NUMERICAL SOLUTIONS

TO SOME INDUSTRIAL PROBLEMS

A thesis submitted for the degree

of Doctor of Philosophy



Rosemary Ann Cartwright
Balliol College
Oxford

September 1984

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr D F Mayers, for his help and guidance throughout my time of research and thesis preparation.

I am grateful to Balliol College for providing a grant towards the cost of producing this thesis and to the Science and Engineering Research Council for their financial support during the last three years.

Thanks are also due to Wendy Norbury for typing this thesis.

Finally, I would like to express my gratitude to my family for their interest and encouragement.

Rosemary Cartwright

Numerical solutions to some industrial problems

Rosemary Ann Cartwright
Balliol College
Oxford

A thesis submitted for the degree of Doctor of Philosophy

Michaelmas Term 1984

Abstract

The particular type of industrial problem discussed is that which reduces to a hyperbolic partial differential equation with singularities.

Two main problems are considered. The first is that of finding the position of a neutrally buoyant cable being towed behind a ship which is executing horizontal transverse motions. Both the method of characteristics and a finite difference method are used to find a solution and both give satisfactory results. The algorithms are explained and the results compared.

The second problem is concerned with the occurrence of thermal runaway when a solid medium is penetrated by a solid or immiscible liquid foreign body. Several versions, ranging in difficulty, of the problem are studied both numerically and analytically.

There is also a study of the error in the solution when certain simple problems involving singularities are solved by finite difference methods. An expression is obtained for the generating function of the finite difference solution. The required coefficient in the generating function is expressed as a contour integral, for which an asymptotic form is obtained by the saddle point method.

CONTENTS

CHAPTER 1 - INTRODUCTION

1.1 Introduction 1

1.2 Notation 2

1.3 Hyperbolic equations and characteristics 3

1.4 Finite differences 6

1.5 The characteristic method 18

1.6 A comparison of finite difference and characteristic methods 28

CHAPTER 2 - THE SADDLE POINT METHOD AND ERRORS IN FINITE DIFFERENCE SCHEMES

2.1 Introduction 29

2.2 The saddle point method 30

2.3 The test problem 34

2.4 Results and observations 44

2.5 Other schemes applied to the test problem 50

2.6 A second order version of the test problem and a scheme to solve it 52

2.7 The second order Lax Wendroff scheme 59

2.8 A discontinuous problem 64

2.9 Summary 70

CHAPTER 3 - THE CABLE PROBLEM

3.1 Description of the cable problem 88

3.2 Numerical solution using characteristics 95

3.3 Numerical solution using finite differences 102

3.4 Results and a comparison of the finite difference and characteristic methods 106

3.5 Minor adaptations and their effects 110

3.6 Concluding comments 113

CHAPTER 4 - A SIMPLIFIED VERSION OF THE THERMAL RUNAWAY PROBLEM

4.1 Introduction 129

4.2 The exact solution 130

4.3 A numerical algorithm to solve the equation 135

4.4 Results and conclusions 142

4.5 An extension to the problem 148

CHAPTER 5 - GEODYNAMIC THERMAL RUNAWAY

5.1 Introduction 162

5.2 Thermal runaway 162

5.3 A model for geodynamic thermal runaway 167

5.4 The steady solution 174

5.5 The scaling of the equations 178

5.6 Links between the thermal runaway problem and Chapter 4 179

5.7 Perturbations of the problem 181

5.8 A numerical algorithm 187

5.9 Results and conclusions 197

REFERENCES 215

CHAPTER 1

Introduction

§1.1 Introduction

This thesis is concerned with the study of hyperbolic partial differential equations which involve singularities.

Some physical situations which lead to such equations and their distinguishing features are given in Section 1.3.

Sections 1.4 and 1.5 then explain two numerical methods for solving such equations. The first is the method of finite differences which can, in fact, be used for any type of partial differential equation. Because it is so widely applicable it does not utilise special information which is available in the case of hyperbolic equations. The second method, the method of characteristics,does make use of this special information and is thus only applicable to hyperbolic equations. A comparison of the two methods is given in Section 1.6.

Both the method of characteristics and finite difference methods are used extensively throughout the thesis. The latter half of the work, Chapters 3, 4 and 5, consists of the solution of three different partial differential equations which arise from physical situations. The equation of Chapter 4 is, though, presented in a purely theoretical way. However, it does have links with the problem studied in Chapter 5 which

describes a physical process; namely the movement of an immiscible liquid through a solid medium by means of thermal runaway. Both of these equations are solved using a characteristic method.

The much simpler problem given in Chapter 3 which describes the movement of a neutrally buoyant cable being towed behind a ship is solved by both the method of characteristics and by finite differences and the results are compared.

Chapter 2 addresses a somewhat more theoretical problem. For continuous problems the leading term of the error which occurs as a result of using a finite difference scheme is easily expressed as a multiple of some derivative of the solution. However, in discontinuous situations such derivatives do not exist and so it is not possible to use this method to find the error at a discontinuity. In Chapter 2 a procedure is suggested which could be used in these situations. It is applied to several very simple problems with a range of finite difference schemes

§1.2 Notation

Two shorthand notations for differentiation have been used. The first is the prime notation to denote differentiation of a function with respect to its argument. Thus $u'(t)$ means $\frac{du}{dt}$.

The second is the use of subscript notation to denote partial differentiation with respect to the subscripted variable and so, for instance, $u_x \equiv \frac{\partial u}{\partial x}$, $f_{xt} \equiv \frac{\partial^2 f}{\partial x \partial t}$.

It should be noted that subscript notation may also be used for other reasons, for example, evaluation at a point. So in this case $u_A = u(\underline{a})$ where $\underline{a}$ is the position vector of the point A.

The relevant interpretation of the subscript should be clear from the context.

§1.3 Hyperbolic equations and characteristics

The simplest hyperbolic equation is the wave equation

$$u_{xx} = u_{tt} \,. \tag{1.1}$$

The general solution to this equation is

$$u(x,t) = u_1(x+t) + u_2(x-t)$$

where the nature of the functions u_1 and u_2 (which are functions of a single variable) depends on the initial and boundary conditions. u_1 is a function which in some sense propagates information along the line $x + t =$ constant whereas u_2 sends it along the line $x - t =$ constant.

These two families of curves along which information is propagated occur in all hyperbolic problems and are called characteristics. As a result of the propagating quality of characteristics, these lines are also places in the solution across which discontinuities may occur. For instance, if the initial data for a problem is discontinuous then the characteristic running through that initial point will convey the discontinuity to later times. Even if the initial data is continuous then for some problems the characteristics of one family may eventually converge and hence indicate a discontinuity in the solution since such a point in space time receives

'conflicting information'. A physical example of this is the formation of a shock wave in gas dynamics when a piston is pulled out of a semi-infinite tube with a decelerating motion.

Mathematically the characteristic curves for a second order hyperbolic partial differential equation can be calculated as follows.

Let $p = u_x$, $q = u_t$ and consider the equation

$$au_{xx} + bu_{xt} + cu_{tt} = e \tag{1.2}$$

where a, b, c, e are functions of x, t, u, p and q but not of u_{xx}, u_{xt} or u_{tt}.

Consider the differentials of p and q along a curve C in the x,t plane on which the equation is satisfied

$$dp = \frac{\partial p}{\partial x}\,dx + \frac{\partial p}{\partial t}\,dt$$

$$= u_{xx}dx + u_{xt}dt, \tag{1.3}$$

$$dq = \frac{\partial q}{\partial x}\,dx + \frac{\partial q}{\partial t}\,dt$$

$$= u_{xt}dx + u_{tt}dt, \tag{1.4}$$

where $\frac{dx}{dt}$ is the slope of the tangent to C.

The system of equations (1,2), (1.3) and (1.4) gives a unique solution for u_{xx}, u_{xt}, u_{tt} in terms of x, t, u, p and q unless the three equations are linearly dependent.

The condition that the equations are linearly dependent is

$$a\left(\frac{dt}{dx}\right)^2 - b\left(\frac{dt}{dx}\right) + c = 0 \tag{1.5}$$

which is obtained by forcing the determinant of the coefficients of the system (1.2) - (1.4) to vanish.

Curves whose tangents satisfy (1.5) are called characteristics.

Along this characteristic the solution of the differential equation must satisfy

$$a\frac{dp}{dx} \cdot \frac{dt}{dx} + c\frac{dq}{dx} + e\frac{dy}{dx} = 0. \tag{1.6}$$

Equation (1.5) can be considered as a quadratic in $\frac{dt}{dx}$. If this quadratic has two real roots there are two families of characteristics and the partial differential equation (1.2) is said to be hyperbolic. This occurs if and only if $b^2 > 4ac$.

Similarly, if $b^2 = 4ac$ there is one real root of (1.5) and so (1.2) is considered parabolic, and finally the situation of two complex conjugate roots, that is $b^2 < 4ac$, corresponds to an elliptic equation.

This analysis also explains why discontinuities propagate along the characteristics in hyperbolic problems. For characteristics are the curves along which the higher order derivatives of u are not uniquely determined and hence discontinuities may occur across these curves.

A numerical method which uses the characteristics of a hyperbolic system to determine its solution (termed the method of characteristics) is given in section 1.5.

We now consider a single first order equation for which the situation is, as might be expected, much simpler. The most general quasilinear first order equation is

$$ap + bq = c \tag{1.7}$$

where a, b, c are functions of x, t and u only and, as before, $p = u_x$, $q = u_t$.

If C is a curve in the x,t plane whose tangent is given by $\frac{dt}{dx}$ then along C the solution satisfies

$$pdx + qdt = du. \tag{1.8}$$

This 2x2 system, given by (1.7) and (1.8), for p and q does not have a unique solution if

$$adt - bdx = 0. \tag{1.9}$$

This is an ordinary differential equation for the curve C and in order for there to be any solutions along this curve (called a characteristic) the solution must satisfy the ordinary differential equation

$$cdx - adu = 0 \tag{1.10}$$

along the characteristic.

In the first order case there is, therefore, always one family of characteristics which consists of curves along which the partial differential equation reduces to an ordinary differential equation.

§1.4 Finite differences

Finite difference methods deal with partial differential equations by finding approximate solutions to the differential equation at discrete points (referred to as grid points) in the x,t plane.

Three questions can immediately be posed about this approximate solution. How accurate is it? If more grid points are used, how much more accurate will it become? Do the errors

grow so quickly that they swamp the true solution in a finite time? The first question is, obviously, the accuracy of the method whilst the second and third are its rate of convergence and stability.

To demonstrate how such questions are investigated in general an analysis of some finite difference schemes which may be used to solve quasilinear equations will be given. Generally the schemes are applied to an equation which arises in Chapter 3, although we begin by describing a scheme for a first order equation (which is also discussed in Chapter 2).

Consider the first order equation

$$U_t + \lambda U_x = 0 \,, \tag{1.11}$$

where λ is a positive constant, and suppose that the solution $U(x,t)$ is required in the region $0 < x < 1$, $t \geq 0$ and that an initial condition $U(x,0) = U_0(x)$ $0 < x < 1$, and a boundary condition $U(0,t) = U_1(t)$ $t > 0$, are given .

We cover the relevant semi-infinite strip of the x,t plane by a mesh where h is the grid size in the x direction and k that in the t direction so that the $(i,j)^{th}$ mesh point is given by $x = ih$, $t = jk$.

If $U_{i,j}$ is the exact solution at (ih,jk) then a Taylor series expansion for $U_{i,j+1}$ about $U_{i,j}$ gives

$$U_{i,j+1} = U_{i,j} + k\left(\frac{\partial U}{\partial t}\right)_{i,j} + \frac{k^2}{2}\left(\frac{\partial^2 U}{\partial t^2}\right)_{i,j} + O(k^3)$$

and using the equation (1.11) we can replace the t derivatives by those in the x direction.

This gives

$$U_{i,j+1} = U_{i,j} - \frac{k\lambda}{2h}(U_{i+1,j}-U_{i-1,j}) + \frac{k^2\lambda^2}{2h^2}(U_{i+1,j}-2U_{i,j}+U_{i-1,j}) + O(k^3+h^2k). \quad (1.12)$$

(The relationships

$$U_x = \frac{U_{i+1,j} - U_{i-1,j}}{2h} + O(h^2)$$

and

$$U_{xx} = \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} + O(h^2)$$

can be obtained from Taylor expansions in exactly the same way as for the t derivatives.)

If the $O(k^3+h^2k)$ terms are ignored then equation (1.12) gives a formula for calculating an approximation $u_{i,j}$ to the true solution $U_{i,j}$.

This formula is

$$u_{i,j+1} = u_{i,j} - \frac{k\lambda}{2h}(u_{i+1,j}-u_{i-1,j}) + \frac{k^2\lambda^2}{2h^2}(u_{i+1,j}-2u_{i,j}+u_{i-1,j}) . \quad (1.13)$$

This is known as the Lax-Wendroff scheme and was first presented in a paper by Lax and Wendroff [17]. It is easy to see from (1.12) that the amount by which the true solution does not satisfy the formula (1.13) is $kT_{i,j}$ where $T_{i,j}$ is $O(h^2+k^2)$. $T_{i,j}$ is called the local truncation error and since it is quadratic in h and k the scheme is said to be second order accurate. It should be noted that $T_{i,j}$ involves third derivatives of the exact solution and will, therefore, not be known exactly.

The rate of convergence of the scheme is quadratic. That is, if we refine the mesh so that the step sizes in both the t and x directions are halved then the local truncation error will be approximately divided by four.

The final question concerns stability. It is required to check if an error made on the j^{th} time step will increase due to the difference formula. If this is so then eventually the errors will increase without bound. A von Neumann investigation will be carried out. (See O'Brien et al [4].)

The error on any time step can be expressed as a linear combination of Fourier components $e^{i\beta h\ell}$, which then propagates in the form $e^{i\beta h\ell}\xi^m$. Since (1.13) is linear the error satisfies the same difference equations as $u_{\ell,m}$ and we have

$$e^{i\beta h\ell}\xi^{m+1} = e^{i\beta h\ell}\xi^m - \frac{k\lambda}{2h}(e^{i\beta h(\ell+1)} - e^{-i\beta h(\ell-1)})\xi^m + \frac{k^2\lambda^2}{2h^2}(e^{i\beta h(\ell+1)} - 2e^{i\beta h\ell} + e^{i\beta h(\ell-1)})\xi^m .$$

Dividing by $e^{i\beta h\ell}\xi^m$ gives

$$\xi = 1 - \frac{k\lambda}{h} i \sin \beta h - \frac{k^2\lambda^2}{h^2} 2 \sin^2 \frac{\beta h}{2} .$$

The error does not increase provided $|\xi| \leq 1$ for all βh and this is so if and only if

$$0 \leq \frac{k\lambda}{h} \leq 1 . \qquad (1.14)$$

So provided the chosen time step is not too large with respect to the grid size in the x direction the scheme is stable.

Lax-Wendroff schemes are often used for systems of first order equations which can be expressed in conservation form. That is, they can be written as

$$\frac{\partial \underline{U}}{\partial t} + \frac{\partial \underline{F}(\underline{U})}{\partial x} = 0.$$

Their value lies in the fact that they are both second order and explicit (the difference equation relates one point at the next time level to values at the old time level), whereas many second order schemes are implicit (the difference equation gives a relationship between more than one point at the new time level). Also, they do not necessarily have a prohibitively restrictive stability condition.

A similar analysis will now be carried out for a second order equation (namely the equation of Chapter 3) and some of the possible difference schemes which could be used to solve it. For simplicity we begin by considering the equation

$$ay_{xx} + cy_{tt} + e = 0 \tag{1.15}$$

where a, c and e are functions of x, t, y, y_x and y_t only.

A simple central difference formula can be obtained by using the following approximations for the derivatives at the point (ih,jk).

$$y_{xx} = \frac{y_{i+1,j} - 2y_{i,j} + y_{i-1,j}}{h^2}, \tag{1.16}$$

$$y_{tt} = \frac{y_{i,j+1} - 2y_{i,j} + y_{i,j-1}}{k^2}, \tag{1.17}$$

$$y_x = \frac{y_{i+1,j} - y_{i-1,j}}{2h}, \tag{1.18}$$

$$y_t = \frac{y_{i,j+1} - y_{i,j-1}}{2k} \tag{1.19}$$

(y_x and y_t may occur in a, c or e).

These approximations are justified, as in the first order example, by taking Taylor expansions about the point (ih,jk).

If the solution is known along the lines t = jk and t = (j-1)k then it can be calculated explicitly at (ih,(j+1)k) from this formula. If a, c and e are constants or depend only on x, t, y and y_x (which is the case of the equation in Chapter 3), the equation reduces to

$$y_{i,j+1} = 2y_{i,j} - y_{i,j-1} + \frac{k^2}{c_{i,j}}\left(-e_{i,j} - a_{i,j}\left(\frac{y_{i+1,j} - 2y_{i,j} + y_{i-1,j}}{h^2}\right)\right) \tag{1.20}$$

where $a_{i,j} = a\Big|_{x=ih,\,t=jk,\,y=y_{i,j}}$

and similarly for $e_{i,j}$, $c_{i,j}$.

If y is given along boundaries x = 0, x = nh then this may be easily incorporated by using (1.20) for $1 \leq i \leq n-1$ and the boundary conditions for i = 0, n.

If, instead of y, some function of y, y_x and y_t is given at the boundary then the situation is a little more complicated but basically finite differences are used to approximate the boundary condition and hence give an equation for $y_{n,j}$ or $y_{0,j}$.

For example, in the cable problem of Chapter 3 the boundary condition at x = nh is $y_x + y_t = 0$. This gives the difference equation

$$\frac{y_{n,j} - y_{n-1,j}}{h} + \frac{y_{n,j+1} - y_{n,j}}{k} = 0 \tag{1.21}$$

which can be used to find $y_{n,j+1}$.

← If y and y_t are prescribed on $t = 0$ then by setting $y_{i,0} = y(ih,0)$ and finding $y_{i,-1}$ from the difference equation

$$y_t(ih,0) = \frac{y_{i,0} - y_{i,-1}}{k} \quad , \tag{1.22}$$

then equation (1.20) now has enough initial information to find y successively at all further time steps.

The local truncation error can be examined for the case where a and c do not involve derivatives of y. Taylor expansions about (ih,jk) give

$$\frac{y_{i+1,j} - 2y_{i,j} + y_{i-1,j}}{h^2} = (y_{xx})_{i,j} + \frac{h^2}{12}\left(\frac{\partial^4 y}{\partial x^4}\right)_{i,j} + O(h^4), \tag{1.23}$$

$$\frac{y_{i,j+1} - 2y_{i,j} + y_{i,j-1}}{k^2} = (y_{tt})_{i,j} + \frac{k^2}{12}\left(\frac{\partial^4 y}{\partial t^4}\right)_{i,j} + O(k^4), \tag{1.24}$$

$$\frac{y_{i+1,j} - y_{i-1,j}}{2h} = (y_x)_{i,j} + \frac{h^2}{6}\left(\frac{\partial^3 y}{\partial x^3}\right)_{i,j} + O(h^4), \tag{1.25}$$

$$\frac{y_{i,j+1} - y_{i,j-1}}{2k} = (y_t)_{i,j} + \frac{k^2}{6}\left(\frac{\partial^3 y}{\partial t^3}\right)_{i,j} + O(k^4). \tag{1.26}$$

The truncation error is the amount by which the exact solution, Y, does not satisfy (1.20) and this yields a local truncation error

$$\frac{ah^2}{12}\left(\frac{\partial^4 Y}{\partial x^4}\right) + \frac{ck^2}{12}\left(\frac{\partial^4 Y}{\partial t^4}\right) + E + O(h^4+k^4) \tag{1.27}$$

where E is the leading term of the error arising from the approximation to e at (ih,jk). E contains multiples of

$$\frac{h^2}{6}\left(\frac{\partial^3 Y}{\partial x^3}\right)_{i,j} \text{ and } \frac{k^2}{6}\left(\frac{\partial^3 Y}{\partial t^3}\right)_{i,j} .$$

For the case of the cable problem of Chapter 3 there is also a term in the differential equation containing y_{xt}. In fact

the equation is

$$(A-x)y_{xx} - 4\alpha(y_{xt}+y_{tt}) = (y_x+y_t)(1+2\beta|y_x+y_t|) \qquad (1.28)$$

where A, α, β are constants.

The cross derivative term y_{xt} may be approximated in a number of ways. Here two different approaches will be analysed. The first is

$$y_{xt} = \frac{y_{i+1,j+1} - y_{i-1,j+1} + y_{i-1,j-1} - y_{i+1,j-1}}{4hk}. \qquad (1.29)$$

If this, together with approximations for y_{xx}, y_{tt}, y_x, y_t given by (1.16)-(1.19),are substituted into (1.28) then the scheme becomes implicit as the i^{th} equation contains three unknowns $y_{i+1,j+1}$, $y_{i,j+1}$ and $y_{i-1,j+1}$ for $1 \leq i \leq n-1$.

In this case y_{xx} at (ih,jk) can be approximated by

$$\tfrac{1}{4}(y_{xx})_{i,j-1} + \tfrac{1}{2}(y_{xx})_{i,j} + \tfrac{1}{4}(y_{xx})_{i,j+1} \qquad (1.30)$$

where $(y_{xx})_{i,j}$, $(y_{xx})_{i,j-1}$ and $(y_{xx})_{i,j+1}$ are as given in (1.16). This causes no extra work in solving the system of equations since the unknowns are still $y_{i-1,j+1}$, $y_{i,j+1}$, $y_{i+1,j+1}$ but stability is improved by this refinement. (This is discussed in more detail below.)

An alternative is to replace $(y_{xt})_{i,j}$ by

$$\frac{y_{i+1,j} - y_{i-1,j} + y_{i-1,j-1} - y_{i+1,j-1}}{2hk} \qquad (1.31)$$

and leave the approximation for y_{xx} as defined by (1.16). Each equations now contains only one unknown and so the i^{th} equation can be solved explicitly for $y_{i,j+1}$ as before.

Taylor series expansions give

$$\frac{y_{i+1,j+1} - y_{i-1,j+1} + y_{i-1,j-1} - y_{i-1,j+1}}{4hk}$$

$$= (y_{xt})_{i,j} + \tfrac{1}{6}(h^2 (\frac{\partial^4 y}{\partial x^3 \partial t})_{i,j} + k^2 (\frac{\partial^4 y}{\partial x \partial t^3})_{i,j})$$

$$+ O(h^4+k^4) \qquad (1.32)$$

and

$$\frac{y_{i+1,j} - y_{i-1,j} + y_{i-1,j-1} - y_{i+1,j-1}}{2hk}$$

$$= (y_{xt})_{i,j} - \frac{k}{2}(\frac{\partial^3 y}{\partial x \partial t^2})_{i,j} + O(h^2+k^2). \qquad (1.33)$$

Thus in the first case the local truncation error consists only of second order terms whereas in the second case it contains the first order term $2\alpha k(\frac{\partial^3 Y}{\partial x \partial t^2})_{i,j}$ where Y is the exact solution.

The second scheme is, therefore, far less accuratethan the first and this is one justification for solving the system of equations resulting from the implicit method. Another, more pressing, justification is from a consideration of stability.

A von Neumann investigation can be carried out just as in the case of the first order equation (1.11) provided that a, b and c are constants. Of course in the cable problem a is not constant, it depends on x, but Richtmyer and Morton [28] assert that a local stability analysis obtained by freezing the coefficients is valuable as in many cases of practical interest instabilities start as local phenomena.

As previously, consider a single error component of the form $e^{i\beta h\ell}\xi^m$.

← The error satisfies the homogeneous equation

$$a y_{xx} - 4\alpha y_{xt} - 4\alpha y_{tt} = 0$$

where a is some constant which approximates A-x.

This gives, with (1.31),

$$\frac{a}{h^2} e^{i\beta h} \ell \xi^m (e^{i\beta h} - 2 + e^{-i\beta h}) - \frac{2\alpha}{hk} e^{i\beta h} \ell \xi^m (e^{i\beta h} - e^{-i\beta h})(1 - \frac{1}{\xi})$$

$$- \frac{4\alpha}{k^2} e^{i\beta h} \ell \xi^m (\xi - 2 + \frac{1}{\xi}) = 0. \qquad (1.34)$$

Dividing by $e^{i\beta h} \ell \xi^{m-1}$ and rearranging yields

$$\xi^2 f + \xi(h - 2f + ig) + f - ig = 0, \qquad (1.35)$$

where

$$f = \frac{-4\alpha}{k^2}, \quad h = \frac{-4a}{h^2} \sin^2 \frac{\beta h}{2}, \quad g = \frac{-8\alpha}{hk} \sin \frac{\beta h}{2} \cos \frac{\beta h}{2}$$

(f, h and g are all real).

Let the roots be ξ_1, ξ_2.

For stability we need $|\xi_1| \leq 1$ and $|\xi_2| \leq 1$ for all β. The product of roots $= \xi_1 \xi_2 = \frac{f - ig}{f} = 1 - \frac{ig}{f}$.

So $|\xi_1||\xi_2| = |1 - \frac{ig}{f}| = (1 + g^2/f^2)^{\frac{1}{2}} > 1$ unless $g = 0$ (which implies $\alpha = 0$ which is not the case).

← This method is, therefore, always unstable for any choice of step size h and k.

Similarly using (1.29) gives the stability equation

$$\frac{a}{4h^2}(e^{i\beta h} - 2 + e^{-i\beta h})(\xi + 2 + \frac{1}{\xi}) - \frac{\alpha}{hk}(e^{i\beta h} - e^{-i\beta h})(\xi - \frac{1}{\xi}) - \frac{4\alpha}{k^2}(\xi - 2 + \frac{1}{\xi}) = 0.$$

That is,

$$\xi^2(f + ig) + \xi h + f - ig = 0, \qquad (1.36)$$

where

$$f = \frac{a}{2h^2}(\cos\beta h - 1) - \frac{4\alpha}{k^2} ,$$

$$h = \frac{a}{h^2}(\cos\beta h - 1) + \frac{8\alpha}{k^2} ,$$

$$g = \frac{-2\alpha}{hk}\sin\beta h .$$

If the roots are ξ_1, ξ_2 then

$$\xi_1\xi_2 = \frac{f-ig}{f+ig} \text{ and so } |\xi_1||\xi_2| = \left|\frac{f-ig}{f+ig}\right| = 1 .$$

The scheme is von Neumann stable, therefore, if and only if a root can be found with unit modulus and $\xi_1 \neq \xi_2$. That is, if and only if, there exists ϕ such that $e^{i\phi}$ satisfies (1.36).

That is,

$$e^{i\phi}(f+ig) + h + (f-ig)e^{-i\phi} = 0$$

which implies

$$2f\cos\phi + ig(2i\sin\phi) = -h$$

$$\cos(\phi+\psi) = \frac{-h}{2(f^2+g^2)^{\frac{1}{2}}} \quad \text{where } \tan\psi = \frac{g}{f} .$$

ϕ exists if and only if

$$\left|\frac{-h}{2(f^2+g^2)^{\frac{1}{2}}}\right| \leq 1 \quad \text{for all } \beta$$

that is

$$\left(2f+\frac{16\alpha}{k^2}\right)^2 \leq 4(f^2+g^2) \quad \text{for all } \beta .$$

$$-16\alpha a \leq 16\alpha^2\cos^2\frac{\beta h}{2} \quad \text{for all } \beta .$$

The condition that $\xi_1 \neq \xi_2$ forces a strict inequality.

That is, the von Neumann condition for stability is satisfied if and only if

$$\alpha a > 0 \ . \tag{1.37}$$

Since this condition is satisfied for the constants of the cable problem this analysis indicates that the scheme should be stable for any choice of h and k.

In contrast if y_{xx} is not averaged over the three time steps the method is only locally stable for

$$\frac{k}{h} \leq (\frac{4\alpha}{a})^{\frac{1}{2}} \ .$$

An alternative attempt to find an explicit scheme for (1.28) can be used since (1.28) does not contain y explicitly. This is to split the second order equation into two first order equations in p and q where $p = y_x$ and $q = y_t$.

This gives,

$$(A-x)p_x - 4\alpha q_x - 4\alpha q_t + e = 0 \tag{1.38}$$

and

$$p_t = q_x \ . \tag{1.39}$$

(1.39) merely stipulates that $y_{xt} = y_{tx}$.

A finite difference scheme for these equations is

$$a_{i,j}(\frac{p_{i+1,j}-p_{i-1,j}}{2h}) - 4\alpha(\frac{q_{i+1,j}-q_{i-1,j}}{2h}) - 4\alpha(\frac{q_{i,j+1}-q_{i,j}}{k}) = -e_{i,j} \tag{1.40}$$

and

$$\frac{p_{i,j+1}-p_{i,j}}{k} = \frac{q_{i+1,j+1}-q_{i-1,j+1}}{2h}. \tag{1.41}$$

$$1 \leq i \leq n-1$$

(1.40) is solved for $q_{i,j+1}$ $(1 \leq i \leq n-1)$ and then since y is a given function on the boundary x = 0, q is the derivative of this function with respect to t and so $q_{0,j}$ is known for all j.

The other boundary condition, at x = nh, is simply p + q = 0.

Unfortunately a stability analysis of this scheme with the coefficient of p_x in (1.38) assumed constant gives unconditional instability.

Equations (1.38) and (1.39) may also be solved using the Lax-Wendroff scheme for a system of equations. This is a conditionally stable second order accurate scheme giving $q_{i,j+1}$ and $p_{i,j+1}$ explicitly in terms of values on the j^{th} time level. However, difficulties arise due to insufficient information at the ends. At x = 0 some other approximation must be used as the Lax-Wendroff scheme requires $p_{-1,j}$ and $q_{-1,j}$ to calculate $p_{0,j+1}$ and similarly at x = nh values are required outside the boundary.

Thus, the Lax-Wendroff scheme, which is explicit, is in fact more difficult to implement than the implicit second order scheme using (1.29).

§1.5 The Characteristic Method

The characteristic method makes use of the characteristics discussed in section 1.3 to solve hyperbolic partial differential equations. The characteristic method outlined below is a general approach but simplifications may be possible for particular equations. Characteristic methods are used in this thesis to solve the cable problem of Chapter 3, the

thermal runaway problem of Chapter 5 and the problem presented in Chapter 4.

Consider a general quasi-linear second order equation as given in (1.2). Then the tangents to the characteristics are the roots of (1.5), namely,

$$\frac{dt}{dx} = \frac{b + (b^2-4ac)^{\frac{1}{2}}}{2a} = f \qquad (1.42)$$

$$\frac{dt}{dx} = \frac{b - (b^2-4ac)^{\frac{1}{2}}}{2a} = g \qquad (1.43)$$

where (1.42) and (1.43) define f and g.

Along these curves the characteristic relation (1.6) becomes

$$afdp + cdq + edt = 0 \qquad (1.44)$$

and

$$agdp + cdq + edt = 0 \ , \qquad (1.45)$$

where $p = u_x$, $q = u_t$ as previously.

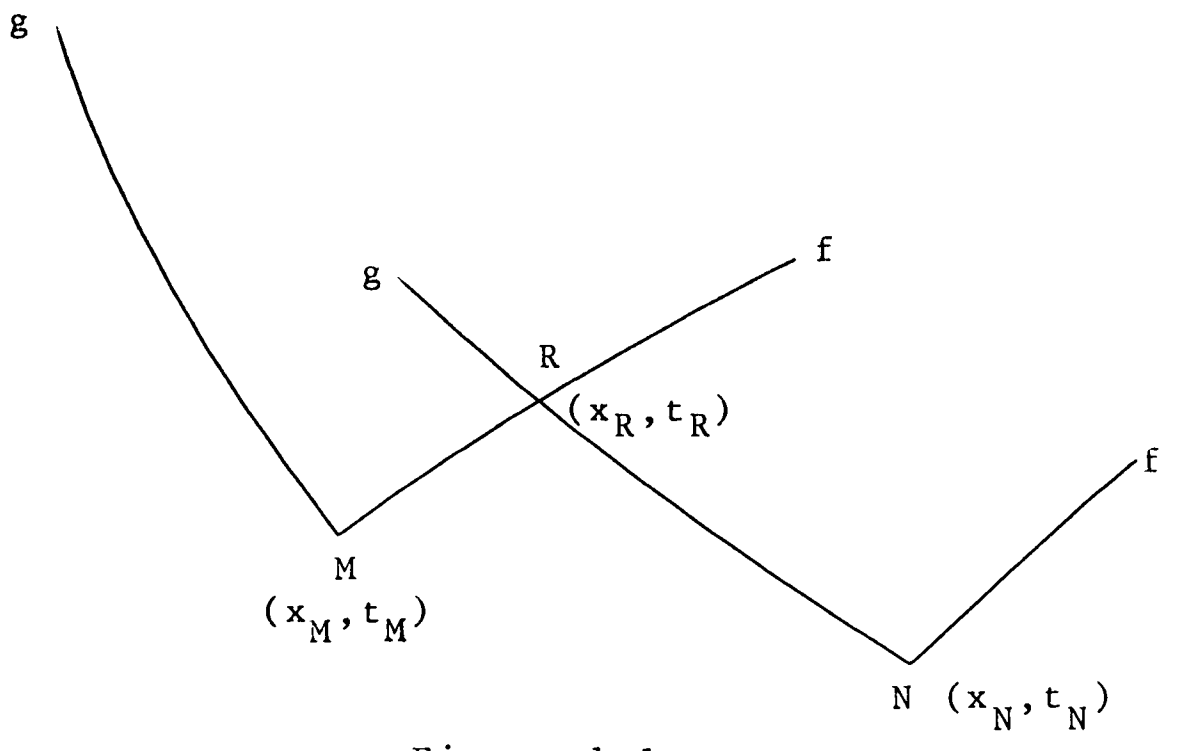


Figure 1.1

Suppose that u, p and q are known at two points M and N close together in the x,t plane. Let R be the point where the f characteristic through M and the g characteristic through N intersect. (See Figure 1.1.)

Then x, t, u, p, q can be calculated from the four equations (1.42)-(1.45), together with

$$du = pdx + qdt. \tag{1.46}$$

The equations are written in difference form as

$$t_R - t_M = (x_R - x_M) f_M \tag{1.47}$$

$$t_R - t_N = (x_R - x_N) g_N \tag{1.48}$$

$$a_M f_M (p_R - p_M) + c_M (q_R - q_M) + e_M (t_R - t_M) = 0 \tag{1.49}$$

$$a_N g_N (p_R - p_N) + c_N (q_R - q_N) + e_N (t_R - t_N) = 0 \tag{1.50}$$

$$u_R - u_M = \tfrac{1}{2}(p_R + p_M)(x_R - x_M) + \tfrac{1}{2}(q_R + q_M)(t_R - t_M) . \tag{1.51}$$

Then

1) x_R and t_R are calculated from (1.47) and (1.48);
2) p_R and q_R are found from (1.49) and (1.50);
3) u_R is obtained from (1.51) or the equivalent relation in R and N.

This appoximation for u_R can now be improved by replacing the values f_M, g_N, a_M, c_M etc., by averages to give

$$(t_R - t_M) = \tfrac{1}{2}(x_R - x_M)(f_R + f_M) \tag{1.52}$$

$$(t_R - t_N) = \tfrac{1}{2}(x_R - x_N)(g_R + g_N) \tag{1.53}$$

$$\tfrac{1}{2}(a_R+a_M)\tfrac{1}{2}(f_R+f_M)(p_R-p_M)+\tfrac{1}{2}(c_R+c_M)(q_R-q_M)+\tfrac{1}{2}(e_R+e_M)(t_R-t_M)=0 \quad (1.54)$$

$$\tfrac{1}{2}(a_R+a_N)\tfrac{1}{2}(g_R+g_N)(p_R-p_N)+\tfrac{1}{2}(c_R+c_N)(q_R-q_N)+\tfrac{1}{2}(e_R+e_N)(t_R-t_N)=0 \quad (1.55)$$

and (1.51) which is unaltered.

These equations are solved and the process repeated until succeeding approximations to u_R are sufficiently close. (Specific tolerances are given later when the algorithm is applied to particular problems.)

← Once the solution of R and S has been calculated the process can be used to extend the solution to T. (See Figure 1.2.)

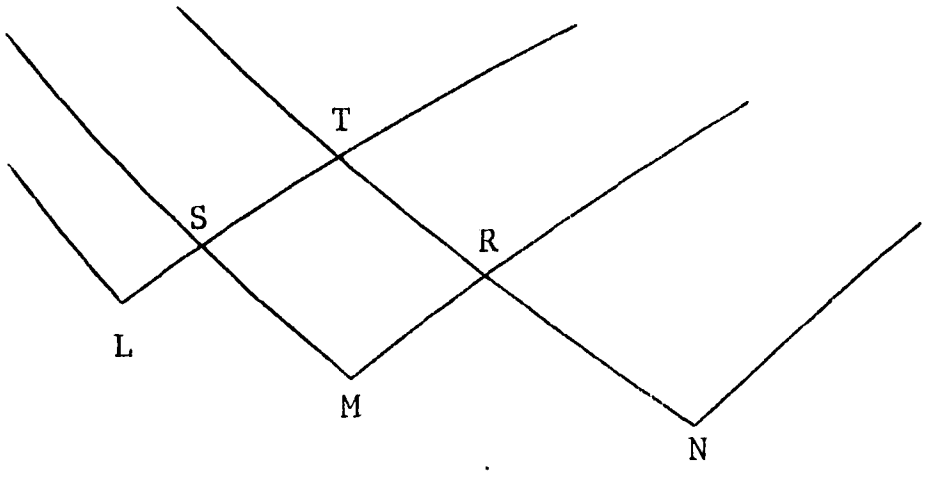


Figure 1.2

In this way the partial differential equation may be solved in a region of the x,t plane provided that the solution is known along an initial non-characteristic curve. If the curve of initial values is characteristic, as it is in Chapters 4 and 5, then there are restrictions about how much data can be supplied on the initial curve. There must be sufficient degrees of freedom left for the characteristic equation to be satisfied along the initial line.

Also, since only one characteristic from each initial point will cross the initial curve the solution is not determined unless there is enough boundary information. (See Courant and Hilbert [7].)

A disadvantage of using the characteristic method as it has been presented is that it is not known in advance where the characteristics will intersect and so it is not possible to calculate the solution at prescribed points. (For example, on a rectangular grid.)

One alternative would be to interpolate the solution obtained in order to find it at the desired points. This is the method used for the thermal runaway problems when one set of characteristics are the lines t = constant.

Hartree [12], however, proposed a scheme which is more convenient in many situations and is the method used in the cable problem.

Suppose the solution is known at grid points on the line $t = t_0$ and it is desired along the next t grid line. (See Figure 1.3.)

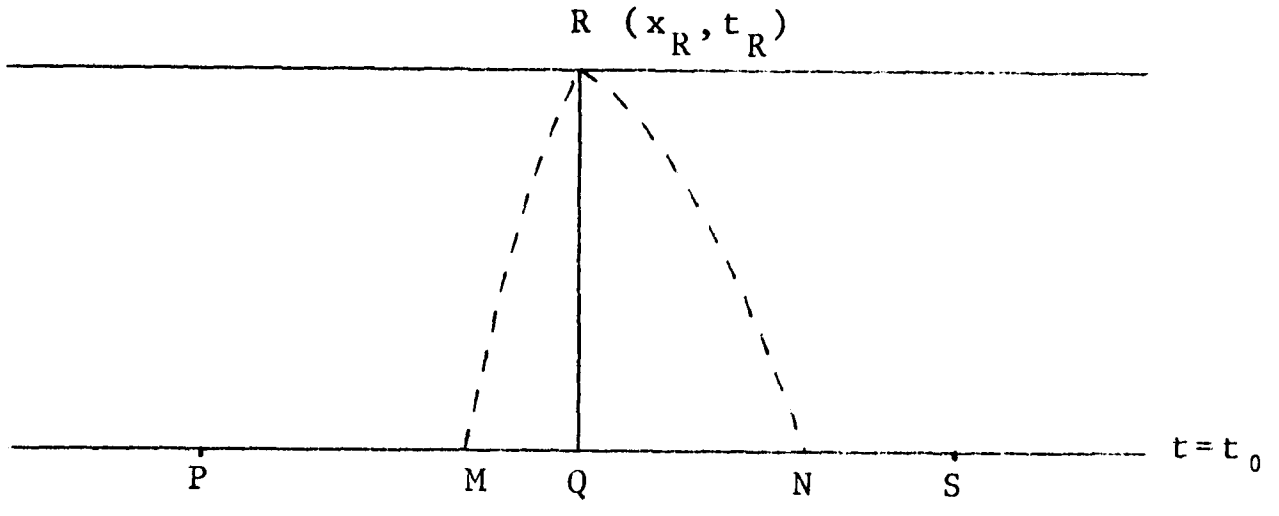


Figure 1.3

The solution has been calculated at P, Q and S and the point (x_R, t_R) is prescribed. The equations (1.42)-(1.45) still apply but now, because the unknowns are x_M, x_N, p_R, q_R, u_R the approach must be slightly different.

In general when a, b or c (and hence f and g) depend on the solution or its first derivatives, the following procedure is adopted.

1) Values for x_M and x_N are estimated.

2) Estimates for u_M, p_M, q_M, u_N, p_N, q_N are obtained by interpolating linearly between the two points adjacent to M and N. (In the diagram the characteristics cross the $t = t_0$ line in PQ and RS but step lengths may have been chosen so that M lies to the left of P for example.)

It is also possible to use quadratic interpolation, again using the solution at neighbouring points. When this method is applied to the industrial problem in Chapter 3, linear interpolation is normally used but quadratic interpolation has been considered and is discussed.

3) These estimates are used to evaluate the coefficients a, b, c, e at M and N and hence f_M and g_N.

4) (1.47) and (1.48) are now solved for x_M and x_N and these points compared with the original guesses.
If they are not tolerably close the process is repeated.

5) Once x_M and x_N have been found to an acceptable accuracy, (1.49) and (1.50) are solved for p_R and q_R and then u_R is obtained from (1.51).

6) Further refinement may be carried out by using the 'averaged' equations (1.52)-(1.55) until the desired accuracy has been achieved.

In many cases the calculations are much simpler than this. In the cable problem, for example, f and g depend only on x. This enables x_N and x_M to be calculated either directly or from (1.52) and (1.53) since both f_R and g_R are now known. The coupled equations (1.49) and (1.50) give p_R and q_R leaving (1.51) to give u_R explicitly once x_N, x_M, p_R and q_R have been found.

The boundary condition may be incorporated in two ways. Firstly, to calculate the solution at the boundary points, the boundary condition and the conditions obtained from the characteristic emanating from a boundary point into the

solution region are used.

For example, suppose that the boundary is the line x = constant, on which $Z(u,p,q) = 0$, where Z is a given function. This equation is solved together with (1.47), (1.49) and (1.51) to find either t_R, u_R, p_R and q_R if the standard method is used or x_M, u_R, p_R, q_R if the Hartree method is used. (It is assumed that RM is the characteristic from the boundary into the region of interest.) (See Figure 1.4.)

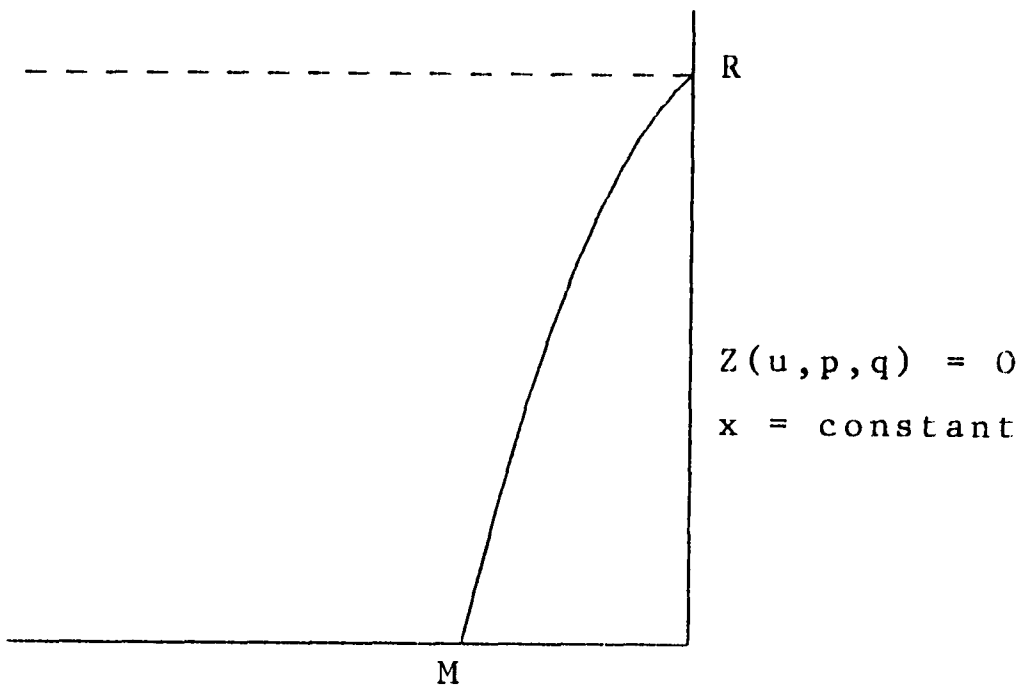


Figure 1.4

The second way in which the boundary is incorporated is applicable only in the Hartree method. If the characteristic from the internal point R' crosses the boundary before it crosses the line $t = t_0$ then neighbouring points on the boundary are used for the interpolation instead of the points on the line $t = t_0$. (See Figure 1.5.).

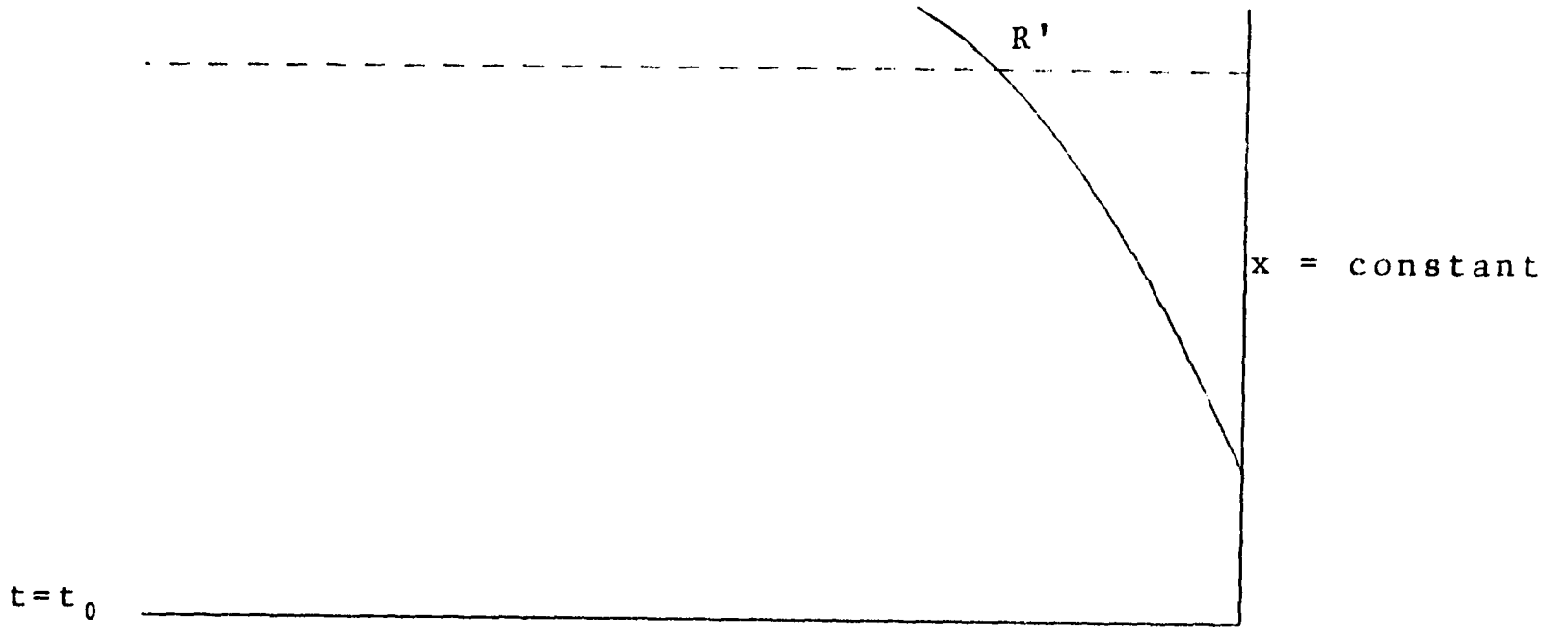


Figure 1.5

The accuracy of the characteristic method depends on two factors. The first is the accuracy of the approximation to the position of the characteristics and the relationships along them given by the ordinary differential equations (1.42)-(1.45). So if Euler's method is used, as in (1.47)-(1.50) the method is first order whereas if the trapezoidal rule is used it becomes second order. (Note that using (1.47)-(1.50) as a predictor followed by (1.52)-(1.55) as a corrector corresponds to using Euler's method and the trapezoidal rule as a predictor-corrector pair. This has unrestricted partial stability provided that the trapezoidal corrector is used to convergence.)

