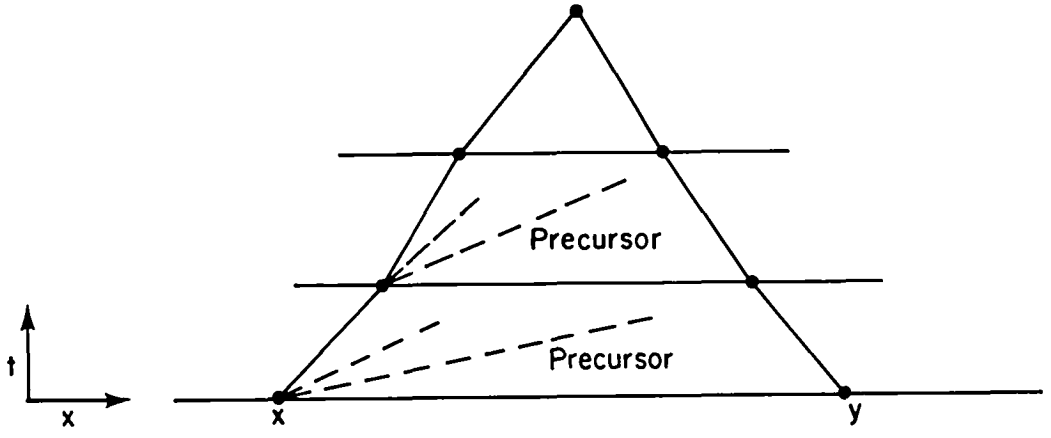


Fig. 3.5. Collision and precursors.

After several partial time steps, caused by collision of the precursor rarefaction wave with the shock, the rarefaction wave will become too weak to be seen by the Riemann solver and the main collision will take place. There will be a small error in a conserved variable such as the mass. The program controls this error by two devices. If the error exceeds a pre-set value the time step is repeated with a smaller strength threshold in the Riemann solver. This works well for constant area, but is not enough in other geometries. For them, we must force additional partial time steps that will reduce R_2.

At the end of the time step (partial or complete) the precursor waves are deleted. At the end of a full time step the separator waves are also removed. Thus each step starts fresh with the main moving points. However, if a major collision has occurred, points may have been created or destroyed. If we wish to keep track of all shocks and contacts, then we must include the resulting transmitted and reflected shocks and residual contact produced by a collision of two shocks or of a shock and a contact. The entire process that we have described works remarkably well, particularly if collisions are rare.

4. THE TEST PROBLEMS.

Three test problems are presented. The first is Sod's problem with reflecting boundaries [1]. The second is a problem posed by Paul Woodward [5] involving the interaction of the solution of two Riemann problems. The third is an elegant spherical shock

problem with a simple exact solution due to Bill Noh [6].

In Figure 4.1 we give our solution (density only) of Sod's problem at t =.175. The initial data define a Riemann problem centered at $x = .5$. The left state has density 1.0, pressure 1.0, and velocity 0.0. The right state has density .125, pressure .1, and velocity 0.0. The equation of state is that for a γ - law gas with $\gamma = 1.4$. This initial-value problem resolves into a rarefaction wave, a contact discontinuity, and a shock wave (from left to right).

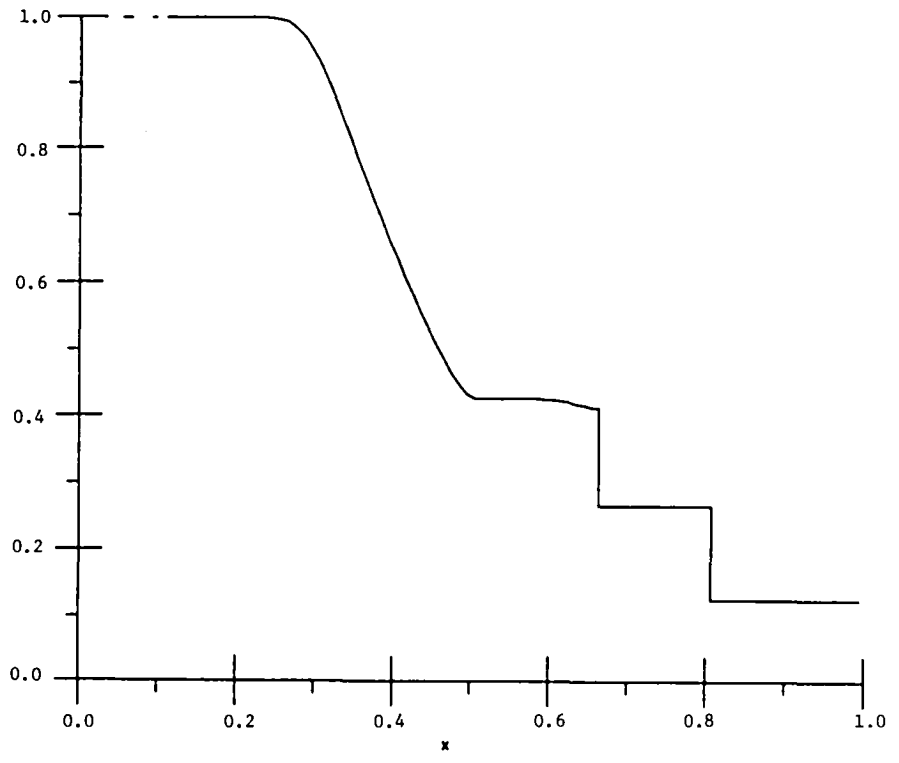


Fig. 4.1 Sod's problem at $t = .175$.

The shock is correct, but the state between the contact and the rarefaction is in error by 5%. This is caused by the presence of the strong rarefaction in close proximity to the contact early in the calculation.

In Figure 4.2 we give the apparently converged computed solution at t = .81. By this time the main shock has reflected off the right boundary and interacted with the contact, producing reflected and transmitted shocks. The rarefaction has reflected off

the left boundary, begun to emerge from the interaction with its image, and is just now beginning to interact with the main pair of reflected/transmitted shocks.

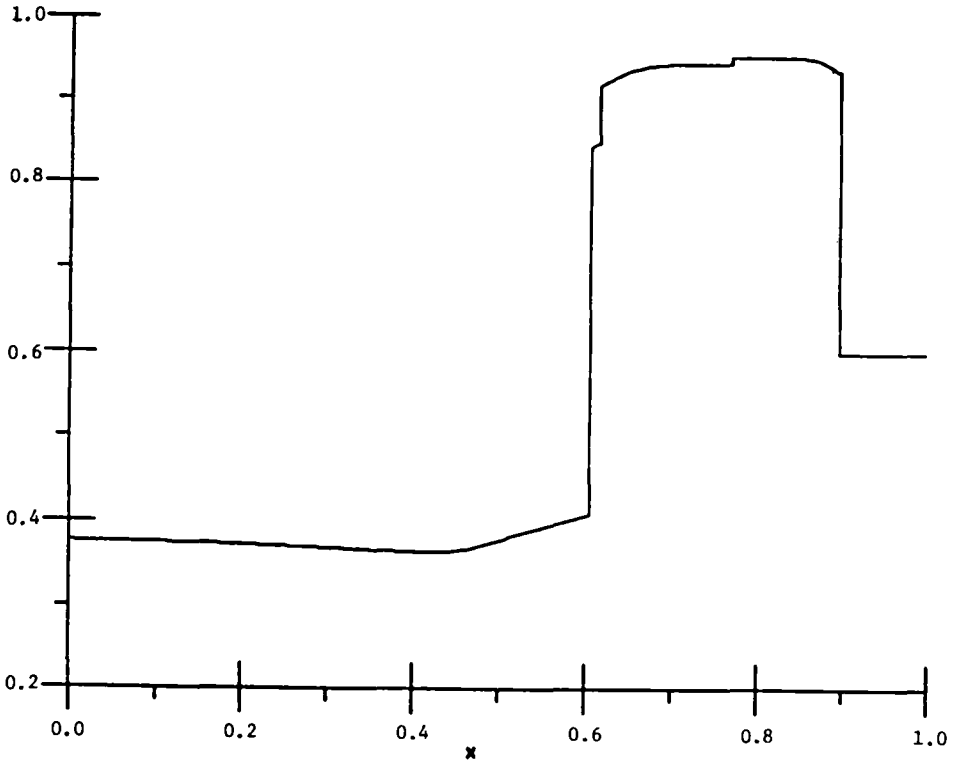


Fig. 4.2 Sod's problem at $t = .81$.

The initial conditions for Woodward's problem are a gas at rest with unit density in a unit interval with reflecting walls. The pressure in the left-most 1/10-th of the interval is 1000 and the pressure in the right-most 1/10-th is 100; it is .01 otherwise. The initial rarefaction waves moving toward the boundaries reflect and quickly catch up to the contacts and the shocks. The collision of the shocks and their trailing waves at about $t = .028$ initiates a complex sequence of intense interactions localized within five to twenty percent of the interval. The computed density is shown in Figure 4.3 at $t = .038$. Woodward has computed this with a very fine grid, but he only gives a graph of the solution. We differ from his solution only in the magnitude of the peak density, which he finds to be 6.5 while ours is 7. Both calculations locate the

discontinuities in the same places.

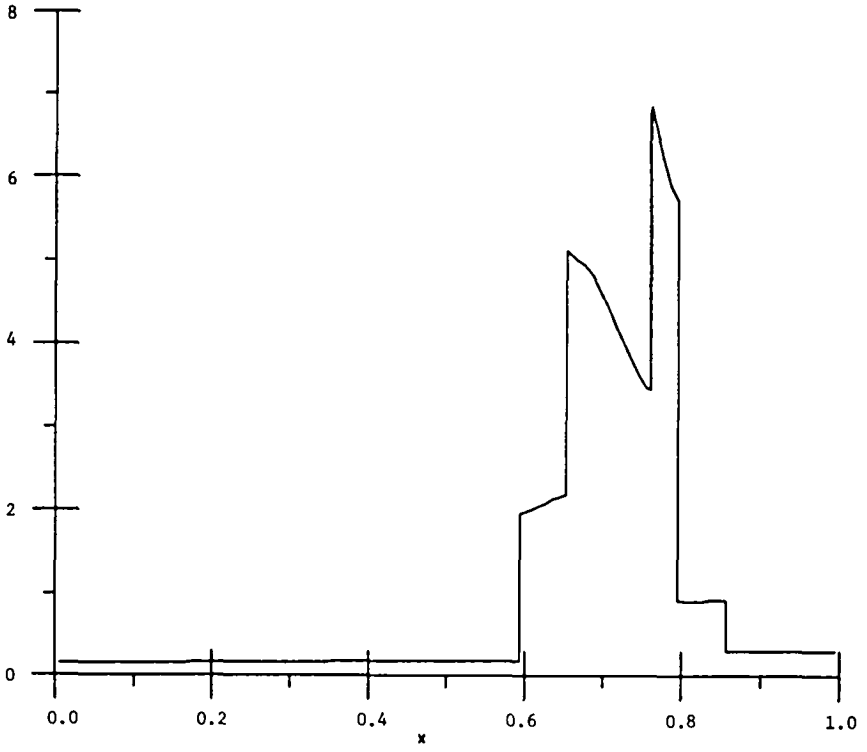


Fig. 4.3. Woodward's problem at $t = .038$.

For Noh's problem we have a sphere of unit radius filled with a γ-law gas, $\gamma = 5/3$, at zero pressure and internal energy, and with velocity = -1. At t = .6 the solution consists of a shock located at x = .2 moving with speed 1/3. Behind the shock the pressure is 64/3 and the density is 64. Ahead of the shock the density is $1 + t/r^2$. The computed density is given in Figure 4.4.

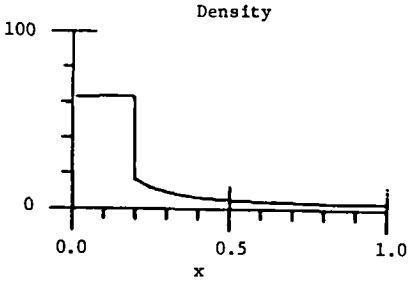


Fig. 4.4. Noh's problem.

5. THE PIECEWISE LINEAR EQUATION OF STATE.

The approximation of arbitrary functions by piecewise linear ones has a long and distinguished history. The value of this approximation in the theory of conservation laws seems to have been first recognized by Dafermos [7], who combined a piecewise linear flux function and piecewise constant initial data to obtain an elegant existence theorem for scalar conservation laws. The crucial property of the piecewise linear scalar flux is that the solution of the Riemann problem has only constant states. Hedstrom [8] observed that if the pressure expressed as a function of specific volume and entropy is piecewise linear in the volume, then again the solution of the Riemann problem has only constant states. Hedstrom used this as a computational device to obtain numerical solutions of the equations of isentropic flow, by tracking the shock-like boundaries of the constant states. In principle, AZTEC can obtain the exact solution of the full gas dynamic equations with such a piecewise linear pressure and piecewise constant initial data simply by having no fixed grid points, only moving ones. If we also take a very large time step, then the collisions determine the intermediate time steps. Each collision will be exact in the sense defined in section 3.

In Figure 5.1 we show the solution of Sod's problem for a piecewise linear approximation to the γ - law gas.

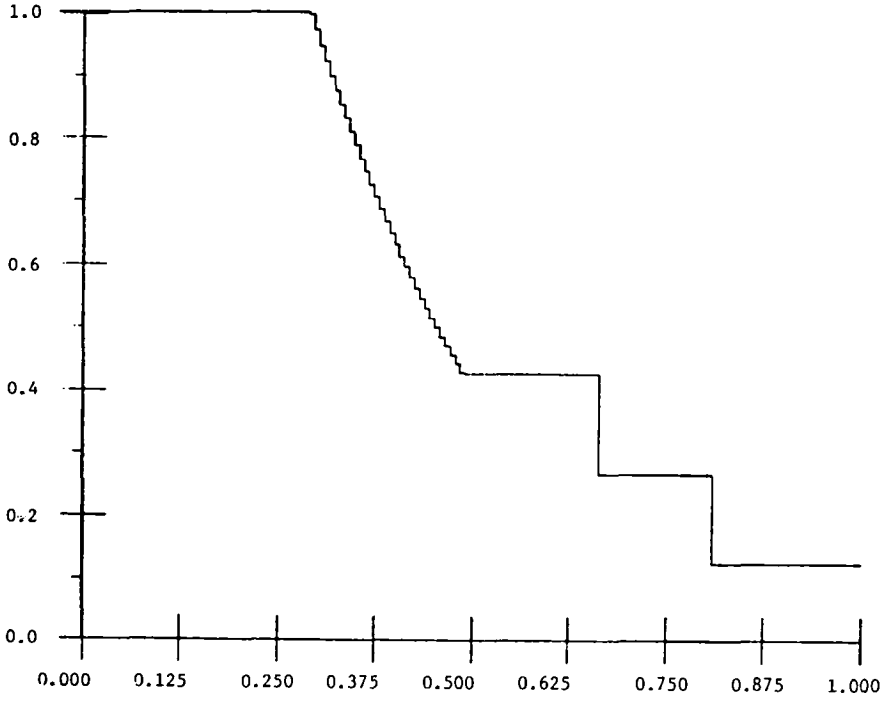


Fig. 5.1. Sod's problem for a piecewise linear equation of state, $t = .175$.

There are 80 nodes per decade in density. The shock and contact are now exact (for the given equation of state), and the rarefaction wave has become piecewise constant. Of course, this is a trivial application of the method as there have been no collisions.

In Figure 5.2 we have a more interesting example, namely, Sod's problem with reflecting boundaries at t=.81, computed with 80 nodes per decade in density. Now we have a rarefaction wave reflected off the left boundary and interacting with the waves reflected from the other boundary. This result should be compared with Fig. 4.2. The solution in Fig. 5.2 was obtained about 30 times faster than the one in Fig. 4.2.

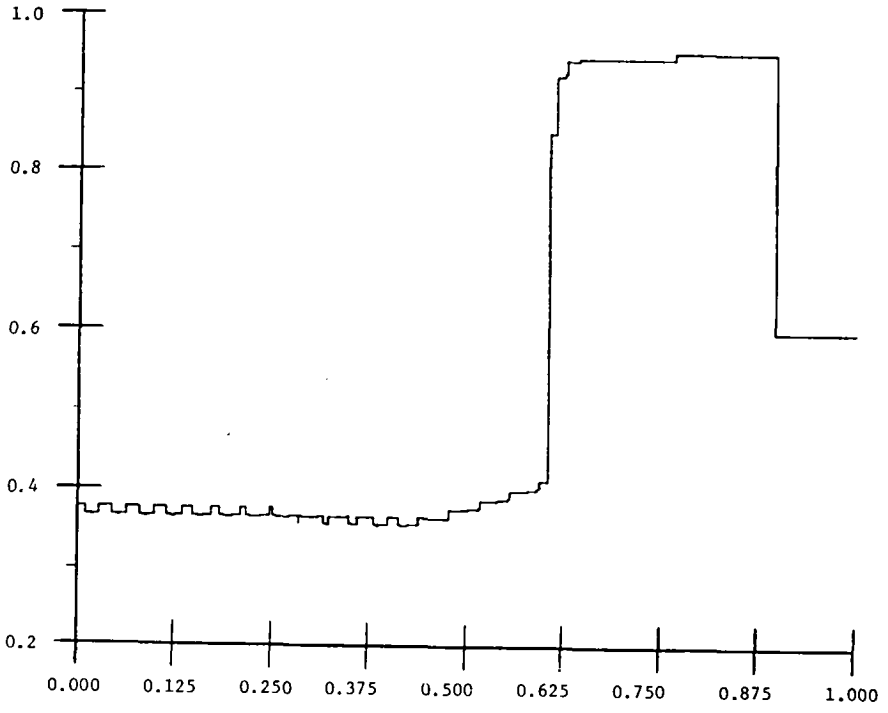


Fig. 5.2. Sod's problem, $t = .81$, piecewise linear equation of state.

There are serious difficulties with the piecewise linear method which seem to prevent it from being more than a curiosity. One problem is that, in general, a collision of two waves will produce at least three outgoing waves, leading to a rapid proliferation of waves and collisions. This can even happen with the interaction of rarefaction waves, such as occurs in the reflecting Sod problem. In the case of the interaction of rarefaction waves, this difficulty is overcome by the following device. Suppose the pressure p at fixed entropy is continuous, and that it is linear in the intervals (τ_i, τ_{i+1}), $i = 1,2 \cdots ,n$, where $\tau = 1/\rho$. We constrain the nodal values p_i to satisfy the condition that $(p_i - p_{i+1})(\tau_{i+1} - \tau_i)$ is a constant that depends only on the entropy, not i. It

follows from this that the velocity jump across each internal ray in a rarefaction fan is constant. Moreover, an examination (as in [8]) of the rarefaction curves in the pressure-velocity space shows that the number of collisions in the complete interaction of two rarefaction fans is now of order n^2 if there are n nodal pressure values in the fans. This constraint was used to generate Figures 5.1 and 5.2.

We anticipate reporting additional detail on our experiences with piecewise linear equations of state.

REFERENCES.

[1] G. Sod, "A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws," J. Comp. Phys., Vol. 27 (1978), pp. 1-31.

[2] B.K. Swartz, "Courant-like conditions limit reasonable mesh refinement to order h^2," Los Alamos National Laboratory Preprint LA-UR-81-2037, 1981.

[3] S.K. Godunov, A.V. Zabrodin, and G.P. Prokopov, "A computational scheme for two-dimensional non stationary problems of gas dynamics and calculation of the flow from a shock wave approaching a stationary state," USSR Comp. Math. Math. Phys., Vol. 1 (1962), pp. 1187-1219.

[4] G. Browning, H.-O. Kreiss, and J. Oliger, "Mesh refinement," Math. Comp., Vol. 27 (1973), pp. 29-39.

[5] P.R. Woodward, "Trade-offs in designing explicit hydrodynamic schemes for vector computers," Livermore National Laboratory Preprint UCRL-85813, 1981.

[6] W.F. Noh, Artificial viscosity (Q) and artificial heat flux (H) errors for spherically divergent shocks, Lawrence Livermore National Laboratory Preprint UCRL-89623, 1983.

[7] C.M. Dafermos, "Polygonal approximations of solutions of the initial-value problem for a conservation law," J. Math. Anal. Appl., Vol. 38 (1972), pp. 640-658.

[8] G.W. Hedstrom, "Some numerical experiments with Dafermos's method for nonlinear hyperbolic equations," Lecture Notes in Math., Vol. 267 (1972), Springer, Berlin.

LEAST SQUARES FINITE ELEMENT SIMULATION OF TRANSONIC FLOWS

T. F. Chen
Carnegie-Mellon University

G. J. Fix
Carnegie-Mellon University

ABSTRACT

Finite difference approximation of transonic flow problems is a well-developed and largely successful approach. Nevertheless, there is still a real need to develop finite element methods for applications arising from fluid-structure interactions and problems with complicated boundaries. In this paper we introduce a least squares based finite element scheme. It is shown that, if suitably formulated, such an approach can lead to physically meaningful results. Bottlenecks that arise from such schemes are also discussed.

Research was supported in part by the National Aeronautics and Space Administration under NASA Contract Nos. NAS1-17070 and NAS1-18107 while the second author was in residence at the Institute for Computer Applications in Science and Engineering, NASA Lanley Research Center, Hampton, VA 23665-5225. Partial support was also provided by Army Research Office under Contract No. DAAG29-83-K0084.

1. INTRODUCTION

In this paper we consider the approximation of transonic flows by finite element methods based on a variational method of the least squares type. The objective here is purely computational. In particular, we have sought to fully exploit the ideas arising from mathematical analysis of such methods (see, for example, [1] - [6]) and directly apply them to a nontrivial transonic flow problem. The major conclusion drawn from this work is that finite element methods--suitably formulated--can give physically meaningful results.

There is a significant and largely successful array of finite difference techniques for transonic flows (e.g., [17]). Nevertheless, an assumption implicit in this work is that there is still a need for stable and accurate finite element approaches. First, there are applications from fluid-structure interactions that would benefit from the availability of a finite element flow model. Second, there is the issue of complicated boundaries in the flow field. The importance of the finite element ideas in such a context--while largely untested--is still promising.

Variational principles of the least squares types have a number of valuable computational properties. For example, the algebraic system generated is always Hermitian semidefinite. In addition, such schemes, if properly formulated, are insensitive to equation type, be it hyperbolic (supersonic flows) or elliptic (subsonic flows). In fact, the majority of the finite element ideas that have been used for hyperbolic problems to date tend to be either implicitly or explicitly of the least squares type.

Least squares based schemes do have, however, some major computational defects. First, they tend to be sensitive to singularities and

discontinuities in the flow variables. Moreover, mesh refinement alone does not overcome these defects [7]. Based on the work in [7] we introduce weighted least squares variational principles, which in combination with mesh refinement is capable of dealing with shocks in the flow field.

In Section 2 we describe the basic numerical formulation, and outline the essential computational properties associated with the approach. A key feature is the proper choice of weighting functions to use in the least squares functional. A closely allied issue is the density modifications needed to rule out nonphysical expansion shocks.

In Section 3 we present sample numerical results. As a model problem we select the planar potential flow over a cylinder.

Other authors have considered finite element approximation of transonic flows. Selected references are [18] - [21].

2. THE LEAST SQUARES FORMULATION

We consider the potential flow over a body $\hat{\Omega}$. Let $\underline{u}$ denote the velocity and ρ the density. Then a mass balance yields

$$\operatorname{div}[\rho \underline{u}] = 0. \tag{2.1}$$

In addition, we have

$$\underline{u} = \operatorname{grad} \phi \tag{2.2}$$

for the velocity potential ϕ. The density ρ is given as a function of $\underline{u}$ by the Bernoulli equation. The system is closed by specifying the normal velocity

$$\underline{u}\cdot\underline{n} = v \tag{2.3}$$

at the boundaries of the flow region. On the body $\hat{\Omega}$ the no flow condition

$$\underline{u}\cdot\underline{n} = 0$$

applies. We assume that the flow region is contained in a box B and that (2.3) is specified on the boundary of B. Thus

$$\Omega = B/\hat{\Omega} \tag{2.4}$$

defines the flow region, and (2.1) - (2.2) hold in Ω with (2.3) holding on the boundary Γ and Ω.

Since the flow is assumed to be irrotational, (2.1) - (2.2) can be replaced with

$$\text{div}(\rho\underline{u}) = 0 \quad \text{in} \quad \Omega \tag{2.5}$$

$$\text{curl}(\underline{u}) = 0 \quad \text{in} \quad \Omega \tag{2.6}$$

$$u\cdot n = v \quad \text{on} \quad \Gamma. \tag{2.7}$$

A least squares scheme based on this system takes the form

$$\int_{\Omega} \{|\text{div}(\rho\underline{u})|^2 + |\text{curl}(\underline{u})|^2\} = \min, \tag{2.8}$$

where the variation is taken $\underline{u}$ in some finite element space satisfying the

boundary conditions (2.7). Such a div - curl system has proven to be very effective for elliptic systems (subsonic flows) in cases where the density $\rho = \rho(\underline{u})$ and the velocity field $\underline{u}$ are smooth [8].

Preliminary results indicate that with appropriate weighting functions on the terms in (2.8), the nonsmooth cases can be treated as well. Nevertheless, in this paper we shall focus attention on (2.1) - (2.2) and least squares schemes of the form

$$\int_{\Omega} \left\{ \left| \frac{\underline{v}}{\rho} - \text{grad } \phi \right|^2 + w|\text{div } \underline{v}|^2 \right\} = \min, \tag{2.9}$$

where $\underline{v} = \rho\underline{u}$ is the mass flow and w is a weighting function to be chosen. In this setup the variables are the potential ϕ and the mass flow $\underline{v}$.

The density in (2.9)

$$\rho = \rho(|\text{grad } \phi|)$$

is obtained from Bernoulli's equation, i.e.,

$$\rho^{\gamma-1} = \left\{ 1 - \left(\frac{\gamma-1}{2}\right) M_{\infty}^2 \left(|\text{grad } \phi|^2 - 1 \right) \right\}.$$

Thus, (2.9) is a nonlinear least squares formulation, which is appropriate since it reflects the nonlinear character of transonic flow. Once a grid is selected (specific examples are given in the next section), the minimization of (2.9) over the associated finite element space leads to a nonlinear system

$$K(\underline{\Phi})\underline{\Phi} = \underline{F}. \tag{2.10}$$

In all of the numerical examples reported in the next section, (2.10) was solved by a combination of Newton's method and elimination. Issues related to this choice for the equation solver will be discussed in the next section.

There are three main cases that are considered in this paper:

Case 1: smooth subsonic flows,

Case 2: smooth transonic flows,

Case 3: transonic flows with shocks.

In the first case (2.9) can be used without modification, and in particular no weighting function is needed (i.e., $w \equiv 1$ can be used). One does need special grids to obtain optimal accuracy (see [1]), and the criss-cross grid pattern which satisfies the grid decomposition property of [1] is used.

In the second case a hyperbolic region appears but the flow field remains smooth. In this case there is a loss of accuracy in the hyperbolic region. In particular, with linear elements the pointwise accuracy in the mass flow $\underline{v}$ drops from $O(h^2)$--in a generic mesh spacing--to $O(h)$. This can be corrected with a suitable choice of weighting function w, and details are given in [8]. This modification was not used in the results reported in this paper since the hyperbolic regions in question were too small for the suboptimal accuracy to have a major effect on the qualitative features of the flow.

