UNIVERSITY OF READING

MOVING FINITE ELEMENTS AND OVERTURNING SOLUTIONS

by

Carol Patricia Reeves

Department of Mathematics

This thesis is submitted for the degree of
Doctor of Philosophy
at the University of Reading

Submission date : September 1991

Abstract

In this thesis we examine overturned solutions of scalar partial differential equations in one and two dimensions using moving finite element methods with particular emphasis on scalar conservation laws. These equations are the simplest nonlinear equations to exhibit the formation of shocks and expansions as their solutions evolve with time.

Both analytic and numerical techniques are examined in one and two dimensions, analytic techniques being considered as a background to the numerical methods, which are adaptive and finite element in nature. They include the classical moving finite element method (MFE) of Miller in its various forms and Lagrangian methods.

The analytic and numerical solution to these equations yield multivalued curves and surfaces. Weak solutions however exist in which shocks feature and these can be obtained from the multivalued solutions by applying a recovery technique to locate the shock position.

To enable this technique to be implemented, in the case of MFE methods, they must be rewritten as a two stage procedure, so as to properly define the method mathematically.

The shock recovery techniques used are based on conservation. One method is based directly on conservation of area, while a second method uses this indirectly after the application of a Legendre transformation. A third method considered is based on the Transport Collapse operator of Brenier. All methods are entropy satisfying. In 2-D the shock recovery method used is 1-D in nature normal to the shock and involves several applications of the 1-D construction in order to find the shock position.

Acknowledgements

I would like to thank Dr. M.J. Baines for his enthusiastic supervision, the many hours of his time spent discussing this work and for his constant good humour during the past three years, without all of which I would never have completed this work.

I am also grateful to Dr. P.K. Sweby for both his help with the computational work and for his friendship.

Many thanks are also given to members of staff and to fellow students in the Department of Mathematics who I have known and who have made my six years at Reading so enjoyable. I would like to take this opportunity to express my gratitude to my family and friends for their support during this period.

I acknowledge the financial support of the Science and Engineering Research Council for the past three years.

Contents

Chapter 1

Introduction

Partial Differential Equations (PDE's) and the techniques for their solution have been widely studied for more than 200 years. One of the reasons that there is such great interest in these equations stems from the fact that they can be used to describe physical systems. PDE's may be used to model the behaviour of many natural phenomena such as the weather, the motion of the sea or the flow of a river. They may also be used to describe the interaction between natural phenomena and man-made structures such as the flow in harbours or around aircraft.

Initially interest in this type of equation occurred after Newton and Leibniz independently introduced the ideas of the calculus in the late 17th century. Once the relationship between differentiation and integration became known, this marked the beginning of the study of solutions of Ordinary Differential Equations (ODE's). ODE's grew in popularity throughout the 18th century as their ability to represent physical situations became known. By the mid 18th century partial differential equations had been introduced and simple methods of solution were being developed. For example, whilst studying the properties of vibrating strings, D'Alembert was led to the PDE $\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$ and subsequently gave its solution as $u = f(x+t) + g(x-t)$ where f, g are arbitrary functions. Similar advances were made by other mathematicians, of which one of the most famous was Euler. Euler was responsible for the introduction of systematic methods of solution of differential equations and introduced techniques such as the use of integrating factors.

It is therefore no secret that over the past 200 years a great deal of progress has been made in the classical solution of in particular, linear PDE's. Unfortunately, the same is not true of the nonlinear equations, which have many more practical applications. Disregarding numerical techniques, the classes of equations for which analytic solutions are available remains distressingly small and still only the simplest nonlinear problems have analytic solutions.

PDE's may be classified in many ways so that their behaviour may be generalized. For example second order linear PDE's may be divided into three classes, hyperbolic, parabolic and elliptic equations depending on the coefficients of the second order derivatives. For each of these classes not only do solution techniques vary, but different boundary conditions are needed for the formation of well posed problems. Other types of behaviour are categorised by physical analogies, such as a u_{xx} term being regarded as a diffusion term whereas a u_x term is an advection term. Other examples of this qualitative behaviour include the well known phenomena of blow-up (see chapter 2, section 2.5).

Although during the last 30-40 years with the advent of numerical techniques, the solution of PDE's has become increasingly practical, it however remains highly desirable to know something about the type of behaviour of the solution of the equation. This knowledge of the behaviour usually comes from the analytical or classical theory and may be used to facilitate the choice of an appropriate numerical method.

In this thesis we shall concentrate on first order scalar PDE's and their properties, and in particular on conservation laws. We shall be concerned with the study of the discontinuities via overturning solutions and the numerical representation of this phenomena using moving elements.

Consider the general first order scalar PDE in n-dimensions,

$$F(\mathbf{x}, u, \mathbf{m}) = 0 \tag{1.1}$$

where $\mathbf{x} = (x_1, ..., x_n)$, $u = u(\mathbf{x})$ and $\mathbf{m} = (m_1, ..., m_n)$ where $m_i = \frac{\partial u}{\partial x_i}$ $(i = 1, ..., n)$. This equation is too general to be solved analytically although it is possible to associate it with a system of ODE's or characteristics (Courant &

Hilbert (1962)),

$$\begin{aligned} \frac{dx_i}{ds} &= F_{m_i} \quad i = 1, ..., n && (1.2)\\ \frac{du}{ds} &= \sum_{i=1}^{n} m_i F_{m_i} && (1.3)\\ \frac{dm_i}{ds} &= -(F_u m_i + F_{x_i}) \quad i = 1, ..., n. && (1.4) \end{aligned}$$

This gives a system of $2n + 1$ ODE's and the equation $F = 0$ for the $2n + 1$ unknowns $x_1, ..., x_n, u, m_1, ..., m_n$. Also, it may be shown that F is an integral of the characteristics, the only requirement being that $F = 0$ at some initial point $t = 0$, hence F is satisfied for all t (John (1971)). These characteristic equations may be used to examine the underlying structure of the PDE (1.1), but little progress may be made analytically until it is simplified.

Equation (1.1) may be simplified by considering x_n=time t as a preferred or distinguished variable. This separates t from the other variables x_i leaving $\mathbf{x}$ containing only spatial variables. If t does not occur anywhere else in the equation, (1.1) reduces to the important practical form

$$u_t + H(\mathbf{x}, u, \mathbf{m}) = 0 \tag{1.5}$$

where $\mathbf{x} = (x_1, ..., x_{n-1})$, $\mathbf{m} = (m_1, ..., m_{n-1})$, u is as above and where $n-1$ is the number of space dimensions. The characteristics can be simplified very slightly for this equation because the t time variable can be extracted from the equations to give

$$\frac{dt}{ds} = 1 \quad \text{and} \quad \frac{du_t}{ds} = -H_u u_t. \tag{1.6}$$

The remaining characteristics are given by

$$\begin{aligned} \frac{dx_i}{ds} &= H_{m_i} \quad i = 1, ..., n-1\\ \frac{du}{ds} &= \sum_{i=1}^{n-1} m_i H_{m_i} + u_t && (1.7)\\ \frac{dm_i}{ds} &= -(H_u m_i + H_{x_i}) \quad i = 1, ..., n-1. \end{aligned}$$

From (1.6), t may replace s in (1.7). The problem is characterized by data given on some initial line. This equation is still too general to be solved analytically and the characteristics (1.6), (1.7) have not been simplified very much from those of equation (1.1).