The second factor affecting accuracy is the way in which any interpolation is carried out when a rectangular mesh is used (as in Hartree's method). Linear interpolation gives first order accuracy; quadratic second order.

It is interesting that the consideration of characteristics helps to give guidelines for the stability of explicit finite difference schemes, via the Courant-Friedrichs-Lewy condition. For the domain of dependence of the true solution at a point, P, is bounded by the two characteristics through P, (see Figure 1.6), which are given by the lines BP and CP.

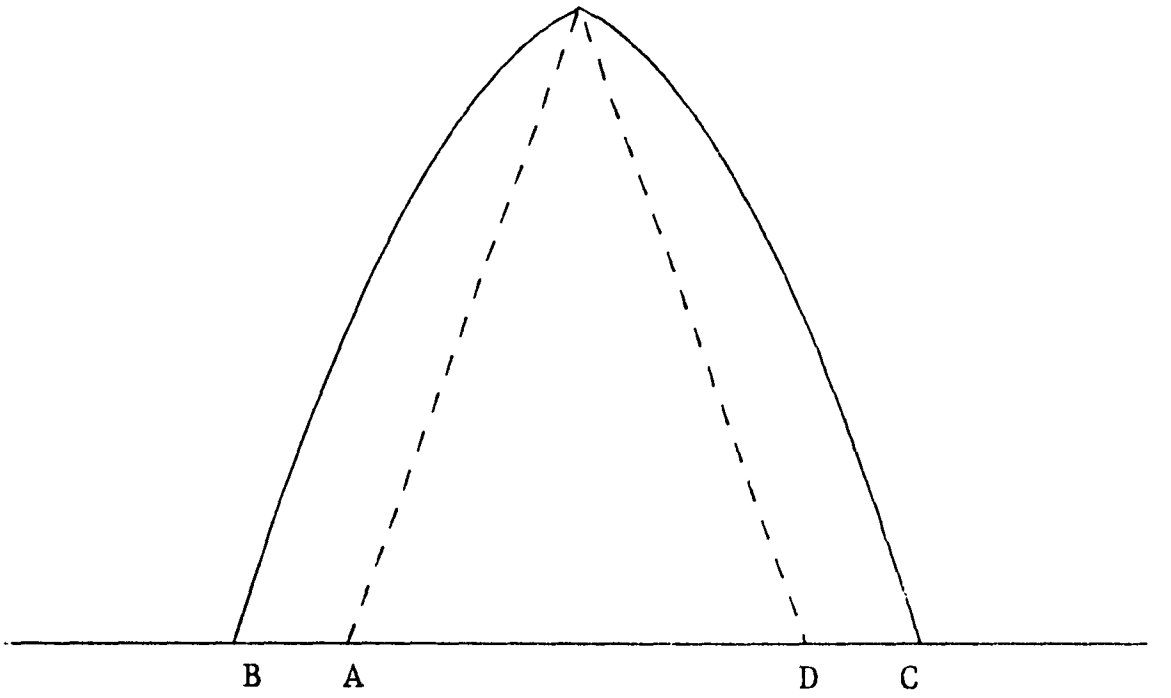


Figure 1.6

If the domain of dependence of the finite difference scheme is bounded by the lines AP and DP then changes in the data along BA or DC will not affect the numerical solution but they do affect the analytical solution and so the numerical solution cannot converge to the analytical one. Hence the domain of

dependence of the numerical scheme must include that of the differential equation. This, for example, leads immediately to the stability condition $k/h \leq 1$ for the explicit method for the solution of the wave equation $u_{xx} = u_{tt}$.

§1.6 A comparison of finite difference and characteristic methods

The main advantage of characteristics over finite differences is that discontinuities in the initial data are propagated as discontinuities in the solution region along the characteristics. A characteristic can thus separate two different solutions and the discontinuity is not dissipated.

In contrast finite difference methods smooth out discontinuities so that the solution is continuous but changes very rapidly over the discontinuity. This is obviously a less accurate approach.

However, compensating for this drawback finite difference methods have an advantage in terms of ease of computation. For characteristic methods the calculations must be done separately on each side of a discontinuity and so for a problem with many singularities this may be a prohibitively difficult and expensive method.

In contrast finite difference schemes do not become significantly more complicated with the nature of the problem and they may give adequate accuracy in problems where the general behaviour of the solution is required, but the details of the solution near the discontinuity are unimportant.

CHAPTER 2

The Saddle Point Method and Errors in Finite Difference Schemes

§2.1 Introduction

In this chapter a method for evaluating the error when a finite difference scheme is used to find the solution to a problem with discontinuities is discussed.

Using Taylor expansions it is a relatively simple procedure to calculate the leading term of the error for continuous solutions. However, as the error expansion is in the form of the derivative of the solution at some point this method breaks down if the solution is discontinuous. As Rozhdestvenskii [30] shows, for quasi linear hyperbolic systems discontinuities may appear despite initially smooth data and if the initial data is not smooth then discontinuities will propagate along characteristics as explained in Chapter 1. In this chapter then, an alternative approach for these situations is presented.

In section 2.2 the saddle point method, which is the basic tool of the method is outlined and then in 2.3 the method is applied to a simple test problem. The remaining sections contain extensions of the method to other simple problems and a variety of finite difference schemes.

§2.2 The Saddle Point Method

This section is concerned with the solution of

$$\int_C g(z)e^{sh(z)}dz \tag{2.1}$$

asymptotically as $s\to\infty$ where $g(z)$ and $h(z)$ are analytic in some region D in the complex plane.

The need for an asymptotic approximation to this integral arises very shortly (see, for example, equation (2.9)) and in all the cases to be considered $g(z)$ and $h(z)$ are both analytic in C except for a finite number of poles.

This integral can be evaluated using the saddle point method (also known as the method of steepest descents).

There are two main stages to the method. The first is the replacement of C by a suitably chosen new contour C' so that the major contribution to the integral is from only a small part of the path. This contribution should become increasingly dominant as $s\to\infty$.

The second stage is to approximate the integral by an asymptotic expression for the contribution from this short part of the contour.

As an intermediate stage we consider

$$F(s) = \int_p e^{s\psi(z)}dz \tag{2.2}$$

where ψ is an analytic function of z in D.

We wish to find an asymptotic approximation to this integral as $s\to\infty$.

Since $|e^{\psi(z)}| = \exp(s\mathrm{Re}(\psi(z)))$ we must try to find a path which passes through a maximum of $\mathrm{Re}(\psi(z))$ and such that

$Re(\psi(z))$ decreases as quickly as possible away from this maximum.

We must also ensure that $Im(\psi(z))$ remains constant along the path, otherwise the increasingly dense oscillations of $\exp(s(\psi(z))$ as s increases may negate the relative increase in the real part near to the maximum.

If we can find such a path then the major contribution to the integral will be from the part of the contour passing through the maximum and this contribution will become increasingly dominant as s increases, as required.

Since $\psi(z)$ is analytic in D, $Re(\psi(z))$ and $Im(\psi(z))$ satisfy the Cauchy-Riemann equations and hence $grad(Re(\psi(z)))$ is perpendicular to $grad(Im(\psi)))$. That is, the gradient vector (or line of steepest descent) of $Re(\psi(z))$ is tangential to the line of constant $Im(\psi(z))$.

The required contour, then, must be chosen so that it passes along a line of constant $Im(\psi(z))$, that is such that $grad(Re(\psi(z)))$ is always tangential to the path, and also such that it passes through a maximum of $Re\psi(z)$ along the path.

At this maximum $grad(Re(\psi(z)))$ is zero and hence by the Maximum Modulus Theorem this maximum point must in fact be a saddle point of $Re(\psi(z))$.

The saddle points of $Re(\psi(z))$ are the points where $\frac{d}{dz}(Re(\psi(z))) = 0$ and by the Cauchy-Riemann conditions this is equivalent to $\psi'(z) = 0$.

Suppose z_0 is a saddle point of $\psi(z)$. Then expanding about z_0 gives

$$\psi(z) = \psi(z_0) + (z-z_0)^2\frac{\psi''(z_0)}{2!} + O(z-z_0)^3.$$

If $z - z_0 = re^{i\theta}$ and $\psi''(z_0) = \alpha e^{i\phi}$ then

$$\psi(z) = \psi(z_0) + \frac{r^2\alpha}{2!}e^{i(2\theta+\phi)} + O(r^3)$$

and hence z lies on the line of steepest descent through z_0 if and only if $e^{i(2\theta+\phi)}$ is real and negative.

That is, $\theta = \frac{\pi}{2} - \frac{\phi}{2}$.

So the required contour crosses the saddle point in the direction

$$\frac{\pi}{2} - \tfrac{1}{2}\arg\psi''(z_0) \ . \qquad (2.3)$$

We now consider

$$\int_c g(z)e^{sh(z)}dz.$$

Provided that g(z) does not behave violently near to the chosen contour it is sufficient to base the choice of contour on the saddle points of h(z). As s increases (and thus exp(sh(z)) behaves more and more violently near to the saddle points if the contour passes along the line of steepest descent), this assumption will become more and more valid.

The first stage of the method, therefore, consists of finding the points where $h'(z) = 0$ and then finding a contour which passes through at least one of these saddle points and which crosses any saddle point, z_k, through which it passes, in the direction $\frac{\pi}{2} - \tfrac{1}{2}\arg h''(z_k)$.

We now move to the second stage of the method which is to find an asymptotic expression for

$$\int_A^B g(z)e^{sh(z)}dz \tag{2.4}$$

where AB is the path of steepest descent through the saddle point z_0 of $h(z)$.

Changing variables,

$$z = z_0 + \omega\chi$$

$$a \leq \chi \leq b$$

$$a < 0 < b$$

$$\omega = \exp(\tfrac{1}{2}\pi i - \tfrac{1}{2}i\arg h''(z_0))$$

$$A = z_0 + a\omega$$

$$B = z_0 + b\omega$$

we obtain

$$\omega\int_a^b g(z_0+\omega\chi)e^{sh(z_0+\omega\chi)}d\chi. \tag{2.5}$$

$h(z_0+\omega\chi)$ can be expanded about z_0 to give

$$h(z_0+\omega\chi) = h(z_0) + \tfrac{1}{2}h''(z_0)\omega^2\chi + O(\omega^3)$$

where $h''(z_0)<0$.

For sufficiently large values of s the main contribution to the integral comes from near x = 0; the value of the integral does not therefore depend on the limits of integration, which can be replaced by $(-\infty,\infty)$.

Then using Laplace's method gives the asymptotic expansion

$$\sqrt{\frac{2\pi}{s}}\ \omega\ \frac{1}{\sqrt{|h''(z_0)|}}\ g(z_0)e^{sh(z_0)}\{1+O(s^{-1})\} \tag{2.6}$$

as $s\to\infty$

for the contribution of a saddle point z_0, through which the contour C passes, to the integral (2.1).

There is still ambiguity about the sign of the contribution which must be investigated.

Since $a < 0 < b$ the contour passes into the region where the axis has slope $\frac{\pi}{2} - \frac{1}{2}\arg h''(z_0)$ from the region with slope $-(\frac{\pi}{2}+\frac{1}{2}\arg h''(z_0))$.

Since $-\pi < \arg h''(z_0) < \pi$ we have

$$0 < \tfrac{1}{2}\pi - \tfrac{1}{2}\arg h''(z_0) < \pi$$

and similarly $0 > -(\pi/2+\frac{1}{2}\arg h''(z_0)) > -\pi$, and so for a contribution as given by (2.6) (referred to as a 'positive' contribution despite the fact that it may not necessarily be positive in sign) the contour must cross the saddle point in a direction such that Im(z) is increasing.

Obviously if the saddle point is crossed in the reverse direction then the negative of the contribution given by (2.6) must be taken.

§2.3 The Test Problem

In this section I shall explain how equation (2.6) can be used to obtain an asymptotic approximation for the error in finite difference approximations when applied to discontinuous problems.

The simple test problem to be considered is

$$u_t + u_x = 0 \qquad (2.7)$$

with initial conditions

$$u(x,0) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The solution to this equation is a square wave moving to the right with speed 1 and height 1, that is u(x,t) = u(x-t,0). (See Figure 2.1).

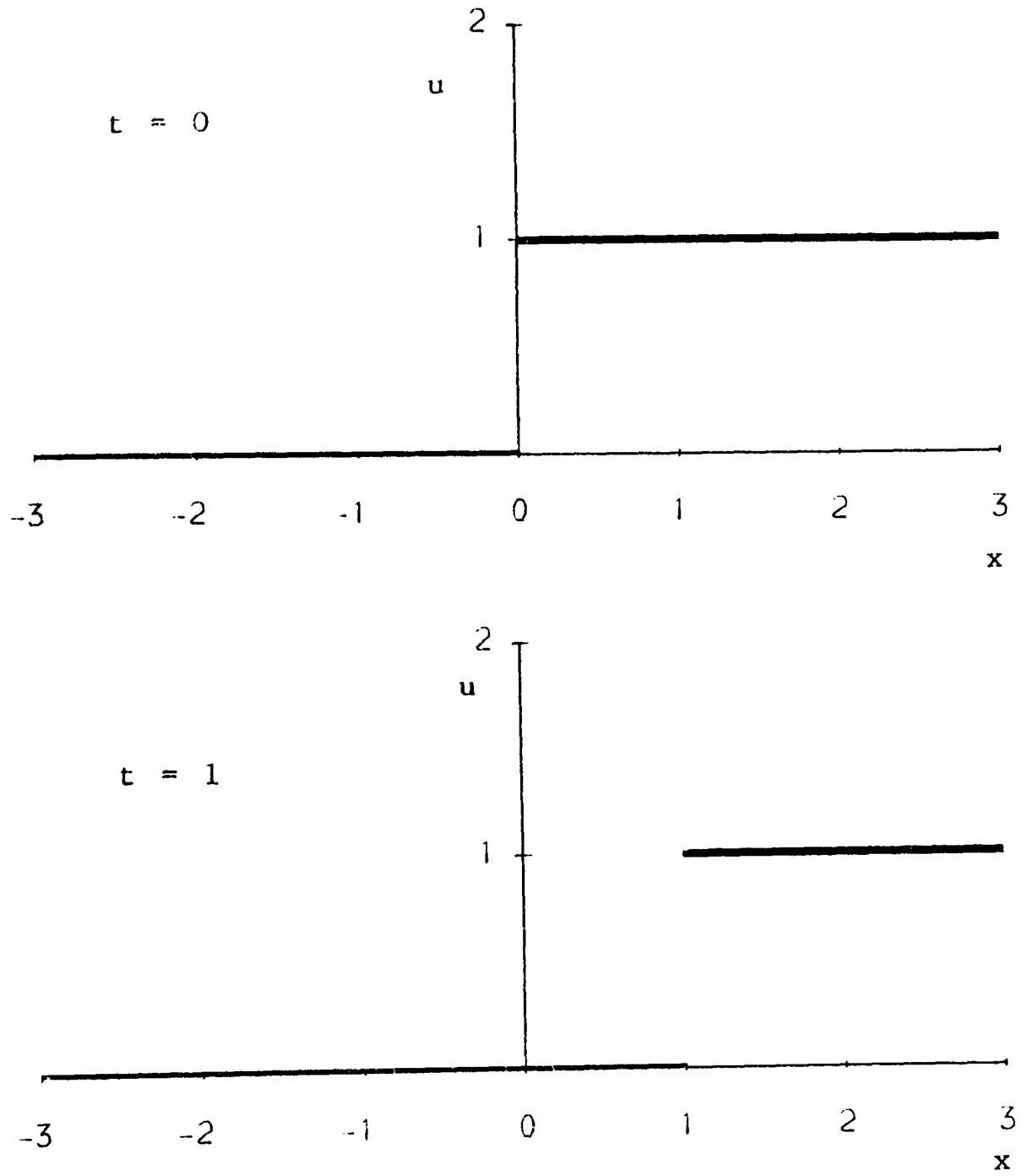


Figure 2.1

If $u_{r,s}$ is an approximation to the solution at (rh,sk) then using the Lax-Wendroff finite difference scheme we obtain the equation

$$u_{r,s+1} = au_{r-1,s} + bu_{r,s} + cu_{r+1,s} \tag{2.8}$$

where $a = \frac{1}{2}p(1+p)$

$b = 1-p^2$

$c = \frac{1}{2}p(p-1)$

and $p = k/h$

k is the step size in the t direction

h is the step size in the x direction.

Equation (2.8) enables us to find approximations (explicitly) for u at successive time steps as we are given values for u initially.

Provided the exact solution U(x,t) has continuous third derivatives, a Taylor series expansion gives a local truncation error for the Lax-Wendroff scheme of

$$\tfrac{1}{6}k^2\frac{\partial^3 U}{\partial t^3} + \tfrac{1}{2}h^2\frac{\partial^3 U}{\partial x^2} + O(h^3) + O(k^3).$$

In our case since the initial data and hence the exact solution U are discontinuous at x = t, we must try to find an alternative expression for the error.

We define

$$f_s(z) = \sum_{r=-\infty}^{\infty} u_{rs} z^r$$

and then multiply (2.8) by z^r and sum from $-\infty$ to ∞ to obtain

$$f_s(z) = (az+b+\frac{c}{z}) f_{s-1}(z)$$

$$= (az+b+\frac{c}{z})^s f_0(z).$$

However,

$$f_0(z) = \sum_{r=-\infty}^{\infty} u_{r,0} z^r$$

$$= 1 + z + z^2 + \ldots$$

from given initial conditions

$$= \frac{1}{1-z}$$

so

$$f_s(z) = (az+b+\frac{c}{z})^s \frac{1}{1-z}.$$

By our definition of $f_s(z)$, u_{rs} is the coefficient of z^r in $(az+b+\frac{c}{z})^s \frac{1}{1-z}$.

That is, u_{rs} is the residue at the origin of

$$\frac{a^s(z+\alpha)^s(z-\beta)^s}{z^{r+s+1}(1-z)}$$

where $-\alpha,\beta$ are the roots of $az^2 + bz + c = 0$.

So,

$$u_{rs} = \frac{1}{2\pi i} \int_C \frac{a^s(z+\alpha)^s(z-\beta)^s}{z^{r+s+1}(1-z)} dz \qquad (2.9)$$

where C is a contour which encircles the origin but includes no other poles. We wish to consider the behaviour of u_{rs} as $r,s\to\infty$ with r/s fixed.

That is, we fix x and t and let the step size tend to zero. This corresponds to just looking at the leading term of the truncation error in the continuous case.

Define $\gamma = \frac{r}{s}$ and keep γ fixed.

Equation (2.9) can be considered as

$$\int_C g(z)e^{sh(z)}dz \qquad (2.10)$$

where $g(z) = \frac{1}{z(1-z)}$

and $h(z) = \log a + \log(z+\alpha) + \log(z-\beta) - (\gamma+1)\log z$ and we require an approximation as $s\to\infty$.

(2.10) is in a suitable form for use for the saddle point method since g(z) and h(z) are meromorphic functions as required.

The function g(z) has a pole at z = 1 and so the contour C must be chosen carefully so that it encircles the origin and does not pass through z = 1.

h(z) has saddle points when h'(z) = 0. That is, when

$$\frac{1}{z+\alpha} + \frac{1}{z-\beta} - \left(\frac{\gamma+1}{z}\right) = 0.$$

This gives a quadratic in z.

Let the roots be θ_+, θ_-.

Then,

$$\theta_{\pm} = \frac{\gamma}{1-\gamma}\left(\frac{1-p}{p}\right) \pm \frac{1}{1-\gamma}\left(\frac{1-p}{1+p}\right)^{\frac{1}{2}}\left(\frac{\gamma^2}{p^2}-1\right)^{\frac{1}{2}} .$$

So $\gamma > p \Rightarrow$ two real distinct roots

$\gamma = p \Rightarrow$ one real double root

$\gamma < p \Rightarrow$ a complex conjugate pair of roots.

These three cases lead to three different situations in the type of error obtained.

We first look at the error when $\gamma > p$, that is, there are two distinct roots.

The contour map of Re(h(z)) is given in Figure 2.2. The path passing through a saddle point and encircling the origin is also shown. It should be noted that this path only passes through the θ_+ saddle point, any path passing through the θ_- saddle point such that the saddle point is the highest point on the path will not encircle the origin.

The contour also contains the point z = 1 which is a pole of g(z). It is a simple pole and so the residue can easily be calculated and is given by

$$-\frac{a^s(1-\alpha)^s(1-\beta)^s}{1^{(\gamma+1)s+1}}$$

$$= -(a+b+c)^s$$

$$= -1.$$

The asymptotic approximation to the contribution from the saddle point is

$$\sqrt{\frac{2\pi}{s}}\,\omega\,|h''(\theta_+)|^{-\frac{1}{2}}g(\theta_+)e^{sh(\theta_+)}$$

$$= \sqrt{\frac{2\pi}{s|D_+|}}\,i\frac{(a\theta_+^2+b\theta_++c)^s}{\theta_+^{(\gamma+1)s}\theta_+(1-\theta_+)}$$

where $D_+ = h''(\theta_+)$.

The contour crosses the saddle point in an 'upwards' (that is, Im(z) increasing) direction and so the positive sign is taken for the contribution. We, therefore, have the result that

$$u_{rs} \sim 1 + \frac{1}{2\pi i}\sqrt{\frac{2\pi}{s|D_+|}}\, i\frac{(a\theta_+^2+b\theta_++c)^s}{\theta_+^{(\gamma+1)s}\theta_+(1-\theta_+)} \quad \text{as } s\to\infty$$

$$= 1 + \frac{(a\theta_+^2+b\theta_++c)^s}{\theta_+^{s(1+\gamma)+1}(1-\theta_+)}\,\frac{1}{\sqrt{2\pi sD_+}} \quad . \tag{2.11}$$

In the case $\gamma < p$ there are two complex conjugate saddle points and the contour map of Re(h(z)) looks like Figure 2.3.

Again, the path which encircles the origin and which has saddle points as its highest points is shown. It should be noticed that in this case the path chosen passes through both saddle points but does not encircle z = 1, the pole of g(z). Again, the saddle points are crossed in an upwards direction and so the contribution is positive.

The θ_+ saddle point gives a contribution of

$$\sqrt{\frac{2\pi}{s}}\,\frac{e^{i/2(\pi-\phi)}}{\sqrt{|D_+|}}\; g(\theta_+)e^{sh(\theta_+)}$$

and the θ_- one

$$\sqrt{\frac{2\pi}{s}}\,\frac{e^{i/2(\pi+\phi)}}{\sqrt{|D_-|}}\; g(\theta_-)e^{sh(\theta_-)} \quad .$$

Since $\theta_+ = \bar{\theta}_-$, $D_+ = \bar{D}_-$ and $\phi = \arg D_+ = -\arg D_-$,

we have

$$u_{rs} \sim 2\mathrm{Re}\left\{\frac{1}{\sqrt{2\pi sD}}\,\frac{(a\theta^2+b\theta+c)^s}{(1-\theta)\theta^{s(1+\gamma)+1}}\right\} \text{ as } s\to\infty \tag{2.12}$$

where $\theta = \theta_+$, $D = D_+$ or $\theta = \theta_-$, $D = D_-$.

It remains to consider the case when $\gamma = p$. There is a difficulty in this situation because the saddle point of h(z)

occurs at z = 1 which is a pole of g(z). All the previous work assumed that g(z) was analytic in a region D of the complex plane which contained the contour C.

To cope with this we split the integral up into four sections as follows

$$\int_C g(z)e^{sh(z)}dz = \int_\Gamma g(z)e^{sh(z)}dz + \int_A^B g(z)e^{sh(z)}dz$$

$$+ \int_\Delta g(z)e^{sh(z)}dz + \int_C^D g(z)e^{sh(z)}dz.$$

These four sections are marked on the contour map for $\gamma = p$ which is given in Figure 2.4.

Γ is the part of the contour, away from the saddle point, which can be ignored as previously.

The remainder of the path consists of AB, the part leading up to the saddle point; Δ, a section of the contour looping around the saddle point (but not passing through it) and finally CD the part of the path just after the saddle point.

AB and CD make an angle of $\frac{2\pi}{3}$ with each other and so Δ consists of a third of the circumference of a circle of radius δ.

Thus,

$$\begin{aligned}\int_\Delta g(z)e^{sh(z)}dz &= 2\pi i\tfrac{1}{3}.(\text{residue at } z=1)\\ &= 2\pi i.\tfrac{1}{3}.-1\\ &= \frac{-2\pi i}{3}.\end{aligned} \qquad (2.13)$$

For the AB integral we can parametrize the path by t and let t run from $-\infty$ to δ and then let $\delta \to 0$.

Then, $z = 1 + t(1-\sqrt{3}i)$ and the integral becomes

$$-\int_{\infty}^{0} \frac{e^{s[\frac{1}{6}H'''(0)t^3(1-i\sqrt{3})^3 + \frac{1}{24}H^{iv}(0)t^4(1-i\sqrt{3})^4]}(1-\sqrt{3}i)dt}{(t+1-\sqrt{3}ti)t(1-\sqrt{3}i)}$$

where $H(w) = h(z)$ and $z = w + 1$

and so

$$H(w) = H(0) + wH'(0) + \frac{w^2}{2!}H''(0) + \frac{w^3}{3!}H'''(0) + \frac{w^4}{4!}H^{iv}(0) + O(w^5)$$

and $H(0) = 0$ since $h(1) = 0$

$H'(0) = 0$ since $z = 1$ is a root of $h'(z)$

$H''(0) = 0$ since $z = 1$ is a double root of $h'(z)$.

All terms after the w^4 term are ignored.

Similarly the CD integral gives

$$-\int_{0}^{\infty} \frac{e^{s[\frac{1}{6}H'''(0)t^3(1+\sqrt{3}i)^3 + \frac{1}{24}H^{iv}(0)t^4(1+\sqrt{3}i)^4]}(1+\sqrt{3}i)dt}{(t+1+t\sqrt{3}i)t(1+\sqrt{3}i)} .$$

After some rearrangement we obtain

$$\int_{0}^{\infty} \frac{e^{-s/3[4H'''(0)t^3+H^{iv}(0)t^4]}}{t(4t^2+2t+1)} . 2i\sin(\frac{s\sqrt{3}}{3}H^{iv}(0)t^4+\theta(t))dt \qquad (2.14)$$

where $\cos\theta = \frac{1+t}{(1+t)^2+3t^2}$, $\sin\theta = \frac{t\sqrt{3}}{(1+t)^2+3t^2}$.

We now make the substitution

$$x^3 = \frac{4s}{3}H'''(0)t^3$$

that is,

$$t = x\left[\frac{3}{4sH'''(0)}\right]^{\frac{1}{3}}$$

$H'''(0) > 0$ so that x is real.

Consider $\theta(t)$.

$$\theta = \tan^{-1}\left(\frac{t\sqrt{3}}{t+1}\right) = \tan^{-1}\left(t\sqrt{3}(1-t+t^2-t^3+O(t^4)\right)$$

$$= \sqrt{3}t(1-t+t^2) - \frac{3\sqrt{3}}{3}t^3 + O(t^4)$$

$$= \sqrt{3}t(1-t) + O(t^4)$$

and so

$$\theta = \sqrt{3}x\left[\frac{3}{4sH'''(0)}\right]^{\frac{1}{3}}\left(1-x\left[\frac{3}{4sH'''(0)}\right]^{\frac{1}{3}}\right). \qquad (2.15)$$

In the above argument higher order terms in t or w have been ignored. This is justified by the fact that the major contribution to the integral comes from the part of the path close to the saddle point; that is, with t or w small, and so higher powers of these variables are negligible.

Substituting for t, approximating $\sin\xi$ by ξ and only keeping $O(s^{-\frac{1}{3}})$ terms we obtain from (2.13) and (2.14)

$$\int_{\infty}^{0} e^{-x^3} 2ix\left[\frac{3}{4sH'''(0)}\right]^{\frac{1}{3}}\left[\frac{H^{iv}(0)x^3}{4H'''(0)}+1\right]\sqrt{3}\,\frac{dx}{x}$$

which is

$$2i\left[\frac{3}{4sH'''(0)}\right]^{\frac{1}{3}}\sqrt{3}\left\{\frac{\Gamma(\frac{4}{3})}{3}\,\frac{H^{iv}(0)}{4H'''(0)} + \frac{\Gamma(\frac{1}{3})}{3}\right\}$$

and so,

$$u_{rs} - 1 = \frac{1}{2\pi i}\left[\int_A^B + \int_C^D - \frac{2\pi i}{3}\right]$$

which implies

$$u_{rs} \sim \frac{2}{3} + \frac{1}{\pi\sqrt{3}}\left[\frac{3}{4sH'''(0)}\right]^{\frac{1}{3}}\left\{\frac{H^{iv}(0)}{4H'''(0)} \cdot \frac{1}{3} + 1\right\}\Gamma(\tfrac{1}{3}) \text{ as } s\to\infty \quad (2.16)$$

and $H'''(0) = h'''(1)$, $H^{iv}(0) = h^{iv}(1)$

since $H(w) = h(z)$ and $z = w + 1$.

In this section, then, three different estimates for u_{rs} have been obtained depending on whether $\gamma > p$, $\gamma < p$ or $\gamma = p$; that is, $x > t$, $x < t$ or $x = t$ since $\gamma = r/s$ and $p = k/h$.

In the next section some results are given for particular cases and some general observations are made.

§2.4 Results and Observations

In this section results are presented for the above scheme for selected values of t, p, γ and s. An explanation of the behaviour of these results is given and this leads to the investigation of schemes other than the Lax-Wendroff one, in the next section.

Figure 2.5 shows a graph of the results for $t = 1$, $p = .5$, $0 < \gamma < 1$, $\gamma \neq .5$ for $s = 50$, 100 and 200. Numerical results for the situation $s = 50$ are given in Table 2.1. Three observations can be made.

Firstly, both the estimate of the error and the error itself become worse near to the discontinuity at $\gamma = \frac{1}{2}$. Secondly the complex roots correspond to an oscillating error

and the real roots to a smooth error. Finally the error is far less when the roots are real.

An explanation for the second observation is straightforward.

If we put

$$\frac{(a\theta^2+b\theta+c)}{\theta^{(1+\gamma)}} = r_1 e^{i\phi_1}$$

and

$$\frac{1}{(1-\theta)\theta\sqrt{D}} = r_2 e^{i\phi_2}$$

where a, b, c, θ, γ and D are as given in equation (2.12).

Then from (2.12), for $\gamma < p$,

$$u_{rs} = \frac{2.}{\sqrt{2\pi s}} r_1^s r_2 \cos(s\phi_1+\phi_2)$$

which is oscillatory. As s increases so does the frequency of the oscillations and this fact is supported by the results shown in Figure 2.5.

It should be noted that this is different from the basis upon which the saddle point method is built; namely that the integral should not oscillate as the contour crosses the saddle point. The oscillations in the results are oscillations as γ changes and oscillations at the saddle point as s changes but <u>not</u> oscillations along the path close to the saddle point for a fixed γ and s. In the case of real saddle points ϕ_1 and ϕ_2 are zero and so there is no oscillatory behaviour.

The other two observations can be explained as follows. From equation (2.12) and (2.13) it can be seen that both in

the case of two real saddle points and that of complex conjugate saddle points the error, E, in u_{rs} is of the form

$$|E| = k_\gamma \cdot \left|\frac{a\theta^2+b\theta+c}{\theta^{\gamma+1}}\right|^s$$

where k_γ depends on γ but not on s.

Thus, the smaller

$$\left|\frac{a\theta^2+b\theta+c}{\theta^{\gamma+1}}\right|$$

is the better the solution will be and it will become increasingly better as s increases.

Simple calculations give the results in Table 2.2 for a range of γ when $p = \frac{1}{2}$.

It is clear that

$$\left|\frac{a\theta^2+b\theta+c}{\theta^{\gamma+1}}\right|$$

decreases rapidly away from $\gamma = .5$ as γ increases but far more slowly as γ decreases. The former is the case of two real saddle points and the latter two complex conjugate saddle points. Thus the error decreases more sharply as γ decreases from 0.5 than when it increases.

Table 2.2

The value of $\left|\frac{a\theta^2+b\theta+c}{\theta^{\gamma+1}}\right|$ to four decimal places for a range of γ when $p = \frac{1}{2}$

γ	$\left\|\frac{a\theta^2+b\theta+c}{\theta^{\gamma+1}}\right\|$
.1	.9104
.2	.9473
.3	.9757
.4	.9897
.5	1.0000
.6	.9449
.7	.8405
.8	.7074
.9	.5558

We must also consider the results obtained at the point $\gamma = p$. The numerical and analytical results for the Lax-Wendroff scheme with $p = \gamma = \frac{1}{2}$ are given in Table 2.3.

Table 2.3

The Numerical Results for $\gamma = p = \frac{1}{2}$

s	Numerical result	Asymptotic approximation
50	.7324	.7295
100	.7184	.7166
200	.7075	.7063
400	.6989	.6981

Three checks will be carried out to show that (2.16) is of the right form for the error at this point.

Firstly a check is made that the error in the numerical results is of the form $c_1 + c_2 s^{-\frac{1}{3}}$ where c_1 and c_2 are constants.

Let ε_s be the error when the number of time steps is s.

Assuming $\varepsilon_s = c_1 + c_2 s^{c_3}$ + higher order in s gives

$$\frac{\varepsilon_{2s} - \varepsilon_s}{\varepsilon_{4s} - \varepsilon_{2s}} \simeq \frac{2^{c_3} - 1}{2^{c_3}(2^{c_3}-1)} = 2^{-c_3} .$$

The numerical results give for s = 50:

$$\frac{\varepsilon_{100} - \varepsilon_{50}}{\varepsilon_{200} - \varepsilon_{100}} = 1.2844 \Rightarrow c_3 = -.3611$$

s = 100:

$$\frac{\varepsilon_{200} - \varepsilon_{100}}{\varepsilon_{400} - \varepsilon_{200}} = 1.2674 \Rightarrow c_3 = -.3419$$

and so $c_3 \simeq -\frac{1}{3}$ as required.

Secondly a check is made that the solution tends to $\frac{2}{3}$ as $s \to \infty$. (That is, that $c_1 = \frac{1}{3}$.) We can use the fact that $\varepsilon_s = c_1 + c_2 s^{-\frac{1}{3}}$ to extrapolate the numerical results given in Table 2.3. This gives the results in Table 2.4 which are tending towards $\frac{2}{3}$ as required.

Table 2.4

Extrapolated Numerical Results

s	Numerical results		Extrapolated numerical results
50	.7324		
		>	.6645
100	.7184		
		>	.6656
200	.7075		
		>	.6658
400	.6989		

Finally it is possible to estimate c_2 from the numerical results and compare it with the analytical value.

The numerical estimates of c_2 are given in Table 2.5.

Table 2.5

Numerical Estimates of c_2

s	c_2 (numerical estimate)
50	-.2420
100	-.2401
200	-.2380
400	-.2375

The analytical value of c_2 is -.2316 which is close to the numerical estimates.

The above results and the discussion of them can be summarised as follows.

The analytical estimates of the error in the Lax-Wendroff scheme when applied to the test problem (2.7) are supported by the numerical results.

The error oscillates when $\gamma < p$ (that is, $x<t$ and the true solution is $U(x,t)=0$) and this is connected qualitatively with complex saddle points, whereas a monotonically changing error is linked to real saddle points.

The rate at which the error decreases away from the line $x = t$ depends upon how quickly $\left|\frac{a\theta^2+b\theta+c}{\theta^{\gamma+1}}\right|$ diminishes which itself is dependent on how rapidly the saddle points move away from the unit circle.

Finally, the results obtained at the point $\gamma = p$ agree both quantitatively and qualitatively with numerical results. The point $\gamma = p$ corresponds to $x = t$ and so this is the point where the true solution jumps from $U(x,t) = 0$ to $U(x,t) = 1$. It is possible that better results close to $\gamma = p$, when the saddle points are close to poles of $g(z)$, could be obtained by using some weighted average of the result at $\gamma = p$ (when the pole of $g(z)$ is taken into account) and the result at γ close to p (when the pole of $g(z)$ is ignored).

§2.5 Other Schemes applied to the Test Problem

The analysis given in section 2.3 can be readily extended to schemes other than the Lax-Wendroff method. For example, it is easy to adapt the approach to the Wendroff implicit formula (also known as the "Box Scheme").

When applied to the test problem, equation (2.7), this scheme gives the formula

$$u_{r,s+1} = u_{r-1,s} + \frac{(1-p)}{(1+p)} (u_{rs} - u_{r-1,s+1})$$

where u_{rs} and p are as defined for (2.8).

Putting $f_s(z) = \sum_{r=-\infty}^{\infty} u_{rs} z^r$ as before gives

$$f_{s+1}(z) = \left[\frac{(1+p)z+(1-p)}{(1+p)+z(1-p)}\right]^s \frac{1}{1-z} \quad .$$

It is possible to proceed in exactly the same way as before but in this case $\gamma < p$ corresponds to two real distinct saddle points and $\gamma > p$ to a complex conjugate pair of saddle points.

Again the real saddle points correspond to a smooth error and the conjugates to an oscillating error and again the smooth error is far smaller than the oscillating one. (See Table 2.6 for results.)

This now suggests another use for this way of finding the error in finite difference approximations; namely the adaptation of finite difference schemes so that the error is smooth rather than oscillatory. The analysis of the previous section together with the numerical evidence of the Lax-Wendroff scheme and Wendroff's implicit formula suggest that the error will be smooth if and only if the saddle points are real.

We therefore try to modify the Lax-Wendroff scheme so that it has real saddle points.

If we take

$$u_{r,s+1} = u_{rs} - \frac{p}{2}(u_{r+1,s} - u_{r-1,s}) + \mu(u_{r+1,s} - 2u_{r,s} + u_{r-1,s})$$

where we are free to choose μ (when $\mu = \frac{p^2}{2}$ this is just the Lax-Wendroff scheme), then stability considerations give the following restriction on μ

$$\frac{1 - (1-p^2)^{\frac{1}{2}}}{2} < \mu < 1 + \frac{(1-p^2)^{\frac{1}{2}}}{2}$$

and so imposing stability for $0 < \gamma < 1$ together with the requirement of real saddle points gives $\mu = \frac{p}{2}$.

Unfortunately this is just the standard explicit formula

$$u_{r,s+1} = (1-p)u_{rs} + pu_{r-1,s}.$$

It does indeed have smooth errors but the order of the scheme is lower.

Even when second order schemes with distinct real saddle points have been found the results near to the discontinuity have been unsatisfactory as excessive 'smearing' of the solution has taken place.

Clearly we need to find a scheme with real saddle points and such that the saddle points move away quickly from the unit circle as γ moves away from p but it is not obvious how the coefficients of a difference scheme may be chosen to satisfy all these criteria.

§2.6 A Second Order version of the test problem and a scheme to solve it

The above sections have dealt with one test problem, namely equation (2.7) and analysed the error when various finite difference schemes are used to solve it. In the remaining parts of this chapter the saddle point method is extended to two other test problems with discontinuities. The first of these problems is the second order version of (2.7). That is,

$$v_t = w_x$$

$$w_t = v_x$$

where v and w are functions of x and t only,

with initial conditions

$$v(x,0) = 0 \quad \text{for all } x$$

$$w(x,0) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \,. \end{cases} \tag{2.17}$$

This is of course the wave equation $v_{tt} = v_{xx}$ expressed as two first order equations. The exact solution to this problem is

$$v(x,t) = \begin{cases} 0 & x < -t \\ \frac{1}{2} & -t \leq x < t \\ 0 & x \geq t \end{cases}$$

$$w(x,t) = \begin{cases} 0 & x < -t \\ \frac{1}{2} & -t \leq x < t \\ 1 & x \geq t. \end{cases}$$

In the next section I shall analyse the error when the Lax-Wendroff formula is used to solve the problem but first I shall consider the following scheme

$$\frac{1}{k}[v_{r,s+1} - \tfrac{1}{2}(v_{r+1,s} + v_{r-1,s})] = \frac{1}{2h}[w_{r+1,s} - w_{r-1,s}] \tag{2.18}$$

$$\frac{1}{k}[w_{r,s+1} - \tfrac{1}{2}(w_{r+1,s} + w_{r-1,s})] = \frac{1}{2h}[v_{r+1,s} - v_{r-1,s}] \tag{2.19}$$

where h, k, v_{rs} and w_{rs} follow the notation of earlier sections of the chapter.

Proceeding in the same fashion as in the first order case, let

$$f_s(z) = \sum_{r=-\infty}^{\infty} v_{rs} z^r \quad \text{and} \quad g_s(z) = \sum_{r=-\infty}^{\infty} w_{rs} z^r .$$

Let $p = k/h$ and impose $0 < p < 1$ for stability.

Multiply both (2.18) and (2.19) and sum over r from $-\infty$ to ∞ to obtain the equations

$$f_{s+1}(z) = \tfrac{1}{2}(\tfrac{1}{z}+z) f_s(z) + \tfrac{p}{2}(\tfrac{1}{z}-z) g_s(z)$$

$$g_{s+1}(z) = \tfrac{1}{2}(\tfrac{1}{z}+z) g_s(z) + \tfrac{p}{2}(\tfrac{1}{z}-z) f_s(z).$$

Now defining $\underline{F}_s(z) = \begin{pmatrix} f_s(z) \\ g_s(z) \end{pmatrix}$ and

$$A = \tfrac{1}{2} \begin{bmatrix} \frac{1}{z}+z & p(\frac{1}{z}-z) \\ p(\frac{1}{z}-z) & \frac{1}{z}+z \end{bmatrix}$$ we have

$$\underline{F}_{s+1}(z) = A\underline{F}_s(z).$$

That is, $\underline{F}_s(z) = A^s F_0(z)$ (2.20)

and from the initial conditions,

$$\underline{F}_0(z) = \begin{pmatrix} 0 \\ \frac{1}{1-z} \end{pmatrix} .$$

The eigenvalues of A are

$$\lambda_1 = \tfrac{1}{2}((1-p)\tfrac{1}{z}+(1+p)z)$$

$$\lambda_2 = \tfrac{1}{2}((1+p)\tfrac{1}{z}+(1-p)z) \qquad (2.21)$$

and it is easily shown that

$$A = R\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}R^{-1}$$

where

$$R = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix} .$$

Hence

$$A^s = R\begin{pmatrix}\lambda_1^s & 0\\ 0 & \lambda_2^s\end{pmatrix}R^{-1}$$

$$= \tfrac{1}{2}\begin{pmatrix}\lambda_1^s+\lambda_2^s & -\lambda_1^s+\lambda_2^s\\ -\lambda_1^s+\lambda_2^s & \lambda_1^s+\lambda_2^s\end{pmatrix}$$

and since v_{rs} and w_{rs} are the coefficients of z^r in $f_s(z)$ and $g_s(z)$ respectively, from (2.20)

$$\begin{aligned} v_{rs} &= \text{coefficient of } z^r \text{ in } \tfrac{1}{2}(-\lambda_1^s+\lambda_2^s)\frac{1}{1-z}\\ w_{rs} &= \text{coefficient of } z^r \text{ in } \tfrac{1}{2}(\lambda_1^s+\lambda_2^s)\frac{1}{1-z} . \end{aligned} \qquad (2.22)$$

To calculate both v_{rs} and w_{rs}, then, we need to find the coefficient of z^r in $\frac{\lambda^s}{1-z}$ where λ is of the form $az + \frac{b}{z}$ and since $0 < p < 1$ for stability then from (2.21) $a > 0$ and $b > 0$ for both λ_1 and λ_2.

This can be considered, as in the first order case, as a contour integral.

It is required to find,

$$\int_C \frac{\lambda^s}{z^{r+1}(1-z)}\,dz = \int_C \frac{(az+b/z)^s}{z^{\gamma s+1}(1-z)}\,dz$$

where as before $\gamma = \frac{r}{s}$

$$= \int_C g(z)e^{sh(z)}dz$$

where $g(z) = \frac{1}{z(1-z)}$,

$$h(z) = \log(az^2+b) - (\gamma+1)\log z.$$

The saddle points, $\theta_\pm$, of $h(z)$ are defined by $h'(z) = 0$. That is,

$$z^2a(1-\gamma) - b(\gamma+1) = 0$$

$$\theta_\pm = \pm\sqrt{\frac{b(\gamma+1)}{a(1-\gamma)}} .$$

These are real and distinct provided $-1 < \gamma < 1$ and real with $\theta_+ = \theta_-$ when $\gamma = -1$.

Since the calculation of $v_{r,s+1}$ and $w_{r,s+1}$ involves values at $((r+1)h,sk)$ and $((r-1)h,sk)$ only, the finite difference formulae (2.18) and (2.19) give exact results for the given initial conditions when $-s > r$ or $r \geq s$. That is, when $\gamma < -1$ or $\gamma \geq 1$. Therefore, it is only necessary to calculate an estimate of the error when $-1 \leq \gamma < 1$.

It is interesting that the contour map of $\mathrm{Re}(h(z))$ when γ is not in this range (shown in Figure 2.7) does not have a contour which passes through either saddle point and encircles the origin. This is linked with the fact that the error is zero in this case.

I shall restrict my attention to the case $-1 \leq \gamma < 1$. The contour map of $\mathrm{Re}(h(z))$ for this case is given for λ_2 in Figure 2.6 and a suitable path for integration, crossing the saddle points as required, is marked. (The contour map for λ_1

is very similar.)

We must consider whether the contour contains the pole of $g(z)$ at $z = 1$. From the contour map it can be seen that it does so if and only if $\theta_+ > 1$.

That is,

$$\frac{b(\gamma+1)}{a(1-\gamma)} > 1$$

or equivalently,

$$\gamma > \frac{a-b}{a+b} . \tag{2.23}$$

In the case of λ_2, $a = \frac{1}{2}(1-p)$, $b = \frac{1}{2}(1+p)$ and so equation (2.23) gives

$$\theta_+ > 1 \iff \gamma > -\frac{2p}{2} = -p. \tag{2.24}$$

The pole at $z = 1$ is, therefore, included for $\gamma > -p$ but excluded for $-1 \leq \gamma < -p$. For the case λ_2, $a = \frac{1}{2}(1+p)$, $b = \frac{1}{2}(1-p)$ and so

$$\theta_+ > 1 \iff \gamma > \frac{2p}{2} = p. \tag{2.25}$$

Thus, in this case, the pole is included for $\gamma > p$ but excluded for $-1 \leq \gamma < p$.