The third case is, by a wide margin, the most important as well as the most challenging. Here we use a weight w so that the term

$$\int_{\Omega} \left\{ w|\operatorname{div} \underline{u}|^2 + \left| \frac{\underline{u}}{\rho} - \operatorname{grad} \phi \right|^2 \right\} \tag{2.11}$$

remains meaningful. In addition, modification to the density $\rho = \rho(|\text{grad } \phi|)$ must be introduced so that nonphysical expansion shocks are eliminated.

For the choice of the weight w, we follow the developments introduced in [7]. For most flows, $\underline{v} = \rho\underline{u}$ is continuous across the shock [10]. Nevertheless, it does not follow that div $\underline{v}$ is square integrable, and the primary rule derived from [7] is that w be chosen so that

$$\int_{\Omega} w|\text{div } \underline{v}|^2 < \infty. \tag{2.12}$$

This requires that w vanishes appropriately on the shock, which in turn means that (2.11) is a least squares principle in a degenerate L^2 norm. A point of significance, on the other hand, is the fact that if w vanishes to minimal order on the shock (in that (2.12) still holds), then optimal $O(h^2)$ can be achieved in unweighted L^2 norms provided appropriate mesh refinement is introduced. This has been proved rigorously only in special cases (see [7]), yet the numerical results in the next section seem to indicate that the principle is general.

These modifications alone do not yield an accurate simulation of the flow problem. To do this one must deal with the presence of nonphysical expansion shocks. In effect, (2.9) does not have a unique minimum, neither over infinite-dimensional function spaces nor over the finite-dimensional finite-element spaces. One can have expansion shocks, compression shocks, or both. What is interesting is the results in the next section tend to indicate that the case where both type of shocks appear tends to be the stable mode for (2.10). That is, an arbitrary choice of starting vector for Newton's methods applied to (2.10) tends to converge to this solution.

To eliminate expansion shocks we consider density biasing which in effect introduces streamwise diffusion into (2.1) - (2.2). Following [11] (see also [12] - [14]) the modified density takes the form

$$\bar{\rho} = \rho - \mu \rho_s \, \Delta s, \tag{2.13}$$

where ρ_s is the derivative of the density ρ along the streamwise direction. Since the density has the form

$$\rho = \rho(|\text{grad } \phi|),$$

the derivative ρ_s formally involves second derivatives of ϕ. Since ϕ is expanded in terms of linear elements, it is necessary to replace ρ_s with a streamwise difference quotient; i.e.,

$$\bar{\rho} = \rho - \mu \, \Delta\rho \, \Delta s, \tag{2.13'}$$

in the least squares formulation.

3. NUMERICAL RESULTS

To illustrate the above ideas we selected the classic problem of a planar flow past a cylinder. The flow region plus boundary conditions are given in Figure 3.1. The configuration shown in this figure assumes that both the outflow and inflow remain subsonic. Figure 3.2 contains a typical grid. For economy only the top part of the flow region is shown, and the special refinement needed for the shocks is not shown.

The first set of results shows a typical subsonic flow pattern. The results are given in Figure 3.3 for a free stream Mach number of

$$M_\infty = 0.1.$$

Convergence studies at such Mach numbers are reported in [5] - [6]. These results indicate, with the type of grid shown in Figure 3.2, one can readily achieve L^2 error of 1% or less for the velocity field.

The next set of results deal with the smooth transonic case. Of special interest here is the ability of the scheme to detect the onset of supersonic flow. Analytical techniques (see [15] and [16]) have given accurate values for the critical free stream Mach number M_* as a function of d/D, where d is the diameter of the cylinder and D is the width of the channel. These results are reproduced in Figure 3.4. Numerical results from the least squares scheme are given in Figures 3.5 - 3.7 for $M_\infty = .42$, $.45$, and $.50$, respectively. The d/D ratio used for this case is $1/6$. Extrapolation based on these results indicates that the critical Mach number is approximately .41, which is good agreement with Figure 3.4.

The next set of results show what least squares based schemes produce when diffusion via density modification is not used. These are shown in Figure 3.8 which contains plots of the velocity $q = |\underline{u}|$ versus angle θ along the cylinder and at a radius slightly above the cylinder. The free stream Mach number is $M_\infty = .5$. The shock at the front of the cylinder is an expansion shock and is nonphysical. The one at the rear is a compression shock. A remarkable feature of this approximation is that the physically relevant compression shock is approximately in its correct position and is apparently unaffected by the spurious shock. (Compare Figures 3.8 and 3.9.)

The solution shown in Figure 3.8 is apparently a stable mode for the nonlinear system (2.10). Indeed, Newton's method converged to this solution rather rapidly for a wide variety of initial conditions.

In this regard, it is interesting to note that for the least squares formulation the Jacobian is not singular near the solution shown in Figure 3.8. Density modifications are needed to remove the spurious shock shown at the front of the cylinder. However, they are not needed to obtain nonsingular Jacobians.

The final results deal with the complete least squares system with the density modification discussed in the previous section. Figures 3.9 - 3.11 show the velocity field over the cylinder, at a radius slightly larger that that of the cylinder, and at a radius in the free stream. Note that the spurious expansion shock has been totally eliminated. Moreover, the shock location and strength as well as the velocity profile appear to be correct as is the supersonic bubble shown in Figure 3.12.

While we regard these numerical experiments as successful, there are a number of areas where the approach could be improved. The first issue concerns the equation solver. Once the density modification were introduced, the number of iterations increased by a factor of 2 to 3. Moreover, the solution shown in Figure 3.9 tended to be less "attractive" to the Newton iterations than that shown in Figure 3.8 (without density modifications). In fact, it was not difficult to find starting vectors where nonconvergence was seen, in the former case, although the starting state of a uniform flow always leads to convergence. This suggests that an alternative equation solver (e.g., preconditioned conjugate gradient) might be a more efficient choice for the equation solver.

A second issue concerns post-shock oscillations. These are seen in Figure 3.10, which is the radius where the oscillations were found to be the most significant. These oscillations were not seen on the body of the cylinder (Figure 3.9) and disappeared rather rapidly away from the cylinder (Figure 3.11). This is clearly a grid effect due to the slight misalignment of shock and grid.

4. CONCLUSIONS

Finite difference approximations to transonic flow problems are well-developed and have been successfully used for a wide range of problems. Nevertheless, there is still a need to develop finite element approaches for such problems for a variety of applications. We feel that the results presented here do show that such schemes can give physically meaningful simulations.

On the other hand, our experience has tended to indicate that straightforward application of the basic finite element idea may not always be successful. Key computational issues are as follows:

(i) There is a need to carefully develop the spaces in which the approximations are formulated. Classical L^2 spaces are generally inappropriate.

(ii) Some form of diffusion (via density modifications or otherwise) appears to be needed. Moreover, care is needed in the way this diffusion is introduced.

(iii) The geometrical pattern of the grid selected is of importance. Some

patterns are definitely superior to others.

Finally, there are some important "bottlenecks" associated with the scheme employed in this paper, which, if properly addressed, could lead to an even more efficient approach. These include the following:

(i) There is a need for an equation solver that is more efficient than the Newton method used in this paper.

(ii) There is a need for adaptive grid refinement techniques that would lead to a better shock grid alignment than that achieved in this paper.

REFERENCES

[1] G. J. Fix, M. D. Gunzburger, and R. A. Nicolaides: Least squares finite element methods, NASA-ICASE Report No. 77-18, revised version published in Comput. Math. Appl., Vol. 5, 1979, pp. 87-98.

[2] G. J. Fix and M. Gurtin: On patched variational methods, Numer. Math., Vol. 28, 1977, pp. 259-271.

[3] G. J. Fix and M. D. Gunzburger: On least squares approximation to indefinite problems of the mixed types, Internat. J. Numer. Methods Engrg., Vol. 12, 1978, pp. 453-470.

[4] C. L. Cox, G. J. Fix, and M. D. Gunzburger: A least squares finite element scheme for transonic flow around harmonically oscillating wings, J. Comp. Phys., Vol. 51, No. 3, September 1983, pp. 387-403.

[5] T.-F.Chen: On finite element approximations to compressible flow problems, Ph.D. Thesis, Carnegie-Mellon University, May 1984.

[6] T. F. Chen: Least squares approximation to compressible flow problems, submitted to Comput. Math. Appl.

[7] C. L. Cox and G. J. Fix: On the accuracy of least squares methods in the presence of corner singularities, Comput. Math. Appls., Vol. 10, No. 6, 1984, pp. 463-476.

[8] G. J. Fix and M. E. Rose: A comparative study of finite element and finite difference methods for Cauchy-Riemann type equations, SIAM J. Numer. Anal., Vol. 22, No. 2, 1985, pp. 250-260.

[9] G. J. Fix: Least squares approximation o hyperbolic systems, submitted to SIAM J. Numer. Anal.

[10] P. D. Lax: Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves, SIAM Regional Conf. Series Lectures in Appl. Math., Vol. 11, 1972.

[11] S. Osher, M. Hafez, and W. Whitlow, Jr.: Entropy conditions satisfying approximations for the full potential equation of transonic flow, Math. Comp., Vol. 44, No. 169, January 1985, pp. 1-29.

[12] A. Eberle: Eine Method Finiter Elements Berechnung der Ttanssonicken Potential--Strimung un Profile, MBB Berech Nr. UFE 1352(0), 1977.

[13] M. M. Hafez, E. M. Murman, and J. C. South: Artificial compressibility methods for numerical solution of transonic full potential equation, AIAA Paper 78-1148, Seattle, Washington, 1978.

[14] M. Hafez, W. Whitlow, Jr., and S. Osher: Improved finite difference schemes for transonic potential calculations, AIAA Paper 84-0092, Reno, Nevada, 1984.

[15] I. Imai: On the flow of a compressible fluid past a circular cylder, II, Proc. Phys. Math. Soc. Japan, Vol. 23, 1941, pp. 180-193.

[16] Z. Hasimoto: On the subsonic flow of a compressible fluid past a circular cylinder between two parallel walls, Proc. Phys. Math. Soc. Japan, Vol. 25, 1943, pp. 563-574.

[17] A. Jameson: Numerical solutions of nonlinear partial differential equations of mixed type, Numerical Solutions of Partial Differential Equations III, Academic Press, New York, 1976, pp. 275-320.

[18] M. O. Bristeau, R. Glowinski, Periaux, J., P. Perrier, O. Pironneau, G. Poirier: A Finite Element Method for the Numerical Simulation of Transonic Potential Flows, Finite Element Handbook, McGraw-Hill, 1983.

[19] R. Pelz and A. Jameson: Transonic flow calculations using triangular finite elements, AIAA J., Vol. 23, No. 4, 1985, pp. 569-576.

[20] W. G. Habashi and M. M. Hafez: Finite element solution of transonic flow problems, AIAA Paper 81-1472.

[21] H. Deconinck and C. Hirsch: Finite element methods for transonic flow calculations, Proc. Conference on Numerical Methods in Fluid Mechanics, 3rd, Cologne, West Germany, October 10-12, 1979, Braunschweig, Friedr. Vieweg und Sohn, Verlagsgesellschaft mbH, 1980, pp. 66-77.

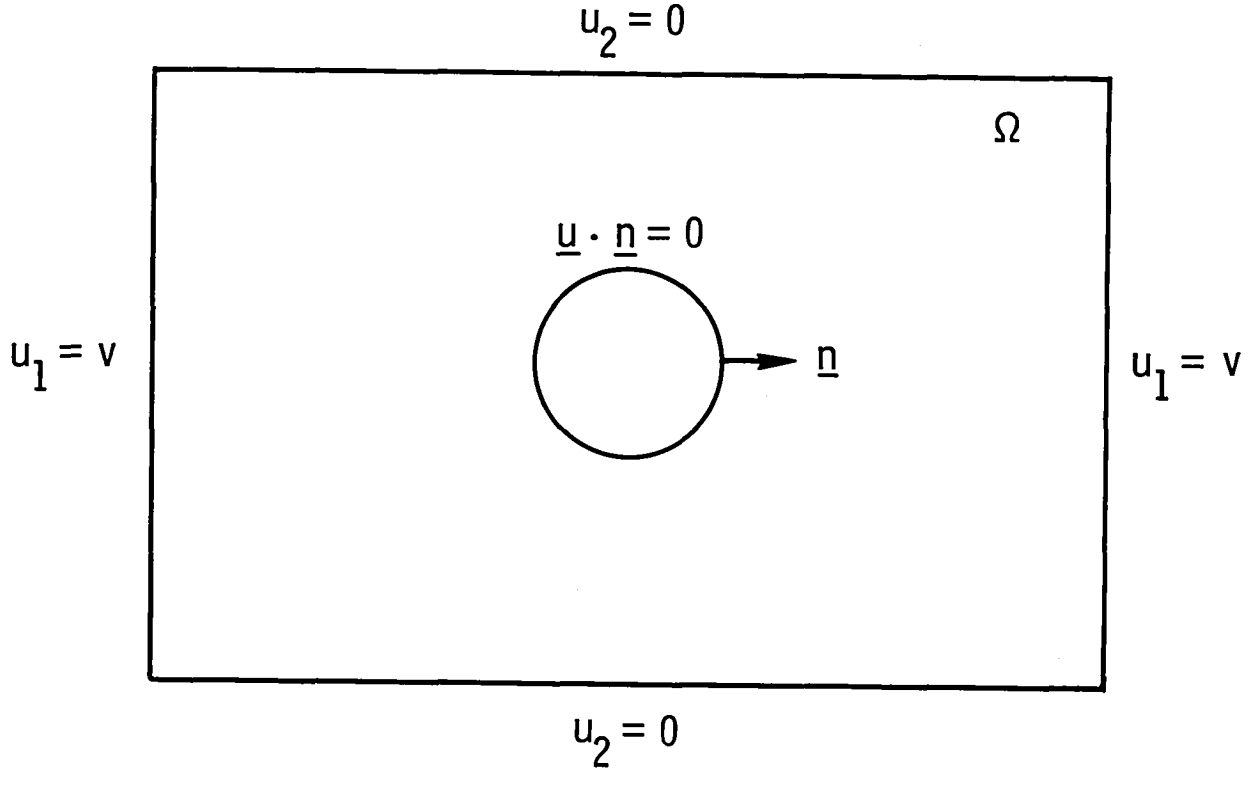


Figure 3.1. The flow region Ω.

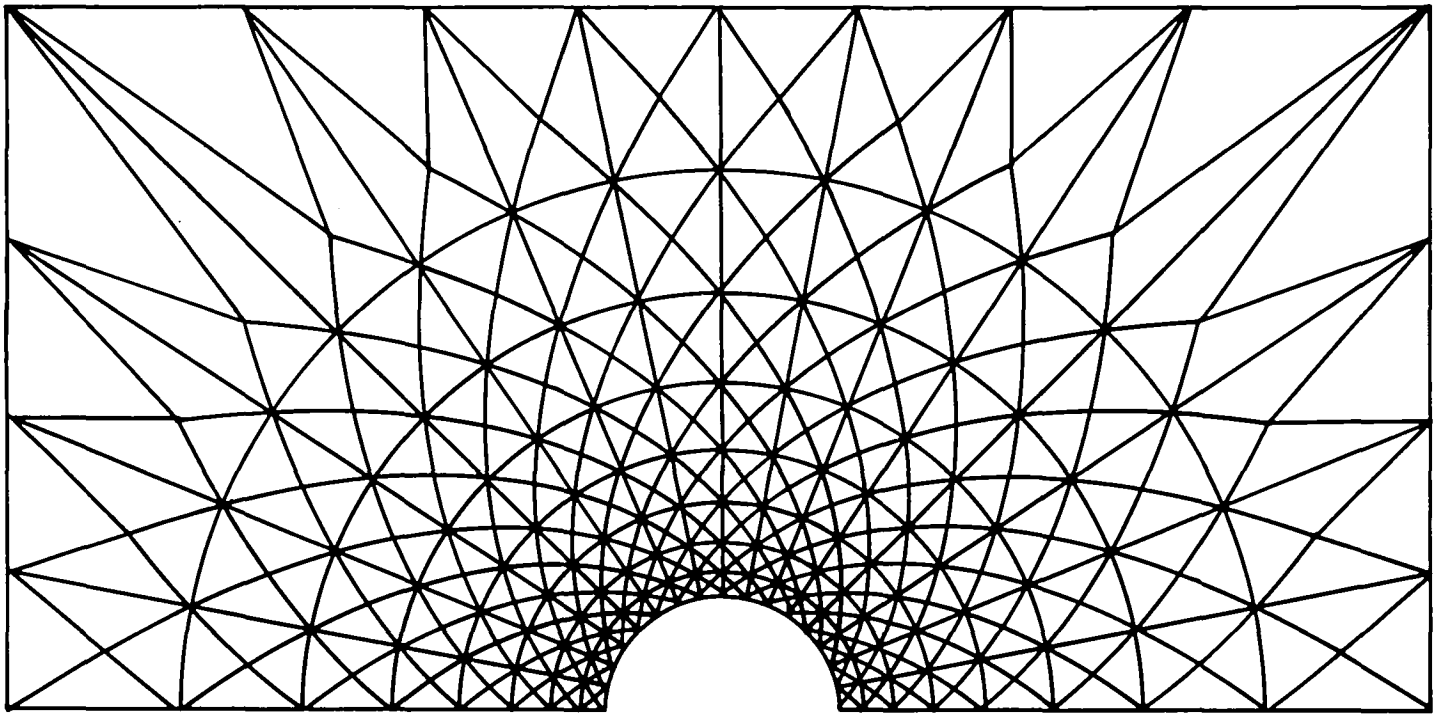


Figure 3.2. 512 elements, 281 nodes, $h = 0.30907 \times 10^{-1}$.

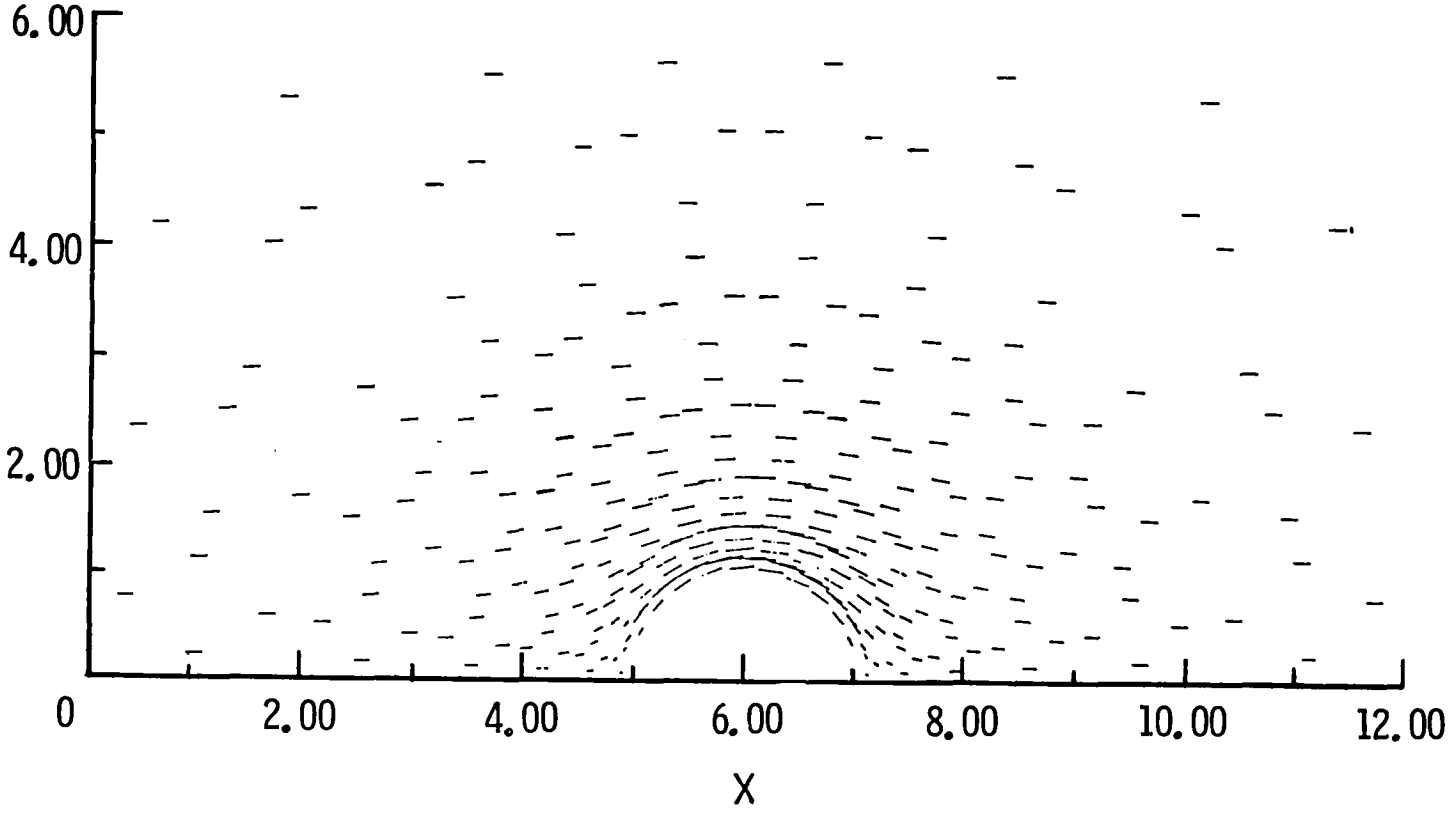


Figure 3.3. Flow pattern for the free stream Mach number $M_\infty = 0.1$.

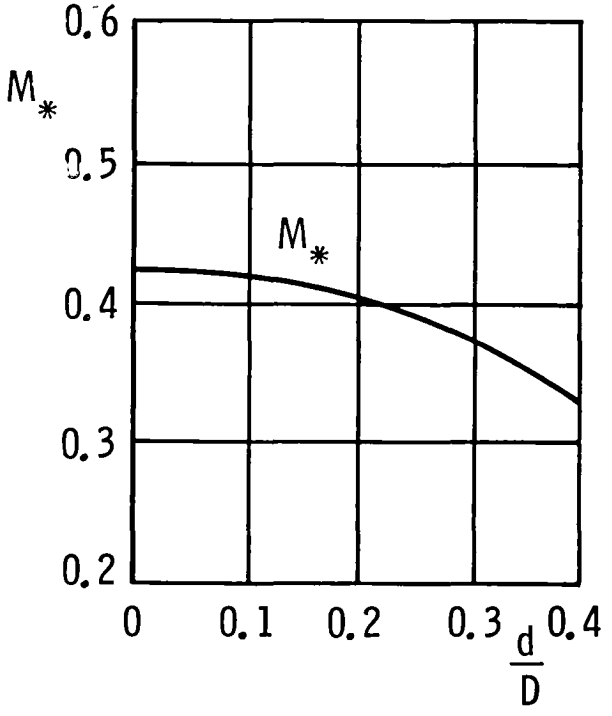


Figure 3.4. Critical Mach number versus d/D.

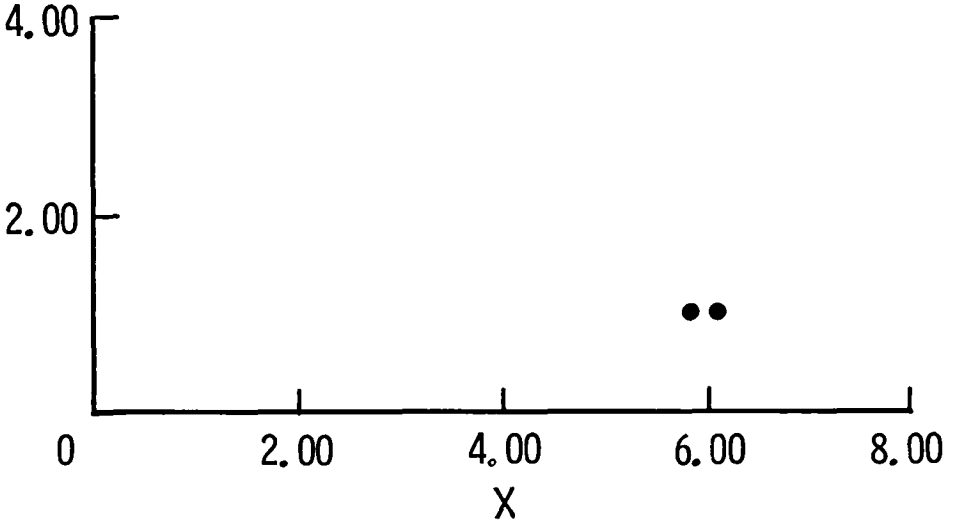


Figure 3.5. Plots of the supersonic pocket for $M_\infty = 0.42$.

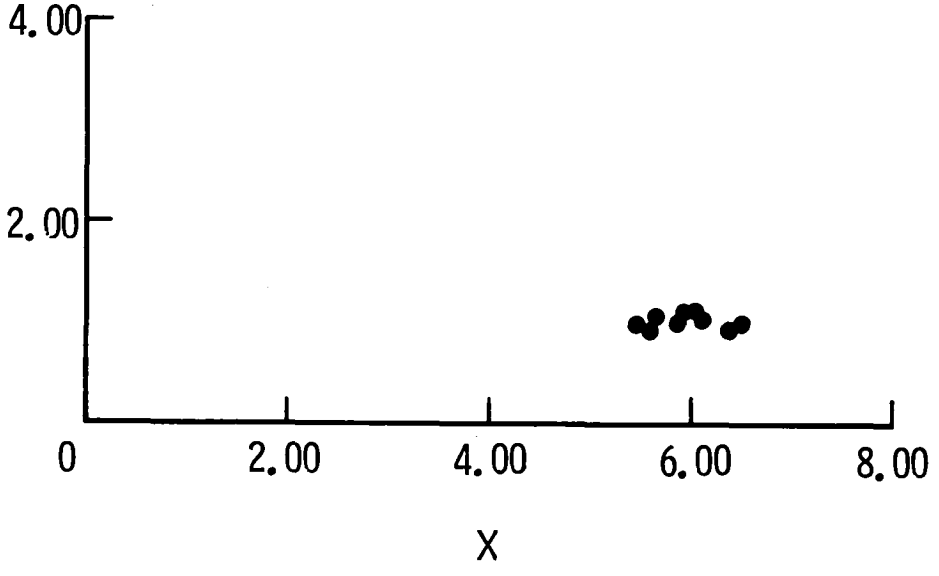


Figure 3.6. Plots of the supersonic pocket for $M_\infty = 0.45$.