Equation (1.1) may be simplified further by assuming only linear dependence of H on u_{x_i} so as to separate it into two new functions $\mathbf{a}$ and b which are dependent on $\mathbf{x}$ and u. The equation is now given by

$$u_t + \mathbf{a}(\mathbf{x}, u).\boldsymbol{\nabla} u = b(\mathbf{x}, u). \tag{1.8}$$

The characteristics can be simplified further to give

$$\begin{aligned}
\frac{dx_i}{ds} &= \mathbf{a}_i(\mathbf{x}, u) \quad i = 1, ..., n-1 && (1.9)\\
\frac{du}{ds} &= \sum_{i=1}^{n-1} \mathbf{a}_i(\mathbf{x}, u) m_i + u_t = b(\mathbf{x}, u) && (1.10)\\
\frac{dm_i}{ds} &= -((\mathbf{a}_u.\boldsymbol{\nabla} u - b_u)m_i + \mathbf{a}_{x_i}.\boldsymbol{\nabla} u - b_{x_i}) \quad i = 1, ..., n-1. && (1.11)
\end{aligned}$$

We can simplify equation (1.8) yet again by allowing $\mathbf{a}$ to be dependent only on u and $b \equiv 0$. This gives the equation

$$u_t + \mathbf{w}(u).\boldsymbol{\nabla} u = 0. \tag{1.12}$$

This equation is a conservation law if $\mathbf{w}$ is integrable so that $\mathbf{w}(u).\boldsymbol{\nabla} u$ may be written as $L(u)_{x_i}$ where $\nabla L(u) = \mathbf{w}(u)$. The characteristics of this equation are given by

$$\begin{aligned}
\frac{dx_i}{ds} &= \mathbf{w}(u)_i \quad i = 1, ..., n-1 && (1.13)\\
\frac{du}{ds} &= \sum_{i=1}^{n-1} \mathbf{w}(u)_i m_i + u_t = 0 && (1.14)\\
\frac{dm_i}{ds} &= -\mathbf{w}(u).\boldsymbol{\nabla} u \quad i = 1, ..., n-1. && (1.15)
\end{aligned}$$

The conservation laws exhibit the types of behaviour which we are investigating in this thesis and, since they are much simpler than (1.1), they provide our main set of examples. Conservation laws form shocks and/or expansions, which allow multivalued solutions to be formed by following the characteristics. It is this type of behaviour that we are interested in and so we concentrate on the conservation laws.

In chapter 2 we examine conservation laws in 1-D, discussing both their properties and their method of solution using characteristics. In the section on characteristics, weak forms are introduced in order to allow discontinuous solutions

to exist (shocks). As a consequence, the notion of entropy (Lax (1972)) has to be introduced as a solution to the problem of non-uniqueness arising from the use of such weak solutions. A second approach is then introduced using the method of characteristics in which the solution is allowed to follow the characteristics through their intersection point (i.e. initial time of forming a shock) so that a multivalued solution is obtained. From this multivalued curve, the shock position can then be calculated in several ways. One method is based upon the ideas of conservation, in conjunction with transformations to Hamilton-Jacobi equations and ODE's. A second method, proposed by Brenier, uses an averaging technique which is applied to the multivalued curve.

Chapter 3 contains the background to numerical methods used in 1-D to solve conservation laws. Initially we focus on finite difference and finite element methods appropriate for the solution of such conservation laws. This leads to a discussion of the possible use of adaptive finite element methods. The moving finite element method of (Miller & Miller (1981)) and its derivatives are examined together with Lagrangian methods. Since the adaptive element methods permit multivalued solutions to form, numerical techniques analogous to those in chapter 2 can be used to find the shock position. This is the main aim of this thesis.

In chapter 4, it is noted that the formal procedure used in the MFE method may not be valid when the solution overturns. This occurs because the norm used in the L_2 minimisation is no longer a true norm. In this chapter the norm is rewritten as the sum of two norms, each of which remains valid. A variety of MFE type methods are then examined in the context of multivalued solutions and their implementations are explained. The final section of this chapter is devoted to the numerical methods of recovery of the shock position from the multivalued curve.

Chapter 5 concludes the discussion of the problem in 1-D by giving numerical examples and results to highlight the points made in chapters 2-4. Examples are given to show both the methods of obtaining overturned solutions and the methods of recovering the shock position.

Chapters 6-9 cover the same subjects as in chapters 2-5, but for higher dimensions. In chapter 6 conservation laws in two and higher dimensions are considered

analytically, together with possible methods of solution. In 2-D only a few equation have known analytic solutions, although it has been shown that for general initial data solutions to conservation laws in 2-D exist. The general behaviour and properties of these equations are discussed before the numerical methods are introduced in chapter 7.

Corresponding to chapter 3, chapter 7 introduces MFE methods in higher dimensions. In general, the MFE method has the same formal structure as in 1-D but a few important differences do occur. For example in 1-D the global and local implementation of the MFE method are identical since their basis functions span the same space, however in two and higher dimensions the spaces spanned by the basis functions are different and consequently the methods are not identical. There are also some problems which occur in higher dimensions due to the increased complexity of the discretised algebraic equations. For example there are more cases of singularity. As in 1-D, the L_2 norm in higher dimensions becomes invalid as the solution overturns which again leads to a need for the methods to be rewritten in a valid form.

Chapter 8 introduces the modifications needed to the MFE methods in order that they may be used to generate multivalued solutions. The concluding sections in chapter 8 discuss a method for obtaining the shock position in 2-D from an overturned solution. The technique introduced is 1-D in nature, based upon the ideas described in chapter 4, but is capable of giving a realistic 2-D shock in many cases.

Numerical results and examples are given for 2-D problems in chapter 9 to illustrate the points made in chapters 6-8. Examples are given of both overturned solutions and of the shock position recovered from the multivalued solutions using the 1-D technique described in chapter 8.

Chapter 10 contains a summary of the work given in this thesis. Ideas on how this work may be extended are also discussed.

Chapter 2

Analytical Properties Of Conservation Laws In 1-D

2.1 Introduction

The types of problem considered here arise from solving one-dimensional scalar, nonlinear, partial differential equations. This general heading encompasses an extremely large field and includes equations of many different types with widely varying properties. In order to examine both analytic and numerical methods of solution it is necessary to reduce this range to a class of equations which show the type of properties that are of particular interest here. We shall concentrate on nonlinear equations which give rise to shocks or discontinuities since the calculation of shocks or discontinuities occur in many physical simulations and pose a challenge both analytically and numerically. Taking this into consideration, a class of equations which are both widely used and exhibit the above properties is the conservation laws.

In 1-D these are equations of the form

$$u_t + f_x = 0 \tag{2.1}$$

where f is a function of u, and u is a function of x and t. Equation (2.1) is given on a 1-D region R with initial conditions specified and boundary conditions given where necessary.

Initially we shall concentrate on the method of characteristics, which has

been the most effective tool for analytic solutions of these equations in 1-D (Courant & Hilbert (1962)). The possibility of the occurrence of shocks then leads to the introduction of jump conditions and weak solutions. However the use of weak solutions means that the solution to the problem is not uniquely defined, hence another (entropy) condition is imposed for uniqueness (Smoller (1983)). Several examples are given to demonstrate the formation of shocks and expansions, using both the entropy and the jump conditions.