It should be noted that the saddle point at $z = \theta_+$ is crossed 'upwards' and so its contribution is that given by (2.6) but that at $z = \theta_-$ is crossed in a 'downwards' direction and so its contribution is the negative of the expression (2.6). Using (2.6) together with the fact that

$$\omega = \exp(\tfrac{1}{2}\pi i - \tfrac{1}{2}\arg h''(\theta))$$

$$= \exp(\tfrac{1}{2}\pi i) \text{ since } h''(\theta) \text{ is real}$$

$$= i$$

gives

$$\frac{1}{2\pi i} \int \frac{\lambda^s}{z^{r+1}(1-z)}\, dz$$

$$\sim \frac{1}{2\pi i} \sqrt{\frac{2\pi}{s}}\, i\, \frac{1}{\sqrt{|h''(\theta)|}} \left\{ \left(\frac{a\theta_+^2+b}{\theta_+^{\gamma+1}}\right)^s \frac{1}{\theta_+(1-\theta_+)} - \left(\frac{a\theta_-^2+b}{\theta_-^{\gamma+1}}\right)^s \frac{1}{\theta_-(1-\theta_-)} \right\}$$

where $h''(\theta) = h''(\theta_+) = h''(\theta_-)$

since h'' is a function of θ^2 and we know that $\theta_+^2 = \theta_-^2$.

This expression can be further simplified using the fact that $\theta = \theta_+ = -\theta_-$ and $h''(\theta) = \frac{a}{b}(1-\gamma)^2$ to give

$$\frac{1}{2\pi i} \int_C \frac{\lambda^s}{z^{r+1}(1-z)}\, dz$$

$$\sim \sqrt{\frac{b}{2\pi s a}}\, \frac{1}{(1-\gamma)}\, \frac{1}{\theta} \left(\frac{2b}{1-\gamma} \cdot \frac{1}{\theta^{\gamma+1}}\right)^s \left\{ \frac{1}{1-\theta} + \frac{(-1)^{(\gamma+1)s}}{1+\theta} \right\} \qquad (2.26)$$

remembering that $\gamma s = r$ which is an integer.

The residue at the simple pole at $z = 1$ is

$$-\left(\frac{a+b}{1^{\gamma+1}}\right)^s \frac{1}{1} = -(a+b)^s = -1. \qquad (2.27)$$

Thus from (2.22), (2.24), (2.25), (2.26), (2.27) we can calculate v_{rs} and w_{rs} for $-1 < \gamma < 1$.

The case $\gamma = -1$ is easily dealt with since $h(z) = \log(az^2+b) - (\gamma+1)\log z = \log(az^2+b)$ when $\gamma = 1$.

Thus

$$\int_C g(z)e^{sh(z)}dz = \int_C \frac{(az^2+b)^s}{z(1-z)}\,dz$$

and now the singularity at the origin is a simple pole and hence the residue is given by b^s.

Substituting for b from (2.21),(2.22) gives

$$v_{-s,s} = \tfrac{1}{2}\left(-\frac{(1-p)^s}{2^s}+\frac{(1+p)^s}{2^s}\right)$$

$$w_{-s,s} = \tfrac{1}{2}\left(\frac{(1-p)^s}{2^s}+\frac{(1+p)^s}{2^s}\right)$$

and it is easy to show by induction that this is exactly the answer obtained when (2.18) and (2.19) are used to solve (2.17).

The results for other values of γ are given for s = 50, 100 and 200 with t = 1 and p = .25 in Tables 2.7 and 2.8. Table 2.7 gives the numerical errors and the analytical estimate of those errors for v and Table 2.8 the errors for w.

In both cases agreement is good except for values of γ close to $-\frac{1}{4}$ or $\frac{1}{4}$. This, of course, is when $\gamma = \pm p$ and so the saddle point coincides with the pole of g(z) at z = 1 for either the λ_1 integral or the λ_2 one. So, as in the previous problem, a special evaluation which takes account of the residue at z = 1 could be used to improve the accuracy of the asymptotic estimate at these points.

§2.7 The Second Order Lax-Wendroff Scheme

In this section a slightly more complicated difference scheme, the Lax-Wendroff method, is used to solve (2.17). The basic approach to find the asymptotic estimate of the error is very similar to that of the previous section and will

not be dealt with in detail here. The special difficulty of this scheme lies in the choice of the sign of the contribution of a saddle point. As the saddle points move, careful thought must be given as to whether they are crossed in an 'upwards' or a 'downwards' direction. This is dealt with in detail after a brief outline of the implementation of the saddle point method to this scheme.

The Lax-Wendroff scheme is given by the following two equations

$$v_{r,s+1} = v_{r,s} + \frac{p}{2}(w_{r+1,s} - w_{r-1,s}) + \frac{p^2}{2}(v_{r+1,s} - 2v_{r,s} + v_{r-1,s}) \quad (2.28)$$

$$w_{r,s+1} = w_{r,s} + \frac{p}{2}(v_{r+1,s} - v_{r-1,s}) + \frac{p^2}{2}(w_{r+1,s} - 2w_{r,s} + w_{r-1,s}) \quad (2.29)$$

where again $p = \frac{k}{h}$ and for stability $0 < p < 1$.

With $f_s(z)$, $g_s(z)$ and $\underline{F}_s(z)$ as defined in the previous section, multiplying (2.28) and (2.29) by z^r and summing over r from $-\infty$ to ∞ gives

$$\underline{F}_s(z) = A^s \underline{F}_0(z) \text{ where } \underline{F}_0(z) = \begin{pmatrix} 0 \\ \frac{1}{1-z} \end{pmatrix}$$

$$\text{and } A = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$$

$$\text{with } a = \tfrac{1}{2}\frac{p^2}{z} + (1-p^2) + \tfrac{1}{2}p^2 z$$

$$b = \tfrac{1}{2}p\left(\frac{1}{z} - z\right).$$

The eigenvalues of A are $\lambda_1 = a + b$, $\lambda_2 = a - b$. That is,

$$\lambda_1 = \tfrac{1}{2}\frac{p}{z}(p+1) + (1-p^2) + \tfrac{1}{2}pz(p-1)$$

$$\lambda_2 = \tfrac{1}{2}\frac{p}{z}(p-1) + (1-p^2) + \tfrac{1}{2}pz(p+1)$$

and,

$$A^s = \tfrac{1}{2}\begin{pmatrix} \lambda_1^s+\lambda_2^s & \lambda_1^s-\lambda_2^s \\ \lambda_1^s-\lambda_2^s & \lambda_1^s+\lambda_2^s \end{pmatrix} .$$

Thus, it is required to calculate

$$\int_C \frac{\lambda^s dz}{z^{r+1}(1-z)} = \int_C \frac{(\frac{m}{z}+n+\ell z)^s}{z^{\gamma s+1}(1-z)}$$

$$\text{where } \gamma = \frac{r}{s}$$

$$m = \tfrac{1}{2}p(p\pm 1)$$

$$n = 1 - p^2$$

$$\ell = \tfrac{1}{2}p(p\mp 1).$$

The choice of sign in n and ℓ depends on whether λ_1 or λ_2 is under consideration.

The integral can be written as $\int_C g(z)e^{sh(z)}dz$

where $g(z) = \frac{1}{z(1-z)}$

and $h(z) = \log(m+nz+\ell z^2) - (\gamma+1)\log z.$

After some algebraic manipulation we have the following results.

The saddle points are at

$$\theta_{\pm} = \frac{\gamma n \pm ((\gamma^2-p^2)(1-p^2))^{\frac{1}{2}}}{2\ell(1-\gamma)} \qquad (2.30)$$

and the function h(z) has poles $P_{\pm}$ at

$$P_{\pm} = \frac{-n\pm(1-p^2)^{\frac{1}{2}}}{2\ell} .$$

As before we are only concerned with $-1 \leq \gamma < 1$ since the scheme is errorless outside these bounds (for the test problem).

Since $0 < p < 1$ for stability, $\ell m < 0$ for both λ_1 and λ_2, and $n > 0$. Hence the poles are always of different signs.

It can be shown that the following two equalities are satisfied when $\gamma > p$.

For λ_1:

$$\theta_+ < \theta_- < P_+ < 0 < P_-$$

and for λ_2:

$$P_- < 0 < P_+ < \theta_- < \theta_+.$$

When $\gamma < -p$ the corresponding two inequalities are

for λ_1:

$$P_+ < 0 < \theta_+ < \theta_- < P_-$$

and for λ_2:

$$P_- < \theta_- < \theta_+ < 0 < P_+.$$

When $-p < \gamma < p$ the saddle points are complex.

Some contour maps are shown in Figures 2.8-2.10 with suitable integration paths.

$g(z)$ has a pole at $z = 1$ which is included in the contour for λ_1 if and only if $\gamma > -p$ and for that of λ_2 if and only if $\gamma > p$.

It now remains to consider the direction in which the saddle point is crossed (so that the correct sign can be assigned to the contribution).

When $\gamma > p$ or $\gamma < -p$ the situation is quite straightforward; if the saddle point is negative it is crossed in a downwards direction and the contribution is negative, likewise a positive saddle point results in a positive contribution. The case of complex conjugate saddle points causes a little more difficulty.

The case of λ_1 will be considered although exactly the same argument can be used for the λ_2 case. Since $\ell < 0$ for λ_1, $\mathrm{Im}(\theta_+) < 0$. When $\gamma = -p$ there is a double saddle point at $z = 1$ and when $\gamma = p$ there is a double saddle point at $\frac{p+1}{p-1}$. As γ changes from $-p$ to p, θ_+ moves from 1 to $\frac{p+1}{p-1}$ with $\mathrm{Im}(\theta_+) < 0$.

The axis of the saddle point is determined by $\arg(h''(\theta_+))$ and, in fact, is given by $\frac{\pi}{2} - \frac{1}{2}\arg h''(\theta_+)$. $h''(\theta_+)$ is a continuous function of γ and so the axis of the saddle point changes in a continuous way from lying in the range $(0,\pi/4)$ when $\gamma = -p$ to lying in the range $(-\pi/2, \frac{-3\pi}{4})$ when $\gamma = p$, moving through $(0, -\pi/2)$ as γ moves from $-p$ to p.

Initially, (that is when γ is slightly greater than $-p$) the θ_+ saddle point is crossed in the direction of $\mathrm{Re}(z)$ increasing and so the positive sign of the contribution is taken.

Since $\mathrm{Re}(h(z))$ is a continuous function of γ the contours are transformed in a continuous way as γ changes and so the direction of the path across the saddle point must change continuously. This fact, combined with the knowledge of the direction of the axis of the saddle point indicates that although the contribution is positive initially, as soon as

$\mathrm{Im}(h''(\theta_+))$ becomes negative the contribution changes sign and remains negative even though eventually $\mathrm{Im}h''(\theta_+)$ becomes positive again (as γ approaches p). This slightly unexpected conclusion is supported by the contour map in Figure 2.8.

An entirely analagous argument can be used to show that the contribution of the θ_+ saddle point in the λ_2 case is positive unless $\mathrm{Im}h''(\theta_+) > 0$ and $\gamma < 0$.

Once the sign of the contribution has been determined the procedure of the previous section can be used to calculate the asymptotic estimate of the error. Table 2.9 gives results at a range of values of γ for both v and w with s = 50 and p = .25. As with the method of section 2.6 the asymptotic estimate of the error is less good when close to $\gamma = \pm p$ as no account has been taken of the poles of g(z) at these points.

As in the first order situation, and for entirely the same reasons, real saddle points give a monotonically changing error whereas complex conjugate ones correspond to an oscillating error.

§2.8 A discontinuous problem

The final problem to be considered is one in which not only are the initial conditions, and hence the solution, discontinuous but also the partial differential equation itself has a discontinuity.

The problem to be considered is

$$u_t + \lambda u_x = 0$$

$$\text{where } \lambda = \begin{cases} \mu \neq 0 & \text{for } x \leq t \\ 0 & x > t \end{cases}$$

with initial conditions

$$u(x,0) = \begin{cases} 1 & x \leq 0 \\ 0 & x > 0 \end{cases} . \qquad (2.31)$$

This has the exact solution

$$u(x,t) = \begin{cases} 1 & x \leq \mu t \\ \text{undefined} & \mu t < x \leq t \\ 0 & x > t \end{cases}$$

and so the restriction $\mu \leq 1$ must be imposed for a single valued solution.

This problem is linked to the equation of motion of the one dimensional, unsteady flow of an incompressible fluid, $u_t + uu_x = 0$ which has characteristics $\frac{dx}{dt} = u$ along which u is constant.

In this case when the initial conditions are

$$u(x,0) = \begin{cases} 1 & x \leq 0 \\ 0 & x > 0 \end{cases}$$

these characteristics become the lines such that $\frac{dx}{dt} = 1$ in the region $x - t \leq 0$ and such that $\frac{dx}{dt} = 0$ when $x - t > 0$ and thus the solution is $u \equiv 1$ when $x \leq t$ and $u \equiv 0$ when $x > t$.

So if u is replaced by its exact solution in these two regions the equation of motion reduces to

$$u_t + u_x = 0 \quad x \leq t$$
$$u_t = 0 \quad x > t$$

which is (2.31) with $\mu = 1$.

A simple finite difference approximation to (2.31) will be analysed since the analysis of the Lax-Wendroff method becomes very complicated.

Also, because it becomes complicated to handle both discontinuities on $x = \mu t$ and $x = t$, it is convenient to use only a fixed ration k/h and to simplify the algebra, $p = k/h = \frac{1}{2}$ is chosen.

The finite difference approximation to be considered is

$$u_{r,s+1} = (1-\tfrac{\mu}{2})u_{rs} + \tfrac{\mu}{2}u_{r-1,s} \quad \text{when } r \leq \tfrac{1}{2}s \tag{2.32}$$

$$u_{r,s+1} = u_{rs} \qquad r > \tfrac{1}{2}s \ . \tag{2.33}$$

This gives

$$u_{rs} = \begin{cases} 1 & r \leq 0 \\ 0 & r \geq \tfrac{1}{2}s \quad s \neq 0 \\ & r > 0 \quad s = 0 \end{cases} \tag{2.34}$$

which is the exact solution at these points.

It remains to analyse the error when $0 < r < \frac{1}{2}s$. For these values of r and s, (2.32) is the appropriate solution.

Let

$$G_n(z) = \sum_{r=1}^{\infty} u_{r,2r+n} z^r \ .$$

Then,

$$G_0(z) = \sum_{r=1}^{\infty} u_{r,2r} = 0 \quad \text{from (2.34)}$$

and $G_n(0) = 0$.

(2.32) can be rewritten as

$$u_{r,2r+n+1} = (1-\tfrac{\mu}{2})u_{r,2r+n} + \tfrac{\mu}{2}u_{r-1,2r+n} \qquad (2.36)$$

where $n = s - 2r$ and $n \geq 0$ since $2r \leq s$.

Multiplying (2.36) by z^r and $\sum_{r=1}^{\infty}$ gives

$$G_{n+1}(z) = (1-\tfrac{\mu}{2})G_n(z) + \tfrac{\mu}{2}u_{0,n+2}z + \tfrac{\mu}{2}zG_{n+z}(z).$$

The general solution to this difference equation is

$$G_n(z) = A(\mu,z)m_+^n + B(\mu,z)m_-^n - \frac{z}{z-1} \qquad (2.37)$$

where

$$m_{\pm} = \frac{1 \pm (1-\mu z(2-\mu))^{\frac{1}{2}}}{\mu z}$$

$$G_0(z) = 0 \Rightarrow A(\mu,z) + B(\mu,z) - \frac{z}{z-1} = 0 \qquad (2.38)$$

$$G_n(0) = 0 \Rightarrow \lim_{z\to 0}(A(\mu,z)m_+^n + B(\mu,z)m_-^n) = 0.$$

Since

$$m_+^n \to (\frac{1}{\mu z})^n \text{ as } z \to 0$$

we take $A(\mu,z) = 0$.

Then (2.37) gives $B(\mu,z) = \frac{z}{z-1}$

and

$$G_n(z) = \frac{z}{z-1}(m_-^n - 1) \qquad (2.39)$$

u_{rs} is the coefficient of z^r in $G_{s-2r}(z)$.

If we make the substitution $w = \frac{1}{z}$, then $u_{r,s}$ is the coefficient of w^{-r} in $F_{s-2r}(w)$ where

$$F_{s-2r}(w) = G_{s-2r}(z)$$

$$= \frac{1/w}{1/w-1}(m^{s-2r}-1)$$

$$= \frac{1}{1-w}(m^{s-2r}-1)$$

$$\text{and } m = \frac{1-(1-\frac{\mu}{w}(2-\mu))^{\frac{1}{2}}}{\mu/w}$$

$$= \frac{w}{\mu}(1-(1-\frac{\mu}{w}(2-\mu))^{\frac{1}{2}}).$$

Since $\frac{1}{1-w} = 1 + w + w^2 + \ldots$, and therefore has no negative powers of w we can ignore thispart of $F_{s-2r}(w)$ and just find the coefficient of w^{-r} in $\frac{m^{s-2r}}{1-w}$.

So, as before,

$$u_{rs} + \begin{matrix}\text{residues}\\ \text{within C}\end{matrix} = \frac{1}{2\pi i}\int_C \left(\frac{m^{s-2r}}{1-w}\right)\frac{1}{w^{-r+1}}\,dw$$

$$= \frac{1}{2\pi i}\int_C \frac{1}{1-w}w^{s-r-1}\left(\frac{1-(1-\frac{\mu}{w}(2-\mu))^{\frac{1}{2}}}{\mu}\right)^{s-2r}dw$$

$$= \frac{1}{2\pi i}\int_C \frac{w^{s(1-\gamma)}}{w(1-w)}\,\psi^{s(1-\gamma)}\,dw$$

$$\text{where } \gamma = \frac{r}{s},\ \psi = \frac{1-(1-\frac{\mu}{w}(2-\mu))^{\frac{1}{2}}}{\mu}$$

$$= \frac{1}{2\pi i}\int_C g(w)e^{sh(w)}\,dw$$

$$\text{where } g(w) = \frac{1}{w(1-w)} \qquad (2.40)$$

$$h(w) = (1-\gamma)\log w + (1-2\gamma)\log\psi.$$

So

$$\int_C g(w)e^{sh(w)}dw = \sqrt{\frac{2\pi}{s}}\alpha\,|h''(\theta)|^{-\frac{1}{2}}g(\theta)e^{sh(\theta)} \qquad (2.41)$$

$$\text{where } \alpha = \exp(\tfrac{1}{2}\pi i - \tfrac{1}{2}i\arg h''(\theta))$$

$$= \exp(\tfrac{1}{2}\pi i) = i,$$

and at the real saddle point $\theta = \frac{\mu(2-\mu)}{4\gamma(1-\gamma)}$, $\psi(\theta) = \frac{2\gamma}{\mu}$.

$$(0<\mu<1,\ 0<\gamma<\tfrac{1}{2})$$

The residue at 1 is -1 and this pole is included in the contour if, and only if, $\mu > 2\gamma$.

$$h''(\theta) = \frac{\gamma^3(1-\gamma)^3 16}{\mu^2(2-\mu)^2(1-2\gamma)^2} \qquad (2.42)$$

and so using (2.40), (2.41) and (2.42) the coefficient of u_{rs} can be calculated except when $\mu = 2\gamma$ and the contour passes through the pole of g(z).

Results for the error in the finite difference scheme and the asymptotic estimate of this error are given at $\mu = .4$ for $s = 50$, 100 and 200 in Table 2.10. The errors at $s = 50$ for $\mu = .6$ and $\mu = .8$ are shown in Table 2.11.

Normally the numerical solution tends to the exact solution as the step size tends to zero. Although the exact solution is undefined for $\mu t < x \leq t$, that is for $\frac{\mu}{2} < \gamma \leq \frac{1}{2}$, the numerical solution still tends to a limit, namely zero. The 'error' in the numerical solution is defined to be the difference between this limit and the numerical solution and it is this quantity which the asymptotic analysis provides an

estimate for.

As expected the error and the asymptotic estimate of it are poor when γ is close to the discontinuity at $x = \mu t$ but there is good agreement between the two as x and t move away from the line $x = \mu t$. The error is monotonic as expected since the asymptotic evaluation uses one real saddle point.

§2.9 Summary

The aim of this chapter has been to analyse the error which occurs when a finite difference formula is used to calculate the solution to a problem which has discontinuities.

The analysis provides both qualitative and quantitative information about the error. For example, the existence of complex conjugate saddle points through which the relevant contour passes indicates that the error is oscillating in nature.

An empirical observation is that the poles of g(z) (that is, the part of the integrand which is independent of s) are linked to discontinuities in the solution. For the values of x and t when the saddle point of h(z) coincides with a pole of g(z) are points where the solution has a discontinuity.

Information of a quantitative nature is provided by terms in the error expansion such as the $\frac{a\theta^2+b\theta+c}{\theta^{\gamma+1}}$ term of (2.12) or (2.11) which indicates how quickly the error decreases on either side of a discontinuity.

This qualitative and quantitative knowledge can be used not only to calculate the error for given finite difference schemes but also to design difference schemes with a specific

type of behaviour.

In this chapter the saddle point method has been used to evaluate the error actually at a discontinuity. The method used, namely using a combination of the residue theorem and the saddle point method to find the contribution of a saddle point as the saddle point approaches a pole of g(z), can be used in any situation where the saddle point method could be used if the pole at g(z) did not exist. That is, the pole at g(z) does not cause any insurmountable analytical difficulties although it must, of course, be taken into account.

The applications have covered a range of problems, albeit of a very simple nature, with different types of discontinuities. Both first order and second order equations have been considered and also problems where not only the starting condition (and hence the solution) but also the governing partial differential equations, have discontinuities.

TABLE 2.1

	s = 50		s = 100		s = 200	
γ	(a)	(b)	(a)	(b)	(a)	(b)
.24	-.1841E-1	-.1777E-1	.8535E-3	.8096E-3	-.1273E-4	-.1295E-4
.30	.4814E-1	.4573E-1	.5505E-2	.5113E-2	-.3024E-3	-.3136E-3
.36	-.6629E-1	-.5331E-1	-.1889E-1	-.2188E-1	-.9869E-2	-.9705E-2
.42	-.1100	-.1761	-.1380	-.1265	.4454E-1	.3968E-1
.48	.5500	.6622	.4008	.3403	.1586	.3797E-1
.52	-.1435	-.3261	-.1033	-.1828	-.5580E-1	-.8130E-1
.58	-.1214E-1	-.1582E-1	-.1312E-2	-.1530E-2	-.1859E-4	-.2022E-4
.64	-.4278E-3	-.4965E-3	-.2534E-5	-.2747E-5	-.1140E-9	-.1189E-9
.70	-.6230E-5	-.6896E-5	-.7807E-9	-.8236E-9	0.0000	-.1661E-16
.76	-.3529E17	-.3814E-7	-.3517E-13	-.3677E-13	0.0000	-.4834E-25

(a) TRUE ERROR AND (b) ASYMPTOTIC ESTIMATE OF ERROR AT t=1

WHEN THE LAX-WENDROFF SCHEME IS USED ON (2.7) WITH p=.5

TABLE 2.6

	s = 50		s = 100		s = 200	
γ	(a)	(b)	(a)	(b)	(a)	(b)
.24	.1605E-7	.1634E-7	.1161E-14	.1172E-14	.8484E-29	.8523E-29
.30	.4283E-5	.4394E-5	.8127E-10	.8237E-10	.4065E-19	.4093E-19
.36	.4326E-3	.4525E-3	.7674E-6	.7863E-6	.3315E-11	.3357E-11
.42	.1682E-1	.1866E-1	.9860E-3	.1046E-2	.4507E-5	.4656E-5
.48	.2385	.3833	.1428	.1988	.6091E-1	.7561E-1
.52	-.2595	-.2069	-.1054	.3299E-1	.1358	.2575
.58	.2436	.2627	-.1855	-.2038	.3022E-1	.2354E-1
.64	-.1658	-.1697	-.1286	-.1344	.3486E-1	.3323E-1
.70	.5845E-1	.5397E-1	-.6224E-1	-.6115E-1	-.6736E-1	-.6752E-1
.76	.8572E-1	.9135E-1	.8850E-1	.8966E-1	-.1175E-1	-.1239E-1

(a) TRUE ERROR AND (b) ASYMPTOTIC ESTIMATE OF ERROR AT t=1

WHEN THE WENDROFF IMPLICIT SCHEME IS USED ON (2.7) WITH p=.5

TABLE 2.7

	s = 50		s = 100		s = 200	
γ	(a)	(b)	(a)	(b)	(a)	(b)
-.5	.1555E-1	.1863E-1	0,2784E-2	0.3015E-2	.5930E-4	.6200E-4
-.4	.8577E-1	.1160	0.3627E-1	0.4414E-1	.8022E-2	.9045E-2
-.3	.1808	.5107	0.1713	.3564	.1283	.2203
-.2	-.1513	-.4856	-.1330	-.3211	-.1051	-.1989
-.1	-.7151E-1	-.1046	-.2537E-1	-.3424E-1	-.6217E-2	-.7456E-2
0.	-.3693E-1	-.4495E-1	-.5635E-2	-.6332E-2	-.1646E-3	-.1777E-3
.1	-.7151E-1	-.1046	-.3766E-1	-.4657E-1	-.8925E-2	-.1017E-1
.2	-.2046	-.5392	-.1686	-.3567	-.1272	-.2210
.3	.1808	.5107	.1348	.3198	.1058	.1977
.4	.5284E-1	.8284E-1	.2369E-1	.3153E-1	.5442E-2	.6461E-2
.5	.1555E-1	.1863E-1	.1448E-2	.1675E-2	.3134E-4	.3445E-4

(a) TRUE ERROR AND (b) ASYMPTOTIC ESTIMATE OF ERROR AT t=1 FOR v

WHEN EQUATION (2.18) AND (2.19) ARE USED TO SOLVE (2.17) WITH p=.25

TABLE 2.8

	s = 50		s = 100		s = 200	
γ	(a)	(b)	(a)	(b)	(a)	(b)
-.5	-.1555E-1	.1863E-1	.2784E-2	.3015E-2	.5930E-4	.6200E-4
-.4	.8577E-1	.1160	.3627E-1	.4414E-1	.8022E-2	.9045E-2
-.3	.1809	.5107	.1713	.3564	.1283	.2203
-.2	-.1503	.4845	-.1330	-.3211	-.1051	-.1989
-.1	-.6554E-1	-.9789E-1	-.2509E-1	-.3395E-1	-.6217E-2	-.7456E-2
0.	.1118E-1	.1124E-1	.1580E-2	.1583E-2	.4631E-4	.4441E-4
.1	.6554E-1	.9789E-1	.3753E-1	.4642E-1	.8928E-2	.1017E-1
.2	.2043	.5388	.1686	.3567	.1272	.2210
.3	-.1809	-.5107	-.1348	-.3198	-.1058	-.1977
.4	-.5284E-1	-.8284E-1	-.2369E-1	-.3153E-1	-.5441E-2	-.6461E-2
.5	-.1555E-1	-.1863E-1	-.1448E-2	-.1675E-2	-.3093E-4	-.3445E-4

(a) TRUE ERROR AND (b) ASYMPTOTIC ESTIMATE OF ERROR AT t=1 FOR w WHEN EQUATION (2.18) AND (2.19) ARE USED WITH p=.25 TO SOLVE (2.17)

TABLE 2.9

	v		w	
γ	(a)	(b)	(a)	(b)
-.5	.7089E-7	.5112E17	.7089E-7	.5112E-7
-.4	.1638E-3	.1326E-3	.1638E-3	.1327E-3
-.3	.3962E-1	.4051E-1	.3960E-1	.4053E-1
-.2	.6375E-1	.1227	.6459E-1	.1236
-.1	-.3313E-1	-.3321E-1	-.3842E-1	-.3860E-1
0	-.1731E-1	-.1680E-1	.844E-1	0.1866E-1
.1	-.5847E-1	-.5961E-1	.6010E-1	.6119E-1
.2	-.6175E-1	-.4153E-2	.6244E-1	.4833E-2
.3	.1613E-1	.2734E-1	-.1614E-1	-.2734E-1
.4	.4129E-4	.5609E-4	-.4125E-4	-.5609E-4
.5	.4340E-9	.1558E-7	0.0000	-.1558E-7

(a) TRUE ERROR AND (b) ASYMPTOTIC ESTIMATE OF ERROR AT t=1 FOR v AND w WHEN THE LAX-WENDROFF SCHEME IS USED WITH p=.25 AND s=50 TO SOLVE (2.17)

TABLE 2.10

	s = 50		s = 100		s = 200	
γ	(a)	(b)	(a)	(b)	(a)	(b)
.02	-.2230E-4	-.2593E-4	-.8136E-8	-.8803E-8	-.1318E-14	-.1435E-14
.06	-.1879E-2	-.2153E-2	-.2674E-4	-.2880E-4	-.7009E-8	-.7288E-8
.10	-.2522E-1	-.3089E-1	-.3101E-2	-.3488E-2	-.5894E-4	-.6290E-4
.14	-.1313	-.1928	-.5770E-1	-.7385E-1	-.1320E-1	-.1532E-1
.18	-.3641	-1.049	-.3091	-.6958	-.2387	-.4328
.22	.3512	1.061	.3023	.7062	.2371	.4220
.26	.1398	.2166	.6803E-1	.9051E-1	.1876E-1	.2234E-1
.30	.3933E-1	.5058E-1	.7526E-3	.8746E-2	.3397E-3	.3698E-3
.34	.7726E-2	.9175E-2	.4065E-3	.4482E-3	.1435E-5	.1513E-5
.38	.1041E-2	.1187E-2	.1066E-4	.1146E-4	.1455E-8	.1512E-8
.42	.9167E-4	.1024E-3	.1310E-6	.1391E-6	.3519E-12	.3631E-12
.46	.4480E-5	.4951E-5	.6456E-9	.6811E-9	.1773E-16	.1823E-16

(a) TRUE ERROR AND (b) ASYMPTOTIC ESTIMATE FOR THE ERROR

WHEN μ=.4 FOR THE DISCONTINUOUS PROBLEM (2.31)

TABLE 2.11

	μ = .6		μ = .8	
γ	(a)	(b)	(a)	(b)
.02	-.3670E-7	-.4130E-7	-.2245E-10	-.2494E-10
.06	-.8536E-5	-.9219E-5	-.1232E-7	-.1305E-7
.10	-.3087E-3	-.3362E-3	-.1043E-5	-.1101E-5
.14	-.4174E-2	-.4672E-2	-.3267E-4	-.3467E-4
.18	-.2836E-1	-.3359E-1	-.5058E-3	-.5437E-3
.22	-.1132	-.1518	-.4494E-1	-.4944E-2
.26	-.2952	-.5549	-.2512E-1	-.2879E-1
.30	.4538	——	-.9416E-1	-.1171
.34	.2232	.4931	-.2492	-.3740
.38	.7980E-1	.1238	-.4872	-1.432
.42	.7939E-2	.1046E-1	.2612	1.187
.42	.6404E-3	.8166E-3	.8122E-1	1.850

(a) TRUE ERROR AND (b) ASYMPTOTIC ESTIMATE FOR THE ERROR WHEN s=50 FOR THE DISCONTINUOUS PROBLEM (2.31)

Figure 2.2

Contour map for (2.10) with $\gamma=.625$, $p=.5$

CONTOUR KEY	
1	-0.3308
2	-0.2808
3	-0.2308
4	-0.1808
5	-0.1308
6	-0.0808
7	-0.0308
8	0.0192
9	0.0692
10	0.1192
11	0.1692

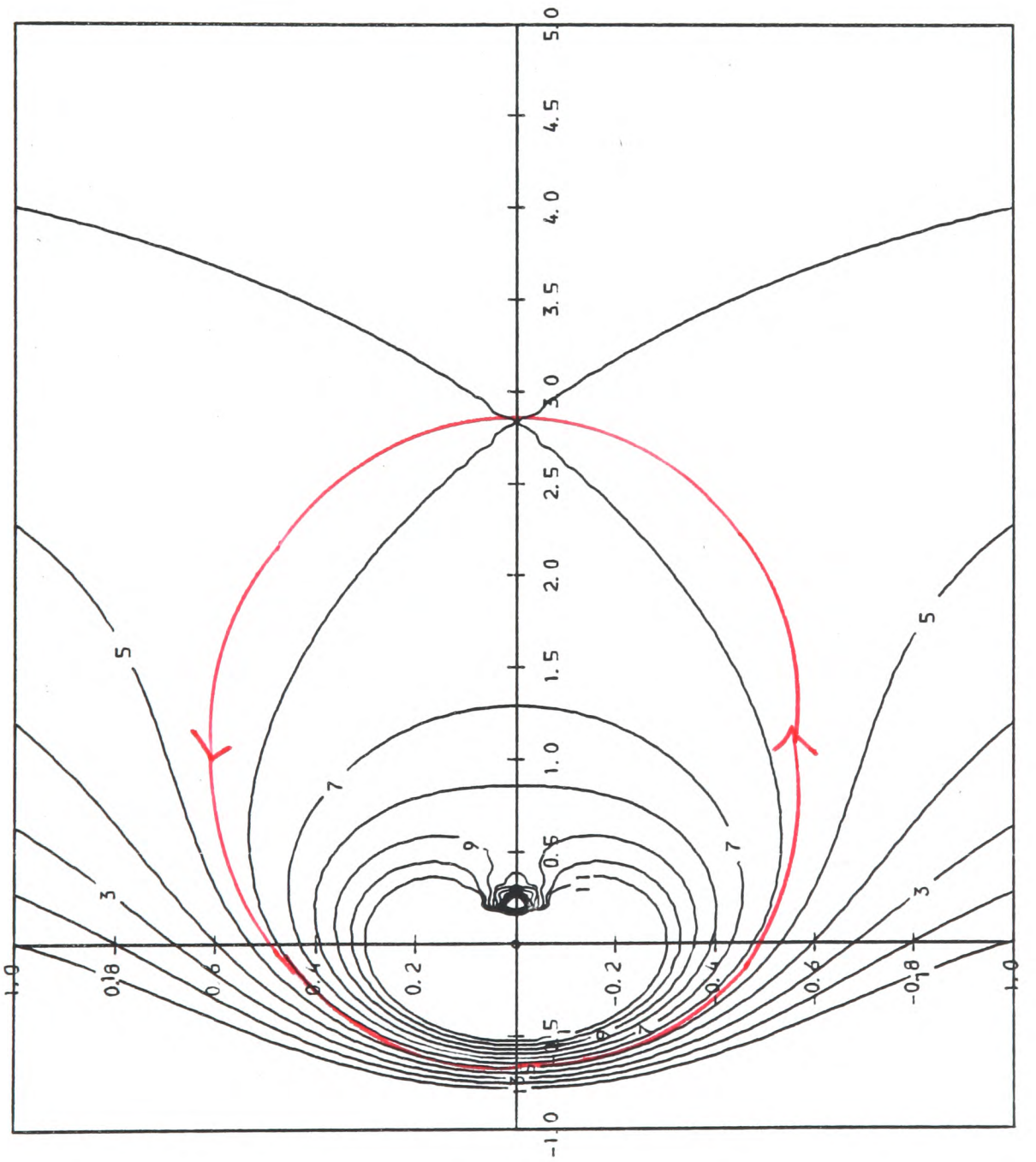


Figure 2.3

Contour map for (2.10) with $\gamma=.4$, $p=.5$

CONTOUR	KEY
1	-0.0314
2	-0.0264
3	-0.0214
4	-0.0164
5	-0.0114
6	-0.0064
7	-0.0014
8	0.0036
9	0.0086
10	0.0136
11	0.0186

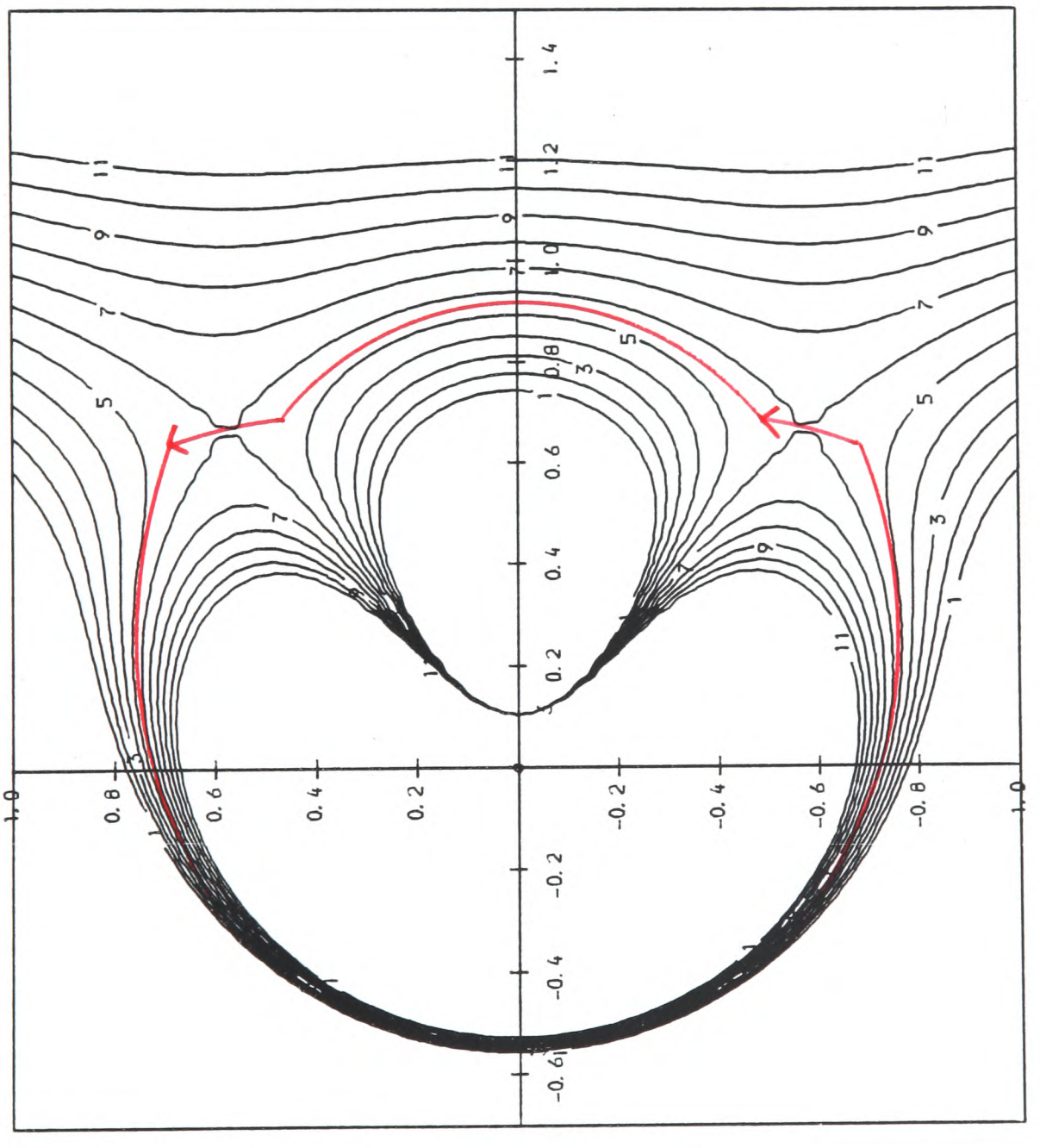


Figure 2.4

Contour map for (2.10) with $\gamma=p=.5$

CONTOUR KEY	
1	-0.0050
2	-0.0040
3	-0.0030
4	-0.0020
5	-0.0010
6	0.0000
7	0.0010
8	0.0020
9	0.0030
10	0.0040
11	0.0050

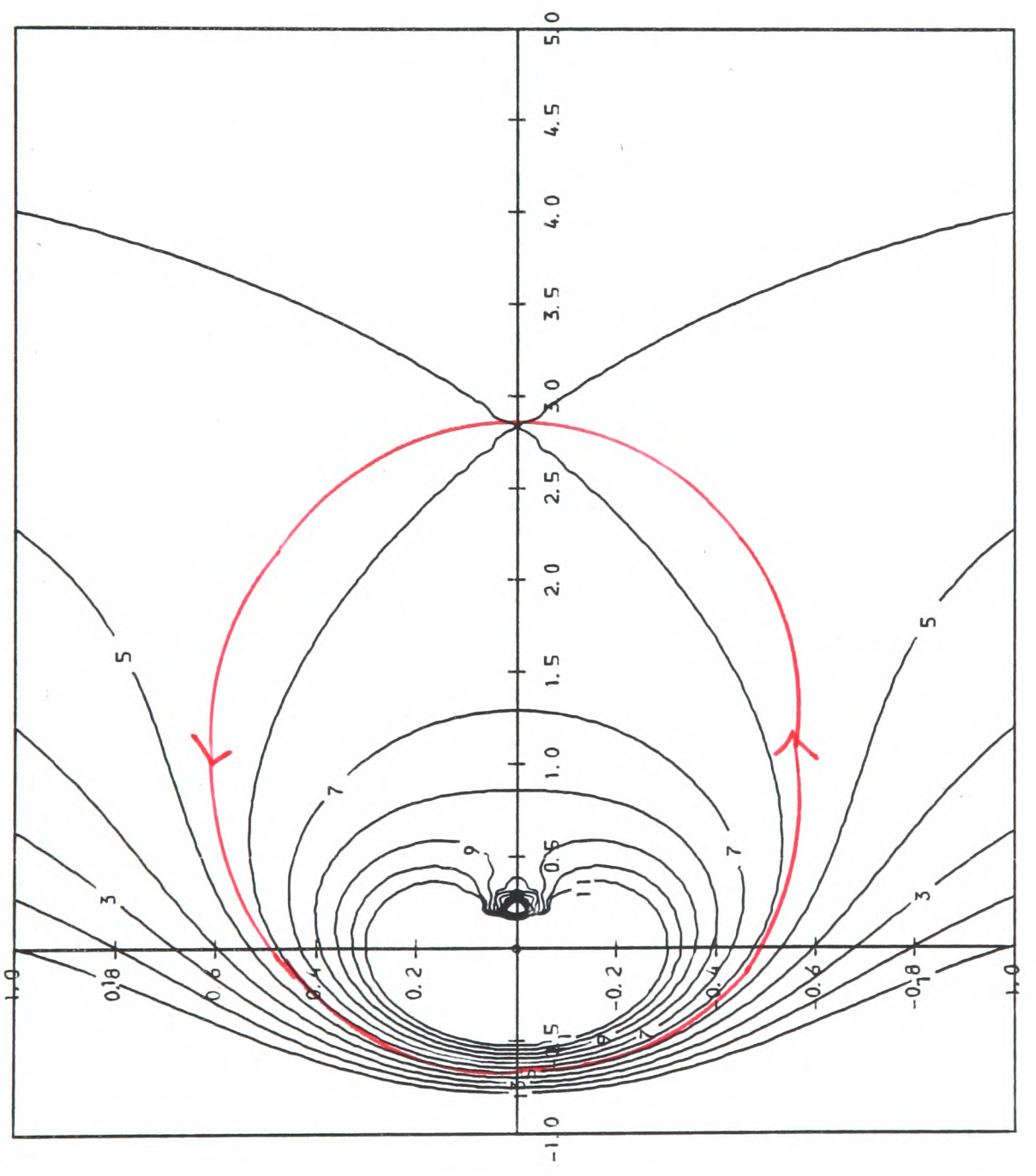
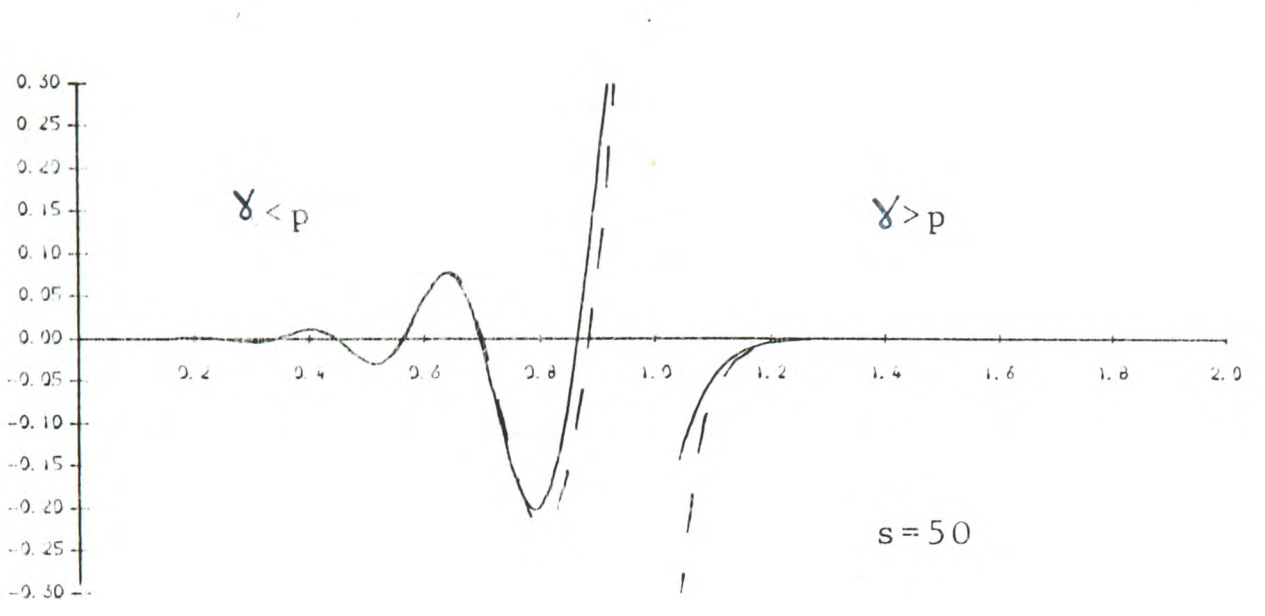
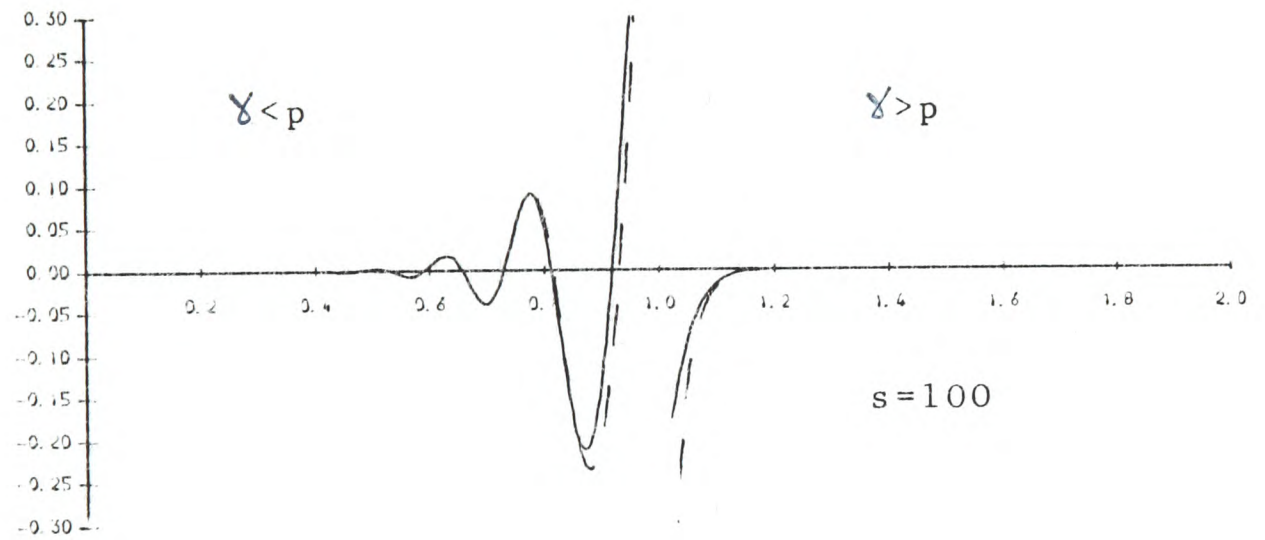