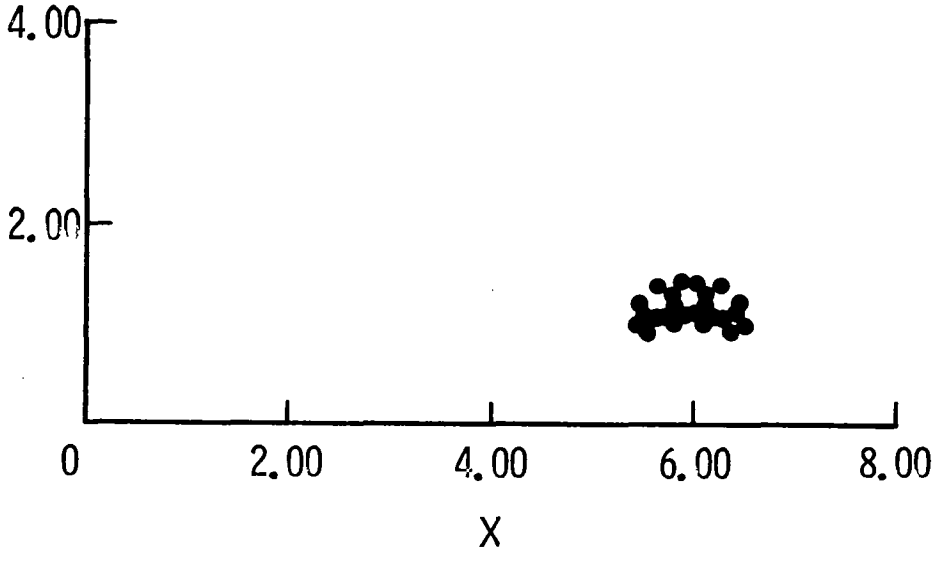
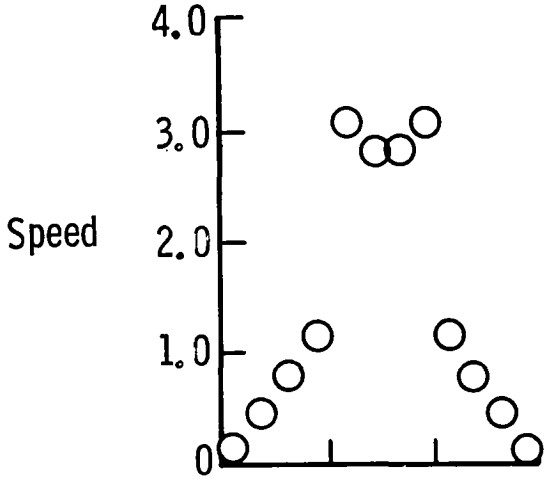


Figure 3.7. Plots of the supersonic pocket $M_\infty = 0.50$.

(a)



(b)

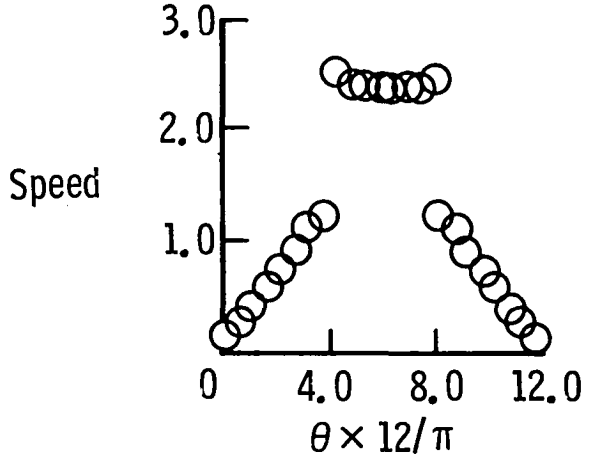


Figure 3.8. Velocity as a function of angle: (a) on cylinder, (b) slightly off cylinder — $M_\beta = .51$.

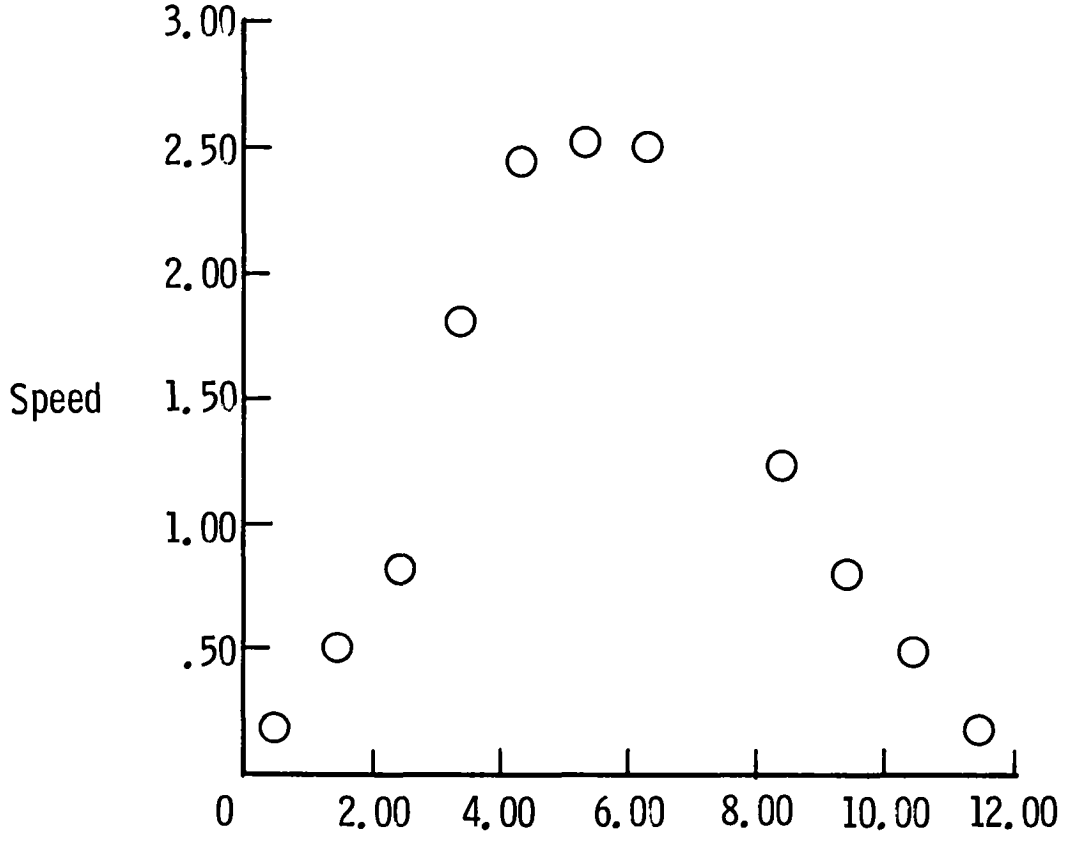


Figure 3.9. Velocity as a function of angle on the cylinder -- full least squares scheme with density modification -- $M_\infty = .5$.

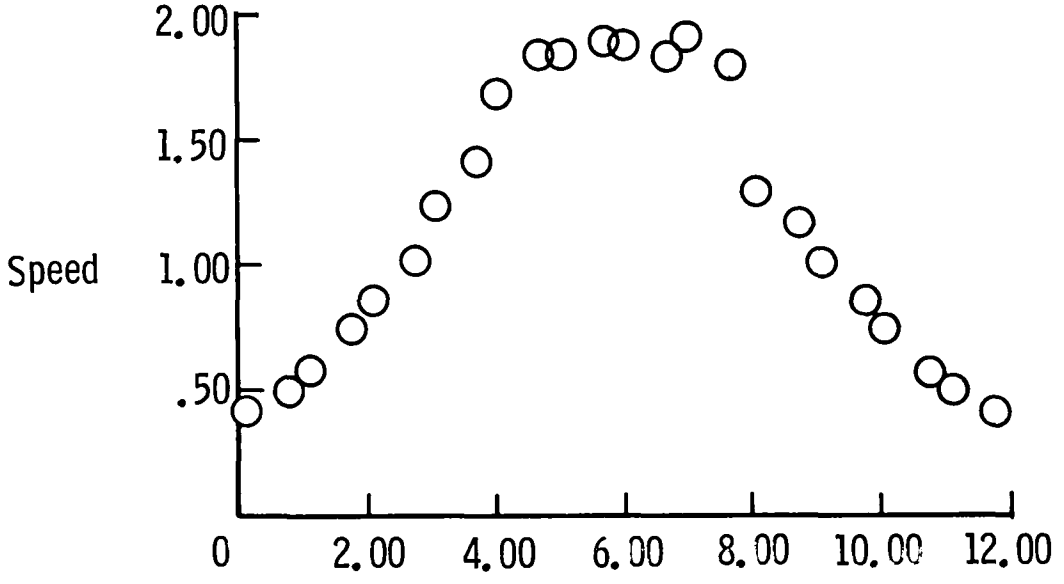


Figure 3.10. Velocity as a function of angle slightly off cylinder -- full least squares scheme with density modification -- $M_\infty = .5$.

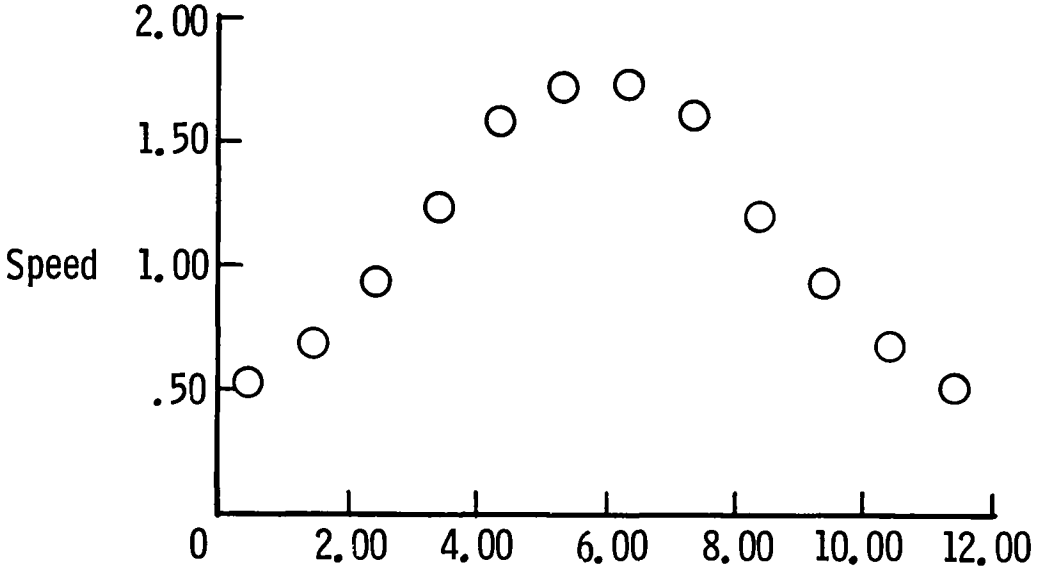


Figure 3.11. Velocity as a function of angle half radius above cylinder -- full least squares scheme with density modification -- $M_\infty = .5$.

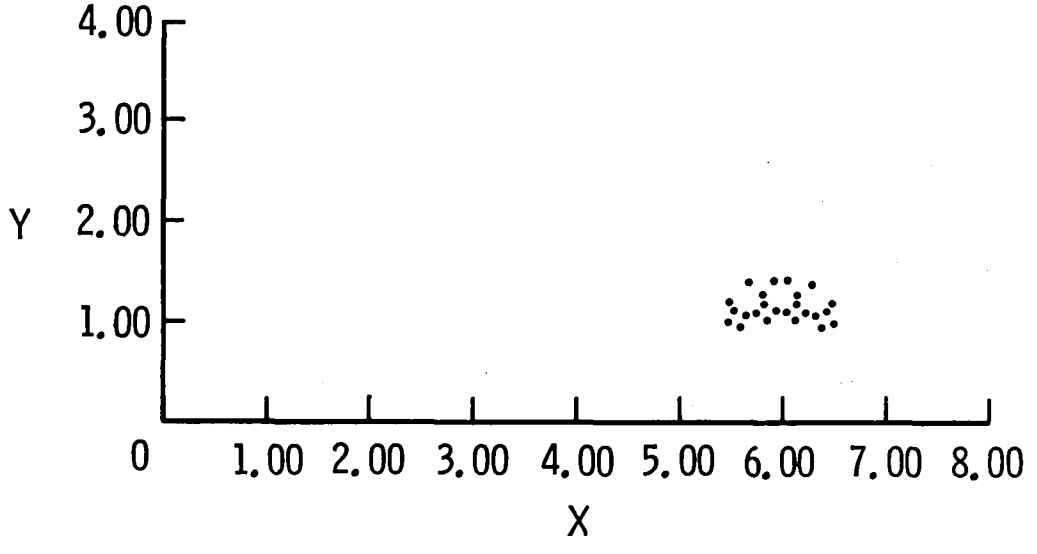


Figure 3.12. Supersonic bubble -- full least squares scheme with density modification -- $M_\infty = .5$.

THE WEAK ELEMENT METHOD APPLIED TO HELMHOLTZ TYPE EQUATIONS

Charles I. Goldstein
Department of Applied Mathematics
Brookhaven National Laboratory
Upton, NY 11973

ABSTRACT

Helmholtz type boundary value problems are important in a variety of scattering and diffraction problems. Standard numerical schemes based on finite difference, finite element, or integral equation methods are generally not well suited for these problems in the "intermediate frequency range" since the oscillatory solution is not accurately approximated by piecewise polynomials. In this paper, a version of the weak element method is employed to numerically solve these problems in two dimensions. This method consists of partitioning the domain into small "elements" and locally approximating the solution in each element by a sum of exponentials. These piecewise approximations are joined together at interelement boundaries by continuity conditions for certain functionals of the approximate solution. The method is analyzed using a complementary variational formulation. It is shown that the weak element method is considerably more accurate than standard discretization methods when the solution is adequately approximated locally by the exponential basis functions. These results are validated by numerical experiments.

The submitted manuscript has been authored under contract DE-AC02-76CH00016 with the U. S. Department of Energy.

1. INTRODUCTION

It is the purpose of this paper to analyze and numerically investigate the weak element method applied to Helmholtz type boundary value problems in multi-dimensional domains. Scalar and vector Helmholtz type equations, $(\Delta + K^2 n)u = 0$, with an appropriate radiation condition and spatially dependent index of refraction, n, are of importance in a variety of stationary wave propagation problems occurring in acoustics, optics, seismology, and electromagnetic theory. Since the solution is rarely known in closed form, it is important to approximately solve these problems numerically in the intermediate frequency range, where asymptotic methods can be unreliable.

When applying typical discretization methods such as finite difference and finite element methods as well as integral equation methods, one is faced with the "resolution problem". This means that in order to approximate the solution accurately when the wave number, K, is not small, one must decrease the grid size, h, and hence solve a prohibitively large number of linear equations. This problem arises from the use, in the usual discretization methods, of piecewise polynomial functions to approximate a highly oscillatory solution. Methods for overcoming this difficulty have been developed in [1] and [2] by combining the finite element method with functions satisfying the desired oscillatory behavior. The method in [1] was developed for one-dimensional problems. The method in [2] was designed to treat multi-dimensional problems for which most of the propagation occurs in a narrow angle band about a fixed direction.

An alternative approach for discretizing boundary value problems is given by the weak element method developed in [3]. This method is based on partitioning the domain into small subdomains (elements) and approximating the solution in each element by a solution of a localized approximation of the differential equation. These piecewise approximations are joined together at interelement boundaries by continuity conditions for certain functionals of the approximate solution. See [4] and [5] as well as references cited there for a discussion of related methods. In this paper we consider a version of the weak element method in which the approximate solution consists of piecewise exponential basis functions joined together at interelement boundaries by imposing continuity conditions on the average values of the approximate solution and its normal derivative. This method is described briefly in Section 2 and in detail in [3].

In Section 3 we analyze this weak element method for a model problem in a rectangle. The analysis employs a complementary variational principle developed in [4] in connection with the Laplace equation. Here we extend the arguments in [4] to a non-selfadjoint Helmholtz boundary value problem. We prove that when K^2h is sufficiently small, the resulting discrete problem is well-posed and the mean-square discretization error is of order $O(K^3h^2)$ as $h\to 0$. This is analogous to the situation for standard second order finite element or finite difference schemes. We also show that when the phase of the solution is adequately approximated locally by the exponential basis functions, the weak element method is much more accurate than standard discretization schemes as K increases. This is the main advantage of

the weak element method. Some techniques for approximating the phase of the exact solution are described in [1] and [2]. In Section 4 we demonstrate the results of some numerical experiments with the weak element method. We summarize our conclusions in Section 5.

2. THE WEAK ELEMENT METHOD

In this section we outline briefly the weak element method described in [3]. We employ the following notational convention. Suppose that $\bar{a}=(a_1,a_2,\ldots,a_n)$ and $\bar{b}$ denote vectors with n components, and $\Phi=(\phi_j^{\ i})$ denotes an nxn matrix whose ith column is $\bar{\phi}^i$ and whose jth row is $\bar{\phi}_j$. We denote the inner product of $\bar{a}$ and $\bar{b}$ by $\bar{a}\cdot\bar{b}$ and the norm of $\bar{a}$ by $|\bar{a}|=(\bar{a}\cdot\bar{a})^{1/2}$. No notational distinction is made between row and column vectors. Hence $\bar{a}$ in $\bar{a}\Phi$ is a row vector, but $\bar{a}$ in $\Phi\bar{a}$ is a column vector.

We consider the following differential operator acting in a bounded domain D in the $\bar{x}=(x_1,x_2)$ plane with a piecewise smooth boundary, ∂D. Suppose that P and A are 2x2 matrices (P being positive definite symmetric) and b and q are scalars. Let $\bar{n}$ denote the outward directed unit normal to D and let $\nabla=(\partial/\partial x_1,\partial/\partial x_2)$ denote the gradient. The linear elliptic operator L is defined in D by

$$L=-\nabla\cdot P\nabla+q, \tag{2.1}$$

and the boundary operator B is defined on ∂D by

$$B=\bar{n}\cdot A\nabla+b. \tag{2.2}$$

Before proceeding further we require the following additional notation. Let $\Pi_N(D)$ denote a partition of D into N elements (subdomains), $\{\pi_i\}$. We use $\sigma_j(\pi)$, $j=1,2,\ldots,\ell(\pi)$, to denote one of the $\ell(\pi)$ smooth sides of the element π. The vector $\bar{\sigma}(\pi)=(\sigma_1(\pi),\sigma_2(\pi),\ldots,\sigma_{\ell(\pi)}(\pi))$ denotes the sides of π oriented in a counterclockwise manner about π. A side $\sigma_j(\pi)$, which is incident to another subdomain, π', is an interior side and is denoted by $\sigma(\pi,\pi')$. Otherwise $\sigma_j(\pi)$ lies on ∂D and is denoted by $\sigma'(\pi)$. (See Figure 1 for the case of rectangular elements.)

The area of π (length of σ_j) is denoted by $|\pi|(|\sigma_j|)$. Let $\theta(\pi)$ denote the smallest angle between the centroid, $\bar{x}^o$, and any two distinct vertices of π. In order for the resulting system of linear equations to be well conditioned, we assume that $\theta(\pi)\geq\theta_o>0$ for each element $\pi\epsilon\Pi_N(D)$, where θ_o is independent of N.

We define localizations, $L(\pi)$ and $\hat{L}(\pi)$, of the operator L given by (2.1) with respect to the element π as follows:

$$L(\pi)=-\nabla\cdot(P_o\nabla)-(\nabla P_o)\cdot\nabla+q_o \tag{2.3}$$

and

$$\hat{L}(\pi)=-\nabla\cdot(P_o\nabla)+q_o \tag{2.4}$$

where P_o denotes P evaluated at $\bar{x}_o$, etc. Finally, if u(x) is a smooth (possibly vector-valued) function and $\sigma_j(\pi)$ is an arbitrary side of π, we define

$$u(\sigma_j(\pi))=\frac{1}{|\sigma_j(\pi)|}\oint_{\sigma_{j(\pi)}} u(\bar{x})ds$$

and

$$\bar{u}(\bar{\sigma}(\pi))=(u(\sigma_1(\pi)),u(\sigma_2(\pi)),\ldots,u(\sigma_{\ell(\pi)}(\pi))).$$

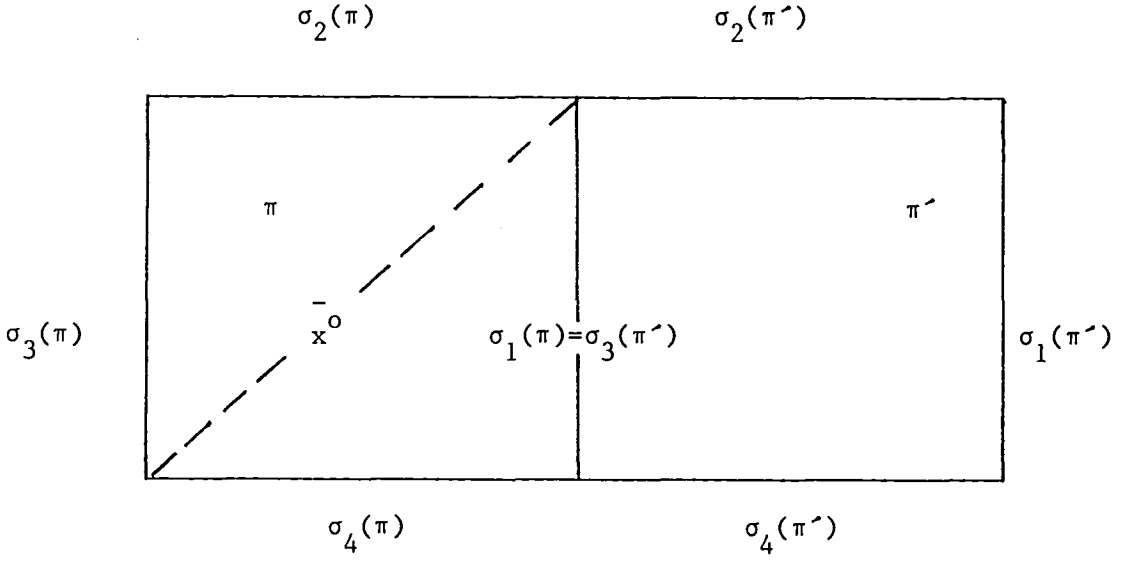


Figure 1

We are now ready to describe the weak element method employed here to solve the boundary value problem Lu=0 in D, Bu=g on ∂D. For each element π, let $\phi_1(\bar{x},\pi)$, $\phi_2(\bar{x},\pi),\ldots,\phi_{\ell(\pi)}(\bar{x},\pi)$ denote a linearly independent set of solutions of the localized equation

$$\hat{L}(\pi)\phi_i=0 \tag{2.5}$$

and define

$$\bar{\phi}(\bar{x},\pi)=(\phi_1(\bar{x},\pi),\phi_2(\bar{x},\pi),\ldots,\phi_{\ell(\pi)}(\bar{x},\pi)).$$

Our approximate solution on π is now defined by the equation

$$w(\bar{x},\pi)=\bar{\phi}(\bar{x},\pi)\cdot\bar{a}(\pi), \tag{2.6}$$

where the coefficient vector $\bar{a}(\pi)=(a_1(\pi),a_2(\pi),\ldots,a_{\ell(\pi)}(\pi))$ is unknown.

Now suppose that $\sigma_j(\pi)$ is incident to π' at the side $\sigma_{j'}(\pi')$ and let $\sigma'(\pi)$ be a side of π on the boundary ∂D. We impose the following continuity and boundary conditions on $w(x,\pi)$:

$$w(\sigma_j(\pi))=w(\sigma_{j'}(\pi')), \tag{2.7a}$$

$$(\bar{n}_j\cdot P\nabla w)(\sigma_j(\pi))=(\bar{n}_j\cdot P\nabla w)(\sigma_{j'}(\pi')) \tag{2.7b}$$

on interior sides, where $\bar{n}_j$ is the outward directed unit normal to $\sigma_j(\pi)$, and

$$(Bw)(\sigma'(\pi))=g(\sigma'(\pi)) \tag{2.7c}$$

on boundary sides. Substituting (2.6) into (2.7), we obtain a system of linear equations for the N vectors

$$\bar{a}(\pi_i)=(a_1(\pi_i),a_2(\pi_i),\ldots,a_{\ell(\pi)}(\pi_i)),i=1,\ldots,N.$$

It is shown in [3] that the weak element approximation given by (2.6) may be obtained by solving an equivalent smaller system of equations for the average values of w on all sides, $\sigma_j(\pi)$, of the partition.

Remark 2.1: As described in [3], the weak element method can be generalized as follows. In (2.7), we impose boundary (continuity) conditions on the average value of the function (and an appropriate derivative) for each boundary (interior) side, $\sigma_j(\pi)$. To generalize the method we can replace the average value on $\sigma_j(\pi)$ by a set of linear functionals on $\sigma_j(\pi)$, denoted by $\langle \Lambda^m(\sigma_j(\pi)),u\rangle, m=1,\ldots M$. We would then require $M\ell(\pi)$ local basis functions in each element. (For example, these linear functionals might consist of the average value of higher order moments of u on each side.) This could lead to higher order methods than the method discussed in this paper for which M=1 and $\langle \Lambda^1(\sigma_j(\pi)),u\rangle = u(\sigma_j(\pi))$.

For the sake of simplicity, in the remainder of the paper we consider the special case in which each element $\pi\in\Pi_N(D)$ is a rectangle with sides parallel to the x_1,x_2 coordinate axes. As will be seen in the next two sections, a key to the success of the weak element method lies in the choice of the local basis functions, $\phi_i(\bar{x},\pi), i=1,\ldots,4$. We employ an exponential basis defined as follows. Let $\bar{x}^o=(x_1^o,x_2^o)$ denote the center of π and define the unit vectors $\bar{e}_1=(1,0)$ and $\bar{e}_2=(0,1)$. We now set

$$\left.\begin{aligned} &\phi_1(\bar{x},\pi)=e^{\rho_1(x_1-x_1^o)}, \phi_2(\bar{x},\pi)=e^{\rho_2(x_2-x_2^o)}, \phi_3(\bar{x},\pi)=e^{-\rho_1(x_1-x_1^o)}, \\ &\text{and } \phi_4(\bar{x},\pi)=e^{-\rho_2(x_2-x_2^o)} \end{aligned}\right\} \quad (2.8)$$

where ρ_1 and ρ_2 are chosen so that each $\phi_j(x,\pi)$ satisfies (2.5). A simple calculation yields

$$\rho_j = q_o^{1/2}(\bar{e}_j \cdot \rho_o \bar{e}_j)^{-1/2}, j=1,2. \tag{2.9}$$

Basis functions analogous to those given by (2.8) and (2.9) can be obtained by solving the equation $L(\pi)\phi_i=0$ instead of (2.5).

The basis functions in (2.8) can be generalized as follows. Define the unit vectors $\bar{e}_{1\alpha}=(\cos\alpha,\sin\alpha)$ and $\bar{e}_{2\alpha}=(\sin\alpha,\cos\alpha)$ with $0<\alpha<\frac{\pi}{4}$. Now set

$$\left.\begin{aligned} &\phi_{1\alpha}(\bar{x},\pi)=e^{\rho_{1\alpha}\overline{e_{1\alpha}}\cdot(\bar{x}-\overline{x^o})}, \phi_{2\alpha}(\bar{x},\pi)=e^{\rho_{2\alpha}\overline{e_{2\alpha}}\cdot(\bar{x}-\overline{x^o})}, \\ &\phi_{3\alpha}(\bar{x},\pi)=e^{-\rho_{1\alpha}\overline{e_{1\alpha}}\cdot(\bar{x}-\overline{x^o})}, \text{ and } \phi_{4\alpha}(\bar{x},\pi)=e^{-\rho_{2\alpha}\overline{e_{2\alpha}}\cdot(\bar{x}-\overline{x^o})}. \end{aligned}\right\} \tag{2.10}$$

The constants $\rho_{1\alpha}$ and $\rho_{2\alpha}$ can be determined as before by substituting (2.10) into (2.5). Note that α can have different values in different elements. This can be useful when some knowledge is available concerning the phase of the exact solution (see Remark 3.1 below).