Allowing the introduction of multivalued solutions calculated by the method of characteristics leads to a discussion of how the physically based discontinuity or shock is to be recovered. One of the simplest methods applied to this problem is the calculation of the shock position by application of the principle of conservation (Courant & Hilbert (1962)). This type of method may be looked at in several forms which are equivalent and based upon this physical condition. A second type of method considered is the use of the so-called Transport Collapse Operator given by (Brenier (1984)). This involves replacing the multivalued curve by a single-valued solution using an averaging technique. The solution obtained is entropy satisfying.

2.2 Basics Of The Problem

We shall start by concentrating on equations of the form

$$u_t + f_x = 0 \tag{2.2}$$

where $f = f(u)$, $u = u(x,t)$. For a well posed problem equation (2.2) is given on a 1-D region R with initial conditions $u_0 = u(x,0)$ and boundary conditions given where data enters the region along characteristics. The equation is known as a conservation law, since it comes from the property that some quantity u is conserved, as follows.

Suppose there is some material of density u distributed along a line, then the rate of change of the 'mass' $\int u dx$ in a fixed interval is balanced by the flux $f(u)$ through the boundaries of the interval. Then the conservation of mass within the

interval $[x, x+\Delta x]$ is expressed by the equation

$$\frac{d}{dt}\int_x^{x+\Delta x} u dx + [f(u(x+\Delta x)) - f(u(x))] = 0. \tag{2.3}$$

i.e.

$$\int_x^{x+\Delta x} (\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x}) dx = 0. \tag{2.4}$$

By the mean value theorem for integrals (2.4) becomes

$$0 = \Delta x \left(\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x}\right)\Bigg|_{x=\theta} \quad \text{for some} \quad \theta \in (x, x+\Delta x). \tag{2.5}$$

Then, since (2.5) holds for arbitrary small Δx, this means that it is equivalent to the differential equation (2.2). (This argument is known as Lagrange's Lemma.)

Equation (2.2) may also be written in the form

$$u_t + a(u)u_x = 0, \tag{2.6}$$

where $a(u) = f'(u)$, and a is known as the wave speed. The behaviour of solutions of these equations is wavelike (see e.g. Whitham (1974)). The presence of the nonlinearity gives rise to solutions which blow up in finite time. However the methods investigated here will permit multivalued solutions to form. Subsequently, techniques will be investigated for obtaining a physically based single valued solution from the multivalued curve.

2.3 Characteristics

Before considering numerical methods for the solution of conservation laws (see chapters 3 and 4), we shall discuss the analytic techniques available. This will enable us to see the form of the solution and to examine some of the problems which may occur. First consider the total time derivative of $u = u(x,t)$ when x is allowed to depend on t,

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + \frac{dx}{dt}\frac{\partial u}{\partial x}. \tag{2.7}$$

Equation (2.6) may be substituted in (2.7) to give

$$\frac{du}{dt} = \left\{\frac{dx}{dt} - a(u)\right\}\frac{\partial u}{\partial x}. \tag{2.8}$$

From (2.8) it can be seen that, if

$$\frac{dx}{dt} = a(u), \quad \text{then} \quad \frac{du}{dt} = 0. \tag{2.9}$$

Equations (2.9) are known as the equations of characteristics of (2.6). The characteristics of (2.6) may also be obtained by writing (2.2) in a Lagrangian moving framework. A coordinate transform is defined (assumed non-singular) between x, t and new independent variables ξ, τ by

$$x = \hat{x}(\xi, \tau), \quad t = \tau, \quad u(x,t) = \hat{u}(\xi, \tau). \tag{2.10}$$

Using the new variables (2.2) may now be written in the Lagrangian form as

$$\frac{\partial \hat{u}}{\partial \tau} - \frac{\partial u}{\partial x}\frac{\partial \hat{x}}{\partial \tau} + f(u) = 0. \tag{2.11}$$

Hence using the notation

$$\dot{u} = \frac{\partial \hat{u}}{\partial \tau}, \quad \dot{x} = \frac{\partial \hat{x}}{\partial \tau}, \quad u_x = \frac{\partial u}{\partial x} \tag{2.12}$$

equation (2.11) becomes

$$\dot{u} - u_x \dot{x} + f(u) = 0. \tag{2.13}$$

The function $f(u)$ is now replaced by $a(u)u_x$ from (2.6) to give

$$\dot{u} - u_x \dot{x} + a(u)u_x = 0. \tag{2.14}$$

If we now compare the coefficients of u_x we obtain the equations

$$\dot{u} = 0 \quad \text{and} \quad \dot{x} = a(u) \tag{2.15}$$

(c.f. (2.9)) which are also the equations of characteristics of (2.6). (It should be noted that the Lagrangian method and the method of characteristics are only the same for certain special cases, which include the conservation laws. If we consider the equation $u_t + H(x, u, u_x) = 0$ which becomes $\dot{u} - u_x\dot{x} + H = 0$, the Lagrangian method has $\dot{u} = 0$ giving $\dot{x} = \frac{H}{u_x}$. For the method of characteristics we obtain the equations $\dot{x} = \frac{\partial H}{\partial u_x}$ and $\dot{u} = -H + u_x \frac{\partial H}{\partial u_x}$ (see (1.7)) which are obviously different.)

The equations of characteristics may be partially solved to give

$$u = u_0 \ \text{(constant)} \quad \text{on} \ \frac{dx}{dt} = a(u) \tag{2.16}$$

as in (2.9). Since u is constant along the curve, the second equation of (2.16) may be integrated to give the straight lines

$$x = a(u)t + x_0, \tag{2.17}$$

where x_0 is a constant (the intersection point on the $t = 0$ axis, see Fig. 2.1), and the first equation of (2.16) gives $u = u_0(x_0)$, where u_0 is the initial data function. The general solution is therefore

$$u(x,t) = u_0(x - a(u)t) \tag{2.18}$$

which gives u implicitly. Note: One of the reasons that numerically calculated solutions are required for this type of problem is that often the solution may only be obtained implicitly.

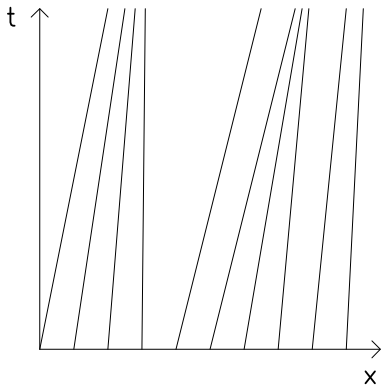

Figure 2.1: Characteristics of a nonlinear PDE.

It is interesting to note that 'blow-up' occurs in the solution to (2.14). Differentiating (2.18) with respect to x gives

$$u_x = u_0'(x - a(u)t)(1 - a'(u)u_x t) \tag{2.19}$$

$$\Rightarrow u_x = \frac{u_0'(x - a(u)t)}{(1 + a'(u)u_0' t)}. \tag{2.20}$$

As $t \to \frac{-1}{a'(u)u_0'}$, $u_x \to \infty$, which shows that u_x blows-up in finite time if $\frac{-1}{a'(u)u_0'}$ is positive. For further discussion see section 2.5.

2.3.1 Overturning Solutions

Fig. 2.1 shows a family of characteristics, crossing $t = 0$ at x_0. The characteristics have slope $1/a(u_0)$, different for each characteristic, so that for a general nonlinear

conservation law with arbitrary initial data, the lines may intersect in finite time (see Fig. 2.2).