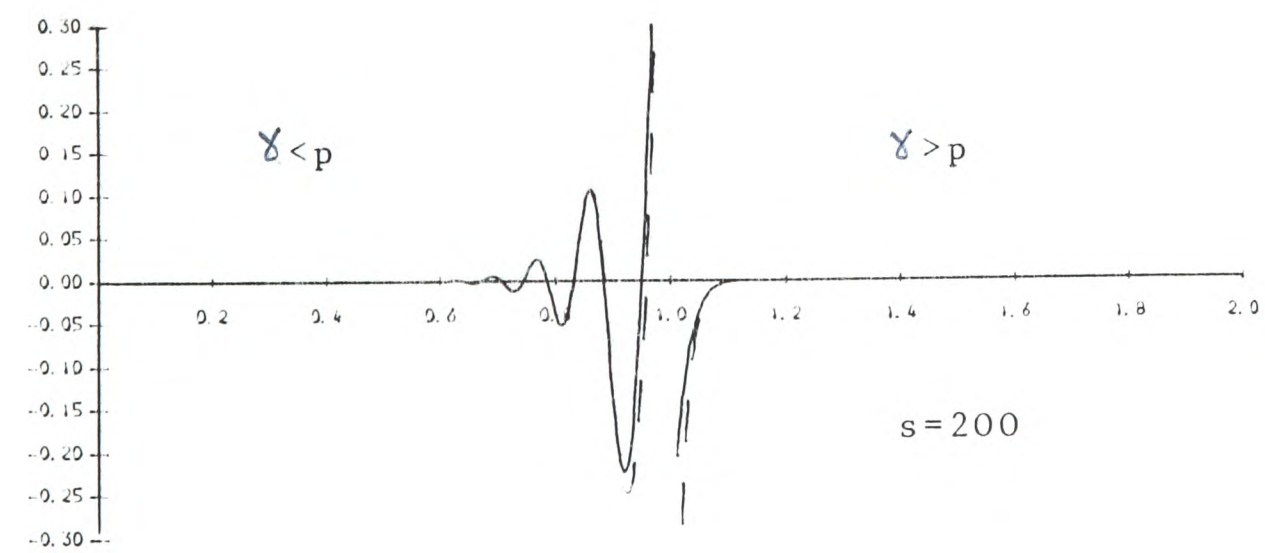
Figure 2.5

——————— Lax Wendroff Error

– – – – – – – Asymptotic approximation of error







Comparison of the error in the Lax Wendroff scheme and the asymptotic approximation to it

Figure 2.6

Contour map for λ_2, γ=-.5, p=.25

CONTOUR	KEY
1	-0.5354
2	-0.4354
3	-0.3354
4	-0.2354
5	-0.1354
6	-0.0354
7	0.0646
8	0.1646
9	0.2646
10	0.3646
11	0.4646

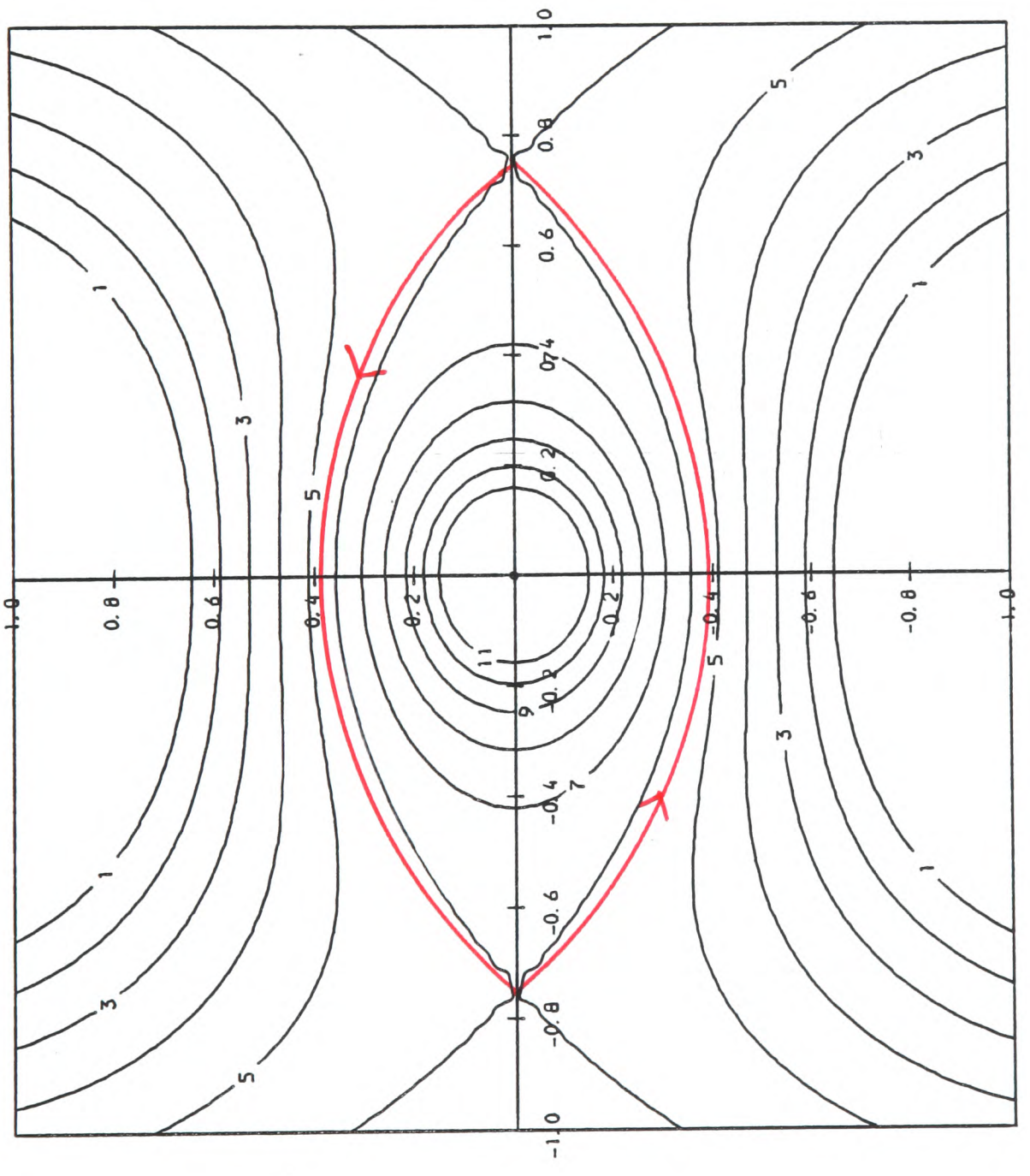


Figure 2.7

Contour map for λ_1, γ=-1.5, p=.75

CONTOUR KEY	
1	-3.67767
2	-3.57767
3	-3.47767
4	-3.37767
5	-3.27767
6	-3.17767
7	-3.07767
8	-2.97767
9	-2.87767
10	-2.77767
11	-2.67767

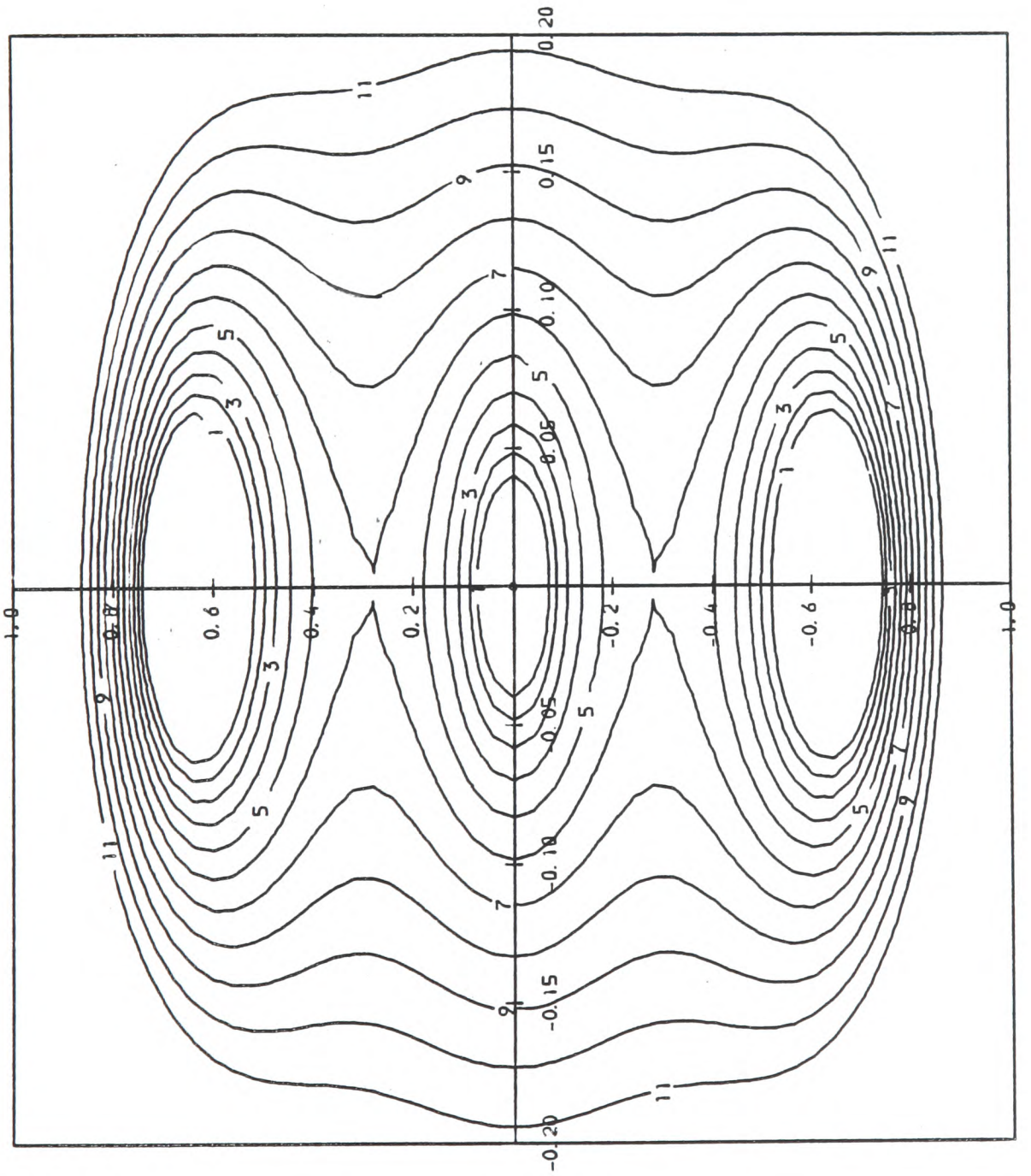


Figure 2.8

Contour map for λ_1, γ=.24, p=.25

CONTOUR	KEY
1	-0.17727
2	-0.16727
3	-0.15727
4	-0.14727
5	-0.13727
6	-0.12727
7	-0.11727
8	-0.10727
9	-0.09727
10	-0.08727
11	-0.07727

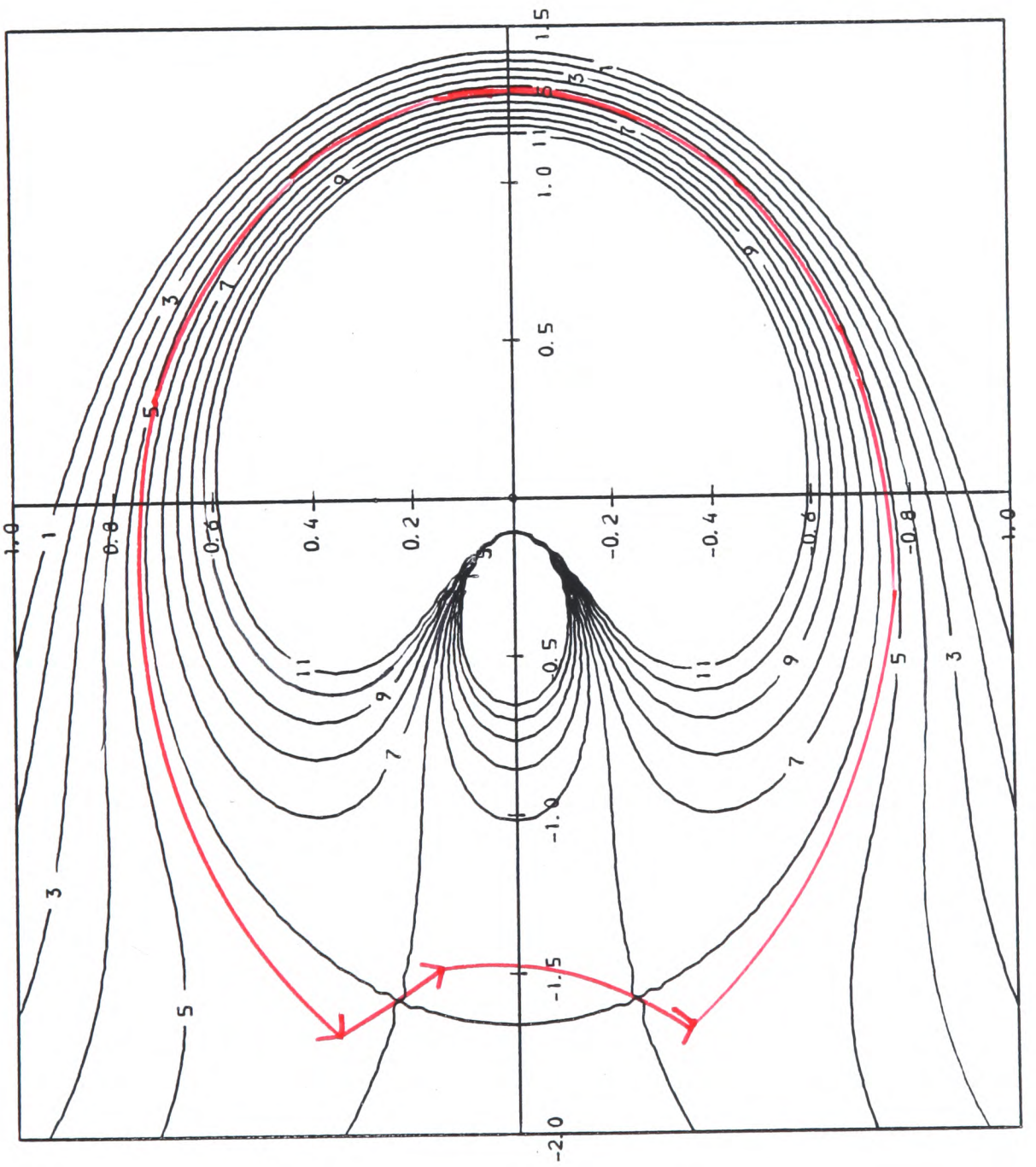


Figure 2.9

Contour map for λ_1, γ=.27, p=.25

CONTOUR	KEY
1	-0.1936
2	-0.1836
3	-0.1736
4	-0.1636
5	-0.1536
6	-0.1436
7	-0.1336
8	-0.1236
9	-0.1136
10	-0.1036
11	-0.0936

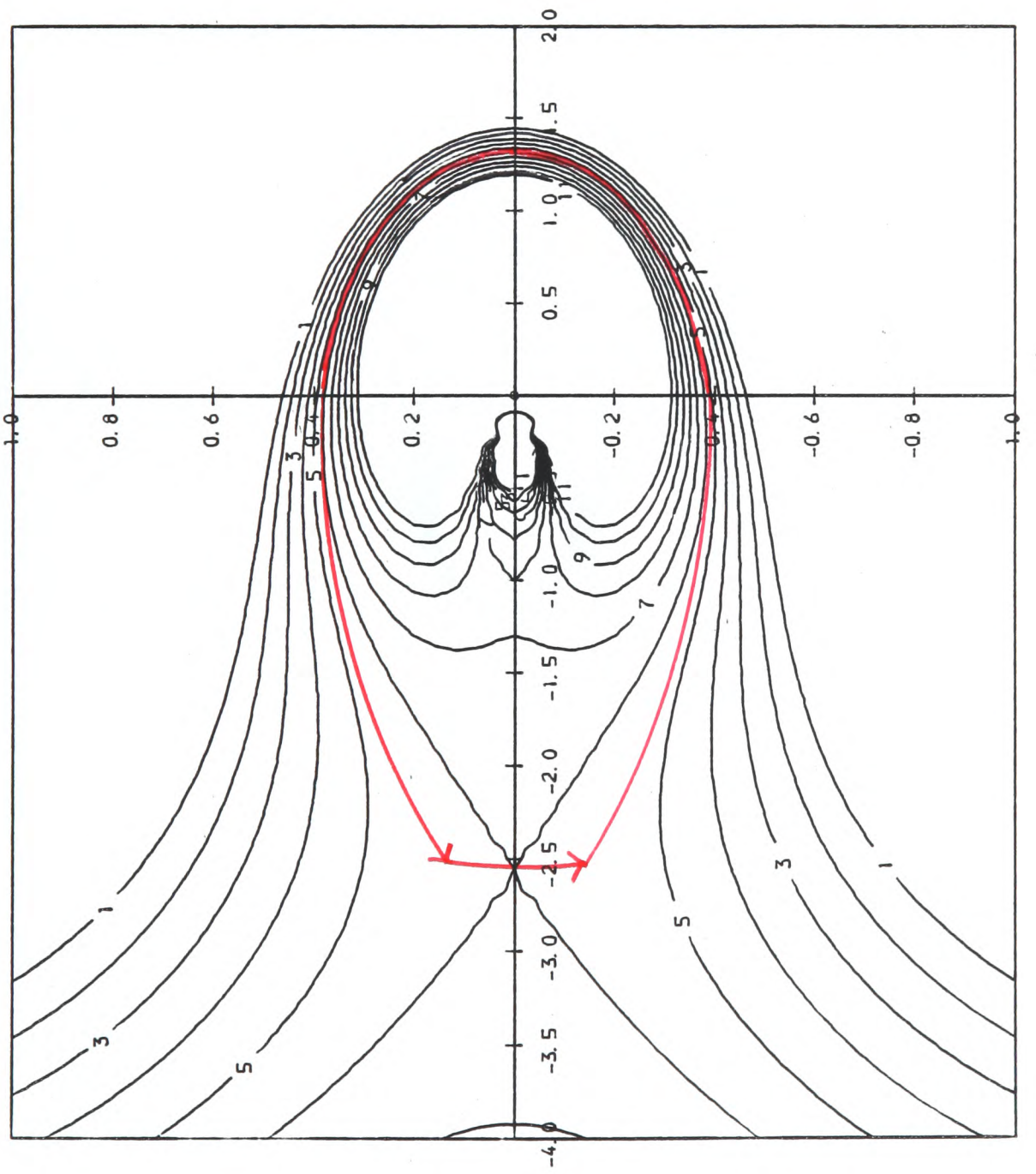
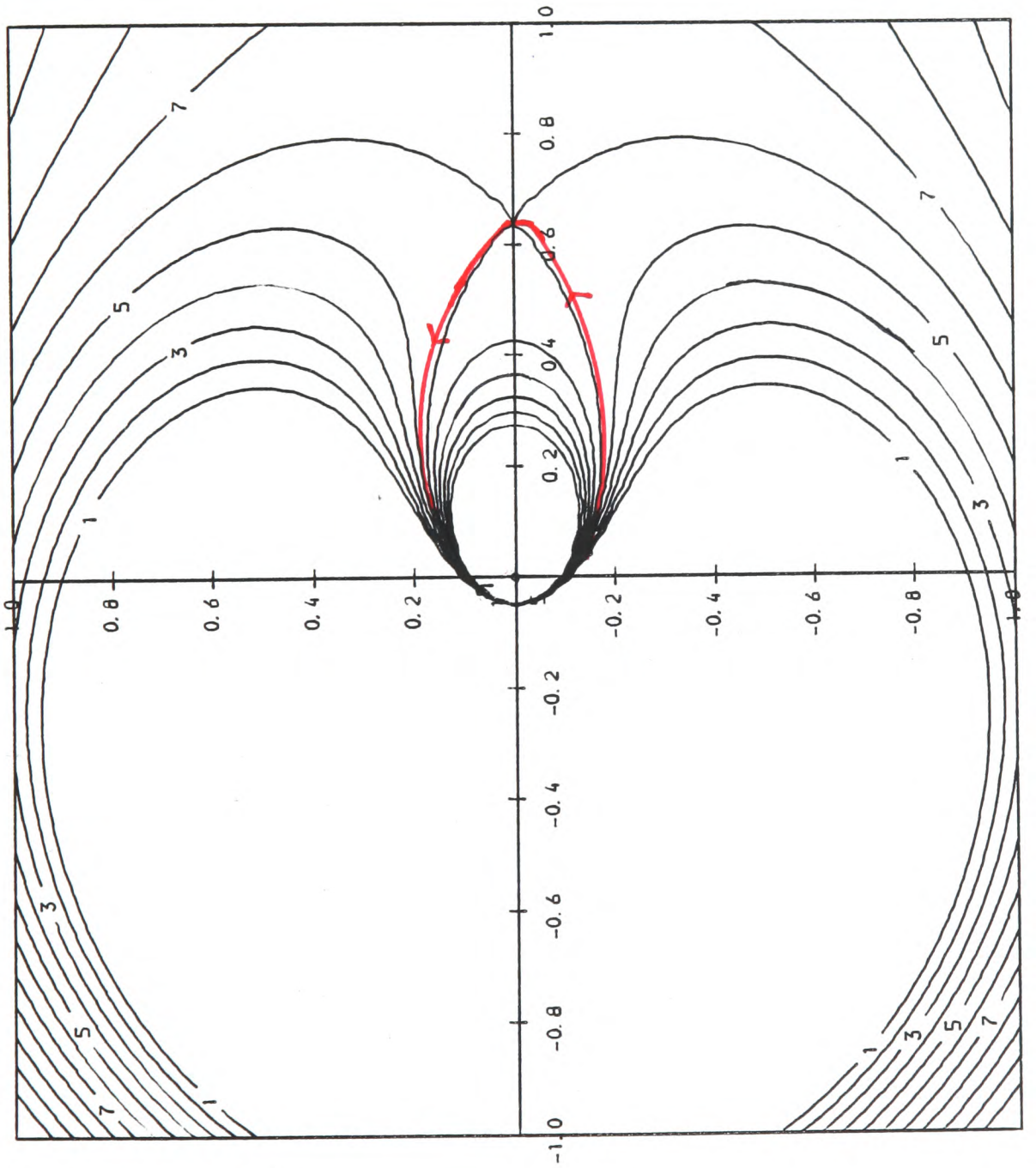


Figure 2.10

Contour map for λ_1, γ=-.27, p=.25

CONTOUR	KEY
1	-0.0557
2	-0.0457
3	-0.0357
4	-0.0257
5	-0.0157
6	-0.0057
7	0.0043
8	0.0143
9	0.0243
10	0.0343
11	0.0443



CHAPTER 3

The Cable Problem

§3.1 Description of the cable problem

This problem is concerned with the motion of a neutrally buoyant cable being towed behind a ship. Our task is to find the position of the cable at any time given the motion of the ship. This is part of a much larger package which is to detect the position of other sea vessels by the use of sonic hydrophones. These instruments, which are attached to the cable, cannot be used effectively unless the position of the cable is accurately known.

The configuration of the situation is as follows. The cable is not attached directly to the ship but to another cable which is itself connected to the ship. (See Figure 3.1.)

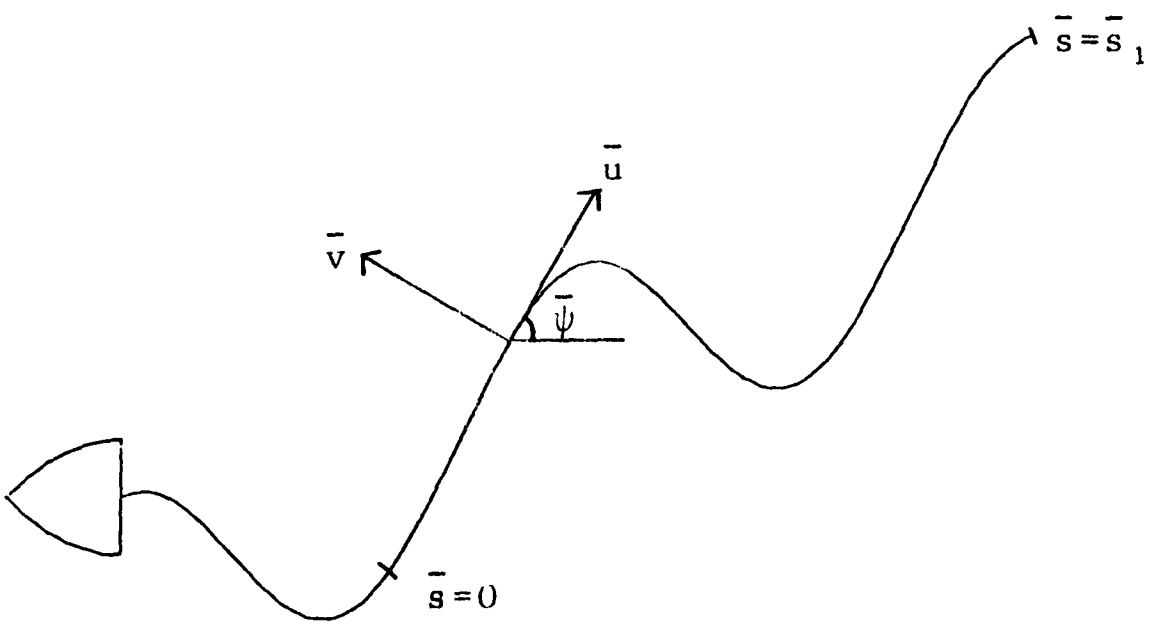


Figure 3.1

Sea velocities relative to the cable

At the far end of the cable a rope is attached in order to prevent a whiplash effect. The ship is moving with speed Vm/s in still water and is executing horizontal transverse motions (with amplitude a $\sim$ 100m, time scale $\tau \sim$ 1000 sec.). In this chapter the following notation is used.

$\bar{t}$ time

$\bar{s}$ distance along the cable from the reference point

$\bar{\psi}$ cable inclination to the oncoming flow direction

D diameter of the cable

$\bar{T}(\bar{s},\bar{t})$ load on the cable

m mass/unit length of the cable.

Along the cable:

$F_{//}\,\delta\bar{s}$ drag force on an element $\delta\bar{s}$

$\bar{u}$ sea velocity relative to the cable.

Transverse to the cable:

$F_{\perp}\,\delta\bar{s}$ drag force on an element $\delta\bar{s}$

$\bar{v}$ sea velocity relative to the cable

and $\bar{Q}^2 = \bar{u}^2 + \bar{v}^2 \simeq V^2$ since the speed of the ship is much greater than its horizontal motions.

C_F and C_D are the frictional and form drag coefficients respectively. In intrinsic variables, relative to a frame fixed in the sea, the momentum equations are tangentially:

$$m\left(-\frac{\partial\bar{u}}{\partial\bar{t}} - \bar{v}\frac{\partial\bar{\psi}}{\partial\bar{t}}\right) = \frac{\partial\bar{T}}{\partial\bar{s}} + F_{//} \quad . \tag{3.1}$$

Transversely:

$$m\left(\frac{\partial \bar{v}}{\partial \bar{t}} - \bar{u}\frac{\partial \bar{\psi}}{\partial \bar{t}}\right) = \bar{T}\frac{\partial \bar{\psi}}{\partial \bar{s}} - F_{\perp} \tag{3.2}$$

and from the compatibility conditions of Cartesian displacements,

$$\frac{\partial \bar{u}}{\partial \bar{s}} + \bar{v}\frac{\partial \bar{\psi}}{\partial \bar{s}} = 0 \tag{3.3}$$

$$\frac{\partial \bar{v}}{\partial \bar{s}} = \bar{u}\frac{\partial \bar{\psi}}{\partial \bar{s}} + \frac{\partial \bar{\psi}}{\partial \bar{t}} \quad . \tag{3.4}$$

In addition some equations are required for the drag force. In general, the drag force acting on a smooth body is given by

$$\text{Drag force} = C_D \tfrac{1}{2}\rho u^2 s$$

where ρ is the density of the fluid

u is the fluid velocity

s is the projected area of the body perpendicular to the stream

C_D is the drag coefficient. (See Milne-Thomson [21].)

Taylor [31] shows that normal drag and longitudinal drag are independent, that is the drag force in the normal direction is proportional to the velocity in that direction. He also investigates the effect of friction on the drag force. As Paidoussis [26] explains, a new drag coefficient is introduced which is called C_F; coefficient of frictional drag of the cylinder.

For a cylinder with a number of projections equally in all directions, which is the model used for the cable, the drag forces per unit length of the cable in normal and longitudinal

directions are as follows.

$$F_N = \tfrac{1}{2}\rho D\bar{Q}^2(C_D\sin^2\bar{\psi}+C_F\sin\bar{\psi})$$
$$F_{//} = \tfrac{1}{2}\rho D\bar{Q}^2 C_F\cos\bar{\psi} \ .$$

The drag force in a longitudinal direction contains only a frictional drag term as $\bar{\psi}$, $\frac{\partial\bar{\psi}}{\partial s}$ are assumed small so that each element of the cylinder can be considered as an element of an infinitely long straight cylinder inclined at an angle $\bar{\psi}$ to the incoming flow.

Since the cable is neutrally buoyant

$$\tfrac{1}{2}\rho D = \frac{2m}{\pi D}$$

and so the longitudinal drag force becomes

$$F_{//} = \frac{2m}{\pi D}C_F\bar{Q}\bar{u} \qquad (3.5)$$

whilst

$$F_N = \frac{2m}{\pi D}\bar{v}(C_F\bar{Q}+C_D|\bar{v}|)$$

where $\bar{Q}^2\sin^2\bar{\psi}$ is replaced by $\bar{v}|\bar{v}|$ so that the sign of the force is fixed. (See Huffman and Genin [15].)

Since the cable displaces liquid when it moves in a transverse direction there is also an added mass term in this direction. (See Breslin [3] and Goodman and Breslin [10].) The added mass is the mass of liquid displaced by the movement. For a circular cylinder (which we assume each element of the cable to be), the added mass per unit length, $M = \rho D = m$, since the cable is neutrally buoyant. (See Milne-Thomson [21].)

Hence the force needed due to the presence of the liquid is

$$M(\frac{\partial \bar{v}}{\partial \bar{t}} + \bar{u}\frac{\partial \bar{v}}{\partial \bar{s}})$$

where $\frac{\partial \bar{v}}{\partial \bar{t}} + \bar{u}\frac{\partial \bar{v}}{\partial \bar{s}}$ is the convective derivative of $\bar{v}$ at speed $\bar{u}$ along the cable.

Hence, the total drag force transverse to the cable is given by

$$F_{\perp} = F_N + m(\frac{\partial \bar{v}}{\partial \bar{t}} + \bar{u}\frac{\partial \bar{v}}{\partial \bar{s}}) \ .$$

That is,

$$F_{\perp} = \frac{2m}{\pi D}\bar{v}(C_F\bar{Q}+C_D|\bar{v}|) + m(\frac{\partial \bar{v}}{\partial \bar{t}} + \bar{u}\frac{\partial \bar{v}}{\partial \bar{s}}) \ . \qquad (3.6)$$

Inserting (3.5) and (3.6) into (3.1) and (3.2) and using (3.4) in (3.2) gives

$$\frac{\partial \bar{T}}{\partial \bar{s}} + \frac{2m}{\pi D} C_F\bar{Q}\bar{u} = -m(\frac{\partial \bar{u}}{\partial \bar{t}} + \bar{v}\frac{\partial \bar{\psi}}{\partial \bar{t}}) \qquad (3.7)$$

and

$$(\bar{T}-m\bar{u}^2)\frac{\partial \bar{\psi}}{\partial \bar{s}} - \frac{2m}{\pi D}C_D\bar{v}(\frac{C_F}{C_D}\bar{Q}+|\bar{v}|) = m(\frac{\partial \bar{v}}{\partial \bar{t}}-\bar{u}\frac{\partial \bar{v}}{\partial \bar{s}}) + m(\frac{\partial \bar{v}}{\partial \bar{t}}+\bar{u}\frac{\partial \bar{v}}{\partial \bar{s}}) \ . \qquad (3.8)$$

The equations to be solved are (3.3), (3.4), (3.7) and (3.8) with suitable boundary and initial conditions. The equations are scaled and non-dimensionalised by transforming to new variables given by

$$\bar{t} = \tau t, \quad \bar{s} = V\tau s, \quad \bar{v} = \varepsilon V v, \quad \bar{u} = Vu,$$
$$\bar{Q} = VQ, \quad \bar{\psi} = \varepsilon\psi, \quad \bar{T} = \frac{4m}{\pi D}C_F V^3 \tau T = \frac{mV^2}{\alpha}T$$

where $\varepsilon = \frac{a}{V\tau}$ is small and $\alpha = \frac{D\pi}{4V\tau C_F}$.

This transforms the equations as follows:

(3.3) becomes $\frac{\partial u}{\partial s} = - \varepsilon^2 v \frac{\partial \psi}{\partial s}$, (3.9)

(3.7) leads to $\frac{\partial T}{\partial s} = -\frac{1}{2} Qu - \alpha \frac{\partial u}{\partial t} - \varepsilon^2 \alpha v \frac{\partial \psi}{\partial t}$, (3.10)

(3.4) becomes $\frac{\partial v}{\partial s} = u \frac{\partial \psi}{\partial s} + \frac{\partial \psi}{\partial t}$, (3.11)

(3.8) leads to $(T - \alpha u^2) \frac{\partial \psi}{\partial s} - \frac{1}{2} v (Q + 2\beta |v|) = 2\alpha \frac{\partial v}{\partial t}$ (3.12)

where $Q^2 = u^2 + \varepsilon^2 v^2$ and $\beta = \frac{2\alpha C_D a}{D\pi} = \frac{\varepsilon C_D}{2C_F}$

β and α are non dimensional parameters.

We reduce these equations by considering $\varepsilon^2 \sim 10^{-4}$ to be negligible in the first approximation. α is not treated as a small parameter since although $\alpha << 1$, 2α is not negligible compared to $T - \alpha u^2$.

To leading order, (3.9) shows that u is a function of t only and since $\bar{u} >> \bar{v}$ we may take $u = 1$ as the simplest model. Then $Q = 1$ so that (3.10) gives

$$T(s,t) = -\tfrac{1}{2}(s - s_1) + T_1(t) \qquad (3.13)$$

where $T_1(t)$ is the tension at the end of the cable.

In (3.12) it is seen that the coefficient of $\frac{\partial \psi}{\partial s}$ (that is, $T - \alpha u^2$) becomes negative near $s = s_1$ unless $T_1 > \alpha$. If $T_1 < \alpha$ the cable would become unstable, giving rise to a whiplash effect. In practice a rope is attached to $s = s_1$ experiencing a drag which we postulate to be constant. This gives the boundary condition

$$T(s_1,t) = T_1 = \text{constant} \qquad (3.14)$$

and so the tension is a linear function of s. Also, assuming there is a zero crossflow v, at the end gives

$$v(s_1,t) = 0 \quad . \tag{3.15}$$

Since u = 1, (3.11) can be rearranged as

$$\frac{\partial}{\partial s}(v-\psi) = \frac{\partial \psi}{\partial t}$$

and putting $\psi = \frac{\partial y}{\partial s}$, where to this approximation y(s,t) is a dimensionless displacement normal to the flow, gives

$$v = \frac{\partial y}{\partial t} + \frac{\partial y}{\partial s} \tag{3.16}$$

Then in (3.12) this gives

$$(A-s)y_{ss} = (y_s+y_t)(1+2\beta|y_s+y_t|) + 4\alpha(y_{st}+y_{tt}) \tag{3.17}$$

where $A = 2(T_1-\alpha) + s_1$ and the subscripts denote partial differentiation with respect to that variable.

The boundary conditions are

$$y(0,t) = \bar{f}(t) \quad \text{(specified)} \tag{3.18}$$

$$y_s + y_t = 0 \text{ at } s = s_1 \tag{3.19}$$

(that is, v = 0)

and initially

$$y(s,0) = \bar{g}(s), \quad y_t(s,0) = \bar{h}(s) \tag{3.20}$$

$$0 < s < s_1$$

We aim to predict the motion into which the cable settles down and so the initial conditions are not crucial. The model examined below uses $\bar{g}(s) \equiv 0$, $\bar{h}(s) \equiv 0$ for simplicity but it would be possible to adapt the schemes considered for different initial conditions.

$\bar{f}(t)$ should approximate the motion of the ship. As this is considered to be periodic $\bar{f}(t) = 1 - \cos(t)$ was used.

This gives $y_t = \sin(t)$ on the left hand boundary and so the first derivative is then continuous (as compared with $\bar{f}(t) = \sin(t)$ which would give a singularity in the first derivative at the origin). However $y_{tt}(0,t) = \cos(t)$ whereas $y_{tt}(s,0) = 0$ and so there is a discontinuity in y_{tt} at the origin. As explained in Chapter 1 this discontinuity will be propagated along the characteristic which passes through the point (0,0). This discontinuity will be damped with time as shown in the next section, which deals with the use of characteristics to solve the equation.

Section 3.3 gives a numerical solution using finite differences and then in Section 3.4 results are given and the two methods are compared.

The boundary condition $\bar{f}(t) = 1 - \cos(t)$ gives an oscillation with period length 2π. This may be longer than that experienced in reality and so the effect on the solution of altering the boundary condition to $1 - \cos(2\pi t)$ (giving a period length of 1) has been investigated too. This, and other minor modifications are presented in Section 3.5.

The constants for the problem are given to be $A = .3$, $\beta = 3$, $\alpha = \frac{1}{30}$ and $s_1 = .25$.

§3.2 Numerical solution using characteristics

Throughout this section, the abbreviation $p = y_s$ and $q = y_t$ has been used. Using this notation (3.17) may be written as

$$(A-s)y_{ss} - by_{st} - cy_{tt} = f(p,q) \qquad (3.21)$$

where $f(p,q) = (p+q)(1+2\beta|p+q|)$

$b = c = 4\alpha.$

This gives the characteristic relation

$$(A-s)\left(\frac{dt}{ds}\right)^2 + b\frac{dt}{ds} - c = 0 \qquad (3.22)$$

and the consistency relation

$$(A-s)dp\frac{dt}{ds} - fdt - cdq = 0 \qquad (3.23)$$

which is satisfied along the characteristics.

From (3.22) the characteristics are given by

$$t = \int^{s} \frac{-b \pm (b^2+4c(A-s'))^{\frac{1}{2}}}{2(A-s')} ds'$$

which gives

$$t = \mp(b^2+4c(A-s))^{\frac{1}{2}} + b\log(\pm b+(b^2+4c(A-s))^{\frac{1}{2}}) + \text{constant}. \qquad (3.24)$$

The constant is determined by specifying a point in the (s,t) plane through which the characteristic passes.

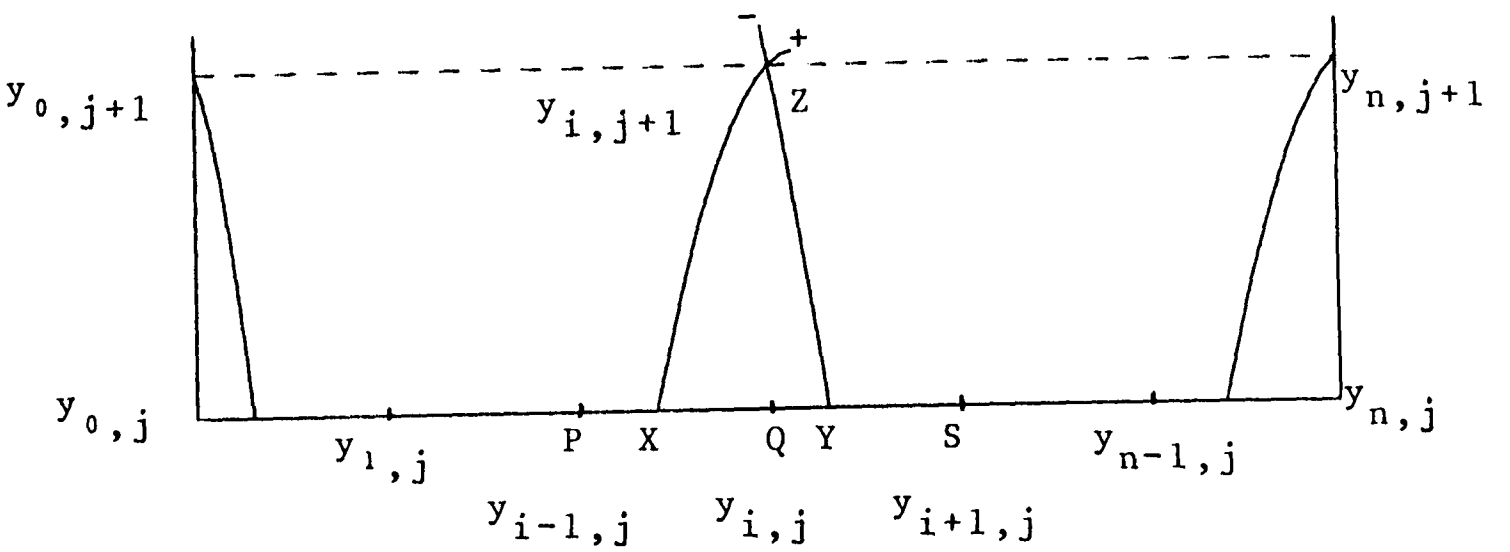


Figure 3.2

If characteristics of different families passing through points X and Y in the (s,t) plane meet at Z then (3.23) gives the two equations, (see Figure 3.2)

$$\tfrac{1}{2}(p_Z-p_X)((A-s_X)\phi_X + (A-s_Z)\phi_Z) - \tfrac{1}{2}(t_Z-t_X)(f(p_X,q_X) + f(p_Z,q_Z)) - c(q_Z-q_X) = 0 \qquad (3.25)$$

and

$$\tfrac{1}{2}(p_Z-p_Y)((A-s_Y)\psi_Y + (A-s_Z)\psi_Z) - \tfrac{1}{2}(t_Z-t_Y)(f(p_Y,q_Y) + f(p_Z,q_Z)) - c(q_Z-q_Y) = 0 \qquad (3.26)$$

where

$$\psi_Z = \frac{-b - (b^2+4c(A-s_Z))^{\frac{1}{2}}}{2(A-s_Z)}$$

$$\phi_z = \frac{-b + (b^2+4c(A-s_Z))^{\frac{1}{2}}}{2(A-s_Z)}$$

and the subscripts denote evaluation at that point.

An equation for y_Z is

$$y_Z = y_X + \tfrac{1}{2}(s_Z-s_X)(p_X+p_Z) + \tfrac{1}{2}(t_Z-t_X)(q_X+q_Z) \; . \qquad (3.27)$$

It is equally valid to use the values at Y instead of X and in fact the program uses both series and then checks that the two values obtained for y_Z are sufficiently close.

Two slightly different methods for dealing with these equations have been used. For both algorithms the following notation is adopted.

$y_{i,j} = y(ih,jk)$, that is the approximation to the solution at the point s = ih, t = jk where h is the step length in the s direction and k is the step length in the t direction.

The first algorithm (henceforth referred to as the 'standard' approach) is as follows. (See Figure 3.2.)

'Standard' algorithm

1) Assume y, p, q are given (or have been calculated) at the j^{th} time step and need to be calculated at the $(j+1)^{th}$ time step.

2) Calculate the constants C_X and C_Y corresponding to the + and - characteristic passing through (s_i, t_{j+1}) by substituting $t = (j+1)k$, $s = ih$ in (3.24).

3) Solve (3.24) with these two constants at $t = jk$ to find s_X and s_Y. This is done by using Newton iteration

$$t = g(s) + C$$

where $g(s) = \pm(b^2+4c(A-s))^{\frac{1}{2}} - b\log(\pm b+(b^2+4c(A-s))^{\frac{1}{2}})$

and $C = C_X$ or C_Y depending on the sign chosen.

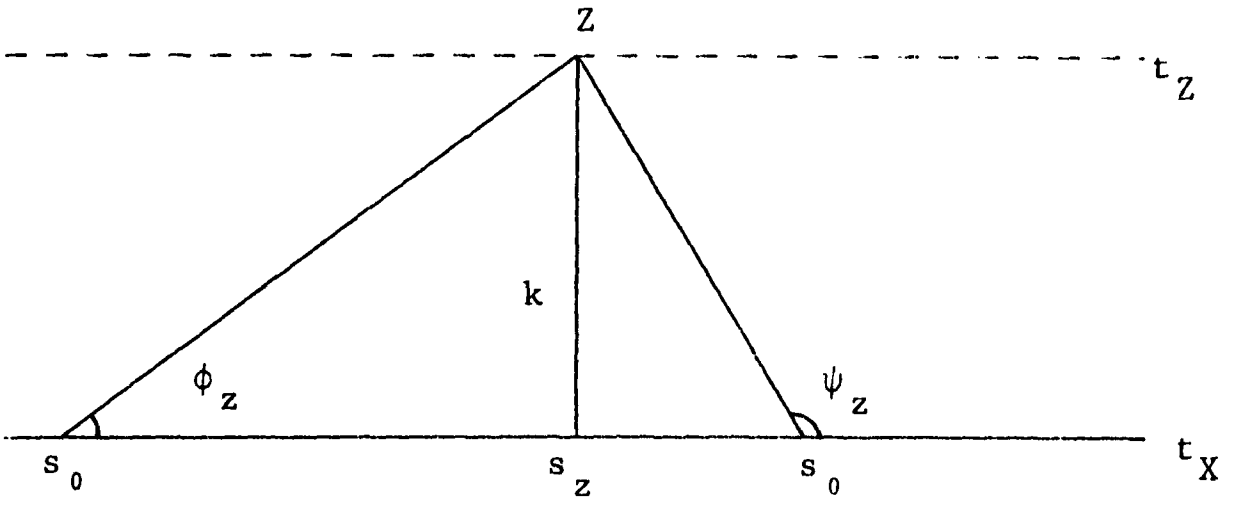


Figure 3.3

It is known that

$$g'(s) = \frac{dt}{ds} = \frac{-b \pm (b^2+4c(A-s))^{\frac{1}{2}}}{2(A-s)}$$

and so we use the following iteration

$$s^{r+1} = s^r - \frac{G(s^r)}{G'(s^r)}$$

with $G(s) = g(s) + C - t_Z + k$

$G'(s) = g'(s)$

and s^r is the approximation after r iterations.

The initial guess, s^0, is given by the straight line through (s_Z, t_Z) with slope $g'(s_Z) = \phi_Z$ or ψ_Z. (See Figure 3.3.)

4) Interpolate linearly between the two grid points adjacent to X to find p_X, q_X and y_X and similarly for Y. In most cases X lies in PQ and Y in QR because of the step sizes k and h chosen. However, the effect of changing the step size so that interpolation takes place between other grid points has been investigated and the results are given later.

