Finite Difference Methods
with application to the
Cavity Problem

by

Gregory Woodford

A thesis submitted to the
Australian National University
for the degree of Doctor of Philosophy
February 1975

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

This work was carried out at the Computer Centre, the Australian National University. I gratefully acknowledge the financial assistance of a Commonwealth Postgraduate Award during this period.

I wish to thank my supervisor and the head of the centre, Mike Osborne, for the general support and encouragement that he has lent me and the constructive criticism that he has lent my work.

I wish to thank all the members of the centre for the welcome and friendliness that they have extended to me during my stay, and for the seminar, talks and discussions that they have contributed to my education.

I would also like to thank Lee Stirzaker for checking the manuscript for small errors and the typists Cheryl Riddell and Anne Clugston for the care and effort that they have devoted to this script.

PREFACE

Part of the work in Chapter 4 in this thesis was done in collaboration with Mike Osborne especially that relating to first order formulations of the problem.

Chapter 2 in this thesis has been published as Woodford [46] and the text of that paper has been followed closely.

Elsewhere in this thesis, unless another source is acknowledged, the work described is my own.

Gregory Woodford

ABSTRACT

This thesis examines the use of graded meshes and extrapolation techniques in finite difference methods of solution for firstly, two-point boundary value problems and secondly, the prototype cavity problem. For equations of the form

$$\frac{d^2y}{dx^2} + a(x)\frac{dy}{dx} + b(x)y = 0$$

subject to the conditions

$$y(0) = y(1) = 1 ,$$

it is shown how to construct graded meshes that give optimum numerical properties to the finite difference scheme and also allow extrapolation processes, something that is not usually available when using graded meshes. Both first and second order formulations of the two-point boundary value problem are examined.

Chapter 3 examines the cavity problem using a regular mesh while Chapter 5 uses graded meshes. Only the case of a Reynolds number of 50 is discussed for both divergence and convective forms of the vorticity transport equation. Extrapolation is used in all cases. For the regular mesh cases and one graded mesh cases, convergence is attained by the extrapolated results to within two significant digits. A mesh is demonstrated that is too severe for the problem in hand to emphasise the dangers involved in the choice of graded mesh for problems where a quantitative method is not known.

CHAPTER 1

THE PROTOTYPE CAVITY PROBLEM

1.1. INTRODUCTION

The prototype cavity problem concerns the fluid motion generated in a rectangular cavity by the uniform translation of the upper surface. The fluid in the cavity is viscous and incompressible. In this thesis, numerical solutions to the Navier-Stokes equations of fluid motion are sought to describe the fluid motion for middle range to high Reynolds numbers where the Reynolds number is defined as $Re = UL/\nu$ where U is the velocity of the upper surface, L the width of the cavity and ν is the kinematic viscosity.

The cavity problem is part of a larger class of problems of steady separated flows. This class of flows and in particular the cavity problem has been studied by Burggraf [8] both analytically and numerically. The fluid dynamic features of the cavity flows and of closely related flows (e.g., with thermal effects added) have been extensively studied in the literature (Kawaguti [24], Mills [29], Burggraf [8], Pan and Acrivos [33], Greenspan [22], Donovan [14], Torrance et al. [43], Runchal, Spalding and Wolfshtein [38], Marshall and Van Spiegel [28], Bozeman and Dalton [5]). Experimental visualisations (Mills [29], Pan and Acrivos [33]) have been attempted but mainly for low to middle range (50 - 3,200) Reynolds numbers.

Many authors (see §1.4) have attempted to find accurate numerical solutions to this problem for various ranges of the Reynolds number with the

general aim of obtaining a solution for high Reynolds numbers. Though questions may still be asked (see Bozeman and Dalton [5], Torrance [42]) about the finite-difference representations of the non-linear terms in the Navier-Stokes equations, it is my feeling that the major finite-difference approach via the use of evenly spaced grids has been fully extended and the problem of a solution for high Reynolds numbers is a matter of truely excessive computer time. It is felt that finite differences using graded meshes may provide an improvement in the solution of the problem. One of the aims of this thesis is to explore the applicability of graded meshes as an alternative approach to the solution of the problem by difference methods.

The cavity problem has special interest as a prototype problem on which to test numerical schemes. This special interest stems basically from the simplicity of formulation which implies a reduced complexity for the implementation of new numerical schemes and allows easy testing of the many parameters in the models in the search for improvements in convergence, etc. The problem's simple formulation certainly does not imply triviality or even ease of solution as will be easily seen by the later discussion.

1.2. FORMULATION OF THE CAVITY PROBLEM

Consider the cavity problem in terms of the physical variables. The Navier-Stokes equations for an incompressible fluid are

$$\frac{\partial \underset{\sim}{u}}{\partial t} + \underset{\sim}{u}\cdot\nabla\underset{\sim}{u} = \nu\nabla^2\underset{\sim}{u} - \frac{1}{\rho}\nabla p \qquad (1.2.1a)$$

and the incompressibility condition

$$\nabla\cdot\underset{\sim}{u} = 0 \qquad (1.2.1b)$$

where

$\underset{\sim}{u}^T = (u,v)$ is the velocity of a fluid particle,

ν = kinematic viscosity,

p = pressure,

ρ = mass density,

t = time.

Suppose the cavity with which we are dealing has width L, height D and the upper surface of the cavity is moving with uniform velocity U from left to right. We define the non-dimensional variables

$$\begin{aligned} x' &= x/L\ , \\ y' &= y/L\ , \\ (\underset{\sim}{u}')^T &= (u/U,\ v/U)\ , \\ p' &= p/\rho U^2\ , \\ t' &= tU/L\ , \\ a &= D/L \quad \text{(the aspect ratio)}\ , \\ Re &= UL/\nu \quad \text{(the Reynolds number)}. \end{aligned}$$

Substituting the variables into the equations (1.2.1), we obtain the equations (1.2.2) given below, the non-dimensional equations of fluid motion in a rectangular cavity of aspect ratio a. Hence we have

$$\frac{\partial \underset{\sim}{u}}{\partial t} + \underset{\sim}{u}\cdot\nabla\underset{\sim}{u} = \frac{1}{Re}\nabla^2\underset{\sim}{u} - \nabla p \qquad (1.2.2a)$$

and

$$\nabla\cdot\underset{\sim}{u} = 0 \qquad (1.2.2b)$$

where the primes have been dropped from all variables because there can be no ambiguity as from this point, non-dimensional quantities are assumed. In equations (1.2.2), the boundary values are known only for the two velocity components, namely the velocity $\underset{\sim}{u} = 0$ at all walls except the upper surface where $\underset{\sim}{u}^T = (1,0)$. The boundary values for the pressure are not known.

If the streamfunction-vorticity formulation of the equations (1.2.2) is used instead of the physical equations in the three unknowns (u, v and p), then only two unknown variables are sought, the stream function ψ and the vorticity ω. The boundary values for this formulation are found naturally without the imposition of any extra conditions. Hence we consider the stream-function-vorticity formulation of the cavity problem.

Eliminating the pressure variable p by taking the curl of equation (1.2.2a) we obtain the vorticity transport equation in convective form

$$\frac{\partial \omega}{\partial t} + \underset{\sim}{u}\cdot\nabla\omega = \frac{1}{Re}\nabla^2\omega \qquad (1.2.3)$$

where ω represents the two-dimensional vorticity

$$\omega = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\,. \qquad (1.2.4)$$

Using the condition (1.2.2b), equation (1.2.3) may be rewritten in divergence form as

$$\frac{\partial \omega}{\partial t} + \nabla\cdot(\omega\underset{\sim}{u}) = \frac{1}{Re}\nabla^2\omega\,. \qquad (1.2.5)$$

We introduce a streamfunction ψ such that

$$u = \frac{\partial \psi}{\partial y} \tag{1.2.6a}$$

and

$$v = -\frac{\partial \psi}{\partial x} . \tag{1.2.6b}$$

From equations (1.2.6) we see that the incompressibility condition (1.2.2b) is automatically satisfied. Substituting the equations (1.2.6) into the definition of vorticity (1.2.4), we obtain a simple relationship between the streamfunction and the vorticity

$$\omega = -\nabla^2 \psi \quad . \tag{1.2.7}$$

Since the boundary of the cavity is a stream-line, the streamfunction is constant there. From the differential definition (1.2.6) of the stream-function, its value is only determined to within an additive constant. We set $\psi = 0$ on the boundary. This choice is helpful in some of the later computation. All the boundary conditions for equations (1.2.3) and (1.2.7) are available, namely the values of the streamfunction and of its normal derivative at the boundary,

$$\begin{aligned}
&x = 0, \quad 0 < y < a, \quad \psi = 0, \quad \frac{\partial \psi}{\partial x} = 0 , \\
&y = 0, \quad 0 < x < 1, \quad \psi = 0, \quad \frac{\partial \psi}{\partial y} = 0 , \\
&x = 1, \quad 0 < y < a, \quad \psi = 0, \quad \frac{\partial \psi}{\partial x} = 0 , \\
&y = a, \quad 0 < x < 1, \quad \psi = 0, \quad \frac{\partial \psi}{\partial y} = 1 .
\end{aligned} \tag{1.2.8}$$

Because the derivatives of the streamfunction tangential to the walls are all zero at the walls, the vorticity there is defined by a reduced form of equation (1.2.7)

$$\omega = - \frac{\partial^2 \psi}{\partial n^2} \qquad (1.2.9)$$

where n is the direction normal to the wall.

1.3. DESCRIPTION

The major flow in the cavity consists of a large primary eddy which fills most of the square cavity and depending on the Reynolds number, there may be two smaller and very much weaker counter-rotating eddies in the lower corners. If the cavity is deeper then large secondary counter-rotating eddies may develop (see Pan and Acrivos [33]) in the lower portions, again with smaller and weaker eddies in the lower corners.

For a square cavity (the only sort covered in this work) and a very low Reynolds number, the vortex centre of the large primary eddy is located at about three-quarters of the cavity height and along the vertical centre line of the cavity. A pair of smaller and very much weaker counter-rotating eddies develop in the lower corners of the cavity. As the Reynolds number increases the vortex centre moves away from the centre line in the direction of the local flow (left to right in this region) and down. With further increases of the Reynolds number, the vortex centre moves further down and towards the centre of the cavity.

As the Reynolds number increases and the vortex centre moves towards the centre of the cavity, the value of the vorticity across the central region of the cavity becomes approximately uniform. This behaviour is explained by Batchelor's [3] proposal that as the viscosity tends to zero (or Reynolds number increases), the flow consists of a recirculating eddy having uniform vorticity over an inviscid core with all the viscous effects being confined to small shear regions near the boundaries.

Batchelor's prediction of uniform viscosity in the central regions of the cavity for high Reynolds numbers is the major motivating force behind the use of graded meshes in the numerical solution. The placement of most of the

mesh points near the boundaries will result in a more efficient collection of information about the important boundary layer features. An evenly spaced grid has to be correspondingly finer overall to describe the same boundary layer phenomena as adequately with consequent excessive calculation in the central regions because of the density of mesh points.

1.4. PREVIOUS INVESTIGATIONS

Burggraf [8] has completed an extensive numerical investigation of the cavity problem. Using a modified relaxation method, he presents results for Reynolds numbers of 0, 100 and 400 using an evenly spaced grid in both directions with grid spacings of 1/10, 1/20, 1/30 and 1/40. His solutions demonstrate the movement of the vortex centre of the primary eddy towards the centre of the cavity as the Reynolds number increases and the development of the secondary eddies in the lower corners of the cavity. Burggraf notes that there is good agreement between the self-similar solution for Stokes' flow in a corner as presented by Dean and Montagnon [13] and modified by Moffatt [30] and the calculated result for the large corner eddy at $Re = 400$. The secondary vortex pattern is completely viscous in nature even though the primary eddy is relatively inviscid.

Pan and Acrivos [33] used the same relaxation technique as Burggraf to obtain numerical solutions to the problem of creeping flow ($Re = 0$) in a cavity of various aspect ratios. Numerical solutions were presented for cavities with aspect ratios of 0.25, 0.5, 1, 2, and 5 using mesh sizes from 0.01 to 0.025. These solutions and those of Burggraf complement the earlier results of Kawaguti [24] for aspect ratios of 0.5, 1 and 2. Unfortunately Kawaguti's results are somewhat inaccurate because of the rather coarse (1/10) mesh size used. In Kawaguti's work, the primary vortex centre moved downstream towards the wall as the Reynolds number increased and secondary eddies did not develop in the lower corners.

Pan and Acrivos found the primary vortex to be symmetrical for all cavity depths considered. In the corners they found a sequence of counter-rotating eddies of decreasing vortex strength and size. Moffatt's work was used to calculate the structure of the secondary eddies once a converged

solution had been obtained for the primary flow. Because of numerical instability problems encountered by Burggraf for Reynolds numbers greater than 400, Pan and Acrivos did not extend their numerical experiments past the creeping flow problem.

Reported in the same paper are flow visualisation studies attempted by Pan and Acrivos for cavity flow over a wide range of Reynolds numbers. For a square cavity the Reynolds number attempted ranged from 80 to 4,000, the upper limit being the point at which flow instability began to appear. The experiments produced flows that are consistent with Batchelor's proposals. The numerical results available agreed with the flow visualisation studies carried out for the parameters involved.

Mills [29] is reported to have examined the cavity problem both numerically and experimentally for a Reynolds number of 100 and aspect ratios of 0.5, 1 and 2. Full copies of his work were not able to be obtained by this author though Donovan [14] includes some of Mills' flow visualisations in his paper. His work is also mentioned in Burggraf [8] and the reference is included for completeness.

Greenspan [22] considers the cavity problem numerically using a generalized Newton's method with over-relaxation. He obtained solutions for Reynolds numbers of 200, 500, 2,000 and 15,000 using a mesh spacing of 1/20 and for Reynolds number of 50, 10^4 and 10^5 using a mesh spacing of 1/40. Secondary eddies in the lower corners were not found for the calculations performed using the mesh size 1/20 for any Reynolds numbers yet other authors (Pan and Acrivos [33], Bozeman and Dalton [5]) have found such eddies. In Dorr [16], a one-dimensional analogue of the Navier-Stokes equations is studied with a view to examining the convergence of different finite difference

representations of this analogue. Dorr found that the resultant algebraic equations could become badly ill-conditioned if care were not taken. He demonstrates an example that shows one cannot always determine whether an iterative method has converged by simply looking at the difference between successive iterates. This convergence criterion is used by Greenspan and is not felt to be adequate for accurate solutions of the cavity problem by his method.

Donovan [14]solves the time dependent physical equations rather than using the usual streamfunction-vorticity formulation. He uses a combination of an explicit time stepping method and an over-relaxation technique to solve the coupled equations for pressure and velocity. Solutions were obtained using a mesh width of 1/20 for a Reynolds number of 100 and aspect ratios of 0.5, 1 and 2. For a square cavity he also obtains solutions for Reynolds numbers varying from 100 to 500. The solutions demonstrate the movement of the vortex centre of the primary eddy towards the centre of the cavity as described by Burggraf but there is no indication of the development of secondary eddies in the lower corners for any Reynolds numbers.

Marshall and Van Spiegel [28] attack the streamfunction-vorticity formulation of the problem by perturbing the streamfunction equation into a time dependent equation where the time derivative of the streamfunction is multiplied by a small positive parameter. The vorticity equation and the perturbed equation are solved explicitly on an even spatial grid of mesh width 1/10 for Reynolds numbers between 0 and 200 and mesh widths 1/20 and 1/40 for a Reynolds number of 400. Excessive computing time made solution of the equations impractical for Reynolds numbers greater than 400. The development of the flow is close to that of Burggraf but the development of the secondary eddies in the lower corners is not well pronounced for the

lower Reynolds numbers possibly because of the coarse mesh size.

Bozeman and Dalton [5] have compared the effects of different methods of differencing the vorticity equation written in either divergence or convective form. The two methods of differencing the non-linear terms ($\nabla\cdot(\omega\underset{\sim}{u})$ and $\underset{\sim}{u}\cdot\nabla\omega$ respectively) in either form of the equation were central differences using second order correct difference quotients and unidirectional differences using first order correct difference quotients which are backwards with respect to the local direction of flow. The boundary values of the vorticity were calculated using a third order correct formula in preference to the usual first or second order correct formulas. The equations were solved by the strongly implicit procedure (SIP) of Stone [40].

The central difference, divergence form of the equation demonstrated clear superiority for a Reynolds number of 100 and mesh sizes between 1/20 and 1/50. Both central difference formulas failed to satisfy the convergence criterion (residual less than a specified value) for $Re = 1{,}000$ while both unidirection forms did.

The solution obtained with unidirectional differences and divergence form is consistent with Batchelor's model and similar to previous flows reported while that in convective form was inconsistent with the expected flow, there being two large vorticies instead of one occupying the cavity. The superiority of the divergence form is also mentioned in Torrance et al. [43]. Convergence was not obtained for Reynolds numbers greater than 1,000 for any method.

Related work has been done on the cavity problem with thermal

convection added. Thermal boundary conditions treated have been, after non-dimensionalizing, (A) zero on all walls except the moving upper surface where the temperature is unity, and (B) zero on the bottom, unity on the top with a continuous linear variation along the side walls. Condition (B) was considered by Runchal, Spalding and Wolfshtein [38]. The equations were written in a finite difference form involving unidirectional derivatives that led to conservation of momentum and energy over the grid and to positive definite equations. These equations were solved by relaxation techniques on a 13 x 13 non-uniform grid for Reynolds numbers of 1 and 10^3. Unfortunately the method for choosing the graded mesh was not explained and general rules were not proposed. The results agree quite well with earlier work and with Batchelor's model for large Reynolds numbers.

Torrance et al. [43] examine the combined effects of a moving wall and natural convection via case (A) for aspect ratios of 0.5, 1 and 2 and for Reynolds numbers of 100, a Prandtl number of 1 and for various Grashof numbers including zero. The equations were written in divergence form. Forward time and central space differences were used for all terms except the convection terms for which special three point non-central differences were employed. Explicit time stepping is used to solve the time dependent equations but only the steady state solutions are presented. The Poisson equation for the stream function is solved by over-relaxation. A mesh spacing of 0.05 is used for an aspect ratio of unity. Velocity profiles and streamfunction values are in close agreement with Donovan [14] and in fair agreement with the works of Mills [29] and Kawaguti [24] who used comparatively coarser mesh spacings. It is suggested that a mesh interval of 0.05 by Torrance et al yields results comparable to a mesh spacing of 0.028 using Burggraf's method. The principle reason suggested for the better

results is the use of a finite difference representation of the convection term (written in divergence form $\nabla\cdot(\omega\underset{\sim}{u})$) that conserves vorticity within the grid. Vorticity patterns were not presented and could not be compared.

For completeness and also to give an example of a possible application of this work, Fromm[19]considers case (B) but without a moving upper surface. He considers the time dependent vorticity and energy equations with the Boussinesq approximation. These equations are differenced to fourth order accuracy and solved explicitly. The Poisson equation for the stream function is solved by the Buneman direct method (see the review by Dorr [15]) over a 65 x 65 grid. Batchelor's model as discussed in Burggraf [8] implies the temperature will be uniform to first order in a closed cavity even for a non-circular eddy. Fromm's solutions for a range of Grashof and Prandtl numbers bear out this result.

Chorin [9] has examined the convergence of discrete approximations to the Navier-Stokes equations. Besides excluding turbulence from the range of application of difference methods (see Chorin[10]) his discussion in [9] suggests that there is no good reason for always casting the non-linear terms of the Navier-Stokes equations in *"conservation form"*, i.e. in a form which implies the existence of identities for the momentum similar in appearance to those which hold for the solutions of the differential equations. Chorin in fact has not endeavoured to do so but approximates his equations by the analytically most accurate formula compatible with the solution scheme being used.

CHAPTER 2

DIRECT SOLUTION OF THE POISSON EQUATION

2.1. INTRODUCTION

The streamfunction-vorticity relationship takes the form of a Poisson equation with Dirichlet boundary conditions,

$$\nabla^2\psi = -\omega \qquad (2.1.1)$$

$$\text{and} \quad \psi = 0 \quad \text{on the boundary .}$$

In the iteration scheme to solve the two coupled equations describing the cavity problem flow, equation (2.1.1) needs to be solved for ψ given the values of ω inside the rectangle. Details of the complete numerical scheme are given in Chapter 3. In this scheme equation (2.1.1) is solved repeatedly; considerations of efficiency in its numerical solution become of paramount importance.

Consider Poisson's equation on an $N \times N$ evenly spaced grid over the unit square. The relative computing costs of several different methods are displaced in Table 2.1.1. In this table and in the ensuing work, one computer operation is defined to be a floating-point multiply and add. Some of the estimates used there can be found in [27]. The computational costs of two iterative methods are displaced in Table 2.1.1. The operation counts for these methods have been calculated for asymptotic rates of convergence to the same relative accuracy as the difference equation approximates the differential equation. It is assumed that optimal parameters are used in these iterative methods.

TABLE 2.1.1

Operation counts for the solution of Poisson's equation on a rectangle.

Method of Solution	*Operation Count*
Optimal Successive Over-relaxation	$14N^3 \log_{10} N$
Bickley-McNamee [4] (PP)	$4N^3 + O(N^2)$
Tensor Product [27] (PP)	$4N^3 + O(N^2)$
Alternating Direction Implicit	$40N^2 \log_{10}^2 N$
Direct Method (PP)	$2N^3 + O(N^2)$
Direct method using Winograd's algorithm for a UNIVAC 1108 computer. (PP)	$1.76N^3 + O(N^2)$
Direct method using conventional FFT*	$4N^2 \log_2 N + O(N^2)$
Direct method using new FFT*	$O(N^2 \log_2(\log_2 N)) + O(N^2)$

(PP) This method has a pre-processing overhead not shown in the operation count (see text).

* FFT = Fast Fourier Transform.

From Table 2.1.1, it is obvious that the algorithms using the FFT are significantly more efficient than other methods. However, in this work we are specially concerned with graded meshes and the special forms that allow the FFT algorithm to be used depend critically on the formalism of the problem when an even grid is used. If a graded mesh is used in this problem then the direct method suggested in this work, including the use of Winograd's matrix multiplication algorithm, is the most efficient method available.

A detailed explanation of the notation and the structure of the finite difference scheme used is given in §2.2. In §2.3 a formal analytic solution of the Poisson equation is derived and it is then shown that this formal solution has a computationally advantageous counterpart in the finite difference formulation. For an even mesh the special advantages gained by the use of Fast Fourier Transform are demonstrated in §2.4 while in §2.5, Winograd's method for matrix multiplication is examined for efficiency. This method is advantageous in the multiplication of medium size full matrices and such matrices are of special concern in the cavity problem.

2.2. NOTATION FOR THE POISSON EQUATION

The problem considered here is slightly more general than the streamfunction-vorticity equation with its zero boundary conditions. Without loss of any generality we restrict attention to a square region and consider the Poisson equation

$$\frac{\partial^2 u}{\partial x^2}(x,y) + \frac{\partial^2 u}{\partial y^2}(x,y) = f(x,y) \qquad (2.2.1)$$

for $(x,y) \in R = (0,1) \times (0,1)$ with the Dirichlet boundary conditions

$$u(x,y) = g(x,y)$$

for $(x,y) \in \partial R$.

Consider a mesh where the lines parallel to the y-axis have abscissae

$$0 = x_0 < x_1 < x_2 < \ldots < x_n < x_{n+1} = 1 , \qquad (2.2.2)$$

and the lines parallel to the x-axis have ordinates

$$0 = y_0 < y_1 < y_2 < \ldots < y_m < y_{m+1} = 1 . \qquad (2.2.3)$$

The following method does not require evenly spaced grid lines but for simplicity of presentation, we consider only an evenly spaced grid in this chapter. Where differences would arise because of unequally spaced grid lines, those differences will be noted in the text.

Let h and k be the spacings between the grid lines in the x and y directions respectively. If the value of a function $u(x,y)$ at the mesh point (x_i,x_j) is represented by the element $u_{ij} = u(x_i,y_i)$ of a matrix for appropriate i and j , then all the function values at mesh points in

the interior of the unit square are in the matrix

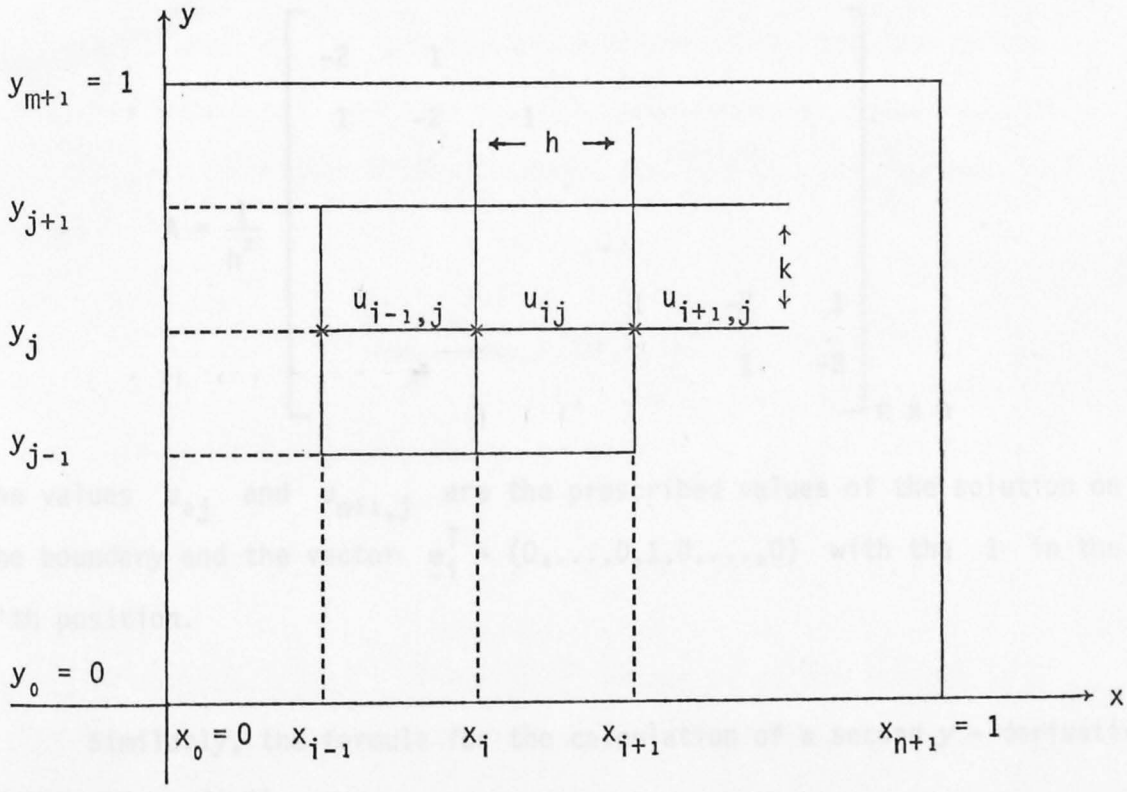
$$U = [\underset{\sim}{u}_1, \underset{\sim}{u}_2, \ldots, \underset{\sim}{u}_m] ,$$

where

$$\underset{\sim}{u}_j^T = (u_{1j}, u_{2j}, \ldots, u_{nj})$$

for $j=1,2,\ldots,m$. The components of this vector can be identified with the values of the function u at the mesh points along the line $y = y_j$ as in Figure 2.1.1.

FIGURE 2.1.1



The standard central difference representation of $\frac{\partial^2 u}{\partial x^2}$ at the grid point (i,j) is

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{(i,j)} = \frac{u_{i-1,j} - 2u_{ij} + u_{i+1,j}}{h^2} . \qquad (2.2.4)$$

This is an operation on the elements of the vector $\underset{\sim}{u}_j$. Hence the central difference analogue of the operator $\left.\frac{\partial^2}{\partial x^2}\right|_{y = y_j}$ can be represented by a matrix pre-multiplication with some correction terms to account for the known boundary values:

$$\left(\frac{\partial^2 \underset{\sim}{u}}{\partial x^2}\right)_j = A\underset{\sim}{u}_j + \frac{u_{0,j}}{h^2}\ \underset{\sim}{e}_1 + \frac{u_{n+1,j}}{h^2}\ \underset{\sim}{e}_n$$

where the vector $\left(\frac{\partial^2 \underset{\sim}{u}}{\partial x^2}\right)_j$ contains the values of $\frac{\partial^2 u}{\partial x^2}(x_i, y_j)$ as components for $i=1,2,\ldots,n$. The matrix A is defined by

$$A = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{bmatrix}_{n \times n}$$

The values u_{0j} and $u_{n+1,j}$ are the prescribed values of the solution on the boundary and the vector $\underset{\sim}{e}_i^T = (0,\ldots,0,1,0,\ldots,0)$ with the 1 in the i'th position.

Similarly, the formula for the calculation of a second y - derivative at the point (i,j),

$$\left(\frac{\partial^2 u}{\partial y^2}\right)_{(i,j)} = \frac{u_{i,j-1} - 2u_{ij} + u_{i,j+1}}{k^2},$$

can be written in vector notation as

$$\left(\frac{\partial^2 \underset{\sim}{u}}{\partial y^2}\right)_j = \frac{1}{k^2}\left(\underset{\sim}{u}_{j-1} - 2\underset{\sim}{u}_j + \underset{\sim}{u}_{j+1}\right) \qquad (2.2.5)$$

where the vector $\left(\frac{\partial^2 u}{\partial y^2}\right)_j$ contains the values of $\frac{\partial^2 u}{\partial y^2}(x_i, y_j)$ for $i=1,2,\ldots,n$. For $j = 1$ and $j = m$, the prescribed boundary values $\underset{\sim}{u}_0$ and $\underset{\sim}{u}_{m+1}$ are used in equation (2.2.5).

The discretised Poisson equation can be written as

$$A\underset{\sim}{u}_j + \frac{1}{k^2}(\underset{\sim}{u}_{j-1} - 2\underset{\sim}{u}_j + \underset{\sim}{u}_{j+1}) = \underset{\sim}{f}'_j \qquad (2.2.6)$$

for $j=2,3,\ldots,m-1$, where

$$\underset{\sim}{f}'_j = \underset{\sim}{f}_j - \frac{1}{h^2} u_{0,j}\, \underset{\sim}{e}_1 - \frac{1}{h^2} u_{n+1,j}\, \underset{\sim}{e}_n$$

where $u_{0,j}$ and $u_{n+1,j}$ are the known values of the solution on the boundary. For $j = 1$ and $j = m$, the known boundary values $\underset{\sim}{u}_0$ and $\underset{\sim}{u}_{m+1}$ must be taken into account. Hence for $j = 1$, the equation (2.2.6) becomes

$$A\underset{\sim}{u}_1 + \frac{1}{k^2}(-2\underset{\sim}{u}_1 + \underset{\sim}{u}_2) = \underset{\sim}{f}'_1$$

where

$$\underset{\sim}{f}'_1 = \underset{\sim}{f}_1 - \frac{u_{01}}{h^2}\underset{\sim}{e}_1 - \frac{u_{n+1,1}}{h^2}\underset{\sim}{e}_n - \frac{1}{k^2}\underset{\sim}{u}_0 ,$$

and similarly for $j = m$.

In equation (2.2.6) linear combinations of column vectors from the matrix U are used in the calculation of the second y-derivative. The above notation allows this operation to be written as a matrix post-multiplication. Thus the equations (2.2.6) for $j=1,2,\ldots,m$ may be written as

$$AU + U\cdot\frac{1}{k^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & \cdot & & \\ & 1 & \cdot & & \\ & & & \cdot & \\ & & & & 1 \\ & & & & -2 \end{bmatrix} = F' \qquad (2.2.7)$$

where

$$F' = [f_1', f_2', \ldots, f_m'] \ .$$

Without ambiguity, the primes may be dropped from the f_i' and the F' in equation (2.2.7). Defining the matrix B as

$$B = \frac{1}{k^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & . & & \\ & 1 & . & & \\ & & & . & \\ & & & 1 & \\ & & & -2 & \end{bmatrix}_{m \times m} ,$$

the discretised Poisson equation with Dirichlet boundary conditions on the given grid becomes

$$AU + UB = F \ . \qquad (2.2.8)$$

Note that both of the matrices A and B are symmetric and tridiagonal. In fact for an evenly spaced grid in the $x(y)$ direction, the matrix $A(B)$ has a constant diagonal and constant and equal sub- and super-diagonals. The special properties that follow from this situation are discussed in §2.4. For a more general graded mesh, the weights in the three term relationship used to calculate the second derivatives change from the simple $\left(\frac{1}{h^2}, \frac{-2}{h^2}, \frac{1}{h^2}\right)$ to a more complicated and unsymmetric pattern.

If we have the mesh (2.2.2) and (2.2.3) and we define

$$h_i = x_{i+1} - x_i \ , \quad i=0,1,2,\ldots,n \ ;$$

then the second x-derivative can be approximated at the (i,j) mesh point by the three term relation

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{(i,j)} = \frac{2u_{i+1,j}}{h_i(h_i + h_{i-1})} - \frac{2u_{ij}}{h_i h_{i-1}} + \frac{2u_{i-1,j}}{h_{i-1}(h_i + h_{i-1})} . \quad (2.2.9)$$

An equivalent expression is used for the second y-derivative. Note that in the central difference formulas (2.2.4) and (2.2.5) for the calculation of a second derivative on an even mesh, the discretisation error is $O(h^2)$. For a graded mesh, the accuracy with which the difference equations approximate the differential equation is degraded to first order. At the (i,j) mesh point, the formula (2.2.9) approximates a second derivative with a minimum discretisation error of $O(h_i - h_{i-1}) + O(h_i^2 + h_{i-1}^2)$. When expanded about the central point by a Taylor series expansion, any other central three point formulas have a larger discretisation error.

With graded meshes even if formulas with minimum discretisation error are used in the approximation of the differential equations, there is an overall loss of accuracy in the approximation of the continuous problem by the difference equations compared with the use of central difference formulae on an equivalent even mesh. This loss of accuracy must be reflected in the solution of this problem. In §4.1, an example is given where a well behaved equation is solved on an even mesh and on a randomly chosen graded mesh. A comparison of these two solutions with the analytic solution clearly demonstrates the effects of careless use of graded meshes.

If a systematic scheme can be found for choosing the mesh points of a graded mesh, then mesh refinement becomes a simple task and extrapolation procedures may possibly be brought into action to improve the accuracy. For appropriate problems such methods may be superior to the use of central

differences on an evenly spaced mesh with a large number of points. For discussion of such problems and for the suggestion of one such technique see Chapter 4.

When graded meshes are used in the solution of Poisson's equation on a rectangle, then central three-point formulas like formula (2.2.9) have the advantage that the resulting matrices A and B in equation (2.2.8) have their tridiagonal form preserved. In general though the symmetry and the constancy of the three diagonals will be lost, resulting in serious consequences for the computational efficiency with which the problem may be solved. These consequences are explored in §2.3 and §2.4.

2.3. DIRECT SOLUTION OF THE POISSON EQUATION

Pre-multiplication of the solution matrix U by the matrix A is equivalent to the application of the operator $L_x = \frac{\partial^2}{\partial x^2}$ to the function $u(x,y)$. Similarly, post-multiplication of the matrix U by the matrix B is equivalent to the application of the operator $L_y = \frac{\partial^2}{\partial y^2}$ to the function $u(x,y)$. With these correspondences in mind, let us examine one method of solution for each of equations (2.2.1) and (2.2.8).