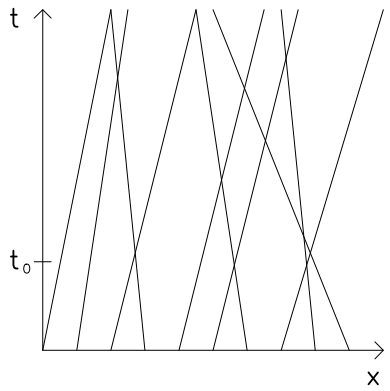

Figure 2.2: Characteristics intersecting after finite time, $t = t_0$.

If the solution to (2.6) is allowed to overturn (i.e. the characteristics are followed beyond the blow-up time) a smooth multivalued solution may be obtained. See Fig. 2.3. Suppose the initial data is a monotonic decreasing function $u_0 = u(x_0)$. Then since (2.2) is a nonlinear advection equation the point at which $u = u_0$ is moved spatially by a distance $a(u_0)t$. If $a(u)$ is strictly increasing, then for large u_0 points on the curve move further than for small u_0. This gives a multivalued solution after $t > t_I$ where t_I is the time the first of intersection of the characteristics.

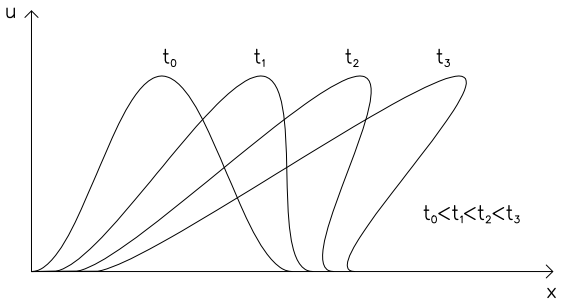

Figure 2.3: Solution overturning as time increases.

Conservation laws are used to model physical processes where, although a smooth overturning solution can be found, it will be multivalued and this does not occur physically (nor does it satisfy (2.2), incidentally). If the solution is required for a time t_n, where $t_n > t_I$ then the conservation law breaks down as

a description of the physical process but a discontinuous solution may be found, using the conservation principle to give a jump condition at a discontinuity. We describe this in the context of an example.

2.3.2 Example 1 - Part 1

The first example uses the inviscid Burgers' equation as the conservation law. The initial data is given by a ramp, which steepens to form a shock. We solve

$$u_t + \left(\frac{u^2}{2}\right)_x = 0 \qquad x \in \mathbb{R} \tag{2.21}$$

with initial data

$$u_0 = u(x,0) = \begin{cases} 1 & x < 0 \\ 1 - x & 0 \le x \le 1 \\ 0 & 1 < x \end{cases} \tag{2.22}$$

shown in Fig. 2.4.

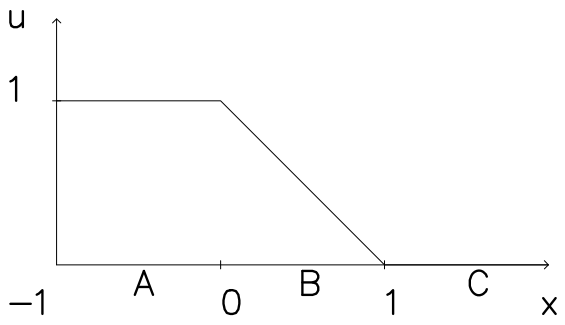

Figure 2.4: Initial data, $u_0 = u(x,0)$.

The characteristics of (2.21) are given by

$$\begin{aligned} \frac{dx}{dt} &= a(u) = u(x_0,0) \\ \Rightarrow x &= u(x_0,0)t + x_0 \qquad (2.23) \\ \Rightarrow x &= u_0(x_0)t + x_0 \qquad (2.24) \end{aligned}$$

where x_0 is the point that the characteristic cross the x-axis. This leads to three distinct regions of characteristics (see Fig. 2.5). In region A, the characteristics all have slope 1, while in region C , the characteristics all have infinite slope. In

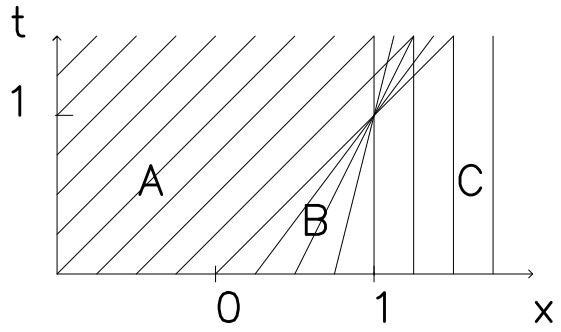

Figure 2.5: Regions of characteristics.

the centre region B, $0 \leq x \leq 1$ hence the characteristics are not parallel and have slope $\frac{1}{1-x_0}$ which is dependent upon x_0.

From Fig. 2.5, it can be seen that the characteristics will first cross at $x = 1$ $t = 1$ and it is at this point the shock is formed. In general, it will not be as easy to find the shock position (see below).

The solution up to the point $t_I = 1$, where the characteristics cross, may be found by tracing back along the characteristics. In region A, and region C, the solutions are $u = 1$ and $u = 0$ respectively, however in region B, more calculation is needed, viz

$$\begin{aligned} u(x,t) &= u(x_0,0) \\ &= 1 - x_0. \end{aligned} \tag{2.25}$$

But $x = u(x_0,0)t + x_0$ from (2.23), hence

$$\begin{aligned} x &= (1-x_0)t + x_0 \\ \Rightarrow x_0 &= \frac{x-t}{1-t}. \end{aligned}$$

Now returning to (2.25), x_0 can be substituted in (2.25) to give the solution

$$u(x,t) = \frac{1-x}{1-t}. \tag{2.26}$$

Therefore until $t = 1$, the solution is

$$u = \begin{cases} 1 & x < 0 \\ \dfrac{1-x}{1-t} & 0 \leq x \leq 1 \\ 0 & 1 < x \end{cases} . \tag{2.27}$$

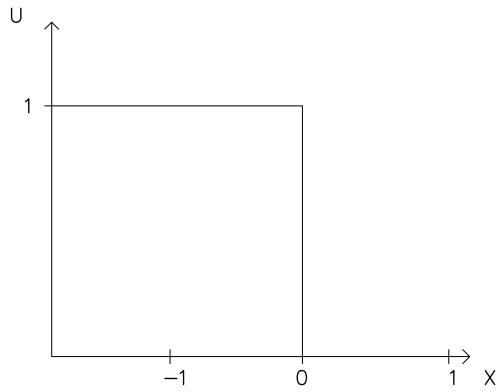

Figure 2.6: Solution at t=1.

See Fig. 2.6. This example shows how the solution until the characteristics cross can be found by tracing back along the characteristics. If we ignore the fact that the characteristics have crossed and continue to follow them after $t_I = 1$ then the curve overturns and we do not have a physical solution. Note that the characteristic equations still have a solution, however.

2.3.3 Weak Solutions

To consider discontinuous solutions of (2.2) we need to generalise (2.2) and another concept is needed. If (2.2) is multiplied by a test function $\phi(x,t)$ of compact support, where $\phi \in C^1$ the space of once-differentiable functions, then after integration by parts over the region $\mathbb{R}$, this gives

$$\int\int_{\mathbb{R}\times[0,\mathrm{T}]}(u\phi_t + f(u)\phi_x)dxdt + \int_{\mathbb{R},\mathrm{t}=0} u_0\phi(x,0)dx = 0 \qquad (2.28)$$

where T is the final time. This is known as the weak form of (2.2) (Smoller (1983)). This equation allows discontinuous solutions u to be admitted since, from (2.28) it can be seen that derivatives of u are no longer present. It can be shown that if (2.28) holds $\forall\ \phi \in C^1$, and u_0 is bounded and measurable then u is the classical solution of the differential equation (Smoller (1983)). In (2.28), u is known as a weak solution of (2.2).