5) Equations (3.25) and (3.26) can be rearranged to give

$$p_Z m_X + n_X - \tfrac{1}{2}kf(p_Z,q_Z) - cq_Z = 0 \qquad (3.28)$$

$$p_Z m_Y + n_Y - \tfrac{1}{2}kf(p_Z,q_Z) - cq_Z = 0 \qquad (3.29)$$

with

$$m_X = \tfrac{1}{2}((A-s_X)\phi_X + (A-s_Z)\phi_Z)$$

$$n_Y = -\tfrac{1}{2}p_Y((A-s_Y)\psi_Y + (A-s_Z)\psi_Z) - \tfrac{1}{2}kf(p_Y,q_Y) + cq_Y$$

$$= -p_Y m_Y - \tfrac{1}{2}kf(p_Y,q_Z) + cq_Y \quad \text{etc.}$$

Subtracting (3.29) from (3.28) gives

$$p_Z(m_X - m_Y) + n_X - n_Y = 0$$

that is,

$$p_Z = \frac{n_Y - n_X}{m_X - m_Y} \quad . \tag{3.30}$$

So, p_Z can be calculated directly and q_Z is calculated from (3.28) using Newton's method and substituting for p_Z from (3.30). The initial guess for q_Z is $\frac{1}{2}(q_X+q_Y)$.

6) Using (3.27) calculate y_Z.

This algorithm can be used to find y, p, q at all the internal points of the grid. However, at the boundaries a slightly modified procedure must be adopted.

At the left hand boundary the condition $y = 1 - \cos(t)$ is given. This implies $q = \sin(t)$ along this boundary. The only undetermined quantity, therefore, is p. This is obtained by calculating s_Y and interpolating linearly as before and then using Newton iteration in (3.29) (after having substituted for q_Z) to find p_Z.

At the right hand boundary the given condition is $p + q = 0$. We also have equation (3.28) with $f(p_Z,q_Z) = 0$.

(Since $f(p,q) = (p+q)(1+2\beta|p+q|) = 0$ when $p + q = 0$.)

That is,

$$p_Z m_X + n_X - cq_Z = 0$$

and $$p_Z + q_Z = 0.$$

This gives

$$q_Z = \frac{n_X}{m_X + c} \quad , \quad p_Z = - \frac{n_X}{m_X + c}$$

for points on the right hand boundary. y_Z can be calculated as before.

'Alternative' algorithm

In this algorithm, use has been made of the fact that the characteristic equation can be integrated exactly to give a nonlinear equation which is solved by Newton's method. In

general, however, this is not the case. The more usual approach is that instead of finding s_X and s_Y via (3.24) the characteristic relation (3.22) can be used to give the difference relations

$$\frac{t_Z - t_X}{s_Z - s_X} = \tfrac{1}{2}(\phi_X+\phi_Z) \tag{3.31}$$

and

$$\frac{t_Z - t_Y}{s_Z - s_Y} = \tfrac{1}{2}(\psi_Y+\psi_Z) \ . \tag{3.32}$$

These equations can be solved iteratively for s_X and s_Y in the following manner.

$$s_X^{r+1} = s_Z - \frac{(t_Z-t_X)}{\tfrac{1}{2}(\phi_X^r+\phi_Z)} \tag{3.33}$$

$$s_Y^{r+1} = s_Z - \frac{(t_Z-t_Y)}{\tfrac{1}{2}(\psi_Y^r+\psi_Z)} \tag{3.34}$$

where $\phi_X^r = \phi(s_X^r)$, $\psi_Y^r = \psi(s_Y^r)$.

Initially, take $\phi_X = \phi_Z$, $\psi_Y = \psi_Z$. This approach (termed the 'alternative' method) is compared with the standard method in section 3.4.

The damping of the weak singularity at the origin as it propagates along the characteristic through (0,0) can be seen as follows.

The characteristic relationship (3.23) satisfied along characteristics, can be linearised since locally we can treat the characteristic as a straight line. This gives $\frac{dt}{ds} = \mu$ (constant). Also locally we can approximate (A-s) by a

constant, A*, and f(p,q) by $\lambda(p+q)$ where λ is a constant.

This gives

$$A^*\mu dp - \lambda(p+q)dt - cdq = 0$$

$$\text{along } \frac{dt}{ds} = \mu.$$

This can be differentiated to give a relationship between dq_t and dp_t,

$$A^*\mu dp_t - \lambda(p_t+q_t)dt - cdq_t = 0.$$

Then subtracting the relationships obtained by considering characteristics an infinitesimal distance to either side of the characteristic through the origin gives

$$A^*\mu d(\delta p_t) - \lambda(\delta p_t+\delta q_t)dt - cd(\delta q_t) = 0$$

where δp_t δq_t are the jumps in p_t, q_t across the characteristic.

Continuity along the characteristic implies $\delta p_t = \mu\delta q_t$ and so we have

$$A^*\mu^2 d(\delta q_t) - \lambda(1+\mu)\delta q_t dt - cd(\delta q_t) = 0$$

which gives $\delta q_t = \delta q_t(0)e^{-kt}$ where $k = \dfrac{\lambda(1+\mu)}{c-A^*\mu^2}$.

The singularity is damped, therefore, provided $k > 0$, that is $\mu^2 < c/A^*$.

It is easy to show that since b = c this condition is always satisfied.

§3.3 Numerical solution using finite differences

This algorithm is a straightforward application of the implicit finite difference method given in Chapter 1. The notation is as in the previous section with the following additions.

$$\underline{y}_{j+1} = (y_{1,j+1}, y_{2,j+1}, \ldots, y_{n-1,j+1})^T$$

$$(y_{ss})_{i,j} = \frac{y_{i+1,j} - 2y_{i,j} + y_{i-1,j}}{h^2}$$

$$(y_s)_{i,j} = \frac{y_{i+1,j} - y_{i-1,j}}{2h}$$

(and similarly for derivatives in the t direction).

$$(y_{st})_{i,j} = \frac{y_{i+1,j+1} - y_{i-1,j+1} + y_{i-1,j-1} - y_{i+1,j-1}}{4hk}$$

$$f(y)_{i,j} = (\tfrac{1}{4}(y_s)_{i,j+1} + \tfrac{1}{2}(y_s)_{i,j} + \tfrac{1}{4}(y_s)_{i,j-1} + (y_t)_{i,j}).$$

Assuming y is known at the j^{th} and $(j-1)^{th}$ time levels and is to be calculated at the $(j+1)^{st}$ level we define

$$g_i(\underline{y}_{j+1}) = (A-ih)(\tfrac{1}{4}(y_{ss})_{i,j+1} + \tfrac{1}{2}(y_{ss})_{i,j} + \tfrac{1}{4}(y_{ss})_{i,j-1})$$
$$- b(y_{st})_{i,j} - c(y_{tt})_{i,j}$$
$$- f(y)_{i,j}(1+2\beta|f(y)_{i,j}|) \qquad \text{for } 1 \leq i \leq n-1$$

where $b = c = 4\alpha$.

From the boundary condition we can put

$$y_{0,j} = 1 - \cos(jk)$$

$$y_{n,j+1} = y_{n,j} - \frac{k}{h}(y_{n,j} - y_{n-1,j}) \quad \text{for all } j > 0$$

which corresponds to $y_t + y_s = 0$ at the right hand end.

This is an implicit scheme and so cannot be solved directly. The system of equations is solved by Newton-Raphson iteration. That is,

$$\underline{y}_{j+1}^{r+1} = \underline{y}_{j+1}^{r} - \delta\underline{y}_{j+1}^{r}$$

where the superscript refers to the iterative process and $\delta\underline{y}_{j+1}^{r}$ is the solution to

$$M_{j+1}^{r}\,\delta\underline{y}_{j+1}^{r} = \underline{g}_{j+1}^{r}$$

$$\text{with } M_{j+1}^{r} = \left(\begin{array}{ccc} \dfrac{\partial g_1}{\partial y_{1,j+1}} & \cdots\cdots & \dfrac{\partial g_1}{\partial y_{n-1,j+1}} \\ \vdots & & \vdots \\ \dfrac{\partial g_{n-1}}{\partial y_{1,j+1}} & \cdots\cdots & \dfrac{\partial g_{n-1}}{\partial y_{n-1,j+1}} \end{array} \right)_{\underline{y}_{j+1}=\underline{y}_{j+1}^{r}}$$

and

$$\underline{g}_{j+1}^{r} = (g_1(\underline{y}_{j+1}^{r}),\ldots,g_{n-1}(\underline{y}_{j+1}^{r})).$$

Since $\dfrac{\partial g_1}{\partial y_{m,j+1}} = 0$ when $m < i-1$ or $m > i+1$, M_{j+1}^{r} is tridiagonal

$$M_{j+1}^{r} = \left(\begin{array}{cccc} C_1 & B_1 & & \\ A_2 & C_2 & B_2 & \\ & \ddots & \ddots & \ddots \\ & & A_{n-2} & C_{n-2} \quad B_{n-2} \\ & & & A_{n-1} \quad C_{n-1} \end{array} \right)$$

where

$$B_i = \left.\frac{\partial g_i}{\partial y_{i+1,j+1}}\right|_{\underline{y}_{j+1}=\underline{y}_{j+1}^{r}}$$

$$= \frac{(A-ih)}{4h^2} - \frac{b}{4hk} - (1+4\beta|f(y)_{i,j}|)\cdot\frac{1}{4.2h}\Bigg|_{\underline{y}_{j+1}=\underline{y}_{j+1}^{r}}$$

$$C_i = \left.\frac{\partial g_i}{\partial y_{i,j+1}}\right|_{y_{j+1}=y^r_{j+1}}$$

$$= \left.\frac{(A-ih)}{-2h^2} - \frac{c}{k^2} - (1+4\beta|f(y)_{i,j}|)\cdot\frac{1}{2k}\right|_{y_{j+1}=y^r_{j+1}}$$

$$A_i = \left.\frac{\partial g_i}{\partial y_{i-1,j+1}}\right|_{y_{j+1}=y^r_{j+1}}$$

$$= \left.\frac{(A-ih)}{4h^2} + \frac{b}{4hk} - (1+4\beta|f(y)_{i,j}|)\cdot\frac{-1}{4.2h}\right|_{y_{j+1}=y^r_{j+1}} .$$

The iterative process is initiated by taking y^0_{j+1} to be the solution to $N\underline{x} = b$ where

$$N = \begin{pmatrix} C'_1 & B'_1 & & & & \\ A'_2 & C'_2 & B'_2 & & & \\ & & & A'_{n-2} & C'_{n-2} & B'_{n-2} \\ & & & & A'_{n-1} & C'_{n-1} \end{pmatrix}$$

and

$$C'_i = \frac{-1}{k},\quad A'_i = \frac{1}{2h},\quad B'_i = \frac{-1}{2h}$$

and

$$b_i = \begin{cases} \dfrac{-y_{1,j}}{k} - \dfrac{y_{0,j+1}}{2h} & i = 1 \\ \dfrac{-y_{i,j}}{k} & 1 \leq i \leq n-1 \\ \dfrac{-y_{n,j}}{k} + \dfrac{y_{n,j+1}}{2h} & i = n-1 . \end{cases}$$

That is, $\underline{y}^0_{j+1}$ is the solution to the fully implicit finite difference approximation for the equation $y_t + y_s = 0$. This scheme cannot be used at the first time step unless values are assigned to $y_{i,-1}$. The initial condition $y_t = 0$ suggests making $y_{i,-1} = 0$ for $i \neq 0$ and $y_{0,-1} = 1 - \cos(k)$. The results for this algorithm are discussed in the following section.

§3.4 Results and a comparison of the finite difference and characteristic methods

In this section I shall discuss the accuracy of the schemes, their rate of convergence and stability and also compare the length of time each of them takes to run.

All the calculations were carried out on the University VAX 11/780 computer in double precision arithmetic. The definition of convergence for the characteristic algorithm is that the difference between two successive iterates should be less than 0.0001.

For the finite difference scheme this is changed to

$$\frac{\max_i (y^{r+1}_{i,j} - y^r_{i,j})}{\max_i y^{r+1}_{i,j}} \leq 0.00001$$

because in this case the solution along all of the j^{th} time step is found at once.

Where there are two alternative formulae for calculating y from p and q a check is made that the difference between the two values is less than this tolerance.

An attempt was made to decrease this tolerance to 0.00001 for the characteristic scheme but for n = 8; k = 0.01 it was found that no matter how accurately s_X and s_Y were

calculated the two formulae for y disagreed by more than 0.00001 because of the large step size which governs the accuracy of the approximation for the relation along the characteristics.

a) Accuracy

Figure 3.4 shows the general shape of the solution at times $\pi/2$, π, $3\pi/2$ and 2π. The cable remains virtually in a straight line with varying slope depending on the position and velocity of the end attached to the ship.

The exact results to 4 decimal places (predicted from the results for the characteristic and finite difference methods with n = 64; k = 0.00125) are given in Table 3.1 for t = 3, 6, and 9. Tables 3.2 and 3.3 give the errors at these times for the 8, 16 and 32 cases.

The errors for the characteristic and finite difference schemes decrease steadily in magnitude as the step size decreases. When the results for the 8, 16 and 32 cases increase for the characteristic method and decrease for the finite difference scheme (or vice versa), a good estimate of the exact solution is that it lies between the values obtained by the two schemes with step size 64. This gives the exact solution to at least 4 decimal places. Alternatively when both methods give oscillating results or when both sets of results decrease (or increase) agreement is so good that the exact solution can still be predicted to four decimal places.

b) Rate of convergence and stability

Each of the methods was also studied in terms of the convergence of the solution as the grid size tends to zero.

The results obtained are very similar in that the difference between the 8 and 16 results is approximately double that of the 16 and 32 results. The general pattern is shown in Figures 3.5 and 3.6 but the main points to observe are that in each case the difference is 2π periodic (although the phase differs between the schemes). For the 8-16 results the amplitude is approximately 1.6×10^{-3} and for the 16-32 results this has been halved, suggesting that the truncation error is of the order of the step size. This is as expected for the characteristic scheme since equations (3.25) and (3.26) use approximations for dp and dq which have leading terms $O(h+k)$ in the error.

For the finite difference method though, an error with leading term $O(h^2+k^2)$ is to be expected since central differences were used for all internal points. In fact the results for small t do show a quadratic rate of convergence. This does not continue because a lower order approximation for the boundary condition at the right hand end has been used (with truncation error $O(h+k)$) and this error soon permeates across the whole solution.

Stability does not cause problems with either of the two algorithms. A von Neumann analysis given in Chapter 1 suggests that the finite difference method should be unconditionally stable provided

$$-4\alpha(0.35-s) < 0 \quad \text{for} \quad 0 \leq s \leq .25$$

which it is.

The characteristic scheme also copes acceptably with large values of k/h. (For example n = 16; k = .02.) Details of the errors in these situations are given in Table 3.4.

c) Running times

An important consideration is the time taken to run the program as one of the original stipulations was that the calculation should not take a prohibitively long time. Details of running times are given in Table 3.5. Of the two schemes, the characteristic method is slightly quicker than the finite difference one when k/h is such that the characteristics cross the previous time level within h of the s coordinate of the new point. If the characteristic passes outside this point, though, the program becomes more complicated as tests must be carried out to find out which points the characteristic passes between.

At the boundaries further complications arise because there are no values for y, p, q outside (0, .25) and so the values where the characteristic crosses these boundaries must be used instead. This immediately destroys the uniformity in the time steps for the calculations at these points, complicates the equations to be solved and thus increases the time required. The finite difference method encounters none of these difficulties and so this is now the slightly faster scheme.

As discussed in the previous subsection, stability considerations do not enter into the choice of h and k for either of the two schemes. Because of this it is possible

by taking, for example, n = 16; k = 0.2 to achieve much lower CPU times. The answers obtained in this case are still very accurate and are given in Table 3.4; the running time is shown in Table 3.5.

As detailed in section 3.3 two different approaches for calculating the position of the characteristics are available. The results for the standard method are indistinguishable from those of the alternative method to the desired accuracy but the alternative method is faster than the standard one. (See Table 3.5.). This speed is due to the simplicity of the calculation of s_X and s_Y in this case. Each step of the iteration requires 7 multiplications/divisions in both schemes but only 6 additions/subtractions are needed per iteration step for the alternative scheme as opposed to 10 additions/subtractions and a logarithm for the standard scheme. A graph of some of the characteristics is shown in Figure 3.7. From this it can be seen that the characteristics are almost straight lines. Thus, the quadratic convergence of the Newton method used in the standard calculations would not have a significant advantage over the linear method used in the other scheme. Therefore the number of iterations required in both cases is comparable (and usually about 2). This explains the difference in the time taken by the two approaches.

§3.5 Minor adaptations and their effects

Three slight changes to the model or its method of solution are discussed below. They include the use of quadratic, instead of liner, interpolation, a refinement of

the way in which the right hand boundary condition is tackled and changes in the initial conditions or one of the key parameters in the model.

a) The use of quadratic interpolation

The characteristic program can be adapted so that the interpolation carried out to find y, p and q on the previous time step is quadratic rather than linear. For both of the characteristics through the i^{th} point on the new time level, the interpolation points used are the (i-1), i and (i+1) points on the old time level. Obviously at the boundaries the interpolation points had to be the 0, 1 and 2 points or the (n-2), (n-1) and n points (as applicable). The errors for n = 8; k = 0.01 and n= 16; k = 0.005 at t = 3 are given in Table 3.6.

A comparison with Table 3.3 shows that these errors are very similar to those of the linear interpolation results at this time. This reinforces the suggestion of section 3.4 that the dominant error does not arise from the calculation of s_X and s_Y or values at these points but from the approximation to the relationships along the characteristics, which is essentially linear. Even if this were not the case it is unlikely that quadratic interpolation would be worthwhile as the solution is almost a straight line.

b) Improvement of the boundary condition at the right hand end

It was felt that for the difference scheme a large part of the error was due to the first order accuracy of the approximation for the right hand boundary condition.

In an attempt to improve this the new approximation

$$\tfrac{1}{4}(y_s)_{n,j+1} + \tfrac{1}{2}(y_s)_{n,j} + \tfrac{1}{4}(y_s)_{n,j-1} + (y_t)_{n,j} = 0$$

was used. This not only involves the unknowns $y_{n,j+1}$ and $y_{n-1,j+1}$ but also $y_{n+1,j+1}$, which lies outside the boundary.

However, putting i = n in the implicit finite difference equation gives an extra independent equation for these three unknowns. The system is now, therefore, n+1 equations for n+1 unknowns. Unfortunately the matrix is no longer tridiagonal since the last equation involves three unknowns. However, $y_{n+1,j+1}$ is easily eliminated to give a tridiagonal system as before and so the program is no more complicated. The results are shown in Table 3.7. The table shows the error in the results at t = 6 with n = 8; k = 0.01 and n = 8; k = 0.05.

The errors with the less accurate boundary approximation for n = 8; k = 0.01 are given for comparison. As can be seen this adaptation does indeed improve the accuracy significantly and the leading term in the error is now $0(h^2) + 0(k^2)$.

c) Small changes to the model

As indicated in section 3.1 there is a possibility that the oscillations of the ship might be faster than modelled. To study the effect of this the boundary condition was changed to $y = 1 - \cos(2\pi t)$ at s = 0. The results of this change are shown in Figure 3.8. As expected the cable at the far end is now lagging behind the motion at the front end. The shape is also more curved.

The effect of decreasing the amplitude of this new boundary condition is shown in Figure 3.9. The motion of the

ship is now considered to be $y = .1(1-\cos(2\pi t))$. Again no difficulties arise.

Finally in Figure 3.10 the results have been plotted for the boundary condition $y = 1 - \cos(2\pi t)$ with $A = .25$ (instead of .35). The program copes adequately with this adaptation and there seems to be no reason why a general boundary condition of the form $y = y^{(0)}[1-\cos\lambda t]$, where $y^{(0)}$ and λ are given parameters, should cause any difficulty.

§3.6 Concluding comments

Both the method of characteristics and a finite difference method have been used to solve the cable problem. Despite the weak singularity in the second derivative of the solution at $s = 0$, $t = 0$, the implicit finite difference method gives acceptable answers. The problem also has a singularity when $|y_s+y_t| = 0$. This term is multiplied by y_s+y_t and hence a discontinuity occurs if $(y_s+y_t)|y_s+y_t|$ is differentiated twice. This gives a discontinuity in the fourth derivative of y which is such a weak singularity that no problems are caused by using the finite difference scheme.

Both methods, therefore, cope adequately with this problem and the results of using either program are similar in accuracy, rate of convergence, stability and running time.

The only difference between the two schemes seems to be the ease with which they can be adapted to cope with variations in the model of the problem. The finite difference method is easily adapted if the boundary conditions or initial conditions are changed, for instance; or indeed if the fundamental partial differential equation is changed.

In contrast, the alterations to the characteristic method may be more difficult. For example, if the boundary condition at the right hand end is investigated and changed the characteristic equations need to be re-examined in detail as at present use has been made of the fact that $p + q = 0$ at the right hand end when simplifying the equations.

However, if changes to the model introduced more serious singularities into the equations it is possible that the finite difference method would cope badly since it smoothes out discontinuities. In this case it would be likely that the characteristic method would be the more suitable of the two.

TABLE 3.1

s	t=3	t=6	t=9
0.00000	1.9900	0.0398	1.9111
0.03125	1.9815	0.0524	1.8947
0.06250	1.9724	0.0655	1.8778
0.09375	1.9628	0.0792	1.8604
0.12500	1.9527	0.0933	1.8425
0.15625	1.9420	0.1079	1.8243
0.18750	1.9310	0.1228	1.9057
0.21875	1.9196	0.1381	1.7868
0.25000	1.9080	0.1536	1.7678

THE EXACT RESULTS y(s,t) TO FOUR DECIMAL PLACES

where s is the dimensionless distance along the cable

and y(s,t) is the dimensionless horizontal displacement of the cable.

TABLE 3.2

s	t=3			t=6			t=9		
	n=8	16	32	n=8	16	32	n=8	16	32
0.00000	0	0	0	0	0	0	0	0	0
0.03125	0	0	0	0	0	0	0	0	0
0.06250	-1	0	0	0	0	0	0	0	0
0.09375	-1	0	0	1	0	0	0	0	0
0.12500	0	0	0	0	0	0	-1	-1	-1
0.15625	-2	-1	-1	1	1	0	-1	-1	-1
0.18750	-2	-1	-1	1	0	0	-2	-1	-1
0.21875	-2	-1	-1	2	1	0	-2	-2	-2
0.25000	-3	-2	-1	2	1	1	-2	-1	-1

ERRORS FOR THE FINITE DIFFERENCE METHOD IN UNITS OF 10^{-4}

TABLE 3.3

s	t=3			t=6			t=9		
	n=8	16	32	n=8	16	32	n=8	16	32
0.00000	0	0	0	0	0	0	0	0	0
0.03125	1	0	0	-1	-1	0	1	0	0
0.06250	1	1	0	-2	-1	-1	2	1	1
0.09375	2	1	0	-2	-1	0	2	1	1
0.12500	3	2	1	-3	-1	-1	2	1	0
0.15625	3	1	0	-3	-1	-1	3	1	0
0.18750	3	1	1	-3	-1	-1	3	1	0
0.21875	3	1	0	-3	-2	-1	2	1	0
0.25000	3	1	0	-3	-2	-1	3	1	1

ERRORS ($\times 10^4$) FOR THE CHARACTERISTIC METHOD

TABLE 3.4

	t=3		t=6		t=9	
	(i)	(ii)	(i)	(ii)	(i)	(ii)
0.00000	0	0	0	0	0	0
0.03125	0	1	0	-2	0	1
0.06250	-1	2	0	-2	0	2
0.09375	-1	2	1	-2	-1	2
0.12500	-2	4	2	-3	-2	2
0.15625	-3	3	3	-4	-3	3
0.18750	-4	4	3	-5	-4	3
0.21875	-6	5	5	-5	-5	3
0.25000	-8	5	7	-5	-6	4

ERRORS ($\times 10^4$) FOR

(i) FINITE DIFFERENCES n=16; k=0.02 (ii) CHARACTERISTICS n=16; k=0.02

TABLE 3.5

	n=8	n=16	n=32
	k=0.01	k=0.005	k=0.0025
Finite differences	3.65	12.82	47.28
Characteristics (Standard)	3.36	10.93	37.80
Characteristics (Alternative)	2.73	7.18	26.63

	n=8	n=16		n=32
	k=0.02	k=0.02	k=0.01	k=0.005
Finite differences	2.09	3.36	5.77	19.19
Characteristics (Standard)	2.54	5.11	9.86	33.16

CPU - TIME (IN SEC) TO RUN THE PROGRAMS AS FAR AS t=2

	n=8	n=16
	k=0.01	k=0.005
0.00000	0	0
0.03125	1	0
0.06250	1	1
0.09375	2	1
0.12500	3	1
0.15625	2	2
0.18750	3	1
0.21875	4	1
0.25000	4	2

ERRORS ($x10^4$) AT t=3 FOR THE CHARACTERISTIC SCHEME USING QUADRATIC INTERPOLATION

TABLE 3.7

s	Exact solution	New boundary approximation		Old boundary approximation
		n=8; k=0.01	n=8; k=0.05	n=8; k=0.01
0.00000	0.0398	0	0	0
0.03125	0.0524	0	0	0
0.06250	0.0655	0	1	0
0.09375	0.0792	0	0	1
0.12500	0.0933	0	1	0
0.15625	0.1079	0	1	1
0.18750	0.1228	1	2	1
0.21875	0.1381	0	2	2
0.25000	0.1536	0	2	2

ERRORS ($x10^4$) FOR THE FINITE DIFFERENCE SCHEME

Figure 3.4

Cable shape at $t=\pi/2, \pi, 3\pi/2$ and 2π

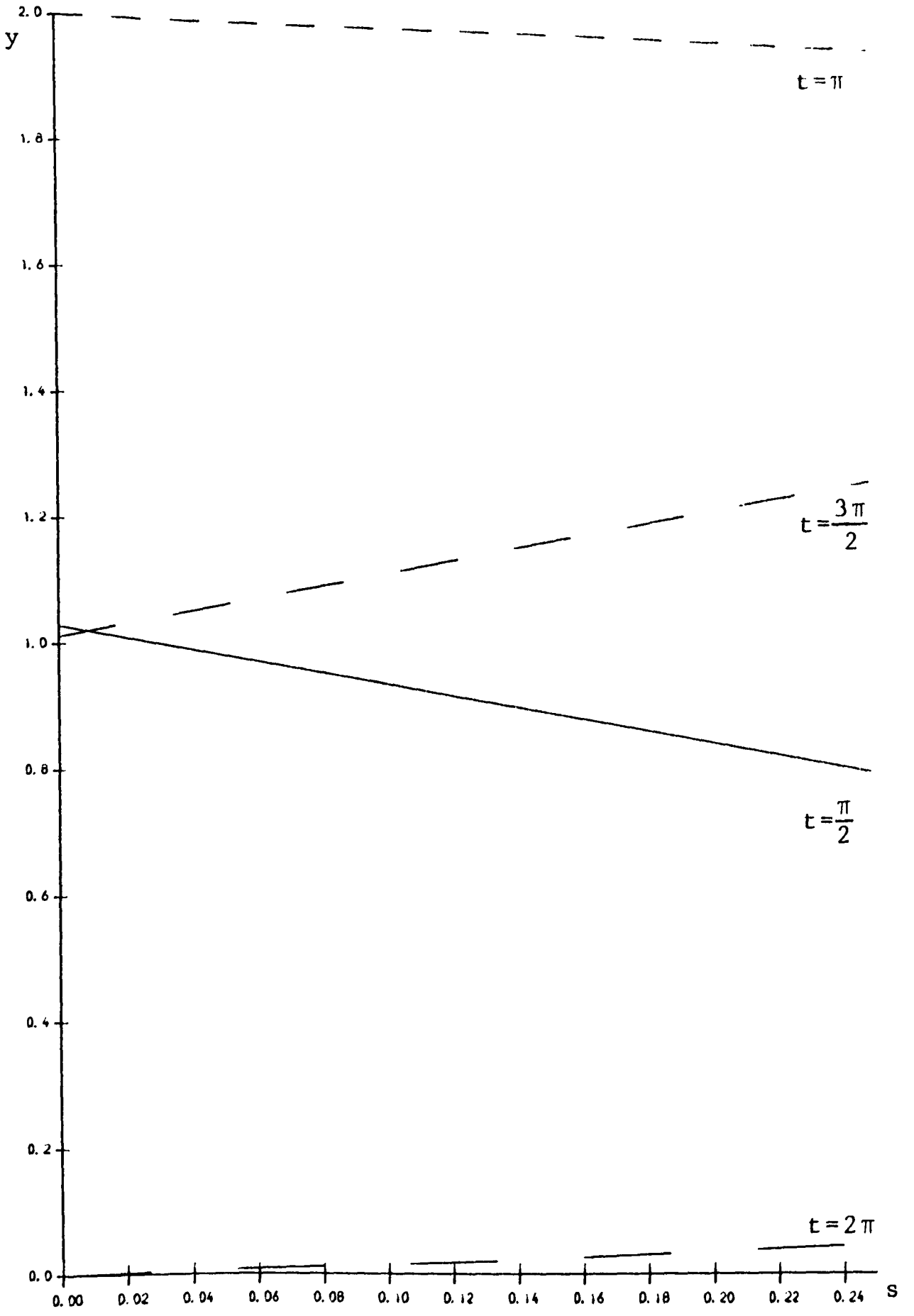


Figure 3.5

Differences between results for different step sizes for s=0.21875

—————— n = 8 and 16
- - - - - - - n = 16 and 32
} Characteristic method

— — — — n = 8 and 16
— — — n = 16 and 32
} Finite difference method

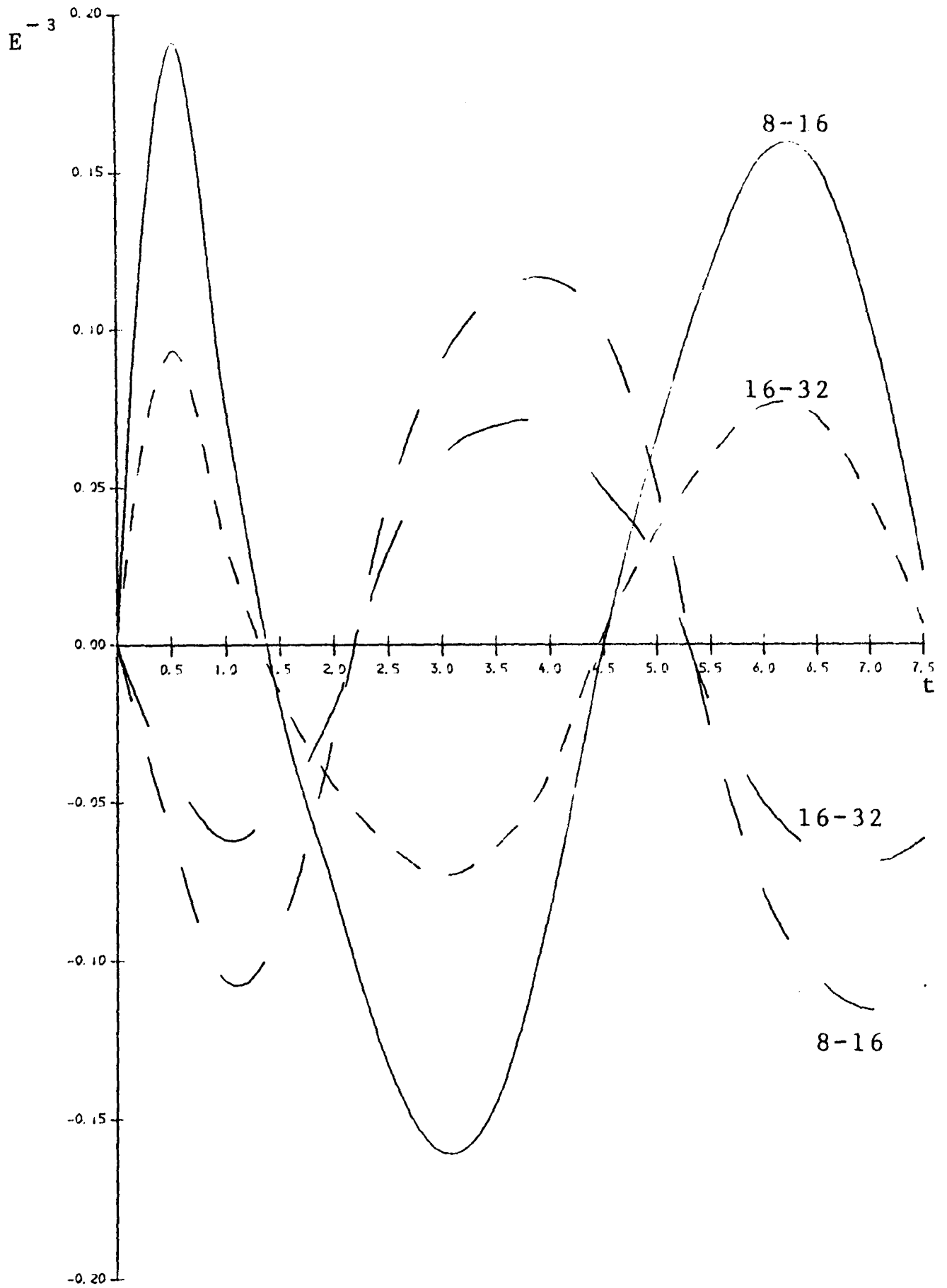


Figure 3.6

Differences between results for n=8 and n=16 at t=3.5

———— Characteristic method

— — — — Finite difference method

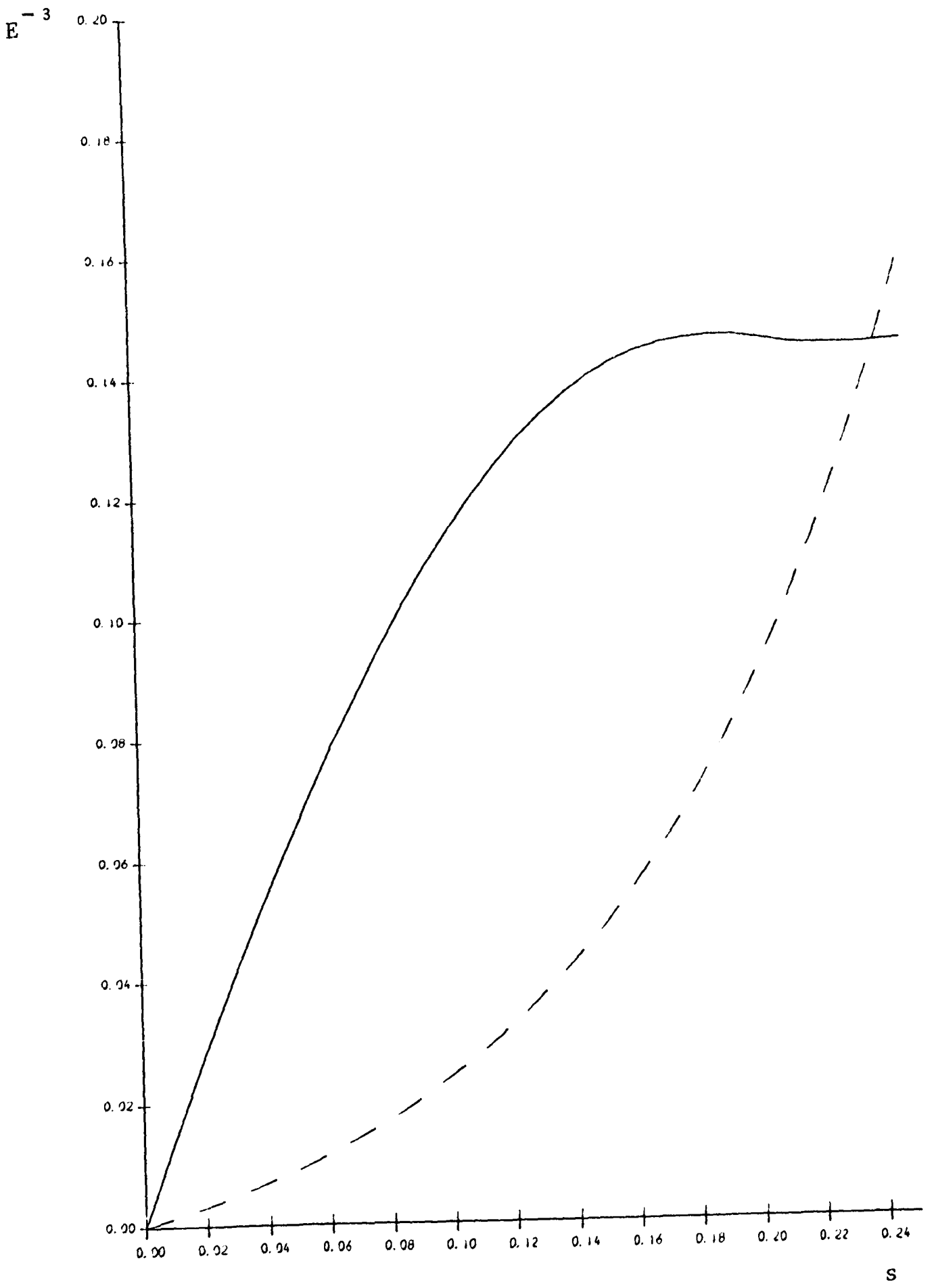


Figure 3.7

Characteristics

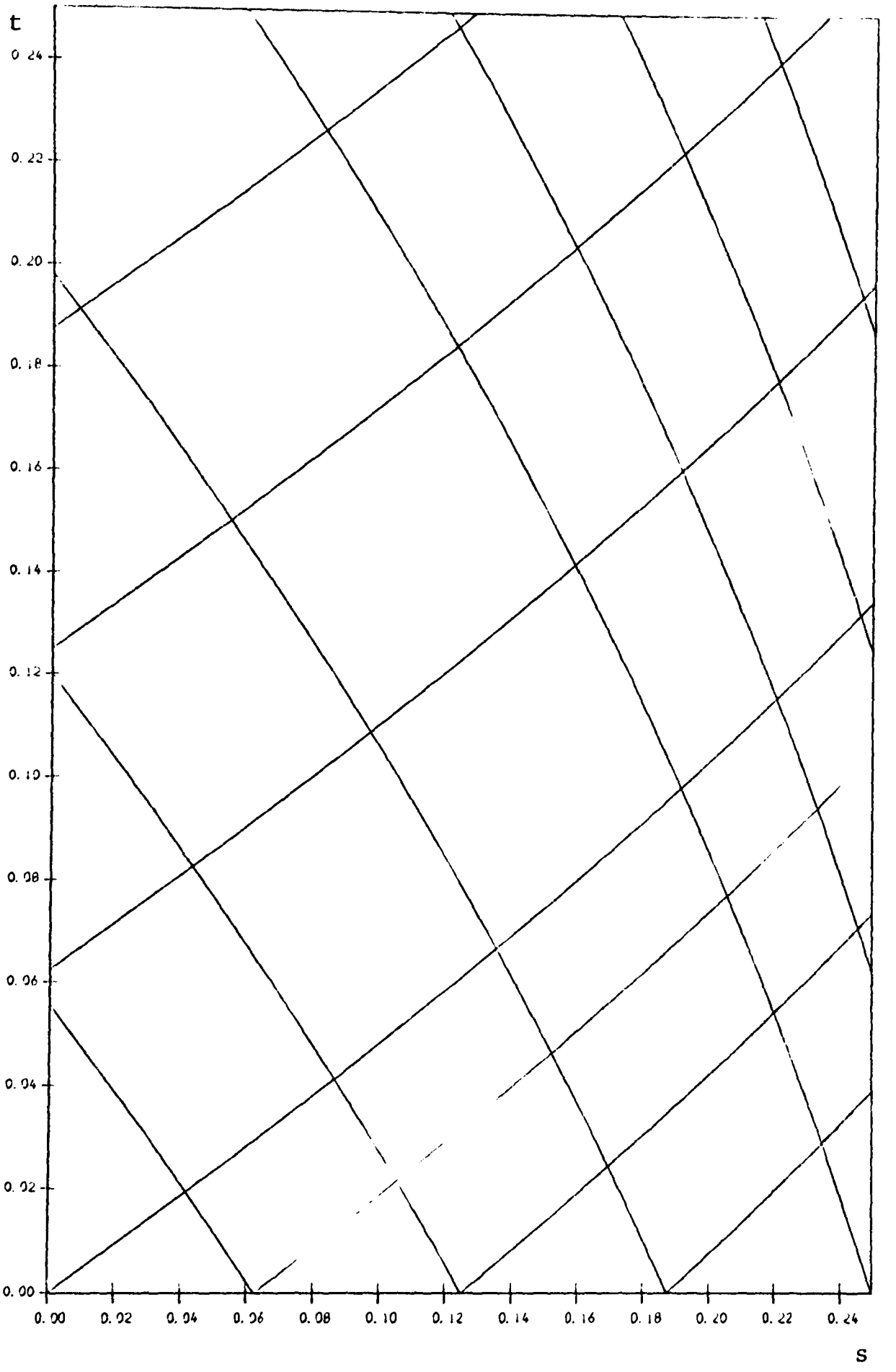


Figure 3.8

Solution with boundary condition $y=1-\cos(2\pi t)$

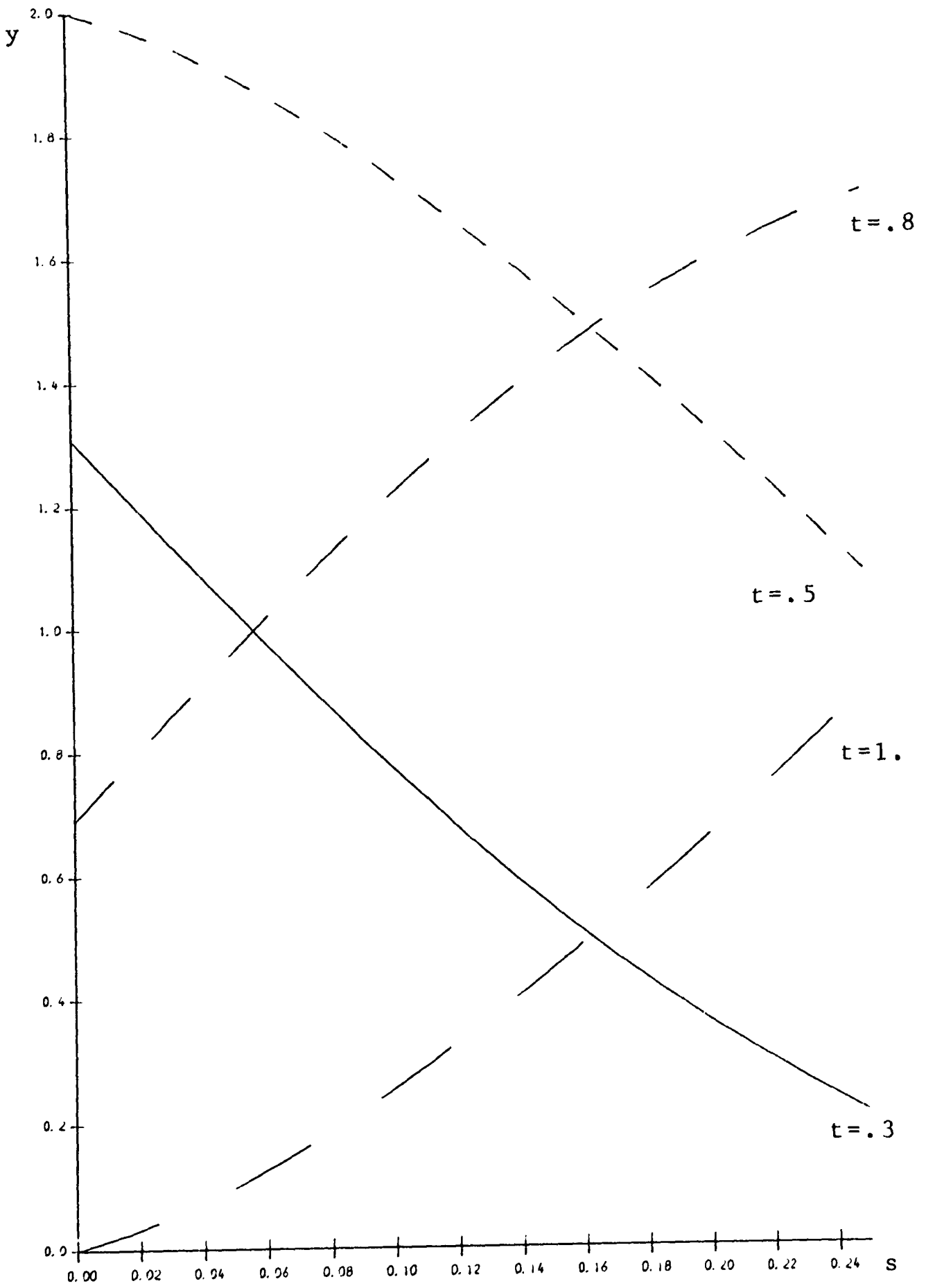


Figure 3.9

Solution with boundary condition $y=.1(1-\cos(2\pi t))$

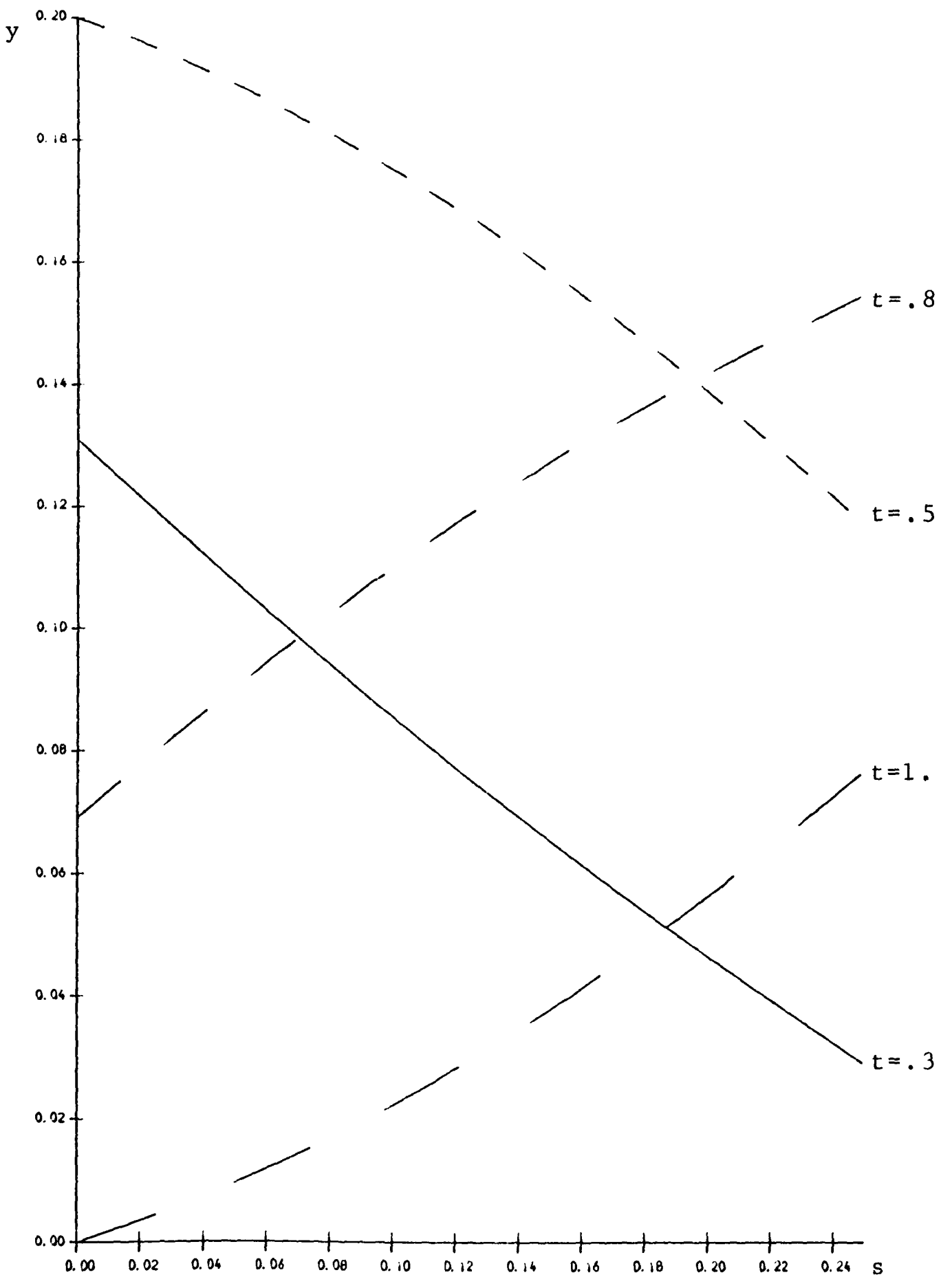
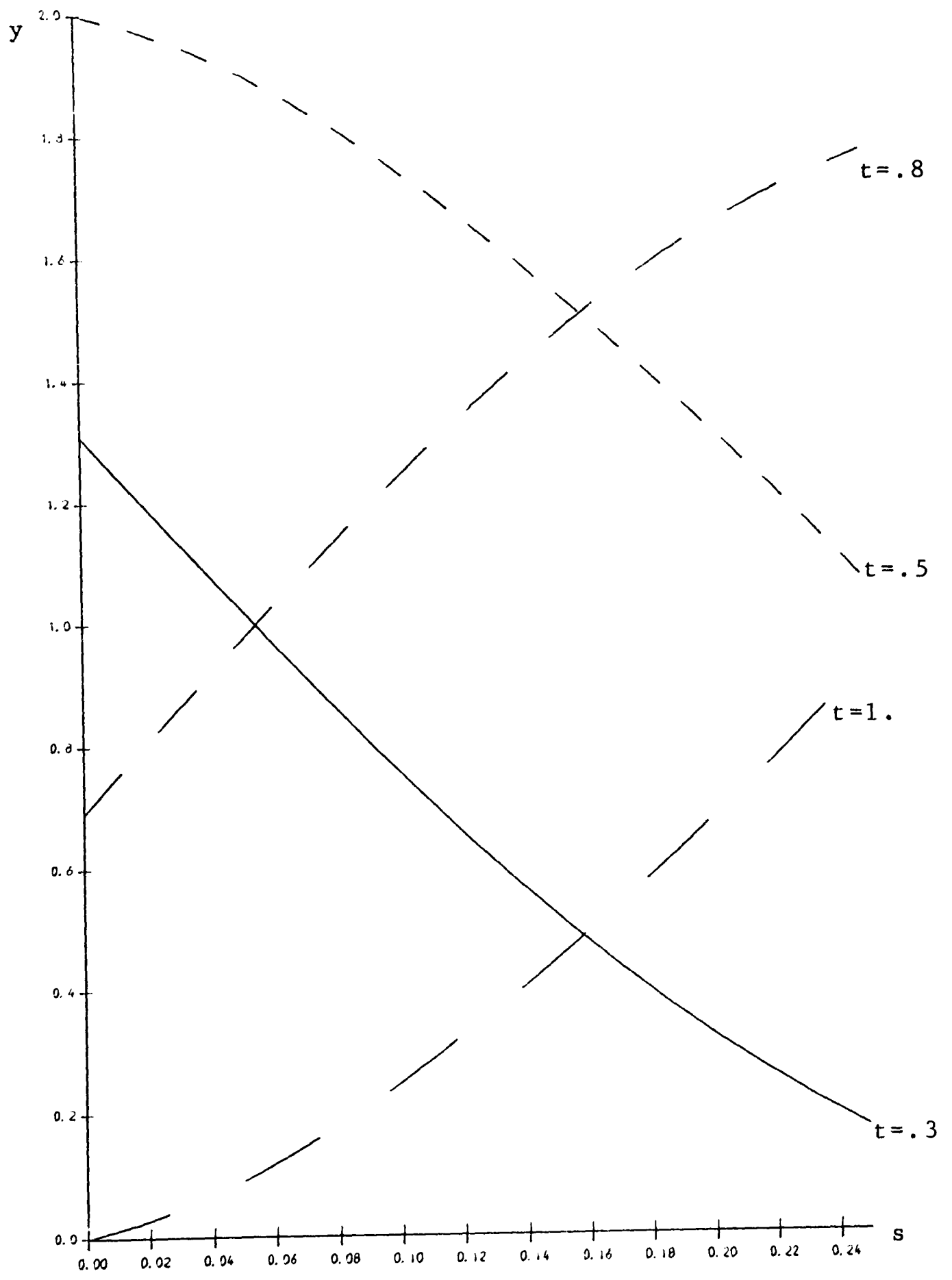


Figure 3.10

Solution with boundary condition $y=1-\cos(2\pi t)$ and A=.25



CHAPTER 4

A simplified version of the thermal runaway problem

§4.1 Introduction

In this chapter I shall consider the solution of the following equation

$$x(u_{xx}+u_{xt}) + (\alpha+1)u_x + u_t = 0 \qquad (4.1)$$

where $\alpha > 0$ is a constant

with initial condition $u(x,0) = u_0(x)$, $x \geq 0$ (a bounded given function with $u_0(0)=0$) and the boundary condition that $u(0,t)$ is regular for $t \geq 0$.