Suppose $\{\phi_k\}_{k=1}^{\infty}$ and $\{\lambda_k\}_{k=1}^{\infty}$ are the eigenfunctions and eigenvalues of the operator L_y, that is,

$$L_y \phi_k = \lambda_k \phi_k , \quad k=1,2,\ldots .$$

The functions $u(\cdot,y)$ and $f(\cdot,y)$ can be expanded in the eigensystem as

$$u(x,y) = \sum_{k=1}^{\infty} \bar{u}_k(x)\, \phi_k(y) \tag{2.3.1a}$$

and

$$f(x,y) = \sum_{k=1}^{\infty} \bar{f}_n(x)\, \phi_k(y) \tag{2.3.1b}$$

for appropriate coefficients $\bar{u}_k(x)$ and $\bar{f}_k(x)$. Substituting the equations (2.3.1) into equation (2.2.1), we obtain

$$L_x\, u(x,y) + L_y\, u(x,y) = \sum_{k=1}^{\infty} \{(L_x + \lambda_k)\, \bar{u}_k(x)\}\, \phi_k(y)$$

$$= \sum_{k=1}^{\infty} \bar{f}_k(x)\, \phi_k(y) . \tag{2.3.2}$$

Since L_y is self-adjoint and the ϕ_k are mutually orthogonal and hence linearly independent, the coefficients of ϕ_k in equation (2.3.2) must be identically zero. Hence we obtain the set of ordinary differential equations

$$(L_x + \lambda_k)\, \bar{u}_k(x) = \bar{f}_k(x) , \quad k=1,2,\ldots .$$

These equations can be solved by any suitable method, then knowing the solution functions $\{\bar{u}_k(x)\}_{k=1}^{\infty}$, equation (2.3.1a) may be used to reform the solution $u(x,y)$. Note that the eigenfunctions of the operator $L_y = \frac{\partial^2}{\partial y^2}$ are sines in this case; this will be important in §2.4 in the discussion of the use of the Fast Fourier Transform.

Consider now the finite difference equivalent of the above scheme. The eigensystem of the matrix B is

$$B = Q \Lambda Q^T \tag{2.3.3}$$

where the matrix Q is orthogonal, that is $QQ^T = I$ and Q^T means the transpose of Q. Let Q be partitioned into column vectors as

$$Q = [\underset{\sim}{q}_1, \underset{\sim}{q}_2, \ldots, \underset{\sim}{q}_m]$$

where $\underset{\sim}{q}_i^T = (q_{1i}, q_{2i}, \ldots, q_{mi})$ for $i=1,2,\ldots,m$. If equation (2.3.3) is substituted into equation (2.2.8), then we have

$$AU + UQ \Lambda Q^T = F .$$

Post-multiply this equation by the matrix Q and defining

$$\bar{U} = UQ \tag{2.3.4a}$$

and

$$\bar{F} = FQ , \tag{2.3.4b}$$

we obtain

$$A\bar{U} + \bar{U} = \bar{F} . \tag{2.3.5}$$

Note that equation (2.3.4) is the finite difference equivalent of finding the coefficients $\bar{u}_k(x)$ and $\bar{f}_k(x)$ from the formula

$$\bar{u}_k(x) = \int_0^1 u(x,y)\, \phi_k(y)\, dy$$

for the continuous case. Compare this with the expansion of equation (2.3.4a) as

$$\bar{u}_{ik} = \sum_{j=1}^{m} u_{ij}\, q_{jk}$$

where the $\bar{u}_{ik}$ are the coefficients of the eigenvectors $\underset{\sim}{q}_j$ in the expansion of U.

Take the j'th column of equation (2.3.5) to obtain the separated equations

$$(A + \lambda_j I)\underset{\sim}{\bar{u}}_j = \underset{\sim}{\bar{f}}_j \qquad (2.3.6)$$

for $j=1,2,\ldots,m$. Each equation of (2.3.6) is the finite difference equivalent of an ordinary differential equation in x for the transformed functions $\bar{u}_k$. Each is a simple symmetric tridiagonal linear system of equations in n variables. Using simple Gaussian elimination on the tridiagonal system requires $5n$ operations for its solution. These are m systems in equation (2.3.6) hence a total of $5nm$ operations are necessary to solve for the matrix $\bar{U}$. Then from equation (2.3.4a) we find

$$U = \bar{U}Q^T .$$

This matrix multiplication takes nm^2 operations as does the multiplication in equation (2.3.4b). These two matrix multiplications and the solutions of equations (2.3.6) make a total of

$$2nm^2 + 5nm \text{ operations}$$

to solve equation (2.2.8). The computational overhead of finding the eigensystem of the matrix B is not included in this total. Since the matrix B depends only on the operator L_y and on the y-spacing in the grid, if the y-mesh is not changed, the eigensystem of the matrix B need be found only once. For the problem discussed in this work, as the Poisson equation is solved for many right hand sides on the same grid, the overhead

for finding the eigensystem of B is considered a preprocessing overhead. Note that for an evenly spaced y-mesh the FFT algorithm can be used as explained in §2.4, and the eigensystem of B is never explicitly found so that this preprocessing overhead does not exist in that case.

The method of solution of equation (2.2.1) using an evenly spaced mesh in at least one direction as discussed above is not adequate if unequally spaced meshes are used in both directions. In this case the finite difference equations arising from equation (2.2.1) has a matrix B which is still tridiagonal but is not now symmetric. The previous method of solution hinged on the fact that the left and right eigenvectors of B were identical. This is a consequence of the symmetry of the matrix B. For unsymmetric B with positive off-diagonal elements, there exists a diagonal similarity transformation that changes B to a symmetric tridiagonal matrix S, namely

$$B = DSD^{-1}$$

where D is a diagonal matrix.

If we substitute this into equation (2.2.8), then after post-multiplying by the diagonal matrix D, we have the equation

$$AUD + UDS = FD .$$

By defining $U_1 = UD$ and $F_1 = FD$, this last equation becomes

$$A U_1 + U_1 S = F_1 .$$

The matrix S is symmetric so this equation has the same form as equation (2.2.8). Hence the previous method of solution can be used to solve for U_1 from which U is found by the diagonal matrix multiplications $U = U_1 D^{-1}$. The extra computational cost caused by the use of graded meshes is the $2nm$

multiplications for the two diagonal matrix multiplications. This makes the total operation count only slightly larger. In fact, it is dwarfed by the major cost of the solution - the $2nm^2$ operations for the two full matrix multiplications.

We now consider the possibilities to reduce this cost for the two separate cases. In §2.4, we examine the application of the Fast Fourier Transform when even grids are used in at least one direction. In §2.5 we examine an efficient method for the multiplication of full matrices.

2.4. USE OF THE FAST FOURIER TRANSFORM

For a mesh with equally spaced grid lines in the y-direction, the matrix B in the discretised Poisson equation (2.2.8) has the form

$$B = \frac{1}{k^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & . & & \\ & 1 & . & & \\ & & & . & \\ & & & 1 & \\ & & & 1 & -2 \end{bmatrix}_{m \times m} . \qquad (2.4.1)$$

The eigenvalues of this matrix B are

$$\lambda_j = \frac{1}{k^2}\left(-4 \sin^2 \frac{j\pi}{2(m+1)}\right) , \quad j=1,2,\ldots,m ;$$

and the eigenvectors $\underset{\sim}{q}_j$ have components

$$q_{ij} = c \sin \frac{ij\pi}{m+1} , \quad \begin{matrix} i=1,2,\ldots,m ; \\ j=1,2,\ldots,m ; \end{matrix}$$

where c is a common normalising factor. For the method of solution introduced in §2.3, let us examine one of the matrix multiplications FQ or $\overline{UQ}^T$, say $\overline{F} = FQ$. For one element of $\overline{F}$ we have

$$\overline{f}_{ij} = \sum_{\ell=1}^{m} f_{i\ell}\, q_{\ell j}$$

$$= c \sum_{\ell=1}^{m} f_{i\ell} \sin \frac{\ell j\pi}{m+1} . \qquad (2.4.2)$$

Let the rows of the matrix F be the vectors $\underset{\sim}{f}_i^*$ so that

$$\underset{\sim}{f}_i^* = (f_{i1}, f_{i2}, \ldots, f_{im}) , \quad i=1,2,\ldots,n .$$

If these row vectors are considered as vectors of data, then equation (2.4.2) is just a discrete sine transform of that data. This should not be surprising since the eigenfunctions of the operator $\frac{\partial^2}{\partial y^2}$ are sines in this case and we are here discussing the finite difference equivalent of that operator, the matrix B.

A discrete sine transform can be performed very quickly just by taking the FFT of the $\underset{\sim}{f}_i^*$ as real data and taking the imaginary component of the complex result or, as explained in [11], the FFT algorithm [12] can be used to perform just a sine transform on real data for little extra work. Hence the matrix multiplications FQ and $\overline{U}Q^T$ can each be performed for a computational cost of

$$2nm \log_2 m \text{ operations} .$$

This is a distinct reduction from the nm^2 operations needed for those matrix multiplications by the usual inner product method. There is the added bonus in the solution of the Poisson equation of no preprocessing overhead to find the eigensystem Q and Λ of the matrix B as there is in the method in §2.3.

With an even grid in the y-direction, the Poisson equation over a square region with Dirichlet boundary conditions can be solved in

$$4nm \log_2 m + 5nm \text{ operations} .$$

There are reports [21] of a new stable method for the FFT algorithm that takes $O(m \log_2(\log_2 m))$ operations for m data points. The use of this new algorithm instead of the usual FFT algorithm ($O(m \log_2 m)$ operations for m data points) would decrease the above operation count even further. Since at least one operation must be performed for each of the m data points,

the theoretical lower limit for the FFT algorithm is $O(m)$ operations which would imply $O(nm)$ operations for the Poisson equation. Methods which achieve this are not currently known.

For the Poisson equation as above but also with an even grid in the x-direction, the matrix A also has the form of (2.4.1). Hence for solving the set of equations

$$(A + \lambda_i I)\underset{\sim}{\bar{u}}_i = \underset{\sim}{\bar{f}}_i \ , \quad i=1,2,\ldots,m \ ;$$

the same use may be made of the FFT algorithm. Because the FFT algorithm takes $O(n \log_2 n)$ for these equations and simple Gaussian elimination takes $O(n)$, the use of the FFT algorithm in this case is not recommended.

The FFT algorithm is not limited only to the Poisson equation. The method described in §2.3 may be applied to any second order linear separable elliptic operator and the FFT algorithm may be used to perform the appropriate matrix multiplications if the resultant matrix B has the form of (2.4.1). This will occur for example if the operator $L_y = \frac{\partial^2}{\partial y^2} + c$ where c is a constant and an even mesh is used in the y-direction.

2.5. WINOGRAD's METHOD OF MATRIX MULTIPLICATION

Since the major computational cost in the solution of equation (2.2.1) is in the matrix multiplications FQ and $\overline{U}Q^T$, a more efficient method of matrix multiplication than the usual inner product method would be welcome. We will demonstrate an algorithm for matrix multiplication which uses Winograd's identity [45] for an inner product. This algorithm performs the multiplication of two $n \times n$ matrices in less than n operations.

For n even (extend the vectors to length $n + 1$ by adding a zero as the last component if n is odd), Winograd's identity for the inner product of two n vectors $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ is: if

$$\alpha = \sum_{j=1}^{n/2} x_{2j} \cdot x_{2j-1}$$

and

$$\beta = \sum_{j=1}^{n/2} y_{2j} \cdot y_{2j-1}$$

are known, then Winograd's identity is

$$\sum_{i=1}^{n} x_i\, y_i = \sum_{j=1}^{n/2} (x_{2j} + y_{2j-1})(x_{2j-1} + y_{2j}) - (\alpha + \beta) .$$

On the left of this identity is the normal inner product formula which requires n multiplications and $n - 1$ additions (we make no distinction between addition and subtraction). On the right of the identity is Winograd's form of the inner product. This requires $\frac{n}{2}$ multiplications and $\frac{3n}{2} + 1$ additions. In Winograd's formula, half $\left(\frac{n}{2}\right)$ of the multiplications are exchanged for slightly more than half $\left(\frac{n}{2} + 2\right)$ additions. Hence the extra speed of Winograd's algorithm over the usual method depends on the relative times for the multiply and add operations on a computer for

whatever data types being used, that is, single or double precision, real or complex numbers.

For the matrix multiplication $Z = XY$ say, of two $n \times n$ matrices X and Y, we compute a number α_i for each row of the matrix X and a number β_j for each column of the matrix Y at a computational cost of $\frac{n}{2}$ multiplications and $\frac{n}{2} - 1$ additions for each of these $2n$ numbers. We then use these numbers in the computation of the n^2 inner products needed to form the matrix Z,

$$z_{ij} = \sum_{k=1}^{n/2} (x_{i,2k} + y_{2k-1,j})(x_{i,2k-1} + y_{2k,j}) - (\alpha_i + \beta_j)$$

for $i=1,2,\ldots,n$ and $j=1,2,\ldots,m$. This method takes $\frac{n^3}{2} + n^2$ multiplications and $\frac{3n^3}{2} + 2n(n - 1)$ additions compared with the usual n^3 multiplications and $n^3 - n^2$ additions.

If on a certain computer

$$f = \frac{\text{time for multiply}}{\text{time for add}},$$

since an operation is one multiply plus one add then one operation is $f + 1$ adds. If W is the computational cost of an $n \times n$ matrix multiplication using Winograd's identity and if IP is that computational cost with the usual inner product, then neglecting terms of order n for simplicity, we have

$$W = \frac{f\left(\frac{n^3}{2} + n^2\right) + \left(\frac{3n^3}{2} + 2n^2\right)}{f + 1}$$

and

$$IP = \frac{fn^3 + n^3 - n^2}{f + 1}.$$

Both W and IP are in units of operations. The condition for some savings in time by the use of Winograd's identity is

$$\frac{W}{IP} < 1 .$$

For a given f this condition is satisfied if the order of the matrices satisfies

$$n > 2\,\frac{f + 3}{f - 1} .$$

If we let $W/IP = R$ the relative efficiency of the two methods, then Figure 2.5.1 presents a graph of relative efficiency versus order of the matrices for various values of the machine constant f.

For a Univac 1108 computer with $f = 1.625$ for single precision floating point arithmetic, we find that n must be greater than 14 for some saving to be made. For larger n on a Univac 1108, the cost for Winograd's method is

$$W = 0.881n^3 + 1.381n^2$$

compared with the inner product method's cost of

$$IP = n^3 - 0.381n^2 .$$

Winograd's method can be important when matrix multiplications have to be performed on very large matrices or a number of times on medium sized matrices. But a word of warning is necessary. The compiler used and the machine scheduling algorithms can affect both times by a large variable amount. As well, Brent [6] has shown that unless the matrices being multiplied by Winograd's method have first been balanced, even if only roughly, then disasterous rounding errors can occur. With balancing, the accuracy of Winograd's method is about the same as the usual inner product method with double precision accumulation. This factor only marginally affects the computational costs because the necessary scaling is a process that takes $O(n^2)$ operations with a small constant multiplying the n^2.

Hence the break-even point for the use of Winograd's method is just raised slightly.

One application of Winograd's method is in the iterative use of equation (2.2.1) since each time the equation is solved, there are two matrix multiplications to perform. Assuming that for an $n \times n$ system, the size n is large enough to satisfy the requirements of Winograd's algorithm on the computer being used, then some small but significant saving may be had. For example, on a Univac 1108 the cost of solving equation (2.2.1) reduced to

$$1.77n^3 + O(n^2) \text{ operations}$$

compared with the usual

$$2n^3 + O(n^2) \text{ operations.}$$

Another method for the efficient multiplication of very large full matrices is that of Strassen [41]. The computations for the multiplication of 2×2 matrices are rearranged to take 7 (usually 8) multiplications and 18 (usually 4) additions. This rearrangement does not depend on the commutativity of the objects being multiplied (as does Winograd's method) and so may be applied recursively to block matrices of size 2^k to perform matrix multiplication in $O(n^{\log_2 7}) \sim O(n^{2.8})$ operations. Strassen's method with one depth of recursion is faster than the normal method for $n \sim 100$ or higher depending on the machine. Significant gains of 7% to 13% are realised by Winograd's method for matrices of this order so that Strassen's method which is much more difficult to code only becomes advantageous for much larger matrices. Brent [7] discusses efficient methods for matrix multiplication and for an IBM 360/67, his formulas suggest that Strassen's method with one depth of recursion overtakes Winograd's method for real arithmetic at about $n \sim 440$.

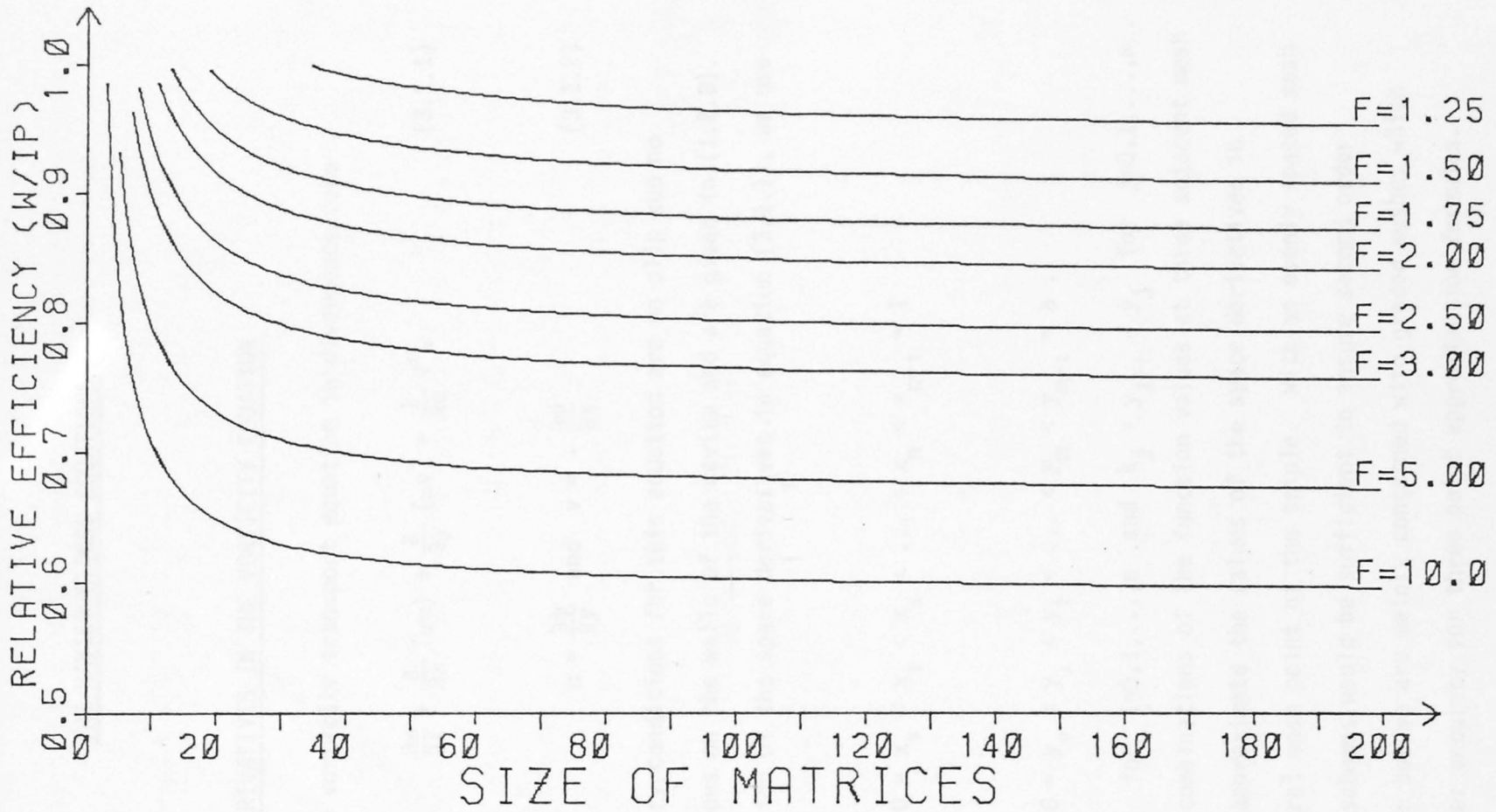


Figure 2.5.1 Graphs of the theoretical relative efficiency (W/IP) of Winograd's algorithm for matrix multiplication to the usual inner product method versus the size (n) of the matrices for various values of the machine parameter F = time for multiply / time for addition.

CHAPTER 3

THE NAVIER-STOKES EQUATIONS

3.1. THE SPACE DERIVATIVES IN THE VORTICITY EQUATION

Consider the vorticity transport equation in divergence form

$$\frac{\partial \omega}{\partial t} + \frac{\partial}{\partial x}(\omega u) + \frac{\partial}{\partial y}(\omega v) = \frac{1}{Re}\nabla^2 \omega \qquad (3.1.1)$$

where

$$u = \frac{\partial \psi}{\partial y} \quad \text{and} \quad v = -\frac{\partial \psi}{\partial x} . \qquad (3.1.2)$$

The physical boundary conditions for this equation are no slip and no penetration conditions on the walls of the cavity and are given in (1.2.8). For the discretisation of the space derivatives in equation (3.1.1), we use the graded mesh

$$0 = x_0 < x_1 < x_2 < \ldots < x_n < x_{n+1} = 1$$

and

$$0 = y_0 < y_1 < y_2 < \ldots < y_m < y_{m+1} = a .$$

Let $h_i = x_{i+1} - x_i$ for $i=0,1,\ldots,n$ and $k_j = y_{j+1} - y_j$ for $j=0,1,\ldots,m$. Appropriate linear combinations of the function values at three adjacent mesh points are used to approximate the values of the space derivatives in (3.1.1) at the central mesh point of the triple. With an evenly spaced mesh central difference schemes would be sufficient to ensure second order discretisation error but we are mainly concerned with graded meshes which have only first order accuracy for three point approximation formulas.

The first term approximated from equation (3.1.1) is the difference term $\nabla^2\omega$. From equation (2.2.9) the three point approximation formula for a second derivative over a general graded mesh leads to the formula

$$(\nabla^2\omega)_{i,j} = \frac{2\omega_{i+1,j}}{h_i(h_i + h_{i-1})} - \frac{2\omega_{i,j}}{h_i\, h_{i-1}} + \frac{2\omega_{i-1,j}}{h_{i-1}(h_i + h_{i-1})}$$

$$+ \frac{2\omega_{i,j+1}}{k_j(k_j + k_{j-1})} - \frac{2\omega_{i,j}}{k_j\, k_{j-1}} + \frac{2\omega_{i,j-1}}{k_{j-1}(k_j + k_{j-1})} \qquad (3.1.3)$$

The discretisation error for this formula is minimum in the sense that, for a general function that has fourth order continuous derivatives over a general graded mesh, any other pair of three point formulas approximate second derivatives to a lower order of accuracy. For the above formula (3.1.3), the discretisation error is $O(h_i - h_{i-1}) + O(k_j - k_{j-1})$. If the mesh is equispaced in both directions, the above formula reduces to the standard five point approximation to the Laplacian

$$(\nabla^2\omega)_{i,j} = \frac{\omega_{i-1,j} - 2\omega_{i,j} + \omega_{i+1,j}}{h^2}$$

$$+ \frac{\omega_{i,j-1} - 2\omega_{i,j} + \omega_{i,j+1}}{k^2} \qquad (3.1.4)$$

which has discretisation error $O(h^2) + O(k^2)$ where h and k are the mesh spacings in the x and y directions respectively.

At points of the mesh for which $i = 1$ or n or $j = 1$ or m, the values of the vorticity on the boundary are needed in the calculation of the diffusion term. The boundary values for the problem do not include the vorticity on the boundary but the wall vorticity can be approximated numerically from the streamfunction values and the boundary values for the

normal velocity component. From equation (1.2.9) the vorticity at the walls is expressed by

$$\omega = - \frac{\partial^2 \psi}{\partial n^2} \tag{3.1.5}$$

where n is the normal to the wall. Suppose the streamfunction values are known at all mesh points. The first normal derivative of the streamfunction $\frac{\partial \psi}{\partial n}$ is known at the walls from the no slip condition. The wall vorticity (3.1.5) can be approximated by a three point relation with second order discretisation error

$$\omega_0 = \alpha \psi_0 + \beta \psi_1 + \gamma \psi_2 + \delta \left(\frac{\partial \psi}{\partial n} \right)_0$$

where the subscripts denote function values at different mesh lines away from the wall, 0 being the wall line, and where

$$\left.\begin{aligned} \alpha &= - (\beta + \gamma) , \\ \beta &= - \frac{2(h_0 + h_1)}{h_0^2 h_1} , \\ \gamma &= \frac{2h_0}{(h_0 + h_1)^2 h_1} , \\ \delta &= \frac{-2(2h_0 + h_1)}{(h_0 + h_1) h_0} \end{aligned}\right\} \tag{3.1.6}$$

where h_0 is the distance of the first mesh line from the wall and h_1 is the distance of the second mesh line from the first.

The streamfunction is determined only to within an additive constant and the choice of the streamfunction $\psi = 0$ on the boundary fixes this constant and means that the formula for the wall vorticity reduces to the expression

$$\omega_0 = \beta\,\psi_1 + \gamma\,\psi_2 + \delta\left(\frac{\partial\psi}{\partial n}\right)_0 .$$

From the boundary conditions (1.2.8), the normal derivative of the stream-function at the boundary is zero except for $y = 1$ where $\frac{\partial\psi}{\partial n} = 1$. The wall vorticity calculations simplify to

$$\left.\begin{aligned} \omega_{0,j} &= \beta\,\psi_{1,j} + \gamma\,\psi_{2,j} , \\ \omega_{n+1,j} &= \beta\,\psi_{n,j} + \gamma\,\psi_{n-1,j} + \delta \\ &\text{for } j=1,2,\ldots,m \text{ , and} \\ \omega_{i,0} &= \beta\,\psi_{i,1} + \gamma\,\psi_{i,2} , \\ \omega_{i,m+1} &= \beta\,\psi_{i,m} + \gamma\,\psi_{i,m-1} \\ &\text{for } i=1,2,\ldots,n , \end{aligned}\right\} \qquad (3.1.7)$$

for appropriate β's , γ's and δ as given in the equations (3.1.6). The formulas (3.1.7) have second order discretisation error. This error is either the same order as or higher order than the approximation of vorticity in the interior of the cavity. So approximations to the boundary values do not degrade the accuracy of the approximations in the interior.

For the special case of an even mesh in both directions, the equations (3.1.5) simplify to

$$\beta = -\frac{4}{k^2} ,$$

$$\gamma = \frac{1}{2k^2} ,$$

$$\delta = -\frac{3}{k^2} ,$$

for the walls $y = 0$ and $y = 1$ where k is the y mesh length. Similar

formulas hold for the walls $x = 0$ and $x = 1$.

Consider the convection term in divergence form

$$\nabla\cdot(\omega\underset{\sim}{u}) = \frac{\partial}{\partial x}(\omega u) + \frac{\partial}{\partial y}(\omega v) \ .$$

The velocity components are obtained by numerical differentiation of the streamfunction then the products ωu and ωv are differentiated numerically and added to obtain the final value. In the calculation of a first derivative, a three point formula is used to approximate the derivative value at the central point. The coefficients in this formula are chosen so that if the terms are expanded about the central point by a Taylor series expansion, then as many as possible lower order contributions to the error cancel. For example, the first x-derivative of the streamfunction is calculated from the formula

$$\left(\frac{\partial\psi}{\partial x}\right)_{i,j} = \frac{h_{i-1}}{h_i(h_i + h_{i-1})}\cdot\psi_{i+1,j} + \frac{(h_i - h_{i-1})}{h_i\, h_{i-1}}\cdot\psi_{i,j}$$

$$- \frac{h_i}{h_{i-1}(h_i + h_{i-1})}\cdot\psi_{i-1,j} \qquad (3.1.8)$$

The discretisation error for this approximation is $0\ (h_i\ h_{i-1})$, a first order error. Similar expressions hold for first derivatives with respect to y . Where even meshes are used, equation (3.1.8) reduces to the usual two point central difference approximation

$$\left(\frac{\partial\psi}{\partial x}\right)_{i,j} = \frac{\psi_{i+1,j} - \psi_{i-1,j}}{2h} \qquad (3.1.9)$$

which has $0(h^2)$ discretisation error. Similar expressions to either formula (3.1.8) or (3.1.9), as appropriate to the grid, are used for the numerical differentiation of the streamfunction to obtain the velocity components and of the products ωu and ωv to obtain the value of the

convection term itself.

The above discussion has covered central difference formulas for modelling the convection contribution to the fluid flow. Another class of techniques used by some authors (Bozeman and Dalton [5], Godaux [20], Torrance [42]) is that of unidirectional differencing. Bozeman and Dalton in fact compare two different schemes for differencing the non-linear term: (1) central differences using second order correct difference quotients and, (2) unidirectional differences using first order correct difference quotients which are backward with respect to the local direction of the fluid velocity. Both the divergence form and the convective form of the non-linear term are discussed but only an even mesh is used. The divergence form gives rise to the non-linear term

$$\frac{A_1(\omega u)_{i+1,j} + A_2(\omega u)_{i,j} + A_3(\omega u)_{i-1,j}}{h} + O(h)$$

$$+ \frac{A_4(\omega v)_{i,j+1} + A_5(\omega v)_{i,j} + A_6(\omega v)_{i,j-1}}{k} + O(k)$$

where

$$A_1 = +1, \quad A_2 = -1, \quad A_3 = 0 \quad \text{when} \quad u_{ij} < 0 ,$$

$$A_1 = 0, \quad A_2 = +1, \quad A_3 = -1 \quad \text{when} \quad u_{ij} \geq 0 ,$$

$$A_4 = +1, \quad A_5 = -1, \quad A_6 = 0 \quad \text{when} \quad v_{ij} < 0 ,$$

$$A_4 = 0, \quad A_5 = +1, \quad A_6 = -1 \quad \text{when} \quad v_{ij} \geq 0 .$$

A similar formula is used for the convective form of the term.

Godaux uses a similar differencing of the divergence form of the non-linear term except that the velocity values are evaluated at the half-mesh points according to some rule. Again even meshes are used. Only low Reynolds

numbers are examined but the results are inconclusive because of the coarseness of the mesh used. Torrance compares various methods of differencing the convection term including backward unidirectional differences. Torrance suggests that this method is one of a number of preferred methods because vorticity is conserved within the grid system (an even grid). The method is free from mesh size restrictions and it is recommended that the method be used when restrictions on other conservative methods cannot be satisfied. A warning is given, however, that the results must be interpreted carefully because of the truncation errors.

Chorin [9] has stated that he knows of no good reason for casting the non-linear terms in the Navier-Stokes equations into "conservation law" form. Such forms often use unidirectional differences and have a much larger truncation error than central differences. Chorin's policy is followed in this thesis.

A mention must be made of the work of Barrett [2] and Dorr [16]. Each has treated a one-dimensional analogue of the singular perturbation problem of very high Reynolds numbers. The conclusions reached suggest that for such problems central differences for the non-linear term may not be appropriate in the limit of high Reynolds numbers. Various schemes including some similar to unidirectional differencing are proposed.

3.2. THE TIME DERIVATIVE

Since the subject of this work is finite difference methods not the solution of non-linear equations as are the steady state Navier-Stokes equations, a time step method of solving the time dependent Navier-Stokes equations was felt to be more illustrative of the finite difference methods. Equations parabolic in time (the vorticity equation) can be solved by one of two major methods - explicit or implicit time stepping from some initial conditions to steady state.

In an explicit method the time derivative is replaced by a forward difference formula and the spatial portion of the equation is evaluated at the earlier time when all quantities are assumed known. An explicit calculation of the new function values can be made as some combination of the old function values at the mesh points. The numerical stability of this method usually implies some restrictions on the size of the time step and perhaps also on the size of the mesh widths.

In an implicit method the time derivative is replaced by a backward difference formula and the spatial portion of the equation is evaluated at the later time when the function values are not known. This results in a set of algebraic equations for the new function values. In simple cases the equations are sparse linear systems but in the case of the coupled equations of fluid flow, they are a set of non-linear algebraic equations. The advantages of implicit methods is that such schemes usually have unconditional numerical stability so that a large time step can be used.

Firstly we demonstrate an implicit finite difference scheme. Let a superscript n denote a function value at the n'th time level t_n.

Let double subscripts on an expression refer to the calculated value of that expression at the referenced grid point and let a subscript h be used when a value at an arbitrary mesh point is to be referenced.

For the implicit scheme, the time derivative in equation (3.1.1) is approximated at the $(n + 1)$'th time level by a backward difference formula which has first order discretisation error

$$\left(\frac{\partial \omega}{\partial t}\right)^{n+1} = \frac{\omega^{n+1} - \omega^{n}}{\Delta t_n} + O(\Delta t_n) \qquad (3.2.1)$$

where Δt_n is the time step between the n'th and $(n + 1)$'th time levels. If we substitute this approximation into equation (3.1.1) and discretise the spatial terms as explained in §3.1, we obtain

$$\omega^{n+1} + \Delta t_n \cdot (\nabla \cdot (\omega \underset{\sim}{u}) - \frac{1}{Re} \nabla^2 \omega)_h^{n+1} = \omega^{n} \ . \qquad (3.2.2)$$

The associated Poisson equation for the streamfunction is

$$\nabla^2 \psi^{n+1} = -\omega^{n+1} \qquad (3.2.3)$$

where both sides of the equation are at the same time level t_{n+1}. The velocity components which are needed in the calculation of the convection term $(\nabla \cdot (\omega \underset{\sim}{u}))_h^{n+1}$ are obtained by numerical differentiation of the streamfunction ψ^{n+1}. However the streamfunction depends on the vorticity via equation (3.2.3). Hence in equation (3.2.2), the coefficients of the vorticity values ω_{ij}^{n+1} are functions of those values. Obviously equation (3.2.2) is not a linear equation in the vorticity values ω_{ij}^{n+1}. If the scheme (3.2.2) is to be used, the set of non-linear equations, (3.2.2), (3.2.3) and the equation for the velocity components in terms of the streamfunction must be solved simultaneously.

A semi-implicit method which avoids the solution of a set of non-linear coupled equations may be obtained by evaluating the velocity components in the convection term at the n'th time level where they are known instead of the (n + 1)'th time level. This introduces the scheme

$$\omega^{n+1} + \Delta t_n(\nabla\cdot(\omega^{n+1}\underset{\sim}{u}^n) - \frac{1}{Re}\nabla^2\omega^{n+1})_h = \omega^n$$

for the vorticity equation. Splitting parts of the spatial portion of the equation onto different time levels destroys the time centering of the whole equation with a consequent increase in the discretisation error. The above equation is a large sparse linear system for the unknowns ω_{ij}^{n+1} which has to be solved at every time step.

A fully explicit scheme for the problem can be obtained by replacing the time derivative in equation (3.1.1) with a forward time difference formula with first order discretisation error

$$\left(\frac{\partial\omega}{\partial t}\right)^n = \frac{\omega^{n+1} - \omega^n}{\Delta t_n} + O(\Delta t_n) \ . \qquad (3.2.5)$$

Substituting this approximation into equation (3.1.1), we obtain the explicit scheme

$$\omega^{n+1} = \omega^n + \Delta t_n\cdot(\frac{1}{Re}\nabla^2\omega - \nabla\cdot(\omega\underset{\sim}{u}))_h^n \ . \qquad (3.2.6)$$

In the explicit scheme (3.2.6) the new vorticity values ω_{ij}^{n+1} are calculated from the vorticity and velocity values at the n'th time level t_n. The velocity values at time t_n are calculated by numerical differentiation ((3.1.8) or (3.1.9)) of the streamfunction values at time t_n. The streamfunction ψ^n is easily obtained from the known vorticity ω^n by the solution of the Poisson equation (Chapter 2)

$$\nabla^2\psi^n = -\omega^n \ . \qquad (3.2.7)$$

Once the streamfunction is known, the values of the vorticity along the walls may be calculated using the formulae (3.1.7). The wall vorticity values are needed in equation (3.2.6) in the calculation of the new vorticity values along mesh lines that are one mesh point inside the boundary.

Once the velocity values $\underset{\sim}{u}^n_{ij}$ are known, all the quantities on the right hand side of equation (3.2.6) are known and the equation may be used to calculate the new voriticity at the interior points are needed in the solution of equation (3.2.7) for the streamfunction then the iteration is continued as above. The time step Δt_n for each iteration is chosen by methods mentioned in §3.3.

3.3. CONSISTENCY, STABILITY AND CONVERGENCE

For an explicit method of solving an equation parabolic in time such as the vorticity equation, numerical stability requirements impose restrictions on the size of the time step and on the x and y mesh spacings. For simplicity in the following discussion, suppose that equi-spaced meshes are used in both directions and that $\Delta x = \Delta y = h$. As will be seen from the later presentation of the method of choosing a graded mesh for the cavity problem, a stability analysis for the graded mesh case will only be a small extension of the principle of this case.