We will now describe how, when characteristics meet, (2.2) can be replaced by a jump condition which allows discontinuous solutions to be found. The argument follows (Smoller (1983)). Suppose that Γ is a smooth curve in (x,t) space and that u has a discontinuity across Γ, u is smooth away from Γ and has well defined

limits on both sides of Γ. Let P be a point on Γ and D be a small ball centred at P. Let $x = x(t)$ be the equation of the curve Γ in D and let D_1, D_2 be parts of D split by Γ. Let Q_1, Q_2 be the points which lie on both Γ and D. See Fig. 2.7 below.

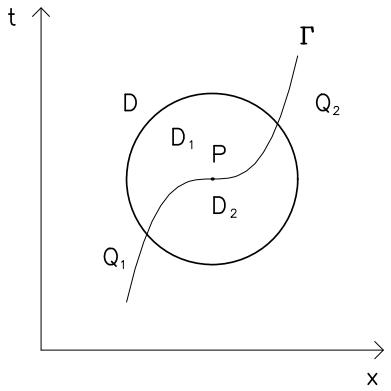

Figure 2.7: The test region D.

Let ϕ be a once differentiable function with compact support in D and let $\phi = 0$ on the boundary ∂D. Multiplying (2.2) by ϕ and integrating by parts over D gives

$$0 = \int\int_D (u\phi_t + f(u)\phi_x)dxdt, \tag{2.29}$$

since $\phi = 0$ on ∂D. Equation (2.29) can be written as the sum of integrals over the two regions D_1, D_2. Now use the divergence theorem on each region D_1, D_2 to give

$$\begin{aligned} \int\int_{D_i} (u\phi_t + f(u)\phi_x)dxdt &= \int\int_{D_i} (u\phi)_t + (f\phi)_x dxdt \\ &= \int_{\partial D_i} \phi(-udx + fdt), \end{aligned} \tag{2.30}$$

where $i = 1, 2$.

Since $\phi = 0$ on the boundary, these line integrals are non-zero only along Γ. Let $u_L = u(x(t) - 0, t)$ and $u_R = u(x(t) + 0, t)$ be the values of u on Γ from each side, hence (2.30) becomes

$$\int_{\partial D_1} \phi(-udx + fdt) = \int_{Q_1}^{Q_2} \phi(-u_L dx + f(u_L)dt) \tag{2.31}$$

$$\int_{\partial D_2} \phi(-udx + fdt) = -\int_{Q_1}^{Q_2} \phi(-u_R dx + f(u_R)dt). \tag{2.32}$$

From (2.29), (2.31) and (2.32)

$$\begin{aligned} 0 &= \int\int_D (u\phi_t + f(u)\phi_x)dxdt \\ \Rightarrow 0 &= \int\int_{D_1} (u\phi_t + f(u)\phi_x)dxdt + \int\int_{D_2} (u\phi_t + f(u)\phi_x)dxdt \\ \Rightarrow 0 &= \int_\Gamma \phi(-[u]dx + [f(u)]dt) \end{aligned} \tag{2.33}$$

where $[u]$ denotes the jump $[u] = u_L - u_R$ and $[f(u)] = f(u_L) - f(u_R)$. Since ϕ is arbitrary, then

$$\frac{dx}{dt}[u] = [f(u)], \tag{2.34}$$

which is known as the jump condition. The quantity $s = \frac{dx}{dt}$ is known as the shock speed (i.e. the speed of the discontinuity).

2.3.4 Example 1 - Part 2

Using the initial data and equation given in (2.21), (2.22) the solution has been calculated to be

$$u = \begin{cases} 0 & x < 0 \\ \dfrac{1-x}{1-t} & 0 \le x \le 1 \\ 1 & 1 < x \end{cases} \tag{2.35}$$

before the crossing of the characteristics at $t_I = 1$. For a discontinuous solution, after t_I the jump condition may be used to find out the speed s of the shock,

$$\begin{aligned} s = \frac{[f]}{[u]} = & \frac{\frac{u_L^2}{2} - \frac{u_R^2}{2}}{u_L - u_R} \\ s = & \tfrac{1}{2}(u_L + u_R) \\ s = & \tfrac{1}{2}(1+0) = \tfrac{1}{2}. \end{aligned}$$

After t_I, therefore the solution becomes

$$u(x,t) = \begin{cases} 1 & x < 1 + \frac{1}{2}(t-1) \\ 0 & x > 1 + \frac{1}{2}(t-1) \end{cases} \tag{2.36}$$

which is a shock moving with speed $\frac{1}{2}$. The characteristics now move into the shock. See Fig. 2.8. This now gives a weak solution for equation (2.2).

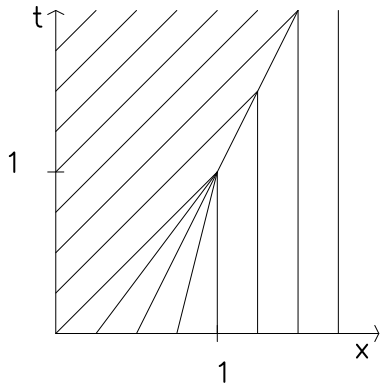

Figure 2.8: Shock position and characteristics as t increases.

Although it is possible to consider discontinuous solutions of conservation laws, these have only been weak solutions. Discontinuous functions are not differentiable hence they cannot strictly satisfy the partial differential equation. The differential equation is only satisfied in the sense of distributions (2.28).

The problem with permitting weak solutions of the PDE is that the uniqueness of the solution is lost and and a so-called entropy condition is required to pick out the physical solution (Smoller (1983)). This is most easily shown by considering another example.

2.3.5 Example 2 - Part 1

Consider the equation

$$u_t + \left(\frac{u^2}{2}\right)_x = 0 \tag{2.37}$$

with initial data

$$u_0(x,0) = \begin{cases} -1 & x < 0 \\ 1 & x > 0 \end{cases}. \tag{2.38}$$

The characteristics of this region are shown in Fig. 2.9. There are several possible solutions two of which are given below.

$$u(x,t) = u_0(x) \tag{2.39}$$

is a solution, with characteristics shown in Fig. 2.10.

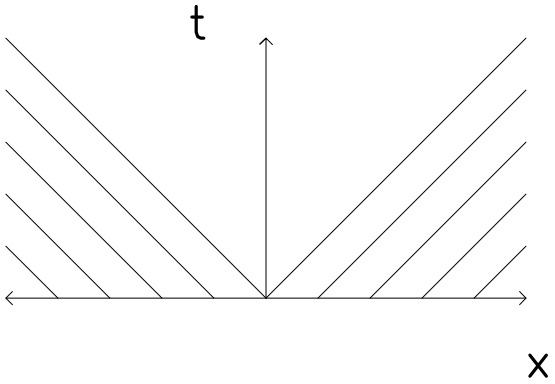

Figure 2.9: Characteristics of the initial data.

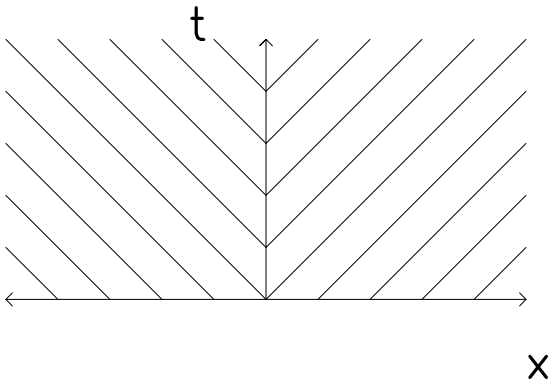

Figure 2.10: Characteristics of solution.