The finite difference equations obtained using basis (2.8) were derived in [3]. See [6] for a detailed investigation of the resulting finite difference formulas using both (2.8) and (2.10) and for various aspects of the implementation of the method. The resulting system of equations may then be solved for the unknowns, $a_j(\pi_i), j=1,\ldots,4, i=1,\ldots,N$. The weak element approximation, $w(\bar{x})$, is obtained from (2.6). Hence we obtain $w(\bar{x})$ at each point $\bar{x}$ in D instead of only at nodal points. Observe that the resulting matrix is highly sparse. Furthermore, the corresponding large system of

equations is nonselfadjoint with indefinite symmetric part for problems of the kind considered in this paper. The preconditioned iterative method developed in [7] is well suited for solving this system of equations. This iterative solver has not been implemented in connection with the weak element method at the present time.

3. ERROR ANALYSIS

In this section we consider, for the sake of simplicity, the following model problem:

$$\left.\begin{array}{ll} \text{(a)} & (-\Delta-(K^2+i\delta K))u=0 \text{ in } D, \\ \text{(b)} & u=g \text{ on } \partial D, \end{array}\right\} \quad (3.1)$$

where D is the unit square, $\delta>0$, and we assume that the solution $u\in C^2(D)$. The term $i\delta K$ is chosen to simulate a radiation condition as in [8]. Furthermore, it is easily seen that this term ensures the well-posedness of (3.1). We set $K'=\sqrt{K^2+i\delta K}$ and note that $q=iK'$ in (2.1) and $b=1$ in (2.2). Furthermore, P(A) is the 2x2 identity (null) matrix in (2.1) ((2.2)).

We shall employ the weak element method described in Section 2 with local basis functions given by (2.8) and (2.9). Hence we have a partition of $D, \Pi_N=\Pi_N(D)$, into small rectangular elements, π_i, $i=1,\ldots,N$, such that the local basis functions defined on π_i are given by

$$e^{\pm iK'(x_1-x_1^i)}, \quad e^{\pm iK'(x_2-x_2^i)}, \quad (3.2)$$

where (x_1^i,x_2^i) denotes the centroid of π_i. Denote the lengths of the horizontal and vertical sides of π_i by h_1^i and h_2^i, respectively, and define

$$h=\max_{\pi_i\in\Pi_N} \max(h_1^i,h_2^i). \quad (3.3)$$

We shall analyze the discretization error using a complementary variational formulation developed in [4] for the Laplace equation.

Before describing the variational formulation, we introduce some additional notation. For a fixed element $_i\varepsilon\Pi_N$, let $\sigma_{j,i}=\sigma_j(\pi_i)$, $j=1,\ldots,4$, denote the four sides of $_i$ (see Figure 1 above) and set $\rho(\sigma_{j,i})=1$ (-1) if $\sigma_{j,i}$ is to the right or top of (to the left or bottom of) π_i. If $\sigma_{j,i}=\sigma_{j',i'}$ is a common side of π_i and π_i', $v\varepsilon H^1(\pi_i)\cap H^1(\pi_i')$, and $v_i(v_i')$ is the restriction of v to $\pi_i(\pi_i')$, we define

$$\left.\begin{array}{ll} \text{(a)} & \delta v_{\sigma_{j,i}} \equiv \rho(\sigma_{j,i})v_i+\rho(\sigma_{j',i'})v_i' \text{ for each interior side } \sigma_{j,i}=\sigma_{j',i'} \text{ and} \\ \text{(b)} & \delta v_{\sigma_{j,i}} \equiv \rho(\sigma_{j,i})v_i-\rho(\sigma_{j,i})g \text{ for each boundary side } \sigma_{j,i}. \end{array}\right\} \quad (3.4)$$

We next define some Sobolev and piecewise Sobolev spaces that are important in the variational formulation. Suppose $B=D$. By $H^m(B)$, we denote the space of functions v such that

$$||v||^2_{m(B)} \equiv \sum_{|\alpha|\leq m} ||D^\alpha v||^2_{L^2(B)} < \infty,$$

where m is a non-negative integer and D^α denotes a derivative of order $|\alpha|$. Let $H_0^1(D)$ denote the closure of $C_0^\infty(D)$ with respect to the norm, $||\ ||_{H^1(D)}$. We define

$$H^m_h \equiv \{v\varepsilon L^2(D): ||v||^2_{H^m_h} \equiv \pi_i\Sigma_{\varepsilon\Pi_N} ||v_i||^2_{H^m(\pi_i)} < \infty\}.$$

We also define

$$|v|^2_{H^m_h} \equiv \sum_{i=1}^{N} |v|^2_{H^m(\pi_i)} \text{ for each } v\varepsilon H^m_h,$$

where the seminorm, $|\ |_{H^m(\pi_i)}$, is defined by

$$|v|^2_{H^m(\pi_i)} \equiv \sum_{|\alpha|=m} ||D^\alpha v||^2_{L^2(\pi_i)} \quad \text{for each } v\in H^m(\pi_i).$$

Finally, set

$$H_K^{1^h} \equiv \{v\in H^{1^h} : (-\Delta-(K^2+i\delta K))v_i=0 \text{ for each } \pi_i\in\Pi_N\}.$$

We now define the subspace $H^E\subset H_K^{1^h}$ by

$$\left\{ H^E\equiv v\in H_K^{1^h} : v \text{ has continuous normal derivatives on } \partial\pi_i \text{ for each } \pi_i\in\Pi_N. \right\}$$

Furthermore, we define the following bilinear form:

$$A_K^h(v,w)\equiv\sum_{i=1}^{N}\int_{\pi_i}(\nabla v.\nabla w^*-(K^2+i\delta K)vw^*)d\bar{x} \quad \forall v,w\in H^E, \tag{3.5}$$

where w^* denotes the complex conjugate of w. It is easily seen that the solution, u, of (3.1) satisfies the following variational problem:

$$\left.\begin{aligned} &\text{Find } u\in H^E \text{ such that} \\ &A_K^h(u,v)=\Gamma_u(v)\equiv\sum_{\sigma'_{j,i}}\oint_{\sigma'_{j,i}} g\frac{\partial v^*}{\partial n}ds \text{ for each } v\in H^E, \end{aligned}\right\} \text{(VP)}$$

where the summation is taken over all element sides, $\sigma'_{j,i}$, contained in ∂D, ds denotes arc length, and $\frac{\partial}{\partial n}$ denotes the outward directed normal derivative to ∂D.

We discretize (VP) by defining a finite dimensional subspace, $S_K^h \subset H^E$, as follows. Let Λ^h denote the functions, ψ^h, defined on the element sides $\sigma_{j,i}$ such that ψ^h is some constant, $c_{j,i}$ on $\sigma_{j,i}$. For each ψ^h in Λ^h, let v^h in $H_K^{1^h}$ satisfy $\frac{\partial v^h}{\partial n_j} = \rho(\sigma_{j,i})\psi^h$ on each side $\sigma_{j,i}$, where $\frac{\partial}{\partial n_j}$ denotes the outward directed normal derivative to $\sigma_{j,i} = \sigma_j(\pi_i)$ from π_i. Hence v^h is the solution of a well-posed Neumann problem in each element. Let S_K^h consist of all such functions v^h. By the construction of S_K^h, we see that $S_K^h \subset H^E$. We now formulate our discrete variational problem.

$$\left.\begin{array}{l} \text{Find } u^h \in S_K^h \text{ such that} \\ \\ A_K^h(u^h, v^h) = \Gamma_u(v^h) \quad \forall v^h \in S_K^h. \end{array}\right\} \quad \text{(DVP)}$$

Note that the weak element and finite element methods are based on complementary variational principles in the sense that essential boundary or interface conditions for one are natural conditions for the other.

We next show that (DVP) is equivalent to the weak element method described in the previous section. Suppose that u^h satisfies (2.7) with local basis functions given by (3.2). In view of the definitions of P and B corresponding to problem (3.1), it follows from (2.7) and (3.4) that

$$\left.\begin{array}{l} \text{(a)} \quad \oint_{\sigma_{j,i}} \delta\left(\frac{\partial u^h}{\partial n_j}\right)_{\sigma_{j,i}} ds = 0 \text{ for each interior side } \sigma_{j,i} \\ \text{and} \\ \text{(b)} \quad \oint_{\sigma_{j,i}} \delta u^h_{\sigma_{j,i}} ds = 0 \text{ for each side } \sigma_{j,i}. \end{array}\right\} \quad (3.6)$$

The local basis coefficients, $\bar{a}(\pi_i)=(a_1(\pi_i),\ldots,a_4(\pi_i))$, are determined in each π_i so that (3.6) holds. It follows from (3.2) that $\frac{\partial u^h}{\partial n_j}$ is constant on each σ_j. Hence it follows from (3.2) and (3.6)(a) that $u^h \epsilon S_K^h$. Furthermore, it is seen from (3.5), (3.6)(b), and integration by parts that (DVP) holds. Conversely, it is easily seen that if u^h satisfies (DVP), then (3.6) and consequently (2.7) (with w replaced by u^h) are also satisfied. The basis coefficients $\bar{a}(\pi_i)$ may be readily obtained as in [3].

It thus suffices to prove that (DVP) is well-posed and to estimate the error, $e^h=u-u^h$, where u and u^h satisfy (VP) and (DVP), respectively. To this end we first state the following result.

Lemma 3.1: Suppose that ψ satisfies the following boundary value problem:

$$(-\Delta-(K^2-i\delta K))\psi=z \text{ in } D,\ \psi=0 \text{ on } \partial D. \tag{3.7}$$

Then

$$||\psi||_{H^2(D)} \leq CK||z||_{L^2(D)}, \tag{3.8}$$

where C is independent of K and z.

Note that we shall often use the same letter C to denote different constants when there is no danger of confusion. Lemma 3.1 was established in [8] and shows how the norm of the resolvent operator for (3.7) depends on K. This Lemma was also established with the Dirichlet condition replaced by a radiation boundary condition on part of the boundary. In such cases K is replaced by $K^{2-\alpha}$ with $0\leq\alpha\leq 1$, where α depends on various factors. See [8] for a more complete discussion of these issues.

We are now ready to prove our error estimates. We first prove the following Lemma using a duality argument as in [4].

Lemma 3.2: Suppose that u satisfies (VP), u^h satisfies (DVP), and $e^h=u-u^h$. Then there exists a constant, C, independent of K and h such that

$$||e^h||_{L^2(D)} \leq CKh|e^h|_{H^1 h}.$$

Proof: First observe that

$$||e^h||_{L^2(D)} = \sup_{z\in C^\infty(D)} \frac{|\int_D e^h z^* d\bar{x}|}{||z||_{L^2(D)}}. \tag{3.9}$$

Let $\psi\in H_0^1(D)\cap C^\infty(D)$ denote the solution of (3.7). For each vertical (horizontal) element side, $\sigma_{j,i}$, let $\frac{\partial}{\partial n}$ denote $\frac{\partial}{\partial x_1}(\frac{\partial}{\partial x_2})$. It follows from (3.4), (3.5), (3.7), and integration by parts that

$$\left.\begin{aligned} &\int_D e^h z^* d\bar{x} = \int_D e^h(-\Delta-K'^2)\psi^* d\bar{x} \\ &= -\sum_{i,j} \oint_{\sigma_{j,i}} \delta e^h_{\sigma_{j,i}} \frac{\partial \psi^*}{\partial n} ds + A_K^h(e^h,\psi). \end{aligned}\right\} \tag{3.10}$$

Since $e^h\in H^E\subset H_K^{1\,h}$, it follows readily using (3.5) and integration by parts that

$$A_K^h(e^h,\psi)=\sum_{i=1}^N \int_{\pi_i}(-\Delta-K'^2)e^h\psi^* d\bar{x}+\sum_{i,j}\oint_{\sigma_{j,i}} \delta(\frac{\partial e^h}{\partial n})_{\sigma_{j,i}}\psi^* ds=0.$$

Combining this with (3.10), we deduce

$$\int_\Omega e^h z^* d\bar{x} = -\sum_{i,j}\oint_{\sigma_{j,i}} \delta e^h_{\sigma_{j,i}} \frac{\partial \psi^*}{\partial n} ds. \tag{3.11}$$

To estimate (3.11), we first divide each rectangular element, π_i, into two right triangles as shown in Figure 1 (with π and π' replaced by π_i and π_i'). For simplicity, consider the triangle containing sides $\sigma_{1,i}$ and $\sigma_{4,i}$. Denote this triangle by t_i and set $\psi_i=\psi|_{t_i}$. Set

$$b_{1,i}=-\frac{1}{|\sigma_{1,i}|}\oint_{\sigma_{1,i}}\frac{\partial\psi_i}{\partial x_1}dx_2 \text{ and } b_{4,i}=-\frac{1}{|\sigma_{4,i}|}\oint_{\sigma_{4,i}}\frac{\partial\psi_i}{\partial x_2}dx_1, \tag{3.12}$$

where $|\sigma_{j,i}|$, j=1 or 4, denotes the length of side $\sigma_{j,i}$. Define

$$\psi_i'\equiv\psi_i+b_{1,i}x_1+b_{4,i}x_2$$

and note that

$$\oint_{\sigma_{j,i}}\frac{\partial\psi_i'}{\partial n}ds=0, \quad j=1 \text{ or } 4. \tag{3.13}$$

Since $u\in C(D)$, it follows from (3.6)(b) that

$$\oint_{\sigma_{j,i}}\delta e^h_{\sigma_{j,i}}\,ds=0 \text{ for each side } \sigma_{j,i}. \tag{3.14}$$

Using (3.14) and a scaling argument as in [4], it may be seen that

$$\oint_{\sigma_{j,i}}|\delta e^h_{\sigma_{j,i}}|^2ds\leq Ch(|e^h|^2_{H^1(\pi_i)}+|e^h|^2_{H^1(\pi_i')}), \tag{3.15}$$

where C is independent of h and π_i (and j=1 in Figure 1).

Note that $\frac{\partial\psi_i}{\partial n}-\frac{\partial\psi_i'}{\partial n}$ is constant on $\sigma_{j,i}$. Combining this with (3.14), we obtain

$$\oint_{\sigma_{j,i}} \delta e^h_{\sigma_{j,i}} \frac{\partial \psi_i^*}{\partial n} ds = \oint_{\sigma_{j,i}} \delta e^h_{\sigma_{j,i}} \frac{\partial \psi_i'^*}{\partial n} ds. \tag{3.16}$$

It follows from Schwarz' inequality and (3.15) that

$$\left.\begin{aligned} &\left|\oint_{\sigma_{j,i}} \delta e^h_{\sigma_{j,i}} \frac{\partial \psi_i'^*}{\partial n} ds\right| \leq \left\| \delta e^h_{\sigma_{j,i}} \right\|_{L^2(\sigma_{j,i})} \left\| \frac{\partial \psi_i'}{\partial n} \right\|_{L^2(\sigma_{j,i})} \\ &\leq Ch^{1/2} \left(|e^h|_{H^1(\pi_i)} + |e^h|_{H^1(\pi_i')} \right) \left\| \frac{\partial \psi_i'}{\partial n} \right\|_{L^2(\sigma_{j,i})}. \end{aligned}\right\} \tag{3.17}$$

Using an argument in [4] (see Lemma 2.2.6) and (3.13), we deduce

$$\left\| \frac{\partial \psi_i'}{\partial n} \right\|_{L^2(\sigma_{j,i})} \leq Ch^{1/2} |\psi_i|_{H^2(t_i)} \leq Ch^{1/2} \|\psi\|_{H^2(\pi_i)}. \tag{3.18}$$

Combining (3.8) with (3.16)-(3.18) and the Schwarz inequality, we conclude that

$$\left| \sum_{i,j} \oint_{\sigma_{j,i}} \delta e^h_{\sigma_{j,i}} \frac{\partial \psi^*}{\partial n} ds \right| \leq CKh \|z\|_{L^2(D)} |e^h|_{H^1 h}. \tag{3.19}$$

Finally, we combine (3.9), (3.11) and (3.19) to complete the proof. Q.E.D.

We now prove our main result.

Theorem 3.1: Suppose that u satisfies (3.1) and $u \in C^2(D)$. Then for

$K^2 h$ sufficiently small, (3.20)

there exists a unique solution u^h satisfying (DVP) and

$$|e^h|^2_{H^1{}^h} \leq C \inf_{v^h \in S_K^h} (|u-v^h|^2_{H^1{}^h} + K^2||u-v^h||^2_{L^2(D)}). \tag{3.21}$$

Furthermore,

$$Kh|e^h|_{H^1{}^h} + ||e^h||_{L^2(D)} \leq CKh^2||u||_{W^2_\infty(D)}. \tag{3.22}$$

The constant, C, in (3.21) and (3.22) is independent of K, h, and u.

Proof: First, assume that u^h exists. Using (3.5), we immediately obtain

$$|e^h|^2_{H^1{}^h} \leq |A_K^h(e^h,e^h)| + |(K^2+i\delta K)|||e^h||^2_{L^2(D)}. \tag{3.23}$$

Employing Lemma 3.2 and condition (3.20), we may "kickback" the last term in (3.23) to obtain

$$|e^h|^2_{H^1{}^h} \leq C|A_K^h(e^h,e^h)|.$$

In view of (VP) and (DVP), the last estimate yields

$$|e^h|^2_{H^1{}^h} \leq C|A_K^h(e^h,u-v^h)| \text{ for each } v^h \text{ in } S_K^h. \tag{3.24}$$

It follows readily using (3.5) that

$$|A_K^h(e^h,u-v^h)| \leq |e^h|_{H^1{}^h}|u-v^h|_{H^1{}^h} + (K^2+i\delta K)||e^h||_{L^2(D)}||u-v^h||_{L^2(D)}$$

$$\leq \eta(|e^h|^2_{H^1{}^h} + K^2||e^h||^2_{L^2(D)}) + C_\eta(|u-v^h|^2_{H^1{}^h} + K^2||u-v^h||^2_{L^2(D)})$$

for arbitrarily small $\eta>0$. Combining this estimate with (3.24) and again applying Lemma 3.2, (3.20), and a "kickback" argument, we obtain (3.21).

We next prove (3.22). In view of (3.24), we construct a function v^h in S_K^h satisfying

$$|A_K^h(e^h,u-v^h)|\leq C_\eta h^2||u||^2_{W^2_\infty(D)}+\eta|e^h|^2_{H^1{}^h} \tag{3.25}$$

for $\eta>0$ arbitrarily small. Suppose π_i in Π_N has sides $\sigma_{j,i}$, $j=1,\ldots,4$. We define four constants $c_{1,i},\ldots,c_{4,i}$ such that

$$\oint_{\sigma_{j,i}} c_{j,i}ds=\oint_{\sigma_{j,i}}\rho(\sigma_{j,i})\frac{\partial u}{\partial n_i}ds, j=1,\ldots,4, \tag{3.26}$$

where $\bar{n}_i$ is the outward directed unit normal to $\partial\pi_i$. It is easily seen using the Mean Value Theorem for integrals that for each $j=1,\ldots,4$, we have

$$c_{j,i}=\rho(\sigma_{j,i})\frac{\partial u}{\partial n_i}(P_{j,i})$$

for some point $P_{j,i}\varepsilon\sigma_{j,i}$. It follows readily using the Mean Value Theorem for derivatives that

$$||\frac{\partial u}{\partial n_i}-\rho(\sigma_{j,i})c_{j,i}||_{L^2(\sigma_{j,i})}\leq Ch(\oint_{\sigma_{j,i}}||\frac{\partial}{\partial s}\frac{\partial u}{\partial n_i}||^2_{L^\infty(\sigma_{j,i})}ds)^{1/2}, \tag{3.27}$$

$j=1,\ldots,4$. We now define ψ^h in Λ^h by $\psi^h|_{\sigma_{j,i}}=c_{j,i}$ for each side $\sigma_{j,i}$.

It follows immediately from (3.26) and (3.27) that for each $\sigma_{j,i}$, we have

$$\left.\begin{aligned} &\text{(a)}\quad \oint_{\sigma_{j,i}} \frac{\partial u}{\partial n_i} ds = \oint_{\sigma_{j,i}} \rho(\sigma_{j,i})\psi^h ds \text{ and} \\ &\text{(b)}\quad \left|\left|\frac{\partial u}{\partial n_i} - \rho(\sigma_{j,i})\psi^h\right|\right|_{L^2(\sigma_{j,i})} \leq Ch^{\frac{3}{2}} ||u||_{W^2_\infty(D)}. \end{aligned}\right\} \quad (3.28)$$

To construct v^h satisfying (3.25), we solve the Neumann problem for equation (3.1)(a) in each π_i with Neumann data given by $\rho(\sigma_{j,i})\psi^h$ on each side $\sigma_{j,i}$, $j=1,\ldots,4$. We denote the solution by v_i. Set $v^h=v_i$ in each π_i and note that $v^h \epsilon S_K^h$. We now employ integration by parts to deduce

$$A_K^h(e^h, u-v^h) = \sum_I \oint_{\partial\pi_i} e^h \frac{\partial}{\partial n_i}(u-v^h)^* ds. \quad (3.29)$$

It follows immediately from the construction of v^h and (3.28)(a) that

$$\oint_{\partial\pi_i} \frac{\partial}{\partial n_i}(u-v^h)^* ds = 0 \text{ for each } \pi_i.$$

In view of this, we may replace e^h on the right hand side of (3.29) by

$$e_i^h = e^h - c^i \text{ in } \pi_i \quad (3.30)(a)$$

with $c^i = \frac{1}{|\partial\pi_i|}\oint_{\partial\pi_i} e^h ds$, so that

$$\oint_{\partial\pi_i} e_i^h ds = 0 \text{ for each } \pi_i. \quad (3.30)(b)$$

Applying (3.28)(b), (3.29) (with e^h replaced by e_i^h on the right side), and the Schwarz inequality, we obtain

$$|A_K^h(e^h,u-v^h)| \leq \sum_{i=1}^{N} \sum_{j=1}^{4} ||e_i^h||_{L^2(\sigma_{j,i})} ||\frac{\partial u}{\partial n_i} - \rho(\sigma_{j,i})\psi^h||_{L^2(\sigma_{j,i})}$$

$$\leq Ch^{\frac{3}{2}} ||u||_{W_\infty^2(D)} \sum_{i=1}^{N} ||e_i^h||_{L^2(\partial\pi_i)}. \tag{3.31}$$

To estimate $||e_i^h||_{L^2(\partial\pi_i)}$, we map π_i onto the unit square, π, and employ the following well-known estimate:

$$||w||^2_{H^1(\pi)} \leq C(|w|^2_{H^1(\pi)} + |\oint_{\partial\pi} w ds|^2) \text{ for each } w \text{ in } H^1(\pi).$$

As in [4], we combine this estimate with (3.30)(b) and map π back onto π_i to obtain

$$||e_i^h||_{L^2(\partial\pi_i)} \leq Ch^{1/2} |e_i^h|_{H^1(\pi_i)} = Ch^{1/2} |e^h|_{H^1(\pi_i)}, \tag{3.32}$$

using (3.30)(a) in the last step. We now combine (3.31) and (3.32) to conclude that for arbitrarily small $\eta>0$:

$$\left.\begin{aligned} |A_K^h(e^h,u-v^h)| &\leq Ch^2 \sum_{i=1}^{N} ||u||_{W_\infty^2(D)} |e^h|_{H^1(\pi_i)} \\ &\leq \sum_{i=1}^{N} (C_\eta h^4 ||u||^2_{W_\infty^2(D)} + \eta |e^h|^2_{H^1(\pi_i)}). \end{aligned}\right\} \tag{3.33}$$

Estimate (3.25) now follows from (3.33) since $N=O(h^{-2})$. Combining (3.24) and (3.25) with $\eta>0$ sufficiently small, we deduce

$$|e^h|_{H^1h} \leq Ch||u||_{W_\infty^2(D)}. \tag{3.34}$$

Estimate (3.22) now follows from (3.34) and Lemma 3.2. Finally, to prove that (DVP) is well-posed, it suffices to prove uniqueness since S_K^h is finite dimensional. If $g \equiv 0$ in D, then $u \equiv 0$ in D since (3.1) is well-posed. Hence it follows from (3.22) that $u^h \equiv 0$ in D and we have proved that (DVP) has a unique solution. Q.E.D.

Remark 3.1: Typically, the solution of (3.1) satisfies $||u||_{W_\infty^2(D)} = 0(K^2)$.

Hence it follows from (3.22) that $||e^h||_{L^2(D)} = 0(K^3h^2)$. This is analogous to results obtained for standard second order finite element or finite difference schemes (see [8]). However, it follows from (3.21) that the weak element method is clearly superior for moderate to large values of K when the oscillatory behavior of the solution is well approximated in each element by the local basis functions, assuming the "stability" constraint (3.20) holds. This constraint also occurs in connection with standard discretization schemes. Our numerical results indicate that this stability constraint does not cause serious computational problems for the weak element method when the oscillatory behavior of the solution is well-approximated by functions in S_K^h. We shall see in Section 4 that in such cases the discretization error is quite small even when K^2h is large.

Remark 3.2: It follows from the previous remark that the main computational advantage of the weak element method occurs when the oscillatory behavior of the solution is well-approximated by functions in S_K^h. The determination of this oscillatory behavior can be difficult for realistic physical models. This question was investigated in [1] and [2]. In [1], asymptotic methods were employed in connection with a one-dimensional scattering problem. In [2], multi-dimensional models were

treated for which it is known that most of the propagation occurs in a narrow angle band about a fixed direction. This condition is closely related to the "paraxial approximation" and holds in a variety of application areas.

Remark 3.3: The weak element method described in Section 2 can be extended in various ways (see Remark 2.1). Alternatively, the variational formulation described in this section can be generalized by employing higher order approximating subspaces S_K^h. See [4] for a detailed discussion of this in connection with the Laplace equation. Furthermore, more general boundary value problems can be treated than (3.1). This includes more general domains, variable coefficients, and radiation boundary conditions. We intend to investigate some of these questions in the future.

4. NUMERICAL ANALYSIS

In this section, we demonstrate the effectiveness of the weak element method described in Section 1 for simple two-dimensional test problems whose solutions are known in closed form. Our measure of error is given by

$$E_2 \equiv \frac{||u-u^h||_{\ell^2(D)}}{||u||_{\ell^2(D)}}$$

where $u(u^h)$ is the exact (approximate) solution, $||\ \ ||_{\ell^2(D)}$ denotes the discrete mean-square norm, and D is a rectangle in either Cartesian or polar coordinates. D is partitioned into rectangular elements as described in Section 2 such that the grid points are equally spaced in each direction. We denote the number of intervals in the x_1 and x_2 directions by N_1 and N_2, respectively. The differential operator is given by (3.1)(a) with $\delta=0$.

Our boundary condition for the first two examples is the Dirichlet condition, (3.1)(b), although we have obtained analogous results for various combinations of Dirichlet, Neumann, and impedance boundary conditions. In Example 3, we consider the Helmholtz equation in polar coordinates in the exterior of the unit circle with a radiation boundary condition imposed on an artificial outer boundary. Our main purpose in all of these examples is to evaluate the discretization error for different values of N_1, N_2, and K. The calculations were performed on a CDC 7600 at Brookhaven National Laboratory. The system of equations were solved using a standard conjugate gradient iterative method applied to the normal equations as well as a direct solver based on Gaussian elimination. Both methods

resulted in essentially the same discretization errors. It is expected that more recently developed preconditioned iterative methods, such as that discussed in [7], would be considerably more efficient.