From equation (3.2.6), the unknown vorticity ω_{ij}^{n+1} may be written as an explicit linear combination of the computed values of vorticity at the time level t_n :

$$\omega_{ij}^{n+1} = a_1 \omega_{i+1,j}^{n} + a_2 \omega_{i-1,j}^{n} + a_3 \omega_{ij}^{n} + a_4 \omega_{i,j+1}^{n} + a_5 \omega_{i,j-1}^{n} \qquad (3.3.1)$$

where the a_k denote coefficients that vary in time but are constant over a time step. The coefficients a_k implicitly refer to only one mesh point and are not constant over the mesh.

In the notation of Richtmeyer and Morton [36] (p.42ff), if we let $\underset{\sim}{\omega}$ be the vector the components of which are all the vorticity values ω_{ij} at mesh points in the interior of the cavity, then equation (3.3.1) describes one component of the matrix-vector equation

$$\underset{\sim}{\omega}^{n+1} = C_n(\Delta t_n)\, \underset{\sim}{\omega}^{n} \, . \qquad (3.3.2)$$

Along any particular row of the matrix $C_n(\Delta t_n)$, the appropriate set of coefficients a_k are the only non-zero elements.

If we let $h = \lambda(\Delta t)$ for some function λ such that $h \to 0$ as

$\Delta t \to 0$, then the family of operators $C_n(\Delta t_n)$ provides a *consistent* approximation to the initial value problem if for every $\omega(t)$ in some class of genuine solutions whose initial elements are dense in some appropriate solution space,

$$\left|\left|\left\{\frac{C(\Delta t) - I}{\Delta t} - A\right\} \omega(t)\right|\right| \to 0 \text{ as } \Delta t \to 0, \qquad (3.3.3)$$

$$0 \leq t \leq T.$$

Here I stands for the identity operator and the operator A represents the operator formed from the spatial portion of the original equation (3.1.1). Since $C(\Delta t) = I + \Delta t S_h$ where S_h is a finite difference operator that approximates the continuous operator A, the norm in (3.3.3) becomes $|| \{S_h - A\} \omega(t)||$. From the approximation formulas (3.1.4) and (3.1.9) the difference between $S_h \omega$ and $A \omega$ is $O(h^2)$. Then from the restriction $h = \lambda (\Delta t)$, the difference $(S_h - A)\omega$ tends to zero as Δt tends to zero, but just convergence to zero is not enough. Consider the n-vector

$$\underset{\sim}{v}_n^T = (\tfrac{1}{n}, \tfrac{1}{n}, \ldots, \tfrac{1}{n})$$

and the l_1, l_2 and l_∞ norms of that vector. We find that

$$||\underset{\sim}{v}_n||_1 = \sum_{i=1}^{n} |(\underset{\sim}{v}_n)_i|$$

$$= \sum_{i=1}^{n} \frac{1}{n}$$

$$= 1$$

which is constant, independent of the size of the vector ,

$$||\underset{\sim}{v}_n||_2 = \left\{ \sum_{i=1}^{n} (\underset{\sim}{v}_n)_i^2 \right\}^{\frac{1}{2}}$$

$$= \frac{1}{\sqrt{n}} \to 0 \text{ as } n \to \infty$$

and

$$||\underset{\sim}{v}_n||_\infty = \max_{1 \leq i \leq n} |(\underset{\sim}{v}_n)_i|$$

$$= \frac{1}{n} \to 0 \text{ as } n \to \infty$$

Obviously not just the form of the components but also the norm used can influence the convergence of a norm of the vector. Condition (3.3.3) must not only imply conditions on the functional form of the dependence $h = \lambda(\Delta t)$ but also on the norms used. For example in the case of an even grid and using the maximum (l_∞) norm, the relationship $h^2 = \lambda \Delta t$ for some constant λ would be an appropriate relationship. The choice of the constant λ is explained below. Note that the relationship between time and spatial mesh sizes can also be written as an inequality relationship for if

$$\Delta t = c\, h^2$$

satisfies (3.3.3) for some constant c then so also does any time step satisfying

$$\Delta t \leq c\, h^2$$

The Lax equivalence theorem (see Richtmeyer and Morton [36] for proof) states: *"Given a properly posed initial value problem and a finite difference approximation to it that satisfies the consistency condition, stability is a necessary and sufficient condition for convergence."* Since the scheme (3.3.1) has been shown to satisfy the consistency condition, only stability conditions must be examined in order that conditions may be found under which the scheme (3.3.1) converges.

From Richtmeyer and Morton [36], the stability of a difference scheme

(3.3.2) means that no component of the solution becomes unbounded, independent of the initial values. It is required that the product of the operators

$$C_n(\Delta t_n)\ C_{n-1}(\Delta t_{n-1})\ \ldots,\ C_0(\Delta t_0)$$

be uniformly bounded where $0 < \Delta t_i < \tau$ for some τ and for $0 \leq \sum_i \Delta t_i \leq T$ for some maximum time T. The essence of the definition is that the product of operators is still bounded as $\Delta t_i \to 0$ and $t = \sum \Delta t_i$ fixed. This condition will be satisfied if the norm of each operator in turn is bounded by unity in the limit as $\Delta t \to 0$, i.e. if

$$\|C(\Delta t)\| \leq 1 + O(\Delta t)\ .$$

If we choose the row sum norm for the matrix operator then from above, we require that the norm of the matrix of the coefficients a_k is bounded by unity for all times under consideration. We obtain the condition that for stability of the scheme (3.3.1), it must satisfy the condition

$$\max_{i,j} \left\{ \sum_{k=1}^{5} |a_k| \right\} \leq 1 \qquad (3.3.4)$$

where $\max_{i,j}$ denotes the maximum value of the row sum norm over all values of i,j in the grid system, that is, for all rows of the matrix $C(\Delta t)$.

Consider the scheme (3.2.6) using a grid equally spaced in each direction with step lengths of h and k in the x and y directions respectively. The divergence form of the vorticity equation is used. Supposing that enough conditions are satisfied so all the coefficients a_k are non-negative, then we have that if

$$a_k \geq 0 \quad , \quad k = 1,2, \ldots 5$$

for all mesh points, then

$$\sum_{k=1}^{5} |a_k| = \sum_{k=1}^{5} a_k$$

$$= 1 - \Delta t \cdot \left\{ \frac{u_{i+1,j} - u_{i-1,j}}{2h} + \frac{v_{i,j+1} - v_{i,j-1}}{2k} \right\}$$

The term in braces in the last equation is a finite difference expression for $\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}$ at the (i,j) mesh point. By the continuity equation (1.2.16) for incompressible fluid flow, the continuous equivalent of this term vanishes. The finite difference expression should be at most an $O(h^2+k^2)$ term over the grid because the velocity components are only correct to that order of approximation. This would mean that

$$\sum_{k=1}^{5} |a_k| = 1 + \Delta t \cdot O(h^2+k^2) .$$

But a sharper result may be had. We find that the grid system identically conserves fluid, that is, the finite difference equivalent of the continuity equation is identically zero over all interior mesh points.

$$\frac{u_{i+1,j} - u_{i-1,j}}{2h} + \frac{v_{i,j+1} - v_{i,j-1}}{2k}$$

$$= \frac{1}{2h} \left\{ \frac{\psi_{i+1,j+1} - \psi_{i+1,j-1}}{2k} - \frac{\psi_{i-1,j+1} - \psi_{i-1,j-1}}{2k} \right\}$$

$$- \frac{1}{2k} \left\{ \frac{\psi_{i+1,j+1} - \psi_{i-1,j+1}}{2h} - \frac{\psi_{i+1,j-1} - \psi_{i-1,j-1}}{2h} \right\}$$

$$= 0$$

Hence for all mesh points in the cavity, the last equation reduces to

$$\sum_{k=1}^{5} |a_k| = 1$$

This result is also obtained if the convective form $(\underset{\sim}{u} \cdot \nabla \omega)$ instead of

the divergence form $(\nabla\cdot(\omega\underset{\sim}{u}))$ of the convection term is used in equation (3.2.6). The non-negativity conditions on the coefficients a_k that enable (3.3.4) to be satisfied are

$$h < \frac{2}{Re \cdot \max\limits_{i,j} \{|u_{ij}|,|v_{ij}|\}} \qquad (3.3.5)$$

and

$$\Delta t < \frac{Re\ h^2}{4} . \qquad (3.3.6)$$

These conditions in practice may be more restrictive than necessary but indicate the general method. Special cases for individual schemes will be given where such schemes are discussed.

3.4. SOLUTION OF THE CAVITY PROBLEM USING EQUISPACED MESHES

In this section the cavity problem is solved by the explicit time step method (3.2.6) using both the divergence and the convective forms of the vorticity transport equation. The Poisson equation for the streamfunction is solved by a direct method (Chapter 2). Equispaced meshes are used in both x and y directions and the mesh widths Δx and Δy are both equal to h. Only a square cavity will be considered. Since this is the case considered in §3.3. as an example, the stability conditions are given as (3.3.5) and (3.3.6).

The algorithm used for the solution is as follows:

1. At time t_n, the vorticity ω^n is known at mesh points inside the cavity. Since only these values of vorticity are used by the Poisson equation (3.2.7), a direct method from Chapter 2 may be applied to obtain the solution ψ^n. For initial conditions, $\omega^\circ = \psi^\circ = 0$ are used.

2. From the streamfunction ψ^n, the velocity component values u^n and v^n are calculated at every mesh point by the central difference formula (3.1.9).

3. A check is made that the stability conditions are satisfied. If they are not, the mesh is refined appropriately and values of vorticity are interpolated at the new mesh points, then return to stage 2.

4. Using the formulas (3.1.7), the values of the vorticity at mesh points along the walls are calculated. The vorticity ω^n at all points of the grid is known except for the corner points. At the two lower corners the vorticity is zero but at the upper corners there are singularities in the vorticity caused by the discontinuity of the sliding wall.

These singularities are of no explicit concern to the above numerical scheme since the corner points do not enter into any of the calculations. However, near the corner points and in general near the boundary, the higher derivatives of the vorticity become large. So that these locally large changes can be well modelled by the finite difference scheme, the mesh in these regions must be correspondingly finer. In this way the truncation error at all points of the grid is kept relatively constant at an acceptable level. The problem of having to use increasingly finer meshes to well model the faster changing flow in the boundary regions for higher Reynolds numbers and the resultant massive increases in computation necessary for a solution, if obtainable at all, is central to the motivation for using graded meshes.

5. The new vorticity values are calculated from equation (3.2.6) or its equivalent for the convective form of the vorticity equation.

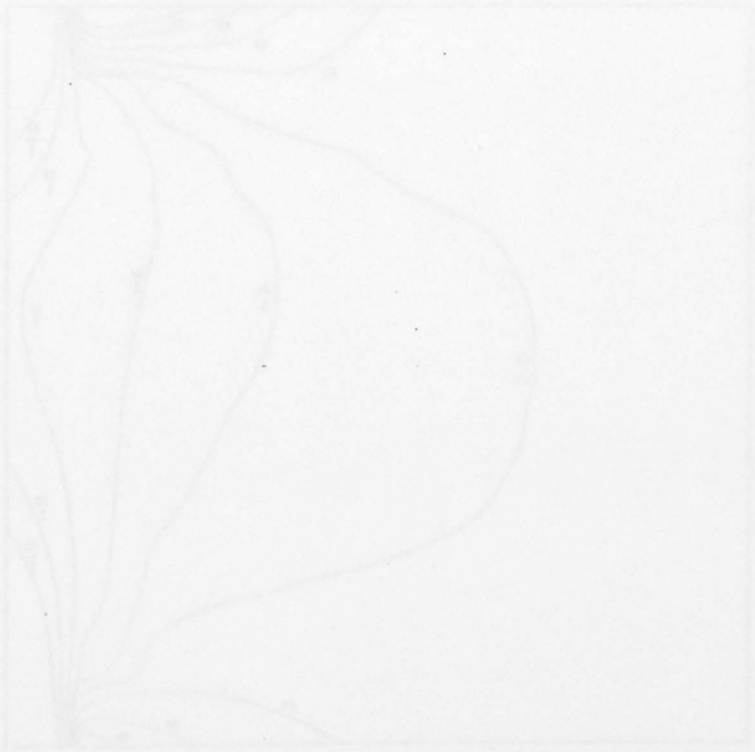
6. The convergence criterion is checked to see if steady state has been reached, if not then return to stage 1. The convergence criterion used to determine if steady state has been reached is

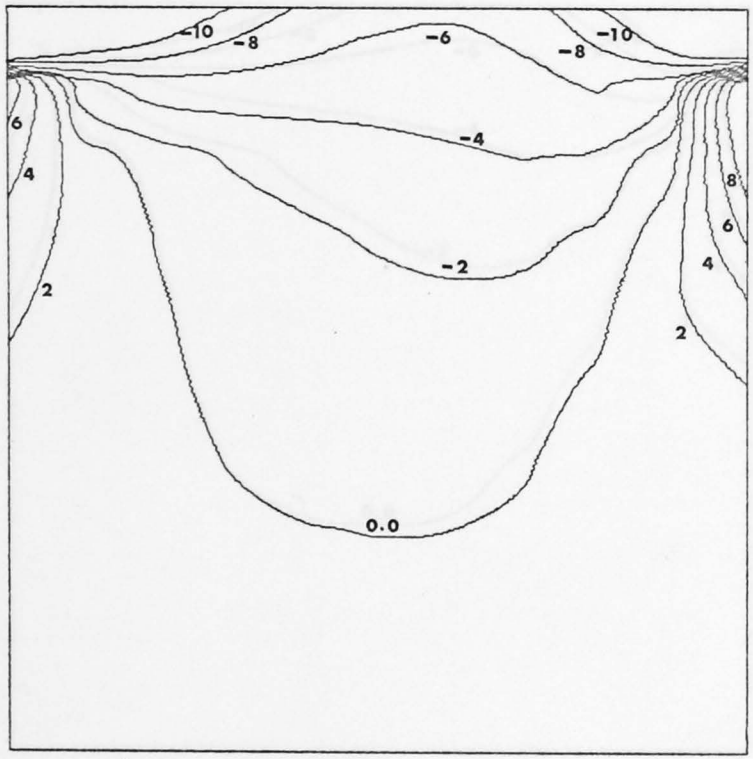
$$\max_{ij} \left| \frac{\omega_{ij}^{n+1} - \omega_{ij}^{n}}{\omega_{ij}^{n}} \right| < \varepsilon$$

for some small positive constant ε. If for example $\varepsilon = 10^{-3}$ is chosen, this implies that the vorticity is constant to within three significant decimal digits everywhere in the cavity.

The results of this scheme are plotted in figures 3.4.1 (a) - (d), 3.4.2 (a) - (d), etc. The (a) and (b) sub-figures refer to the results using the divergence form of the vorticity equations, the (c) and (d) sub-figures refer to the results using the convective form of the vorticity equation. The (a) and (c) sub-figures are vorticity fields and the (b) and (d) sub-figures are streamfunction fields.

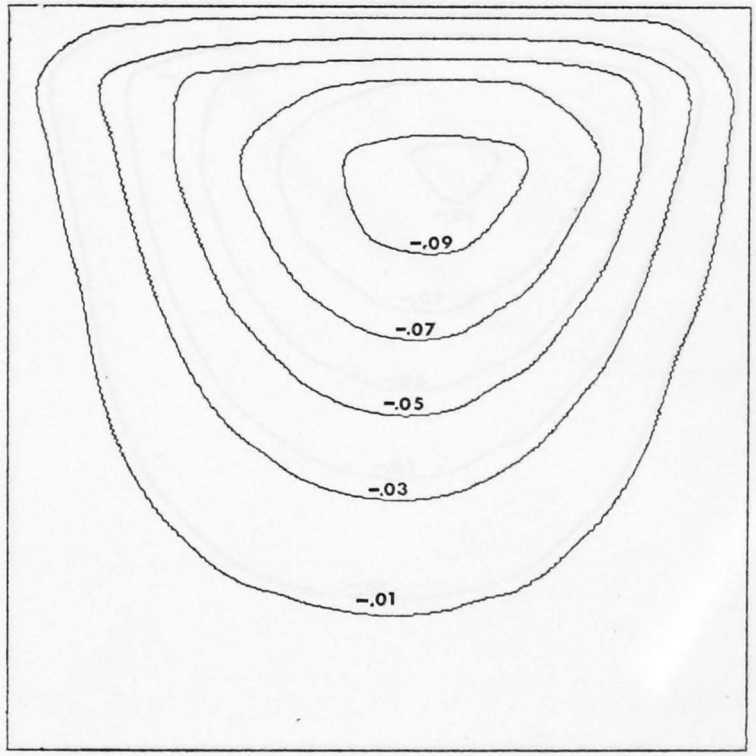
Discussion of these results and a comparison with those obtained using graded meshes can be found in Chapter 6. The equivalent results for graded meshes can be found in Chapter 5.





Vorticity

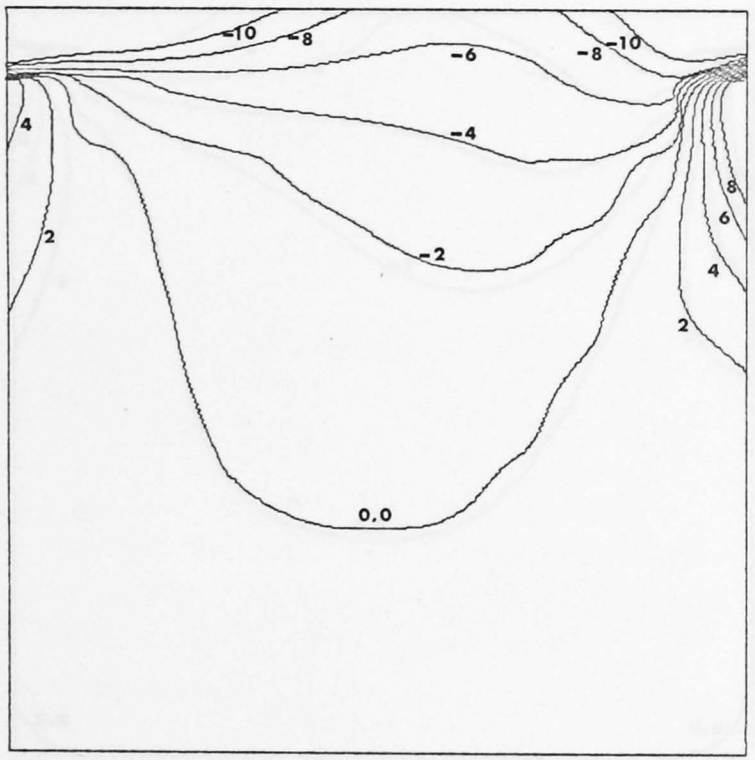
Figure 3.4.1a



Streamfunction

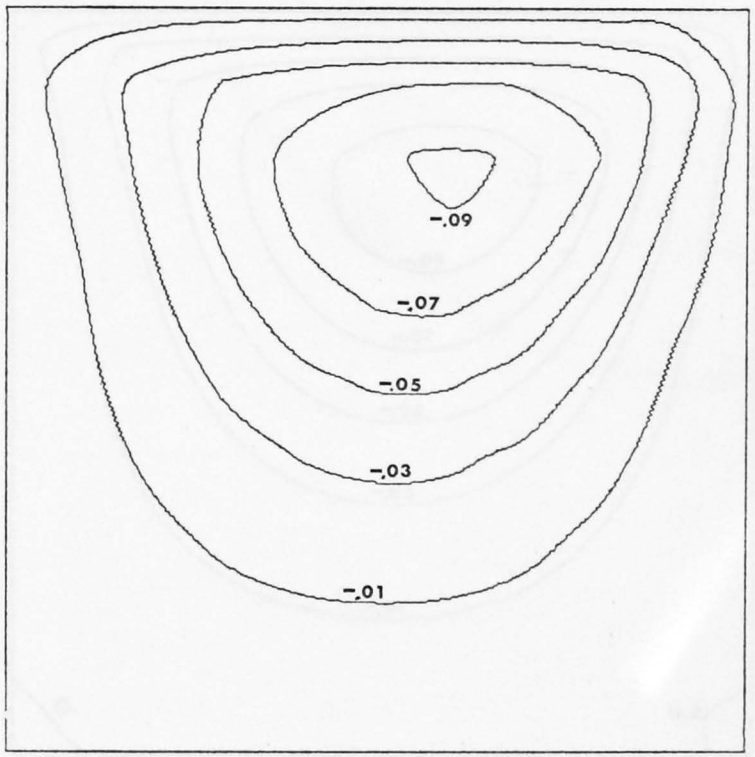
Figure 3.4.1b

Example RD11



Vorticity

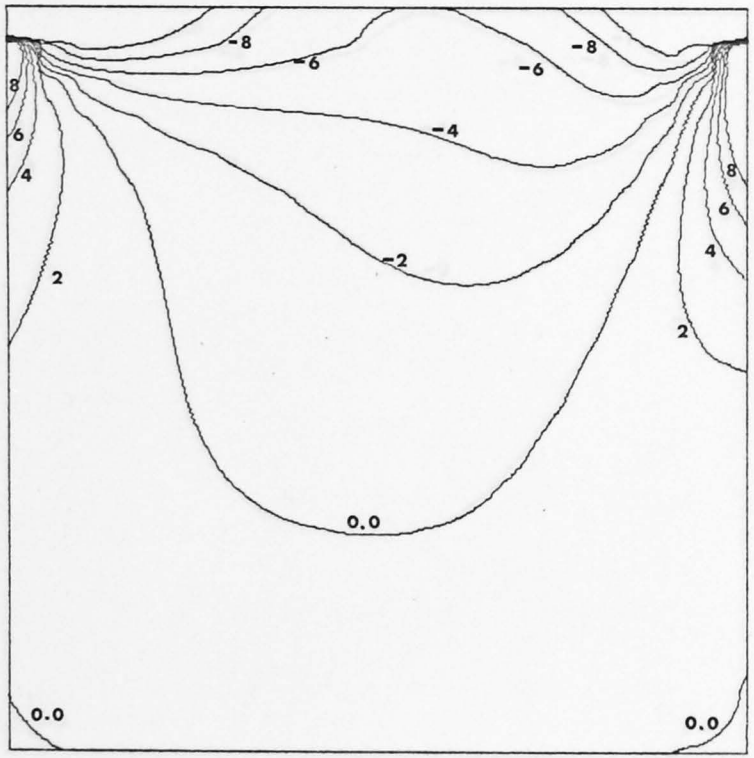
Figure 3.4.1c



Streamfunction

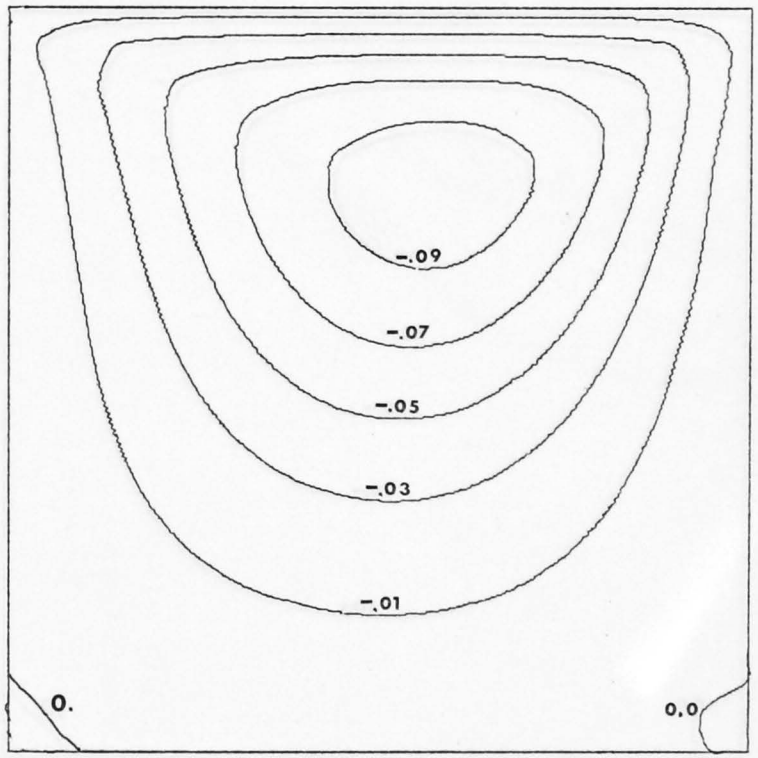
Figure 3.4.1d

Example RC11



Vorticity

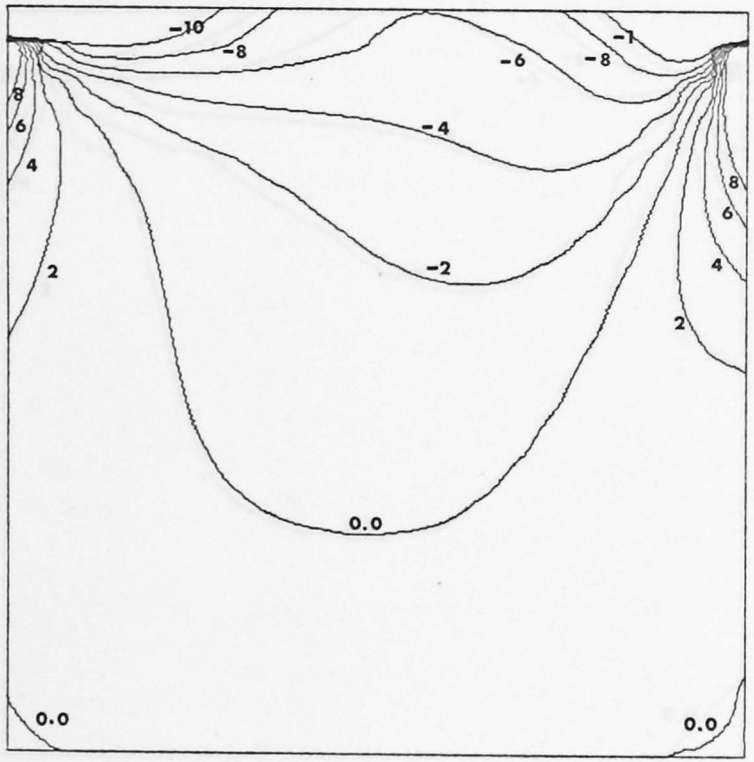
Figure 3.4.2a



Streamfunction

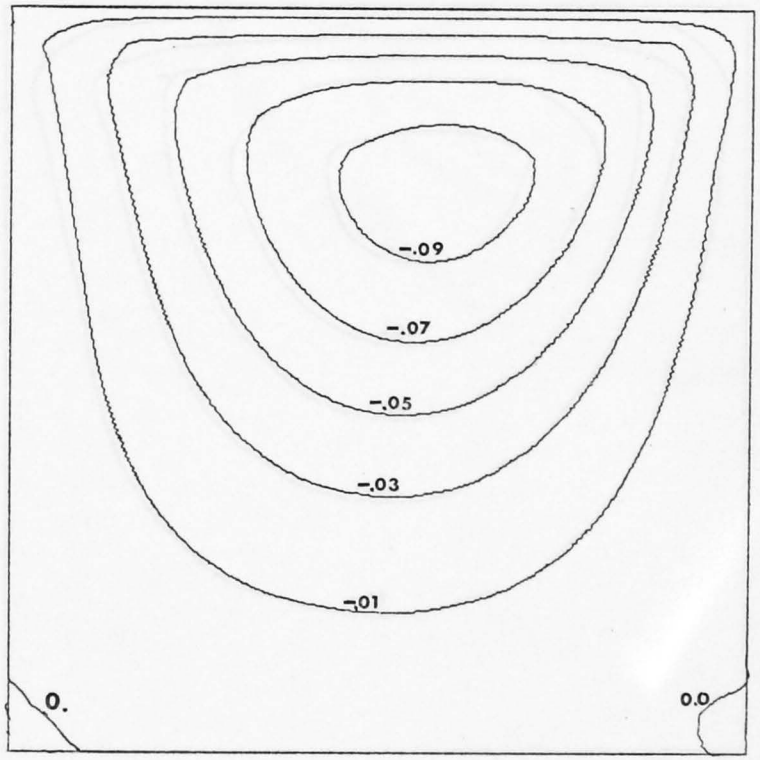
Figure 3.4.2b

Example RD21



Vorticity

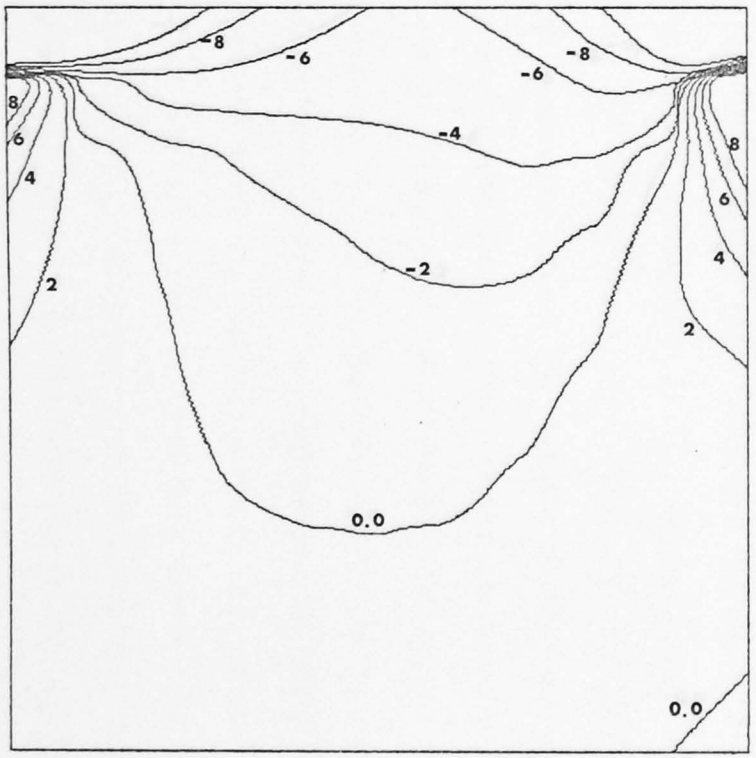
Figure 3.4.2c



Streamfunction

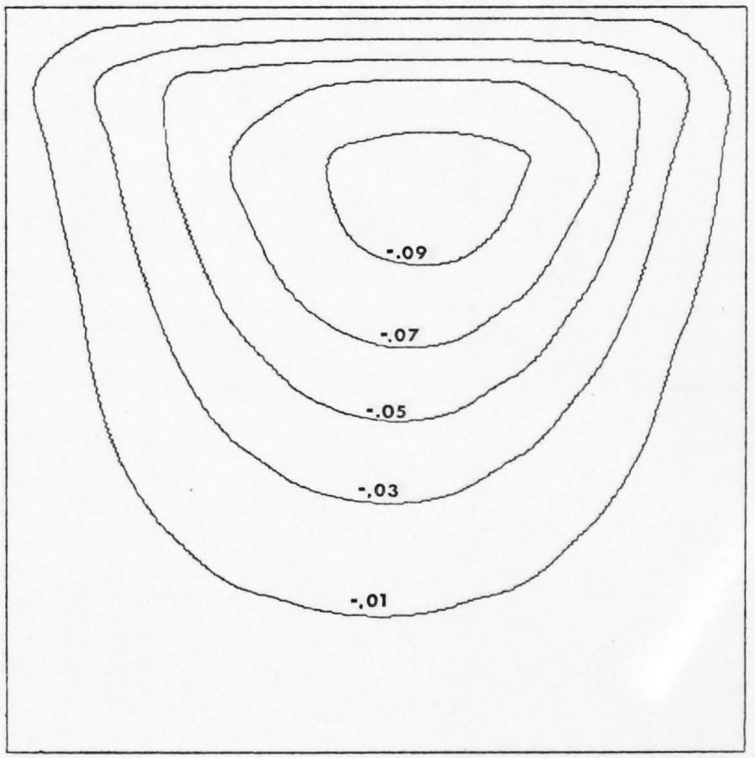
Figure 3.4.2d

Example RC21



Vorticity

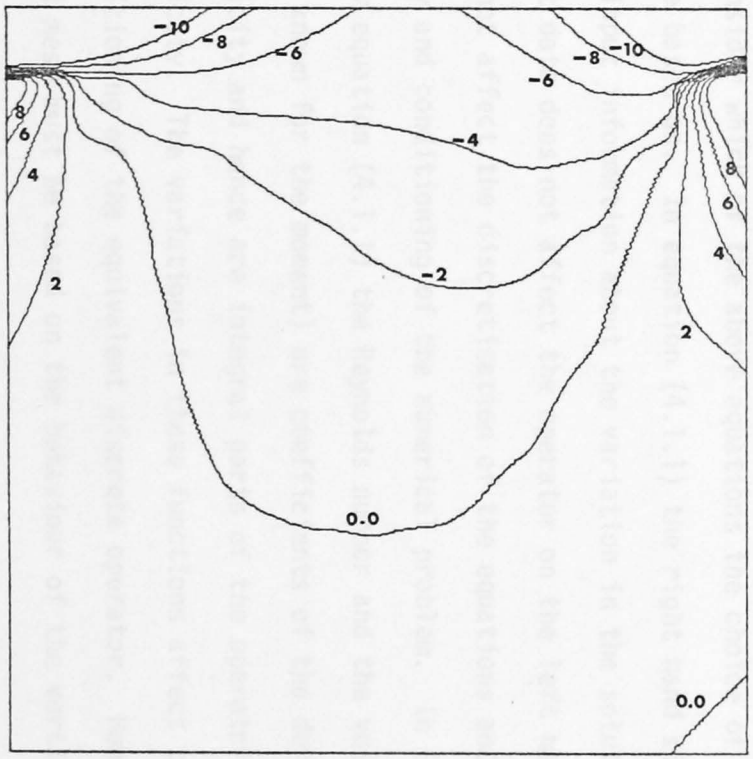
Figure 3.4.3a



Streamfunction

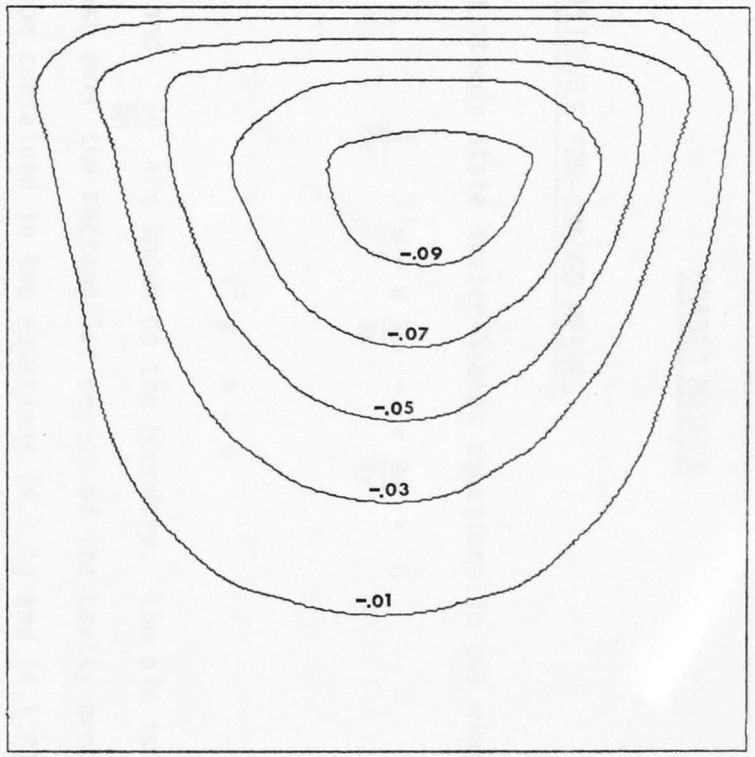
Figure 3.4.3b

Example RDE



Vorticity

Figure 3.4.3c



Streamfunction

Figure 3.4.3d

Example RCE

CHAPTER 4

GRADED MESHES

§4.1 RATIONALE FOR GRADED MESHES

The steady state Navier-Stakes equations in two dimensions are

$$\frac{1}{Re} \nabla^2\omega - u \frac{\partial \omega}{\partial x} - v \frac{\partial \omega}{\partial y} = 0 \qquad (4.1.1)$$

and

$$\nabla^2 \psi = - \omega \qquad (4.1.2)$$

where ψ and $\frac{\partial \psi}{\partial n}$ are known on the boundary. The aim is to choose a graded mesh over the rectangular region of the cavity problem using the information contained in the equations (4.1.1) and (4.1.2). The criteria for the choice of graded mesh, as yet undetermined, must imply numerical stability and good conditioning for the resulting discrete problem.