Similarly

$$u(x,t) = \begin{cases} -1 & x < -t \\ \dfrac{x}{t} & -t < x < t \\ 1 & t < x \end{cases} \tag{2.40}$$

is a solution of the equation (2.37), which may be checked by substituting $u(x,t)$ into the equation. This solution has characteristics shown in Fig. 2.11.

This now gives us two different solutions which both satisfy the equation (2.37). It is now clear that for a unique solution another condition must be applied.

2.3.6 Entropy Condition - 1

One way of choosing the correct physical solution may be obtained by considering the viscous form of equation (2.2),

$$u_t + f(u)_x = \epsilon u_{xx} \tag{2.41}$$

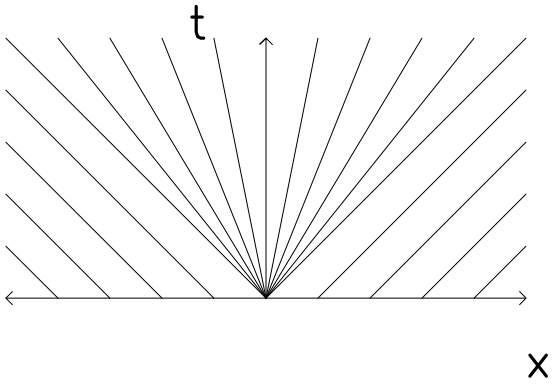

Figure 2.11: Characteristics of solution.

where $\epsilon > 0$, in the limit as $\epsilon \to 0$. This equation is used because it has a continuous solution with a very steep front, steepening as $\epsilon \to 0$ to approach the discontinuous solution of (2.2). The idea that the viscous equation may be used to derive a condition to impose on the non-viscous equation, so that in the limit (as $\epsilon \to 0$) the two solutions would be identical, is due to (Oleinik (1957)) and is known as the entropy condition.

This condition may be expressed in many forms, one of the simplest being given by the inequality (Lax (1972)),

$$f^{'}(u_L) > s > f^{'}(u_R) \tag{2.42}$$

where s is the jump speed, $f^{''} > 0$ and u_L, u_R are the values of u just to the left and right of the shock position. This condition forces the characteristics to go into the shock as t increases hence giving the correct type of solution. Another less restrictive form of the entropy condition which holds $\forall f$ was given in 1957 by Oleinik as

$$\frac{f(u_L) - f(u)}{u_L - u} \geq s \geq \frac{f(u_R) - f(u)}{u_R - u} \tag{2.43}$$

$\forall u \in \{u : u_L < u < u_R\}$.

2.3.7 Entropy Condition - 2

An alternative entropy condition (Lax (1972)) may also be given by considering a convex function $V(u)$ and associating it with a function $F(u)$ defined by

$$F^{'} = V^{'} f^{'}. \tag{2.44}$$

V is known as the entropy function and F is its associated entropy flux. Now return to the viscous equation (2.41), where $\epsilon > 0$, and multiply (2.41) by $V^{'}$ to give

$$\begin{aligned} V^{'}(u)u_t + V^{'}f(u)_x &= \epsilon V^{'}u_{xx} && (2.45)\\ \Rightarrow V^{'}(u)u_t + V^{'}f^{'}(u)u_x &= \epsilon V^{'}u_{xx} && (2.46)\\ \Rightarrow V(u)_t + F^{'}(u)u_x &= \epsilon((V^{'}u_x)_x - V^{''}u_x^2) && (2.47)\\ \Rightarrow V_t + F_x &= \epsilon(V_{xx} - V^{''}u_x^2) && (2.48)\\ \Rightarrow V_t + F_x &\leq \epsilon V_{xx} && (2.49) \end{aligned}$$

since V is convex $(V^{''} > 0)$. Now let $\epsilon \to 0$, then

$$V_t + F_x \leq 0 \tag{2.50}$$

which holds in the weak sense. Note: For smooth solutions equality holds whereas for discontinuous solutions, strict inequality holds. If f is convex, and if (2.50) is satisfied by any convex V, then (2.50) will be satisfied for all V. If f is not convex, (2.50) must be checked for all convex V, (see Kružkov (1970)).

As we shall see, the solution of equations of the form of (2.2) where shocks are formed may also be effected by following the solution along the characteristics and then using the jump condition to find the discontinuous solution. When expansions are formed the entropy condition is used to obtain a valid solution of the original equation.

2.3.8 Example 2 - Part 2

Returning to example 2 part 1 (section 2.3.5), we found that there was a non-uniqueness of solution. Let us now apply one of the simplest entropy conditions described above to determine if either of the solutions (2.39) or (2.40) are correct. If we let the value of u at the left of the shock to be -1, then from the problem the value of u at the right of the solution is 1. The speed of the shock can now be calculated to be

$$s = \frac{\frac{1}{2}(u_L^2 - u_R^2)}{u_L - u_R} = 0. \tag{2.51}$$

Now calculating the derivative of f to be u, so that $f^{'}(u_L) = -1$ and $f^{'}(u_R) = 1$, consequently the entropy condition (2.42) is not satisfied for the solution $u =$

u_0. It should also be noted that this solution cannot be acceptable as there are characteristics emerging from a line which carries no data.

2.3.9 Envelopes And The Shock Position

An alternative, more general, way of finding the shock position, by calculating where the characteristics first cross, is to obtain their envelope (Courant & Hilbert (1962)). Two neighbouring characteristics will cross at $t = t_I$ provided both x values are the same. If a small change in the initial position gives no change in the subsequent location of the characteristic then neighbouring characteristics must have crossed. Mathematically this can be expressed in terms of the solution of the equations of the characteristics, $x(x_0)$, when this equation occurs

$$\frac{dx}{dx_0} = 0 \tag{2.52}$$

for the first time. The example below uses this technique to find the shock position.

2.3.10 Example 3

In this example, the conservation law chosen is again the inviscid Burgers' equation (2.21) but this time the initial data is the smooth curve (see Fig. 2.12),

$$u_0 = \tanh(5 - 10x) \qquad 0 \leq x \leq 1. \tag{2.53}$$

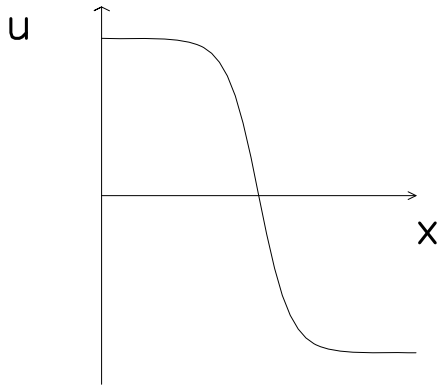

Figure 2.12: Initial data.