The motivation behind this study is that this equation shares some features with the equation of Chapter 5, but is very much easier to deal with than that equation. The experience gained in solving this equation provides an insight into possible ways of tackling the problem presented in the next chapter.

In the next section an exact solution to the equation is considered and then the following two sections present a numerical method for the solution of the equation followed by a discussion of the results obtained when using this method.

By considering the characteristics of (4.1), we see that the line $x=t$ is special, and that on the line $x=0$ (4.1) is degenerate. We are interested only in solutions of (4.1) that are smooth across $x=t$.

§4.2 The Exact Solution

Equation (4.1) has the following exact solution

$$u(x,t) = \begin{cases} \frac{1}{x^{\alpha}} \left[\int_0^t f(\tau)(\tau+x-t)^{\alpha-1} d\tau + u_0(x-t)(x-t)^{\alpha} \right] & x > t \quad (4.2) \\ \frac{1}{x^{\alpha}} \int_{t-x}^t f(\tau)(\tau+x-t)^{\alpha-1} d\tau & x < t \quad , \quad (4.3) \end{cases}$$

where $f(\tau)$ is an arbitrary smooth function for $\tau \geq 0$ with $f(0) = u_0(0) = 0$.

It is easy to check that this solution satisfies equation (4.1).

It now remains to check the initial conditions and the condition along $x = 0$.

If $x > 0$,

$$u(x,0) = \lim_{t \to 0} \frac{1}{x^{\alpha}} \left[\int_0^t f(\tau)(\tau+x-t)^{\alpha-1} d\tau + u_0(x-t)(x-t)^{\alpha} \right]$$

$$= u_0(x) \text{ since } f(0) \text{ is bounded.} \quad (4.4)$$

We must now show that the initial condition is satisfied at the origin.

If we approach the origin along any line $x - mt = 0$ with $m > 1$ then (4.2) gives

$$u(0,0) = \lim_{x \to 0} \frac{1}{x^{\alpha}} \left[\int_0^{x/m} f(\tau)(\tau+x-\frac{x}{m})^{\alpha-1} d\tau + u_0(x-\frac{x}{m})(x-\frac{x}{m})^{\alpha} \right]$$

$$= 0 \, . \quad (4.5)$$

We, therefore, also require that $\lim_{x \to 0} u(x, \frac{x}{m}) = u_0(0)$ when $m < 1$ (that is, $t > x$).

Equation (4.3) gives

$$\lim_{x\to 0} u(x,\frac{x}{m}) = \frac{1}{x^{\alpha}} \int_{\frac{x}{m}-x}^{\frac{x}{m}} f(\tau)(\tau+x-\frac{x}{m})^{\alpha-1} d\tau$$

$$= \frac{1}{x^{\alpha}} . x . f(\frac{x}{m}) x^{\alpha-1} = f(\frac{x}{m}) = f(0).$$

From (4.5) it is seen that we require $f(0) = u_0(0) = 0$.

We now consider the solution at the left hand boundary, $x = 0$.

From (4.3),

$$u(0,t) = \lim_{x\to 0} \frac{1}{x^{\alpha}} \int_{t-x}^{t} f(\tau)(\tau+x-t)^{\alpha-1} d\tau$$

$$= \lim_{x\to 0} \frac{1}{x^{\alpha}} x . f(t) x^{\alpha-1} = f(t) . \qquad (4.6)$$

In order to satisfy the condition along this boundary we need the condition that $f(t)$ is a regular function of t.

There is particular interest in the case when $\alpha = 1$ and so all further analysis will confine itself to this situation. The numerical work presented in sections 4.3 and 4.4 limits itself to the case $\alpha = 1$ although the method could be easily extended to other values of α.

In this case, equation (4.1) becomes

$$x(u_{xx}+u_{xt}) + 2u_x + u_t = 0 \qquad (4.7)$$

with $u(x,0) = u_0(x)$ $x \geq 0$, $u_0(0) = 0$

and $u(0,t)$ regular for $t \geq 0$.

This has exact solution

$$u(x,t) = \begin{cases} \frac{1}{x}[F(t)+u_0(x-t)(x-t)] & x > t \\ \\ \frac{1}{x}[F(t)-F(t-x)] & x < t \end{cases} \tag{4.8}$$

where $F(t) = \int_0^t f(\tau)d\tau$ and $f(\tau)$ is an arbitrary smooth function with $f(0) = u_0(0) = 0$.

⟵ Since $f(\tau)$ is smooth we have $F(0) = 0$, $F'(0) = f(0) = u_0(0)$ and $F(t)$ is a continuous function of t.

⟵ Thus, the condition along $x = 0$ ensures that there are no discontinuities in $u(x,t)$ when $t > x$.

⟵ Since $F(0) = 0$ and $u_0(0)$ is finite (because $u_0(0) = f(0)$ which is finite) it is clear that $u(x,t)$ is continuous across the line $x = t$.

In the region $x > t$, discontinuities in the solution can only arise as a result of discontinuities in the initial condition $u_0(x)$. A discontinuity in $u_0(x)$ at $x = x^*$ is propagated along the line $x - t = x^*$.

The solution given by (4.8) exists in the upper right-hand quarter plane, $x \geq 0$, $t \geq 0$. We now consider the effect of confining the solution to $1 \geq x \geq 0$, $t \geq 0$ and imposing a boundary condition $u(1,t) = u_1(t)$. It will be shown that the chosen boundary condition must satisfy certain requirements if the condition along $x = 0$ is not to be violated.

Since there must be no discontinuities in $u(x,t)$, when $t \geq x$ we must have $u_1(t)$ continuous for $t \geq 1$. When $t < 1$,

equation (4.8) gives

$$u_1(t) = F(t) + u_0(1-t)(1-t). \tag{4.9}$$

The continuity of $F(t)$ forces $u_1(t)$ to have discontinuities to 'match' the discontinuities of $u_0(1-t)$. That is, a discontinuity at $x = x^*$ is carried along the line $x - t = x^*$ and manifests itself as a discontinuity on the boundary at $t = 1 - x^*$.

Finally when $t = 0$ we have (from (4.9)) $u_1(0) = u_0(1)$ (since $F(0)=0$).

← There are thus three conditions on $u_1(t)$. Namely

$$\begin{aligned} &\text{(i)} \quad u_1(0) = u_0(1) \\ &\text{(ii)} \quad u_1(t) \text{ continuous for } t \geq 1 \\ &\text{(iii)} \quad u_1(t) - u_0(1-t)(1-t) \text{ continuous for } t < 1. \end{aligned} \tag{4.10}$$

Once these three conditions are satisfied the choice of $u_1(t)$ determines $F(t)$ and hence $u(x,t)$ uniquely.

← For, from (4.8)

$$\begin{aligned} F(t) &= u_1(t) - u_0(1-t)(1-t) \quad \text{for } t < 1 \\ \text{and,} & \\ F(t) &= u_1(t) + F(t-1) \qquad\quad \text{for } t \geq 1 . \end{aligned} \tag{4.11}$$

It is interesting to note that because we have specified $F'(0) = u_0(0)$ too, then $f(t)$ is uniquely determined also.

We can also consider the uniqueness of the solution in a slightly different light. If we return to the infinite problem (that is, without the right hand boundary), then provided that $u_0(x)$ has a convergent Taylor series expansion for all x and if

we stipulate that the solution $u(x,t)$ must be infinitely differentiable across the line $x = t$, then it is easy to show inductively that $F(0)$ and all its derivatives at the origin are uniquely determined by (4.8). Hence, if F is limited to the set of functions which have convergent Taylor series expansions it is uniquely determined and so the solution $u(x,t)$ of the form given by (4.8) is unique. In fact from (4.8) we see that $F(x) = xu_0(-x)$, ($u_0(-x)$ is defined since we stipulate that u_0 has a convergent Taylor series expansion for all x).

Returning to the bounded problem and combining this with (4.11) we see that if we want a solution which is infinitely differentiable across the line $x = t$ we must put

$$u_1(t) = tu_0(-t) + u_0(1-t)(1-t).$$

This choice of $u_1(t)$ satisfies the three conditions (4.10) given above as follows:

(i) $u_1(0) = 0 + u_0(1),\ 1 = u_0(1)$,

(ii), (iii) $u_1(t)$ is continuous for all t since u_0 is continuous everywhere .

In the numerical work of section 4.4 we have applied the method given in section 4.3 to the problem with initial condition $u_0(x) = -xe^{-x}$ and have taken $u_1(t) = t^2e^t - (1-t)^2e^{t-1}$ so that the exact solution, $u(x,t) = \frac{1}{x}[t^2e^t-(x-t)^2e^{t-x}]$ is infinitely differentiable across the line $x = t$.

§4.3 A numerical algorithm to solve the equation

In this section equation (4.1) will be solved using the method of characteristics. This is a straightforward application of the method outlined in Chapter 1 although care must be taken at the left hand boundary. Using the notation of Chapter 1 we have

$$au_{xx} + bu_{xt} + cu_{tt} + e = 0$$

where $a = b = x$

$c = 0$

$e = (\alpha+1)u_x + u_t$.

This has characteristics whose slope is given by

$$\frac{dt}{dt} = + \frac{b \pm \sqrt{b^2-4ac}}{2a} = \frac{x \pm x}{2x} = 1, 0.$$

So the two families of characteristics are

t = constant and $t - x$ = constant,

along which we have the relationships

$$x\left(\frac{dp}{dx}+\frac{dq}{dx}\right) + (\alpha+1)p + q = 0$$

and $x\frac{dp}{dx} + (\alpha+1)p + q = 0$ respectively.

(Here, as elsewhere, $p = u_x$, $q = u_t$.)

Rearranging these expressions we obtain

$$\frac{d}{dx}(xp) + \frac{d}{dx}(xq) + \alpha p = 0 \quad \text{along } t = \text{constant} \qquad (4.9)$$

and

$$\frac{d}{dx}(xp) + \alpha p + q = 0 \quad \text{along } t - x = \text{constant}. \qquad (4.10)$$

We should now consider the way in which we can incorporate the boundary and initial conditions into the method of solution.

The solution is specified along $t = 0$ for $x \geq 0$, (that is, $u(x,0) = u_0(x)$, $x \geq 0$ is a given function). Differentiating this with respect to x gives p along the line $t = 0$, $x \geq 0$. We also notice that the initial line $t = 0$ is a characteristic in the family t = constant and so equation (4.9) holds along this line. This enables us to calculate q initially provided that a value for q is prescribed at some point along the initial line.

The condition along $x = 0$, $t \geq 0$, namely that u is infinitely differentiable along this line and hence u_{xx} and u_{xt} are bounded gives

$$(\alpha+1)p + q = 0 \text{ when } x = 0,\ t \geq 0 \text{ from (4.1)}.$$

We are interested in finding a solution which is smooth across the line x = t and so, in particular, at the origin we must have no discontinuities in q and p and so we must specify $q(0,0) = -(\alpha+1)p(0,0)$.

As shown in the previous section, if we limit ourselves to infinitely differentiable solutions then we have enough information to solve the equation in the quarter plane $t \geq 0$, $x \geq 0$. In practice, however, this would be difficult for two reasons. Firstly, we would need to differentiate equation (4.1) repeatedly and evaluate along the line x = 0 to obtain expressions for all the higher derivatives of the solution. Also, for any practical method we must insert a right hand

boundary. For the results presented in the next section this boundary is taken to be the line x = 1. The method could, however, be extended for larger x without difficulty.

As shown above, if we wish to find the infinitely differentiable solution we must take $u(1,t) = tu_0(-t) + u_0(1-t)(1-t)$ along the right hand boundary.

The algorithm presented below is based on a grid in which the step size, h, is the same in both the x and t directions. This is because the sloping characteristics (which are the straight lines t - x = constant) will then pass through grid points. This simplifies the explanation of the algorithm although the same procedure could be adapted with some modifications if the step size were not the same in the x and t directions. Basically an interpolative step after calculations have been done at characteristic grid points on the new time step to find values at standard grid points would be necessary.

In the equations below the subscripts i,j denote evaluation at the point x = ih, t = jk (that is, $p_{i,j}$ is the approximation to p at x = ih, t = jk, for example). We suppose that the solution is known along the line t = jk and try to calculate it along t = (j+1)k.

Discretizing (4.9) gives

$$\frac{(i+1)hp_{i+1,j+1} - ihp_{i,j+1}}{h} + \frac{(i+1)hq_{i+1,j+1} - ihq_{i,j+1}}{h}$$

$$+ \frac{\alpha}{2}(p_{i+1,j+1}+p_{i,j+1}) = 0, \qquad (4.11)$$

and similarly from (4.10) we obtain

$$\frac{(i+1)hp_{i+1,j+1} - ihp_{i,j}}{h} + \frac{\alpha}{2}(p_{i+1,j+1}+p_{i,j}) + \tfrac{1}{2}(q_{i+1,j+1}+q_{i,j}) = 0. \tag{4.12}$$

Subtracting (4.12) from (4.11) gives

$$i(p_{i,j}-p_{i,j+1}) + (i+1)q_{i+1,j+1} - iq_{i,j+1} + \frac{\alpha}{2}(p_{i,j+1}-p_{i,j}) - \tfrac{1}{2}(q_{i+1,j+1}+q_{i,j}) = 0,$$

which can be rearranged to give

$$q_{i+1,j+1} - \frac{i}{i+\frac{1}{2}}q_{i,j+1} - \frac{(i-\frac{\alpha}{2})}{i+\frac{1}{2}}p_{i,j+1} = -\frac{(i-\frac{\alpha}{2})}{i+\frac{1}{2}}p_{i,j} + \frac{\frac{1}{2}}{i+\frac{1}{2}}q_{i,j}, \tag{4.13}$$

and from (4.12) we have

$$p_{i+1,j+1} + \frac{\frac{1}{2}q_{i+1,j+1}}{i+1+\frac{\alpha}{2}} = \frac{i-\frac{\alpha}{2}}{i+1+\frac{\alpha}{2}}p_{i,j} - \frac{\frac{1}{2}q_{i,j}}{i+1+\frac{\alpha}{2}}. \tag{4.14}$$

Equations (4.13) and (4.14) are valid for $0 \leq i \leq N-1$ where $Nh = 1$.

From (4.6), the boundary condition at the left hand end we have

$$(\alpha+1)p_{0,j+1} + q_{0,j+1} = 0, \tag{4.15}$$

and using (4.15) together with $(\alpha+1)p_{0,j} + q_{0,j} = 0$ in (4.13) when $i = 0$ gives

$$q_{1,j+1} - \frac{\alpha}{\alpha+1}q_{0,j+1} = -p_{0,j} \tag{4.16}$$

and similarly from (4.14) we have

$$p_{1,j+1} + \frac{q_{1,j+1}}{\alpha+2} = \frac{p_{0,j}}{\alpha+2} \quad . \tag{4.17}$$

Since $u(1,t) = u_1(t)$ is a given function we know $q(1,t)$, that is $q_{N,j}$ for all j.

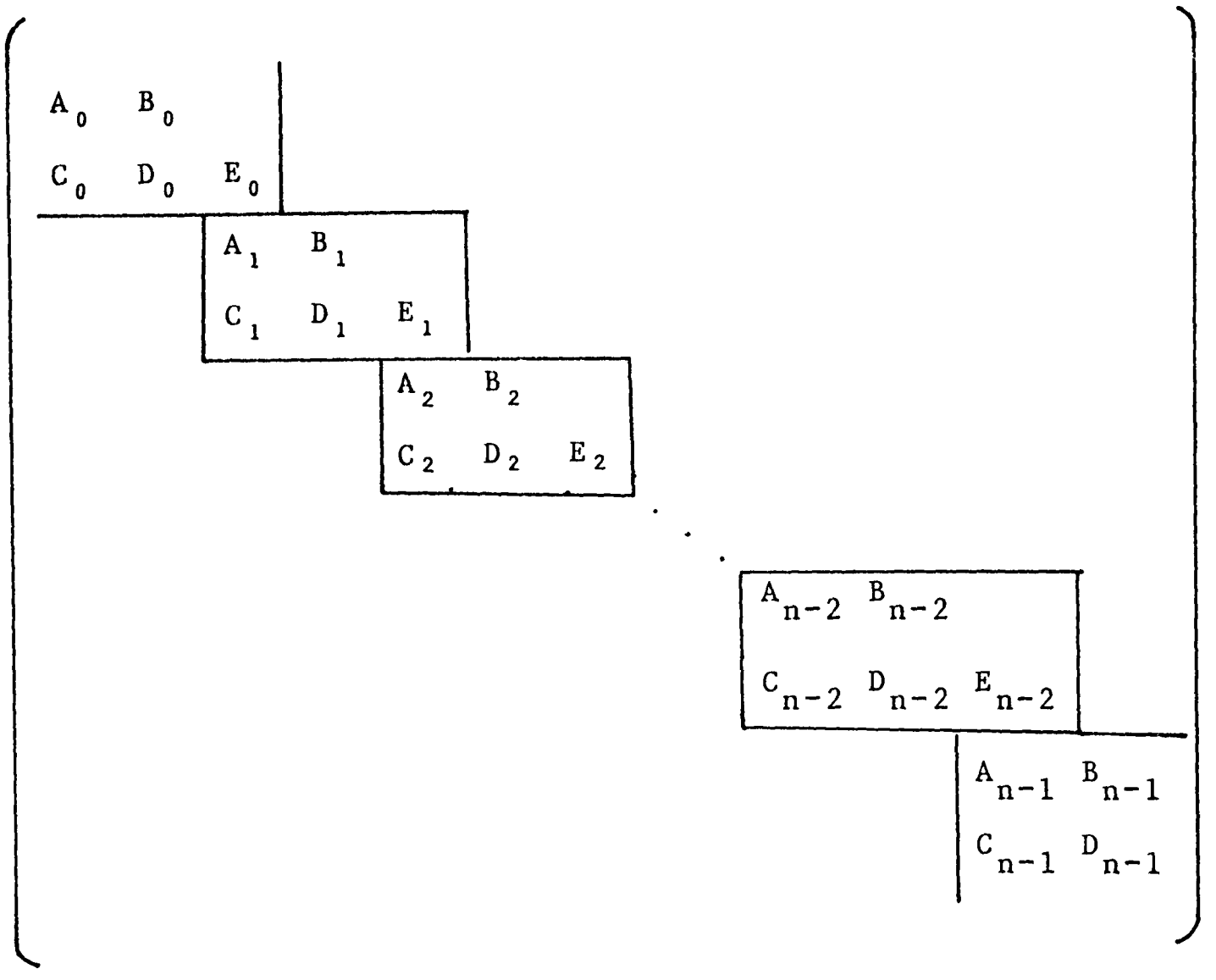
Using (4.14) with $i = N-1$ we can use $q_{N,j+1}$ to find $p_{N,j+1}$.

We now need to find $p_{0,j+1}, p_{1,j+1} \cdots, p_{N-1,j+1}$ and $q_{0,j+1}, q_{1,j+1}, \cdots, q_{N-1,j+1}$. That is, we have 2n unknowns. We use 2n equations to calculate these unknowns, namely (4.15), (4.16) and (4.17) together with (4.13) and (4.14) for $1 \leq i \leq N-2$ and (4.13) with $i = N-1$.

This sytem can be written as a matrix equation

$$M' \underline{x}_{j+1} = \underline{b}'$$

where M' is a 2n x 2n matrix and has the following structure.



and $\underline{x}_{j+1} = (q_{0,j+1}, p_{0,j+1}, q_{1,j+1}, p_{1,j+1}, \ldots, q_{n-1,j+1}, p_{n-1,j+1})^T$

$$\underline{b}' = (R_0, S_0', R_1, S_1', \ldots, R_{n-1}, S_{n-1}')^T$$

with

$$A_0 = \frac{1}{\alpha+1},\ B_0 = 1,\ C_0 = \frac{-\alpha}{\alpha+1},\ D_0 = 0,\ E_0 = 1$$

$$R_0 = 0,\ S_0' = -p_{0,j}$$

and

$$A_1 = \frac{1}{2+\alpha},\ B_1 = 1,\ R_1' = \frac{p_{0,j}}{2+\alpha}$$

$$S_{n-1}' = -\frac{(n-1-\frac{\alpha}{2})}{n-\frac{1}{2}} p_{n-1,j} + \tfrac{1}{2}\frac{q_{n-1,j}}{n-\frac{1}{2}} - q_{n,j+1}$$

and

$$A_i = \frac{\frac{1}{2}}{i+\frac{\alpha}{2}},\ B_i = 1,\ C_i = \frac{-i}{i+\frac{1}{2}},\ D_i = -\frac{(i-\frac{\alpha}{2})}{i+\frac{1}{2}},$$

$$E_i = 1,\ R_i = \frac{i-1-\frac{\alpha}{2}}{i+\frac{\alpha}{2}} p_{i-1,j} - \frac{\frac{1}{2}q_{i-1,j}}{i+\frac{\alpha}{2}}$$

and

$$S_i' = -\frac{(i-\frac{\alpha}{2})}{i+\frac{1}{2}} p_{i,j} + \frac{\frac{1}{2}q_{i,j}}{i+\frac{1}{2}} \quad \text{for all other values of } i\ . \quad (4.18)$$

We can perform row transformations on M' so that it becomes upper triangular. In particular we subtract the $(2i+1)^{th}$ row multiplied by $\frac{C_i}{A_i}$ from the $2(i+1)^{th}$ row to obtain blocks of the form

$$\begin{pmatrix} A & B & 0 \\ 0 & D-\frac{BC}{A} & E \end{pmatrix}.$$

The last block is now $\begin{pmatrix} A_{n-1} & B_{n-1} \\ 0 & D_{n-1} - \dfrac{B_{n-1}C_{n-1}}{A_{n-1}} \end{pmatrix}$

The right hand side is replaced by $\underline{b}$ where

$$\underline{b} = (R_0, S_0, R_1, S_1, \ldots, R_{n-1}, S_{n-1})^T \text{ and}$$

$$S_i = S_i' - \frac{C_i R_i}{A_i}$$

and we can now solve the system straightforwardly by back substitution.

It remains to show how q and p can be calculated along $t = 0$, given that $u(x,0) = u_0(x)$.

$p(x,0)$ is simply $\frac{\partial}{\partial x}u(x,0) = u_0'(x)$ and $q(x,0)$ can be calculated by using the relationship along the first characteristic as follows.

As in (4.15) $q_{0,0} = -(\alpha+1)p_{0,0}$. Then $q_{i+1,0}$ can be calculated from (4.11) using

$$q_{i+1,0} = \frac{1}{i+1}\{iq_{i,0} + (i-\frac{\alpha}{2})p_{i,0} - (i+1+\frac{\alpha}{2})p_{i+1,0}\} \text{ for } i \geq 0.$$

As in the previous chapter q and p can now be used to calculate u on the new time step using either $u_{i,j+1} = u_{i,j} + \frac{1}{2}(q_{i,j}+q_{i,j+1})$

or $u_{i,j+1} = u_{i+1,j+1} + \frac{1}{2}(p_{i+1,j+1}+p_{i,j+1})$,

since u is known at the old time step and at the right hand boundary. Thus, since the solution is known initially it is possible to calculate it at all later times using the above algorithm.

§4.4 Results and Conclusions

In this section results are presented and discussed for the method given in the previous section. All the calculations have been based on the situation $\alpha = 1$ with initial condition $u_0(x) = -xe^{-x}$ and boundary condition $u(1,t) = t^2e^t - (1-t)^2e^{t-1}$. This corresponds to the infinitely differentiable exact solution

$$u(x,t) = \frac{1}{x}[t^2e^t-(x-t)^2e^{t-x}].$$

The exact values of $u(x,t)$, $p(x,t)$ and $q(x,t)$ at $t = .125$ and $t = .5$ are tabulated in Table 4.1. The same results are shown graphically in Figure 4.1.

The remainder of this section is concerned with the accuracy, rate of convergence and stability of the method.

a) Accuracy and Rate of Convergence

In Table 4.2 the error at $t = .125$ and $t = .5$ is given for various step sizes. There are two main points to notice about the error. Firstly, it is second order since halving the step size results in the error being divided by a factor of four. Secondly, the error is roughly constant across each time step except that it is zero at the right hand boundary (since the solution is prescribed there) and it changes sign and increases sharply at the left hand end.

The $O(h^2)$ truncation error is as expected since the position of the characteristics is known exactly and the trapezoidal rule is used to discretize all the relationships along the characteristics.

The sharp increase in the error can be explained as follows. The truncation errors for (4.11) and (4.12) respectively have leading term

$$\frac{h^2}{12}p_{xx} \text{ and } \frac{h^2}{12}\left(p_{xx}-\frac{p_{tt}}{3}\right).$$

These are the equations which are used to calculate values at internal points.

In contrast the leading term of the truncation error for (4.16) and (4.17), which are the equations used to calculate the solution at the boundary, are

$$\frac{2}{3}h^2p_{tt} \text{ and } \frac{h^2}{2}\left(3p_{xx}-\frac{4p_{tt}}{3}\right)$$

respectively.

So, even though the truncation error for all points is $O(h^2)$ and the coefficient contains p_{xx} and p_{tt} in both cases, for boundary points the coefficient is very much greater than at internal points which accounts for the loss of accuracy in this region.

b) Stability

The algorithm is based upon solving the system

$$M\underline{x}_{j+1} = \underline{b} \tag{4.19}$$

where M is of the form

$$\begin{pmatrix} J_0 & & & & & \\ & J_1 & & & & \\ & & J_2 & & & \\ & & & \ddots & & \\ & & & & J_{n-2} & \\ & & & & & J_{n-1} \end{pmatrix}$$

with $J_i = \begin{pmatrix} A_i & B_i & 0 \\ 0 & D_i - \frac{B_i C_i}{A_i} & E_i \end{pmatrix}$ $\quad 0 \leq i \leq n-2$

$$J_{n-1} = \begin{pmatrix} A_{n-1} & B_{n-1} \\ 0 & D_{n-1} - \frac{B_{n-1} C_{n-1}}{A_{n-1}} \end{pmatrix}$$

and A_i is the $(2i,2i)^{th}$ element of M.

$\underline{x}_{j+1}, \underline{b}, A_i, B_i, C_i, D_i, E_i$ are as given by (4.18)

$\underline{b}$ can be rewritten as $N\underline{x}_j + \underline{c}_{j+1}$ where

$$N = \begin{pmatrix}
0 & & & & & & & \\
0 & -1 & & & & & & \\
0 & \frac{1}{2}+\alpha & 0 & & & & & \\
C_1 & G_1 & A_1 & D_1 & & & & \\
 & & -A_1 & F_1 & 0 & & & \\
 & & C_2 & G_2 & A_2 & D_2 & & \\
 & & & & -A_2 & F_2 & 0 & \\
 & & & & & \cdot & \cdot & \cdot \\
 & & & & & & -A_{n-2} & F_{n-2} & 0 \\
 & & & & & & C_{n-1} & G_{n-1} & A_{n-1} & D_{n-1}
\end{pmatrix}$$

a lower triangular matrix with $G_i = -\frac{C_i F_i}{A_i}$

and $F_i = \frac{i-1-\frac{\alpha}{2}}{i+\frac{\alpha}{2}}$

and $\underline{c}_{j+1} = (0,0,\ldots,0,-q_{n,j+1})^T$.

N has eigenvalues $-1, D_1, D_2, \ldots D_{n-1}$ and 0 (repeated n times) and M, an upper triangular matrix, has eigenvalues $A_i, D_i - \frac{B_i C_i}{A_i}$ $i = 0,\ldots,n-1$

$$A_i \neq 0, \text{ and } D_i - \frac{B_i C_i}{A_i} = \left\{ \begin{array}{ll} \alpha & i = 0 \\ \frac{2+5\alpha}{3} & i = 1 \\ \frac{-i+\frac{\alpha}{2}+2i^2+i\alpha}{i+\frac{1}{2}} & i \neq 0,1 \end{array} \right\} \neq 0$$

and so M is non singular.

$$M\underline{x}_{j+1} = N\underline{x}_j + \underline{c}_{j+1}$$

implies

$$\underline{x}_{j+1} = M^{-1}N\underline{x}_j + M^{-1}\underline{c}_{j+1}$$
$$= K\underline{x}_j + \underline{m}_{j+1}$$
where $K = M^{-1}N$ and $\underline{m}_{j+1} = M^{-1}\underline{c}_{j+1}$.

Thus,

$$\underline{x}_j = K^j\underline{x}_0 + K^{j-1}\underline{m}_1 + K^{j-2}\underline{m}_2 + \ldots + K\underline{m}_{j-1} + \underline{m}_j.$$

Since $K = M^{-1}N$ the i^{th} eigenvalue of K is given by

$$\frac{i^{th} \text{ eigenvalue of } N}{i^{th} \text{ eigenvalue of } M}$$
$$= \frac{0}{A_i}, \quad \frac{D_i}{(D_i - \frac{B_iC_i}{A_i})} \qquad i = 0,\ldots,n-1$$
$$= 0 \text{ (n times)}$$

and $\dfrac{A_iD_i}{A_iD_i - B_iC_i}$ $\qquad i = 0,\ldots,n-1$

$\left|\dfrac{A_iD_i}{A_iD_i - B_iC_i}\right| < 1$ if and only if

$$|A_iD_i| < |A_iD_i - B_iC_i| \tag{4.20}$$

$i = 0$: $A_0D_0 = 0$

$$A_0D_0 - B_0C_0 = 0 + \frac{\alpha}{\alpha+1}$$

and so (4.20) is satisfied since $\alpha > 0$.

$i = 1$: $A_1D_1 = \dfrac{1}{2+\alpha} \cdot \dfrac{(\frac{\alpha}{2}-1)}{\frac{3}{2}} = \dfrac{(\alpha-2)}{3(2+\alpha)}$

$$A_1D_1 - B_1C_1 = \frac{(\alpha-2)}{3(2+\alpha)} - \frac{1.(-1)}{\frac{3}{2}}$$

$$= \frac{3\alpha+2}{3(2+\alpha)}$$

and so $|A_1D_1| < |A_1D_1-B_1C_1|$ for all $\alpha > 0$.

$$i \neq 0, 1: \quad A_iD_i = \frac{\frac{1}{2}}{i+\frac{\alpha}{2}} \frac{(\frac{\alpha}{2}-i)}{i+\frac{1}{2}} = \frac{(\alpha-2i)}{(2i+\alpha)(2i+1)}$$

$$A_iD_i-B_iC_i = \frac{(\alpha-2i)}{(2i+\alpha)(2i+1)} - \frac{1.(-i)}{i+\frac{1}{2}}$$

$$= \frac{\alpha-2i+4i^2+2i\alpha}{(2i+\alpha)(2i+1)}$$

and so $|A_iD_i| < |A_iD_i-B_iC_i|$ for all $\alpha > 0$.

Thus all the eigenvalues of K are less than 1 in modulus and so the scheme is stable for all $\alpha > 0$.

To summarise, in this section it has been shown that the scheme presented in section 4.3 is stable for all values of $\alpha > 0$ (with a square grid) and has a second order rate of convergence everywhere. However, at the left hand end, even though the rate of convergence is quadratic there is a sharp drop in accuracy.

§4.5 An extension to the problem

In this section an extension of the above work to a more difficult problem is considered. This extended problem is also connected with the model discussed in Chapter 5 and thus provides a link between the two problems.

The equation to be discussed is

$$x(qu_{xx}-3pu_{xt}) + q(q^4+p) = 0 \qquad (4.21)$$

with initial condition

$$u(x,0) = u_0(x), \text{ a given function for } x \geq 0 \qquad (4.22)$$

and the boundary condition that $u(0,t)$ is an infinitely differentiable function. Exact solutions exist for $u_0(x) = -x$, namely $u(x,t) = \pm t - x$. A numerical algorithm is required to find solutions for small perturbations of this initial condition and to decide if the perturbation dies away or continues to grow.

The characteristics of (4.21) are the lines t = constant and $\frac{dt}{dx} = -\frac{3p}{q}$. Unfortunately this means that we cannot, as in the previous work, choose a fixed grid such that the characteristics always pass through grid points and so Hartree's method is used to enable the solution to be found at grid points.

The characteristic relationship along the sloping characteristic $\frac{dt}{dx} = -\frac{3p}{q}$ is

$$\frac{dp}{dx} + \frac{q^4+p}{x} = 0, \qquad (4.23)$$

and along the lines t = constant it is

$$\frac{dp}{dx} - \frac{3p}{q}\frac{dq}{dx} + \frac{q^4+p}{x} = 0. \qquad (4.24)$$

Following the spirit of the earlier work, second order (implicit) approximations are used to find the solution at a new point, from (4.23) and (4.24).

(4.23) becomes

$$\frac{p_C-p_B}{x_C-x_B} + \left[\frac{q_C^4+p_C}{2x_C}\right] + \left[\frac{q_B^4+p_B}{2x_B}\right] = 0 \qquad (4.25)$$

and (4.24) gives

$$\frac{p_C-p_A}{x_C-x_A} - \frac{3}{2}\left(\frac{p_C}{q_C}+\frac{p_A}{q_A}\right)\left(\frac{q_C-q_A}{x_C-x_A}\right) + \frac{q_C^4+p_C}{2x_C} + \frac{q_A^4+p_A}{2x_A} = 0 \qquad (4.26)$$

where the points A, B and C are as indicated in Figure (4.2).

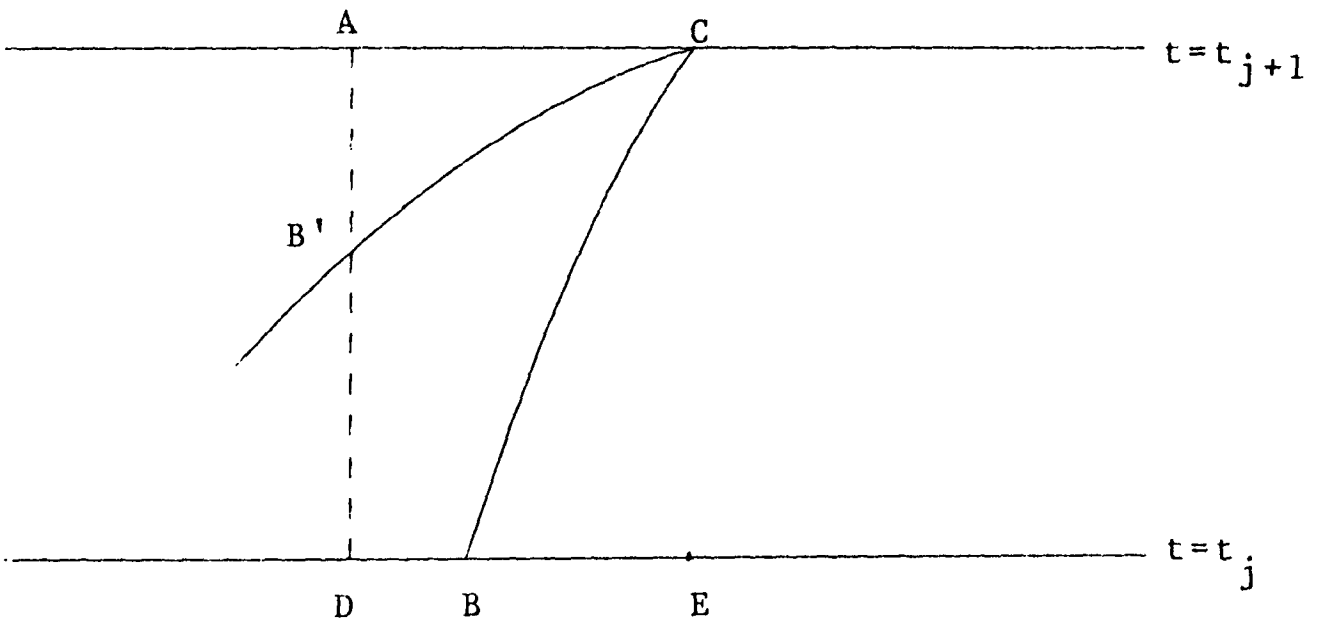


Figure 4.2

As shown in the diagram the sloping characteristics have positive slope if the solution close to $u(x,t) = t - x$ is sought. If this ceases to be the situation the program signals an error. In the above diagram it is assumed that the solution is known at A and along all the previous time level and it is required at C. Unfortunately a second order accurate approximation to the position of the sloping characteristics uses unknown information at C to give

$$\frac{t_C - t_B}{x_C - x_B} = -\frac{3(p_C + p_B)}{q_C + q_B} \quad . \tag{4.27}$$

If (4.27) is used to find the position, x_B, at which the sloping characteristic through C cuts the previous time level then $t_C - t_B = k$ (step size in t direction) and p_B, q_B can be approximated by their linear interpolates to give

$$p_B = \frac{1}{h}[x_C - x_B)p_D + (h - x_C + x_B)p_E]$$

and a similar expression for q_B.

This gives a quadratic for $x_C - x_B$ in terms of known values together with p_C and q_C. In practice, p_C and q_C are guessed so that the position of the sloping characteristic can be found. Then the relationships along the characteristics are used to refine p_C and q_C and the characteristic position is recalculated. This process is repeated to convergence.

If the characteristic cuts the previous time level outside DE then (4.27) is used to find where it cuts AD (the previous x grid point) by putting $x_C - x_B = h$ (step size in x

direction) and solving (4.27) for $t_C - t_B$. Obviously p_B and q_B are now replaced by linear interpolates using the values at A and D. This gives a quadratic for $t_C - t_B$. The situation is shown by the line B'C in Figure 4.2.

In fact the program solves both of these quadratics and signals an error unless there is only one root from either of the two quadratics in the required range.

Once the position of the characteristic from C has been found then (4.25) and (4.26) can be used as an 'inner loop' to refine the values of p_C and q_C.

Subtracting (4.25) from (4.26) gives

$$\frac{p_C-p_A}{x_C-x_A} - \frac{p_C-p_B}{x_C-x_B} - \frac{3}{2}\left(\frac{p_C}{q_C}+\frac{p_A}{q_A}\right)\left(\frac{q_C-q_A}{x_C-x_A}\right) + \left(\frac{q_A^4+p_A}{2x_A}\right) - \left(\frac{q_B^4+p_B}{2x_B}\right) = 0. \quad (4.28)$$

(4.26) gives p_C and $\frac{dp_C}{dq_C}$ in terms of q_C and values at B and so (4.28) can be considered as a nonlinear function of q_C and solved iteratively using Newton's method.

These new values of q_C and p_C are now used to recalculate the position of the characteristics. Provided that the solution is known on an initial time step and at the left hand boundary (x=0), this procedure could be used to find p and q and hence u at all points in the upper right quarter plane.

The left hand boundary, however, poses a problem for two reasons. If $x_A = 0$ then the expression $\frac{q_A^4+p_A}{2x_A}$ which is needed for (4.28) cannot be calculated directly. The other problem is that of using the condition at $x = 0$ to find a solution on this line so that the algorithm described above can be used to extend this solution for positive values of x.

Since $u(x,t)$ is a bounded function (4.21) gives

$$q^4 + p\Big|_{x = 0} = 0 \qquad (4.29)$$

or $$q\Big|_{x = 0} = 0.$$

When the initial conditions are $u_0(x) = -x$ then the exact solution $u(x,t) = \pm t - x$ has $q = \pm 1 \neq 0$ and so since we are considering perturbations of this situation (4.29) is assumed to be true rather than $q(0,t) = 0$.

A second relationship for p and q may be obtained by differentiating (4.21) partially with respect to x and then evaluating at $x = 0$.

This gives

$$(2qu_{xx} - 7pu_{xt})\Big|_{x = 0} = 0 \qquad (4.30)$$

(using (4.29) and the fact that $q(0,t) \neq 0$).

(4.30) can be written in finite difference form to give an expression to calculate p at the left hand end and then (4.29) can be used to find q, up to sign. We are interested in perturbations of the solution $u(x,t) = t - x$ and so $q = (-p)^{\frac{1}{4}}$ (rather than $q = -(-p)^{\frac{1}{4}}$) is used at the left hand boundary.

The remaining problem with the left hand boundary is to find some expression for

$$\lim_{x \to 0} \frac{q^4 + p}{x}$$

This can be written as $\frac{\partial}{\partial x}(q^4+p)\Big|_{x=0}$

$$= 4q^3u_{xt} + u_{xx}$$

$$= -\frac{1}{7}u_{xx}\Big|_{x=0} \qquad \text{from (4.29) and (4.30).}$$

The algorithm can now be used to find the solution at all later time steps, once the solution has been found initially. Since $u(x,0) = u_0(x)$ is a known function, $p(x,0) = \frac{du_0(x)}{dx}$.

Also, the initial line, $t = 0$, is a characteristic and so the relationship (4.24) can be used to find q along $t = 0$. A second order finite difference approximation to (4.24) gives a nonlinear equation for $q((i+1)h,0)$ in terms of $q(ih,0)$ and known functions. This can be solved iteratively for $q((i+h)h,0)$ using Newton's method. The initial value for q, $q(0,0)$ is found from (4.29) since $p(0,0) = u_0'(0)$ is known.

Once p and q have been calculated on a new time level it is easy to find u either from

$$u_{i,j+1} = u_{i,j} + \frac{k}{2}(q_{i,j}+q_{i,j+1})$$

or

$$u_{i,j+1} = u_{i-1,j+1} + \frac{h}{2}(p_{i,j+1}+p_{i-1,j+1}).$$

Using the algorithm described above and the starting function $u_0(x) = -xe^{-x}$, the solution to (4.21) was found for $x\varepsilon[0,1]$ and $t > 0$. The results are shown in Table 4.3 for $t = 5.$ and various values of h and k. It is seen that the results do not

appear to converge. A variety of ratios between h and k were used but in no case did convergence occur. This can be explained by the following perturbation analysis.

Let $u(x,t)$ be a solution to (4.21) and consider $v(x,t) = u(x,t) + \varepsilon(x,t)$ where ε is small so that all quadratic terms in ε may be ignored. (4.21) becomes

$$x(u_t\varepsilon_{xx} - 3u_x\varepsilon_{xt}) + \varepsilon_t(u_t^4 + u_x + xu_{xx} + 4u_t^4) + \varepsilon_x(-3xu_{xt} + u_t) = 0. \quad (4.31)$$

In the case when $u(x,t) = t - x$ this becomes

$$x(\varepsilon_{xx} + 3\varepsilon_{xt}) + 4\varepsilon_t + \varepsilon_x = 0. \quad (4.32)$$

This has a solution $\varepsilon = a(t-4x) + b$ where a and b are arbitrary constants. This suggests that if the initial perturbation is $\varepsilon(x,0) = -4ax + b$ then the solution to (4.21) will not settle back down to $u(x,t) = t - x$ but will be of the form

$$u(x,t) = t - x + a(t-4x) + b.$$

This analytical result is supported by numerical evidence (see Table 4.4). Here $\varepsilon(x,0) = .0001(-4x+1)$ was used and the solution for the first few time steps is almost exactly

$$u(x,t) = t - x + .0001(t-4x+1).$$

Unfortunately, (4.32) not only has solutions such as these which grow linearly with time (which is also the rate at which the unperturbed solution $u(x,t) = t - x$ grows) but

also solutions which grow exponentially.