Example 1: For our first series of numerical experiments, we assume that D is the unit square and choose Dirichlet boundary conditions such that the solution is given by

$$u(x_1,x_2) = \sin K(x_1\cos\alpha + x_2 \sin\alpha),\ 0 \leq \alpha \leq \frac{\pi}{2}. \qquad (4.1)$$

We employ the weak element method with local basis functions given by (3.2) (with K'=K). Our results are demonstrated in Tables 1A - D with $N_1 = N_2 = N = h^{-1}$. We have also employed the five-point finite difference scheme in this and the following example although it is not necessary to demonstrate the results obtained using this scheme. It suffices to observe that, as expected, this finite difference scheme has convergence rate $O(h^2)$ as $h\downarrow 0$ with K fixed. (Our numerical results indicate that this is also the case for the weak element method.) Furthermore, the five-point scheme is not accurate when $Kh = \frac{K}{N} > 1$.

It is readily seen from (3.2) and (4.1) that when $\alpha=0$ or $\alpha=\frac{\pi}{2}$, the solution in each element may be expressed as a linear combination of local basis functions. Hence we would expect the weak element approximation to yield the exact solution, except for accumulated roundoff errors. This is validated in Table 1A for N=4 and various values of K. On the other hand, it follows from (3.2), (3.21), and (4.1) that the more α differs from 0 and $\frac{\pi}{2}$, the less effective the weak element method should be with this basis (see Remark 3.1). In Tables 1B - D, we consider various values of N and K with

$\alpha=\frac{\pi}{150}$, $\frac{\pi}{25}$, and $\frac{\pi}{8}$, respectively. We see from Table 1B that when $\alpha = \frac{\pi}{150}$, the weak element method is accurate even when Kh = 16. From Table 1C we see that for $\alpha = \frac{\pi}{25}$, the method yields accurate results when Kh = 2 and hence is more effective than the five-point finite difference scheme. On the other hand, when $\alpha = \frac{\pi}{8}$ we see from Table 1D that the method does not yield accurate results when Kh > 1. We have observed that in cases such as this for which the phase of the solution is not sufficiently well approximated, there is no advantage in using the weak element method instead of a standard discretization scheme.

Table 1A

(N=4)

K	$\alpha=0$ E_2	$\alpha = \frac{\pi}{2}$ E_2
1	3.9×10^{-13}	3.8×10^{-13}
2	1.7×10^{-13}	1.9×10^{-13}
4	3.8×10^{-13}	7.0×10^{-13}
8	2.9×10^{-14}	1.7×10^{-12}
16	1.2×10^{-13}	6.1×10^{-12}
32	3.4×10^{-14}	1.3×10^{-11}
64	5.2×10^{-13}	2.5×10^{-11}
128	6.8×10^{-14}	4.9×10^{-11}
256	2.1×10^{-14}	9.9×10^{-11}
512	1.9×10^{-12}	2.0×10^{-10}
1024	1.5×10^{-13}	4.0×10^{-10}
2048	1.2×10^{-13}	8.1×10^{-10}

Table 1B

($\alpha=\frac{\pi}{150}$)

K	N	E_2
1	4	5.8×10^{-5}
1	8	1.6×10^{-5}
1	16	5.2×10^{-6}
2	4	2.3×10^{-4}
2	8	6.2×10^{-5}
2	16	2.0×10^{-5}
4	4	8.7×10^{-4}
4	8	2.5×10^{-4}
4	16	8.2×10^{-5}
8	4	3.5×10^{-3}
8	8	9.9×10^{-4}
8	16	3.3×10^{-4}
16	4	1.2×10^{-2}
16	8	4.3×10^{-3}
16	16	3.6×10^{-3}
32	4	2.6×10^{-2}
32	8	1.3×10^{-2}
32	16	5.4×10^{-3}
64	4	4.9×10^{-2}
64	8	2.7×10^{-2}
64	16	1.5×10^{-2}
128	4	9.7×10^{-2}
128	8	6.9×10^{-2}
128	16	5.1×10^{-2}
256	4	1.9×10^{-1}
256	8	1.0×10^{-1}
256	16	5.6×10^{-2}

Table 1C

($\alpha=\frac{\pi}{25}$)

K	N	E_2
1	4	3.7×10^{-4}
1	8	1.0×10^{-4}
1	16	3.2×10^{-5}
2	4	1.4×10^{-3}
2	8	3.9×10^{-4}
2	16	1.3×10^{-4}
4	4	6.0×10^{-3}
4	8	1.7×10^{-3}
4	16	5.6×10^{-4}
8	4	2.3×10^{-2}
8	8	6.5×10^{-3}
8	16	2.1×10^{-3}
16	4	1.1×10^{-1}
16	8	8.6×10^{-2}
16	16	1.2×10^{-1}
32	4	1.8×10^{-1}
32	8	1.2×10^{-1}
32	16	5.9×10^{-2}
64	4	3.1×10^{-1}
64	8	1.8×10^{-1}
64	16	1.4×10^{-1}

Table 1D

($\alpha=\frac{\pi}{8}$)

K	N	E_2
1	4	2.1×10^{-3}
1	8	5.6×10^{-4}
1	16	1.8×10^{-4}
2	4	8.2×10^{-3}
2	8	2.2×10^{-3}
2	16	7.5×10^{-4}
4	4	5.3×10^{-2}
4	8	1.3×10^{-2}
4	16	3.9×10^{-3}
8	4	1.0
8	8	6.9×10^{-2}
8	16	1.9×10^{-2}
16	4	1.3
16	8	1.2
16	16	1.5

Example 2: For our next class of problems, we consider solutions of the form

$$u(x_1,x_2)=\sin Lx_1 \cos\sqrt{K^2-L^2}x_2, L=1,2,\ldots, \tag{4.2}$$

with Dirichlet boundary conditions on the unit square. Our local basis for the weak element method is again given by (3.2). For $Kh<<1$, the convergence rate is again $O(h^2)$. However, for $Kh>>1$, the weak

element method behaves differently for this problem than for the previous example. The reason for this is that the x_1 dependence of u is independent of K. Suppose that K>>L in (4.2), so that $u(x_1,x_2)\tilde{\sim}\sin Lx_1 \cos Kx_2$. Hence the x_2-dependence of u may be reproduced almost exactly by the basis functions for K large and the accuracy will be almost independent of the number of x_2 grid points. We are thus left with approximating $\sin Lx_1$ by constants, yielding an $O(N_1^{-1})$ order approximation to u that is independent of K for large K. We illustrate typical results in Table 2 for L=1, K=128, and various values of N_1 and N_2.

Table 2

(L=1, K=128)

N_1	N_2	E_2
8	1	9.5×10^{-2}
16	1	4.9×10^{-2}
32	1	2.5×10^{-2}
64	1	1.3×10^{-2}
128	1	6.4×10^{-3}
8	2	9.3×10^{-2}
16	2	4.8×10^{-2}
32	2	2.4×10^{-2}
64	2	1.2×10^{-2}
128	2	6.1×10^{-3}
8	4	9.3×10^{-2}
16	4	4.8×10^{-2}
32	4	2.4×10^{-2}
64	4	1.2×10^{-2}
128	4	5.8×10^{-3}

Example 3: The final problem we consider is one treated in [9] using an integral equation approach combined with a finite difference method for K not too large. Introducing polar coordinates, (r,θ), the problem consists of solving the Helmholtz equation, $(\Delta+K^2)u=0$ in the exterior of the unit circle, subject to the boundary conditions $u(\bar{x})=x^2$ on $r=1$ and the radiation condition

$$(\partial u/\partial r)-iKu=o(r^{-1/2}) \tag{4.3}$$

for r large. The solution of this problem is given by

$$u(\bar{x})=\frac{x^2}{r}\,\frac{H_1^{(1)}(Kr)}{H_1^{(1)}(K)} = \frac{\sin\theta H_1^{(1)}(Kr)}{H_1^{(1)}(K)}, \tag{4.4}$$

where $H_1^{(1)}$ denotes the Hankel function of first kind and order 1.

In order to determine the discretization error due to applying the weak element method to this problem, we replace the right-hand side of the radiation condition (4.3) by the exact value obtained by applying $(\partial/\partial r)-iK$ to (4.4) on a circle of radius $R>1$ and denote this function by $g(R,\theta)$. Employing polar coordinates and appropriate symmetry conditions on $u(\bar{x})$, we obtain the following boundary value problem for $u(\bar{x})$:

$$\left.\begin{aligned} &\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right)+\frac{1}{r}\,\frac{\partial^2 u}{\partial\theta^2}+K^2ru=0 \text{ in } D,\\ &\qquad u=\sin\theta \text{ on } r=1,\\ &(\partial u/\partial r)-iKu=g(R,\theta) \text{ on } r=R,\\ &\qquad (\partial u/\partial\theta)=0 \text{ on } \theta=\pi/2 \text{ and}\\ &\qquad u=0 \text{ on } \theta=0, \end{aligned}\right\} \tag{4.5}$$

where D denotes the domain $1<r<R, 0<\theta<\pi/2$.

Problem (4.5) may be placed in the general framework of (2.1) by replacing the (x_1,x_2) coordinates by (r,θ) coordinates, so that

$\nabla=(\partial/\partial r,\partial/\partial\theta)$. Hence the domain D is a rectangle, the matrix P is given by

$$P=\begin{pmatrix} r & 0 \\ 0 & 1/r \end{pmatrix},$$

and $q=-K^2 r$ in (2.1). Since $(\nabla P)_o=(1,0)\neq\bar{0}$, we simplify the problem by making the transformation

$$u(r,\theta)=r^{-1/2}v(r,\theta).$$

We now obtain the following boundary value problem for v:

$$\left.\begin{aligned} &(-\nabla\cdot P'\nabla+q')v=0 \text{ in } D,\\ &v=\sin\theta \text{ on } r=1,\\ &\frac{\partial v}{\partial r}-(iK+\frac{1}{2R})v=R^{1/2}g \text{ on } r=R,\\ &\frac{\partial v}{\partial\theta}=0 \text{ on } \theta=\frac{\pi}{2} \text{ and}\\ &v=0 \text{ on } \theta=0, \end{aligned}\right\} \qquad (4.6)$$

where

$$P'=\begin{pmatrix} 1 & 0 \\ 0 & r^{-2} \end{pmatrix} \text{ and } q'=-(K^2+\frac{1}{4r^2}).$$

Hence $(\nabla P')_o=\bar{0}$. Using (4.4), we see that the solution is given by

$$v(r,\theta)=r^{1/2}u(r,\theta)=\frac{r^{1/2}\sin\theta H_1^{(1)}(Kr)}{H_1^{(1)}(K)}. \qquad (4.7)$$

We now apply (2.8) and (2.9) to obtain the following local basis functions:

$$e^{\pm i(K^2+1/4r_o^2)^{1/2}(r-r_o)},\ e^{\pm ir_o(K^2+1/4r_o^2)^{1/2}(\theta-\theta_o)}. \qquad (4.8)$$

We also note the following asymptotic representation of $H_1^{(1)}(Kr)$ (see [10]):

$$H_1^{(1)}(Kr) \tilde{\sim} \left(\frac{2}{\pi Kr}\right)^{1/2} e^{i(Kr-3\pi/4)} \sum_{j=0}^{J-1} \frac{(-1)^j \Gamma(j+\frac{3}{2})}{j!(2iKr)^j \Gamma(-j+\frac{3}{2})} \qquad (4.9)$$

for Kr large and $J \geq 1$, where Γ denotes the gamma function.

If we compare (4.7)-(4.9), we see that the r-dependence of $v(r,\theta)$ is accurately reproduced by the basis functions for Kr large. This is analogous to the situation in Example 2. In Table 3A, we examine the error, E_2, for R=2 and different values of K and $N=N_1=N_2$, where $N_1(N_2)$ is the number of subintervals in the $r(\theta)$ direction. For $\frac{K}{N}$ small, the weak element method again behaves analogously to the five-point finite difference scheme. On the other hand, E_2 is nearly constant for $\frac{K}{N}$ large and N fixed as K increases.

Furthermore, we have observed that for larger values of R the errors are about the same as for R=2 when $\frac{K}{N}$ is large. When $\frac{K}{N}$ is small, accuracy is destroyed by the coarse grid in the r-direction. This can be remedied by using a graded mesh in which the r-grid sizes are systematically increased as r increases (see [11]). We illustrate the high frequency behavior in Table 3B, where K=R=128. In this case the grid sizes in the r-direction are quite large. We observe that E_2 is essentially constant when N_2 (the number of intervals in the θ-direction) is fixed and N_1 varies. Furthermore, the error with respect to θ is of order $O(N_2^{-1})$. The explanation of these numerical results is the same as that given in example 2 (i.e., the θ-dependence of $v(r,\theta)$ is approximated locally by constants).

Table 3A

(R=2)

K	N	E_2
1	8	8.9×10^{-3}
1	16	3.7×10^{-3}
2	8	1.2×10^{-2}
2	16	3.9×10^{-3}
4	8	2.1×10^{-2}
4	16	5.9×10^{-3}
8	8	4.0×10^{-2}
8	16	1.1×10^{-2}
16	8	2.6
16	16	2.1×10^{-2}
32	8	1.7×10^{-1}
32	16	4.6
64	8	8.0×10^{-2}
64	16	9.2×10^{-2}
128	8	9.3×10^{-2}
128	16	4.2×10^{-2}
256	8	8.9×10^{-2}
256	16	4.7×10^{-2}
512	8	1.0×10^{-1}
512	16	4.6×10^{-2}
1024	8	9.5×10^{-2}
1024	16	4.7×10^{-2}
2048	8	8.8×10^{-2}
2048	16	5.2×10^{-2}

Table 3B

(R=128, K=128)

N_1	N_2	E_2
1	8	9.1×10^{-2}
2	8	9.1×10^{-2}
4	8	9.3×10^{-2}
8	8	9.1×10^{-2}
1	16	4.7×10^{-2}
2	16	5.0×10^{-2}
4	16	4.7×10^{-2}
8	16	4.7×10^{-2}
1	32	3.1×10^{-2}
2	32	2.4×10^{-2}
4	32	2.4×10^{-2}
8	32	2.4×10^{-2}
1	64	1.3×10^{-2}
2	64	1.3×10^{-2}
4	64	1.3×10^{-2}
8	64	1.3×10^{-2}

5. CONCLUSIONS AND COMMENTS

We have analyzed and tested a version of the weak element method developed in [3] in connection with the Helmholtz equation. Mathematical models of this kind occur in various scattering and diffraction problems. Standard discretization schemes based on finite difference, finite element, or integral equation methods are not well suited for these problems when the wave number, K, is not small, since piecewise polynomials are not good approximations to the oscillatory solution. On the other hand, the weak element method is based on piecewise exponentials that satisfy a localized approximation to the differential equation.

We have proved that the particular weak element method outlined in Section 2 with mesh size h has a convergence rate of order $O(h^2)$ as $h \to 0$ for fixed K. Our analytic results also indicate that the method yields a good approximation to the solution, u, when the oscillatory behavior of u is well approximated by the local basis functions. The proof is based on a complementary variational principle developed in [4] in connection with the Laplace equation. It is expected that this proof can be extended to more general boundary value problems and higher order weak element methods.

We have also validated our theoretical results with respect to test problems for which the solution is known in closed form. We have seen from these examples that the weak element method offers no advantage in general compared to the five-point finite difference scheme. However, our theoretical and numerical results demonstrate that the weak element is considerably superior for moderate to large K

when the oscillatory behavior of the solution is adequately approximated locally. For general variable coefficient problems, this oscillatory behavior will vary in different parts of the domain. Consequently, an important practical area of investigation is the development of methods for determining locally the approximate oscillatory behavior of the solution for large K. This was done in [1] in connection with a one-dimensional scattering problem using asymptotic methods. This was also done in [2] for multi-dimensional propagation models for which most of the propagation occurs in a narrow angle band about a fixed direction. The use of error estimators and adaptive discretization methods might also be useful in determining appropriate local basis functions.

ACKNOWLEDGMENTS

The author wishes to express his gratitude to Dr. M. E. Rose for several stimulating discussions during the course of this work. The author is also grateful to H. Berry for his help in the preparation of the numerical results in Section 4.

REFERENCES

[1] A. K. Aziz, R. B. Kellogg, and A. B. Stephens, "A two point boundary value problem with a rapidly oscillating solution," to appear.

[2] C. I. Goldstein, "Finite element methods applied to nearly one-way propagation," J. Comp. Phys., to appear.

[3] M. E. Rose, "Weak element approximations to elliptic differential equations," Numer. Math., 24, 185-204, 1975.

[4] I. Babuska, "The method of weak elements," Tech. Note BN-809, Inst. Fl. Dyn. and Appl. Math., U. Maryland, 1974.

[5] J. Greenstadt, "Cell discretization," in Conf. on Appl. of Num. Anal. Lecture Notes #288, Springer, Berlin, 1971.

[6] C. I. Goldstein and H. Berry, "A numerical study of the weak element method applied to the Helmholtz equation," BNL Report No. 50746, 1977.

[7] A. Bayliss, C. I. Goldstein, and E. Turkel, "The numerical solution of the Helmholtz equation for wave propagation problems in underwater acoustics," Comp. and Math. with Appl., 11, No. 718, 655-665, 1985.

[8] A. Bayliss, C. I. Goldstein, and E. Turkel, "On accuracy conditions for the numerical computation of waves," J. Comp. Phys., 59, 396-404, 1985.

[9] D. Greenspan and P. Werner, "A numerical method for the exterior Dirichlet problem for the reduced wave equation," Arch. Rat. Mech. Anal., 23, 288-316, 1966.

[10] I. S. Gradshteyn and I. M. Ryzik, Table of Integrals, Series, and Products, Academic Press, New York and London, 1965.

[11] C. I. Goldstein, "The finite element method with nonuniform mesh sizes applied to the exterior Helmholtz problem," Numer. Math., 38, 61-82, 1981.

THE LOCAL REDISTRIBUTION OF POINTS ALONG CURVES FOR NUMERICAL GRID GENERATION

Peter R. Eiseman
Department of Applied Physics and Nuclear Engineering
Columbia University
New York, NY 10027

ABSTRACT

A methodology is established to cluster points along curves in a manner which does not change the existing pointwise distribution outside of a specified region containing the cluster. In each instance, points are pulled from the perimeters of the region towards the cluster center. The result is a smooth expansion from each end followed by a compression into the center. Altogether, this represents a local redistribution of points which can be employed either interactively or automatically.

This work was supported by the US Air Force Grant AFOSR-82-0176B and the NASA Langley Research Center Grants NAG1-479 and NAG1-427.

INTRODUCTION

When a pointwise distribution along a curve is acceptable everywhere except in certain local regions, the capability to redistribute points only in those regions becomes important. Our objective, here, is to create a framework from which methods for the desired local redistribution can be developed. This is done by forming elementary operations which are then applied in succession. Each operation smoothly forms a single cluster about a point by attracting only nearby points: the pointwise locations beyond a specified distance on either side remain unchanged. As such, the action occurs in an interval where both the endpoints and the internal cluster point remain fixed: the other points move in from each side while maintaining a globally smooth variation in pointwise spacing.

Upon application, a new pointwise distribution is created from an old one and differs from it only in the chosen interval. The old or "previous" distribution is always viewed as a mapping from a uniformly distributed independent variable to the curve. This variable is often referred to as just the existing parameterization for the curve. With a finite number of points, a uniformly spaced grid along the parametric interval is mapped onto a grid along the curve. The new distribution is simply constructed by composition whereby we first map a new uniformly distributed parametric interval onto the old one and then apply the old map to the result. In terms of grids, a uniformly spaced grid on the new parametric interval is mapped onto the old parametric interval to produce a distribution there which is non-uniform in some local subinterval. On that subinterval, the application of

the old mapping is accomplished with the aid of local interpolation. Off of that subinterval, the points coincide with the old parametric locations and no interpolation is required.

At each stage, there is a progression from old to new corresponding to the application of an elementary operation. As noted above, each such operation can be generated from a local reparameterization which is just a mapping between old and new parameters. The actual construction can be done in either forward or backward directions by the use of weight functions. The forward direction is from new to old while the backward is from old to new.

WEIGHTS AND TRANSFORMATIONS

With the view of larger masses pulling more strongly to a center of gravity, weights are most commonly thought of as being more strongly attractive when they are large. For the application to distribution functions, this view means that points are more strongly attracted to locations of large weight. In correspondence, the pointwise spacing must then shrink to adjust to a large weight. The most simple way to have this happen is to make the spacing vary in an inverse proportion to the weight. In terms of changing the spacing in the old parametric interval, we must then force the product of the weight and the desired spacings to be equal to a constant. When the same interval is taken for the old and new parameters, that constant is just the increment from a uniform spacing. In our development, we will always let s denote the old parameter and t denote the new parameter. In this notation and with our assumption of the same interval,

the forced condition is given by

$$dt = w\, ds \tag{1}$$

and is called an equidistribution of the weight w since equal amounts of weight must appear in each interval ds. For grids, the total weight between every pair of points is then the same.

To construct the elementary operations of local clustering, we recall the basic requirements that the cluster center and the interval endpoints must remain unchanged. For the weights, these requirements become integral statements; namely, that the integrals of wds and ds are the same over both of the intervals from the cluster center to the endpoints of the cluster region. Noting that uniform spacing would occur if w = 1, deviations therefrom are responsible for non-uniform spacing and can be represented as a function f which is added to the unity of uniformity in the weight to get w = 1+f. In terms of f, the preservation of cluster center and endpoints results when the integral of f vanishes over each of the two intervals above. To define these intervals in a clear way, a zero subscript will be employed for the center while a minus and a plus will be used as subscripts to indicate the endpoints in negative and positive directions, respectively, from the center. In this notation, the preservation condition means that new t and old s must satisfy $t_- = s_-$, $t_0 = s_0$, and $t_+ = s_+$ or that the function f which gives variations from uniformity must integrate to zero both from s_- to s_0 and from s_0 to s_+.

To obtain a maximal amount of control over shape, such a function is best created from a piecewise polynomial construction. The sim-

plest of these constructions is accomplished with two adjoining line segments for each of the two intervals. This is depicted in Fig. 1 where it is clearly evident that the first segment from either end

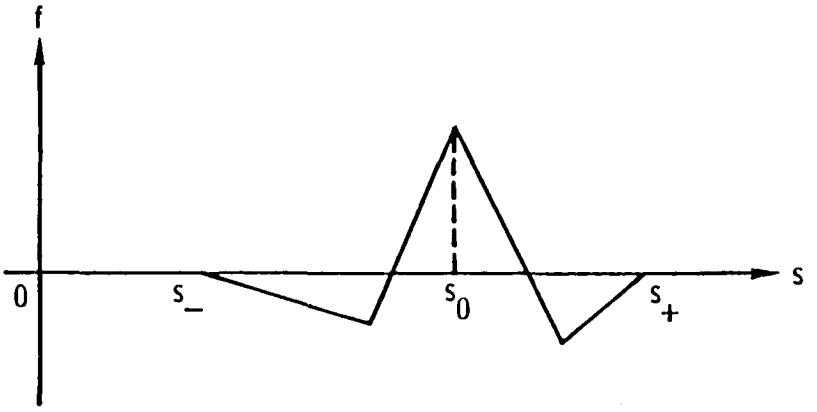


Figure 1: The function which must be added to unity to form a weight for Eq. 1

must lie below the axis to create negative areas which are enough to balance the positive area caused the linear rise to the positive value at the center. The center value determines the intensity of the clustering: the sum with unity gives the weight w_0, and hence, the minimum spacing dt/w_0. In compensation for the smallest spacing at the center, the spacing must first increase and then start to decrease. This starts from each endpoint spacing and linearly increases to a maximum at the end of each segment below the axis. Upon forming the weight w with an addition to unity, the maximum spacing on each side is given by dt/w with w evaluated at each corresponding end. Aside from the obvious limitation on the maximum spacing imposed by the

total interval length, there is the basic limitation that the weight w must be positive: negative weights flip the incremental intervals; thereby, causing a singularity in the mapping and a folded grid. As a consequence, there is then a limitation also on the intensity of clustering at the center. This is caused by the required balancing of positive and negative areas in Fig. 1.

Obeying the intensity limitation, the elementary clustering operation is obtained by a direct integration of Eq. 1 with our weight. The consequent mapping is then expressed with the new parameter t given as a monotone function of the old parameter s. In correspondence with the linear segments of the integrand, t is expressed as a piecewise quadratic function of s. To apply the mapping, a uniformly distributed t must produce the desired non-uniformities in s which in turn are sent to the curve by using the old curve mapping. This is just the composition of going from t to s and then to the curve. By construction, however, we go from s to t which is backward. This means that t(s) must be inverted to obtain s(t) which is forward rather than backward and thus can be used directly. Fortunately, in this piecewise quadratic case, the analytical inversion is possible and is somewhat simple. Since it contains radicals, it is not as simple as the original backward construction.

With the motivation towards more simplicity and higher levels of clustering intensity, we shall consider forward rather than backward constructions. To accomplish this, we must invert our thinking about weights, and thereby, have points attracted to locations where the weight is smaller rather than larger as would have been expected when

compared to the center of gravity shifts for masses. In terms of the piecewise-linear construction depicted in Fig. 1, the inversion results in a rigid reflection about the horizontal axis and a relabeling of that axis to be for the new parameter t in place of the old s. This is displayed in Fig. 2 and as earlier is added to unity to form the weight w = 1+f which is now used in

$$ds = w\, dt \tag{2}$$

For notational consistency, the new parameters $t_- = s_-$, $t_0 = s_0$, and

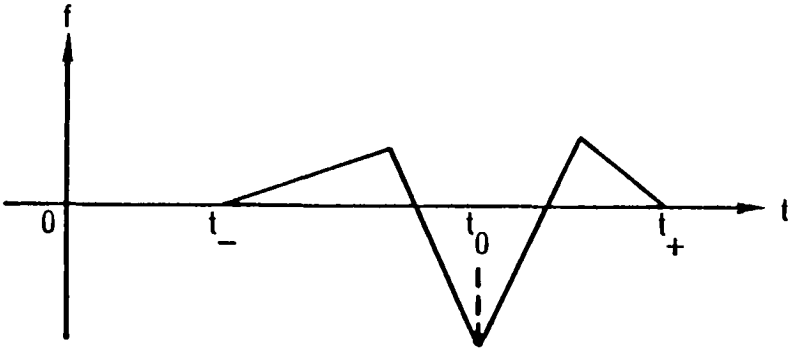


Figure 2: The function which must be added to unity to form a weight for the forward mapping with Eq. 2

$t_+ = s_+$ are used. The equalities also result from the rigidity of the reflection.

Rather than derive the algebraic formulation directly from t_- to t_0 and then from t_0 to t_+, we first re-examine the basic constraint which led to the equalities above; namely, that the integral of f over each interval vanishes. This constraint must be satisfied not only by the function displayed in Fig. 2 but also by any function which is to be employed for the same purpose. To begin our re-examination, we first note that the two integrals still vanish, if we rigidly translate the function along the t-axis. The translation is just the transformation from t to t+c from some c. Moreover, we also note that the vanishing is preserved under a constant dilation or contraction of either vertical or horizontal axes. These are just transformations which scale an axis by scalar multiplication. In the horizontal case, it is the transformation from t to at for some a. The effect of either transformation is to multiply the vanishing integral by a finite constant, and thereby, preserve the vanishing. In more formal terms, the constraint is invariant under the groups of transformations for translation and scaling. As a practical consequence, we can derive our algebraic formulation with standard conditions for height and interval and then apply the transformations to get the formulation for any other conditions that we wish. This also means that the same derivation can be used for the intervals on each side of t_0; and consequently, reduce the complexity of derivation by half. A further reduction comes from selecting the unit interval and a unit height since the arithmetic will be simplified.