The characteristics are given as before by

$$\begin{aligned} \frac{dx}{dt} &= a(u) = u(x_0, 0) \\ &= \tanh(5 - 10x_0) \\ \Rightarrow x &= \tanh(5 - 10x_0)t + x_0. \end{aligned}$$

The characteristics will first cross at the instant t_I where $x(t_I, x_0^2) = x(t_I, x_0^1)$. Using the envelope construction this first occurs when

$$\frac{dx}{dx_0} = 0 \tag{2.54}$$

first holds. This implies that $t = t_I$ satisfies

$$\begin{aligned} 0 &= -10\text{sech}^2(5 - 10x_0)t + 1 \\ \Rightarrow t &= \frac{1}{10\text{sech}^2(5 - 10x_0)} = \frac{\cosh^2(5 - 10x_0)}{10}. \end{aligned} \tag{2.55}$$

To find the minimum t such that (2.55) holds $\cosh^2(5 - 10x_0)$ must be as small as possible. i.e. $\cosh(0) = 1$, hence $5 - 10x_0 = 1 \Rightarrow x_0 = \frac{1}{2}$ and $t = \frac{1}{10}$.

For $t \leq \frac{1}{10}$, the solution can be found by tracing back along the characteristics, i.e., since u remains constant and only x changes,

$$\begin{aligned} u(x, t) &= u(x_0, 0) \\ &= \tanh(5 - 10x_0) \end{aligned} \tag{2.56}$$

where

$$x = u(x_0, 0)t + x_0. \tag{2.57}$$

If x_0 can be found from the implicit equation (2.57) it can be substituted into (2.56) to give the solution for $t < \frac{1}{10}$. An explicit solution for $u(x, t)$ is not obtainable, but a parametric solution is given by (2.56) and (2.57).

Now for $t > \frac{1}{10}$ consider the jump condition to find the speed of the shock. This is given as before by

$$\begin{aligned} s &= \frac{u_L + u_R}{2} \\ &= \frac{1 + -1}{2} = 0 \end{aligned}$$

which means that the shock is stationary (the vertical line in Fig. 2.13).

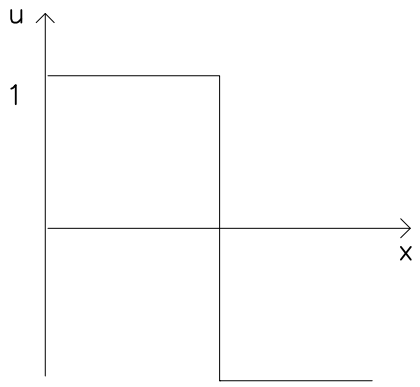

Figure 2.13: Shock position.

2.4 Reduction Of Conservation Laws To Inviscid Burgers' Equation

It is interesting to note that a general conservation law of the form (2.6) may be reduced to the Inviscid Burgers' Equation by a simple substitution. Let $c = a(u)$, then $c_t = a'(u)u_t$ and $c_x = a'(u)u_x$, which when substituted into (2.7) give

$$c_t + cc_x = 0 \tag{2.58}$$

provided that f in (2.2) is either convex or concave. i.e. $f'' = a'(u) \neq 0$. This equation is a considerable simplification of the original, but care must be taken when applying both the jump and entropy conditions over what quantity is being conserved since conservation of c does not imply conservation of u. The following example shows the problems that can arise if care is not taken in deciding which variable is to be conserved.

2.4.1 Example 4

If we consider Burgers' equation

$$u_t + uu_x = 0 \tag{2.59}$$

and multiply it by u to give

$$uu_t + u^2u_x = 0, \tag{2.60}$$

this is equivalent to the equation

$$\left(\frac{1}{2}u^2\right)_t + \left(\frac{1}{3}u^3\right)_x = 0. \tag{2.61}$$

Let $v = \frac{1}{2}u^2$, then this gives a conservation law in v as

$$v_t + \frac{\sqrt{8}}{3} v_x^{\frac{3}{2}} = 0. \tag{2.62}$$

The characteristics for the equations (2.59) and (2.62) are the same, consequently the shock forms at the same time. If we now calculate the shock speed for (2.59) it is

$$s_1 = \frac{\frac{u_L^2}{2} - \frac{u_R^2}{2}}{u_L - u_R} = \frac{u_L - u_R}{2} \tag{2.63}$$

whereas the shock speed of (2.62) is

$$s_2 = \frac{\frac{u_L^3}{3} - \frac{u_R^3}{3}}{\frac{u_L^2}{2} - \frac{u_R^2}{2}} = \frac{2}{3}\left(\frac{u_L^3 - u_R^3}{u_L^2 - u_R^2}\right). \tag{2.64}$$

It is clear that s_1 and s_2 are not equal, hence this demonstrates that care is needed in deciding which variable is being conserved. Expansions are not a problem when variables are changed.

2.5 Blow Up Of Solution

An alternative but equivalent method of solution is given as follows. Consider again the equation (2.6) with characteristic equations

$$\frac{du}{dt} = 0 \qquad \frac{dx}{dt} = a(u). \tag{2.65}$$

Now differentiate (2.6) with respect to x to give

$$u_{tx} + a'(u)u_x^2 + a(u)u_{xx} = 0. \tag{2.66}$$

In a frame of reference moving with speed $\dot{x}$ (c.f. section 2.3), this becomes

$$\dot{u}_x - u_{xx}\dot{x} + a'(u)u_x^2 + a(u)u_{xx} = 0 \tag{2.67}$$

where

$$\dot{u} - u_x\dot{x} + a(u)u_x = 0 \tag{2.68}$$

and if $\dot{x} = a(u)$

$$\dot{u}_x = -a'(u)u_x^2. \tag{2.69}$$

(If $a(u) = u$ this equation is of Hamilton-Jacobi type (Courant & Hilbert (1962)).)
Now let $q = a(u)_x$ (the c_x of section 2.4), giving

$$\dot{q} = -(a'(u)u_x)^2 \tag{2.70}$$

$$\Rightarrow \dot{q} = -q^2. \tag{2.71}$$

This ODE may be readily solved when $q = q_0$ say at $t = 0$ and gives

$$q = \frac{q_0}{q_0 t + 1}. \tag{2.72}$$

This shows that when $t = -\frac{1}{q_0}$ the solution 'blows up' and illustrates analytically the unboundedness in the derivative u_x of solutions to (2.6) within finite time. It is the analytic counterpart of the 'geometrical' characteristics solution breakdown discussed earlier.

The above examples all deal with the case where, once the characteristics have crossed, the shock position is found and a discontinuous solution allowed immediately. Another way of obtaining the solution using the characteristics is to allow them to cross and continue to follow their paths so that a multivalued function is obtained. From this multivalued curve, a single-valued discontinuous solution may then be calculated, as follows.

2.6 Calculation Of Shock Position Using Conservation

The calculation of the shock position from an overturned solution may be done in several ways. One method of replacing the multivalued solution by a discontinuous solution is by using the conservation property (Whitham (1974)). Both the multivalued curve and the discontinuous curve satisfy the conservation property (see Fig. 2.14) so that $\int u dx$ under each curve remains the same for all time.

This means that the overturned part of the curve in Fig. 2.14. may be replaced by the vertical line drawn below, in which area A = area B. This satisfies the conservation of $\int u dx$ and the discontinuity replaces the multivalued solution. The discontinuity only replaces those regions which have overturned leaving the single-valued regions unaffected. This will automatically satisfy the jump condition which is also based on the conservation property.

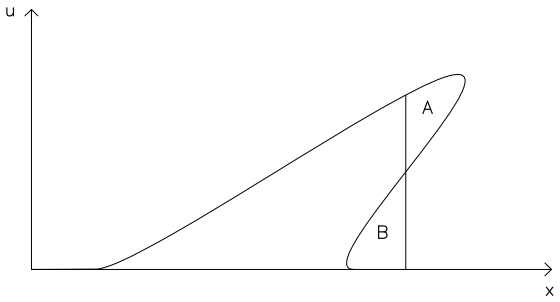

Figure 2.14: Equal area construction.