Consider which of the above equations the choice of graded mesh should be based on. In equation (4.1.1) the right hand side contains all the input information about the variation in the solution ψ. Because the input data does not affect the operator on the left hand side, that data cannot affect the discretisation of the equations and the resultant stability and conditioning of the numerical problem. In the steady state vorticity equation (4.1.1) the Reynolds number and the velocity components (assumed known for the moment) are coefficients of the derivatives of the vorticity and hence are integral parts of the operator that acts on the vorticity. The variations in these functions affect the stability and conditioning of the equivalent discrete operator. Hence the choice of graded mesh must be based on the behaviour of the vorticity equation (4.1.1).

To enable attention to be concentrated on the method of choosing a graded mesh, we restrict the motivating problems in the chapter to one dimension.

A one dimensional analogue of equation (4.1.1) is

$$\varepsilon \frac{d^2y}{dx^2} + f(x) \frac{dy}{dx} = 0 \qquad (4.1.3)$$

for $x \in (0,1)$ where ε is a small positive parameter and the function $f(x)$ and the boundary values $y(0)$ and $y(1)$ were known. This type of analogue of the Navier-Stokes equation has been studied by Barrett [2] and by Dorr [16] in connection with singular perturbation problems. In the study of such problems, interest centres on the solution of (4.1.3) with or without the extra term $g(x)\, y(x)$ as the parameter $\varepsilon \to 0$. In such situations the solution consists of one or more boundary layer type regions of small width and rapid change where the equation is properly second order and in the remainder of the interval, the equation is effectively of lower order.

Pearson [34] has studied numerous examples of such problems and has used finite difference meshes over graded meshes in their solution. Pearson's method is an iterative one and also uses the principle of continuation with respect to the parameter ε. The problem is solved for a relatively large value of ε, then that solution being taken as input to another problem with some smaller ε until the desired value of ε is reached. For a given ε value, the equation is discretised over either an initial even mesh or a previous graded mesh and solved. More points are added where the variation in solution values between adjacent mesh points exceeds some predetermined level. Meshes of up to 25000 points had to be used to solve some equations.

This chapter is concerned with problems of the type

$$\varepsilon \frac{d^2y}{dx^2} + f(x) \frac{dy}{dx} + g(x)\, y = 0$$

with coefficients ε, $f(x)$ and $g(x)$ such that though there may be some concentration of the gross behaviour of the solution $y(x)$ into one or more small subregions, the equation itself is properly of second order throughout the interval. This requires basically that ε is not small where small in the situation can mean only 10^{-2}.

This study is aimed at the two-point boundary value problem

$$\frac{d^2y}{dx^2} + a(x) \frac{dy}{dx} + b(x)\, y = 0 \qquad (4.1.4)$$

with $y(0) = y(1) = 1$ where the parameter ε has been set to unity. The second order problem (4.1.4) is also studied in its equivalent first order formulation.

Let

$$\underset{\sim}{w}^T(x) = \left(y(x) , \frac{dy}{dx}(x) \right) ,$$

then equation (4.1.4) can be written as

$$\frac{d\underset{\sim}{w}(x)}{dx} = A(x)\ \underset{\sim}{w}(x) \qquad (4.1.5)$$

where the coefficient matrix $A(x)$ is

$$A(x) = \begin{bmatrix} 0 & 1 \\ -b(x) & -a(x) \end{bmatrix} \qquad (4.1.6)$$

with the two-point boundary condition

$$B_0 \underset{\sim}{w}(0) + B_1 \underset{\sim}{w}(1) = \underset{\sim}{c}$$

where

$$B_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad B_1 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

and

$$\underset{\sim}{c}^T = (1,1)$$

We wish to examine the use of graded meshes in the finite difference solutions of equations of type (4.1.4) and (4.1.5). A scheme is desired for systematically choosing the points of the graded mesh to have 'optimal' properties (in some sense to be defined) by taking into account the natural structure and the numerical formulation of the problem.

The usual finite difference approximations to the previous equations over a graded mesh are only of first order accuracy. Such low order of approximation must be reflected in the accuracy of the calculated solution. Thus it is of paramount importance to try to choose the mesh points so that the final numerical scheme has optimal numerical performance. If satisfactory accuracy has not been attained, then a refinement of the mesh may either destroy the optimal properties of the numerical scheme or cause the information gained at the previous stage to be rendered useless. For example, extrapolation of the solutions is not possible.

If a systematic scheme is to be designed then its first aim is to pick a graded mesh that automatically endows the numerical scheme with optimal performance. A systematic method of choosing grid points must obviously allow easy mesh refinement. But the scheme's other major aim, in fact, in some ways more general and certainly more important aim is to choose a mesh for the numerical scheme that allows the use of extrapolation

techniques. In this way information gained at one stage of the solution is not lost after mesh refinement and in fact can contribute to a dramatic improvement in the accuracy of the solution. Even if the optimal numerical scheme is not known, as shown later the mesh can still be designed to allow h^2 and h^4 extrapolation to be applied to the solutions with a consequent increase in the accuracy of the final solutions.

Worthwhile contributions are considered to be made by the following attempts to create a systematic method of choosing a graded mesh in a manner that uses the natural structure of the problem specifically to choose a graded mesh so that extrapolation procedures can be used to improve the accuracy.

An example is given below to emphasise that if a graded mesh is to be used in the solution of a two-point boundary value problem, then that mesh must be chosen carefully. A mesh of $(n+1)$ points is chosen, each internal point being chosen randomly from a uniform distribution on the interval $(0,1)$ and then ordered to obtain the mesh

$$0 = x_0 < x_1 < x_2 < \ldots < x_n < x_{n+1} = 1 .$$

As far as is known to the author, the examination of such a mesh in the solution of equations using finite differences has not been considered previously.

Example 4.1.1

Consider the two-point boundary value problem

$$\frac{d^2y}{dx^2} - \frac{1}{(x+\varepsilon)} \cdot \frac{dy}{dx} - \frac{3}{(x+\varepsilon)^2} \cdot y = 0 \qquad (4.1.7)$$

for $x \in (0,1)$ subject to $y(0) = y(1) = 1$ where ε is a small positive parameter. Equation (4.1.7) has a solution of the form

$$y(x) = \frac{A}{(x+\varepsilon)} + B(x+\varepsilon)^3$$

for appropriate constants A and B.

For the even mesh solution central difference formulas (2.2.4) and (3.1.9) were used to approximate the derivative terms while the equivalent central three-point approximations (2.2.9) and (3.1.8) respectively were used for the graded mesh case. The resulting sets of tridiagonal systems of linear equations were solved by Gaussian elimination and compared with the analytic solution at the mesh points. Table 4.1.1 gives the solution and errors for an even mesh and for one example of a graded mesh. Table 4.1.2 contains an examination of the errors in the solution of the equation (4.1.7) for various examples of the random graded mesh.

Since extrapolation is so very important to the following techniques, §4.2 contains a short resumé of the appropriate theory applicable to this work. Some work done by Osborne [31] on shooting methods is summarised in §4.3 as it is the motivation for the scheme proposed for choosing the graded mesh in the solution of the first order system (4.1.5) that is discussed in §4.4. Two different second order examples are studied in §4.5 and §4.6. There is a summary and discussion of the applicability of this work in §4.7.

TABLE 4.1.1

Example 4.1.1 with parameter $\varepsilon = 0.1$ for $n = 11$ mesh points.

Even Mesh			Random Graded Mesh		
mesh points	analytic soln.	error	mesh points	analytic soln.	error
0.1	.5051	.5873E-1	.8333E-2	.9233	-.3514
0.2	.3515	.4926E-1	.1926	.3587	.4564
0.3	.2935	.3931E-1	.2080	.3444	.3218
0.4	.2852	.3157E-1	.3327	.2863	.2489
0.5	.3141	.2537E-1	.4366	.2917	.2141
0.6	.3770	.2003E-1	.5474	.3397	.2028
0.7	.4746	.1511E-1	.7617	.5530	.2939
0.8	.6090	.1027E-1	.7840	.5849	.2221
0.9	.7830	.5293E-2	.9403	.8651	.1037
Sum of squares of errors		.9824E-2			.7307

TABLE 4.1.2

Comparison of results for example 4.1.1 for $\varepsilon = 0.1$ for an even mesh and various random graded meshes.

Even Mesh

Sum of squares of errors = .9824E-2

Random Graded Meshes

Mesh No.	S.SQ. errors	Mesh No.	S.SQ. errors
1	.7307	6	.3858
2	.2922	7	.1598
3	.5745	8	.3640E+2
4	.1417	9	.6695E+3
5	.9367E-1	10	.1414E+1

4.2 DISCRETISATION AND EXTRAPOLATION

Consider the continuous linear problem

$$Lz = 0 \qquad (4.2.1)$$

where we assume that (4.2.1) has a unique solution z. Consider the general approach of Stetter [39] to the discretisation of (4.2.1) which is summarised in figure 4.2.1

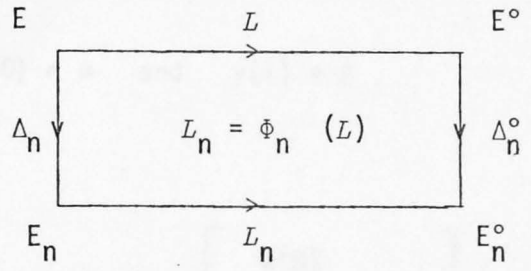


Figure 4.2.1

The space E is the space of allowable functions that constitute the domain of the operator L while the space E° is the null space of the operator L. The linear discretisation functions Δ_n and Δ_n° map the spaces E and E° to finite dimensional spaces E_n and E_n°, the elements of which approximate the elements of the spaces E and E° in the sense

$$\| \Delta_n u \|_{E_n} \rightarrow \| u \|_E$$

and

$$\| \Delta_n^\circ v \|_{E_n^\circ} \rightarrow \| v \|_{E^\circ}$$

for some appropriate norm as $n \rightarrow \infty$ through permitted values $n \in N$. We have the condition

$$\dim E_n = \dim E_n^\circ .$$

The continuous operator L is transformed by the discretisation function Φ_n into the discrete operator L_n.

As an example, consider the linear two-point boundary value problem

$$\frac{d^2y}{dx^2}(x) + f(x)\,y(x) = 0$$

for $x \in [0, 1]$ with the boundary conditions

$$y(0) = \alpha \quad \text{and} \quad y(1) = \beta .$$

We have

$$L : u \rightarrow \begin{bmatrix} u(0) \\ u(1) \\ \dfrac{d^2u}{dx^2} + f\,u \end{bmatrix}$$

$$\Delta_n : u \rightarrow \sum_{i=1}^{n} u(x_i)\; H_{io}$$

where the $\{x_i\}_{i=1}^{n}$ are a set of mesh points and the H_{io} are some interpolatory functions,

$$\Delta_n^{\circ} : \begin{bmatrix} \alpha \\ \beta \\ v \end{bmatrix} \rightarrow \begin{bmatrix} \alpha \\ \beta \\ v\,(j/n)\ ,\ j=2,3,\ \ldots,\ n-1 \end{bmatrix} ,$$

$$L_n : \sum_{i=1}^{n} u_i\, H_{io} \rightarrow \begin{bmatrix} u_1 \\ u_n \\ I_j\ ,\ j=2,3,\ \ldots,\ n-1 \end{bmatrix}$$

Corresponding to the continuous problem $Lz = 0$, there is the discretised problem

$$L_n \zeta_n = 0 \tag{4.2.2}$$

We assume that the problem (4.2.1) has a unique solution. The basic definitions from Stetter [39] are summed up below.

(i) Consistency: A discretisation is consistent at y if

$$\lim_{n\to\infty} \| L_n \Delta_n y - \Delta_n^\circ Ly \|_{E_n^\circ} = 0 \quad (n \in N) \tag{4.2.3}$$

(ii) Convergence: A discretisation method is convergent if

$$\lim_{n\to\infty} \| \Delta_n z - \zeta_n \|_{E_n} = 0 \quad (n \in N) \tag{4.2.4}$$

Though it is not used in the following, the definition of stability is also given.

(iii) Stability: Let $\eta = \{\eta_n\}$, $\eta_n \in E_n$, $n \in N$.

A discretisation is stable at η if $\exists$ S, r such that

$$\| \eta_n^{(1)} - \eta_n^{(2)} \|_{E_n} \leq S \, \| L_n \eta_n^{(1)} - L_n \eta_n^{(2)} \|_{E_n^\circ}$$

$\forall\, \eta_n^{(i)}$, i=1, 2 s.t.

$$\| L_n \eta_n^{(i)} - L_n \eta_n \|_{E_n^\circ} \leq r$$

uniformly $\forall\, n \in N$.

The constants S and r are called the stability bound and the stability threshold respectively. Note that the above definitions are also satisfactory for non-linear problems.

The basic results that follow from the above presentation are:

if

(a) the discretisation is stable at $\{\Delta_n z\}$ and consistent at z, and

(b) L_n satisfies a suitable continuity condition

then

(i) the discretised problem has a unique solution for $n \in N$ and large enough, and

(ii) the discretisation method is convergent.

Compare this with the usual presentation of the result (see Lax Equivalence Theorem, §3.3)

'Consistency + Stability = Convergence' .

Both consistency and stability depend on the choice of E_n° while convergence is defined in terms of the norm on E_n, thus any 'equation' of the above type must imply assumptions about the choice of E_n° .

In the definition of convergence, suppose that the solution ζ_n of the discrete problem possesses an asymptotic expansion in the parameter $h = 1/n$ of the form

$$\zeta_n = \Delta_n \left(z + \sum_{i=k}^{p-1} h^i w_i + O(h^p)\right) \qquad (4.2.4)$$

Such an expansion is valuable not only for estimating the accuracy of the solution but also for refining the solutions. To construct such an expansion (Fox [18], Stetter [39]), we seek a mapping $I : E \to E^\circ$ such that

$$\Delta_n^\circ \, Iy = L_n \, \Delta_n \, y$$

Example:

$$Iy = \begin{bmatrix} y(0) \\ y(1) \\ \dfrac{y(x+h) - 2\,y(x) + y(x-h)}{h^2} + f(x)\,y(x) \end{bmatrix} \qquad (4.2.5)$$

With such a mapping I, we can determine an expansion of the form

$$I(h) = L + \sum_{i=k}^{p-1} h^i Y_i + O(h^p)$$

where the Y_i are independent of h. That the h^0 term is the operator L is a consequence of consistency but it also verifies it.

Example 4.2.1

(a) $$I(h) = \left\{ \frac{y(x+h) - h\,y(x) + y(x-h)}{h^2} + f(x)\,y(x) \right\} \Bigg|_{h=0}$$

$$= \left\{ \frac{d^2y}{dx^2} + O(h^2) + f(x)\,y(x) \right\} \Bigg|_{h=0}$$

$$= \frac{d^2y}{dx^2} + f(x)\,y(x). \quad \text{(Consistency)}$$

(b) $$I(h) = I(-h)$$

$$\Rightarrow I(h) = L + \sum_{i=k}^{p-1} h^{2i} Y_i + O(h^{2p})$$

To verify the expansion for ζ_n we have, assuming stability,

$$\| \zeta_n - \Delta_n z \|_{E_n} \leq S \, \| L_n \Delta_n z \|_{E_n^\circ}$$

$$\leq S \| \Delta_n^\circ I z \|_{E_n^\circ}$$

$$= O(h^k)$$

Choose w_k so that

$$L(z + h^k w_k) + h^k Y_k (z + h^k W_k) = O(h^{k+1}) .$$

This will be so if

$$L'(z) W_k + Y_k(z) = 0$$

where $L'(z)$ is the Frechet derivative of L at z and is a linear operator. Thus we have

$$\| \zeta_n - \Delta_n (z + h^k w_k) \|_{E_n} \leq S \| \Delta_n^\circ I(z + h^k w_k) \| E_n^\circ$$

$$= O(h^{k+1}) .$$

The definition is completed recursively.

Consider the example 4.2.1. There we have a finite difference scheme over an even mesh for the simple equation

$$Ly = \frac{d^2y}{dx^2} + f(x) y = 0 \qquad (4.2.6)$$

As proved in example 4.2.1, the finite difference operator has the expansion

$$I(h) = L + \sum_{i=1}^{p-1} h^{2i} Y_i + O(h^{2p}) .$$

The solution y_h of $I(h) y_h = 0$ must also be even in L, that is $y(h) = y(-h)$, and thus have an expansion of the form

$$y_h(x) = y_0(x) + \sum_{i=1}^{p-1} h^{2i} y_i(x) + O(h^{2p}) . \qquad (4.2.7)$$

This expansion can easily be verified by substituting a general polynomial expansion in h for the solution y_h into the equation $I(h)\, y_h = 0$ and equating coefficients of powers of h. The coefficient $y_0(x)$ of h^0 is the solution of the continuous problem $Ly = 0$ at the point x. All the coefficients of h^k are independent of h.

The mesh transformations are so constructed that the finite difference operator L_h over the graded mesh and the solution have asymptotic expansions in even powers of the parameter h. Thus we have

$$y_h(x) = y_0(x) + h^2\, y_1(x) + h^4\, y_2(x) + O(h^6) \qquad (4.2.8)$$

where $y_0(x)$ is the solution of the true continuous problem at the point x and the $y_i(x)$ are independent of h.

To use such an expansion to refine the accuracy of a solution, consider solutions for parameter values h, $h/2$ and $h/4$ (that is, n, $2n$ and $4n$ mesh points). Then we have the equivalent equation to (4.2.8) for $h/2$,

$$y_{h/2} = y_0(x) + \left(\frac{h}{2}\right)^2 y_1(x) + \left(\frac{h}{2}\right)^4 y_2(x) + O(h^6) \qquad (4.2.9)$$

and again for $h/4$. If we define

$$y_h^{(1)} = \frac{4\, y_{h/2} - y_h}{3}$$

and similarly for $y_{h/2}^{(1)}$, from equations (4.2.8) and (4.2.9), we have the expansion

$$y_h^{(1)} = y_0(x) + h^4\, y_2^{(1)}(x) + O(h^6) \qquad (4.2.10)$$

and a similar expansion for $y_{h/2}^{(1)}$ where the $y_i^{(1)}$ are independent of the parameter h. This process of eliminating the h^2- error term is

called h^2- extrapolation. With a further appropriate linear combination of $y_h^{(1)}$ and $y_{h/2}^{(1)}$, the $O(h^4)$ term may also be eliminated. This process is h^4- extrapolation.

$$y_h^{(2)} = \frac{16\ y_{h/2}^{(1)} - y_h^{(1)}}{15}$$

$$= y_o(x) + O(h^6) .$$

For further references on the more general techniques of extrapolation, see Stetter [39] and the review by Joyce [23] .

The basic method employed in this chapter for systematically choosing the mesh points x_i of a graded mesh is to define a mesh transformation $x(t)$ from a new variable t defined over the unit interval onto the old independent variable x such that an even mesh in the variable t is mapped onto the graded mesh in x. Thus we have

$$x_i = x(ih) , \quad i = 0,1,2, \ldots, n,$$

where

$$h = 1/n .$$

When an equation is discretised over such a graded mesh, it is hoped that an asymptotic expansion in powers of h exists and thus extrapolation can be applied to the results. For consider the equation (4.2.6) discretised as

$$I(h)\ y = \frac{2}{\Delta_+(\Delta_+ + \Delta_-)}\ y_+ - \frac{2}{\Delta_+\Delta_-}\ y_o + \frac{2}{\Delta_-(\Delta_+ + \Delta_-)}\ y_- + f(x_o)\ y_o \qquad (4.2.11)$$

where

$$x_o = x(t) , \ x_+ = x(t+h) \ \text{ and } \ x_- = x(t-h) ,$$

$$\Delta_+ = x_+ - x_o \ \text{ and } \ \Delta_- = x_o - x_- ,$$

and

$$y_0 = y(x_0), \quad y_+ = y(x_+) \text{ and } y_- = y(x_-) .$$

Consider

$$I(-h) .$$

We have

$$\Delta_+(-h) = -\Delta_-(h)$$

$$\Delta_-(-h) = -\Delta_+(h)$$

$$y_+(-h) = y_-(h) \text{ and } y_-(-h) = y_+(h) .$$

Thus equation (4.2.10) does not change so we have

$$I(-h) = I(h)$$

Consistency is an obvious property of the formula (2.2.9) applied in equation (4.2.10) to approximate the second derivative term thus $I(0) = L$. With these results, we have an asymptotic polynomial expansion of y_h in even powers of h as in equation (4.2.7) for some highest power p and extrapolation with respect to the parameter h may be employed.

The important technique used to guarantee an extrapolation principle is the use of a mesh transformation to obtain the points of the graded mesh. The observation may be made that, while in this chapter the mesh transformation is chosen to optimise the performance of the numerical solution scheme, the principles of the above extrapolation results are independent of the reason for the choice of the transform. Thus if suitable mesh transformations that optimise the numerical performance cannot be found or are too difficult to calculate, an analytically known transform that is thought to be intuitively correct in some sense can be used and then extrapolation can be brought to bear to improve the accuracy. This principle is used in Chapter 5.

4.3 SHOOTING METHODS

Osborne [31] has presented a method of choosing the number and position of the shooting points for multiple shooting methods of solving two-point boundary value problems. A short summary of that work is presented below as it has been the motivating force for the later work in this chapter.

Consider the system of ordinary differential equations

$$\frac{d\underset{\sim}{x}}{dt} = \underset{\sim}{f}(\underset{\sim}{x},t) \tag{4.3.1}$$

subject to the boundary conditions

$$\underset{\sim}{g}(\underset{\sim}{x}(0), \underset{\sim}{x}(1)) = \underset{\sim}{0} \tag{4.3.2}$$

where dim $\underset{\sim}{x}$ = dim $\underset{\sim}{f}$ = dim $\underset{\sim}{g}$ = p, and $\underset{\sim}{f}(\underset{\sim}{x},t)$ and $\underset{\sim}{g}(\underset{\sim}{u},\underset{\sim}{v})$ are at least twice continuously differentiable as functions of their arguments.

The boundary value problem (4.3.1) and (4.3.2) can be reduced to the problem of solving a system of equations by noting that if $\underset{\sim}{\Phi}(\underset{\sim}{y},\xi,t)$ satisfies the initial value problem

$$\begin{aligned} \frac{d\underset{\sim}{\Phi}}{dt} &= \underset{\sim}{f}(\underset{\sim}{\Phi},t) \\ \underset{\sim}{\Phi}(\underset{\sim}{y},\xi,\xi) &= \underset{\sim}{y} \end{aligned} \tag{4.3.3}$$

then

$$\underset{\sim}{x}(t) = \underset{\sim}{\Phi}(\underset{\sim}{x}(0), 0, t) \tag{4.3.4}$$

satisfies (4.3.1) and (4.3.2) provided $\underset{\sim}{x}(0) = \underset{\sim}{y}$ satisfies

$$\underset{\sim}{p}(\underset{\sim}{y}) = \underset{\sim}{0} \tag{4.3.5}$$

where

$$\underset{\sim}{p}(\underset{\sim}{y}) = \underset{\sim}{g}(\underset{\sim}{y}, \underset{\sim}{\Phi}(\underset{\sim}{y},0,1)) \tag{4.3.6}$$

The replacement of the original two-point boundary value problem by that of finding an appropriate initial condition such that the solution of the initial value problem also solves the boundary value problem is a characteristic feature of a shooting method.

To complete the specification of the problem (4.3.1) and (4.3.2), conditions which guarantee the existence and uniqueness of the solution are assumed in a form which permits at least in theory the use of Newton's method to solve equation (4.3.5). Thus we assume the existence of a set $S \subset E_p$ and constants K_1 and K_2 such that

(A1) $\exists\, \underset{\sim}{y} \in S$ such that $\underset{\sim}{P}(\underset{\sim}{y}) = \underset{\sim}{0}$,

(A2) $\| \underset{\sim}{P}'(\underset{\sim}{x})^{-1} \| \leq K_1$, $\underset{\sim}{x} \in S$, and

(A3) $\| \underset{\sim}{P}''(\underset{\sim}{x}) \| \leq K_2$, $\underset{\sim}{x} \in S$,

where the norm is the operator norm subordinate to the maximum vector norm, and where the primes denote the appropriate Frechet derivatives. These assumptions guarantee (for example, Luenberger [26]) the existence of a subset $S_1 \subset S$ such that Newton's method is convergent for any choice of initial vector $\underset{\sim}{x}$ from S_1. Implementation of Newton's method to solve (4.3.5) has been extensively discussed in the literature (see Roberts and Shipman [37] and the references there) and it is often found that inordinate care is necessary in the selection of the initial value for the Newton iteration if this is to be convergent. The key step in the Newton iteration is the solution of the system of linear equations

$$\underset{\sim}{P}'(\underset{\sim}{x})\ \underset{\sim}{h} = -\ \underset{\sim}{P}(\underset{\sim}{x}) \tag{4.3.7}$$

which defines the correction $\underset{\sim}{h}$ to the current approximation $\underset{\sim}{x}$. Numerical problems can be anticipated in solving the equation if

$$\chi(\underset{\sim}{P}') = \| \underset{\sim}{P}' \|\ \| \underset{\sim}{P}'^{-1} \|$$

is large. While this does not necessarily imply that the Newton iteration is prejudiced, experience tends to show that it is.

For the class of problems for which A2 applies for a particular K_1

$$X(\underset{\sim}{P}') \leq K_1 \| \underset{\sim}{P}' \| \tag{4.3.8}$$

so that the circumstances in which relative difficulty is encountered in solving (4.3.5) for this class are characterised by $\| \underset{\sim}{P}' \|$ being large. We have

$$\underset{\sim}{P}' = \nabla_u \underset{\sim}{g}(\underset{\sim}{y}, \underset{\sim}{\Phi}(\underset{\sim}{y})) + \nabla_v \underset{\sim}{g}(\underset{\sim}{y}, \underset{\sim}{\Phi}(\underset{\sim}{y})) \underset{\sim}{\Phi}'(\underset{\sim}{y}) \tag{4.3.9}$$

which suggests that the source of the difficulty is, in general, the size of $\| \underset{\sim}{\Phi}' \|$. We have $\underset{\sim}{\Phi}'$ defined by

$$\frac{d\underset{\sim}{\Phi}}{dt} = \nabla_x \underset{\sim}{f}(\underset{\sim}{x},t) \underset{\sim}{\Phi}', \qquad \underset{\sim}{\Phi}'(\underset{\sim}{y},0,0) = I, \tag{4.3.10}$$

so that $\| \underset{\sim}{\Phi}'(\underset{\sim}{y},0,1) \|$ will be large if the differential equation (4.3.1) is unstable about the trajectory $\underset{\sim}{x}(t)$ on the interval [0,1].

Three main approaches have been used in an attempt to overcome these difficulties - (i) a careful choice of starting value to the Newton iteration, (ii) special precautions to stabilise the solution to (4.3.7), and (iii) to reformulate the problem in an attempt to reduce the difficulties caused by ill-conditioning. See Osborne [32] for other references to these approaches. The concern here is with the third approach.

We consider only methods which have the characteristic feature that they determine simultaneously estimates of the solution values $\underset{\sim}{x}_i = \underset{\sim}{x}(t_i)$ at the set of points

$$0 = t_1 < t_2 < \ldots\ldots < t_n = 1.$$

Let $\underset{\sim}{x}^{*T} = (\underset{\sim}{x}_1^T, \underset{\sim}{x}_2^T, \ldots \underset{\sim}{x}_n^T)$ where we denote by * vectors with np components. Then the reformulated problem will lead to the set of equations

$$\underset{\sim}{Q}^{(n)}(\underset{\sim}{z}^*) = \begin{bmatrix} \underset{\sim}{q}_1 \\ \underset{\sim}{q}_2 \\ \vdots \\ \underset{\sim}{q}_n \end{bmatrix} = \underset{\sim}{0} \qquad (4.3.11)$$

Such systems require the properties of consistency and stability; these are discussed by Osborne [32] in detail. If we have

$$\underset{\sim}{q}_1 = \underset{\sim}{g}(\underset{\sim}{z}_1, \underset{\sim}{z}_n)$$

and

$$\underset{\sim}{q}_i = \underset{\sim}{\psi}_i(\underset{\sim}{z}_i, \underset{\sim}{z}_{i-1}), \quad i = 2, 3, \ldots n,$$

then (4.3.11) defines a multiple shooting method if $\underset{\sim}{Q}^{(n)}$ is consistent and stable. For example two algorithms are

1. $\underset{\sim}{\psi}_i(\underset{\sim}{u},\underset{\sim}{v}) = \underset{\sim}{u} - \underset{\sim}{\Phi}(\underset{\sim}{v}, t_{i-1}, t_i)$

and

2. $\underset{\sim}{\psi}_i(\underset{\sim}{u},\underset{\sim}{v}) = \underset{\sim}{u} - \underset{\sim}{v} - \frac{t_i - t_{i-1}}{2}(\underset{\sim}{f}(\underset{\sim}{u},t_i) + \underset{\sim}{f}(\underset{\sim}{v},t_{i-1}))$

Consistency and stability of algorithm 1 are demonstrated by Osborne [32] and we have

$$Q^{(n)'}(\underset{\sim}{z}^*) = \begin{bmatrix} \nabla_u \underset{\sim}{g}(\underset{\sim}{z}_1, \underset{\sim}{z}_n) & & \nabla_v \underset{\sim}{g}(\underset{\sim}{z}_1, \underset{\sim}{z}_n) \\ -X_1(\underset{\sim}{z}_1, t_2) & I & \\ & \ddots & \\ & & -X_{n-1}(\underset{\sim}{z}_{n-1}, t_n) \quad I \end{bmatrix} \tag{4.3.12}$$

where $X_i(\underset{\sim}{z}_i, t)$ satisfies

$$\frac{dX_i}{dt} = \nabla_{\underset{\sim}{x}} f(\underset{\sim}{\Phi}(\underset{\sim}{z}_i, t_i, t), t)\, X_i,$$

$$X_i(\underset{\sim}{z}_i, t_i) = I$$

so that

$$\| Q^{(n)'}(\underset{\sim}{z}^*) \| \leq \max \{ \| \nabla_u \underset{\sim}{g} \| + \| \nabla_v \underset{\sim}{g} \|, \max_{1 \leq i \leq n-1} \| X_i \| + 1 \}.$$

As $\| X_i \| \to 1$ as $|t_{i+1} - t_i| \to 0$, we see that a reasonable choice of bound for $\| Q^{(n)'} \|$ is possible.

(i) provided the boundary conditions are suitably scaled, and

(ii) provided the number and location of the shooting points t_i, $i = 1, 2, \ldots n$ are chosen appropriately.

The norm reduction achieved in this way is the characteristic feature of satisfactory multiple shooting methods.

Consistency for algorithm 2 follows on noting that this procedure is equivalent to integrating (4.3.1) from t_{i-1} to t_i using the trapezoidal rule. Stability is demonstrated by Osborne [32]. Again a reasonable choice

of bound for $\| Q^{(n)'} \|$ is possible provided conditions (i) and (ii) above are satisfied.

For matrices of the form of $Q^{(n)'}$ in (4.3.12), Osborne [32] proves that, under the previous assumptions,

$$\| Q^{(n)'^{-1}} \| \leq K n$$

for some constant K. Thus we have

$$\chi(Q^{(n)'}) \leq L n$$

for some constant L.

The main advantage of multiple shooting is that the number and location of the shooting points are available to reduce the magnitude of $\| Q^{(n)'} \|$ which can be exponentially large in the simple shooting case. The disadvantage is that $\chi(Q^{(n)'})$ grows linearly with n. The implications of this are (a) that optimal strategies exist for the selection of the shooting points and (b) that difficulties with the Newton iteration are again likely for very large values of n.

Note that the spectral radious of $Q^{(n)'}$ is independent of the choice of the shooting points t_i, i = 2, ...n-1. Also for a family H(M) of similar matrices, the matrix U with the smallest condition number will be such that $\| U \| - \rho(U)$ is small compared to $\max \| V \| - \rho(V)$ for $V \in H(M)$. Here $\rho(.)$ denotes the spectral radius for each element of H(M). This is an inverse way of stating the commonly reported phenomenon that numerical difficulties are likely if $\| U \| - \rho(U)$ is large. Provided the boundary conditions are suitably scaled then, for fixed n, the optimum choice of shooting points is proposed to be that for which each $\| X_i(\underset{\sim}{x}_i, t_{i+1}) \|$, i = 1, 2, ..., n-1 is equal to a constant J (say). This choice is hoped to minimise the difference $\| Q^{(n)'} \| - \rho(Q^{(n)'})$.

As a final remark on Osborne's paper [32], note that he suggests that the magnitude of the Lipschitz constant for $\underset{\sim}{f}$ is a good indication of the appropriate number of shooting points required.

§4.4 FIRST ORDER CASE

Consider the two-point boundary value problem (4.1.5) subject to the boundary condition (4.1.6). Suppose that equation is to be integrated using the mid-point rule over each mesh interval of a general mesh

$$0 = x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n = 1 .$$

Then equation (4.1.5) becomes

$$\underset{\sim}{w}_{i-1} - \underset{\sim}{w}_i = \Delta_i A_{i+\frac{1}{2}} \frac{\underset{\sim}{w}_i + \underset{\sim}{w}_{i+1}}{2}$$

where $\Delta_i = x_{i+1} - x_i$ and $A_{i+\frac{1}{2}} = A\left(\frac{x_i + x_{i+1}}{2}\right)$ for appropriate values of the index i. Rearranging this last equation we obtain

$$(-I + \tfrac{1}{2}\Delta_i A_{i+\frac{1}{2}}) \underset{\sim}{w}_{i+1} + (I + \tfrac{1}{2}\Delta_i A_{i+\frac{1}{2}}) \underset{\sim}{w}_i = 0 \qquad (4.4.1)$$

for $i = 0,1,\ldots,n-1$. The boundary conditions (4.1.6) and the equations (4.4.1) can be combined in a block matrix equation for the solution vector $\underset{\sim}{w}^T = (w_1, w_2, \ldots, w_{n-1})$,

$$M \underset{\sim}{w} = \underset{\sim}{b} \qquad (4.4.2)$$

where

$$M = \begin{bmatrix} B_0 & & & B_1 \\ I + \tfrac{1}{2}\Delta_0 A_{\frac{1}{2}} & -I + \tfrac{1}{2}\Delta_0 A_{i+\frac{1}{2}} & & \\ & \ddots & \ddots & \\ & & I + \tfrac{1}{2}\Delta_{n-1} A_{n-\frac{1}{2}} & -I + \tfrac{1}{2}\Delta_{n-1} A_{n-\frac{1}{2}} \end{bmatrix} \qquad (4.4.3)$$

and

$$\underset{\sim}{b}^T = ((1,1), (0,0), \ldots, (0,0)) \ .$$

Note that the matrix (4.4.3) is of the same form as the multiple shooting matrix (4.3.12). In §4.3 it is suggested that norm reduction of the matrix $Q^{(n)'}$ and an easing of the numerical difficulties associated with the matrix $Q^{(n)'}$ will occur by appropriate choice of the multiple shooting points. By analogy the suggestion is made that the application of the techniques for choosing the multiple shooting points for $Q^{(n)'}$ may be made to the matrix M for choosing the mesh points.

The row sum norm of the matrix M is estimated as

$$\|M\| = \max \left\{ \|B_0\| + \|B_1\|, \max_{0 \leq i \leq n-1} \left(\|I + \frac{\Delta_i}{2} A_{i+\frac{1}{2}}\| + \|I - \frac{\Delta_i}{2} A_{i+\frac{1}{2}}\| \right) \right\}$$

As the norms $\|I+\frac{1}{2}\Delta_i A_{i+\frac{1}{2}}\|$ and $\|I-\frac{1}{2}\Delta_i A_{i+\frac{1}{2}}\| \to 1$ as $\Delta_i \to 0$, we see that a reasonable choice of bound for $\|M\|$ is possible

(i) provided the boundary conditions are suitably scaled, and

(ii) provided the locations of the mesh points $x_i, i=1,\ldots,n-1$ are suitably chosen.