With the unit lengths for our standard conditions, we are led to consider the function shown in Fig. 3 where the juncture point loca-

tion $x = \alpha$ is arbitrary. From a given downward unit $f_\alpha(1) = -1$ and the requirement for equal areas above and below the x-axis, we find that f_α must cross the x-axis at $1/(2-\alpha)$ and have a value of $1-\alpha$ at α. This function is then uniquely determined by the value at α which is also indicated in the figure. From the figure, the algebraic formulation is directly seen to be

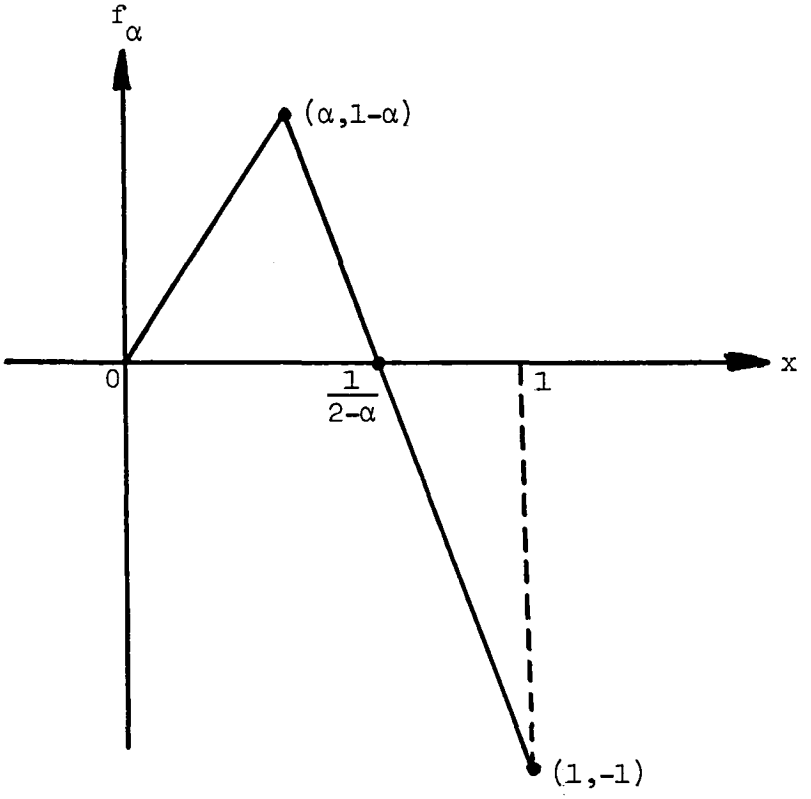


Figure 3: A standard function for the construction of piecewise linear weights

$$f_\alpha(x) = \left\{ \begin{array}{ll} (1-\alpha)\left(\frac{x}{\alpha}\right) & \text{for } 0 \leq x \leq \alpha \\ (1-\alpha) - (2-\alpha)\left(\frac{x-\alpha}{1-\alpha}\right) & \text{for } \alpha < x \leq 1 \\ 0 & \text{otherwise} \end{array} \right\} \qquad (3)$$

As a matter of interpretation, α represents the location of maximal spacing while the intersection point $1/(2-\alpha)$ represents the location where the spacing starts to decrease beyond the original uniform spacing. This means that the desired impact of clustering becomes significant only after the intersection point since we must first recover from the large spacing at α. Thus, $1/(2-\alpha)$ is the break-even point. As α varies through its possible range from 0 to 1, the break-even point varies from $1/2$ to 1. In order to provide a reasonably gentle transition into the smallest spacing, it is preferable to have a large region for the progression in spacing to occur. The largest possible region would have a length of $1/2$ and would occur when α vanishes. This, however, would leave no room for a smooth transition from endpoint spacing to the maximum spacing at α: a reasonably sized region is needed here for the same reasons as in the situation with the smallest spacing. Thus, a compromise is needed. As an example, we consider the case with $\alpha = \frac{1}{3}$. The transition into large spacing then occurs over a third of the length while the final compression after the break-even point occurs over the last 40% of the length. The corresponding function is then

$$f(x) = \left\{ \begin{array}{ll} 2x & \text{for } 0 \leq x < \frac{1}{3} \\ \frac{1}{2}(3-5x) & \text{for } \frac{1}{3} \leq x \leq 1 \\ 0 & \text{otherwise} \end{array} \right\} \qquad (4)$$

where for notational convenience, we have dropped the subscript of $\frac{1}{3}$ which would have been required from the specialization of Eq. 3.

With the function for $\alpha = \frac{1}{3}$, a weight for the forward mapping is given by

$$w = 1 + \beta\{f(\frac{t-t_-}{t_0-t_-}) + f(\frac{t_+-t}{t_+-t_0})\} \tag{5}$$

for $t \neq t_0$ and by $w = 1 - \beta$ for $t = t_0$. The special treatment of t_0 is required to remove a discontinuity caused by a contribution of -1 from f on each side when otherwise only one nonzero value would appear. The coefficient β is a control on the intensity of clustering. In a direct sense, the spacing at the center is scaled by $1-\beta$ to produce a smallest spacing in s. This spacing comes from Eq. 2 which gives $(1-\beta)dt$ at t_0. For an n-point grid, it becomes $(1-\beta)(t_{max}-t_{min})/(n-1)$. As β varies from 0 to 1, the minimum spacing varies from the original spacing down to 0. To avoid singularities, we do not go down to 0 but rather are interested in coming arbitrarily close to 0. Unlike the earlier backward mapping, there is no price for this arbitrary level of clustering intensity. This occurs because the compensating areas for grid expansion are now in the positive direction where there is no limit on size as there previously was when the axis itself was being approached.

By use of the weight of Eq. 5 in Eq. 2, we obtain the forward mapping

$$s = t + \beta\{(t_0-t_-)g(\frac{t-t_-}{t_0-t_-}) + (t_0-t_+)g(\frac{t-t_+}{t_0-t_+})\} \tag{6a}$$

where

$$g(x) = \begin{cases} x^2 & \text{for } 0 \le x < \frac{1}{3} \\ \frac{1}{4}(1-x)(5x-1) & \text{for } \frac{1}{3} \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \qquad (6b)$$

is the integral of f for increasing x. The interval lengths multiplying each application of g result from a change of variable in each corresponding integral. Geometrically, g appears as a simple bump which leaves the axis ($g(0) = 0$) with zero slope ($g'(0) = 0$), monotonically increases in the positive direction to reach a maximum, and then monotonically descends back to the axis ($g(1) = 0$) to enter with a negative slope ($g'(1) = -1$). When assembled in the transformation, β scales a combination of positive and negative bumps which are joined together with matching nonzero slopes. The addition to the line $s = t$ then represents a local distortion of it which causes clustering but which preserves the uniform spacing elsewhere. An illustration of the transformation is given in Fig. 4.

From a geometric viewpoint, we have simply taken the uniform transformation $s = t$ and have given it a local clockwise twist about t_0. The severity of the twist is controlled by the slope at t_0 and to some extent by the location of the maximum displacement from $s = t$. This location corresponds to the break-even point where the spacing from the transformation matches the uniform spacing from $s = t$; or in other words, where the two slopes match. In the case of Eq. 6, the choice of $\alpha = \frac{1}{3}$ led to a maximum displacement at $x = 0.6$. With the more general piecewise construction, it occurs at $x = 1/(2-\alpha)$ and

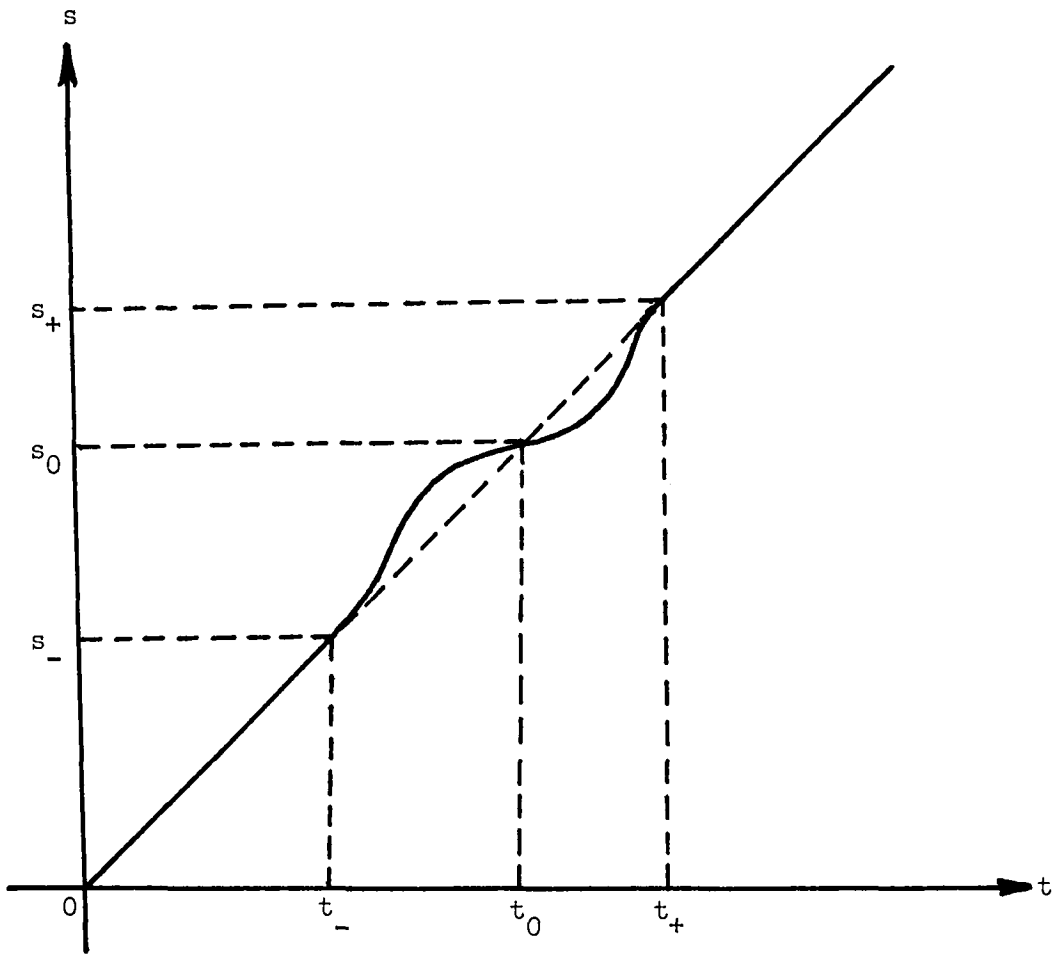


Figure 4: An elementary parameter transformation in the forward direction of new t going into old s

thus can be controlled with the choice of α. The analytical formulation is obtained by repeating the development with the f_α from Eq. 3 in place of f. Moreover, the locations can be controlled separately on each side of t_0 by using corresponding distinct α_1 and α_2. The analytical formulation is only altered by replacing the two applications of g in Eq. 6a with the corresponding generalizations g_{α_1} and g_{α_2} of Eq. 6b.

ALTERNATIVE FORMULATIONS

While further shape control over the forward transformation depicted in Fig. 4 can be exercised with more exotic piecewise constructions, we shall instead examine some alternatives which offer less shape control but which present attractive options because of their simplicity in statement. In this regard, we first note that the previous piecewise constructions achieved a high degree of algebraic simplicity at the expense of doing it in a number of successive defining intervals.

As a first step, we shall consider formulations which reduce the number of defining intervals. Returning to the unit interval on which we established f and then g, we shall consider a replacement. Noting that second- and first-order zeros for g are desired respectively at 0 and 1, we are brought to consider the simplest positive bump function

$$g_\alpha(x) = x^\alpha(1-x) \tag{7}$$

which satisfies the conditions when $\alpha > 1$ and which is defined by one segment. The derivative

$$f_\alpha(\alpha) = x^{\alpha-1}[\alpha - (\alpha+1)x] \tag{8}$$

assumes the value of -1 at x=1 and is also seen to vanish at $x = \alpha/(1+\alpha)$, which by our previous observations is the break-even point with uniform spacing. As earlier, the location can be adjusted with a choice of α. In distinction, this α alters the complexity of the equation by creating larger powers when we wish to push the break-even point towards 1. By contrast, the original piecewise development re-

quired only a shift of juncture point for the same purpose. The application of Eq. 7 to create a forward transformation is direct and is accomplished by simply replacing the g in Eq. 6 with the g_α of Eq. 7. This can be done for either one or both of the intervals about t_0 and each can have a separate adjustable α. Because $f_\alpha(1) = -1$, the control over minimum spacing by β is exactly the same. Thus, while we also retain a capability to separately control the locations of break-even points, we have been able to reduce the number of defining intervals: non-zero values now appear on two rather than four intervals.

In continuation, we seek a further reduction to a single interval with non-zero values. To do this, we shall construct a function which will directly produce symmetric clusters. Rather than considering the unit interval, we will develop the function on the larger interval from -1 to 1. To start, we form a symmetric positive bump with the function

$$h_\alpha(x) = (1-x)^\alpha(1+x)^\alpha \tag{9}$$

which is attached to the axis with vanishing first derivatives when $\alpha > 1$ and which has a single maximum value of unity when $x = 0$. At this stage, a monotonically decreasing function through the origin is needed as a factor to produce a negative slope at the origin and to split the bump into a positive bump before origin and a negative bump after the origin. If u(x) is such a function, then the derivative of uh_α at $x = 0$ is just $u'(0)$ since $h_\alpha(0) = 1$ and $h'_\alpha(0) = 0$. The simplest such choice is to set $u(x) = -x$. The associated function is then

$$g_\alpha(x) = - x(1-x)^\alpha(1+x)^\alpha \tag{10}$$

and satisfies the properties $g_\alpha(\pm 1) = g'_\alpha(\pm 1) = g_\alpha(0) = 0$ and $g'_\alpha(0) = - 1$. For a cluster interval of length 2T about t_0, we take $x = (t-t_0)/T$ in Eq. 10 and obtain the transformation

$$s = \begin{cases} t + \beta T g_\alpha\left(\frac{t-t_0}{T}\right) & \text{for } t_0-T \leq t \leq t_0+T \\ t & \text{otherwise} \end{cases} \tag{11}$$

by vertically scaling the resultant bump pair by βT and then adding it to the uniform mapping s = t. By direct differentiation, we have the weight function of unity everywhere except on the interval about t_0 where it assumes the form

$$w = 1 + \beta\left[(2\alpha+1)\left(\frac{t-t_0}{T}\right)^2 - 1\right]\left[1 - \left(\frac{t-t_0}{T}\right)^2\right]^{\alpha-1} \tag{12}$$

The evaluation at t_0 gives precisely the earlier clustering control β which produces the shrinking factor of 1-β. The motivation to get the same control came from $g'_\alpha(0) = - 1$ and the chain rule contribution of 1/T as a factor. The break-even points with uniform spacing are given when w = 1 in the interval about t_0 and are just $t_0 \pm T/\sqrt{2\alpha+1}$. As α increases, these points then symmetrically approach t_0. The largest possible distance is bounded by $T/\sqrt{3}$ in correspondence with α = 1 which is the lower bound for α. Thus, α can be used to control the distance of break-even points from t_0 over an interval from 0 to $T/\sqrt{3}$. To have at least one-third of the interval for clustering, this choice must be for α between 1 and 4. Undoubtedly, variations on this theme could be executed both to produce larger regions for the final clus-

tering compression and to insert a desired amount of asymmetry.

Rather than pursue these variations, we shall inspect the further possibility of asymptotic approximation with the desire to simply define an elementary clustering transformation in one global statement without having to establish a particular clustering interval. From this viewpoint, such intervals are implicitly defined when the asymptotic decay is essentially complete. The impreciseness here then gives us only a fuzzy definition. By contrast, however, we shall see that the earlier break-even points can be established precisely.

As in the last case, we multiply a positive bump function by the monotonically decreasing function - x which passes through the origin. To start, we consider the bump function $(1+x^2)^{-\alpha}$ and arrive at

$$g_\alpha(x) = - x(1+x^2)^{-\alpha} \tag{13}$$

which decays when $\alpha > 1$. The uniform transformation $s = t$ is now altered for local clustering about t_0 by setting

$$s = t + \beta T g_\alpha\left(\frac{t-t_0}{T}\right) \tag{14}$$

Once a decay rate α is chosen, the length scale T is used to appropriately shrink or expand the region of primary influence. By differentiation, the associated weight is given by

$$w = 1 + \beta \frac{(2\alpha-1)\left(\frac{t-t_0}{T}\right)^2 - 1}{\left[1 + \left(\frac{t-t_0}{T}\right)^2\right]^{\alpha+1}} \tag{15}$$

This reduces to $w = 1-\beta$ at the cluster center t_0 and thereby retains the meaning of the previous intensity controls β. The decay rate α

controls the location of break-even points which from Eq. 15 appear at a distance of $T/\sqrt{2\alpha-1}$ on either side of t_0. An increase in α simply causes a shift towards t_0 relative to the scaling T. At the other extreme, as α approaches 1, the shift is away from t_0 and is bounded by T. Altogether, adjustments in decay rate allow break-even points to be located anywhere between 0 and T units away from the center t_0. At the extreme of T, the effective clustering region is enlarged beyond T. To keep it, say within T units of t_0, a somewhat conservative choice is needed.

In the same spirit, we may also repeat the asymptotic construction with notably different analytical formulas. For example, we may decide that a better bump function would be given by the Gaussian form $e^{-\alpha x^2}$ and would then get

$$g_\alpha(x) = - xe^{-\alpha x^2} \tag{16}$$

in place of Eq. 13. This would correspondingly be used in Eq. 14 with the same interpretations for T and would lead to the weight

$$w = 1 + \beta\left[2\alpha\left(\frac{t-t_0}{T}\right)^2 - 1\right]\exp\left[-\alpha\left(\frac{t-t_0}{T}\right)^2\right] \tag{17}$$

with the same clustering intensity control β. The positive damping rate α is a control over the location of the break-even points relative to T. These are located at a distance of $T/\sqrt{2\alpha}$ on either side of t_0.

THE APPLICATIONS SETTING

To describe the setting in which applications are to be perform-

ed, we take note both of the general topic of grid generation and of the order of application. Grid generation arose as a topic of study in response to the need for numerical simulations of realistic physical systems. It has now been the subject of three general reviews [1] - [3], three major conferences [4] - [6] and one textbook [7]. A fundamental part of grid generation is the determination of pointwise distributions on curves. This occurs because curves are basic constructive elements in virtually any approach to grid generation. At the very least, they represent boundaries of two-dimensional regions and are typically used to create bounding surfaces for three-dimensional regions. The pointwise distribution on them directly influences the regional grid regardless of the method employed to generate that grid. The further redistribution of families of curves or surfaces within a regional grid is also a typical consequence of the redistribution of points along curves.

To accomplish the redistribution of points along curves in a precise manner, we have developed herein the elementary operation of creating a single local cluster about a point. The application of the operation to a succession of points can be ordered in either of two natural ways: the points are taken one at a time or they are done simultaneously. In correspondence, we may view the first as most ideally suited to an interactive graphical environment while the second appears more attractive for an automatic approach.

In the interactive setting, we assume that someone is sitting at a graphics terminal or workstation with the capability to view the pointwise distribution on the curve and to locate or insert pointwise

data by means of a cursor. For simplicity, we will assume that the cluster center and endpoints are taken from the existing grid points on the curve rather than at intermediate locations which would then necessitate an interpolation. With this assumption, the cursor is used to identify those grid points according to their indices. Since the grid on the curve is the result of mapping a uniform grid in a parameter s and since the corresponding uniform spacing can be taken as unity, the indices directly give the parametric distance that the endpoints s_- and s_+ are from the center s_0. If we take $s_0 = 0$, then $-s_-$ and s_+ are respectively the number of grid points below and above the cluster center. In terms of our new uniform parameter t, this becomes $t_- = s_-$, $t_0 = 0$, and $t_+ = s_+$. Next, the desired fractional decrease in spacing $1-\beta$ is chosen for the center. The forward mapping from Eq. 6 (or any of the equivalent variants) is now applied within the interval from t_- to t_+ to produce a local cluster of points about $s_0 = 0$. Unlike the center point and the points outside this interval, the clustering has caused points to fall generally between the old uniformly spaced points in s. If the curve is given analytically in terms of s, then the old mapping is just an evaluation at those in between points. Otherwise, for each new position in s, we must find the unit grid interval that contains it and then linearly interpolate the old map from s to the curve to get the new grid point location on the curve. In this process, there is no need to operate on the points outside of the cluster interval since they remain fixed. Upon application of such an elementary clustering operation, the new distribution is viewed and then a judgment to stop or continue is made. If

the previous distribution is stored, then there is also the option to easily restore it should we not like the result. Altogether, by applying the elementary clustering operations one at a time, we are able to interactively manipulate the pointwise distributions on curves.

In the context where the judgments for clustering are determined automatically for a collection of locations, it is more attractive to perform a single mapping rather than a succession of mappings. Certainly, as the cluster regions overlap each other, the successive mapping approach becomes more repetitious and less efficient. To obtain a single mapping, we may proceed from either of two viewpoints. The first is to consider what would have occurred had we done successive mappings. For any given order of mappings, the single mapping would be a successive composition in the same order. By applying the chain rule at each stage, the weight for the single mapping is just the product of the weights from the elementary cluster maps. We note that the elementary clustering weights are of the form $w_i = 1 + \beta_i C_i$ for cluster functions C_i and intensities β_i where $i = 1, 2, \ldots, n$ and n is the number of clusters. The weight for the single mapping is then

$$w = (1 + \beta_1 C_1)(1 + \beta_2 C_2) \cdots (1 + \beta_n C_n) \tag{18}$$

which is independent of the order of application. Unfortunately, the product is not particularly convenient to integrate. As a consequence, the linear β_i-approximation

$$w = 1 + \beta_1 C_1 + \beta_2 C_2 + \cdots + \beta_n C_n \tag{19}$$

is preferred and is also order-independent. Thus, the forward mapping

results from Eq. 2 by adding to uniform t = s the scaled bump pairs from each $\beta_i C_i$ as i = 1, ..., n. This superposition can then be viewed in the format of Fig. 4 where now the single twist about $s_0 = t_0$ is replaced by n of them. In this context of n simultaneous clusters, we note that a choice of specific intervals for each results in a detailed partition of s that can be avoided if we employ asymptotic approximations of the nature discussed in the section on alternative formulations.

CONCLUSIONS

The capability to locally manipulate pointwise distributions on curves was established through the introduction of an elementary operation for locally clustering points about any single point. The operation was created as a reparametrization where the spacing between new and old parameters is prescribed by means of a weight function. Various constraints upon the weights were established and the corresponding transformations were examined. It was found that the forward transformations from new to old are better because the composition of mappings is simpler and because the clustering intensity control is not limited as it is in the backward case.

The basic elements of construction were done in the most flexible manner by using piecewise linear weights. This gave piecewise quadratic transformations that were nontrivially defined over four intervals, and more importantly, gave the fundamental guidelines for more arbitrary constructions. Rather than pursue the greater degree of shape control that is available from general piecewise polynomial con-

structions, alternatives were presented to reduce the number of intervals of definition and thereby simplify the statement of the transformations. This viewpoint was taken up to the stage where endpoints of the local cluster region were only defined in a fuzzy sense by using asymptotic forms. These are attractive due to their simple global expression in one statement rather than in the previous piecemeal fashion. In summary, we first established a class of transformations that are suitable for elemenatry clustering operations and then we explored a broad range of attractive candidates from that class.

The most obvious demand for the local redistribution of points along curves occurs within the topic of grid generation and to some extent provides a general applications setting. In a more particular sense, the applications are considered to occur in sequence or simultaneously. Cases where only certain parts are simultaneous can be subdivided into either of these two possibilities. The sequential order of application is ideally suited to interactive graphics while the simultaneous application is well suited to automation.

REFERENCES

[1] P.R. EISEMAN, "Grid generation for fluid mechanics computations," Annual Review of Fluid Mechanics, Vol. 17, 1985, pp. 487-522.

[2] J.F. THOMPSON, "Grid generation techniques in computational fluid dynamics," AIAA Journal, Vol. 22, No. 11, 1984, pp. 1505-1523.

[3] J.F. THOMPSON, Z.U.A. WARSI and C.W. MASTIN, "Boundary-fitted coordinate systems for numerical solution of partial differential equations - a review," Journal of Computational Physics, Vol. 47, No. 1, 1982, pp. 1-108.

[4] K.N. GHIA and U. GHIA, Eds., Advances in Grid Generation, FED-Vol. 5, American Society of Mechanical Engineers, New York, 1983.

[5] J.F. THOMPSON, Ed., Numerical Grid Generation, North-Holland, New York, 1982.

[6] R.E. SMITH, Ed., "Numerical grid generation techniques," NASA CP 2166, 1980.

[7] J.F. THOMPSON, Z.U.A. WARSI and C.W. MASTIN, Numerical Grid Generation: Foundations and Applications, North-Holland, New York, 1985.

ON SIMILARITY SOLUTIONS OF A BOUNDARY LAYER PROBLEM WITH AN UPSTREAM MOVING WALL

M. Y. Hussaini
Institute for Computer Applications in Science and Engineering

W. D. Lakin
Old Dominion Univeristy
and
Institute for Computer Applications in Science and Engineering

A. Nachman
Air Force Office of Scientific Research

ABSTRACT

This work deals with the problem of a boundary layer on a flat plate which has a constant velocity opposite in direction to that of the uniform mainstream. It has previously been shown that the solution of this boundary value problem is crucially dependent on the parameter which is the ratio of the velocity of the plate to the velocity of the free stream. In particular, it was proved that a solution exists only if this parameter does not exceed a certain critical value, and numerical evidence was adduced to show that this solution is nonunique. Using Crocco formulation the present work proves this nonuniqueness. Also considered are the analyticity of solutions and the derivation of upper bounds on the critical value of wall velocity parameter.

Abbreviated title: Boundary layer on an upstream moving wall

Key words: non-uniqueness, Blasius equation, similarity solution

AMS classifications: 34B15 (Nonlinear boundary value problems)
76D10 (Boundary layer theory)

Research for the first and second authors was supported by the National Aeronautics and Space Administration under NASA Contract Nos. NAS1-17070 and NAS1-18107 while they were in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665-5225.