For, putting $\varepsilon(x,t) = X(x)T(t)$ in (4.32) gives

$$x(X''(x)T(t)+3X'(x)T'(t)) + 4X(x)T'(t) + X'(x)T(t) = 0$$

which can be rearranged to give

$$\frac{T'(t)}{T(t)} = -\frac{(xX''(x)+X'(x))}{3xX'(x)+4X(x)} .$$

Since the left hand side is a function of t only and the right hand side one of x only both sides must be constant and so

$$\frac{T'(t)}{T(t)} = k \Rightarrow T(t) = c_1 e^{kt} \text{ where } c_1 \text{ is an arbitrary constant} \qquad (4.33)$$

and

$$xX'' + X'(1+3kx) + 4kX = 0. \qquad (4.34)$$

It is required to find smooth solutions of (4.34) and so we put

$$X(x) = \sum_{r=0}^{\infty} a_r x^r . \qquad (4.35)$$

Equating coefficients of x in (4.34) gives

$$a_{r+1} = -\frac{k(3r+4)}{(r+1)^2} a_r . \qquad (4.36)$$

$$\lim_{r\to\infty} \frac{a_{r+1}}{a_r} = \lim_{r\to\infty} \frac{1}{r} = 0$$

and so the series converges for all x and k.

(4.3) thus has a smooth solution for any value of k.

This means that $\epsilon(x,t) = c_1 e^{kt} X_k(x)$ (where $X_k(x)$ is the solution to (4.34) given by (4.35) and (4.36)) is a solution to (4.32) for any value of k and, therefore, there exists an infinite number of exponentially growing solutions.

Any algorithm to solve (4.21) will thus eventually fail as rounding errors will introduce the exponentially growing solutions and these will eventually dominate the true solution. Also if the true solution is not a linear function of x and t (such as for a starting function like $u_0(x) = -xe^{-x}$) then the situation becomes worse much more quickly. The finite difference approximations in this case are not exact and any truncation errors they introduce lead to the exponentially growing solutions quickly dominating the true solution. Since the truncation error depends on the step size used different spurious solutions will be introduced for different step sizes and so convergence is not obtained. The problem (4.21) appears to be, therefore, not well posed and no numerical method can be expected to find solutions of it.

Also, since $X(0) = a_0$, $X'(0) = a_1 = -4ka_0$, then $\frac{X'(0)}{X(0)} = -4k$ which is large and negative if k is large and positive. Therefore, the general solution includes components growing exponentially and which have a large sharp spike at $x = 0$.

A similar problem arises in Chapter 5 and it is possible that it exhibits similar behaviour to the problem discussed in this section.

TABLE 4.1

t = .125

x	u	p	q
0.000	.301	-1.425	2.851
0.125	.142	-1.133	2.408
0.250	.016	- .890	2.031
0.375	-.083	- .688	1.711
0.500	-.158	- .522	1.440
0.625	-.214	- .385	1.209
0.750	-.255	- .273	1.015
0.875	-.283	- .182	0.850
1.000	-.302	- .109	0.711

t = .5

x	u	p	q
0.000	2.061	-3.504	7.007
0.125	1.661	-2.918	6.120
0.250	1.328	-2.422	5.355
0.375	1.052	-2.003	4.693
0.500	.824	-1.649	4.122
0.625	.637	-1.351	3.628
0.750	.485	-1.101	3.202
0.875	.361	- .891	2.834
1.000	.261	- .715	2.516

THE EXACT SOLUTION TO 3 DECIMAL PLACES

CORRESPONDING TO THE INITIAL CONDITION $u(x,0)=-xe^{-x}$

TABLE 4.2

t = .125

x	$h=\frac{1}{16}$	$h=\frac{1}{32}$	$h=\frac{1}{64}$
0.000	-168	-42	-12
0.125	-8	-4	-1
0.250	32	6	2
0.375	40	8	2
0.500	39	9	2
0.625	38	8	2
0.750	36	8	2
0.875	34	8	2
1.000	0	0	0

t = .5

x	$h=\frac{1}{16}$	$h=\frac{1}{32}$	$h=\frac{1}{64}$
0.000	-498	-149	44
0.125	-50	-13	3
0.250	58	14	4
0.375	108	27	7
0.500	142	34	9
0.625	160	39	10
0.750	163	40	10
0.875	160	39	10
1.000	0	0	0

THE ERROR $x10^5$ IN THE NUMERICAL SOLUTION

FOR THE INITIAL CONDITION $u(x,0)=-xe^{-x}$

TABLE 4.3

RESULTS FOR $u(x,t)$ WHEN $t=1$ and $u(x,0)=-xe^{-x}$

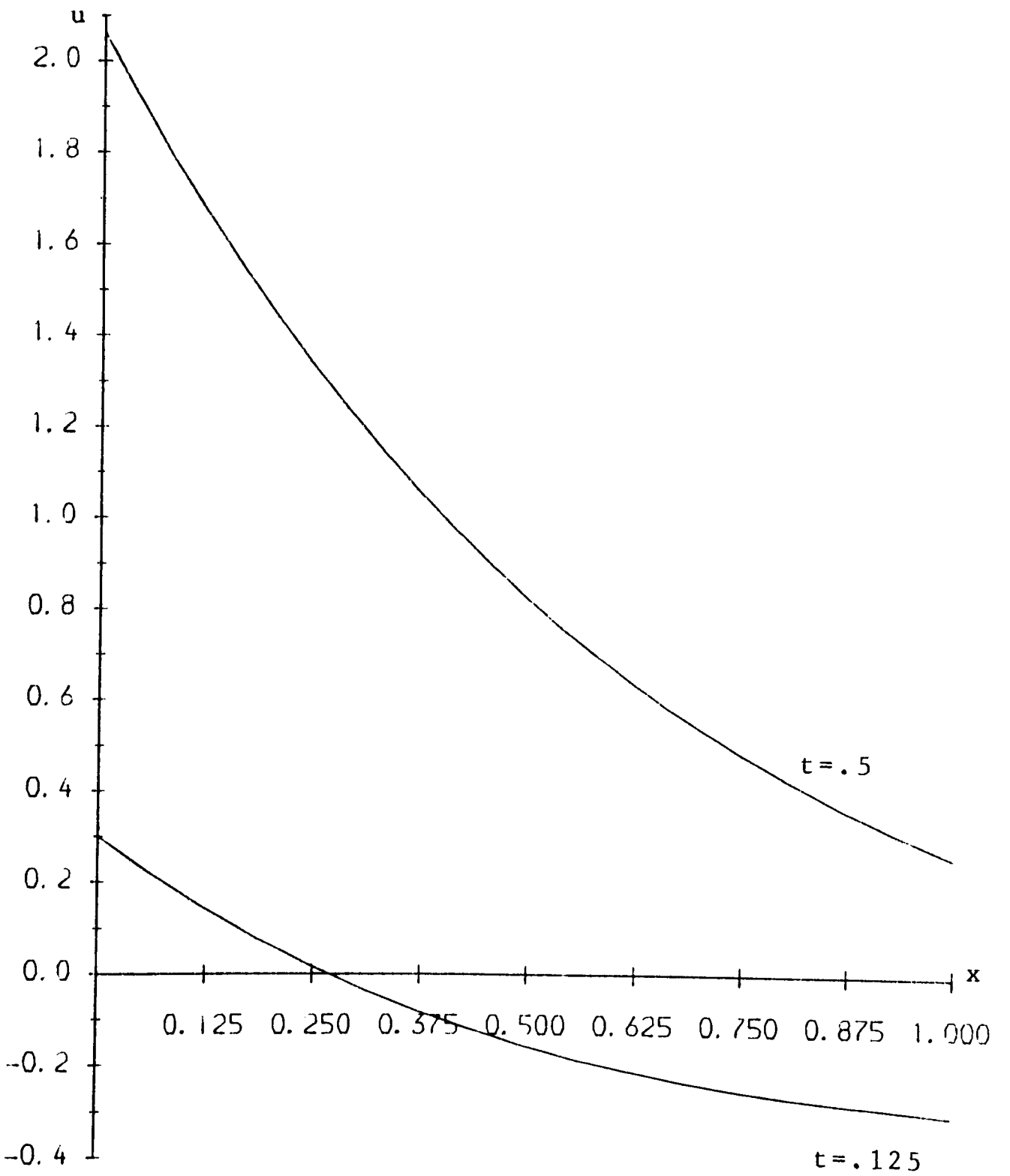
x	Solution	$h=\frac{1}{16}$	$h=\frac{1}{32}$	$h=\frac{1}{64}$
		k=.05	k=.025	k=.0125
0.000		5.777	5.806	5.849
0.125		5.487	5.504	5.528
0.250		5.216	5.226	5.238
0.375		4.964	4.969	4.974
0.500		4.730	4.732	4.733
0.625		4.512	4.513	4.512
0.750		4.311	4.311	4.309
0.875		4.125	4.125	4.123
1.000		3.953	3.953	3.952

TABLE 4.4

x	Analytical solution $u(x,t)=t-x+.0001(t-4x+1)$	Numerical solution $h=\frac{1}{16}$ $k=.05$	$h=\frac{1}{32}$ $k=.025$	$h=\frac{1}{64}$ $k=.0125$
0.000	1.0002	1.0002	1.0002	1.0002
0.125	.8752	.8751	.8751	.8751
0.250	.7501	.7501	.7501	.7501
0.375	.6251	.6250	.6250	.6250
0.500	.5000	.5000	.5000	.5000
0.625	.3750	.3750	.3750	.3750
0.750	.2499	.2499	.2499	.2499
0.875	.1249	.1249	.1249	.1249
1.000	-.0002	-.0002	-.0002	-.0002

SOLUTION FOR $u_0(x)=-x+.0001(-4x+t)$ AT $t=1$

Figure 4.1



The exact solution to (4.1) with initial condition $u_0(x)=-xe^{-x}$ at t=.125 and t=.5

CHAPTER 5

Geodynamic Thermal Runaway

§5.1 Introduction

This chapter is concerned with the investigation and solution of a hyperbolic partial differential equation which describes the movement of a foreign body through a solid medium by means of thermal runaway. The term thermal runaway and some of the physical situations in which it has been considered important are discussed in the next section. Then, in section 5.3 the specific model with which this chapter is concerned is presented.

The remaining sections of the chapter deal with certain special solutions of the equation and describe an algorithm for use in more general cases.

§5.2 Thermal Runaway

Some of the earliest work on thermal runaway was done by Gruntfest [11] in 1963. He considered a slab of incompressible viscous material subject to constant shear stress and discovered that if the viscosity η is of the form

$$\eta = \eta_0 e^{-a(T-T_0)}$$

where T is the temperature and η_0, a and T_0 are constants, then in certain situations the solution to the energy equation

$$\sigma u = c\frac{\partial T}{\partial t} - k\frac{\partial^2 T}{\partial x^2}$$

where $\sigma = \eta u$ is the shear stress (constant)

u is the velocity gradient

c is the volumetric specific heat

and k is the conductivity

has a solution for which the temperature and the strain rate become very high in a finite time. This phenomenon is termed thermal runaway.

Nitsan [23] confirmed this when he too considered a slab of incompressible fluid with a temperature dependent viscosity. He concluded that if constant velocity is imposed on one boundary then a negative feedback occurs for temperature and the situation is stable. However, if a constant stress is applied to one boundary then there is no stabilisation (unless the heat loss through the boundaries is at a sufficiently high rate) and thermal runaway occurs.

Work published in 1979 by Melosh and Ebel [20] and Yuen and Schubert [33] considered whether thermal runaway could result from frictional heating in the upper mantle of the earth. The outermost layer of the earth's surface, termed the lithosphere, consists of rigid plates moving relatively rapidly over the asthenosphere which is a region of high shear stresses since it has the more slowly moving, highly viscous, deep mantle beneath it. There are two possibilities; either an increase in velocity of the lithosphere causes

enough internal dissipative heating to significantly reduce the viscosity of the asthenosphere and hence the shear stresses and thus allow a further increase in velocity or else the thermal effects are minor and cause little change in the viscosity of the asthenosphere in which case an increase in velocity causes an increase in shear stresses, which results in the velocity decreasing once more.

The authors concluded that the second, stable situation, always occurs when the plate velocity is less than some critical velocity (subcritical motion) and also occurs for the supercritical case (plate velocity greater than the critical velocity) when there is any perturbation of constant velocity boundary conditions.

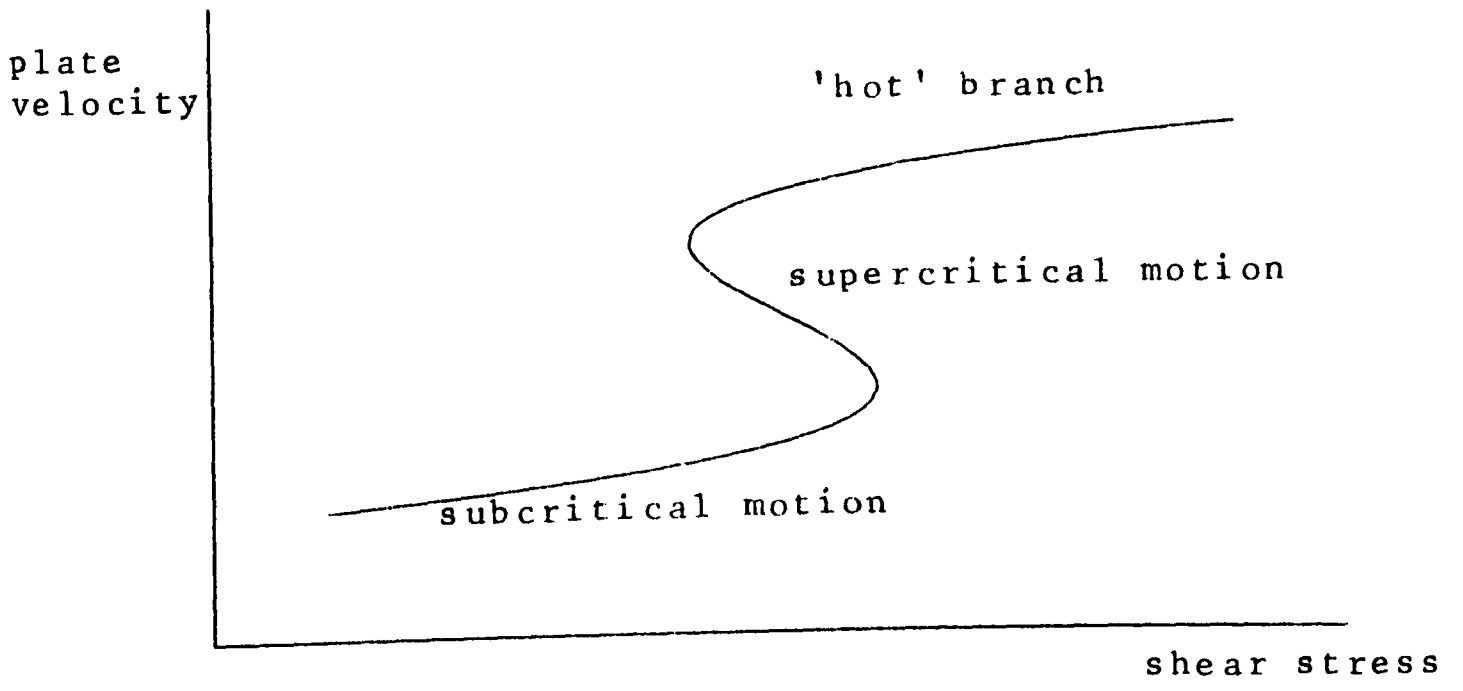


Figure 5.1

Velocity/Shear stress curve

The conditions required for thermal runaway (or the unstable situation) to occur are a perturbation of long wavelength of constant shear stress boundary conditions when the plate velocity is supercritical. Physically this corresponds to occurrences such as plate collisions (as compared with earthquakes which would not cause runaway because they finish too rapidly).

Both sets of authors agree that the velocity/shear stress curve is S shaped as shown in Figure 5.1, but there is disagreement as to the relevance of the 'hot' branch.

Melosh and Ebel believe that the 'hot' branch is not applicable to situations on earth as the temperatures are far too high. In contrast, Yuen and Schubert argue that this is exactly the way in which runaway occurs. A sudden, prolonged increase in shear stress forces a jump from the supercritical branch up to the 'hot' branch and runaway follows.

Both sets of authors, however, agree that runaway systems are linked with a precipitous rise in temperature which grows in a superexponential manner. In neither the case of runaway nor that of the stable situation does an oscillating solution result; either the temperature change dies away or it grows rapidly.

Other work on plate tectonics and continental drift has been carried out by Turcotte et al [32] and Hewitt et al [13].

The model presented in the next section is an attempt to describe a very different type of geological situation. It is concerned with the movement of a solid or immiscible liquid body through a solid medium by means of frictional heating

which is sufficient to melt the medium and maintain a thin liquid layer ahead of the body. This is the situation presented by Ockendon et al [25].

Its direct applicability is to the migration of iron through the mantle of silicates. According to Elsasser [8] the earth accreted homogeneously at a low temperature (of the order of $1000^{o}K$) from lumps of meteoric matter and/or dust. $1000^{o}K$ is well below the melting temperature of the silicates of which most of the earth is made but pure iron melts at $1808^{o}K$ and impure iron at temperatures as low as $1400^{o}K$. Therefore, since the temperature of accretion is 'of the order of' $1000^{o}K$ it is uncertain whether or not the iron would be solid or molten at formation. At the present time almost all of the iron is at the core of the earth and so the model explains how the iron, as a solid or as an immiscible liquid, migrated through the matrix of silicates to the core.

Another possible application is to the study of magma ascent to the earth's surface. This has been studied by Marsh [18] who suggests that the blobs of magma rise through the lithosphere by softening a thin rind of wall rock and causing it to flow past the magma. Morris [22] calculated the time it would take such a sphere to reach the surface assuming there was some way for the sphere to remain at a constant temperature. He acknowledges that magma moving in this way would in fact lose heat rapidly and would have cooled to the surrounding temperature after moving through a distance of only one or two of its own diameters. He suggests that

either successive magma blobs must use the same pathway (it being much easier for later magma to rise through cooled magma than through wall rock) or there must be a trailing stalk which feeds fresh, hot magma to the blob. The model of the next section would suggest that there may be a third alternative; namely that the magma temperature could be maintained by thermal runaway.

§5.3 A model for geodynamic thermal runaway

In this section a model is presented which describes the mechanism by which a solid or an immiscible fluid can melt its way through a host medium. The model was first described by Ockendon et al [25].

The following assumptions are made:

1. The molten host medium is confined to a thin layer around the upstream surface of the body and the wake.
2. The wake, comprising molten host medium just above the melting temperature exerts negligible stress on the magma compared to the lubrication forces upstream.
3. Variable material properties, regelation and surface tension are negligible.

The viscosity of the molten host medium is given by

$$\mu = \mu_0 e^{-\beta T} \tag{5.1}$$

where μ is the viscosity

T is the temperature

μ_0 is the viscosity at $T=0$ (constant)

β is a constant.

This leads to a nonlinearity in the model which is not easy to deal with and so the viscosity of the host medium is considered to be constant in the liquid region but to vary rapidly at the free boundary where the liquid solidifies.

The justification for this is based on the work of McKean [19] and Rinzel and Keller [29] who simplified a similar nonlinearity in a model presented by Hodgkin and Huxley [14] by piecewise constant coefficients and a free boundary. More recently Lacey [16] used the same technique in combustion theory.

Penetration by a solid body

The model consists of two halves. The first is when the penetrating body is assumed to be solid, or more precisely a rigid sphere, which moves buoyantly by melting its way through the host medium. This situation is amenable to analysis and the results are asfollows.

There is a key size parameter γ which is defined as

$$\gamma = \frac{|\rho_1 - \rho_0|}{\rho_0} \frac{ag}{L} \tag{5.2}$$

where ρ_1 is the density of the sphere

ρ_0 is the density of the host medium

a is the radius of the sphere

L is the Latent heat of melting of the host medium

g is the constant of gravitational acceleration.

Steady motion is only possible when $\gamma = \frac{3}{2}$ and in this situation θ (the temperature) can be any constant and the velocity is proportioned to $\theta^{\frac{3}{4}}$. So all possible steady motions

are parametrized by the sphere temperature.

If $\gamma \neq \frac{3}{2}$ then the temperature is monotonically increasing or decreasing to zero in finite time depending on $\mathrm{sgn}(\gamma-\frac{3}{2})$. Steady motion is thus not stable with respect to small changes in γ. Surprisingly, there is apparently no terminal velocity.

The rest of this chapter, including all the numerical work, is concerned with the second half of the model when the foreign body is considered as an immiscible liquid.

Penetration by an immiscible liquid

In this situation, when the penetrating body is assumed to be an immiscible liquid with negligible surface tension, the predictions of the model are significantly different.

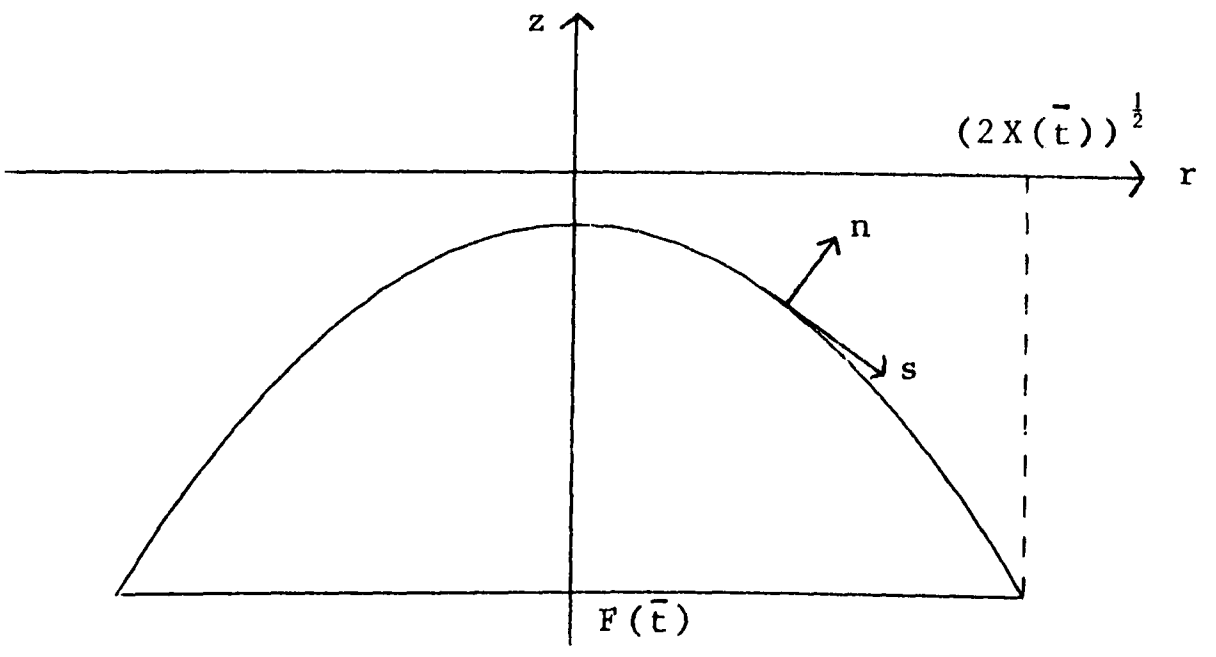


Figure 5.2

An immiscible liquid body

The body is considered to be axisymmetric with shape $z = f(r,\bar{t})$ with respect to axes fixed in space (where z and r are space variables and $\bar{t}$ is time). (See Figure 5.2.) The base of the body is assumed planar at $z = F(\bar{t})$ and the wake is in $z < F(\bar{t})$. Local coordinates (s,n) are defined with n normal to the surface and s tangential in the z, n plane. All variables have been nondimensionalised.

The pressure in the body is assumed to be hydrostatic which allows a local force balance to be imposed on the molten layer. Since the molten layer is thin in comparison to the size of the body, lubrication theory can be used. (See Batchelor [2].) This gives the following viscous flow equations in the molten layer.

$$\frac{\partial p}{\partial s} = \frac{\partial^2 u}{\partial n^2} \tag{5.3}$$

and

$$\frac{\partial p}{\partial n} = 0, \qquad \text{where } p \text{ is the pressure.}$$

The continuity equation is

$$\frac{1}{r}\frac{\partial}{\partial s}(ru) + \frac{\partial v}{\partial n} = 0 \tag{5.4}$$

and that of conservation of energy is

$$\frac{\partial^2 T}{\partial n^2} + \left(\frac{\partial u}{\partial n}\right)^2 = 0 \tag{5.5}$$

where u is the tangential velocity and v the normal velocity. T is, as previously, the temperature in the molten layer.

The boundary conditions are

$$n = 0,\ v = 0 = \frac{\partial u}{\partial n},\ T = \theta(\bar{t}),\ p = \gamma f \tag{5.6}$$

where γ is the size parameter given by (5.2) with a^3 now equal to the volume of the body, and $\theta(\bar{t})$ is the temperature of the penetrating body.

If $\bar{\delta}$ is the thickness of the molten rock, then,

$$n = \bar{\delta},\ u = 0,\ v = \frac{-\partial f}{\partial \bar{t}} \frac{\partial r}{\partial s},\ T = 0,\ \frac{\partial T}{\partial n} = -\frac{\partial f}{\partial \bar{t}} \frac{\partial r}{\partial s} \tag{5.7}$$

so that (5.3) gives

$$u = \frac{\partial p}{\partial s}\left(\frac{n^2}{2} + Bn + C\right)$$

where B and C are arbitrary constants.

From (5.6) and (5.7), $B = 0$ and $C = -\frac{\bar{\delta}^2}{2}$ so that

$$u = \frac{\partial p}{\partial s}\left(\frac{n^2 - \bar{\delta}^2}{2}\right) . \tag{5.8}$$

Then integrating (5.4) over the molten layer gives

$$\int_0^{\bar{\delta}} \frac{1}{r} \frac{\partial}{\partial s}\left(r \frac{\partial p}{\partial s}\left(\frac{n^2 - \bar{\delta}^2}{2}\right)\right) dn + \int_0^{\bar{\delta}} \frac{\partial v}{\partial n}\, dn = 0 .$$

That is,

$$-\frac{1}{r} \frac{\partial}{\partial s}\left(r \frac{\partial p}{\partial s} \frac{\bar{\delta}^3}{3}\right) - \frac{\partial f}{\partial \bar{t}} \frac{\partial r}{\partial s} = 0. \tag{5.9}$$

Since $p = \gamma f$ on $n = 0$ (5.6) and $\frac{\partial p}{\partial n} = 0$ (5.3), then

$$\left.\frac{\partial p}{\partial s}\right|_{n=\bar{\delta}} = \gamma \left.\frac{\partial f}{\partial s}\right|_{n=0}$$

Hence (5.9) becomes

$$\frac{\gamma}{3}\frac{\partial}{\partial r}\left(r\bar{\delta}^3\frac{\partial f}{\partial s}\right) + r\frac{\partial f}{\partial \bar{t}} = 0. \qquad (5.10)$$

Similarly, substituting for u from (5.8) into (5.5) and integrating gives

$$T = -\gamma^2\left(\frac{\partial f}{\partial s}\right)^2 \frac{n^4}{12} + Cn + D$$

where C and D are arbitrary constants.

The boundary conditions $T = \theta(\bar{t})$ when $n = 0$ (5.6) and $T = 0$ when $n = \bar{\delta}$ (5.7) give

$$D = \theta \text{ and } C = \gamma^2\left(\frac{\partial f}{\partial s}\right)^2 \frac{\bar{\delta}^3}{12} - \frac{\theta}{\bar{\delta}}.$$

The equation for T is, therefore,

$$T = -\frac{\gamma^2}{12}\left(\frac{\partial f}{\partial s}\right)^2 (n^4 - \bar{\delta}^3 n) + \theta\left(1-\frac{n}{\bar{\delta}}\right). \qquad (5.11)$$

Applying the condition $\frac{\partial T}{\partial n} = -\frac{\partial f}{\partial \bar{t}}\frac{\partial r}{\partial s}$ when $n = \bar{\delta}$ gives

$$\frac{\partial f}{\partial \bar{t}}\frac{\partial r}{\partial s} = \frac{\gamma^2}{4}\bar{\delta}^3\left(\frac{\partial f}{\partial s}\right)^2 + \frac{\theta}{\bar{\delta}}. \qquad (5.12)$$

The two equations to be solved are (5.10) and (5.12). They can be simplified by the following change of variables.

$$\bar{\delta} = \left(\frac{3\theta}{2\gamma}\right)^{\frac{1}{4}}\delta, \quad \frac{\partial}{\partial \bar{t}} = \theta^{\frac{3}{4}}\left(\frac{2\gamma}{3}\right)^{\frac{1}{4}}\frac{\partial}{\partial t}, \quad x = \frac{r^2}{2}$$

which gives

$$\frac{\partial f}{\partial t} = -\frac{\partial}{\partial x}\left\{\frac{x\delta^3\frac{\partial f}{\partial x}}{\left(1+2x\left(\frac{\partial f}{\partial x}\right)^2\right)^{\frac{1}{2}}}\right\} \qquad (5.13)$$

$$\frac{\partial f}{\partial t} = \frac{3\gamma}{4}\frac{x\delta^3\left(\frac{\partial f}{\partial x}\right)^2}{\left(1+2x\left(\frac{\partial f}{\partial x}\right)^2\right)^{\frac{1}{2}}} + \frac{\left(1+2x\left(\frac{\partial f}{\partial x}\right)^2\right)^{\frac{1}{2}}}{\delta}. \qquad (5.14)$$

f of course is now considered as a function of x and t.

In addition to these two equations there are three global conservation laws to be applied.

The conservation of volume gives

$$-\int_0^{X(t)} x\frac{\partial f}{\partial x}dx = \frac{1}{2\pi} \tag{5.15}$$

where $2\pi X(t)$ is the area of the base of the body $z = F(t)$.

Conservation of momentum gives

$$\theta^{\frac{3}{4}}\frac{\partial}{\partial t}\int_0^{X(t)} xf\frac{\partial f}{\partial x}\,dx = -\frac{U}{2\pi} \tag{5.16}$$

where the non-dimensional constant momentum in the z direction is $(\frac{2\gamma}{3})^{\frac{1}{4}}U$.

Finally, conservation of heat gives

$$\frac{\alpha}{2\pi}\frac{d\theta}{dt} = \int_0^{X(t)}\left\{\frac{\gamma x\delta^3(\frac{\partial f}{\partial x})^2}{(1+2x(\frac{\partial f}{\partial x})^2)^{\frac{1}{2}}} - \frac{\partial f}{\partial t}\right\}dx \tag{5.17}$$

where α is heat capacity of the penetrating body and is taken to be 2π for simplicity. Naturally it would be possible to adapt any analysis or numerical calculations to include a different value for α.

The last condition to be imposed on the solution is a condition at the nose, namely $\delta^3 f_x$ must be bounded at $x = 0$. This helps to eliminate non-physical solutions.

The above system of equations, (5-13)-(5.17) are to be solved for $f(x,t)$, $\delta(x,t)$, $\theta(t)$ and $X(t)$. Thermal runaway is considered to occur if there exists a solution with

$\delta(x,t) > 0$.

As explained below, the system is a second order hyperbolic equation for $f(x,t)$ with a free boundary, $X(t)$, and a singularity at $x = 0$ (which corresponds to the nose of the body).

§5.4 The Steady Solution

It is possible to find an exact solution to (5.13)-(5.17) of the form

$$f(x,t) = Ut - g(x), \quad \theta = 1.$$

(5.13) can now be integrated and using the fact that $\delta^3\frac{\partial f}{\partial x}$ is bounded at $x = 0$ this gives an expression for δ^3. Substitution for δ in (5.14) then gives

$$U^4(1-\frac{3\gamma}{4}xh)^3 = h(1+2xh^2) \tag{5.18}$$

where $h(x) = -\frac{\partial f}{\partial x} = \frac{dg}{dx}$.

If $X(t)$ is a constant, (5.15) and (5.16) reduce to the same relationship and if this constant is $\frac{\gamma}{2\pi}$ all three boundary conditions give

$$\int_0^{\gamma/2\pi} xh\,dx = \frac{1}{2\pi} . \tag{5.19}$$

Equations (5.18) and (5.19) give two relationships between γ the size parameter, U the non-dimensional momentum, and f the shape of the body.

From (5.18) it can be shown that $\frac{d}{dx}(xh) > 0$ for $x, h > 0$ and that $xh \to \frac{4}{3\gamma}$ as $x \to +\infty$. Hence xh is a monotonically increasing function of x.

(5.18) can be considered as giving xh as a function of x and U with γ fixed. Then as $U \to \infty$ with x, h > 0, $xh \to \frac{4}{3\gamma}$ and as $U \to 0$, $xh \to 0$. Thus, the left hand side of (5.19) tends to $\frac{2}{3\pi}$ as $U \to \infty$ and zero as $U \to 0$. So, by continuity, there exists at least one value of U satisfying (5.19) for any γ.

Figure (5.3) shows the relationship between U and γ and and Figure (5.4) gives a graph of $\log_{10}h$ plotted against x for a fixed U and γ. Also plotted on this graph is the function $\log_{10}h$ obtained if the equations (5.13)-(5.17) are slightly simplified.

The simplification is to 'remove' the 1 from the term $(1+2x(\frac{\partial f}{\partial x})^2)^{\frac{1}{2}}$ to give $-(2x)^{\frac{1}{2}}\frac{\partial f}{\partial x}$ instead. As seen below, this has the effect of making $\frac{\partial f}{\partial x}$ singular at the origin rather than nearly singular as previously. It proves easier to deal with the singularity rather than the near singularity of the original equations.

This gives the following system of equations:

$$\frac{\partial f}{\partial t} = \frac{1}{\sqrt{2}} \frac{\partial}{\partial x} (\sqrt{x}\delta^3) \tag{5.20}$$

$$\frac{\partial f}{\partial t} = - \frac{\sqrt{x}}{\sqrt{2}} \frac{\partial f}{\partial x} (\lambda\delta^3 + \frac{2}{\delta}) \tag{5.21}$$

where $\lambda = \frac{3\gamma}{4}$,

together with the boundary conditions

$$- \int_0^{X(t)} x\frac{\partial f}{\partial x}dx = \frac{1}{2\pi} \tag{5.22}$$

$$\theta^{\frac{3}{4}}\frac{\partial}{\partial t}\int_0^{X(t)} xf\frac{\partial f}{\partial x}dx = -\frac{U}{2\pi} \tag{5.23}$$

and

$$\frac{d\theta}{dt} = -\int_0^{X(t)}\left\{\gamma\frac{\sqrt{x}}{\sqrt{2}}\ \delta^3\ \frac{\partial f}{\partial x}+\frac{\partial f}{\partial t}\right\}dx. \tag{5.24}$$

(In (5.24) the substitution $\alpha = 2\pi$ has been made.)

This system of equations has a 'steady' solution (obtained by integrating (5.20) and then substituting for δ (5.21)) as follows

$$f(x,t) = -\frac{3}{2\lambda}\log(2^{\frac{1}{3}}+\lambda x^{\frac{2}{3}}U^{\frac{4}{3}}) + Ut, \tag{5.25}$$

$$p = \frac{\partial f}{\partial x} = \frac{-U^{\frac{4}{3}}}{x^{\frac{1}{3}}(2^{\frac{1}{3}}+\lambda x^{\frac{2}{3}}U^{\frac{4}{3}})}\ ,$$

$$q = \frac{\partial f}{\partial t} = U, \tag{5.27}$$

$$\delta = (\sqrt{2}U)^{\frac{1}{3}}x^{\frac{1}{6}}\ , \tag{5.28}$$

$$X(t) = \frac{\gamma}{2\pi}\ , \tag{5.29}$$

$$\theta = 1. \tag{5.30}$$

This satisfies the condition $\delta^3 f_x$ bounded at $x = 0$, and the boundary conditions (5.22)-(5.24) reduce to the following relationship between γ and U.

$$\int_0^{\gamma/2\pi}\frac{U^{\frac{4}{3}}x^{\frac{2}{3}}}{2^{\frac{1}{3}}+\lambda x^{\frac{2}{3}}U^{\frac{4}{3}}}\,dx = \frac{1}{2\pi}\ . \tag{5.31}$$

This can be integrated exactly to give

$$\frac{1}{2\pi} = \frac{2}{3\pi} - \frac{16}{3\pi^{\frac{1}{3}}} \cdot \frac{1}{\gamma^{\frac{5}{3}}U^{\frac{4}{3}}}$$

$$+ \frac{2^{\frac{3}{2}}16}{3^{\frac{3}{2}}U^{2}\gamma^{\frac{5}{2}}} \tan^{-1}\left(\frac{3\gamma^{\frac{5}{3}}U^{\frac{4}{3}}}{8\pi^{\frac{2}{3}}}\right)^{\frac{1}{2}} = 0.$$

This is a function of γ^5U^4 and so gives the following equation for one variable, N, where $N^3 = \gamma^5U^4$.

$$\tan\left\{\left(-\frac{1}{6\pi} + \frac{16}{3\pi^{\frac{1}{3}}N}\right) \frac{3^{\frac{3}{2}}N^{\frac{3}{2}}}{2^{\frac{3}{2}}16}\right\} = \left(\frac{3N}{8\pi^{\frac{2}{3}}}\right)^{\frac{1}{2}} .$$

The above equation can be solved by Newton's method to give $N \simeq 36.0598$ and thus 'matching' values of γ and U can be found such that (5.31) is satisfied.

Table 5.1 gives corresponding values of U and γ for this system of equations and also the values obtained by solving (5.18) and (5.19), (the equations with the '1' included). It is seen that the two sets of values are very close.

Figure 5.4 shows the graph of $\frac{\partial f}{\partial x}$ for both the set of equations with the '1' and those without. Again, the two curves are very similar, the main difference being that in the former case $\frac{\partial f}{\partial x}$ remains bounded at $x = 0$ whereas in the latter it tends to infinity.

In view of the explicit expressions for f and δ obtained when equations (5.20)-(5.24) are used instead of (5.13)-(5.17) and the fact that the steady solution does not differ greatly in general shape between the two sets of equations all further work is carried out on (5.20)-(5.24).

The steady solution for this, the immiscible liquid case is fundamentally different from penetration by a solid sphere. In both cases the motion is determined by the velocity, U, but in the case of the solid sphere the size parameter γ is always $\frac{3}{2}$ and choosing a velocity fixes the temperature, whereas for the immiscible liquid the temperature is always 1 and choosing a velocity fixes the size parameter γ.

An investigation must now be made as to whether the steady motion in the immiscible liquid case is stable with respect to small changes in the velocity with a fixed γ. Section 5.5 gives a brief discussion of the scaling of the equations which gives some indication of the likely behaviour if the motion is stable. Then section 5.6 links the thermal runaway problem to the problem discussed in the previous chapter. Section 5.7 gives a linear perturbation analysis and the final sections present a numerical algorithm for the solution of equations (5.20)-(5.24) and the results which are obtained.

§5.5 The scaling of the equations

Some indication of the likely behaviour of the solution to the equations (5.20)-(5.24) when the steady solution is slightly perturbed can be obtained by considering the scaling of the problem.

Putting $\bar{x} = \gamma^{-r}x$, $\bar{h} = \gamma^{(5+r)/3}N^{-1}h$, $\bar{f} = \gamma^{(5-2r)/3}N^{-1}f$ gives from (5.20) and (5.21), $\bar{\delta} = [N^{-1}\gamma^{(5-2r)/3}]^{\frac{1}{4}}\delta$, $\bar{t} = [N^{-1}\gamma^{(5-8r)/3}]^{\frac{1}{4}}t$ and $\bar{x}$, $\bar{h}$, $\bar{f}$, $\bar{\delta}$ and $\bar{t}$ are all $O(1)$.

Close to the nose, if $x \sim \gamma^{10}$ (assuming $\gamma < 1$), then

$$\bar{t} = N^{-\frac{1}{4}}\gamma^{-\frac{25}{4}}t$$

whereas further out, when $x \sim \gamma$, then

$$\bar{t} = N^{-\frac{1}{4}}\gamma^{-\frac{1}{4}}t.$$

Since $\gamma^{-\frac{25}{4}} >> \gamma^{-\frac{1}{4}}$ it takes very much longer for any change to occur at the nose.

It seems plausible that if the initial conditions are some perturbation of a steady solution, then the solution further out from the nose will change much more rapidly than that closer in. Thus if the solution to the equations is going to settle back into a steady state it is likely that the solution at the nose will not change rapidly and that the solution further out will adapt to fit in with the nose conditions.

In the next section the scaled equations near to the nose are considered in more detail.

§5.6 Links between the thermal runaway problem and Chapter 4

If the scaling of section 5.5 is used then equation (5.21) becomes

$$\frac{\partial \bar{f}}{\partial \bar{t}} = \frac{1}{\sqrt{2}} \frac{\partial}{\partial \bar{x}} (\sqrt{\bar{x}}\bar{\delta}^3) \qquad (5.32)$$

and (5.20) gives

$$\frac{\partial \bar{f}}{\partial \bar{t}} = \bar{h}\sqrt{\frac{\bar{x}}{2}} \{\varepsilon\delta^3 + \frac{2}{\bar{\delta}}\} , \qquad (5.33)$$

where $\varepsilon = \frac{3}{4}N\gamma^{\frac{2}{3}}(r-1)$.

Close to the nose, when r = 10, $\varepsilon << 1$ and so the term $\varepsilon\bar{\delta}^3$ can be ignored to give

$$\frac{\partial\bar{f}}{\partial\bar{t}} = \bar{h}\sqrt{\frac{\bar{x}}{2}} \cdot \frac{2}{\bar{\delta}} .$$

Substituting for $\bar{\delta}$ in (5.32) gives

$$\bar{f}_{\bar{t}} + 2\frac{\partial}{\partial\bar{x}}\left(\frac{\bar{x}^2\bar{f}_{\bar{x}}^3}{\bar{f}_{\bar{t}}^3}\right) = 0 ,$$

which can be rearranged as

$$6\bar{x}^2\bar{f}_{\bar{x}}^2[\bar{f}_{\bar{t}}\bar{f}_{\bar{x}\bar{x}} - \bar{f}_{\bar{x}}\bar{f}_{\bar{t}\bar{x}}] + \bar{f}_{\bar{t}}^5 + 4\bar{x}\bar{f}_{\bar{x}}^3\bar{f}_{\bar{t}} = 0, \qquad (5.34)$$

and the condition that $\bar{\delta}^3\bar{f}_{\bar{x}}$ should be bounded gives

$$\frac{\bar{f}_{\bar{x}}^4\bar{x}\sqrt{\bar{x}}}{\bar{f}_{\bar{t}}^3} \quad \text{bounded.} \qquad (5.35)$$

(5.34) is an equation which looks very similar in form to (4.21), the later problem of Chapter 4. (5.34) is, of course, a more complicated equation and it also has a more complicated condition at the left hand boundary. (As compared with (4.21) when the condition was that the solution should be infinitely differentiable at x=0.) It is, therefore, not possible to undertake an analysis of the equation in the same way as in Chapter 4. However, the results of Chapter 4 lead to the suspicion that the problem may be ill-posed and for this reason a linear perturbation analysis of the full equation (that is, not just the simpler solution near to the nose), is carried out in the next section.

§5.7 Perturbations of the problem

The steady solution with velocity U and size parameter γ given by (5.25)-(5.30) can be slightly perturbed as follows.

Let

$$f(x,t) = \bar{f}(x,t) + v(x,t) \tag{5.36}$$

$$\delta(x,t) = \bar{\delta}(x,t) + w(x,t) \tag{5.37}$$

$$X(t) = \frac{\gamma}{2\pi} + \xi(t) \tag{5.38}$$

$$\theta(t) = 1 + \varepsilon(t) \tag{5.39}$$

where $\bar{f}$, $\bar{\delta}$ is the steady solution.

If $v(x,t)$, $w(x,t)$, $\xi(t)$, $\varepsilon(t)$ are considered to be small perturbations so that only linear terms involving them are considered to be significant, then (5.20) gives

$$v_t = \frac{5U^{\frac{2}{3}}}{2^{\frac{7}{6}}} \cdot \frac{1}{x^{\frac{1}{6}}} w + \frac{3U^{\frac{2}{3}}}{2^{\frac{1}{6}}} x^{\frac{5}{6}} w_x \tag{5.40}$$

and (5.21) gives

$$v_t = -\frac{x^{\frac{1}{3}}}{U^{\frac{1}{3}}} \{2^{\frac{1}{3}} + \lambda U^{\frac{4}{3}} x^{\frac{2}{3}}\} v_x + \frac{U^{\frac{2}{3}}}{2^{\frac{1}{6}} x^{\frac{1}{6}}} \left\{\frac{-2^{\frac{1}{3}} + 3\lambda U^{\frac{4}{3}} x^{\frac{2}{3}}}{2^{\frac{1}{3}} + \lambda U^{\frac{4}{3}} x^{\frac{2}{3}}}\right\} w \ . \tag{5.41}$$

(5.40) and (5.41) can be rewritten as,

$$v_t = (Bw)_x \tag{5.42}$$

$$w = Cv_t + Dv_x \tag{5.43}$$

where

$$B = \frac{3U^{\frac{2}{3}}}{2^{\frac{1}{6}}}\, x^{\frac{5}{6}}$$

$$C = \frac{2^{\frac{1}{6}}x^{\frac{1}{6}}}{U^{\frac{2}{3}}} \left\{ \frac{2^{\frac{1}{3}}+\lambda U^{\frac{4}{3}}x^{\frac{2}{3}}}{-2^{\frac{1}{3}}+3\lambda U^{\frac{4}{3}}x^{\frac{2}{3}}} \right\}$$

$$D = \frac{2^{\frac{1}{6}}x^{\frac{1}{2}}}{U} \frac{(2^{\frac{1}{3}}+\lambda U^{\frac{4}{3}}x^{\frac{2}{3}})^{2}}{(-2^{\frac{1}{3}}+3\lambda U^{\frac{4}{3}}x^{\frac{2}{3}})} \quad ,$$

and w can be eliminated between (5.42) and (5.43) to give the quasilinear hyperbolic partial differential equation

$$v_t = (BCv_t + BDv_x)_x, \tag{5.44}$$

whose characteristics are the lines t = constant and

$$t = -\frac{3}{2\lambda U}\log(2^{\frac{1}{3}}+\lambda U^{\frac{4}{3}}x^{\frac{2}{3}}) + \text{constant}.$$

The boundary conditions (5.22)-(5.24) result, unfortunately, in slightly more awkward expressions.

(5.22) gives,

$$\xi(t) = -\frac{2\pi}{\gamma \bar{f}_x\Big|_{\gamma/2\pi}} \int_0^{\gamma/2\pi} x v_x dx. \tag{5.45}$$

(5.23) gives

$$\int_0^{\gamma/2\pi} x\{\bar{f}_x v_t + U v_x + \bar{f} v_{xt}\}dx + \xi \frac{\gamma}{2\pi} U \bar{f}_x\Big|_{\gamma/2\pi} + \frac{d\xi}{dt}\frac{\gamma}{2\pi}.\bar{f}\Big|_{\gamma/2\pi} \bar{f}_x\Big|_{\gamma/2\pi} = \frac{3U}{8\pi}\varepsilon(t). \tag{5.46}$$

Since $\xi(t)$ is known in terms of a linear functional of v_x from (5.45) then $\varepsilon(t)$ may be expressed as a linear functional of derivatives of v.