In the multiple shooting case the technique for ensuring condition (ii) was to make the row sum norms of the block matrix $Q^{(n)'}$ equal to a constant. Using the same technique on the matrix M we have in the row sum norm

$$\|I + \frac{\Delta_i}{2} A_{i+\frac{1}{2}}\| + \|I - \frac{\Delta_i}{2} A_{i+\frac{1}{2}}\|$$

$$= 2 + \frac{\Delta_i}{2} \|A^-_{i+\frac{1}{2}}\|$$

where

$$A^-_{i+\frac{1}{2}} = A_{i+\frac{1}{2}} - \text{diag}\,(A_{i+\frac{1}{2}})$$

if

$$\frac{\Delta_i}{2} \|\text{diag}\,(A_{i+\frac{1}{2}})\| \lesssim 1. \qquad (4.4.4)$$

It is now easily seen that this equilibration of the block row sum norms is asymptotically (as $n \to \infty$ and $\max_i \Delta_i \to 0$) equivalent to a favourable change of independent variable. For if we have

$$2 + \frac{\Delta_i}{2} \|A^-_{i+\frac{1}{2}}\| = 2 + \gamma h \qquad (4.4.5)$$

for some small parameter h and some constant γ for all i, then, rearranging (4.4.5) to

$$\frac{\Delta_i}{h} = \frac{\gamma}{\frac{1}{2}\|A^-_{i+\frac{1}{2}}\|},$$

we see that asymptotically as $\Delta_i \to 0$, we must have $h \to 0$ and the mesh transformation equation

$$\frac{dx}{dt} = \frac{\gamma}{\|A^-_{(x)}\|} \qquad (4.4.6)$$

some new independent variable t. This transformation (4.4.6) means that asymptotically the transformed differental equation

$$\frac{d\underset{\sim}{w}(x(t))}{dt} = \frac{\gamma\, A(x)}{\|A(x)\|}\, \underset{\sim}{w}(x(t))$$

is being solved over an even mesh in t.

As it is only a matter of length scaling which can be absorbed by the constant γ since the left hand side of (4.4.6) is homogeneous in t of order 1, let the variable t be over the unit interval. We thus have the initial condition

$$x(0) = 0 \tag{4.4.7}$$

and a scaling condition to fix the constant γ

$$x(1) = 1 . \tag{4.4.8}$$

Once the equation (4.4.6) with the conditions (4.4.7) and (4.4.8) is solved for the mesh transformation $x(t)$, then the graded mesh points $\{x_i\}_{i=1}^{n-1}$ may be calculated by

$$x_i = x(ih)$$

for $i = 1,2,\ldots,n-1$ where the parameter h is defined as

$$h = 1/n .$$

Suppose that the matrix $A(x)$ has the form

$$A(x) = \begin{bmatrix} 0 & 1 \\ f(x) & 0 \end{bmatrix}$$

Condition (4.4.4) is satisfied automatically and we have

$$\|A(x)\| = \max \left\{ 1, |f(x)| \right\} .$$

Solving equation (4.4.5) we have

$$t(x) = \frac{1}{\gamma} \int_0^x \|A(x)\| \, dx \ .$$

Thus from (4.4.7), the constant γ can be determined as

$$\gamma = \int_0^1 \|A(x)\| \, dx \ .$$

For many problems of interest, we would have only

$$\|A(x)\| = |f(x)|$$

with its slight simplification.

Since the finite difference scheme is based on a mesh with a parameter h via the mesh transformation function $x(t)$, asymptotic expansions of the form of (4.2.5) in even powers of h are available and can be used for h^2 and h^4 extrapolation as described in §4.2.

This technique for choosing the graded mesh for the finite difference solution of (4.1.5) is quite general. The same principles can also be applied when the trapezoidal rule instead of the mid-point rule is used for equation (4.4.1). For the case of the trapezoidal rule (4.4.1) would become

$$(- I + \tfrac{1}{2}\Delta_i \, A_{i+1}) \, \underset{\sim}{w}_{i+1} + (I + \tfrac{1}{2}\Delta_i \, A_i) \, \underset{\sim}{w}_i = 0$$

with a block row sum norm for the equivalent matrix M of

$$\|M\| = \max \left\{ \|B_0\| + \|B_1\| \ , \ \max_{0 \le i \le n-1} \|I - \tfrac{1}{2}\Delta_i \, A_{i+1}\| + \|I + \tfrac{1}{2}\Delta_i \, A_i\| \right\} \ .$$

Supposing that the boundary conditions are properly scaled, then we have the bound for the block row sum norm of each row of

$$r_i = 2 + \tfrac{1}{2}\Delta_i \{ \| A_i \| + \| A_{i+1} \| \}$$

by applying the triangle inequality to the norms in the last expression. In an attempt to include the coefficient $a(x)$ in the transformation, we use the bound estimate instead of the actual norm for equilibration by the asymptotic mesh transformation. We would have

$$r_i \sim 2 + \gamma h$$

where the symbols have the same meanings as before leading to the asymptotic differential equation for the mesh transformation

$$\frac{dx}{dt} = \frac{\gamma}{\| A(x) \|} .$$

The same initial condition (4.4.7) and scaling condition (4.4.8) still apply. Asymptotic expansions in even powers of the parameter h again exist and h^2 and h^4 extrapolation may be used.

For Δ_i small enough, if the block row sum norms are equilibrated to a constant then we have the expression

$$r_i = 2 + \tfrac{1}{2}(b_i + b_{i+1}) \Delta_i + O(\Delta_i^2) .$$

Asymptotically we must obtain the same mesh transformation as for the mid-point rule, the coefficient $a(x)$ having no effect. Cases do occur in which the function $b(x)$ is a constant (example 4.6.2). For such cases the above mesh transformation based on the norm bound rather than the norms themselves must be used to take account of the effect of the $a(x)$ coefficient.

Example 4.4.1

In equation (4.1.5) let

$$A(x) = \begin{bmatrix} 0 & 1 \\ \dfrac{2}{(x+\varepsilon)^2} & 0 \end{bmatrix} \qquad \ldots \qquad (4.4.8)$$

for some small positive parameter ε with the boundary conditions

$$w_1(0) = w_1(1) = 1 .$$

This equation has the analytic solution

$$\underset{\sim}{w}(x) = \begin{bmatrix} \dfrac{A_1}{x+\varepsilon} + A_2(x+\varepsilon)^2 \\ \dfrac{-A_1}{(x+\varepsilon)^2} + 2A_2(x+\varepsilon) \end{bmatrix} \qquad (4.4.9)$$

for $x \in [0,1]$ where

$$A_1 = \frac{\varepsilon + 3\varepsilon^2 + 2\varepsilon^3}{1 + 3\varepsilon + 3\varepsilon^2}$$

and

$$A = \frac{1}{1 + 3\varepsilon + 3\varepsilon^2}$$

The solution $w_1(x)$ drops from 1 at $x = 0$ to $\frac{1}{2} + O(\varepsilon)$ at $x = \varepsilon$, has a turning point at $O\left(\left(\frac{\varepsilon}{2}\right)^{\frac{1}{2}}\right)$ and behaves like $(x+\varepsilon)^2$ as $x \to 1$. Thus the behaviour of the solution is concentrated at the left hand end of the interval with increasing severity as $\varepsilon \to 0+$. Intuitively we would expect that a 'good' graded mesh would place a major portion of

the mesh points in the $O(\varepsilon)$ region at the $x = 0$ end of the interval. This intuitive feeling is verified by the mesh transformation. From (4.4.8) we have that

$$\|A^{-}(x)\| = |f(x)| ,$$

thus equation (4.4.5) becomes

$$\frac{dx}{dt} = \frac{\gamma}{2}(x+\varepsilon)^2$$

subject to the conditions

$$x(0) = 0 \quad \text{and} \quad x(1) = 1$$

This last equation can easily be solved for the mesh transformation which is

$$x(t) = \frac{\varepsilon\, t}{1 + \varepsilon - t} \tag{4.4.10}$$

Figure 4.4.1 displays the shape of this mesh transformation for the parameter values $\varepsilon = 0.1$ and $\varepsilon = 0.01$. The diagram vividly illustrates the concentration of the mesh points x_i at the left hand end of the interval (0,1). A quantitative measure of the grading of the meshes may be gathered from Table 4.4.1 which contains the locations of the graded mesh points for $n = 10$ and for parameter values of $\varepsilon = 0.1$, 0.01, 0.001 and 0.0001.

The equations (4.4.3) has been solved for $n = 10$, 20 and 40, that is, for $h = 0.1$, 0.05 and 0.025. Extrapolation procedures (h^2, h^4) have been applied to the results and Table 4.4.2 contains more detailed results for the particular case of $\varepsilon = 0.1$ while Table 4.4.3 summarises the results for $\varepsilon = 0.1$, 0.01, 0.001 and 0.0001 by exhibiting the maximum

relative error between the calculated or extrapolated solutions and the analytic solution. In a sense, this measure indicates the minimum number of correct significant digits in the solutions at any mesh point.

TABLE 4.4.1.

EVEN MESH	GRADED MESHES			
n=10	ε = 0.1	0.01	0.001	0.0001
0.1	.1000E-1	.1099E-2	.1110E-3	.1111E-4
0.2	.2222E-1	.2469E-2	.2497E-3	.2500E-4
0.3	.3750E-1	.4225E-2	.4280E-3	.4285E-4
0.4	.5714E-1	.6557E-2	.6656E-3	.6666E-4
0.5	.8333E-1	.9804E-2	.9980E-3	.9998E-4
0.6	.1200	.1463E-1	.1496E-2	.1500E-3
0.7	.1750	.2258E-1	.2326E-2	.2333E-3
0.8	.2667	.3810E-3	.3980E-2	.3998E-3
0.9	.4500	.8182E-1	.8911E-2	.8991E-3

FIGURE 4.4.1.

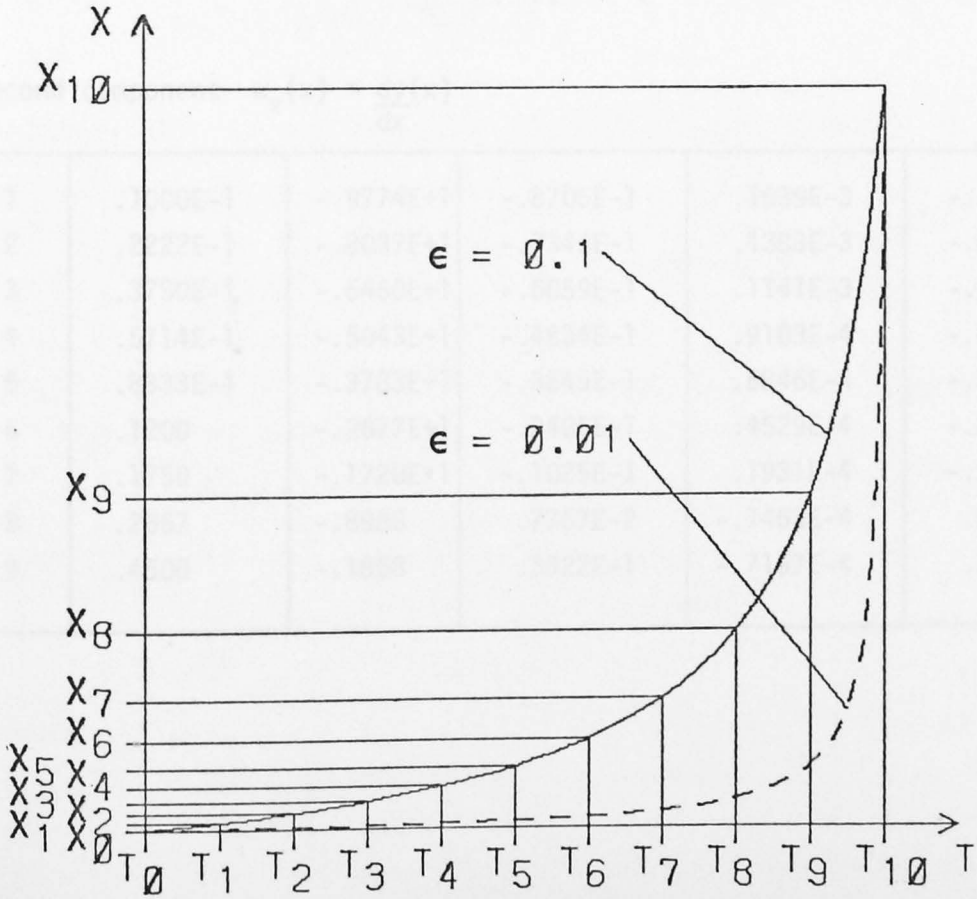


TABLE 4.4.2

Detailed results for solution of example for $\varepsilon = 0.1$ and 4.4.1 for $n = 10$ by first order method.

First component $w_1(x) = y(x)$

No.	Mesh point	Analytic solution	Error in Calc.Soln.	Error in h^2-extrap.	Error in h^4-extrap.
1	.1000E-1	.9114	-.1353E-2	.2549E-5	-.1198E-8
2	.2222E-1	.8233	-.2835E-2	.5340E-5	-.2510E-8
3	.3750E-1	.7360	-.4486E-2	.8448E-5	-.3972E-8
4	.5714E-1	.6501	-.6361E-2	.1198E-4	-.5632E-8
5	.8333E-1	.5666	-.8544E-2	.1609E-4	-.7565E-8
6	.1200	.4875	-.1116E-1	.2101E-4	-.9876E-8
7	.1750	.4178	-.1435E-1	.2703E-4	-.1271E-7
8	.2667	.3718	-.1823E-1	.3433E-4	-.1614E-7
9	.4500	.4079	-.2153E-1	.4055E-4	-.1906E-7

Second component $w_2(x) = \frac{dy}{dx}(x)$

1	.1000E-1	-.9774E+1	-.8706E-1	.1639E-3	-.7708E-7
2	.2222E-1	-.8037E+1	-.7343E-1	.1383E-3	-.6501E-7
3	.3750E-1	-.6460E+1	-.6059E-1	.1141E-3	-.5364E-7
4	.5714E-1	-.5043E+1	-.4834E-1	.9103E-4	-.4279E-7
5	.8333E-1	-.3783E+1	-.3645E-1	.6846E-4	-.3218E-7
6	.1200	-.2677E+1	-.2405E-1	.4529E-4	-.2129E-7
7	.1750	-.1720E+1	-.1025E-1	.1931E-4	-.9074E-8
8	.2667	-.8988	.7767E-2	-.1463E-4	.6882E-8
9	.4500	-.1868	.3822E-1	-.7197E-4	.3384E-7

TABLE 4.4.3

Maximum relative errors in example 4.4.1 for first component $w_1(x) = y(x)$

ε	n	calculated solution	h^2-extrap.	h^4-extrap.
.1	10	.5279E-1	.9941E-4	.4673E-7
	20	.1354E-1	.6367E-5	
	40	.3412E-2		
.01	10	.6965	.1694E-2	.1023E-5
	20	.3949	.2397E-3	
	40	.1092		
.001	10	.9694	.2422E-2	.1620E-5
	20	.9351	.5823E-3	
	40	.8411		
.0001	10	.9970	.2498E-2	.1723E-5
	20	.9939	.6202E-3	
	40	.9875		

Maximum relative errors in example 4.4.1 for second component $w_1(x) = \frac{dy}{dx}$

0.1	10	.7667E-1	.1463E-3	.6878E-7
	20	.3405E-1	.1601E-4	
	40	.4501E-1		
0.01	10	.7609	.1851E-2	.1442E-5
	20	.1601E+1	.9729E-3	
	40	.3996		
0.001	10	.7579E+1	.1893E-1	.2671E-4
	20	.1881E+1	.1158E-2	
	40	.4424		
0.0001	10	.7576E+2	.1898	.3009E-3
	20	.1880E+2	.1158E-1	
	40	.4897		

4.5 SECOND ORDER CASE-I

Consider the two-point boundary value problem

$$\frac{d^2y(x)}{dx^2} - f(x)\, y(x) = 0 \qquad (4.5.1)$$

subject to the conditions

$$y(0) = y(1) = 1 . \qquad (4.5.2)$$

For the equation to have a unique solution (see Keller [25]), it is sufficient for the function $f(x)$ to be continuous and satisfy

$$0 < f_* \leq f(x) \leq f^* \qquad (4.5.3)$$

for $0 \leq x \leq 1$ for some positive constants f_* and f^*. We wish to solve the equation (4.5.1) subject to (4.5.2) by finite difference methods using a graded mesh

$$0 = x_0 < x_1 < x_2 < \quad .... \quad < x_{n-1} < x_n = 1 .$$

If we use a central three-point approximation to the second derivative (see (2.2.9)) in equation (4.5.1), then at the i'th mesh point the finite difference equation is

$$\frac{2}{\Delta_i\,(\Delta_i+\Delta_{i-1})}\, y_{i+1} + \left(\frac{-2}{\Delta_i\Delta_{i-1}} + f_i\right) y_i$$

$$+ \frac{2}{\Delta_{i-1}(\Delta_i+\Delta_{i-1})}\, y_{i-1} = 0$$

for $i = 1,2, \ldots, n-1$ where $\Delta_i = x_{i+1} - x_i$ for $i = 0,1,\ldots n-1$. Multiply the last equation by $\Delta_i\Delta_{i-1}$ to reduce the coefficients of the y_i to

O(1) terms and we obtain

$$\frac{2\Delta_{i-1}}{\Delta_i+\Delta_{i-1}}\ y_{i+1} + (-2 + f_i\ \Delta_i\ \Delta_{i-1})\ y_i + \frac{2\Delta_i}{\Delta_i+\Delta_{i-1}}\ y_{i-1} = 0 \qquad (4.5.4)$$

for $i = 1,2,\ldots,n-1$. The set of equations (4.5.4) can be written as tridiagonal system of linear equations for the vector solution $\underset{\sim}{y}^T = (y_1, y_2,\ldots,y_{n-1})$. Thus we have

$$M\ \underset{\sim}{y} = \underset{\sim}{b} \qquad (4.5.5)$$

where M is an (n-1) x (n-1) square tridiagonal matrix of the form

$$M = \begin{bmatrix} -2 + f_1\Delta_0\Delta_1 & \frac{2\Delta_0}{\Delta_0+\Delta_1} & & & \\ \frac{2\Delta_2}{\Delta_1+\Delta_2} & -2 + f_2\ \Delta_1\Delta_2 & \frac{2\Delta_1}{\Delta_1+\Delta_2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{2\Delta_{n-1}}{\Delta_{n-2}+\Delta_{n-1}} & -2 + f_{n-1}\ \Delta_{n-2}\Delta_{n-1} & \end{bmatrix} \qquad (4.5.6)$$

and where

$$\underset{\sim}{b}^T = \left(\frac{-2\Delta_1}{\Delta_0+\Delta_1} \cdot y(0),\ 0,\ldots,0,\ \frac{-2\Delta_{n-2}}{\Delta_{n-2}+\Delta_{n-1}} \cdot y(1) \right) .$$

In the appendix to Chapter 4, a criterion for the best scaling of real square tridiagonal matrices is given. The criterion is that such a matrix should have all the diagonal elements equal to a constant d (say) if it is to be best scaled.

Suppose we can choose the mesh points x_i so that this criterion is satisfied for the matrix M of (4.5.6). Then we must have

$$\begin{aligned} M_{ii} &= -2 + f_i \Delta_{i-1} \Delta_i \\ &= -2 + K h^2 \end{aligned} \tag{4.5.7}$$

where K is a constant and h is a parameter such that $h \to 0$ as $\delta = \max_{0 \leq i \leq n-1} \Delta_i \to 0$. Thus the condition (4.5.7) is asymptotically equivalent to a favourable change of independent variable from x to t that is defined in the limit as $\delta \to 0$ by the equation

$$f(x) \left(\frac{dx}{dt} \right)^2 = K \tag{4.5.8}$$

Because the left hand side of equation (4.5.8) is homogeneous of order 2 in t, any length scaling of the variable t can be absorbed into the constant K so without loss of generality, we can choose the variable t to be in the unit interval. Then we have the initial condition

$$x(0) = 0 \tag{4.5.9}$$

and the scaling condition

$$x(1) = 1 \tag{4.5.10}$$

to be satisfied. The equation (4.5.8) can be solved analytically as

$$t(x) = \frac{1}{\sqrt{K}} \int_0^x \sqrt{f(x)}\, dx \tag{4.5.11}$$

where K is chosen so that (4.5.10) is satisfied, namely,

$$\sqrt{K} = \int_0^1 \sqrt{f(x)}\, dx \,. \tag{4.5.12}$$

Having found K from (4.5.12), x as a function of t can be found either by inverting $t(x)$ in (4.5.11) or by numerically solving (4.5.8) subject to (4.5.9) as an initial value problem. Very adequate algorithms exist for such problems, for example, Runge-Kutta schemes. The mesh points x_i can be found from this transformation as

$$x_i = x(ih)$$

for $i = 0,1,2,\ldots,n$ where

$$h = 1/n \,.$$

The finite difference scheme (4.5.4) with the mesh points x_i chosen by the mesh transformation is intuitively equivalent to a central difference scheme over an even grid applied to equation (4.5.2) with a favourable change of independent variable. This intuitive view is supported by the structure of the asymptotic diagonal elements of the matrix M (4.5.6), namely,

$$M_{ii} \sim -2 + K h^2 \,.$$

This structure for a diagonal element of a finite difference matrix of the form (4.5.6) results from a central difference scheme applied to an equation of the form

$$\frac{d^2y}{dz^2} + a(z)\frac{dy}{dz} + Ky = 0$$

where K is a constant. For a central difference scheme over an even mesh, the first derivative term $a(z)\frac{dy}{dz}$ makes no contribution to the diagonal

elements. Even if only such were the case it would be an advance. But practical experience with a number of problems indicates that the sub- and super-diagonals are also changed by the mesh transformation such that all their respective elements are almost equal so the final matrix M that results is of a form that would be expected to arise from a transformed equation (4.5.1) of the form

$$\frac{d^2y}{dz^2} + K_1 \frac{dy}{dz} + K_2 y = 0$$

where K_1 and K_2 are constants. Such an equation has a finite difference matrix best scaled by an even mesh.

Because the mesh and hence the finite difference scheme is based on the single parameter h, asymptotic expansions of the solution in that parameter may be sought and can be found.

The solution $y_h(t)$ has an asymptotic expansion of the form

$$y_h(t) = y(x(t)) + h^2 y_1(t) + h^4 y_2(t) + O(h^6) ,$$

where $y(x(t))$ is the exact solution of (4.5.1). The principles of h^2 and h^4 extrapolation as discussed in §4.2 can be applied to the solutions y_h, $y_{h/2}$ and $y_{h/4}$.

Example 4.5.1

Consider the equation

$$\frac{d^2y(x)}{dx^2} - \frac{2}{(x+\varepsilon)^2} y(x) = 0 \qquad (4.5.13)$$

subject to the boundary conditions

$$y(0) = y(1) = 1 .$$

This equation is the second order formulation of the example 4.4.1 and as seen there, the solution is

$$y(x) = \frac{A}{x+\varepsilon} + B(x+\varepsilon)^2$$

for appropriate constants A and B (see example 4.4.1). The mesh transformation equation for equation (4.5.13) is

$$\frac{dx}{dt} = K\,(x+\varepsilon) \qquad (4.5.14)$$

subject to the conditions

$$x(0) = 0 \text{ and } x(1) = 1\ .$$

Equation (4.5.14) has the analytic solution

$$x(t) = \varepsilon\,(\gamma^t - 1)$$

where

$$\gamma = \frac{1+\varepsilon}{\varepsilon}$$

Compare this with the first order transformation (4.4.10).

Equivalent results to those for example 4.4.1 are presented to enable comparison of the methods. Table 4.5.1 contains the locations of the mesh points for n = 10 and parameter values of ε = 0.1, 0.01, 0.001 and 0.0001. Table 4.5.2 contains detailed results for ε = 0.1 and n = 10, 20 and 40 while Table 4.5.3 summarises the results for the parameter values mentioned above.

TABLE 4.5.1

EVEN MESH	GRADED MESHES FOR EXAMPLE 4.5.1			
n=10	ε=0.1	0.01	0.001	0.0001
0.1	.2710E-1	.5865E-2	.9955E-3	.1512E-3
0.2	.6154E-1	.1517E-1	.2982E-2	.5310E-3
0.3	.1053	.2993E-1	.6946E-2	.1485E-2
0.4	.1609	.5335E-1	.1486E-1	.3881E-2
0.5	.2317	.9050E-1	.3064E-1	.9900E-2
0.6	.3235	.1494	.6213E-1	.2502E-1
0.7	.4358	.2429	.1250	.6300E-1
0.8	.5809	.3913	.2504	.1584
0.9	.7655	.6266	.5006	.3980

graded mesh points for second order example 4.5.1 for n = 10.

TABLE 4.5.2

Errors in results for example 4.5.1 with $\varepsilon = 0.1$ for $n = 10$, 20 and 40.

Point	exact soln.	error in calc. soln.	error in h^2-extrap.	error in h^4-extrap.
.2710E-1	.7930	.1421E-10	-.8151E-12	-.3779E-13
.6154E-1	.6340	.3414E-10	-.1958E-11	-.9074E-13
.1053	.5151	.6395E-10	-.3668E-11	-.1700E-12
.1610	.4315	.1102E-9	-.6322E-11	-.2929E-12
.2317	.3820	.1835E-9	-.1052E-10	-.4875E-12
.3215	.3690	.3007E-9	-.1725E-10	-.7989E-12
.4358	.4011	.4891E-9	-.2805E-10	-.1299E-11
.5809	.4944	.7928E-9	-.4547E-10	-.2106E-11
.7655	.6779	.1283E-8	-.7357E-10	-.3408E-11

TABLE 4.5.3

ε	n	solution	h^2-extrap.	h^4-extrap.
0.1	10	.1892E-8	.1085E-9	.5027E-11
	20	.4130E-9	.1212E-10	
	40	.9614E-10		
0.01	10	.6767E-10	.7118E-11	.2084E-12
	20	.1181E-10	.6527E-12	
	40	.2476E-11		
0.001	10	.3228E-10	.4804E-11	.6472E-13
	20	.4489E-11	.3628E-12	
	40	.8512E-12		
0.0001	10	.7657E-11	.1428E-11	.7174E-15
	20	.8445E-12	.8867E-13	
	40	.1447E-12		

Maximum relative errors at the mesh points between the calculated solutions and the analytic solution for example 4.5.1.

Example 4.5.2

Consider the equation

$$\frac{d^2y(x)}{dx^2} - (1 + x^2)\, y(x) = 0$$

subject to the conditions

$$y(0) = 1 \quad \text{and} \quad y(\infty) = 0 .$$

The boundary condition is replaced by $y(10.2) = 0$. This allows some comparison of the results with other attempts on the problem by Osborne [31] (by multiple shooting), Pruess [35] (a method that approximates coefficients) and Allen and Wing [1] (an invariant imbedding approach).

The solution of the mesh transformation equation has the form

$$t(x) = \gamma\, (\sinh^{-1} x + x\sqrt{1+x^2})$$

where γ is a constant such that

$$t(10.2) = 1 .$$

There seems to be no ready method of inverting this function to find $x(t)$ thus the mesh points must be calculated numerically. Table 4.5.4 displays the solutions obtained for $n = 10$ and using h^2 and h^4 extrapolation. A measure of how close the answer is to the true solution may be estimated by comparing the solutions obtained for $n = 10$ with the h^4 extrapolated solution from $n = 10$, 20 and 40.

TABLE 4.5.4

Results for example 4.5.2

mesh point	calc. soln.	h^2 extrap.	h^4 extrap.
0.0000	.1000E+1	.1000E+1	.1000E+1
2.9111	.1591E-1	.2712E-2	.2666E-2
4.3404	.4204E-3	-.6073E-4	.9774E-5
5.4198	.1213E-4	-.3188E-5	.1414E-6
6.3228	.3626E-6	-.1113E-6	.5828E-8
7.1147	.1104E-7	-.3573E-8	.2152E-9
7.8279	.3404E-9	-.1122E-9	.7187E-11
8.5538	.1057E-10	-.3510E-11	.2304E-12
9.0906	.3304E-12	-.1100E-12	.7289E-14
9.6609	.1036E-13	-.3450E-14	.2295E-15
10.2000	.0000	.0000	.0000

4.6 SECOND ORDER CASE-II

Consider the linear two-point boundary value problem

$$\frac{d^2y(x)}{dx^2} + a(x)\,\frac{dy}{dx}(x) + b(x)\,y(x) = 0 \qquad (4.6.1)$$

for $x \in (0,1)$ subject to the conditions

$$y(0) = 1 \quad \text{and} \quad y(1) = 1 \,. \qquad (4.6.2)$$

For the problem (4.6.1) to have a unique solution (see Keller [25]), it is sufficient for the coefficients $a(x)$ and $b(x)$ to be continuous and satisfy

$$|a(x)| < A \quad \text{and} \quad 0 < B_* \leq -b(x) \leq B^* \qquad (4.6.3)$$

for $0 \leq x \leq 1$ and for some positive constants A, B_* and B^*. We wish to solve the equation (4.6.1) subject to (4.6.2) by finite difference methods using a graded mesh

$$0 = x_0 < x_1 < x_2 < \ldots\ldots\ldots < x_{n-1} < x_n = 1 \,.$$

If we use central three-point finite difference approximations to the second derivative (see (2.2.9)) and the first derivative (see (3.1.8)) terms then at the i'th mesh point, the equation (4.6.1) becomes

$$\left(\frac{2}{\Delta_i(\Delta_i+\Delta_{i-1})} + \frac{a_i\ \Delta_{i-1}}{\Delta_i(\Delta_i+\Delta_{i-1})} \right) y_{i+1}$$

$$+ \left(\frac{-2}{\Delta_i\Delta_{i-1}} + \frac{a_i(\Delta_i-\Delta_{i-1})}{\Delta_i\Delta_{i-1}} + b_i \right) y_i$$

$$+ \left(\frac{2}{\Delta_{i-1}(\Delta_i+\Delta_{i-1})} - \frac{a_i\ \Delta_i}{\Delta_{i-1}(\Delta_i+\Delta_{i-1})} \right) y_{i-1} = 0 \,.$$

where $\Delta_i = x_{i+1} - x_i$ for appropriate i. If we multiply this equation by $\Delta_i\Delta_{i-1}$ to reduce the coefficients of the y_i to $O(1)$ terms then we obtain the set of equations

$$\left(\frac{2\Delta_{i-1} + a_i \Delta_{i-1}^2}{\Delta_i+\Delta_{i-1}} \right) \quad y_{i+1}$$

$$+ (-2 + a_i (\Delta_i-\Delta_{i-1}) + b_i \Delta_i\Delta_{i-1}) \quad y_i$$

$$+ \left(\frac{2\Delta_i - a_i \Delta_i^2}{\Delta_i + \Delta_{i-1}} \right) \quad y_{i-1} \quad = 0 . \qquad (4.6.4)$$

for $i = 1,2,\ldots,n-1$. This set of equations can be written as a tri-diagonal system of linear equations for the vector $\underset{\sim}{y}^T = (y_1, y_2,\ldots,y_{n-1})$. Thus we have

$$M \underset{\sim}{y} = \underset{\sim}{b} \qquad (4.6.5)$$

where M is the $(n-1) \times (n-1)$ square tridiagonal matrix

$$M = \begin{bmatrix} -2+a_1(\Delta_1-\Delta_0) + b_1\Delta_1\Delta_0 & \frac{2\Delta_0 + a_1\Delta_0^2}{\Delta_0+\Delta_1} & & \\ \frac{2\Delta_1 - a_2\Delta_2^2}{\Delta_1+\Delta_2} & -2 + a_2(\Delta_1-\Delta_1) + b_2\Delta_2\Delta_1 & \frac{2\Delta_1 + a_2\Delta_1^2}{\Delta_1+\Delta_2} & \\ & \ddots & \ddots & \ddots \\ & & \frac{2\Delta_{n-1} - a_{n-1}\Delta_{n-1}^2}{\Delta_{n-2}+\Delta_{n-1}} & -2+a_{n-1}(\Delta_{n-1}-\Delta_{n-2}) + b_{n-1}\Delta_{n-1}\Delta_{n-2} \end{bmatrix}$$

and where the right hand side vector is

$$\underset{\sim}{b}^T = \left(- \left(\frac{2\Delta_1 - a_1 \Delta_2^2}{\Delta_0 + \Delta_1} \right) y(0),\ 0, \ldots, 0,\ - \left(\frac{2\Delta_{n-1} + a_{n-1} \Delta_{n-2}^2}{\Delta_{n-2}\ \Delta_{n-1}} \right) y(1) \right) .$$

As proved in the Appendix to this chapter, the criterion for a tri-diagonal matrix to be best scaled is that it have all its diagonal elements equal to a constant d (say). Suppose that the mesh point x_i can be chosen so that this criterion is satisfied then we would have

$$M_{ii} = -2 + a_i (\Delta_1 - \Delta_{i-1}) + b_i \Delta_i \Delta_{i-1}$$

$$= -2 + K h^2 \qquad (4.6.6)$$

where K is a constant and h is a parameter such that asymptotically $h \to 0$ as $\delta = \max_{1 \leq i \leq n-1} \Delta_i \to 0$. Thus such a choice of grid points is asymptotically equivalent to a favourable change of independent variable such that the mesh in the new independent variable t is equally spaced in the unit interval. The mesh points x_i can now be expressed as $x_i = x(ih)$ for $i = 0,1,\ldots,n$ where $h = 1/n$.

Assuming that the transformation $x(t)$ satisfies appropriate regularity conditions, if we expand the diagonal element M_{ii} in a Taylor series expansion about the point t_i, then we have the asymptotic identification of (4.6.6) in the form

$$M_{ii} = -2 + a_i \left(h^2 x_i'' + \frac{h^4}{3} x^{(4)} + O(h^6) \right)$$

$$+ b_i \left(h^2 (x_i')^2 + h^4 \left[x' x^{(3)} - (x'')^2 \right]_i + O(h^6) \right)$$

$$= -2 + K h^2$$

where ' is d/dt. Thus we find the differential equation that defines the mesh transformation function is

$$a(x)\ \frac{d^2x}{dt^2}\ +\ b(x)\ \left(\frac{dx}{df}\right)^2\ =\ K \qquad (4.6.7)$$

subject to the initial condition

$$x(0)\ =\ 0 \qquad (4.6.8)$$

and the scaling condition

$$x(1)\ =\ 1\ . \qquad (4.6.9)$$

The left hand side of (4.6.7) is homogeneous in t of order 2, thus condition (4.6.9) only affects the length of the t-interval and the size of the constant K. We lack one condition for the solution of 4.6.7.

Consider the second initial condition

$$\frac{dx}{dt}\ (0)\ =\ \alpha$$

where we have to find a suitable α . Given a value for α , the differential equation (4.6.7) can be integrated numerically for the constant $|K| = 1$ as an initial value problem. The sign of the constant K is that of the function $b(x)$ for compatibility with the case $a(x) \equiv 0$. The integration is continued until the root β of the equation

$$x(\beta)\ =\ 1\ .$$

Thus we have solved the problem

$$a(x)\ \frac{d^2x}{dt^2}\ +\ b(x)\ \left(\frac{dx}{dt}\right)^2\ =\ B.1 \qquad (4.6.10)$$

where $B = \text{sign}\,(b(x))$ over the interval $0 < t < \beta$ subject to the condition

$$x(0) = 0 \text{ and } \frac{dx}{dt}(0) = x .$$

As the left hand side of the equation (4.6.10) is homogeneous in t of order 2, if we make the change of variable $t = \beta t_1$ then $t_1 \in [0,1]$ and we have the rescaled differential equation (4.6.10) to

$$a(x) \frac{d^2x}{dt_1^2} + b(x) \left(\frac{dx}{dt_1}\right)^2 = \beta^2 B$$

subject to the rescaled initial conditions

$$x(0) = 0 \text{ and } \frac{dx}{dt_2} = \beta \alpha .$$

Thus equation (4.6.7) can be solved subject to the conditions (4.6.8) and (4.6.9) but the choice of α is still not fixed.