1. Introduction

The boundary layer on the upstream-moving flat plate at zero incidence admits of the classical similarity transformation which reduces the relevant partial differential equations to the Blasius equation.

$$
\begin{aligned}
f''' + ff'' &= 0 \\
f(0) &= 0 \\
f'(0) &= -\lambda, \quad \lambda > 0 \\
f'(\infty) &= 1,
\end{aligned}
$$

where $f = \psi(x,y)/\sqrt{(2\nu x)}$, ψ being the dimensional stream function, and ν the kinematic viscosity, and $\eta = y/\sqrt{(2\nu x)}$. This equation can be readily integrated once to yield

$$
f''(\eta) = f''(0)\exp\left[-\int_0^{\eta} f(z)dz\right],
$$

i.e.,

$$
f''(\eta) = f''(0)\exp\left[\frac{1}{2}\lambda\eta^2 - \frac{1}{2}\int_0^{\eta} (\eta - z)^2 f''(z)dz\right]
$$

using integration by parts twice. Obviously, the shear stress $f''(\eta)$ has the same sign as the skin-friction at the wall, $f''(0)$. For $\lambda = 0$, Weyl proved the existence and uniqueness using function-theoretical methods. For $\lambda \leq 0$, Callegari and Friedman and Callegari and Nachman found it expedient to work with the Crocco formulation, that is, in terms of shear stress $g(=f'')$ as the dependent variable and tangential velocity $u(=f')$ as the independent variable:

$$g(u)g''(u) + u = 0, \quad -\lambda < u < 1,$$

$$g'(-\lambda) = 0$$

$$g(1) = 0.$$

For $\lambda \leq 0$, they proved existence, uniqueness, and analyticity of solutions to Eq. (2) using an analytical function theory approach. For the case $\lambda > 0$, Hussaini and Lakin proved that a solution exists only for λ less than a critical value λ_c. Their numerical results showed nonuniqueness for $\lambda \leq \lambda_c$, and the numerical value of λ_c was found to be 0.3541... . In this work, the nonuniqueness is established rigorously. Also, proof of analyticity, and absolute monotonicity etc., is given. Certain analytical upper bounds on λ are established.

For convenience, we use the transformation $x = u + \lambda$ to map the interval $-\lambda < u < 1$, to $0 < x < 1 + \lambda$. So we consider the equations

$$g(x)g''(x) + (x - \lambda) = 0, \qquad 0 < x < 1 + \lambda \tag{1.1}$$

$$g'(0) = 0$$

$$g(1 + \lambda) = 0. \tag{1.2}$$

2. Analiticity of Solutions

In this section, the following basic result will be proved:

THEOREM 1: There is a range of positive values of λ such that the positive continuous solution $g(x)$ of the boundary value problem (1.1) and (1.2) is analytic on the closed intervel $[0,1 + \lambda]$.

This theorem will be proved by considering a sequence of lemmas. The first lemma required is:

LEMMA 1: The derivative $g'(x)$ vanishes at one and only one point on the interval $0 < x < 1 + \lambda$. Further, $g(x)$ has its maximum value at this point.

Proof of Lemma 1: Equation (1.1) can be integrated using the initial condition $g'(0) = 0$ to give

$$g'(x) = \int_0^x \frac{\lambda - \xi}{g(\xi)}\, d\xi, \tag{2.1}$$

Thus, as the initial value $\alpha = g(0) > 0$, both $g(x)$ and $g'(x)$ are positive for $0 < x \leq \lambda$. Also,

$$g''(x) = (\lambda - x)/g(x) \tag{2.2}$$

is positive for $0 \leq x < \lambda$ and $g''(\lambda) = 0$. The continuous solution $g(x)$ remains positive for $\lambda < x < 1 + \lambda$, and hence $g''(x)$ is now negative. This gives that $g'(x)$ is a monotone decreasing function for $x > \lambda$. As $g'(1 + \lambda) = -\infty$, there must thus be at least one point on the interval $(\lambda, 1 + \lambda)$ at which $g'(x)$ vanishes. In fact, assuming that $g'(x)$ vanishes at more than one point leads to a contradiction, for suppose that g' vanishes at both x_1 and x_2 with $x_1 < x_2$. Then, g'' would have to vanish at least once between these two points which is impossible as $g'' < 0$ for $x > \lambda$. The proof of Lemma 1 is concluded by noting that $g''(x_1) < 0$ implies that $g(x_1)$ must be the maximum value of $g(x)$.

LEMMA 2: The solution $g(x)$ has a convergent power series expansion on the closed interval $[x_1, 1 + \lambda]$.

Proof of Lemma 2: As $g(x)$ is positive and differentiable for $x_1 \leq x < 1 + \lambda$, equation (2.2) shows that $g(x)$ has derivatives of all orders on this interval. Further, expressions for these derivatives may be obtained directly from the differential equation (1.1). Induction shows that for $n \geq 1$, derivatives of $g(x)$ satisfy the recursion relation

$$g^{(n+3)} = -\frac{1}{g}\left\{(n+1)g' g^{(n+2)} + \frac{1}{2}\sum_{k=2}^{n+1}\left[\binom{n+1}{n-k+3} + \binom{n+1}{k} g^{(k)}\right] g^{(n-k+3)}\right\} \tag{2.3}$$

where $g^{(k)}$ is the k-th derivative of g with respect to x and $\binom{p}{q}$ is the usual combinatorial symbol.

Let $g(x_1) = \beta$, and consider the auxilliary function $G(x)$ defined by

$$G(x) = \beta - g(x). \tag{2.4}$$

Then, as β is the maximum value of $g(x)$, $G(x)$ is non-negative for $x_1 < x < 1 + \lambda$. Also, for all $n \geq 1$, $G^n(x) = -g^{(n)}(x)$. Consequently, equation (2.1) shows that $G'(x)$ is positive on the interval $x \leq x < 1 + \lambda$. From (1.1),

$$G''(x) = \frac{x - \lambda}{g(x)} \quad \text{and} \quad G'''(x) = \frac{1 + G' G''}{g(x)}$$

are also both positive on this interval. The recursion relation (2.3) thus shows that all derivatives of $G(x)$ are non-negative on the closed interval $[x_1, 1 + \lambda - \varepsilon]$ where $1 + \lambda - x_1 > \varepsilon > 0$. Hence, $G(x)$ is absolutely

monotonic on this closed interval. A theorem of Bernstein [4] now gives that $G(x)$ has a convergent Taylor series expansion about the point x_1 whose radius of convergence is not less than $1 + \lambda - x_1$. From the definition of $G(x)$, it immediately follows that for $|x - x_1| < 1 + \lambda - x_1$, $g(x)$ has the convergent expansion

$$g(x) = \sum_{n=0}^{\infty} \frac{g^{(n)}(x_1)}{n!} (x - x_1)^n. \tag{2.5}$$

Application of a Tauberian theorem [5] further shows that the power series (2.5) converges at the singular point $x = 1 + \lambda$ to the value $g(1 + \lambda) = 0$ completing the proof of Lemma 2.

To establish Theorem 1, it must be shown that for a nontrivial range of positive values of λ, the power series (2.5) for the solution $g(x)$ of the boundary value problem (1.1) and (1.2) converges at the left boundary point $x = 0$. This will be accomplished in Lemma 3. A consequence of this convergence will be an expansion for the initial value of $g(x)$ as the series

$$\alpha = \beta + \sum_{n=2}^{\infty} \frac{(-1)^n x_1^n}{n!} g^{(n)}(x_1). \tag{2.6}$$

LEMMA 3: *There exists a positive value* $\bar{\lambda}$ *such that if* $0 < \lambda < \bar{\lambda}$ *then* $x_1 < (1 + \lambda)/2$.

Lemma 3 gives that the left-hand boundary point $x = 0$ lies inside the radius of convergence of the power series expansion (2.5). Consequently, the corresponding solution of the boundary value problem will be analytic. It should be noted that the upper bound on x_1 given in Lemma 3 is a sufficient, but not a necessary, condition for convergence.

Proof of Lemma 3: Equation (1.1) may be integrated from 0 to x using the identity $gg'' = (gg')' - (g')^2$ and the initial condition $g'(0) = 0$. A second integration from 0 to x_1 now gives the result

$$\frac{x_1^2(x_1 - 3\lambda)}{6} = \frac{\alpha^2 - \beta^2}{2} + \int_0^{x_1} (x_1 - \xi)g'^2(\xi)d\xi. \tag{2.7}$$

An upper bound on the right-hand side of (2.7) and a lower bound on the maximum point x_1 are now required to establish the lemma.

A lower bound on x_1 may be obtained by using (2.1) and the fact that $g'(x_1) = 0$ to obtain

$$\int_0^{\lambda} \frac{\lambda - \xi}{g(\xi)}\, d\xi = \int_{\lambda}^{x_1} \frac{\xi - \lambda}{g(\xi)}\, d\xi. \tag{2.8}$$

As $g(x)$ is monotone increasing on $[0,x_1]$, $g(x) \leq g(\lambda)$ on $[0,\lambda]$, but $g(\lambda) \leq g(x)$ on $[\lambda,x_1]$. Equation (2.8) now gives

$$x_1 \geq 2\lambda. \tag{2.9}$$

As $g(x)$ has its only maximum at x_1 by Lemma 1, an immediate lower bound on $g(x_1) = \beta$ is $\beta > \alpha$. A sharper lower bound on β can be obtained from the expression

$$\beta = \alpha + \int_0^{x_1} \frac{(x_1 - \xi)(\lambda - \xi)}{g(\xi)}\, d\xi = \alpha + \int_0^{x_1} \frac{(\lambda - \xi)^2}{g(\xi)}\, d\xi \tag{2.10}$$

obtained by integrating (2.1) from 0 to x_1. As $g(x) \leq \beta$, and by (2.9), $x_1 - \lambda \geq \lambda$, equation (2.10) now gives the quadratic inequality

$$\beta^2 - \alpha\beta - \frac{2\lambda^3}{3} \geq 0 \tag{2.11}$$

which implies

$$\beta \geq \frac{\alpha + \sqrt{\alpha^2 + 8\lambda^3/3}}{2} . \tag{2.12}$$

A lower bound on $\beta^2 - \alpha^2$ which follows from (2.12) is thus

$$\beta^2 - \alpha^2 \geq \frac{2\lambda^3}{3} . \tag{2.13}$$

Consider next bounds on the initial value α. Let $X = 1 + \lambda$. Then, integrating (2.1) from 0 to X and using $g(X) = 0$ gives

$$\alpha = \int_0^X \frac{(x - \xi)(\xi - \lambda)}{g(\xi)} \, d\xi . \tag{2.14}$$

This relation may be rewritten in terms of strictly positive integrals as

$$\alpha = \int_\lambda^X \frac{(x - \xi)(\xi - \lambda)}{g(\xi)} \, d\xi - \int_0^\lambda \frac{(x - \xi)(\lambda - \xi)}{g'(\xi)} \, d\xi \tag{2.15}$$

which shows

$$\alpha \leq \int_\lambda^X \frac{(x - \xi)(\xi - \lambda)}{g(\xi)} \, d\xi . \tag{2.16}$$

The convexity of $g(x)$ on $[\lambda,X]$ implies that on this interval $g(x) \geq g(\lambda)\cdot(X - x)$. Equation (2.16) now gives that $\alpha \leq (2g(\lambda))^{-1}$. As $\alpha < g(\lambda)$, this further implies

$$\alpha^2 \leq 1/2 . \tag{2.17}$$

Equation (2.15) does not lend itself to the derivation of a lower bound on α^2. However, in the present consideration of analyticity, the required bound can be obtained from a relation between α and β which follows from the existence proof of Hussaini and Lakin [3]. This proof shows that if λ is positive and does not exceed a critical value, there is at least one initial value α such that a positive continuous solution of the initial value problem consisting of (1.1) and the conditions $g(0) = \alpha$ and $g'(0) = 0$ exists and has $g(X) = 0$, i.e., it is a solution of the boundary value problem. Further, the solution of the initial value problem will be unique if

$$\beta < 2\alpha. \tag{2.18}$$

It must be noted that a unique solution of the initial value problem does not imply a unique solution of the boundary value problem. This will be shown in section 4.

A lower bound on α^2 follows by using (2.18) in (2.12). The result is

$$\alpha^2 \geq \frac{\lambda^3}{3}. \tag{2.19}$$

The final bound needed for use in equation (2.7) is an upper bound for $g'(x)$ on the interval $[0,x_1]$. From (2.2), $g''(x)$ is a monotone decreasing function on this interval. Further, $g''(\lambda) = 0$ while the third derivative of g is negative when $x = \lambda$. Thus, $g'(x)$ has its maximum value at $x = \lambda$. This implies that on $[0,x_1]$

$$0 \leq g'(x) \leq g'(\lambda) = \int_0^{\lambda} \frac{\lambda - \xi}{g(\xi)}\, d\xi. \tag{2.20}$$

As $g(x) \geq \alpha$ on $[0,\lambda]$, equation (2.20) gives

$$0 \leq g'(x) \leq \frac{\lambda^2}{2\alpha} . \tag{2.21}$$

An upper bound on the integral in equation (2.7) is thus

$$\int_0^{x_1} (x_1 - \xi) g^{-2}(\xi) d\xi \leq \frac{3}{8} \lambda x_1^2. \tag{2.22}$$

Use of (2.13) and (2.22) in equation (2.7) implies

$$x_1^2 (x_1 - \frac{21}{4} \lambda) + 2\lambda^3 \leq 0. \tag{2.23}$$

This relation gives that x_1 will be less than $X/2$ for λ in the range $0 < \lambda < \overline{\lambda} = 0.1176$. The sufficient condition for analyticity is thus satisfied for a range of positive values of λ establishing Lemma 3 and Theorem 1.

Equation (2.9) implies that x_1 cannot be less than $X/2$ if $\lambda > 1/3$. Indeed, direct numerical solution of the boundary value problem shows that $x_1 < X/2$ when $\lambda < \hat{\lambda} = 0.32$ and α lies on the upper branch in Figure 1. The gap between the values of $\overline{\lambda}$ and $\hat{\lambda}$ is associated with fundemental problems in obtaining sharper bounds on the initial value α. For example, equation (2.15) implies

$$\alpha \leq \int_{\lambda}^{x_1} \frac{(x - \xi)(\xi - \lambda)}{g(\xi)} d\xi + \int_{x_1}^{X} \frac{(\xi - \lambda)}{g(\xi)} d\xi - \int_{x_1}^{X} \frac{(\xi - \lambda)^2}{g(\xi)} d\xi. \tag{2.24}$$

Individually, the last two integrals in (2.24) are formally infinite, yet they must cancel so as to give an order one upper bound. Direct numerical calculations show that the upper bound on α^2 is $\alpha^2 < 0.219961$. The upper bound in (2.17) is thus conservative by over a factor of two.

It must again be noted that $x_1 < X/2$ is only a sufficient condition for analyticity. For values of α on the upper branch of Figure 1, solutions of the boundary value problem can thus be expected to remain analytic for λ greater than $\hat{\lambda}$. Further insight can be gained by examining parameter values for which the condition (2.18), which is sufficient for a unique solution of the associated initial value problem, is maintained. Numerical results show that (2.18) holds for all values of α on the upper branch of Figure 1. It also holds for α on the lower branch of Figure 1 in the relatively small range $0.351 < \lambda < \lambda_c$ and is violated over the remainder of the lower branch. The behavior of β as a function of α is given in Figure 2. For values of λ associated with initial values on much of the lower branch of Figure 1, there must thus be serious doubts as to whether solutions of the boundary value problem (1.1) and (1.2) are analytic.

3. An Upper Bound on λ_c

The existence proof of Hussaini and Lakin [3] established the existence of solutions of (1.1) and (1.2) for positive values of λ less than a critical value λ_c. It was shown from (1.1) and (1.2) that $\lambda_c < 1/2$. The value of λ_c was also determined numerically in that work to be

$$\lambda_c = 0.3541079... \tag{3.1}$$

In this section, additional upper bounds for λ_c will be obtained directly from (1.1) and (1.2).

Using the identity that precedes equation (2.7), equation (2.1) can be integrated from 0 to x and the result integrated again from 0 to X. As $g(X) = 0$, this gives

$$\frac{X^2(X - 3\lambda)}{6} = \frac{\alpha^2}{2} + \int_0^X (x - \xi)g^{-2}(\xi)d\xi. \tag{3.2}$$

The right-hand side of (3.2) is intrinsically positive, and thus

$$X - 3\lambda \geq 0. \tag{3.3}$$

This relation immediately implies

$$\lambda \leq 1/2. \tag{3.4}$$

To obtain sharper bounds now requires the use of positive lower bounds for α^2 and the integral in (3.2). While no additional assumptions are required to obtain (3.4), in what follows it will be necessary to assume that $\beta < 2\alpha$. However, as noted previously, this condition is satisfied on the entire upper branch in Figure 1. In particular, it is satisfied in the limiting case when $\lambda = \lambda_c$.

Let the integral $I(x)$ be defined by

$$I(x) = \int_0^x (X - \xi)g^{-2}(\xi)d\xi. \tag{3.5}$$

Then, as $I(X) > 0$, equation (3.2) implies

$$X^2(X - 3\lambda) \geq 3\alpha^2. \tag{3.6}$$

Replacing X by $1 + \lambda$ and using (2.19) now gives the inequality

$$3\lambda^3 + 3\lambda^2 - 1 < 0 \tag{3.7}$$

which yields the improved bound

$$\lambda < 0.47533. \tag{3.8}$$

A slightly sharper bound can be obtained by noting that $I(X) > I(\lambda)$. Let $\delta = g(\lambda)$. Then, $g(x) \leq \delta$ on $[0,\lambda]$, so on this interval

$$g^{-2}(x) \geq \frac{x^2}{4\delta^2}(2\lambda - x)^2. \tag{3.9}$$

This leads to the relation

$$I(\lambda) \geq \frac{\lambda^5}{120\delta^2}(5\lambda + 16). \tag{3.10}$$

An upper bound on δ now follows from the fact that $g(x) \geq \alpha$ on $[0,\lambda]$ and

$$\delta = \alpha + \int_0^\lambda \frac{(\xi - \lambda)^2}{g(\xi)}\, d\xi. \tag{3.11}$$

In particular,

$$\delta^2 \leq \frac{2\lambda^3 + 1}{2}. \tag{3.12}$$

Use of (3.12) in (3.10) then shows

$$I(\lambda) \geq \frac{\lambda^5(5\lambda + 16)}{60(2\lambda^3 + 1)} . \tag{3.13}$$

Equation (3.2) now gives

$$X^2(X - 3\lambda) \geq 3\alpha^2 + 6I(\lambda) \tag{3.14}$$

which leads to the inequality

$$65\lambda^6 + 76\lambda^5 + 10\lambda^3 + 30\lambda^2 - 10 \leq 0. \tag{3.15}$$

The solution of (3.15) is

$$\lambda < 0.46824 \tag{3.16}$$

which is only a marginal improvement over (3.8).

Even if the lower bound on $I(X)$ is further sharpened by considering this integral on the full interval $[0,X]$, a significant decrease in the bound on λ is not obtained. Again, this is due to the difficulties associated with obtaining sufficiently sharp bounds on the initial value α.

4. Non-uniqueness of Solutions for $0 < \lambda < \lambda_c$

Using direct numerical results, Hussaini and Lakin [3] have shown that if λ is positive and less than λ_c then solutions of the boundary value problem are not unique. For a fixed value of λ in this range, as shown in Figure 1 there are two initial values α which lead to solutions of the boundary value problem. The purpose of this section is to prove this non-

uniqueness directly from (1.1) and (1.2). To this end, it is convenient to consider the normalized initial value problem

$$hh'' + t - L = 0, \tag{4.1}$$

$$h(0) = 1, \qquad h'(0) = 0 \tag{4.2}$$

obtained from the initial value problem for $g(x)$ by taking

$$g(x) = \alpha h(t) \quad \text{with} \quad x = \alpha^{2/3}\, t. \tag{4.3}$$

The parameter L in (4.1) is related to α and λ through the expression

$$L = \alpha^{-2/3}\, \lambda. \tag{4.4}$$

If $h(T) = 0$ and $\alpha(\lambda)$ is given by

$$\alpha = \{(1 + \lambda)/T\}^{3/2}, \tag{4.5}$$

then $g(X) = 0$, so the solution of the initial value problem with initial value (4.5) will also be a desired solution of the boundary value problem. Equations (4.3) through (4.5) also imply that in terms of T and L

$$\lambda = \frac{L}{T - L} \tag{4.6}$$

and

$$\alpha = (T - L)^{-3/2}. \tag{4.7}$$

LEMMA 4: Let $h_1(t)$ and $h_2(t)$ be solutions of the initial value problem (4.1) and (4.2) corresponding to L values L_1 and L_2, respectively. Then, if $L_2 > L_1$, $h_2(t) > h_1(t)$.

Proof of Lemma 4: For $t \ll L$, $h(t)$ must be of the form $1 + Lt^2/2$. Thus, the lemma holds for small values of t. That it holds for $0 < t \leq T$ can now be shown by contradiction. Let $\bar{t}$ be the first value of t at which $h_1(\bar{t}) = h_2(\bar{t})$. As h_1 was previously less than h_2, this requires $h_2''(\bar{t}) < h_1''(\bar{t})$. But,

$$h_2''(\bar{t}) = \frac{L_2 - \bar{t}}{h_2(\bar{t})} = \frac{L_2 - \bar{t}}{h_1(\bar{t})} > \frac{L_1 - \bar{t}}{h_1(\bar{t})} = h_1''(\bar{t}). \tag{4.8}$$

This contradiction establishes Lemma 4. Lemma 4 also shows that if $h_1(T_1) = 0$ and $h_2(T_2) = 0$, then $h_2(T_1) > 0$. This implies that:

COROLLARY: $T_2 > T_1$.

The derivative $h'(t)$ is given by an expression analogous to equation (2.1). As $h(0)$ is positive, both $h(t)$ and $h'(t)$ will be positive for $0 < t \leq L$. This shows that $T > L$. Consequently, the denominators in (4.6) are strictly positive. The following lemma gives a sharper result:

LEMMA 5: $T > 3L$.

Proof of Lemma 5: Equation (4.1) may be integrated twice from 0 to t using (4.2) to give

$$\tfrac{1}{2} h^2(t) + \frac{t^3}{6} - \frac{Lt^2}{2} = \tfrac{1}{2} + \int_0^t (t - \xi) h'^2(\xi) d\xi . \tag{4.9}$$

This implies

$$h^2(t) + \frac{1}{3} t^2(t - 3L) \geq 0. \tag{4.10}$$

Setting $t = T$ and $h(T) = 0$ now establishes the lemma.

Consider next the behavior of T as a function of L. It has already been shown in Lemma 4 that T is a monotone increasing function of L.

LEMMA 6: $T(L)$ is superlinear in L.

Proof of Lemma 6: Let t_1 be the point at which $h'(t_1) = 0$. As is the case for the original initial value problem in the variable x, there is one and only one such point, it lies in the interval $L < t < T$, and $h(t_1)$ is a maximum value.

Equation (4.1) may be multiplied by h' and divided by h to give

$$hh'' + \frac{h'(t - L)}{h} = 0. \tag{4.11}$$

Integration from 0 to t produces the result

$$\tfrac{1}{2} h'^2 + (t - L)\ln h(t) - \int_0^t \ln h(\xi) d\xi = 0. \tag{4.12}$$

Evaluating (4.12) at t_1 now shows

$$t_1 = L + \frac{\int_0^{t_1} \ln h(\xi)\,d\xi}{\ln h(t_1)} . \tag{4.13}$$

Next, the expression

$$h(t) = 1 + \int_0^t \frac{(t - \xi)(L - \xi)}{h(\xi)}\,d\xi \tag{4.14}$$

may be evaluated at $t = L$ to give an expression for $h(L)$.

$$h(L) = 1 + \int_0^L \frac{(L - \xi)^2}{h(\xi)}\,d\xi . \tag{4.15}$$

As $h'(t)$ is non-negative on the interval $[0,L]$, $h(t)$ is monotone increasing, so $h(t) \leq h(L)$. Use of this fact in (4.15) gives the quadratic inequality

$$h^2(L) - h(L) - \frac{L^3}{3} \geq 0 \tag{4.16}$$

which implies $h^2(L) \geq L^3/3$. The solution $h(t)$ has its maximum value at t_1. Consequently,

$$h(t_1) > \sqrt{\frac{L^3}{3}} . \tag{4.17}$$

One additional bound is needed before demonstrating the superlinear behavior of $T(L)$. The change of concavity of $h(t)$ on the interval $[0,t_1]$ due to the fact that $h''(L) = 0$ precludes obtaining as a lower bound for h on this interval the straight line which passes through the origin and the point $(t_1,h(t_1))$, i.e., it cannot be shown that $h(t) > h(t_1)\cdot t/t_1$. However, for a given L, it is clear that $h(t)$ can be bounded below on this interval by a curve of the form

$$H(t;k) = \frac{h(t_1)t^k}{t_1^k} \tag{4.18}$$

for a value of $k > 1$. As k increases, these curves become progressively more convex. It should be noted that if $H(t,\overline{k})$ provides a lower bound on $[0,t_1]$ for the solution of (4.1) and (4.2) associated with $L = \overline{L}$, then, by Lemma 4, $H(t,\overline{k})$ also provides a lower bound for solutions associated with larger values of L.

This lower bound for $h(t)$ on $[0,t_1]$ may be used to obtain an lower bound for the integral in equation (4.13). In particular,

$$\int_0^t \ln h(\xi)d\xi > t_1 \ln h(t_1) - kt_1. \tag{4.19}$$

Equation (4.13) now implies that

$$t_1 > \frac{L}{k} \ln h(t_1). \tag{4.20}$$

Use of (4.17) then gives

$$T > t_1 > \frac{L}{2k} \ln \left(\frac{L}{3^{1/3}}\right). \tag{4.21}$$

The superlinear behavior of $T(L)$ is thus established.

THEOREM 2: For positive values of λ in the range $0 < \lambda < \lambda_c$, solutions of the boundary value problem (1.1) and (1.2) are not unique.

Proof of Therem 2: Consider the behavior of L as a function of λ. By (4.4), $L(0) = 0$. Equation (4.6) and the superlinear behavior of T with repect to L shown in Lemma 6 now imply that the graph of λ vs L must be as in Figure 3. In particular, for a fixed positive λ which is less than λ_c, there will be two distinct values of L. By the corollary to Lemma 4, each value of L must correspond to a different value of T. Equation (4.5) now shows that for the fixed value of λ, two distinct values α_1 and α_2 exist such that the solutions of the initial value problems with these α's are solutions of the boundary value problem (1.1) and (1.2). Solutions of the boundary value problem are thus not unique completing the proof of Theorem 2.

References

[1] A. J. Callegari and M. B. Friedman, "An analytical solution of a nonlinear, singular boundary value problem in the theory of viscous flows," J. Math. Anal. Appl., 21 (1968), pp. 510-529.

[2] A. J. Callegari and A. Nachman, "Some singular, nonlinear differential equations arising in boundary layer theory," J. Math. Anal. Appl., 64 (1978), pp. 96-105.

[3] M. Y. Hussaini and W. D. Lakin, "Existence and nonuniqueness of similarity solutions of a boundary-layer problem," Quart. J. Mech. Appl. Math., 39 (1986), in press.

[4] A. F. Tinman, Theory of Approximations of Functions of a Real Variable, Pergamon Press, England, 1963.

[5] N. Wiener, The Fourier Integral and Certain of its Applications, Dover, New York, 1933.