Finally, from (5.24) there is another relationship between $\xi(t)$ and $\varepsilon(t)$.

$$\frac{d\varepsilon}{dt} = -\int_0^{\gamma/2\pi}\{\frac{3\gamma}{\sqrt{2}}\bar{\delta}^2 x^{\frac{1}{2}}\bar{f}_x w+\frac{\gamma}{\sqrt{2}}x^{\frac{1}{2}}\bar{\delta}^3 v_x - v_t\}dx$$

$$-\xi\{\frac{\gamma}{\sqrt{2}}(\frac{\gamma}{2\pi})^{\frac{1}{2}}\bar{\delta}^3\Big|_{\gamma/2\pi}\bar{f}_x\Big|_{\gamma/2\pi} - U\}. \qquad (5.47)$$

So, formally, (5.45)-(5.47) could give expressions for $\xi(t)$, $\varepsilon(t)$ and a condition of the form

$$\int_0^{\gamma/2\pi} G(v)dx = 0 \qquad (5.48)$$

where G(v) is a linear combination of derivatives of v(x,t).

The condition $\delta^3 f_x$ bounded at x=0 gives the corresponding condition for the perturbed problem

$$\bar{\delta}^3 v_x + 3\bar{\delta}^2 f_x w \text{ bounded at } x = 0$$

which reduces to

$$4.2^{\frac{1}{3}}x^{\frac{1}{2}}v_x + 3U^{\frac{1}{3}}x^{\frac{1}{6}}v_t \text{ bounded at } x = 0. \qquad (5.49)$$

Equations (5.44), (5.48) and (5.49) can be considered as a partial differential equation for v(x,t). If we try to find a solution of the form $Bw(x,t) = Y(x)T'(t)$ then from (5.42) we obtain

$$v_t = Y'(x)T'(t),$$

and thus

$$v_x = Y''(x)T(t)$$

and so (5.43) gives

$$Y(x)T'(t) = BCY'(x)T'(t) + BDY''(x)T(t)$$

and thus

$$\frac{T'(t)}{T(t)} = \frac{BDY''(x)}{Y-BCY'(x)} .$$

Since B, C, D are functions of x only, this gives

$$\frac{T'(t)}{T(t)} = k \text{ where } k \text{ is same constant}$$

and so $T(t) = e^{kt}$,

and the partial differential equation becomes an ordinary differential equation in Y(x).

$$BDY''(x) + BCkY'(x) - kY = 0,$$

or

$$Y''(x) + kr(x)Y'(x) - ks(x)Y(x) = 0, \qquad (5.50)$$

where $r(x) = \frac{C}{D}$ is always non-negative

and $s(x) = \frac{1}{BD}$.

The solution $Y(x) \equiv 0$ would satisfy (5.50) and the boundary conditions (5.48) and (5.49). Therefore, any non-zero solution, $Y_k(x)$, must be an eigensolution with eigenvalue k.

$v(x,t) = Y'_k(x)e^{kt}$ is a solution of the partial differential equation (5.40) satisfying the boundary conditions. Since the equation is linear in v, solutions can be added to give a new solution and so the most general solution of this form is

$$v(x,t) = \sum_{k\in\Lambda} \alpha_k Y'_k(x)e^{kt}$$

where Λ is the set of eigenvalues and the α_k are arbitrary constants.

If the initial value of the perturbation is known, that is v(x,0) is a known function of x and $\{Y'_k(x)\}_{k\in\Lambda}$ spans a subset of the real functions containing v(x,0), then the constants α_k can be chosen such that

$$v(x,0) = \sum_{k\in\Lambda} \alpha_k Y'_k(x)$$

and thus

$$v(x,t) = \sum_{k\in\Lambda} \alpha_k Y'_k(x)e^{kt} \quad \text{(with the } \alpha_k \text{ now fixed)}$$

is a solution of the partial differential equation (5.40) satisfying the initial and boundary conditions.

The perturbation v(x,t) dies away to zero if and only if

$$\alpha_k \neq 0 \text{ implies } k < 0.$$

If this solution holds then the steady solution $\bar{f}(x,t)$ is stable for small perturbations.

If the condition $\alpha_k \neq 0$ implies $k \leq 0$, and $\alpha_k \neq 0$ when $k = 0$, then v(x,t) approaches a function of x only as t tends

to infinity. In this situation the solution $f(x,t) = \bar{f}(x,t) + v(x,t)$ settles down to a new steady solution, not back to $\bar{f}(x,t)$.

Unfortunately it has not been possible to find the eigenvalues of (5.46), together with the given boundary conditions. This is partly due to the difficulties imposed by the form of (5.36) which has coefficients which are complicated functions of x and has the eigenvalue k multiplying the $Y'(x)$ term unlike the standard Sturm-Liouville form.

Also the complicated nature of the boundary condition (5.44) precludes any special analysis using the exact nature of the boundary condition.

It seems unlikely that all the eigenvalues will be negative and if there are any positive eigenvalues the initial conditions are crucial to determine which eigenvectors are present. In the light of this it would seem probable that $v(x,t)$ does not tend to zero or a function of x only as t tends to infinity except possibly for very special initial conditions. That is, the steady solution is unstable except maybe for a set of particular initial perturbations.

The perturbation equation (5.40) provides another link between the thermal runaway problem and the equations of the previous chapter. For, replacing BC and BD by βx, (5.44) becomes

$$\beta x(v_{xx}+v_{xt}) + \beta v_x + (\beta-1)v_t = 0, \qquad (5.51)$$

which is very similar to both equation (4.1), considered in the early part of Chapter 4, and the linear perturbation

equation (4.32) which arises as a result of the perturbation analysis in the latter half of Chapter 4.

Unfortunately the boundary conditions of this problem are far more complicated than for the equations of Chapter 4 and so it is not possible to carry across the approach used in that chapter to investigate the equation further analytically or to solve it numerically. However, a numerical solution to equations (5.20)-(5.24) (rather than to the perturbation equation (5.51)) has been attempted and the algorithm and results are given in the final two sections of this chapter.

§5.8 A numerical algorithm

The equations to be considered are (5.20) to (5.24), together with the fact that $\delta^3 f_x$ is bounded on $x = 0$.

(5.21) can be differentiated with respect to x to obtain

$$f_{xt} = -f_{xx}\sqrt{\frac{x}{2}}(\lambda\delta^3+\frac{2}{\delta}) - \frac{f_x}{2\sqrt{2x}}(\lambda\delta^3+\frac{2}{\delta}) - f_x\sqrt{\frac{x}{2}}(3\lambda\delta^2-\frac{2}{\delta^2})\delta_x. \qquad (5.52)$$

(5.20) gives an equation for δ_x in terms of δ, x, f_x and f_t and (5.21) gives δ implicitly in terms of x, f_x and f_t and so formally (5.52) could be rewritten as

$$\alpha f_{xx} + f_{xt} + e = 0, \qquad (5.53)$$

where α and e are functions of x, f_x and f_t.

This is a hyperbolic, quasilinear equation and the characteristics are given by the lines t = constant and $\frac{dt}{dx} = \frac{1}{\alpha}$.

Along t = const the relationship

$$\alpha dp + dq + edx = 0 \qquad (5.54)$$

must be satisfied and along $\frac{dt}{dx} = \frac{1}{\alpha}$, the relationship is

$$\alpha dp + e dx = 0. \tag{5.55}$$

Unfortunately

$$\lim_{x\to 0} \alpha = \lim_{x\to 0} e = 0$$

and so the slope of the characteristics given by $\frac{dt}{dx} = \frac{1}{\alpha}$ becomes steeper and steeper as x gets smaller and x = 0 is a singular line along which characteristics coincide. In its present form (5.55) cannot be used to find the change in p along a sloping characteristic when x is zero. The situation is somewhat simplified, however, by the observation that

$$\alpha = \sqrt{\frac{x}{2}}(\lambda\delta^3 + \frac{2}{\delta}) \quad \text{(from (5.52))} \tag{5.56}$$

and substituting for the right hand side of this expression from (5.21) gives $\alpha = \frac{-q}{p}$.

Hence, the sloping characteristics are given by the lines $\frac{dt}{dx} = \frac{-p}{q}$, that is, the lines such that f(x,t) = constant. So, instead of using (5.55) along the sloping characteristics, we use instead the equivalent equation that

$$f(x,t) = \text{constant}. \tag{5.57}$$

The fact that characteristics coincide along the line x = 0 is linked to the rapidly changing nature of the steady solution. As (5.26) shows $p(x,t) \sim x^{-\frac{1}{3}}$ as $x \to 0$ in this case. Therefore, a much finer grid is needed as x becomes small, in order to retain accuracy. For this reason the grid used in the algorithm consisted initially of the N_1+1 points $\{ih\}_{i=0,N_1}$

where $h = \frac{X(0)}{N_1}$, together with the N_2 grid points $\{\frac{h}{2^k}\}_{k=1,N_2}$. The grid, therefore, consisted of N_1+1 equally spaced points with a further N_2 points inserted into the first interval in a way such that points are clustered close to $x = 0$ and gradually spread out in the interval. At later time levels the boundary point $X(t)$ may not fall exactly on grid points and so it is then considered as an extra grid point.

As yet the condition along the line $x = 0$ has not been considered. Using this fact, together with (5.20) and (5.21), it is possible to obtain series expansions for p, q and δ which are valid for small x.

This gives

$$\delta = a(t)x^{\frac{1}{6}} + \frac{da}{dt}x^{\frac{5}{6}} + \ldots$$

$$p = -\frac{a^4}{2}x^{-\frac{1}{3}} + \ldots \qquad (5.58)$$

$$q = \frac{a^3}{\sqrt{2}} - 3\frac{da}{dt}a^3x^{\frac{2}{3}} + \ldots$$

where $a(t)$ is an unknown function of t only.

The unknown function $a(t)$ proves to be very important as if the solution is known along $x = 0$ (that is, if $a(t)$ has been found), then provided that $pq < 0$ (so that the gradient of the sloping characteristics is positive), knowledge of the solution along a line $t = t_j$ can be extended to give the solution along the line $t = t_{j+1}$.

The problem has three boundary conditions which could be used to solve for the unknowns $X(t)$, $\theta(t)$ and $a(t)$. Unfortunately it proves to be difficult to arrange these three boundary conditions in such a way that the three unknown

quantities can be determined and so the program uses two of the conditions to find $X(t)$ and $\theta(t)$ and then uses the remaining condition to check that a guess for $a(t)$, obtained from other sources, is approximately correct.

The final point about the algorithm which should be stressed is that calculations involving p when x is small are likely to be inaccurate as $p \to \frac{-U^{\frac{4}{3}}}{2^{\frac{1}{3}}x^{\frac{1}{3}}}$ as $x \to 0$. Care has therefore been taken to ensure that p is never used or calculated explicitly, except at the right hand boundary.

The details of the algorithm are as follows:

1. Assume $f, q, \delta, X, \theta, a$ and $\frac{da}{dt}$ are known on an initial time level, $t = t_j$.
2. (5.24) gives

$$\left.\frac{d\theta}{dt}\right|_{t=t_j} = -\int_0^{X(t)} f_t\left(1-\frac{\gamma\delta^4}{\lambda\delta^4+2}\right)dx \qquad \text{(using (5.21))}$$

$$= \int_0^{X(t_j)} f_t \left.\frac{(\frac{1}{4}\gamma\delta^4-2)}{(\frac{3}{4}\gamma\delta^4+2)}\right|_{t=t_j} dx$$

which gives $\theta'(t)\big|_{t=t_j}$.

3. $$\int_0^{X(t)} xff_x dx = \tfrac{1}{2}\left[xf^2\Big|_0^{X(t)} - \int_0^{X(t)} f^2 dx\right]$$

and so (5.23) gives

$$X'(t) = \frac{1}{Xff_x\big|_{X(t)}}\left\{-\frac{U}{2\pi\theta^{\frac{3}{4}}} + \int_0^{X(t)} ff_t dx - Xff_t\Big|_{X(t)}\right\}$$

which can be used to find $X'(t_j)$, where U is the velocity which 'matches' with γ (that is, $\gamma^5 U^4 = N^3$ where N=36.0598).

4. (5.22) is a 'checking' equation. Integrating by parts to remove f_x gives

$$X(t)f\Big|_{X(t)} - \int_0^{X(t)} f\,dx + \frac{1}{2\pi} = 0.$$

This is calculated for $t = t_j$ and if this condition is not approximately satisfied then the assumption that a(t) only changes very slowly is no longer valid and the rest of the calculation is without foundation. The assumption that a(t) will only change slowly is based upon the scale analysis of section 5.5 which suggested that change happens much more quickly away from the nose (that is, x=0) and so a(t) (which determines the solution close to the nose) is likely to be a slowly changing function.

5. Using $X'(t_j)$, $\theta'(t_j)$, $a'(t_j)$ values for X(t), $\theta(t)$ and a(t) can be obtained on $t = t_{j+1}$.

6. The position of the characteristics on the new time level is found by one of three methods.
When x is small $\delta^3 \simeq a^3 x^{\frac{1}{2}}$ is assumed to be small enough to ignore and so from (5.56) and (5.58)

$$\frac{dx}{dt} \simeq \frac{x^{\frac{1}{2}}}{2^{\frac{1}{2}}} \cdot \frac{2}{a x^{\frac{1}{6}}} = \frac{2^{\frac{1}{2}} x^{\frac{1}{3}}}{a}$$

which implies that the equation of the characteristic

through the point (x_i, t_j) is

$$\frac{2^{\frac{3}{2}}}{3a}\left(x^{\frac{2}{3}}-x_i^{\frac{2}{3}}\right) = t - t_j. \tag{5.59}$$

The leading term in x_i of the error when this expression is used to find the position of the characteristic at $t = t_{j+1}$ is given by

$$E_1 = \left[\frac{1}{\lambda a^3}\left\{\exp\left(\frac{2\lambda a^2}{3\sqrt{2}}k\right)(\lambda a^3 x_i^{\frac{2}{3}}+2)-2\right\}\right]^{\frac{3}{2}} - \left(x_i^{\frac{2}{3}}+\frac{2^{\frac{3}{2}}}{3a}k\right)^{\frac{3}{2}}. \tag{5.60}$$

An alternative method of approximating (5.56) is the standard explicit finite difference approximation,

$$\frac{x-x_i}{k} = \frac{x_i^{\frac{1}{2}}}{2^{\frac{1}{2}}}\left\{\lambda\delta^3+\frac{2}{\delta}\right\}\Bigg|_{(x_i,t_j)}. \tag{5.61}$$

The error resulting from this is, to leading order,

$$E_2 = \frac{k^2}{2}\left(\frac{\lambda^2 a^6}{2}x_i+\frac{4\lambda a^2}{3}x_i^{\frac{1}{3}} + \frac{2}{3a^2}x_i^{-\frac{1}{3}}\right). \tag{5.62}$$

The program swops from using (5.59) to using (5.61) to calculate the position of the characteristics, as soon as $|E_1| > |E_2|$.

Finally, (5.56) can be approximated by

$$x^{\frac{1}{6}} = x_i^{\frac{1}{6}} + \frac{k}{6x_i^{\frac{5}{6}}}\frac{x_i^{\frac{1}{2}}}{\sqrt{2}}\left(\lambda\delta^3+\frac{2}{\delta}\right)\Bigg|_{(x_i,t_j)} \tag{5.63}$$

which has an error with leading term

$$E_3 = \frac{k^2}{2}\left[\frac{\lambda^2 a^6}{12}x_i-\frac{\lambda a^2}{3}x_i^{\frac{1}{3}}\frac{x_i^{-\frac{1}{3}}}{a^2}\right]. \tag{5.64}$$

$E_3 < 0$ for small values of x_i and it is easy to show that $E_2 > -E_3$ if and only if

$$x_i > \left(\frac{1}{2\lambda a^4}\right)^{\frac{3}{2}} . \tag{5.65}$$

So (5.63) is used to calculate the position of the characteristics as soon as (5.65) is satisfied.

7. Using the fact that f is constant along characteristics, f can be calculated at all the points at which characteristics cross the new time level. Then, using cubic spline interpolation f can be calculated at grid points on the new time level to the right of the first characteristic.
Provided that the characteristic through $X(t_j)$ passes to the right of the boundary point $X(t_{j+1})$ on the new time level then f can also be calculated at the new boundary point $X(t_{j+1})$ by this method.
If the characteristic through $X(t_j)$ passes inside the new boundary point then an error is signalled.

8. Theoretically q could be approximated by

$$q_{i,j+1} = \frac{2}{k}(f_{i,j+1}+f_{i,j}) - q_{i,j}) \tag{5.66}$$

but in practice this is not satisfactory near to the first characteristic. For, f is changing rapidly when x is small and so slight errors in the position of the characteristics cause large errors in f and even larger ones in q (since they are multiplied by $^1/k$). So if f is changing too rapidly q is calculated using the small x algorithm (5.58).

In practice, q is calculated using (5.58) until (5.65) holds, that is

$$x_{i,j+1} > \left(\frac{1}{2\lambda a^4}\right)^{\frac{3}{2}} .$$

This is the point at which the program switches to the third method of calculating the characteristics and so the error in the characteristic calculation is given by (5.64). If errors resulting from interpolation are ignored then the error in f at (ih,(j+1)k) is given by,

$$\text{error in } f = \begin{matrix}\text{error in position} \\ \text{of characteristic}\end{matrix} \times f_x .$$

Approximating $f_x\Big|_{i,j+1}$ by

$$\frac{f_{i+1,j+1} - f_{i,j+1}}{h}$$

gives an error in q of

$$E_q = \frac{E_3 \cdot 2}{hk} \left(f_{i+1,j+1} - f_{i,j+1}\right)$$

and the program only starts calculating q by (5.66) instead of the small x algorithm (5.58) when $|E_q| < .01$.
It is not possible to calculate q by (5.66) at any points x_i such that $x_i > X(t_j)$. (That is, points which lie outside the old boundary.) In this case f is used to give a finite difference approximation to f_x and then (5.21) is used to find q. Since δ has not been calculated on the new time level the values on $t = t_j$ are used to give a first approximation to q. The calculation of δ uses q on the new time level in (5.20) as explained in 10, and so these two

calculations can be iterated to convergence.

9. Knowledge of q along all of the line $t = t_j$ enables f to be calculated at grid points to the left of the first characteristic using

$$f_{i,j+1} = f_{i,j} + \frac{k}{2}(q_{i,j}+q_{i,j+1}) \quad . \tag{5.67}$$

10. δ can be calculated by using

$$\delta(0,t_{j+1}) = 0$$

$$\delta(x_1,t_{j+1}) = a(t_{j+1})x_1^{\frac{1}{6}}$$

where x_1 is the first grid point.

(5.20) gives

$$\sqrt{x}\delta^3\Big|_{x_{i+1}} - \sqrt{x}\delta^3\Big|_{x_i} = \frac{\sqrt{2}}{2}(x_{i+1}-x_i)(q_{i+1}+q_i)$$

which can be used to calculate δ at all further grid points and the boundary $X(t_{j+1})$.

11. Finally $a'(t_{j+1})$ must be calculated. Due to the physical basis of the problem it is reasonable to suppose that q will be smooth along the line $t = t_{j+1}$. So, if the left-most q to be evaluated from (5.58) (rather than from (5.57)) is q* and its x coordinate is x* we have

$$q^* = \frac{(a(t_{j+1}))^3}{\sqrt{2}} - 3(a(t_{j+1}))^3 \frac{da}{dt}\Big|_{t_{j+1}} (x^*)^{\frac{2}{3}}$$

and

$$a(t_{j+1}) = a(t_j) + \frac{k}{2}\left(\left(\frac{da}{dt}\right)\Big|_{t_j} + \frac{da}{dt}\Big|_{t_{j+1}}\right) .$$

These two equations can be solved for $\left.\frac{da}{dt}\right|_{t_{j+1}}$ and $a(t_{j+1})$. (this refines $a(t_{j+1})$.)

This new value of $a(t_{j+1})$ is used to refine the values of q to the left of q*.

12. All the quantities which are required have now been found and it is possible to return to 1. with $t=t_{j+1}$ and continue onto a new time level.

The remaining consideration is that of initial conditions. The interest in the problem and the above analysis was for a small perturbation in the steady solution. The following initial conditions were therefore considered.

$$f = -\frac{3}{2\lambda}\log(2^{\frac{1}{3}}+\lambda x^{\frac{2}{3}}V^{\frac{4}{3}})$$

$$q = V$$

$$\delta = (\sqrt{2}V)^{\frac{1}{3}}x^{\frac{1}{6}} \quad .$$

This satisfies (5.20) and (5.21) and the condition that $\delta^3 f_x$ is bounded on $x = 0$.

The small x expansion (5.58) gives $a = (\sqrt{2}V)^{\frac{1}{3}}$, $\frac{da}{dt} = 0$ for these initial conditions.

The boundary condition (5.22) can be integrated to give a nonlinear equation for X(0) which can be solved using Newton's method. Since $f_{xt} = q_x = 0$ when $t = 0$, (5.22) also gives $X'(0) = 0$.

$\theta(0)$ can be obtained from (5.23). For, since (5.22) holds

$$\frac{\partial}{\partial t}\int_0^{X(t)} xf\frac{\partial f}{\partial x}dx = -\frac{V}{2\pi}$$

and so taking $\theta(0) = (\frac{U}{V})^{\frac{4}{3}}$ ensures that (5.23) is satisfied.

Finally, using (5.22) to integrate (5.24) gives

$$\left.\frac{d\theta}{dt}\right|_0 = V(\frac{\gamma}{2\pi}-X(0)) \quad . \tag{5.68}$$

The above initial conditions satisfy equations (5.20)-(5.24) and they provide enough information for the solution to be calculated for $t > 0$ using the algorithm 1-12.

The results for this algorithm are discussed in the following section.

§5.9 Results and Conclusions

The algorithm was used with initial conditions

$$f(x,0) = f_0(x,0)$$

$$q = V$$

$$\delta = (\sqrt{2}V)^{\frac{1}{3}}x^{\frac{1}{6}} \tag{5.69}$$

where $f_0(x,t) = -\frac{3}{2\lambda}\log(2^{\frac{1}{3}}+\lambda x^{\frac{2}{3}}V^{\frac{4}{3}}) + Vt$

and $\gamma = 1$ (and, therefore, $\lambda=.75$).

The 'matching' velocity, U, which would give the steady solution is U = 14.715. The two perturbations considered were V = 13 and V = 16. This gives a(0) = 2.639 when V = 13 and a(0) = 2.828 when V = 16 (as compared with a(0) = 2.750 when U = V = 14.715). Also, $\theta(0) = 1.18$ and $X(0) = .1646$ when V = 13 and $\theta(0) = .89$ and $X(0) = .1558$ when V = 16. (When V=U, $\theta(0)=1$ and $X(0)=.1592$.)

The grid used in the x direction had 50 regularly spaced points and 50 points inserted into the first interval (that is, $N_1=N_2=50$). Thus $h = \frac{X(0)}{50} \simeq .003$ for the values of X(0) arising in the above two situations. Three different step sizes in the t direction were considered, k = .002, k = .001 and k = .0005 and in all these cases k < h.

The results for f, δ and q at t = .04 are given in Tables 5.2-5.7. It is easy to see that the numerical solution is tending linearly as $k \to 0$ towards the solution given by (5.69) with $f(x,t) = f_0(x,t)$.

This numerical result can be supported by the observation that (5.69) with $f(x,t) = f_0(x,t)$ satisfies (5.20) and (5.21) for any V. The boundary conditions have no effect on the numerical solution as (5.22) is ignored and (5.23) and (5.24) only provide information about $\theta(t)$ and X(t).

The results for $\theta'(t)$ and $X'(t)$ are given in Tables 5.9 and 5.10 whilst 5.8 gives $a'(t)$. It is seen that $X'(t) < 0$ for both perturbations. All perturbations and step sizes considered and any other alternative numerical algorithms that have been tried give the result that for $t > t_c$, $X'(t) < 0$ where t_c is some critical value of t.

Eventually the boundary moves inwards so quickly that the method breaks down. For example, with the above initial conditions and V = 16 and k = .005 the boundary collapses after 45 steps, that is, at t = .225.

The above algorithm is only valid if the assumption that q is smooth (which replaces (5.22)) results in a solution which

satisfies (5.22). Table 5.11 shows the values of (5.22) for V = 13 and V = 16 at t = .04 for the three different values of k. Whilst the integral is small it is not negligible compared with the truncation errors. Analytically it is obvious that unless $X'(t) = 0$ and $\theta'(t) = 0$ for all $t \geq 0$ then no solution to (5.20)-(5.24) will ever be found by this algorithm since, even if $X'(t) \to 0$ and $\theta'(t) \to 0$ as $t \to \infty$, $X(t) \neq X(0)$ and so (5.22) will not be satisfied.

It is possible that the root of the difficulty is the fact that, as explained in section 5.8, characteristics coincide on the singular line x = 0. The above algorithm assumed that q was smooth and used this fact to obtain a(t). The lack of success of this approach indicates that there may be some sort of weak singularity propagated along the characteristic emanating from the origin. In this case there may be many solutions to the problem depending upon the exact nature of the singularity across this characteristic.

Alternatively the problem may be ill-posed. That this is a possibility is indicated by the evidence of sections 5.6 and 5.7, together with the latter part of Chapter 4. For the equation near to the nose (x=0) was shown in section 5.5 to be likely to be important in determining the solution elsewhere. In section 5.5 and in section 5.6 this equation was shown to be similar to the possibly ill-posed problem of Chapter 4. Also the linear perturbation of the full problem, analysed in section 5.7 gave a similar equation to that of Chapter 4.

Finally, numerical algorithms have difficulties incorporating the boundary conditions into the scheme in a

stable way. Many approaches have been tried. The main difficulty appears to stem from the fact that information must be obtained about a(t) from the boundary conditions. a(t) will eventually influence the whole solution and so it is important that it is calculated accurately. However, all three boundary conditions are almost independent of a(t). This is because all three are integral conditions and the main contribution to the integral is not from the part of the integrand near to x = 0. It is, therefore, difficult to calculate a(t) accurately and any errors which are made will eventually influence the whole solution.

It is interesting to note that with any of the numerical algorithms attempted the boundary moves inwards. This behaviour occurs even when (5.22) is satisfied. (That is, volume is conserved.) As the boundary moves inwards the function $\frac{\partial f}{\partial x}$ increases in modulus close to the origin, (and is, of course, infinite at the origin), so that (5.22) can still be satisfied. This type of behaviour is reminiscent of the 'sharp spike' discussed in section 4.5 of Chapter 4.

All the numerical experiments took as initial conditions a slightly perturbed 'steady state' solution. In every case the computed solution did not tend to a neighbouring 'steady' solution; instead the boundary point moved towards the origin. However, the numerical evidence is not yet entirely conclusive and this behaviour may be due to the accumulation of numerical errors.

The above model assumed that there were no heat sources or sinks. It is interesting to note that if they are introduced

they occur in equation (5.24), the boundary condition which gives conservation of heat.

Therefore, any initial conditions of the form considered above can lead to a steady solution provided that heat is lost or gained in a fashion such that $\frac{d\theta}{dt} = 0$. So from (5.68) it is seen that a heat sink is required if $X(0) < \frac{\gamma}{2\pi}$ and a heat source if $X(0) > \frac{\gamma}{2\pi}$.

TABLE 5.1

γ	U (a)	U (b)
.5	35.04	35.00
.6	27.92	27.87
.7	23.05	22.98
.8	19.53	19.45
.9	16.90	16.79
1.0	14.82	14.72
1.1	13.21	13.06
1.2	11.90	11.72
1.3	10.78	10.60
1.4	9.88	9.66
1.5	9.09	8.86

VALUES OF U AND γ FOR a) EQUATIONS (5.18) AND (5.19)
AND b) THE MODIFIED EQUATION (5.31) OBTAINED BY REMOVING THE '1'

TABLE 5.2

x	$-\frac{3}{2\lambda}\log(2^{\frac{1}{3}}+\lambda x^{\frac{2}{3}}V^{\frac{4}{3}})+V_t$	Numerical Solution for f		
		(a)	(b)	(c)
.1533E-11	.5790E-1	.5784E-1	.5787E-1	.5789E-1
.1608E-5	.5291E-1	.5285E-1	.5288E-1	.5290E-1
.3293E-1	-2.050	-2.049	-2.050	-2.050
.9878E-1	-3.116	-3.115	-3.116	-3.116
.1613	-3.652	-3.652	-3.652	-3.652

ANALYTICAL SOLUTION AND NUMERICAL SOLUTION FOR f WITH STEP SIZES

(a) k=.002, (b) k=.001 AND (c) k=.0005 AT t=.04 AND WITH V=13

TABLE 5.3

x	$-\frac{3}{2\lambda}\log(2^{\frac{1}{3}}+\lambda x^{\frac{2}{3}}V^{\frac{4}{3}})+V_t$	Numerical solution for f		
		(a)	(b)	(c)
.1451E-11	.1779	.1778	.1779	.1779
.1521E-5	.1716	.1715	.1715	.1715
.3116E-1	-2.256	-2.255	-2.255	-2.256
.9348E-1	-3.387	-3.386	-3.387	-3.387
.1527	-3.945	-3.945	-3.945	-3.945

ANALYTICAL SOLUTION AND NUMERICAL SOLUTION FOR f WITH STEP SIZES

(a) k=.002, (b) k=.001 AND (c) k=.0005 AT t=.04 AND WITH V=16

x	Numerical solution for δ $-(\sqrt{2}V)^{\frac{1}{3}}x^{\frac{1}{6}}$		
	(a)	(b)	(c)
.1533E-11	-.2505E-5	-.1213E-5	-.5391E-6
.1608E-5	-.2499E-4	-.1208E-4	-.5361E-5
.3293E-1	.8214E-3	.4127E-3	.1809E-3
.9878E-1	.1101E-2	.5386E-3	.2404E-3
.1613	.9874E-3	.4979E-3	.2221E-3

NUMERICAL SOLUTION FOR $\delta-(\sqrt{2}V)^{\frac{1}{3}}x^{\frac{1}{6}}$ WITH STEP SIZES

(a) k=.002, (b) k=.001 AND (c) k=.0005 AT t=.04 AND WITH $V=13$

TABLE 5.5

x	{Numerical solution for q} - 16.00		
	(a)	(b)	(c)
.1451E-11	-.4296E-2	-.1987E-2	-.9053E-3
.1521E-5	-.4194E-2	-.1941E-2	-.8827E-3
.3116E-1	.4442E-1	.1999E-1	.9050E-2
.9348E-1	.2090E-1	.1118E-1	.5110E-2
.1527	.1873E-1	.7439E-2	.3462E-2

NUMERICAL SOLUTION FOR q -16.00 WITH STEP SIZES

(a) k=.002, (b) k=.001 AND (c) k=.0005 AT t=.04 AND WITH V=16

x	{Numerical solution for q} - 13.00		
	(a)	(b)	(c)
.1533E-11	-.3443E-2	-.1665E-2	-.7379E-3
.1608E-5	-.3387E-2	-.1637E-2	-.7256E-3
.3293E-1	.3654E-1	.1708E-1	.74102-2
.9878E-1	.1698E-1	.1002E-1	.4288E-2
.1613	.1238E-1	.5504E-2	.2785E-2

NUMERICAL SOLUTION FOR q-13.00 WITH STEP SIZES

(a) k=.002, (b) k=.001 AND (c) k=.0005 AT t=.04 AND WITH V=13

x	$\left\{\text{Numerical solution for } \delta\right\} - (\sqrt{2}V)^{\frac{1}{3}}x^{\frac{1}{6}}$		
	(a)	(b)	(c)
.1451E-11	-.2697E-5	-.1249E-5	-.5706E-6
.1521E-5	-.2678E-4	-.1239E-4	-.5646E-5
.3116E-1	.1277E-2	.5789E-3	.2794E-3
.9348E-1	.1365E-2	.6383E-3	.3008E-3
.1527	.1225E-2	.5676E-3	.2711E-3

NUMERICAL SOLUTION FOR $\delta-(\sqrt{2}V)^{\frac{1}{3}}x^{\frac{1}{6}}$ WITH STEP SIZES

(a) k=.002, (b) k=.001 AND (c) k=.0005 AT t=.04 AND WITH V=16

TABLE 5.8

	V = 13.			V = 16.		
t	(a)	(b)	(c)	(a)	(b)	(c)
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
.01	-.4949E-2	-.2370E-2	-.1113E-2	-.4786E-2	-.2284E-2	-.1026E-2
.02	-.5829E-2	-.2853E-2	-.1340E-2	-.6154E-2	-.2836E-2	-.1255E-2
.03	-.7097E-2	-.3286E-2	-.1490E-2	-.7300E-2	-.3353E-2	-.1357E-2
.04	-.7358E-2	-.3696E-2	-.1623E-2	-.1138E-1	-.5184E-2	-.2509E-2

a'(t) FOR V=13. AND V=16. WITH (a) k=.002, (b) k=.001 and (c) k=.0005

TABLE 5.9

	V = 13.			V = 16.		
t	(a)	(b)	(c)	(a)	(b)	(c)
0.00	-.7128E-1	-.7128E-1	-.7128E-1	.5369E-1	.5369E-1	.5369E-1
.01	-.7182E-1	-.7213E-1	-.7197E-1	.5207E-1	.5203E-1	.5258E-1
.02	-.7220E-1	-.7246E-1	-.7197E-1	.5092E-1	.5118E-1	.5245E-1
.03	-.7260E-1	-.7280E-1	-.7199E-1	.4949E-1	.5039E-1	.5230E-1
.04	-.7317E-1	-.7321E-1	-.7202E-1	.4834E-1	.5038E-1	.5224E-1

θ'(t) FOR V=13. AND V=16. WITH (a) k=.002, (b) k=.001 AND (c) k=.0005

	V = 13.			V = 16.		
t	(a)	(b)	(c)	(a)	(b)	(c)
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
.01	-.2862E-1	-.2283E-1	-.2352E-2	-.5129E-1	-.3498E-1	-.6789E-2
.02	-.3220E-1	-.2471E-1	-.2735E-2	-.5873E-1	-.3787E-2	-.7140E-2
.03	-.3648E-1	-.2713E-1	-.3177E-2	-.6788E-1	-.9010E-2	-.7571E-2
.04	-.4168E-1	-.2969E-1	-.3661E-2	-.3907E-1	-.1022E-1	-.8024E-2

X'(t) FOR V=13. AND V=16. WITH (a) k=.002, (b) k=.001 AND (c) k=.0005

TABLE 5.11

	V = 13.			V = 16.		
t	(a)	(b)	(c)	(a)	(b)	(c)
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
.01	0.1E-3	0.1E-3	-.9E-4	0.3E-3	0.2E-3	-.6E-4
.02	0.4E-3	0.4E-3	-.6E-4	0.9E-3	0.6E-3	0.3E-4
.03	0.8E-3	0.7E-3	-.3E-4	0.2E-2	0.1E-2	0.1E-3
.04	0.1E-2	0.1E-2	0.2E-4	0.2E-2	0.1E-2	0.2E-3

THE VALUE OF THE IGNORED BOUNDARY CONDITION (5.22) FOR

V=13. AND V=16. WITH (a) k=.002, (b) k=.001 AND (c) k=.0005

Figure 5.3

Graph of $-\log_{10} U^4$ against γ

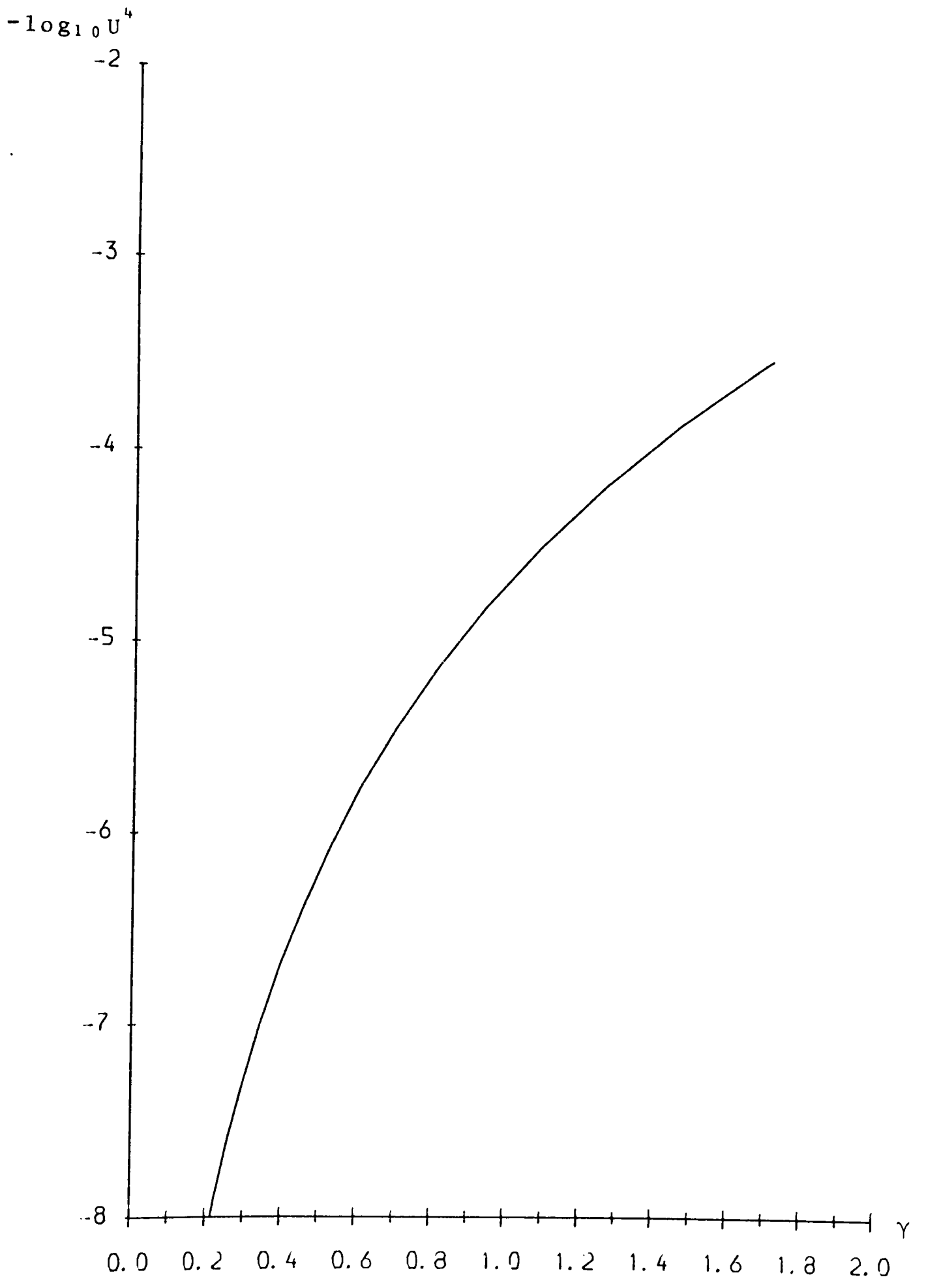
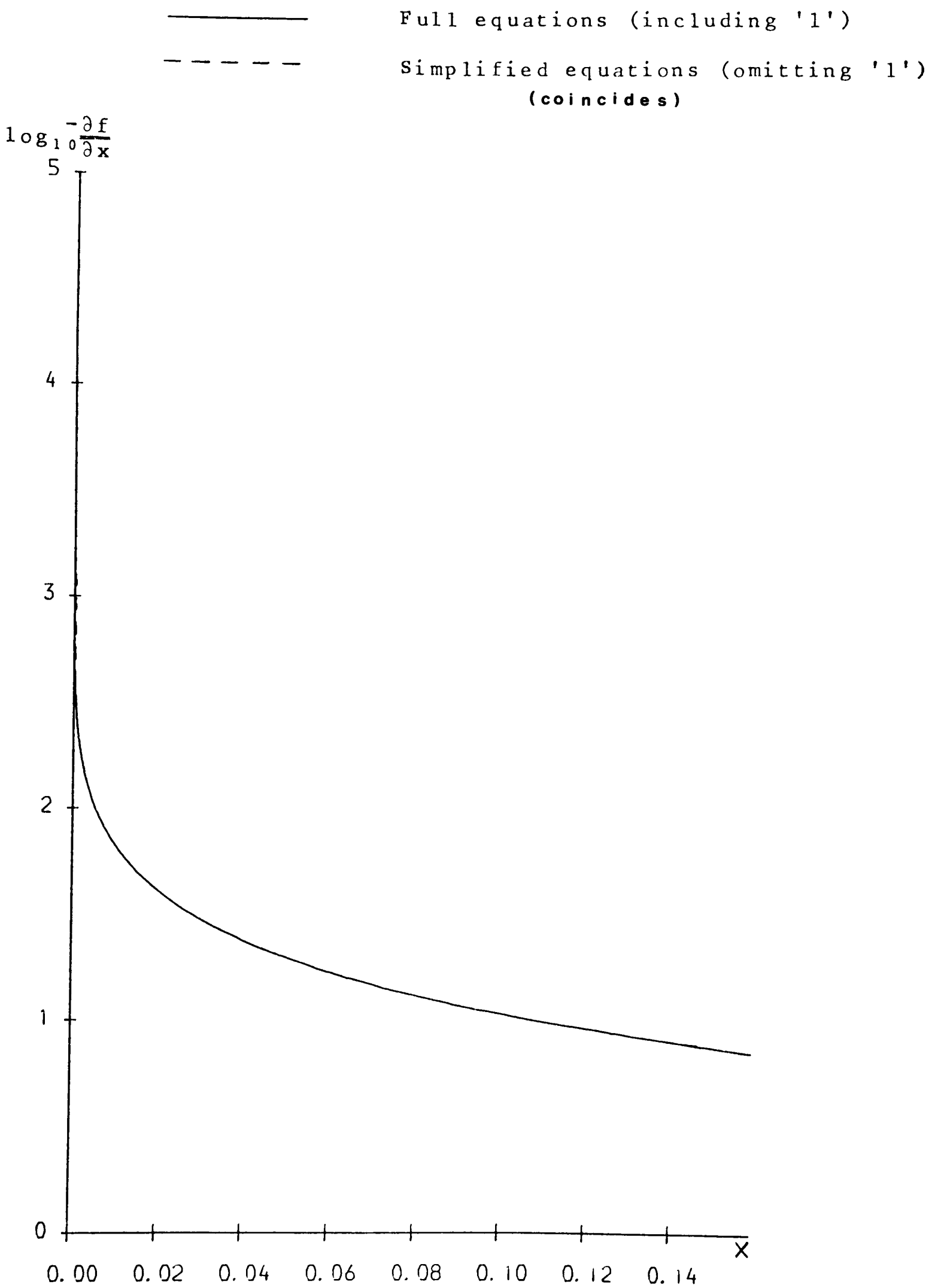


Figure 5.4



$\log \frac{-\partial f}{\partial x}$ against x when $\gamma=1$, U=14.75

REFERENCES

1. Ames, W.F., Numerical methods for partial differential equations (second edition), Academic Press (1977).

2. Batchelor, G.K., An introduction to fluid dynamics, Cambridge University Press (1967).

3. Breslin, J.P., Dynamic forces exerted by oscillating cables, Journal of Hydronautics, Vol.8, 1974, pp.19-31.

4. O'Brien, C.G., Hyman, M.A. and Kaplan, S., A study of the numerical solution of partial differential equations, J. Math. Phys., Vol.29, 1951, pp.223-251.

5. de Bruijn, N.G., Asymptotic methods in analysis, Interscience (1958).

6. Carrier, G.F., Krook, M. and Pearson, C.E., Functions of a complex variable, McGraw-Hill (1966).

7. Courant, R. and Hilbert, D., Methods of mathematical physics, Vol.II, Partial differential equations, Interscience (1962).

8. Elsasser, W.M., Early history of the earth, in 'Earth science and meteorites',pp.1-30, North-Holland (1963).

9. Ferris, D.H., Numerical determination of the configuration of an underwater umbilical subjected to steady hydrodynamic loading. Part 2: Unsteady aspects. National Physical Laboratory, DNACS Report 50/81, 1981.

10. Goodman, T.R. and Breslin, J.P., Statics and dynamics of anchoring cables in waves, Journal of Hydronautics, Vol.10, No.4, 1976, pp.113-120.

11. Gruntfest, I.J., Thermal feedbacks in liquid flow; plane shear at constant stress. Trans. Soc. Rheol. Vol.7, 1963, pp.195-207.

12. Hartree, D.R., Numerical analysis (second edition), Oxford University Press (1958).

13. Hewitt, J.M., McKenzie, D.P. and Weiss, N.O., Dissipative heating in convective flows, J. Fluid Mech., Vol.68, 1975, pp.721-738.

14. Hodgkin, A.L. and Huxley, A.F., A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. Vol.117, 1952, pp.500-544.

15. Huffman, R.R. and Genin, J., The dynamical behaviour of a flexible cable in a uniform flow field. Aeronautical Quarterly, 1971, pp.183-195.

16. Lacey, A.A., The spatial dependence of supercritical reacting systems. IMA J. Appl. Math. Vol.27, 1981, pp.71-84.

17. Lax, P.D. and Wendroff, B., Systems of conservation laws. Comm. Pure Appl. Math. Vol.13, 1960, pp.217-237.

18. Marsh, B.D., On the cooling of ascending andesitic magma. Phil. Trans. R. Soc. Lond. A288, 1978, pp.611-625.

19. McKean, H.P., Nagumo's equation. Advances in mathematics, Vol.4, 1970, pp.209-223.

20. Melosh, M.J. and Ebel, J., A simple model for thermal instability in the asthenosphere. Geophys. J. Roy. Astron. Soc., Vol.59, 1979, pp.419-436.

21. Milne-Thomson, L.M., Theoretical hydrodynamics (fifth edition). Macmillan (1968).

22. Morris, S., The effects of a strongly temperature-dependent viscosity on slow flow past a hot sphere. J. Fluid Mech., Vol.124, 1982, pp.1-26.

23. Nitsan, U., Viscous heat production in a slab. J. Geophys. Res., Vol.78, 1973, pp.1395-97.

24. Ockendon, J.R and Tayler, A.B., Oscillations of a neutrally buoyant cable. Oxford Study Group report (1981).

25. ———————————, Emermann, S.H. and Turcotte, D.L., Geodynamic thermal runaway with melting. To appear. 1985.

26. Paidoussis, M.P., Dynamics of flexible cylinders in axial flow. Part 1. Theory. J. Fluid Mech., Vol.26, 1966, pp.717-736.

27. ———————————, Stability of towed totally submerged flexible cylinders. J. Fluid Mech., Vol.34, 1968, pp.273-297.

28. Richtmyer, R.D. and Morton, K.W., Difference methods for initial value problems. Interscience (1967).

29. Rinzel, J. and Keller, J.B., Travelling wave solutions of a nerve conduction equation. Biophys. J. Vol.13, 1973, pp.1313-1337.

30. Rozhdestvenskii, B.L., Discontinuous solutions of hyperbolic systems of quasi linear equations. Russian Math Surveys, Vol.15, No.6, 1960, pp.53-111.

31. Taylor, G.I., Analysis of the swimming of long and narrow animals. Proc. Roy. Soc. A214, 1952, pp.158-183.

32. Turcotte, D.L., Hsui, A.T., Torrance, K.E. and Schubert, G., Influence of viscous dissipation on Benard convection. J. Fluid Mech., Vol.64, 1974, pp.369-374.

33. Yuen, D.A. and Schubert, G., On the stability of frictionally heated Shaw flows in the asthenophere. Geophys. J. Roy. Astron. Soc., Vol.57, 1979, pp.189-207.