We choose α to minimise the variation in the computed diagonal elements M_{ii} of the finite difference matrix M for some chosen value of n. The measure of variation used is the range

$$r(\alpha) = \max_{1 \leq i \leq n-1} |M_{ii}| - \min_{1 \leq i \leq n-1} |M_{ii}| .$$

Since the derivative of the function $r(\alpha)$ is not uniquely defined at a finite set of points, a minimisation routine that does not require derivative values is useful. An initial estimate of α that may be used is that obtained from the case $a(x) \equiv 0$, i.e.,

$$\alpha_0 = \frac{1}{\sqrt{|b(0)|}} .$$

From Taylor series expansion arguments applied to the equation (4.6.7), another estimate of α is

$$\alpha_1 = \left\{ \frac{1}{|b(0)|} \left(1 - \frac{a(0)\, b'(0)}{4\, b^2(0)} \right) \right\}^{\frac{1}{2}} .$$

It has been found in the author's experience that this last estimate holds no advantage over the much simpler

$$\alpha_1 = 0.9\ \alpha_0$$

Since the finite difference scheme (4.6.4) is effectively based on the parameter h, asymptotic expansions may be sought for the operator M and for the solution $y(x(t))$. Such expansions may be obtained and are of the form of even powers in the parameter h as in (4.2.8). Thus h^2 and h^4 extrapolation techniques may again be used on the solutions obtained.

Example 4.6.1

Consider the two-point boundary value problem

$$\frac{d^2y}{dx^2}(x) - \frac{1}{(x+\varepsilon)} \frac{dy}{dx}(x) - \frac{3}{(x+\varepsilon)^2}\, y(x) = 0 \qquad (4.6.10)$$

subject to the conditions

$$y(0) = y(1) = 1.$$

This equation is the same as in example 4.1.1. It may be noticed by way of comparison with the results in Table 4.1.1 that the specially chosen graded mesh is considerably more accurate than an even mesh of the same number of points.

The equation (4.6.10) has a solution of the form

$$y(x) = \frac{A}{x+\varepsilon} + B(x+\varepsilon)^3$$

where the A and B are chosen to satisfy the boundary conditions. It can be seen that this equation has a similar behaviour to that in example 4.5.1. Table 4.6.1 exhibits details of the solution and the h^2 and h^4 extrapolated answers for $\varepsilon = 0.1$ for $n = 10$ mesh points. Table 4.6.2 displays the maximum relative errors for the solutions and appropriate extrapolated answers for $n = 10$, 20 and 40 for the parameter values $\varepsilon = 0.1$, 0.01, 0.001 and 0.0001. It can be seen that the errors are greater than in the case of the very similar example 4.5.1. This is probably because the mesh transformation is not as well determined in this case. The benefits of extrapolation are well illustrated by this example.

TABLE 4.6.1

Errors in the solution of example 4.6.1 for $\varepsilon = 0.1$.

Point	exact soln.	error in calc.soln.	error in h^2-extrap.	error in h^4-extrap.
.2710E-1	.7876	.1324E-2	.5359E-5	.1207E-7
.6155E-1	.6215	.2068E-2	.7942E-5	.1449E-7
.1053	.4926	.2408E-2	.9034E-5	.1466E-7
.1610	.3951	.2463E-2	.9381E-5	.1368E-7
.2317	.3262	.2306E-2	.9540E-5	.1210E-7
.3215	.2882	.1975E-2	.9931E-5	.1012E-7
.4358	.2916	.1486E-2	.1075E-4	.7649E-8
.5810	.3624	.8643E-3	.1162E-4	.4333E-8
.7655	.5583	.2250E-3	.1037E-4	-.1270E-9
Sum of squares of absolute errors		.3012E-4	.8093E-9	.1084E-14

TABLE 4.6.2

Maximum relative errors for example 4.6.1

ε	n	solution	h^2-extrap.	h^4-extrap.
0.1	10	.7069E-2	.3689E-4	.3710E-7
	20	.1807E-2	.2330E-5	
	40	.4534E-3		
0.01	10	.6459E-1	.7364E-3	.1964E-5
	20	.1650E-1	.4649E-4	
	40	.4148E-2		
0.001	10	.2383E+0	.4773E-2	.1856E-4
	20	.6072E-1	.3054E-3	
	40	.1530E-1		
0.0001	10	.6130E+0	.1699E-1	.1236E-3
	20	.1555E+0	.1233E-2	
	40	.3836E-1		

Example 4.6.2

We now consider an equation which arises in a practical setting. It describes the stress distribution in a spherical membrane with normal and tangential loads and an attempt is made at a solution in Allen and Wing [1] by an invariant imbedding algorithm.

$$\frac{d^2y}{dx^2} + (3 \cot x + 2 \tan x) \frac{dy}{dx} + 0.7\, y = 0$$

subject to the boundary conditions

$$y(30°) = 0. \,, \quad y\,(60°) = 5.0.$$

The value of y rises from zero to about 283 as x changes from 30° to 30.7°. Allen and Wing also give the results

$$y(40°) = 89.07069,$$
$$y(50°) = 21.26790,$$

which are consistent to six digits with other methods in the literature (see Allen and Wing [1]). Table 4.6.3 gives selected results from the case $n = 20$ with h^2 and h^4 extrapolation applied. The improvements from the extrapolation technique is clearly evident

TABLE 4.6.3

Results for example 4.6.2

No.	Point(°)	Calc.Soln. (n=20)	h^2-extrap.	h^4-extrap.
0	30.0000	.0000	.0000	.0000
1	31.2452	.6146E+3	.9343E+2	.2975E+3
2	32.5750	.4002E+2	.2830E+3	.2299E+3
3	33.9692	.4128E+3	.1195E+3	.2005E+3
4	35.4109	.5378E+2	.1978E+3	.1607E+3
5	36.8871	.2673E+3	.8960E+2	.1379E+3
6	38.3877	.5057E+2	.1299E+3	.1094E+3
7	39.9063	.1685E+3	.6346E+2	.9201E+2
8	41.4345	.4150E+2	.8311E+2	.7243E+2
9	42.9716	.1042E+3	.4342E+2	.5996E+2
10	44.5140	.3146E+2	.5229E+2	.4698E+2
11	46.0598	.6378E+2	.2899E+2	.3849E+2
12	47.6080	.2266E+2	.3257E+2	.3006E+2
13	49.1576	.3885E+2	.1903E+2	.2441E+2
14	50.7081	.1578E+2	.2021E+2	.1909E+2
15	52.2593	.2371E+2	.1238E+2	.1544E+2
16	53.8109	.1076E+2	.1257E+2	.1211E+2
17	55.3626	.1458E+2	.8026E+1	.9791E+1
18	56.9147	.7270E+1	.7879E+1	.7725E+1
19	58.4668	.9096E+1	.5229E+1	.6264E+1
20	60.0000	.5000E+1	.5000E+1	.5000E+1

§4.7 NON-OPTIMAL CHOICE OF MESH

Consider the second order two-point boundary value problem studied in §4.5

$$\frac{d^2y}{dx^2} - f(x)\ y = 0 \tag{4.7.1}$$

subject to the conditions

$$y(0) = y(1) = 1 . \tag{4.7.2}$$

If this equation is discretised over a graded mesh by the finite difference scheme of §4.5, the tridiagonal matrix equation

$$M_2\ \underset{\sim}{y} = \underset{\sim}{b} \tag{4.7.3}$$

is obtained for the solution vector $\underset{\sim}{y}$ where the matrix M_2 is displayed in equation (4.5.6). As explained in §4.5, the optimal mesh transformation $x(t)$ for choosing the mesh points is defined by the differential equation

$$\frac{dx}{dt} = \frac{K}{\sqrt{f(x)}} \tag{4.7.4}$$

subject to the conditions

$$x(0) = 0 \quad \text{and} \quad x(1) = 1 . \tag{4.7.5}$$

We wish to examine the effects of a non-optimal choice of mesh transformation $x(t)$ on the condition number of the matrix M_2 of equation (4.7.3) and on the errors in the calculated and h^2 and h^4 extrapolated solutions. The condition number of a non-symmetric matrix M is defined as

$$\kappa(M) = \left\{\frac{\Lambda(M^TM)}{\lambda(M^TM)}\right\}^{\frac{1}{2}}$$

where $\Lambda(A)$ is the maximal eigenvalue and $\lambda(A)$ is the minimum eigenvalue of the symmetric positive definite matrix A.

Consider the mesh transformation $x(t)$ defined by the equation

$$\frac{dx}{dt} = \frac{k}{\{f(x)\}^{\beta}} \qquad (4.7.6)$$

subject to the conditions (4.7.5) for values of β in the range

$$0 \leqslant \beta \leqslant 2 .$$

The numerical scheme used is as described in §4.5. Only the mesh transformation is changed.

The particular example considered is example 4.5.1 for which

$$f(x) = \frac{2}{(x+\varepsilon)^2}$$

and we study only the case $\varepsilon = 0.1$.

The Table 4.7.1 of condition numbers of the matrix M_2 for meshes with $n = 10$, 20 and 40 completely supports the analysis of best conditioned tridiagonal matrices presented in the Appendix to Chapter 4 of this thesis. The criterion for the choice of mesh transformation for the second order case (4.7.1) is based on this analysis. Table 4.7.2 of the maximum relative errors in the solutions and the h^2 and h^4 extrapolated solutions (if applicable) for the cases $n = 10$, 20 and 40 also fully supports the choice $\beta = 0.5$ for the optimal mesh transformation for the second order case (4.7.1).

TABLE 4.7.1

CONDITION NUMBERS - SECOND ORDER CASE

β	n = 10	n = 20	n = 40
0.0	.2394E+2	.9514E+2	.3800E+3
0.1	.2208E+2	.8771E+2	.3503E+3
0.2	.2046E+2	.8123E+2	.3243E+3
0.3	.1921E+2	.7622E+2	.3043E+3
0.4	.1846E+2	.7330E+2	.2927E+3
0.5	.1836E+2	.7298E+2	.2915E+3
0.6	.1984E+2	.7541E+2	.3013E+3
0.7	.2013E+2	.8027E+2	.3209E+3
0.8	.2181E+2	.8701E+2	.3479E+3
0.9	.2382E+2	.9503E+2	.3801E+3
1.0	.2606E+2	.1038E+3	.4153E+3
1.25	.3223E+2	.1271E+3	.5072E+3
1.5	.3809E+2	.1500E+3	.5946E+3
1.75	.4242E+2	.1698E+3	.6735E+3
2.0	.4520E+2	.1848E+3	.7393E+3

TABLE 4.7.2

MAXIMUM RELATIVE ERRORS - SECOND ORDER CASE

β	n = 10 solution	10 h^2 extrap	10 h^4 extrap	20 solution	20 h^2 extrap	40 solution
0.0	.1024	.9387E-2	.5150E-3	.3263E-1	.1070E-2	.8931E-2
0.1	.6387E-1	.3172E-2	.9380E-4	.1835E-1	.2915E-3	.4797E-2
0.2	.3663E-1	.9624E-3	.1308E-4	.9867E-2	.7241E-4	.2525E-2
0.3	.1888E-1	.2352E-3	.1423E-5	.4958E-2	.1613E-4	.1251E-2
0.4	.7771E-2	.4279E-4	.1018E-6	.1975E-2	.2789E-5	.4975E-3
*0.5	.1892E-8	.1085E-9	.5027E-11	.4130E-9	.1212E-10	.9614E-10
0.6	.6307E-2	.6845E-5	.4286E-9	.1581E-2	.4280E-6	.3956E-3
0.7	.1273E-1	.6716E-5	.5500E-8	.3185E-2	.4402E-6	.7994E-3
0.8	.2047E-1	.1526E-4	.4974E-8	.5284E-2	.9496E-6	.1320E-2
0.9	.3370E-1	.4531E-4	.3234E-7	.8390E-2	.2801E-5	.2108E-2
1.0	.5279E-1	.9941E-4	.4674E-7	.1354E-1	.6367E-5	.3412E-2
1.25	.1681	.1906E-1	.8201E-4	.5121E-1	.3153E-3	.1304E-1
1.5	.3014	.1957E-1	.2405E-2	.1534	.6936E-2	.5482E-1
1.75	.3980	.5767E-1	.1332E-1	.2709	.3287E-1	.1480
2.0	.4585	.1030	.3412E-1	.3619	.7520E-1	.2514

Consider the first order equation equivalent to equation (4.7.1), namely

$$\frac{d\underset{\sim}{w}}{dx} = A(x)\underset{\sim}{w} \qquad (4.7.7)$$

where

$$A(x) = \begin{bmatrix} 0 & 1 \\ f(x) & 0 \end{bmatrix}$$

and

$$\underset{\sim}{w}^T = \left(y, \frac{dy}{dx}\right)$$

subject to the same boundary conditions as (4.7.2) in the form

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \underset{\sim}{w}(0) + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \underset{\sim}{w}(1) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

For this equation as formulated in §4.4, the 'optimal' mesh transformation is

$$\frac{dx}{dt} = \frac{K}{f(x)}$$

subject to the conditions (4.7.5). Note that this choice of mesh transformation is based on the proposition that equilibration of the block row sum norms of the first order finite difference matrix M_1 (see (4.4.3)) to a constant enhances the accuracy of the numerical solution.

Again we examine the effects of the mesh transformation (4.7.6) for various values of β on the same example function $f(x) = \frac{2}{(x+\varepsilon)^2}$ as before. The condition numbers of the first-order finite difference matrices M_1 are displayed in Table 4.7.3 for the cases $n = 10$, 20 and 40. The maximum relative errors for the same cases are

TABLE 4.7.3

CONDITION NUMBERS - FIRST ORDER CASE

β	n = 10	20	40
0.0	.1515E+3	.1279E+3	.1449E+3
0.1	.1242E+3	.1083E+3	.1230E+3
0.2	.1061E+3	.9713E+2	.1131E+3
0.3	.9441E+2	.9087E+2	.1143E+3
0.4	.8681E+2	.8758E+2	.1259E+3
0.5	.8176E+2	.8725E+2	.1426E+3
0.6	.7824E+2	.9197E+2	.1620E+3
0.7	.7574E+2	.1008E+3	.1831E+3
0.8	.7429E+2	.1113E+3	.2052E+3
0.9	.7586E+2	.1226E+3	.2280E+3
1.0	.9361E+2	.1363E+3	.2514E+3
1.1	.1263E+3	.1534E+3	.2753E+3
1.25	.2074E+3	.2029E+3	.3146E+3
1.5	.4461E+3	.4072E+3	.4609E+3
1.75	.8215E+3	.7928E+3	.8518E+3
2.0	.1320E+4	.1354E+4	.1517E+4

TABLE 4.7.4

MAXIMUM RELATIVE ERRORS - FIRST ORDER CASE

β	(n = 10) solution	(n = 10) h^2 extrap	(n = 10) h^4 extrap	(n = 20) solution	(n = 20) h^2 extrap	(n = 40) solution
0.0	.3723	.2224E-1	.2213E-3	.7960E-1	.1601E-2	.1890E-1
0.1	.2146	.9481E-2	.2896E-4	.4823E-1	.5691E-3	.1170E-1
0.2	.1305	.3505E-2	.1956E-4	.3113E-1	.2033E-3	.7670E-2
0.3	.8862E-1	.1304E-2	.6581E-5	.2156E-1	.7709E-4	.5348E-2
0.4	.6507E-1	.5262E-3	.1783E-5	.1598E-1	.3158E-4	.3991E-2
0.5	.5167E-1	.2317E-3	.5475E-6	.1282E-1	.1400E-4	.3198E-2
0.6	.4447E-1	.1081E-3	.1751E-6	.1109E-1	.6674E-5	.2770E-2
0.7	.4138E-1	.5482E-4	.6191E-7	.1035E-1	.3477E-5	.2598E-2
0.8	.4071E-1	.6355E-4	.8949E-7	.1056E-1	.3888E-5	.2640E-2
0.9	.4603E-1	.5491E-4	.4414E-7	.1152E-1	.3390E-5	.2897E-2
1.0	.5279E-1	.9941E-4	.4673E-7	.1354E-1	.6367E-5	.3412E-2
1.1	.7130E-1	.7052E-3	.1056E-4	.1730E-1	.5951E-4	.4284E-2
1.25	.1034	.2847E-2	.1828E-3	.2870E-1	.5280E-3	.6778E-2
1.5	.1644	.3266E-2	.1317E-2	.7022E-1	.2735E-2	.2066E-1
1.75	.2214	.1158E-1	.8966E-4	.1284	.1503E-2	.5679E-1
2.0	.2651	.3786E-1	.7996E-2	.1849	.1947E-1	.1087

found in Table 4.7.4.

The results offer some support for the hypothesis that the original choice of mesh transformation ($\beta = 1.0$) for the first order case is near the optimal for the first order formulation of the problem especially for small values of n. What is definitely supported by Table 4.7.4 is the merits of h^2 and h^4 extrapolation. It is seen that as long as the mesh transformation is near optimal (as is the case $\beta = 1.0$) then the usage of extrapolation can mean excellent improvement in the accuracy of the solution.

Table 4.7.4. suggests that if the optimal mesh transformation is not known and is to be chosen intuitively, than a conservative choice of mesh transformation is to be favoured. The increase in the condition numbers in Table 4.7.3. as the parameter β climbs away from the optimum reflects the increasing effect of the large truncation error on the numerical solution of the problem.

A possible reason is as follows. A conservative grid selects information roughly from throughout the interval and the finite difference scheme has an overview of the problem. There is the possibility that small regions of large variation of the coefficients are not well modelled. A mesh transformation that is not near the optimal and is significantly different from the unit transformation (e.g. $\beta > 1$) can place too many points in small regions such that the equation is not well-modelled mesh or they receive scant attention and other regions of less importance are emphasised. With such bad-modelling of the problem by a grid comes large truncation errors in the finite difference equations which are reflected in the numerical accuracy of the solution.

§4.8 SUMMARY AND DISCUSSION

In this chapter a method is discussed for systematically choosing the mesh points of a graded mesh. That graded mesh is to be used in the solution of linear two-point boundary value problems by finite difference methods. The basic technique is to define a mesh transformation that maps the points of an even grid in a new independent variable into an appropriate graded mesh in the old independent variable. A discussion of the meaning of 'appropriate' and of the ramifications of the use of a mesh transformation are discussed in §4.4, §4.5 and §4.6.

Only when a mesh transformation is used is the graded mesh and hence the finite difference scheme based on the single parameter h, the step width of the underlying even mesh. With sufficient regularity conditions on the mesh transformation, asymptotic expansions in the parameter h of the solution of the finite difference scheme can be obtained. If suitable difference schemes are chosen for the equation over the graded mesh, then the asymptotic expansions are polynomial expansions in even powers of the parameter h. With such expansions on hand, if solutions are obtained for parameter values of h, $h/2$ and $h/4$ then h^2 and h^4 extrapolation may be applied to these solutions. This form of extrapolation is just the selective linear combinations of solutions for different parameter values to eliminate the lower order error terms of the asymptotic expansion as discussed in §4.2.

One major concern of this chapter and in fact specific to this chapter is the selection of the mesh transformation for equations of the form $\frac{d^2y}{dx^2} + a(x)\frac{dy}{dx} + b(x)y = 0$ in either first or second order formulation. For this example the mesh transformation may be chosen such that the resultant numerical scheme possesses optimal numerical properties. In the first order case (§4.4), the criterion used was the equilibration

of the block row sum norms of the finite difference matrix (4.4.3). For the second order cases treated (§4.5 and §4.6), choices of mesh points were made that made equal all the diagonal elements of the finite difference matrices (4.5.4) and (4.6.4); that choice resulting in a best-scaled matrix equation as discussed in the Appendix to Chapter 4.

Note that if a second order problem $\frac{d^2y}{dx^2} + f(x)\, y = 0$ is also formulated as a first order problem then the concentration of mesh points is more severe for the first order case than for the second order case. The first order transformation is defined by the differential equation

$$\frac{dx}{dt} = \frac{\gamma}{|f(x)|}$$

while the second order transformation is defined by

$$\frac{dx}{dt} = \frac{\gamma}{\sqrt{|f(x)|}} .$$

It is thus seen that the gradient of the mesh transformation in the second order case is ameliorated by the square root function. The difference in the distribution of the points can be striking. Compare Table 4.4.1 with Table 4.5.1 for an example.

Consider equations of the type $\varepsilon\frac{d^2y}{dx^2} + a(x)\frac{dy}{dx} + b(x)y = 0$ where ε is a positive parameter. Such equations are mentioned in §4.1 as a one-dimensional analogue of the full Navier-Stokes equations and methods were hoped to be developed to efficiently solve such equations for small parameter values. A complete answer has not as yet been obtained. The character of the solutions is very dependent on the size of ε. If the parameter ε is very small ($\varepsilon < 10^{-2}$ say), then the

character of the problem changes from the one discussed in this chapter and it becomes a singular perturbation problem. For such problems usually $\frac{d^2y}{dx^2}$ is not large enough for the term $\varepsilon\frac{d^2y}{dx^2}$ to make a significant contribution to the equation for most of the interval length so the equation is effectively of lower order for most of the interval. Where such is not the case, with very large $\frac{d^2y}{dx^2}$, significant changes must occur in small regions of, very roughly, $O(\varepsilon)$ with the consequent development of boundary layer type regions. Because the coefficient of $\frac{d^2y}{dx^2}$ is a constant, it does not affect the criterion used for choosing the mesh transformation for such cases. Hence it seems that the above criterion for the choice of mesh transformation is not applicable in these cases because of the entirely different character of the solution.

Fortunately the principle of the mesh transformation and the resultant extrapolation principle are still available even if the criteria are not known for the optimal choice of points for the performance of the numerical scheme. Thus, if some rough ideas of the solution behaviour can be gained,for example, a brief asymptotic analysis (Pearson [34]) of the equation, an analytic mesh transformation can be constructed that would place the bulk of any mesh points chosen by this mesh transformation in those regions of most rapid expected change. This is not as satisfactory as the above cases but very importantly, the principles of extrapolation still apply and as has been demonstrated, can dramatically improve the accuracy of the calculated solution.

Appendix

Best Scaled Tridiagonal Matrices

A.1 Introduction

Consider the following characterisation of best scaling of matrices by Varah and Golub [44]. Let A be a n x n real non-singular matrix. Consider

$$\min_{D,E} \quad \kappa_2(DAE) = \frac{\sigma_1}{\sigma_n}$$

$$= \text{ratio of largest to smallest singular value of DAE,}$$

where D and E are diagonal matrices, that is, the best L^2 diagonal scaling characterisation. The matrix DAE is best scaled if, in the singular value decomposition $DAE = U\Sigma V^T$, we have $|u_{i1}| = |u_{in}|$ and $|v_{i1}| = |v_{in}|$ for i = 1, 2, ...n, that is, the first and last columns of U and V have components of equal magnitude.

Proof:

$$\frac{\sigma_1}{\sigma_n}(DAE) = \max_{\underset{\sim}{p},\underset{\sim}{q},\underset{\sim}{t},\underset{\sim}{s}} \left(\frac{\dfrac{|\underset{\sim}{p}^T DAE \underset{\sim}{q}|}{\|\underset{\sim}{p}\|_2 \, \|\underset{\sim}{q}\|_2}}{\dfrac{|\underset{\sim}{r}^T DAE \underset{\sim}{s}|}{\|\underset{\sim}{r}\|_2 \, \|\underset{\sim}{s}\|_2}} \right)$$

$$= \max_{\underset{\sim}{p},\underset{\sim}{q},\underset{\sim}{r},\underset{\sim}{s}} \left(\frac{\dfrac{|\underset{\sim}{p}^T A \underset{\sim}{q}|}{\|D^{-1}\underset{\sim}{p}\|_2 \, \|E^{-1}\underset{\sim}{q}\|_2}}{\dfrac{|\underset{\sim}{r}^T A \underset{\sim}{s}|}{\|D^{-1}\underset{\sim}{r}\|_2 \, \|E^{-1}\underset{\sim}{s}\|_2}} \right)$$

$$\geq \frac{\sigma_1}{\sigma_n}(A) \left[\frac{(\underset{\sim}{r}^T D^{-2} \underset{\sim}{r})\ (\underset{\sim}{s}^T E^{-2} \underset{\sim}{s})}{(\underset{\sim}{p}^T D^{-2} \underset{\sim}{p})\ (\underset{\sim}{q}^T E^{-2} \underset{\sim}{q})} \right]^{\frac{1}{2}}$$

if we take $\underset{\sim}{p}, \underset{\sim}{q}, \underset{\sim}{r}$ and $\underset{\sim}{s}$ to be singular vectors of A.

If $|r_i| = |p_i|$ and $|s_i| = |q_i|$ for i = 1, 2, ...n, then the second term is unity and

$$\kappa_2 (DAE) \geq \kappa_2(A)$$

This condition for best scaling is sufficient but not necessary.

Forsythe and Straus [17] give the same kind of characterisation for the diagonal scaling DAD of a positive definite matrix A.

Let A be a positive definite Hermitian matrix of finite order n, and let Λ and λ be its maximal and minimal eigenvalues respectively. The condition number of A is the ratio $\kappa(A) = \Lambda/\lambda$. Let $\mathcal{T}$ be a class of regular linear transformations and define $A^T = T^* A T$ where $T \in \mathcal{T}$. We say that A is *best conditioned with respect to* $\mathcal{T}$ if $\kappa(A^T) \geq \kappa(A)$ for all $T \in \mathcal{T}$.

For positive definite symmetric matrices with maximum eigenvalue Λ and minimum eigenvalue λ with corresponding eigenvectors $\underset{\sim}{x}^{\Lambda}$ and $\underset{\sim}{x}^{\lambda}$ respectively, Forsythe and Straus state the theorem (theorem 3, p.343 of [17])

Theorem: A sufficient condition for matrix A to be best conditioned with respect to a class $\mathcal{T}$ of transformations is that, for some pair of eigenvectors $\underset{\sim}{x}^{\Lambda}$ and $\underset{\sim}{x}^{\lambda}$ belonging to the eigenvalues Λ and λ respectively, we have

$$|x_i^{\Lambda}| = |x_i^{\lambda}|, \quad i=1,2,\ldots,n. \qquad (A.1)$$

Moreover, if Λ and λ are simple eigenvalues, then (A.1) is also necessary.

Condition (A.1) is the equal magnitude condition but for positive definite symmetric matrices, that condition is both necessary and sufficient for a best conditioned matrix if the eigenvalues are simple. Consider the class S of symmetric tridiagonal matrices. It is claimed that for such matrices, condition (A.1) is equivalent to the matrix having a constant diagonal.

A.2 Theorems

We first prove a subsidiary lemma.

Lemma: If A is a symmetric tridiagonal matrix with all of its diagonal elements equal to b, then the diagonal element

$$b = \frac{\lambda + \Lambda}{2}$$

where λ and Λ are the smallest and largest eigenvalues respectively.

Proof: We first split A so that

$$A = T + bI$$

where T is a tridiagonal matrix with a zero diagonal. The lemma needs to be proven only for the matrix T, for if λ^T and Λ^T are the smallest and largest eigenvalues of T, then $\lambda = \lambda^T + b$ and $\Lambda = \Lambda^T + b$ are the smallest and largest eigenvalues respectively of the matrix A. If the lemma is true for the matrix T, i.e.,

$$\lambda^T + \Lambda^T = 0.$$

then

$$\frac{\lambda + \Lambda}{2} = \frac{\lambda^T + b + \Lambda^T + b}{2}$$

$$= b ,$$

hence the lemma is true for the matrix A.

Consider the maximum eigenvalue of T:

$$\Lambda^T = \max_{\underset{\sim}{v}\in R^n} \frac{\underset{\sim}{v}^T T \underset{\sim}{v}}{\underset{\sim}{v}^T \underset{\sim}{v}}$$

$$= \max_{\underset{\sim}{v}\in R^n} \frac{\sum_{i=2}^{n} A_{i,i-1} v_i v_{i-1} + \sum_{i=1}^{n-1} A_{i,i+1} v_i v_{i+1}}{\sum_{i=1}^{n} v_i^2} . \qquad (A.2)$$

Let $\underset{\sim}{v}^*$ be a vector for which Λ^T is attained by the right hand side of (A.2), then if $\underset{\sim}{v}^{*T} = (v_1, v_2, \ldots, v_n)$, then

$$\Lambda^T = \frac{\sum_{i=2}^{n-1} A_{i,i-1} v_i v_{i-1} + \sum_{i=1}^{n-1} A_{i,i+1} v_i v_{i+1}}{\sum_{i=1}^{n} v_i^2}$$

Let the vector $\underset{\sim}{w}^T = (w_1, w_2, \ldots, w_n)$ have components

$$w_i = (-1)^i v_i \ , \quad i=1,2,\ldots,n . \qquad (A.3)$$

Then we have

$$w_{i-1} w_i = -v_{i-1} v_i$$

for $i = 2,3,\ldots n$, and

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} v_i^2 .$$

Therefore

$$\Lambda^T = - \frac{\sum_{i=2}^{n} A_{i,i-1} w_i w_{i-1} + \sum_{i=1}^{n-1} A_{i,i+1} w_i w_{i+1}}{\sum_{i=1}^{n} w_i^2}$$

$$= - \lambda^T \qquad (A.4)$$

by the definition of $\lambda^T = \min_{\underset{\sim}{v}\epsilon R^n} \frac{\underset{\sim}{v}^T T \underset{\sim}{v}}{\underset{\sim}{v}^T \underset{\sim}{v}}$. For if $\underset{\sim}{v}_2$ were a vector such that

$\frac{\underset{\sim}{v}_2^T T \underset{\sim}{v}_2}{\underset{\sim}{v}_2^T \underset{\sim}{v}_2} < -\Lambda^T$, then by applying the same transformation (A.3) to $\underset{\sim}{v}_2$,

a vector $\underset{\sim}{v}_3$ is obtained such that

$$\frac{\underset{\sim}{v}_3^T T \underset{\sim}{v}_3}{\underset{\sim}{v}_3^T \underset{\sim}{v}_3} > \Lambda^T$$

which is a contradiction, hence (A.4) is true,

$$\lambda^T + \Lambda^T = 0$$

and the lemma is proved.

We now state and prove the major theorem of this appendix.

Theorem: For a symmetric tridiagonal matrix A, if $\underset{\sim}{v}$ is an eigenvector of the maximal eigenvalue Λ of A, then $\underset{\sim}{w}^T = (w_1, w_2, \ldots, w_n)$ with $w_i = (-1)^i v_i$, $i=1,2,\ldots,n$ is an eigenvector of λ, the minimal eigenvalue of A if and only if the matrix A has a constant diagonal.

Proof: We first prove necessity of a constant diagonal.

Let the constant diagonal element be b. Define the diagonal matrix P with diagonal elements $P_{ii} = (-1)^i$, $i=1,2,\ldots,n$. Then

$$\underset{\sim}{w} = P \underset{\sim}{v}$$

and

$$P^2 = I .$$

Now we have

$$(A - \Lambda I)\underset{\sim}{v} = 0 . \qquad (A.5)$$

Splitting $A = T + bI$ as before and pre-multiplying (A.5) by $-P$, we have

$$-P(T + (b - \Lambda)I)P^2\underset{\sim}{v} = 0$$

or

$$(-PTP - (b - \Lambda)I)P\underset{\sim}{v} = 0 . \qquad (A.6)$$

From the above lemma, we have that

$$b = \frac{\lambda + \Lambda}{2}$$

hence

$$b - \Lambda = -(b - \lambda) .$$

Substituting this result into equation (A.6), we obtain

$$(-PTP + (b - \lambda)I)\underset{\sim}{w} = 0 . \qquad (A.7)$$

Consider an off-diagonal element of $-PTP$, say, for appropriate i :

$$(-PTP)_{i,i-1} = -P_{i,i}\ A_{i,i-1}\ P_{i-1,i-1}$$

$$= -(-1)^i\ A_{i,i-1}(-1)^{i-1}$$

$$= A_{i,i-1} .$$

Hence

$$-PTP = T .$$

Therefore, equation (A.7) reduces to

$$(A - \lambda I)\underset{\sim}{w} = 0 , \qquad (A.8)$$

and

$$(A - \Lambda I)\underset{\sim}{v} = 0 . \qquad (A.9)$$

If we define $\underset{\sim 0}{v} = \underset{\sim 0}{w} = \underset{\sim n+1}{v} = \underset{\sim n+1}{w} = 0$ for simplicity of the notation, then on expanding the i'th rows of each of equations (A.8) and (A.9), for $i=1,2,\ldots,n$, we obtain

$$A_{i,i-1}\, w_{i-1} + (A_{ii} - \lambda)w_i + A_{i,i+1}\, w_{i+1} = 0 \qquad (A.10)$$

and

$$A_{i,i-1}\, v_{i-1} + (A_{ii} - \Lambda)v_i + A_{i,i+1}\, v_{i+1} = 0 \,. \qquad (A.11)$$

If equation (A.10) is multiplied by $(-1)^i$ and added to (A.11), we obtain

$$2A_{ii} - (\lambda + \Lambda) = 0 \,, \quad i=1,2,\ldots,n \,,$$

that is,

$$A_{ii} = \frac{\lambda + \Lambda}{2} \quad \text{for} \quad i=1,2,\ldots,n \,,$$

or that the matrix A has a constant diagonal.

Hence the theorem is proved.

Applying the above results to positive definite symmetric tridiagonal matrices from class S, and using the theorem in A.2 from Forsythe and Straus [17], we have that such matrices are best conditioned in the sense of that theorem if they have contant diagonals.

CHAPTER 5

GRADED MESHES IN THE NAVIER-STOKES EQUATIONS

§5.1 SOLUTION SCHEME FOR GRADED MESHES

Consider the general graded mesh over the unit square

$$0 = x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n = 1$$

and

$$0 = y_0 < y_1 < y_2 < \ldots < y_{m-1} < y_m = 1 .$$

Let $h_i = x_{i+1} - x_i$ and $k_j = y_{j+1} - y_j$ for appropriate i and j. For the solution of the Poisson equation for the streamfunction (3.2.7), the discretised Laplacian operator ∇_h^2 for a general rectangular grid has been given in (3.1.3). Thus at the (i,j) mesh point, for the streamfunction we have the equation

$$\nabla_h^2 \ \psi_{ij}^n = -\omega_{ij}^n$$

Given ω^n this equation can be solved for ψ^n by one of the direct methods of Chapter 2.

The explicit time step method that is used to solve the vorticity transport equation is the same as for the regular mesh case - equation (3.2.6). Both the convective $(\underset{\sim}{u}.\nabla\omega)$ and the divergence $(\nabla.(\underset{\sim}{u}\omega))$ forms of the convention term are examined in the numerical scheme. Central three point approximation formulations are used in the finite difference representation of all the derivatives in the equations. The relevant formulas are given in §3.1 which discusses the discretisation of the spatial derivatives.

Although concrete algorithms have not been developed for systematically choosing a graded mesh that is problem dependent even for the one-dimensional analogues of the Novier-Stokes equation, the mesh transformation principle (§4.7) is still valid and can be used to choose a mesh in each direction. One level of extrapolation can be applied to the results for h, k and $\frac{h}{2}, \frac{k}{2}$ mesh spacings in the mesh transformation variables t_x and t_y. Even if different numbers of mesh points are used in each direction (n and m say), h^2 extrapolation in the form $y^{(1)} = \frac{4\, y_{h/2,\, k/2} - y_{h,k}}{3}$ will still be valid if the ratio n/m is kept fixed when the mesh is refined. The actual mesh transformation functions are chosen intuitively as was mentioned in §4.7 and as discussed below. For a particular choice of problem parameters (Re, n, m, etc) the mesh transformations are fixed for that problem.

The existence of singularities in the vorticity at the upper corners of the cavity tends to create difficulties for the extrapolation process near those points as to the author's knowledge, asymptotic expansions, upon which the extrapolation process is based, are not known for the flow past those corners. Elsewhere in the cavity the solution is well-behaved and no difficulties can be expected or are encountered in the extrapolation.

Given the mesh transformations for the problem under consideration, the solution scheme follows the basic pattern laid down in §3.4 except that

1) the original graded meshes are picked by x and y mesh transformations,

2) after convergence has been attained for the mesh parameters h and k the number of mesh points in each direction is doubled ($h \to h/2$, $k \to k/2$). The previous converged vorticity solution is interpolated by any simple scheme onto the new extended grid. The resultant vorticity field is considered as the initial field ω^0 for a new problem and the algorithm is restarted.