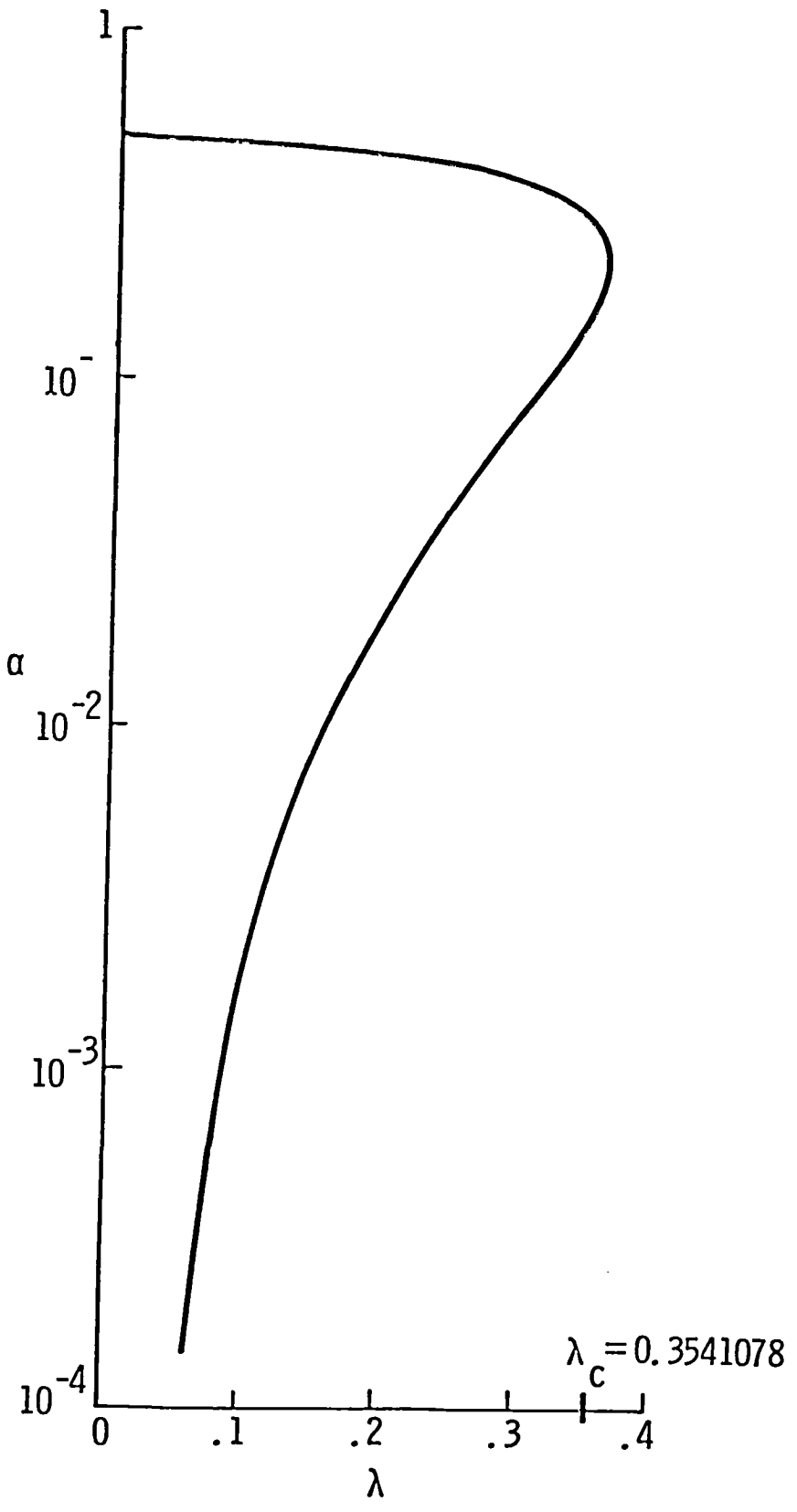


Figure 1. Values of the parameter $\alpha = f''(0)$ for which $f'(\infty) = 1$ as a function of λ.

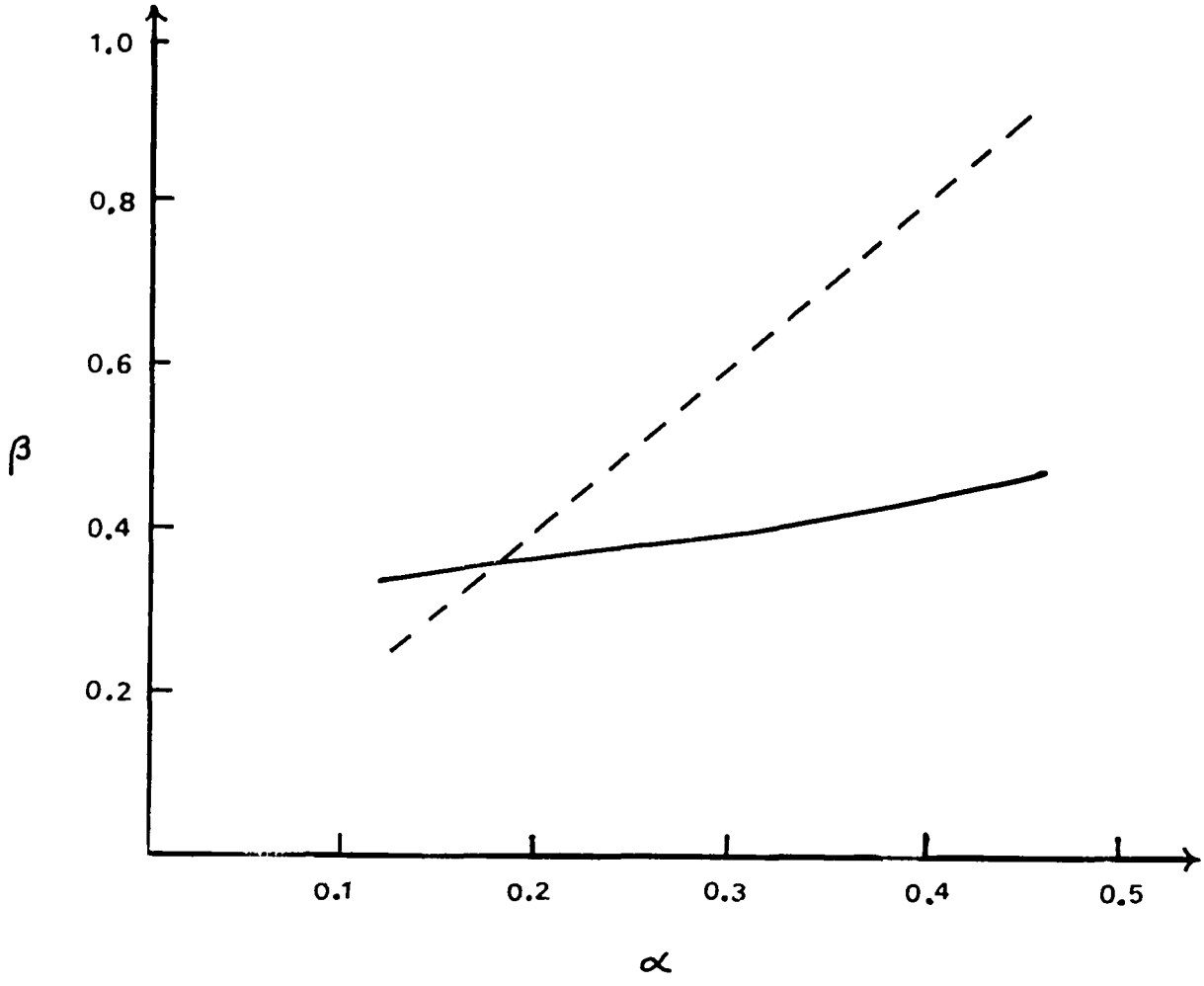


Figure 2. Values of the maximum value β of $g(x)$ as a function of the initial value $g(0) = \alpha$. The dotted line is $\beta = 2\alpha$.

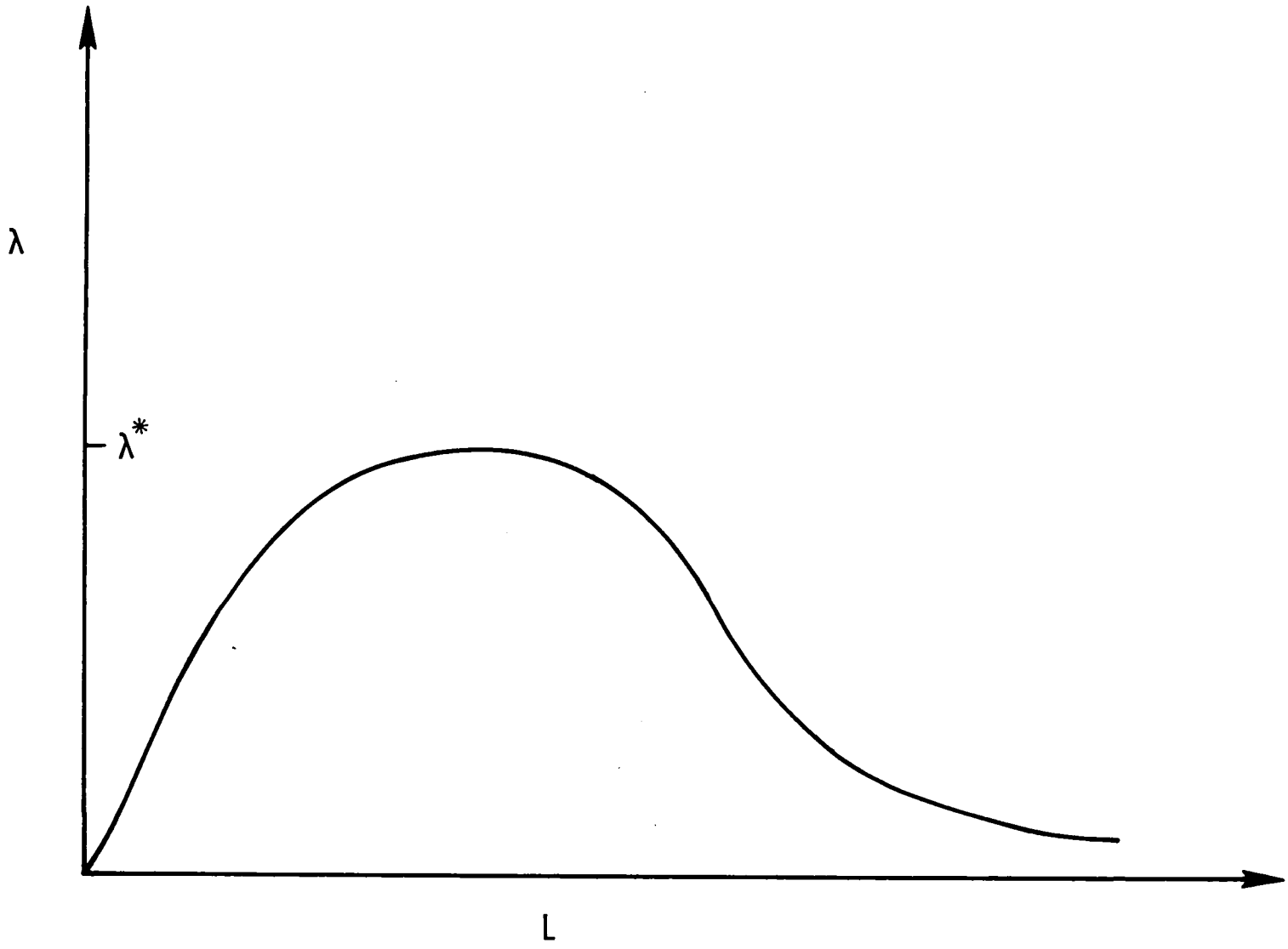


Figure 3. The qualitative behavior of the parameter L in the initial value problem (4.1) and (4.2) as a function of λ.

ON THE ADVANTAGES OF THE VORTICITY-VELOCITY FORMULATION OF THE EQUATIONS OF FLUID DYNAMICS

Charles G. Speziale
Institute for Computer Applications in Science and Engineering
NASA Langley Research Center, Hampton, VA 23665-5225
and
Georgia Institute of Technology, Atlanta, GA 30332

Abstract

The mathematical properties of the pressure-velocity and vorticity-velocity formulations of the equations of viscous flow are compared. It is shown that a vorticity-velocity formulation exists which has the interesting property that non-inertial effects only enter into the problem through the implementation of initial and boundary conditions. This valuable characteristic, along with other advantages of the vorticity-velocity approach, are discussed in detail.

Research was supported by the National Aeronautics and Space Administration under NASA Contract No. NAS1-18107 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665-5225.

Two distinctly different approaches have been utilized in the literature for the numerical solution of the equations of viscous flow in three-dimensions. In the more common approach, the momentum equation, which contains both the velocity and pressure, is solved numerically along with a derived Poisson equation for the pressure (i.e., the pressure-velocity or primitive variable formulation [1-3]). The alternative approach is based on eliminating the pressure from the momentum equation by the application of the curl. In this manner, a vorticity transport equation is solved numerically in lieu of the momentum equation (i.e., the vorticity-velocity formulation [4-6]). The purpose of the present note is to explore in more detail the properties of these disparate numerical approaches. It will be shown that the vorticity-velocity formulation has a striking advantage when applied to problems in non-inertial frames of reference. More specifically, there exists an intrinsic vorticity-velocity formulation wherein all non-inertial effects (arising from both the rotation and translation of the frame of reference relative to an inertial framing) only enter into the solution of the problem through the implementation of initial and boundary conditions. This is in stark contrast to the pressure - velocity formulation where non-inertial effects appear directly in the momentum equation in the form of Coriolis and Eulerian accelerations--a state of affairs which can give rise to a variety of numerical problems [2]. A detailed exposition of this interesting property of the vorticity-velocity formulation will be presented along with a brief discussion of other advantages of this approach.

For simplicity, we will restrict our attention to the analysis of viscous incompressible flow governed by the Navier-Stokes equation and continuity equation which, respectively, take the form

$$\frac{\partial \underset{\sim}{v}}{\partial t} + \underset{\sim}{v} \cdot \underset{\sim}{\nabla v} = - \underset{\sim}{\nabla} p + \nu \nabla^2 \underset{\sim}{v}, \tag{1}$$

$$\underset{\sim}{\nabla} \cdot \underset{\sim}{v} = 0, \tag{2}$$

where $\underset{\sim}{v}$ is the velocity vector, p is the pressure, and ν is the kinematic viscosity of the fluid. Here, the validity of (1) requires that the external body forces be conservative and that the frame of reference be inertial. In an arbitrary non-inertial frame of reference (which can rotate with a time-dependent angular velocity $\underset{\sim}{\Omega}(t)$ and translate with a time-dependent velocity $\underset{\sim}{v}_0(t)$ relative to its origin 0), the Navier-Stokes equation takes the more complex form [7]

$$\frac{\partial \underset{\sim}{v}}{\partial t} + \underset{\sim}{v} \cdot \underset{\sim}{\nabla}\, \underset{\sim}{v} + \dot{\underset{\sim}{\Omega}} \times \underset{\sim}{r} + \underset{\sim}{\Omega} \times (\underset{\sim}{\Omega} \times \underset{\sim}{r}) + \dot{\underset{\sim}{v}}_0 + 2\underset{\sim}{\Omega} \times \underset{\sim}{v} = - \underset{\sim}{\nabla} p + \nu \nabla^2 \underset{\sim}{v}. \tag{3}$$

Here, $\underset{\sim}{r}$ is the position vector and the non-inertial terms on the left-hand side of (3) are, respectively, referred to as the Eulerian, centrifugal, translational, and Coriolis accelerations. The continuity equation still assumes the same form (2) in any non-inertial frame of reference.

By the introduction of a modified pressure P which includes the centrifugal and translational acceleration potentials, the non-inertial form of the Navier-Stokes equation (3) can be simplified considerably. More specifically, (3) can be written in the equivalent form

$$\frac{\partial \underset{\sim}{v}}{\partial t} + \underset{\sim}{v} \cdot \underset{\sim}{\nabla}\, \underset{\sim}{v} + \dot{\underset{\sim}{\Omega}} \times \underset{\sim}{r} + 2\underset{\sim}{\Omega} \times \underset{\sim}{v} = - \underset{\sim}{\nabla} P + \nu \nabla^2 \underset{\sim}{v}, \tag{4}$$

where

$$P = p + \frac{1}{2}\left(\underset{\sim}{\Omega} \cdot \underset{\sim}{r}\right)^2 - \frac{1}{2}\,\Omega^2\, r^2 + \dot{\underset{\sim}{v}}_0 \cdot \underset{\sim}{r}. \tag{5}$$

In the pressure-velocity formulation, equation (4) is solved in conjunction with a Poisson equation for the pressure which is obtained by taking the divergence of (4). Hence, the governing equations to be solved numerically in this approach can be summarized as follows:

$$\frac{\partial \underset{\sim}{v}}{\partial t} + \underset{\sim}{v} \cdot \underset{\sim}{\nabla}\,\underset{\sim}{v} + \dot{\underset{\sim}{\Omega}} \times \underset{\sim}{r} + 2\underset{\sim}{\Omega} \times \underset{\sim}{v} = -\,\underset{\sim}{\nabla} P + \nu \nabla^2\, \underset{\sim}{v}, \tag{6}$$

$$\nabla^2\, P = -\,\mathrm{tr}\left(\underset{\sim\sim}{\nabla v} \cdot \underset{\sim\sim}{\nabla v}\right) + 2\underset{\sim}{\Omega} \cdot \underset{\sim}{\omega}, \tag{7}$$

subject to the initial and boundary conditions

$$\underset{\sim}{v} = \underset{\sim}{v}_0\,, \qquad \text{at} \quad t = t_0, \tag{8}$$

$$\left.\begin{aligned} \underset{\sim}{v} &= \underset{\sim}{v}_B \\ P &= P_B \end{aligned}\right\} \qquad \text{on} \quad B. \tag{9}$$

In (7) and (9), $\mathrm{tr}(\cdot)$ denotes the trace, $\underset{\sim}{\omega}$ is the vorticity vector, and B denotes the boundary surface of the region. Of course, equations (6) and (7) must be solved subject to the continuity equation (2). Since we are considering general three-dimensional flow, a stream function solution does not exist. Hence, the solution for the velocity $\underset{\sim}{v}$ must be projected in some suitable fashion onto the space of solenoidal vectors.

It is quite clear that the form of (6) and (7) (and, hence, their mathematical character) change depending on whether or not the frame of

reference is inertial. Consequently, a particular numerical algorithm which may be optimal for a given class of flows in an inertial frame of reference may not be so for the same class of flows in a non-inertial framing. It will now be demonstrated that the vorticity-velocity formulation does not suffer from this deficiency.

The vorticity-velocity formulation is based on the vorticity transport equation which is obtained by taking the curl of (4). This equation takes the form

$$\frac{\partial \underset{\sim}{\omega}}{\partial t} + \underset{\sim}{v} \cdot \underset{\sim\sim}{\nabla\omega} = \underset{\sim}{\omega} \cdot \underset{\sim\sim}{\nabla v} + \nu\nabla^2 \underset{\sim}{\omega} + 2\underset{\sim}{\Omega} \cdot \underset{\sim\sim}{\nabla v} - 2\dot{\underset{\sim}{\Omega}} \tag{10}$$

in any non-inertial frame of reference where

$$\underset{\sim}{\omega} = \underset{\sim}{\nabla} \times \underset{\sim}{v} \tag{11}$$

is the vorticity vector. It is clear that the velocity and vorticity are also connected through the Poisson equation

$$\nabla^2 \underset{\sim}{v} = - \underset{\sim}{\nabla} \times \underset{\sim}{\omega} \tag{12}$$

which is a direct consequence of the vector identity

$$\underset{\sim}{\nabla} \times (\underset{\sim}{\nabla} \times \underset{\sim}{v}) = \underset{\sim}{\nabla}(\underset{\sim}{\nabla} \cdot \underset{\sim}{v}) - \nabla^2 \underset{\sim}{v}. \tag{13}$$

The intrinsic vorticity $\underset{\sim}{W}$, defined by

$$\underset{\sim}{W} = \underset{\sim}{\omega} + 2\underset{\sim}{\Omega}, \tag{14}$$

can be introduced which represents the vorticity relative to an inertial frame of reference. Since $\underset{\sim}{\Omega}$ is spatially homogeneous (i.e., $\underset{\sim\sim}{\nabla\Omega} = 0$), it is a simple matter to show that the non-inertial form of the vorticity-velocity formulation can be written as follows:

$$\frac{\partial \underset{\sim}{W}}{\partial t} + \underset{\sim}{v} \cdot \underset{\sim\sim}{\nabla W} = \underset{\sim}{W} \cdot \underset{\sim\sim}{\nabla v} + \nu \nabla^2 \underset{\sim}{W} \tag{15}$$

$$\nabla^2 \underset{\sim}{v} = - \underset{\sim}{\nabla} \times \underset{\sim}{W}. \tag{16}$$

Equations (15) - (16) must be solved (in some region R with a boundary surface B) subject to the initial and boundary conditions

$$\underset{\sim}{W} = (\underset{\sim}{\nabla} \times \underset{\sim}{v})_0 + 2\underset{\sim}{\Omega}_0, \qquad \text{at} \quad t = t_0 \tag{17}$$

$$\left.\begin{aligned} \underset{\sim}{v} &= \underset{\sim}{v}_B \\ \underset{\sim}{W} &= (\underset{\sim}{\nabla} \times \underset{\sim}{v})_B + 2\underset{\sim}{\Omega} \end{aligned}\right\} \qquad \text{on} \quad B. \tag{18}$$

Of course, it is well known that the vorticity, as well as the intrinsic vorticity, are solenoidal, i.e.,

$$\underset{\sim}{\nabla} \cdot \underset{\sim}{W} = 0, \tag{19}$$

and, hence, the solutions for $\underset{\sim}{W}$ and $\underset{\sim}{v}$ must, in some suitable fashion, be projected onto the space of solenoidal vectors.

This vorticity-velocity formulation of fluid dynamics represented by equations (15) - (18) has the striking property that non-inertial effects only

enter into the solution of the problem through the implementation of initial and boundary conditions. Consequently, the basic structure of the numerical algorithm (i.e., the numerical formulation of (15) - (16)) will be independent of whether or not the frame of reference is inertial--a situation which greatly enhances the general applicability of any Navier-Stokes computer code which is developed based on this approach.

At this point, a few comments should be made concerning the alternate ways in which the velocity field can be calculated in the vorticity-velocity formulation. Instead of solving the Poisson equation (16), it is possible to solve the defining equation for vorticity directly, i.e.,

$$\underset{\sim}{\nabla} \times \underset{\sim}{v} = \underset{\sim}{\omega} = \underset{\sim}{W} - 2\underset{\sim}{\Omega}, \tag{20}$$

(see Gatski, Grosch, and Rose [6,8]). Of course, for plane or axisymmetric flows, there exists a stream function ψ such that [7]

$$\underset{\sim}{v} = \underset{\sim}{\lambda} \times \underset{\sim}{\nabla}\psi \tag{21}$$

$$\underset{\sim}{\nabla} \times (\underset{\sim}{\lambda} \times \underset{\sim}{\nabla}\psi) = \underset{\sim}{W} - 2\underset{\sim}{\Omega}, \tag{22}$$

where $\underset{\sim}{\lambda} = \underset{\sim}{\nabla}\chi$ and χ is the coordinate that the flow is independent of (for plane flows, (22) reduces to the Poisson equation $\nabla^2 \psi = W - 2\Omega$). While the motion of the frame of reference does enter into the equations of motion in these alternate vorticity-velocity formulations, it does so in a much less significant way than in the pressure-velocity formulation. To be specific, the transport equation which is solved (i.e., equation (15)) does not contain

any frame-dependent terms and, at each time step, the partial differential equation for the determination of the velocity field is only altered by the addition of a constant forcing function in the form of $2\underset{\sim}{\Omega}$ (the added term on the right-hand side of (20) and (22)).

Finally, it would be of value to mention some other advantages of the vorticity-velocity formulation. More difficulties have been known to arise in the implementation of pressure boundary conditions than vorticity boundary conditions [1,2] (of course, both boundary conditions must usually be derived). Difficulties in satisfying the continuity equation in the pressure-velocity formulation have also been known to give rise to numerical instabilities [1]. Furthermore, in the vorticity-velocity approach, the vorticity vector is calculated directly. This is of considerable value since the vorticity field can play an important role in characterizing certain features of turbulence [9]. While it is certainly not being suggested that the pressure-velocity formulation be abandoned, this study does indicate that the vorticity-velocity formulation can have distinct advantages when applied to an important class of viscous flows.

Acknowledgment

The author would like to thank Dr. T. Gatski and Dr. M. Rose for some valuable comments and criticisms of the original draft of this paper.

REFERENCES

[1] A. J. CHORIN, Math. Comp., 22 (1968), 745.

[2] G. P. WILLIAMS, J. Fluid Mech., 37 (1969), 727.

[3] D. A. ANDERSON, J. C. TANNEHILL, and R. H. PLETCHER, "Computational Fluid Dynamics and Heat Transfer," McGraw-Hill, New York, 1984.

[4] S. C. R. DENNIS, D. B. INGHAM, and R. N. COOK, J. Comp. Phys., 33 (1979), 325.

[5] H. F. FASEL, "Numerical Solution of the Complete Navier-Stokes Equations for the Simulation of Unsteady Flows," Lecture Notes in Mathematics, No. 771, Springer-Verlag, Berlin, 1980.

[6] T. B. GATSKI, C. E. GROSCH, AND M. E. ROSE, to be published.

[7] G. K. BATCHELOR, "An Introduction to Fluid Dynamics," Cambridge University Press, London, 1967.

[8] T. B. GATSKI, C. E. GROSCH, and M. E. ROSE, J. Comp. Phys., 48 (1982), 1.

[9] E. LEVICH and A. TSINOBER, Phys. Letters, 93A (1983), 293.

*U.S. GOVERNMENT PRINTING OFFICE: 1986- 625-014:40004

Standard Bibliographic Page

1. Report No. NASA CR-178076 ICASE Report No. 86-18	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle ADVANCES IN NUMERICAL AND APPLIED MATHEMATICS		5. Report Date March 1986 6. Performing Organization Code
7. Author(s) J. C. South, Jr. and M. Y. Hussaini (editors)		8. Performing Organization Report No. 86-18
9. Performing Organization Name and Address Institute for Computer Applications in Science and Engineering Mail Stop 132C, NASA Langley Research Center Hampton, VA 23665-5225		10. Work Unit No. 11. Contract or Grant No. NAS1-17070; NAS1-18107
12. Sponsoring Agency Name and Address National Aeronautics and Space Administration Washington, DC 20546		13. Type of Report and Period Covered Contractor Report 14. Sponsoring Agency Code 505-31-83-01

15. Supplementary Notes

Langley Technical Monitor:
J. C. South, Jr.

Final Report

To appear in Applied Numerical Mathematics

16. Abstract

This collection of papers covers some recent developments in numerical analysis and computational fluid dynamics. Some of these studies are of a fundamental nature. They address basic issues such as intermediate boundary conditions for approximate factorization schemes, existence and uniqueness of steady states for time-dependent problems, pitfalls of implicit time stepping, etc. The other studies deal with modern numerical methods such as total-variation-diminishing schemes, higher-order variants of vortex and particle methods, spectral multidomain techniques, and front-tracking techniques. There is also a paper on adaptive grids. The fluid dynamics papers treat the classical problems of incompressible flows in helically-coiled pipes, vortex breakdown, and transonic flows.

17. Key Words (Suggested by Authors(s))	18. Distribution Statement
Numerical Analysis Computational Fluid Dynamics Transonic Flows Vortex Breakdown Spectral Methods	34 - Fluid Mechanics & Heat Transfer 64 - Numerical Analysis Unclassified - Unlimited

19. Security Classif.(of this report)	20. Security Classif.(of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	597	A25

For sale by the National Technical Information Service, Springfield, Virginia 22161