3) h^2 extrapolation is applied to the solution values for vorticity and streamfunction at each point of the original mesh.

4) the maximum time step that is compatible with numerical stability is used as discussed below.

Stability is guaranteed by the condition (3.3.4). For that condition to be satisfied, the coefficients a_k in the matrix $C_n(\Delta t_n)$ of §3.3 must all be non-negative. Assuming this to be true, by a similar analysis for the graded mesh case to that for the even mesh case in §3.3, condition (3.3.4) may be proved to be identically satisfied. Thus the non-negativity conditions on the coefficients a_k provide the constraints.

For the scheme (3.2.6) using graded meshes where $h_i = x_{i+1} - x_i$ and $k_j = y_{j+1} - y_j$ for appropriate i and j, the coefficients a_k as defined in equation (3.3.1) are set out below.

$$a_1 = \Delta t_n \left\{ \frac{1}{Re} \frac{2}{h_i(h_i+h_{i-1})} - \frac{h_{i-1}\, u_{i+1,j}}{h_i(h_i+h_{i-1})} \right\} \tag{5.1.1a}$$

$$a_2 = \Delta t_n \left\{ \frac{1}{Re} \frac{2}{h_{i-1}(h_i+h_{i-1})} - \frac{h_i\, u_{i-1,j}}{h_{i-1}(h_i+h_{i-1})} \right\} \tag{5.1.1b}$$

$$a_3 = 1 - \frac{\Delta t_n}{Re} \left\{ \frac{2}{h_i h_{i-1}} + \frac{2}{k_j k_{j-1}} \right\}$$

$$- \Delta t_n \left\{ \frac{h_i - h_{i-1}}{h_i\, h_{i-1}} u_{i,j} + \frac{k_j - k_{j-1}}{k_j\, k_{j-1}} v_{i,j} \right\} \tag{5.1.1c}$$

$$a_4 = \Delta t_n \left\{ \frac{1}{Re} \frac{2}{k_j(k_j+k_{j-1})} - \frac{k_{j-1}}{k_j} \frac{v_{i,j+1}}{(k_j+k_{j-1})} \right\} \tag{5.1.1d}$$

$$a_5 = \Delta t_n \left\{ \frac{1}{Re} \frac{2}{k_{j-1}(k_j+k_{j-1})} - \frac{k_j\, v_{i,j-1}}{k_{j-1}(k_j+k_{j-1})} \right\} \tag{5.1.1e}$$

Note that for the case where the convective form of the vorticity equation is used, all the velocity components in the above equations would be centered at the (i,j) mesh point.

The conditions $a_1 \geq 0$ and $a_2 \geq 0$ result in restrictions on the size of the x mesh widths. From (5.1.1a), the condition $a_1 \geq 0$ for the point (i,j) and from (5.1.1b) the condition $a_2 \geq 0$ for the point (i-1,j) give the restriction

$$h_{i-1} \leq \frac{2}{Re} \cdot \frac{1}{\max \{|u_{i+1,j}|, |u_{i-2,j}|\}}$$

This restriction is seen to be a localized version of the even mesh global restriction (3.3.5). Similar results hold in the y direction from the conditions $a_4 \geq 0$ and $a_5 \geq 0$.

The condition $a_3 \geq 0$ results in a restriction on the time step Δt_n. The largest time step compatible with stability can be calculated from the condition $a_3 \geq 0$. If we write (5.1.1c) as

$$a_3 = 1 - \Delta t_n \left\{ \frac{b_1}{Re} + b_2 \right\}$$

then from the condition $a_3 \geq 0$ we have that

$$\Delta t_n \leq \frac{1}{\max\limits_{i,j} \left\{ \frac{b_1}{Re} + b_2 \right\}} \qquad (5.1.2)$$

where the $\max\limits_{i,j}$ is taken overall grid points (i,j) in the mesh. The actual time step used is 0.95 that of the maximum allowed by (5.1.2)

The convergence criteria used is the same as for the regular mesh case in §3.4.

§5.2 CHOICE OF GRADED MESH

As emphasised in §4.8 and §5.1, for second order linear differential equations with the second derivative multiplied by a small constant ε, no method is known for choosing a mesh transformation function that is derived directly from the equation to be solved. Analytic mesh transformations are sought that place the majority of mesh points in regions of rapid variation that are approximately known from other sources.

Batchelor [3] has described the asymptotic behaviour of the solution for large Reynolds numbers (see §1.3). Note that the vorticity is relatively constant in the central region of the cavity. This behaviour has been supported by many of the previous investigations (see §1.4), indicating that relatively few mesh points are needed in the central region. The boundary layers near the walls of the cavity contain most of the variation in the vorticity solution with the corollary that most of the mesh points should be placed near the walls.

A simple analytic mesh transformation is sought and we use a symmetric transformation from the class defined by the equation

$$\frac{dx}{dt} = \gamma \left((x+\varepsilon)\ (1+\varepsilon - x)\right)^{\beta} \qquad (5.2.1)$$

where ε is a small positive parameter and the parameter β is positive, subject to the conditions

$$x(0) = 0 \quad \text{and} \quad x(1) = 1 \qquad (5.2.2)$$

The choice of parameter $\varepsilon = 0.1$ has been used. For smaller values of ε, it was found that any mesh transformation was too severe to be practical yet satisfy stability requirements in the central region of the cavity.

The mesh points used are the same in each direction, though a case may be made for biasing the distribution of y mesh points towards the $y = 1$ end of the cavity, in the interests of simplicity this was not done as there is no concrete procedure to suggest the amount of bias. Two example meshes were tried for a Reynolds number of 50. The transformations are

$$\text{A)} \quad \frac{dx}{dt} = \gamma_1((x+\varepsilon)\ (1+\varepsilon - x))^{\frac{1}{2}}$$

and

$$\text{B)} \quad \frac{dx}{dt} = \gamma_2\ ((x+\varepsilon)\ (1+\varepsilon - x))^2$$

Both equations are subject to the conditions (5.2.2).

The mesh transformation (A) gave only a very slight grading and the stability requirements were satisfied by a 11 x 11 mesh. Both cases 11 x 11 and 21 x 21 were computed and h^2 extrapolation was applied to the results. The transformation (B) is significantly more graded and a 16 x 16 mesh was needed to satisfy the stability requirements. A 31 x 31 mesh was also computed and h^2 extrapolation was applied to the two sets of results. Both the convective and divergence forms of the equations was used. The results are discussed more fully in §5.3.

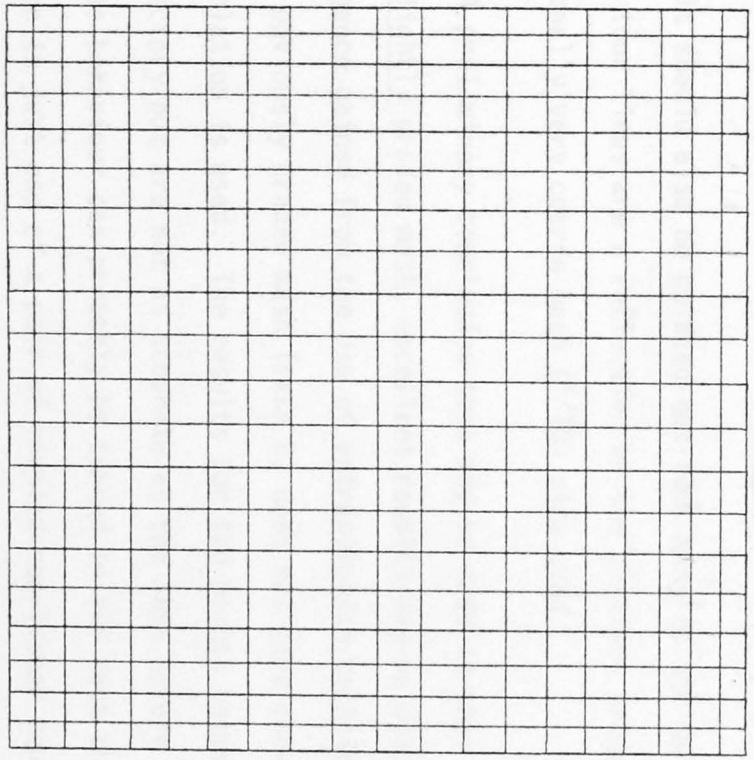
Table 5.2.1 contains the values of the mesh points for 21 points for case A and 31 points for case B.

A qualitative estimate of the amount of grading in the two meshes can be had from figure 5.2.1 which shows plots of the actual grids A and B for the cases 21 x 21 and 31 x 31 respectively.

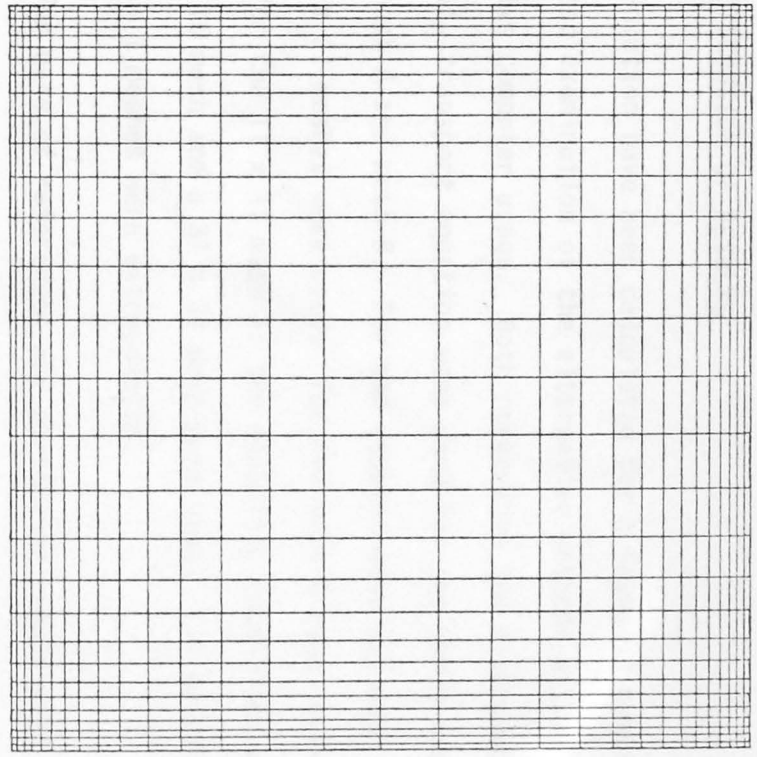
TABLE 5.2.1

MESH POINTS FOR GRADED MESHES

Case A (n = 21)		Case B (n = 31)	
0.0000	0.5000	0.000	0.5776
0.0350	0.5590	0.0079	0.6504
0.0746	0.6174	0.0169	0.7148
0.1183	0.6747	0.0274	0.7696
0.1657	0.7304	0.0396	0.8151
0.2163	0.7837	0.0541	0.8524
0.2696	0.8343	0.0713	0.8830
0.3252	0.8817	0.0920	0.9080
0.3825	0.9254	0.1170	0.9287
0.4410	0.9650	0.1476	0.9459
	1.0000	0.1849	0.9604
		0.2304	0.9726
		0.2852	0.9831
		0.3496	0.9921
		0.4224	1.0000
		0.5000	



Graded Mesh for Case A
(See TABLE 5.2.1 for point values)



Graded Mesh for Case B
(See TABLE 5.2.1 for point values)

FIGURE 5.2.1

§5.3 DISCUSSION OF RESULTS

Solution have been calculated for a Reynolds number of 50. This choice permitted examination of the alternative methods without excessive computer usage. Both convective and divergence forms of the vorticity transport equation were used for the three cases of a regular mesh, mesh A and mesh B. For the regular mesh and mesh A cases, 11 x 11 and 21 x 21 meshes were used: for the mesh B case, convergence did not occur for the 11 x 11 mesh as the stability conditions were violated so a 16 x 16 mesh and a 31 x 31 mesh were used. All solutions for the n + 1 and 2n + 1 meshes were extrapolated.

For ease of reference, any particular result will be referred to by 'mesh type (R, A, B), equation form (D, C), no of points or extrapolated (eg. 11, 21, 16, E)' thus RD21 is the solution for a regular mesh divergence form for a 21 x 21 mesh.

It should also be pointed out that many of the small 'ripples' in the contour lines are a reflection of the plotting program and the occasionally very coarse mesh ($^1/_{10}$) size used.

A preliminary conclusion that may be made is that for a regular mesh or a slightly graded mesh, excellent results may be obtained with good convergence gained from the use of extrapolation techniques. The use of a very obviously graded mesh (case B) does not give convergence when extrapolation is used. The results for the meshes themselves are more satisfactory but are not as accurate as the less severe cases. The reason for this behaviour can probably be traced to the fact that the vorticity equation is just one of a pair of coupled non-linear equations and the simple theory that can be applied to linear second order one dimensional equations does not carry over well to the more complicated case.

Another conclusion that may be drawn from these results is that graded meshes in such problems must be approached with extreme caution and much more work needs to be done in this area. On the other hand, for well behaved cases (R and A meshes) the use of extrapolation leads to good convergence.

Regular mesh

Though significant differences are found between the solutions for RD11 and RC11 and even overall single digit consistency is not found between the RD21 and RC21 solutions, the two extrapolated solutions RDE and RCE agree to 2 significant digits over almost all of the cavity, differing by more only in the top corners where the effects of the vorticity discontinuity along the walls would be felt most. Contour plots of these examples may be found in figures 3.4.1 ff.

A mesh

The A mesh cases exhibit very similar behaviour to the regular mesh as a group again with two significant digits agreement between the ADE and ACE cases. Also there is very close agreement between the contour lines between RDE and RCE (which are identical to the accuracy of the plots) and ADE and ACE (again identical to the accuracy of the plots) in the central regions. There are slight differences near the tops of the cavities.

The 21 x 21 cases for the A meshes gave good agreement with lower corners for the same case for the R meshes. The extra accuracy gained by the use of a slightly graded mesh is reflected in the appearance of a small, weak eddy in the lower left corners of the cavities for the A mesh when the solutions are extrapolated. The relevant contour plots are in figure 5.3.1 ff.

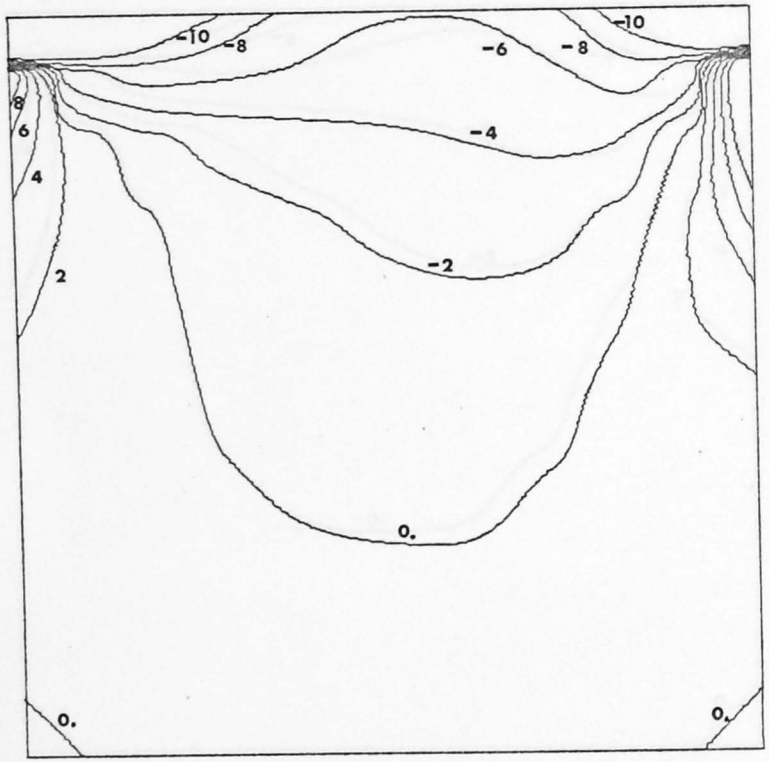
Note that the solutions for the R and A cases agree well with those already in the literature. See Burggraf [8] and Marshall and van Spiegel [28] for example.

B mesh

Only the BD16 and BC16 cases agree even qualitatively with the previous results. Convergence is not seen to be obtained. The contour plots are to be found in figure 5.3.4 ff.

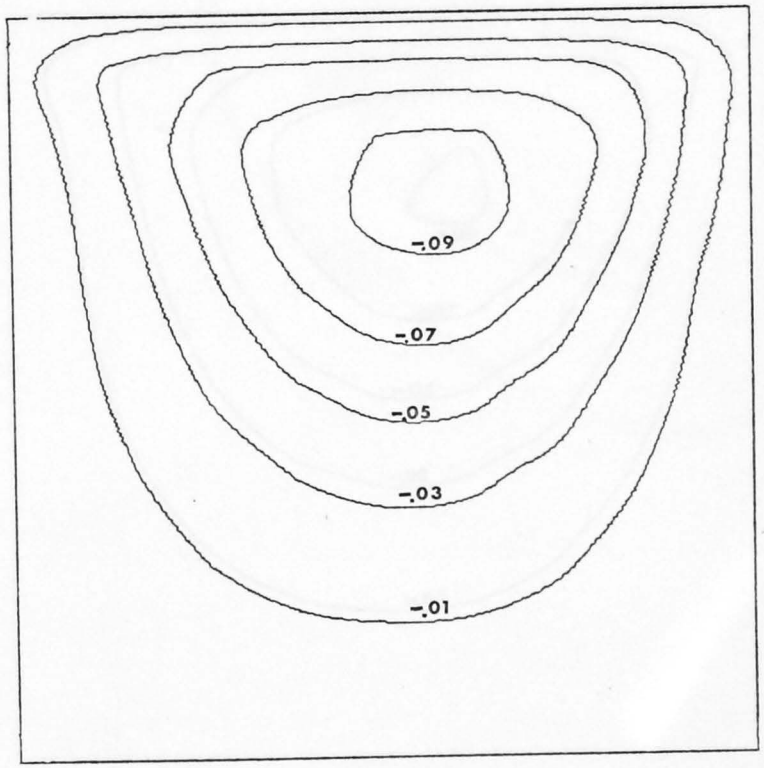
The severe grading of the B mesh, a 10:1 ratio between largest and smallest mesh width, places many parts of the B mesh near the boundaries. This results in good results at the very bottom of the cavity in the modelling of the secondary eddies. This region is far from the effects of the discontinuities in the vorticity at the two upper corners caused by the sliding wall. Because of the severe bunching of points near this region of the B mesh causes an excessive influence for these singularities to be propagated through the mesh. Caution is obviously required when only qualitative methods can be used to choose the graded mesh.

Note that the problem of a too severely graded mesh has occurred in §4.4 concerning the mesh transformation for a first order formulation of two-point boundary value problems.



Vorticity

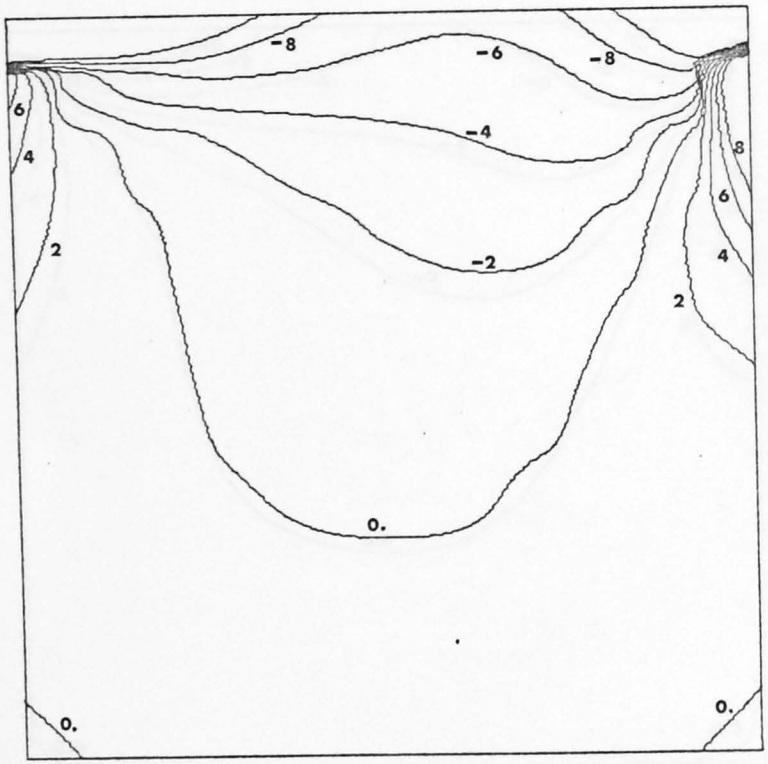
Figure 5.3.1a



Streamfunction

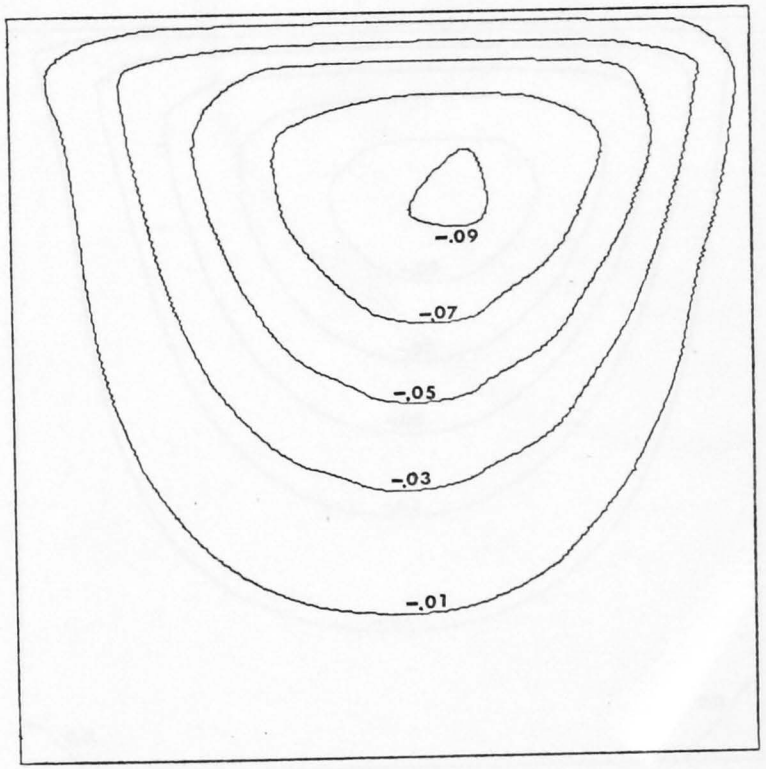
Figure 5.3.1b

Example AD11



Vorticity

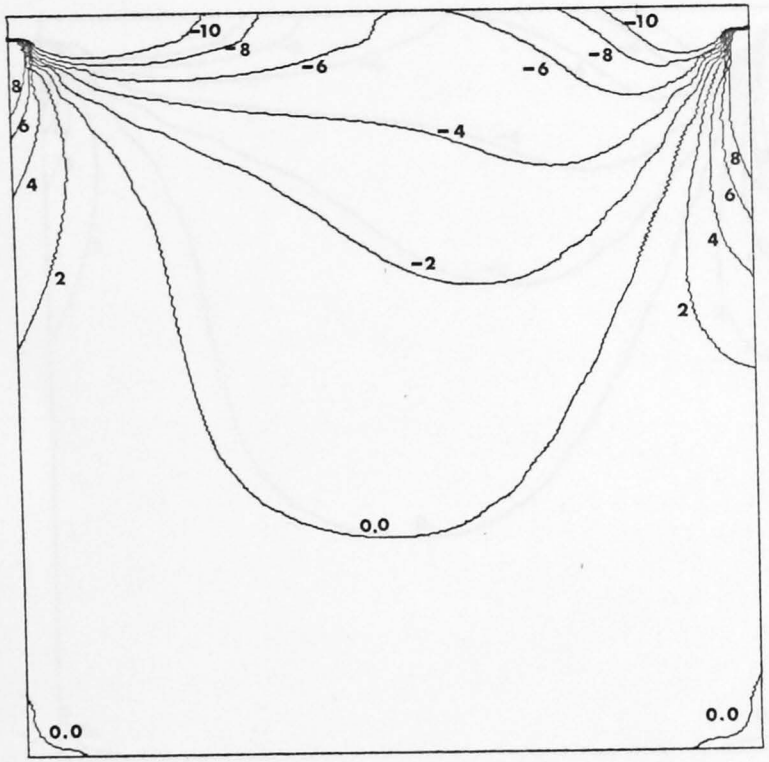
Figure 5.3.1c



Streamfunction

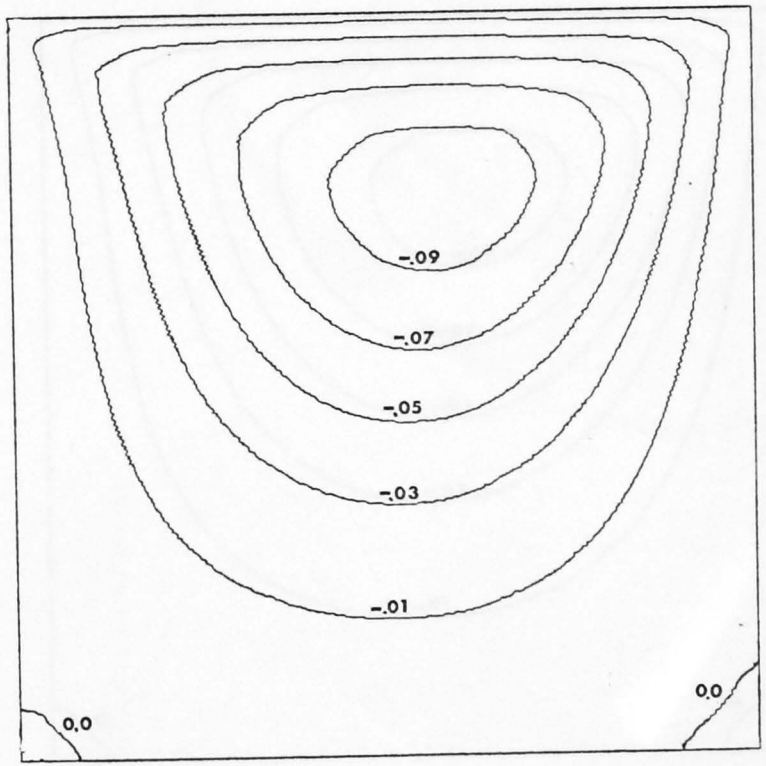
Figure 5.3.1d

Example AC11



Vorticity

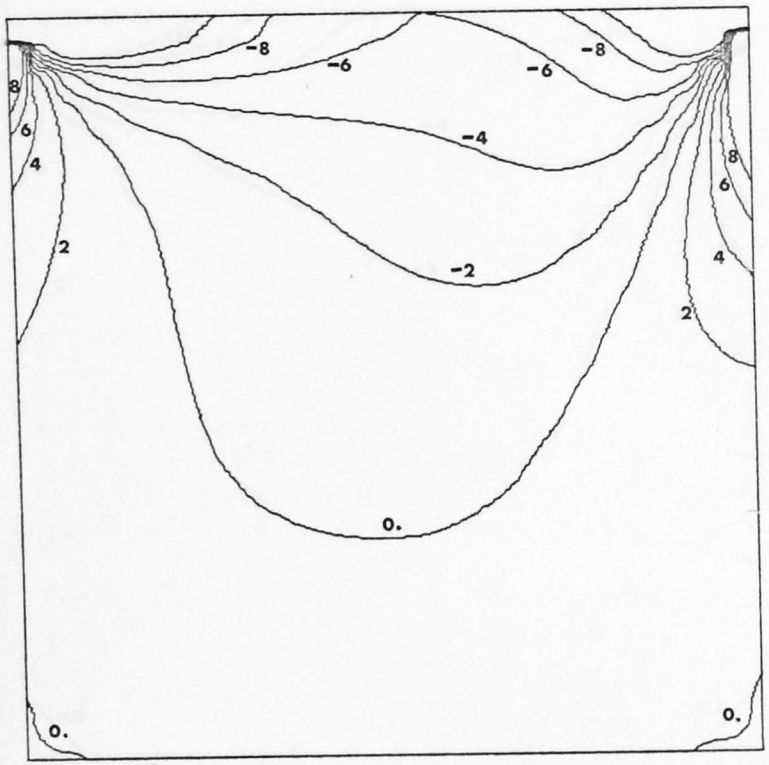
Figure 5.3.2a



Streamfunction

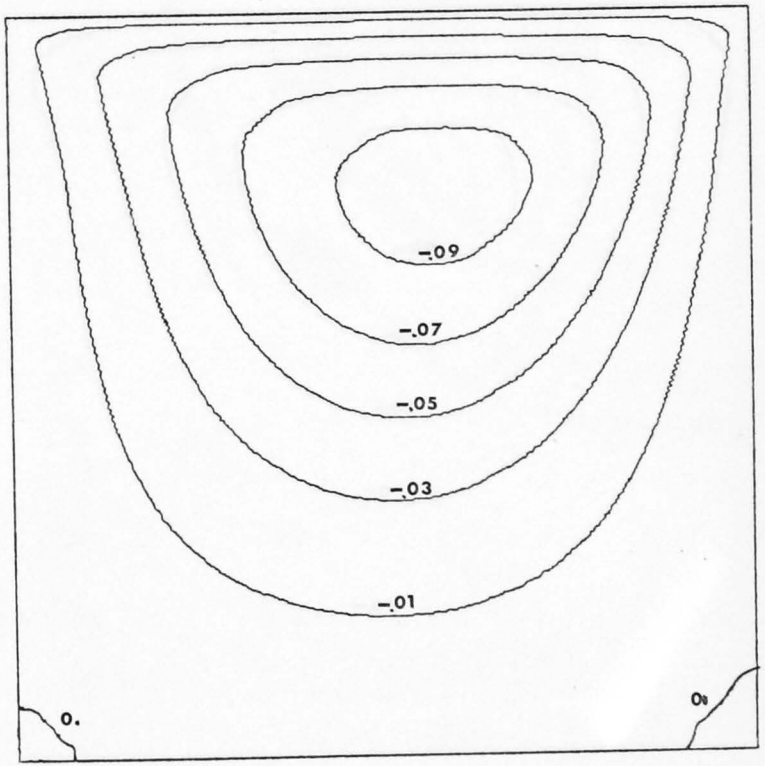
Figure 5.3.2b

Example AD21



Vorticity

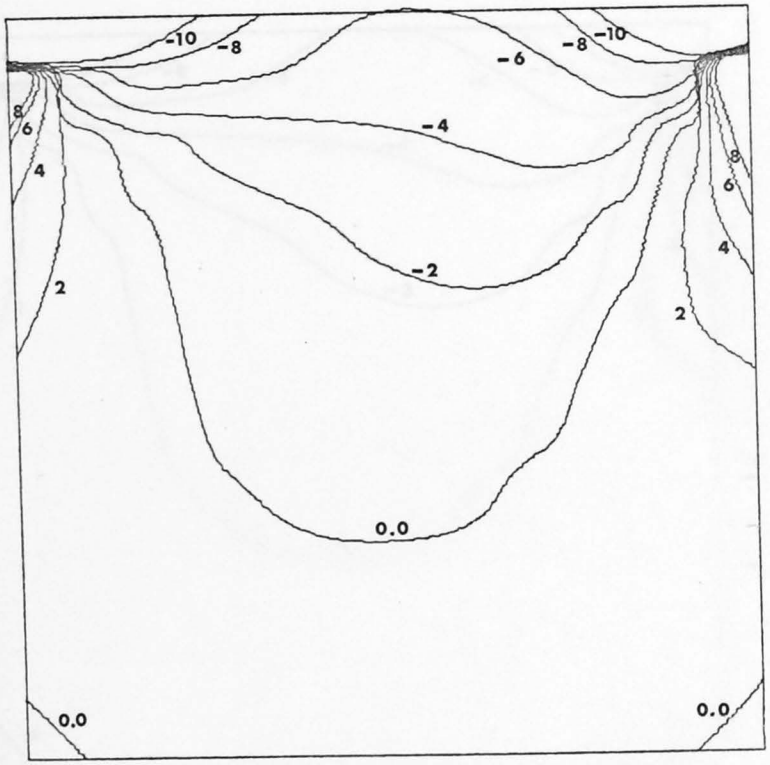
Figure 5.3.2c



Streamfunction

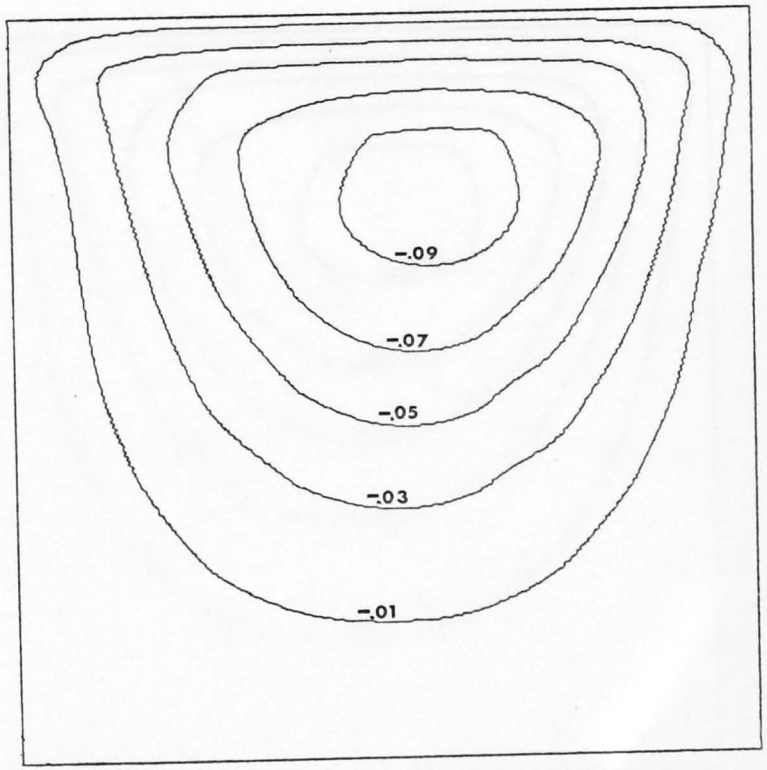
Figure 5.3.2d

Example AC21



Vorticity

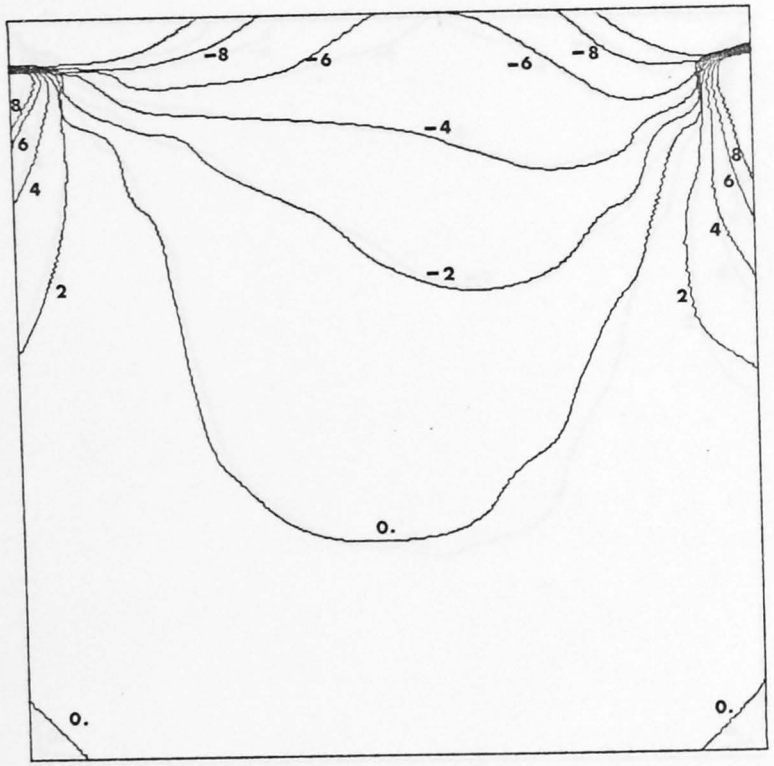
Figure 5.3.3a



Streamfunction

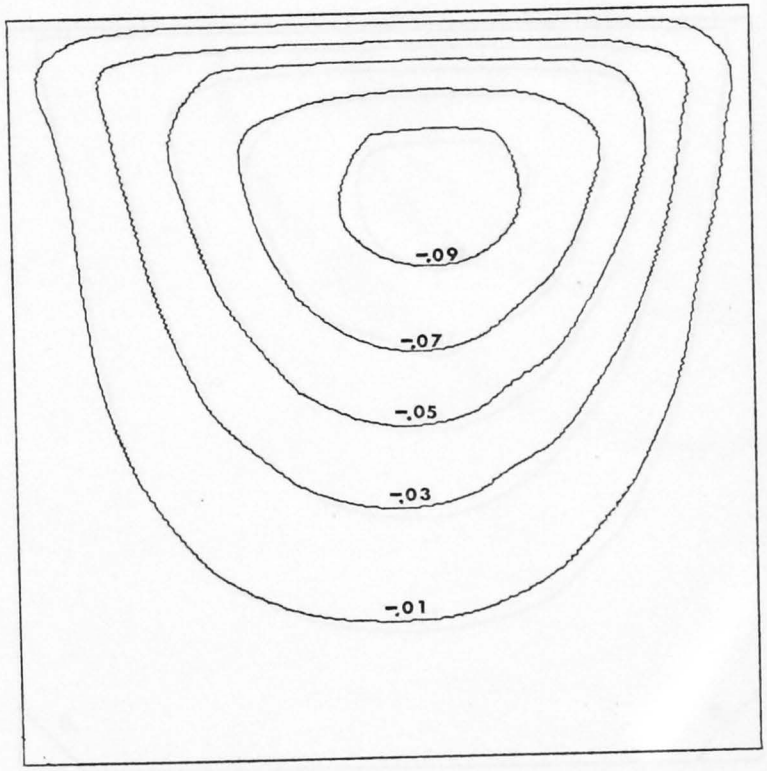
Figure 5.3.3b

Example ADE



Vorticity

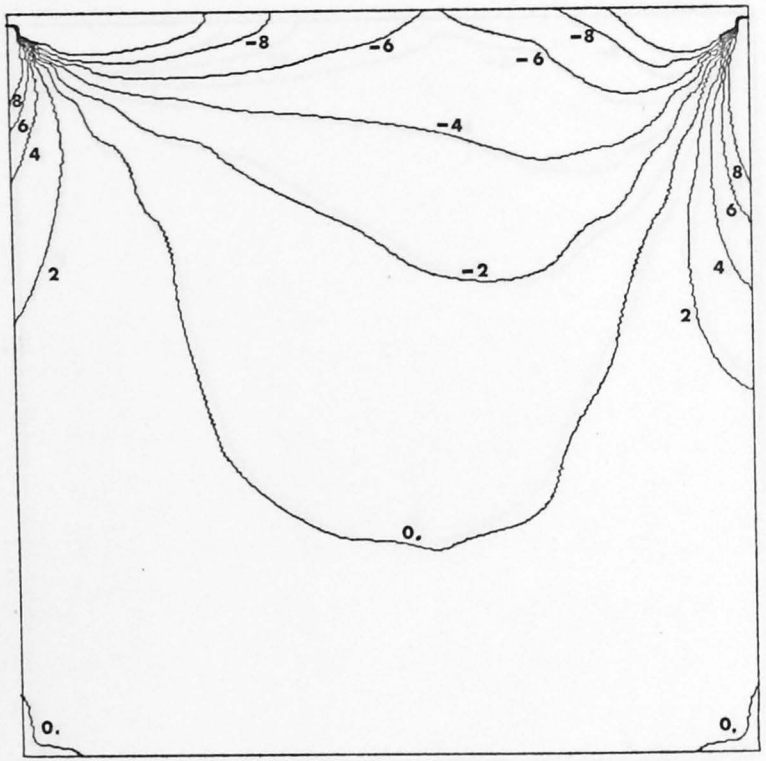
Figure 5.3.3c



Streamfunction

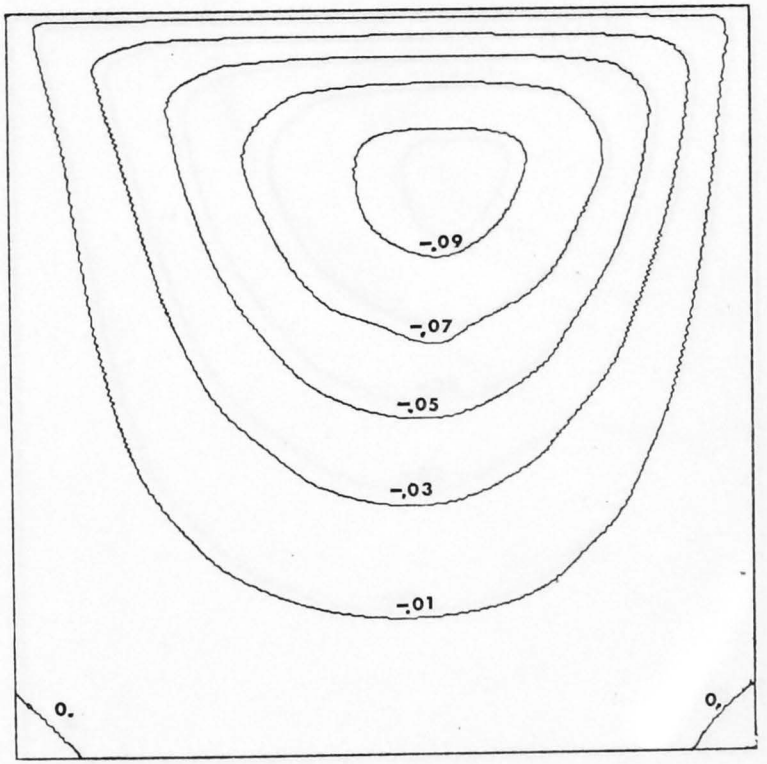
Figure 5.3.3d

Example ACE



Vorticity

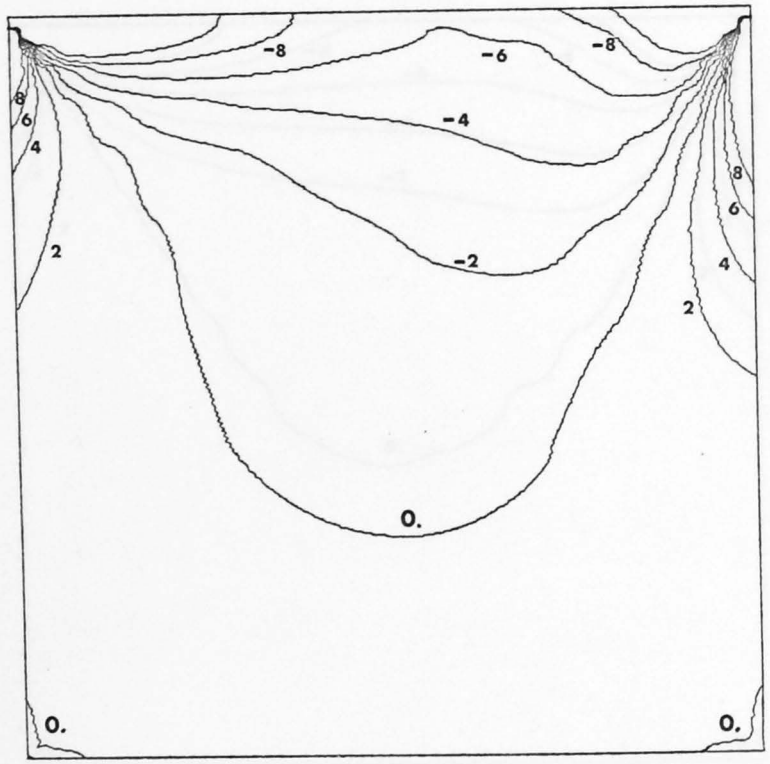
Figure 5.3.4a



Streamfunction

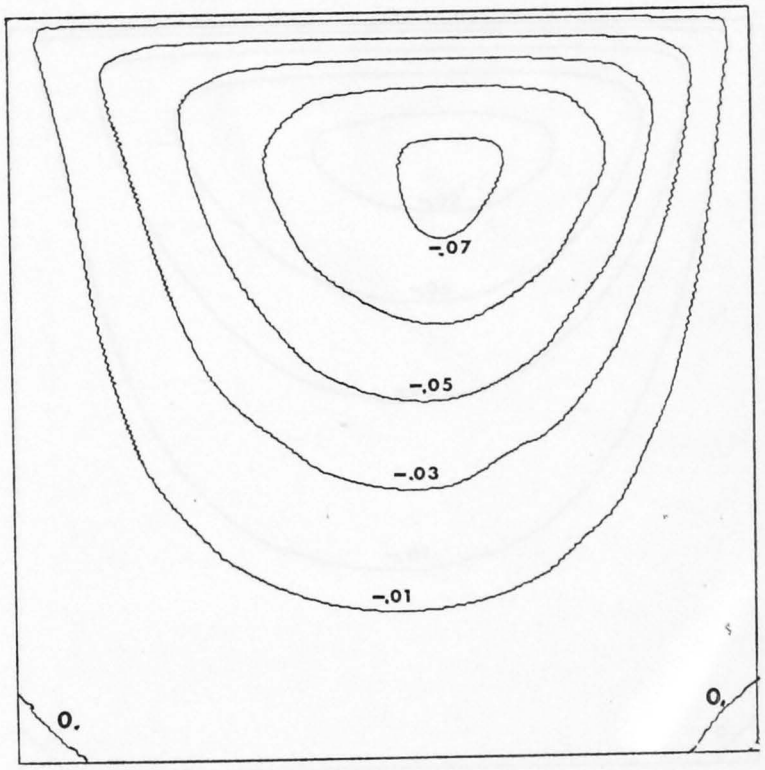
Figure 5.3.4b

Example BD21



Vorticity

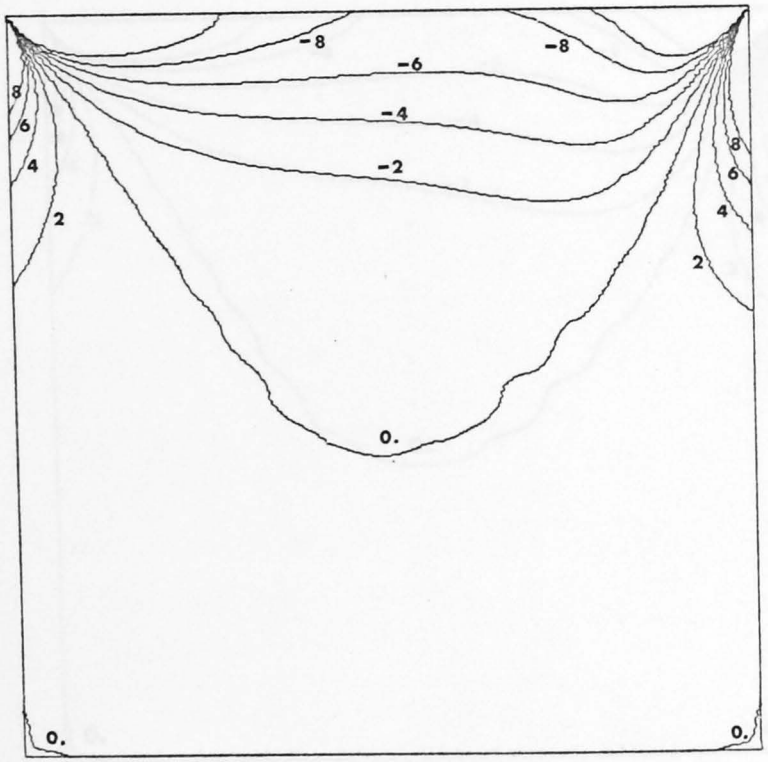
Figure 5.3.4c



Streamfunction

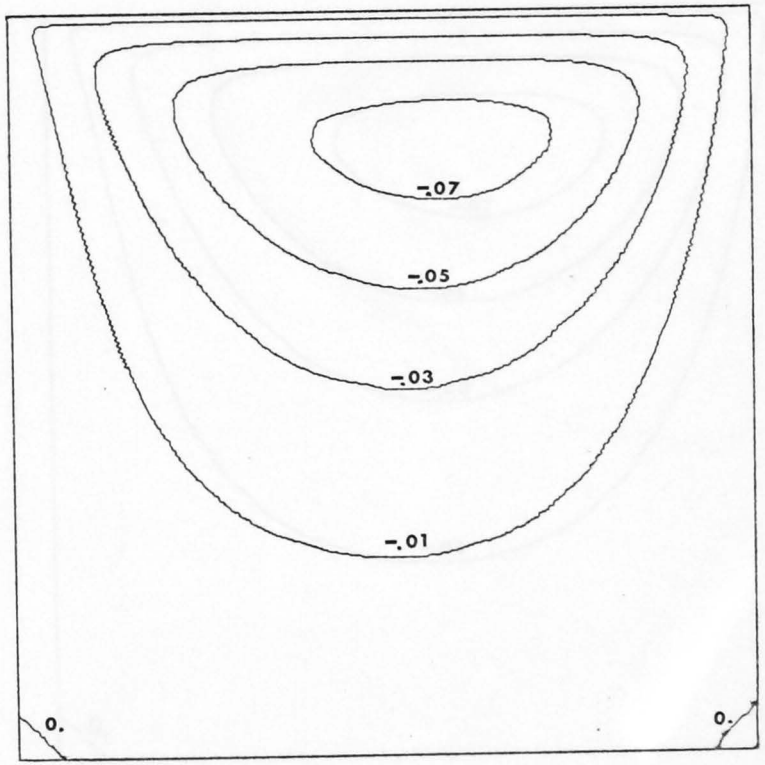
Figure 5.3.4d

Example BC16



Vorticity

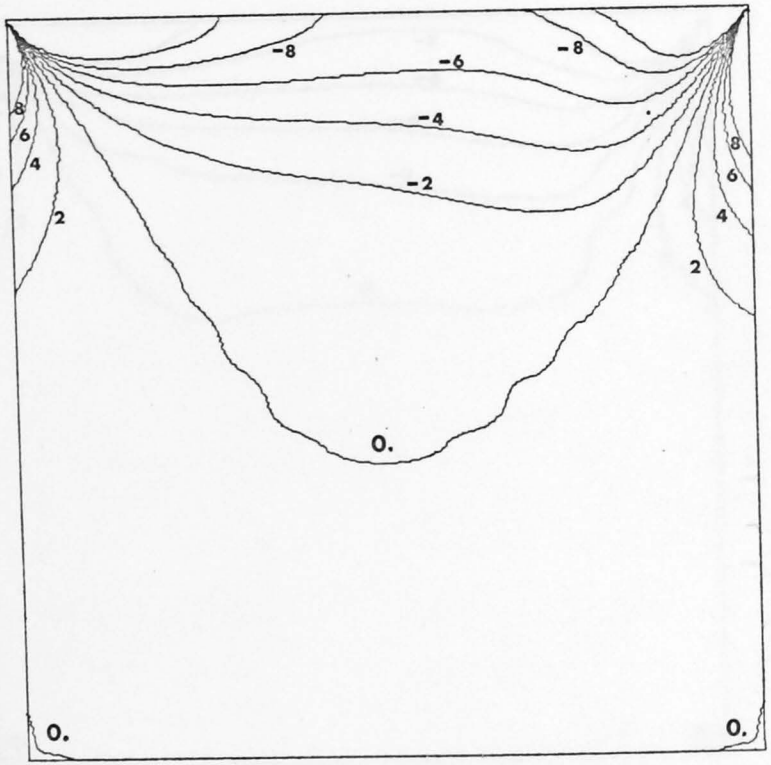
Figure 5.3.5a



Streamfunction

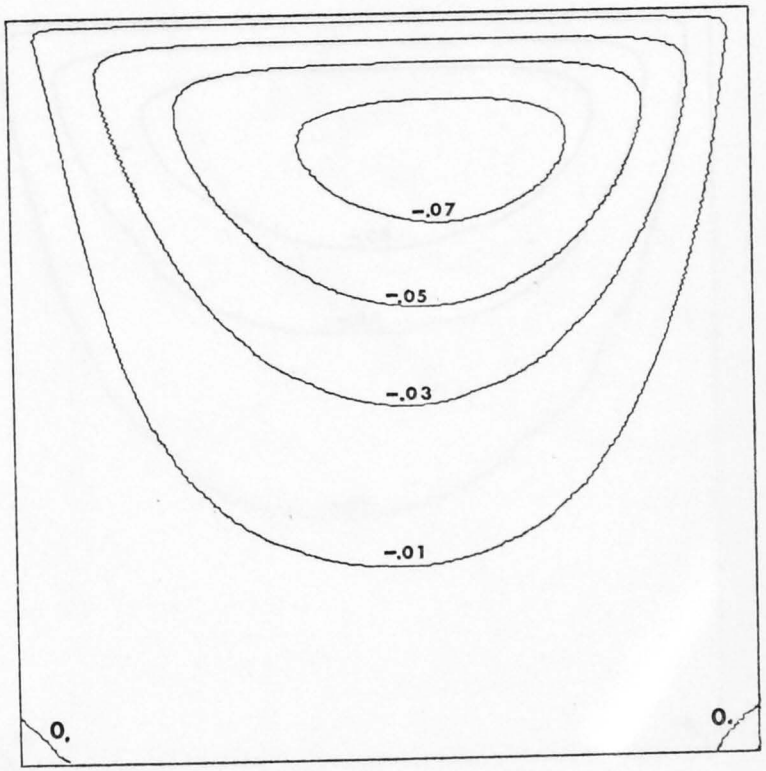
Figure 5.3.5b

Example BD31



Vorticity

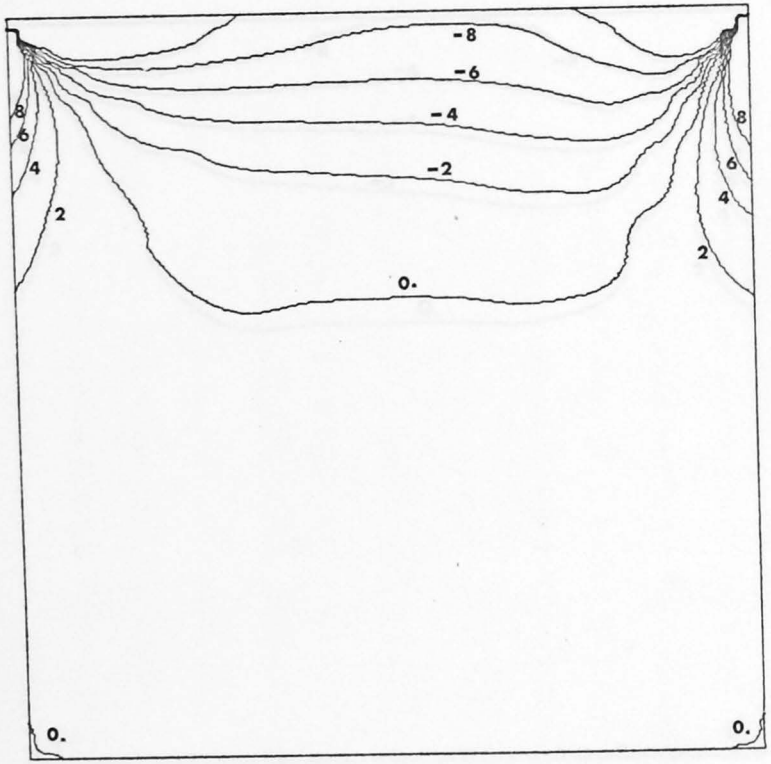
Figure 5.3.5c



Streamfunction

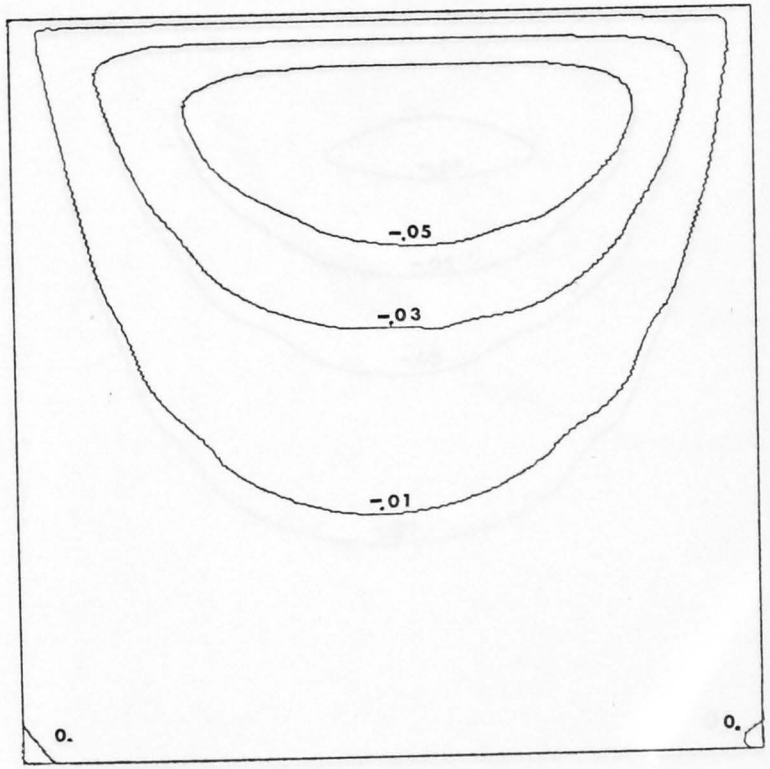
Figure 5.3.5d

Example BC31



Vorticity

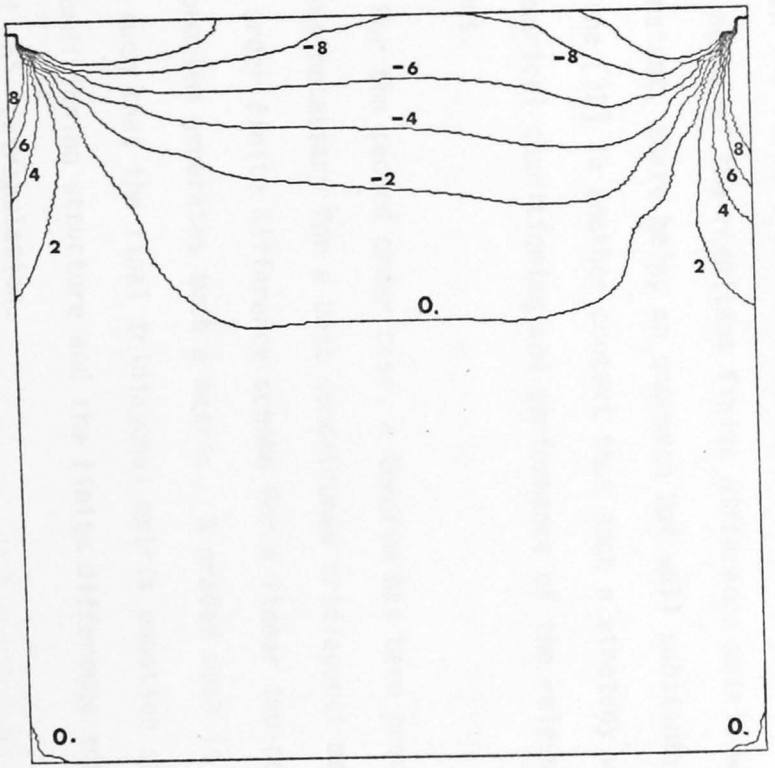
Figure 5.3.6a



Streamfunction

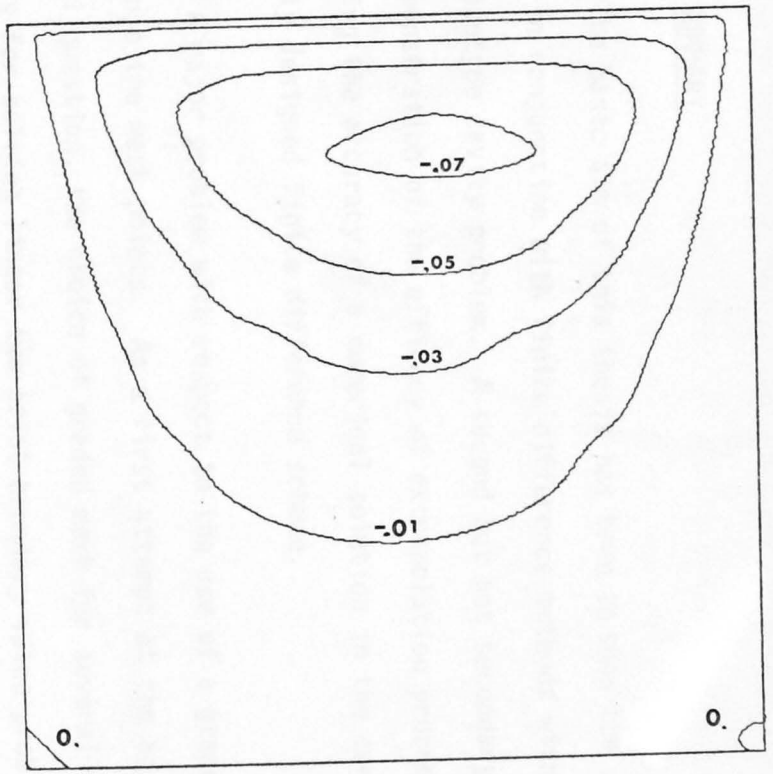
Figure 5.3.6b

Example BDE



Vorticity

Figure 5.3.6c



Streamfunction

Figure 5.3.6d

Example BCE

CHAPTER 6

§6.1 SUMMARY

The basic aim of this thesis has been to show the use of graded meshes in conjunction with finite difference methods with reference to the prototype cavity problem. A second but not secondary aim has been the demonstration of the efficacy of extrapolation processes for improving the accuracy of a numerical solution in the context of a suitably designed finite difference scheme.

A major problem with respect to the use of a graded mesh is how to choose the mesh points. As a first attempt at the answer to this general question, the choice of graded mesh for several finite difference methods for solving linear two-point boundary value problems was examined in Chapter 4. Two formulations of the problem were discussed.

In the first order case the mesh was chosen so that the block row sum norms of the resultant finite difference matrix were equilibrated to a constant, there being an unproven but well substantiated proposition of Osborne [32] in another context that such a strategy would lead to good numerical conditioning and performance of the relevant matrix equations.

For the second order case, a theorem has been proved re the structure necessary for a best conditioned tridiagonal matrix. The second order finite difference scheme for a linear two-point boundary value problem generates such a matrix. A graded mesh is able to be chosen such that the final tridiagonal matrix equation satisfied the best conditioning structure and the finite difference scheme allowed h^2 and h^4 extrapolation.

Excellent results were had for both basic formulations of the problem with good improvements in the accuracy on the application of the extrapolation techniques. The chosen graded meshes are shown to be clearly superior to even meshes with the same number of points and to be at or near the optimum for a certain whole class of graded meshes. That such results can be obtained by proper choice of graded mesh is emphasised by an example using graded meshes for which the mesh points are chosen randomly from a uniform distribution on the unit interval. Very importantly, note that random meshes do not permit any extrapolation process.

The importance of h^2 and h^4 extrapolation and the specific design of all the above-mentioned schemes to allow for extrapolation cannot be over-emphasised.

The non-linear vorticity transport equation can be cast into either the divergence or the convective form. Both forms are studied here. In this thesis, all the terms in the Navier-Stokes equation are approximated by central three point formulations with the weights at the three points chosen to minimise the discretisation error while maintaining consistency. Directional formulas for the first derivative terms were not considered though there are recent indications ([2], [5], [16]) that such formulations may bring some improvements in the solution of this problem. Only the case for a Reynolds number of 50 was considered. This case has many of the features of creeping flow (Re = 50) including small eddies in each lower corner but the expected asymmetry in the vorticity field as the Reynolds number increases is beginning to show through.

An even grid is considered first (Ch. 3) while two examples of graded meshes are considered in Chapter 5. One mesh (mesh A) is only slightly graded; the other (mesh B) is significantly so. The results from all cases were extrapolated.

The h^2-extrapolated results for the regular mesh and the mesh A cases for both forms of the vorticity equation all agree to within two significant digits over almost all of the cavity, differing only by more along the very top of the cavity where the singularity of the sliding wall can be expected to have different influences on different schemes. A small vorticity in the lower left corner was picked up immediately on a 11 x 11 grid by the mesh A a significant difference from the regular mesh case. The mesh B cases being 16 x 16 and 31 x 31 grids and more severely graded than the other cases had many more points closer to the upper singularities. These singularities caused a large deviation from the previous solution especially when extrapolated. All B meshes reproduced well the behaviour of the solution in the bottom of the cavity well away from the upper discontinuities.

The point must be made that care must be taken when using graded meshes where there are singularities in the solution and where a suitable quantitative method for choosing the grading of the mesh is not known. It is obvious from the above-mentioned results that more work is needed to be done to gain a complete understanding of the problems involved and of their solution.

REFERENCES

[1] Allen, R.C. and Wing, G.M. An invariant imbedding algorithm for the solution of inhomogeneous linear two-point boundary value problems. *J. Comp. Phys.* 14, 40-58 (1974)

[2] Barrett, K.E. The numerical solution of singular perturbation boundary-value problems. *Q.J. Mech. Appl. Math.* XXVII, 57-68 (1974)

[3] Batchelor, G.K. On steady laminar flow with closed streamlines at large Reynolds numbers. *J. Fluid Mech.* 1, 177-190 (1956)

[4] Bickley, W.G. and McNamee, J. Matrix and other direct methods for the solution of systems of linear differential equations. *Phil. Trans. Roy. Soc. Ser. A.* 252, 69-131 (1960)

[5] Bozeman, J.D. and Dalton, C. Numerical study of viscous flow in a cavity. *J. Comp. Phys.* 12, 348-363 (1973)

[6] Brent, R.P. Error analyses of algorithms for matrix multiplication and triangular decomposition using Winograd's identity. *Numer. Math.* 16, 145-156 (1970)

[7] Brent, R.P. Algorithms for matrix multiplication. Tech. Report CS157 (March 1970), Computer Sci. Dept., Stanford Uni.

[8] Burggraf, O.R. Analytical and numerical studies of the structure of steady separated flows. *J. Fluid Mech.* 24, 113-151 (1966)

[9] Chorin, A.J. On the convergence of discrete approximations to the Navier-Stokes equations. *Maths. of Comp.* 23, 745-762 (1969)

[10] Chorin, A.J. Numerical study of slightly viscous flow. To appear.

[11] Cooley, J.W., Lewis, P.A.W. and Welch, P.D. The Fast Fourier Transform algorithm: Programming considerations in the calculation of sine, cosine and Laplace transformations. *J. Sound and Vibration* 12, 315-337 (1970)

[12] Cooley, J.W. and Tukey, J.W. An algorithm for the machine computation of complex Fourier series. *Math. Comp.* 19, 297-301 (1965)

[13] Dean, W.R. and Montagnon, P.E. On the steady motion of viscous liquid in a corner. *Proc. Camb. Phil. Soc.* 45, 389-394 (1949)

[14] Donovan, L.F. Numerical solution of unsteady flow in a rectangular cavity with a moving wall. NASA Technical Memorandum TMX-52767 (1970)

[15] Dorr, F.W. The direct solution of the discrete Poisson equation on a rectangle. *SIAM Review* 12, 248-263 (1970)

[16] Dorr, F.W. An example of ill-conditioning in the numerical solution of singular perturbation problems. *Maths. of Comp.* 25, 271-283 (1971)

[17] Forsythe, G.E. and Straus, E.G. On best conditioned matrices. *Proc. Amer. Math. Soc.* 6, 340-345 (1955)

[18] Fox, L. "The Numerical Solution of Two-Point Boundary Value Problems in Ordinary Differential Equations," O.U.P., 1957

[19] Fromm, J.E. Numerical method for computing non-linear time dependent buoyant circulation of air in rooms. *IBM J. Res. & Develop.* 15, 185-196 (1971)

[20] Godaux, F. Ecoulement et transfert de chaleur dans une cavité rectangulaire dont une paroi est mobile - Etude numerique. *Acad. Roy. Belg. Bull. Cl. Sci., V Sér.*, 57, 559-575 (1971)

[21] Golub, G.H. Private communication (1973)

[22] Greenspan, D. Numerical studies of prototype cavity flow problems. *Comp. J.* 12, 89-94 (1969)

[23] Joyce, D.C. Survey of extrapolation processes in Numerical Analysis. *SIAM Review* 13, 435-490 (1971)

[24] Kawaguti, M. Numerical solution of the Navier-Stokes equations for the flow in a two-dimensional cavity. *J. Phys. Soc. Japan* 16, 2307-2318 (1961)

[25] Keller, H.B. "Numerical Methods for Two-Point-Boundary Value Problems". Ginn-Blaisdell, 1968

[26] Luenberger, D.G. "Optimisation by Vector Space Methods". Wiley, 1969

[27] Lynch, R.E., Rice, J.R. and Thomas, D.H. Direct solution of partial differential equations by tensor product methods. *Numer. Math.* 6, 185-199 (1964)

[28] Marshall, G. and Van Spiegel, E. On the numerical treatment of the Navier-Stokes equation for an incompressible fluid. *J. Engin. Math.* 7, 173-188 (1973)

[29] Mills, R.D. Numerical solution of the viscous flow equations for a class of closed flows. *J. Roy. Aeron. Soc.* 69, 714-718 (1965)

[30] Moffatt, H.K. Viscous and resistive eddies near a sharp corner. *J. Fluid Mech.* 18, 1-18 (1964)

[31] Osborne, M.R. On shooting methods for boundary value problems *J. Math. Anol. Appl.* 27, 417-433 (1969)

[32] Osborne, M.R. On the numerical solution of boundary value problems for ordinary differential equations. *Proc. IFIP Conference,* 1974

[33] Pan, F. and Acrivos, A. Steady flows in rectangular cavities *J. Fluid Mech.* 28, 634-655 (1967)

[34] Pearson, C.E. On a differential equation of boundary layer type. *J. Math. & Phys.* 47, 134-154 (1968)

[35] Pruess, S. Solving linear boundary value problems by approximating the coefficients. *Math. Comp.* 27, 551-561 (1973)

[36] Richtmeyer, R.D. and Morton, K.W. "Difference Methods for Initial-Value Problems". 2nd Ed. Interscience, 1967.

[37] Roberts, S.M. and Shipman, J.S. Two Point Boundary Value Problems: Shooting Methods" Elsevier, 1972.

[38] Runchal, A.K., Spalding, D.B. and Wolfshtein, M. Numerical solution of the elliptic equations for the transport

of vorticity, heat and matter in two-dimensional flow. *Phys. of Fluids* 12, 11-21 - 11-28 (1969)

[39] Stetter, H.J. "Analysis of Discretisation Methods for Ordinary Differential Equations". Springer-Verlag, 1973

[40] Stone, H.L. Iterative solution of implicit approximations of multidimensional partial differential equations. *SIAM J. Numer. Aral.* 5, 530-558 (1968)

[41] Strassen, V. Gaussian elimination is not optimal. *Numer. Math.* 13, 354-356 (1969)

[42] Torrance, K.E. Comparison of finite-difference computations of natural convection. *J. Res. Nat'l. Bur. Standards* 72B, 281-301 (1968)

[43] Torrance, K., Davis, R., Eike, K., Gill, P., Gutman, D., Hsui, A., Lyons, S. and Zien, H. Cavity flows driven by buoyancy and shear. *J. Fluid Mech.* 51, 221-231 (1972)

[44] Varah, J. and Golub, G.H. On the best L^2 scaling for matrices. Lecture Notes, Fifth Gatlinburg Symposium on Numerical Algebra, 1972

[45] Winograd, S. A new algorithm for inner products. *IEEE Trans.* C-17 (1968), 693-694

[46] Woodford, G. Fast Direct Methods for the solution of separable elliptic equations on rectangles in 'Computational Methods in Mathematical Physics', edited by R.S. Anderssen and R.O. Watts, University of Queensland Press, Brisbane, 1974.