It is interesting to note that the conservation argument for finding the discontinuity may be written in several forms (Sewell (1987)). This can involve the transformation of the problem into a different set of variables.

2.7 Multivalued Solutions And Transformations

The solution of a nonlinear PDE evolving with time will translate the initial data (single-valued) to a multivalued solution once the 'shock formation' time has been reached. Here we will examine a conservation law and see how the multivalued solution forms, then a variety of transformations will be applied to the conservation law in order to investigate the behaviour of the solutions of the transformed equations at the time the shock is formed.

2.7.1 Catastrophe Form

First consider the PDE

$$u_t + uu_x = 0 \tag{2.73}$$

whose solution may be obtained from the implicit equation

$$u = u_0(x - ut) \tag{2.74}$$

where u_0 is a piecewise differentiable function in x (the initial data). Suppose for example that u_0 is given by

$$u_0 = u_0(x) \quad \text{where} \quad x = -\frac{1}{3}u_0^3. \tag{2.75}$$

This gives the solution of (2.73) as

$$-\frac{1}{3}u^3 = x - ut, \tag{2.76}$$

as the classic 'bifurcation set' Sewell (1987). The figure below shows how the solution changes as time increases, from a single-valued curve to a multivalued curve. In Fig. 2.15, consider the initial data to be the cubic (2.75), then apply (2.73) to see how the solution (2.76) moves with time.

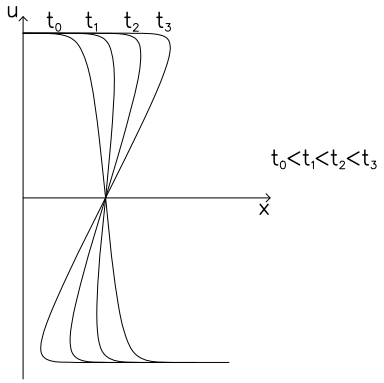

Figure 2.15: Solution becomes multivalued as time increases.

2.7.2 Conservation Laws And Hamilton-Jacobi Equations

Conservation laws and Hamilton-Jacobi equations are related via a simple transformation (Courant & Hilbert (1962)). This transformation is used in order to obtain an insight into the 'overturning' or shock formation of the conservation laws. We then consider the behaviour of the solution to the Hamilton-Jacobi equation, using the initial data described above, in order to see what is the corresponding transformation of the overturned manifold. If (2.73) is integrated with respect to x it gives a Hamilton-Jacobi type equation for the function $a = \int u dx$. For a general conservation law, i.e. $u_t + f(u)_x = 0$, a family of Hamilton-Jacobi equations may be obtained by this transformation.

Consider the conservation law

$$u_t + f_x = 0 \tag{2.77}$$

on a region R with $u_0 = u(x, 0)$ and boundary conditions given where appropriate.

Integrating with respect to x gives

$$\begin{aligned} \int_{x_0}^{x} (u_t + \frac{\partial f}{\partial x}(u))dx &= 0 \\ \Rightarrow \frac{d}{dt}\int_{x_0}^{x} u dx + \int_{x_0}^{x} \frac{\partial f}{\partial x}(u)dx &= 0. \end{aligned} \tag{2.78}$$

Now let $a = \int_{x_0}^{x} u dx$, then (2.78) becomes

$$\begin{aligned} \frac{\partial a}{\partial t} + (f(u(x)) - f(u(x_0))) &= 0 \\ \Rightarrow \frac{\partial a}{\partial t} + f(a_x) &= \text{constant}, \end{aligned} \tag{2.79}$$

where the constant is dependent on the boundary conditions of (2.77). The constant may be included in a new function H to give a form of Hamilton-Jacobi equation

$$\frac{\partial a}{\partial t} + H(a_x) = 0. \tag{2.80}$$

Consider now a specific case of the cubic initial data (2.75) applied to a conservation law which becomes multivalued as time increases. Once the solution becomes multivalued, under the transformation above, the resulting curve is a swallow-tail. See Fig. 2.16.

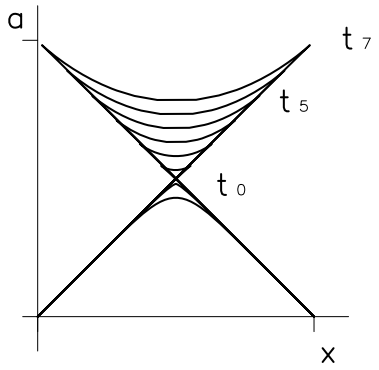

Figure 2.16: Swallow-tail forming as time increases.

2.7.3 Conservation Laws And ODE's

If a slightly different transformation is applied to the conservation law with cubic initial data, a quartic curve results. The transformation involves the integration with respect to the other variable, u, to give a pair of ODE's. This means that there is also a relationship between the conservation laws and a pair of ODE's.

Consider the conservation law

$$\begin{aligned} u_t + f_x &= 0 \qquad (2.81) \\ \Rightarrow u_t + f^{'}(u)u_x &= 0. \qquad (2.82) \end{aligned}$$

Now the characteristics of (2.82) are given by

$$\dot{u} = 0 \qquad (2.83)$$

and

$$\dot{x} = f^{'}(u) \qquad (2.84)$$

where the dot means differentiation in the sense of (2.12). Consider (2.84) and integrate with respect to u to give

$$\begin{aligned} \int \dot{x} du &= \int f^{'}(u) du \\ \Rightarrow \int \dot{x} du &= f(u). \qquad (2.85) \end{aligned}$$

Let

$$\begin{aligned} b &= \int x du \\ \Rightarrow \dot{b} &= \frac{\partial}{\partial \tau} \int x du. \qquad (2.86) \end{aligned}$$

Comparing (2.85) and (2.86) and using (2.83) gives

$$\dot{b} = \frac{\partial}{\partial \tau} \int x du = \int \dot{x} du = f(u) \qquad (2.87)$$

which gives the ODE system of (2.83) and (2.87)

$$\dot{u} = 0 \qquad (2.88)$$

and

$$\dot{b} = f(u). \qquad (2.89)$$

If the initial data is cubic in u, x variables (2.75) and moves with time under (2.73), the solution becomes multivalued. In the case of the variables u, b, the cubic becomes a quartic and changes shape (Sewell (1987)). See Fig. 2.17.

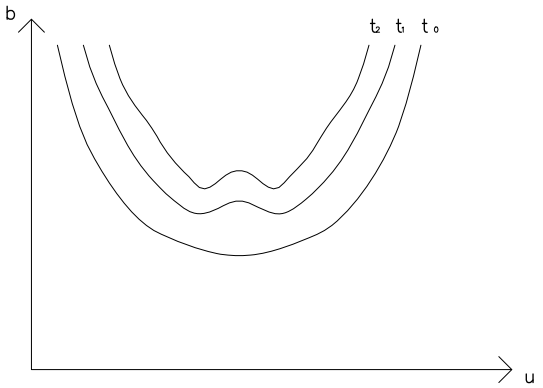

Figure 2.17: Canonical quartic moving with time.

2.7.4 Legendre Transform

There is a way of connecting the conservation law, the Hamilton-Jacobi Equation and the system of ODE's using a Legendre Transform. We show that the Hamilton-Jacobi equation

$$a_t + H(a_x) = 0 \tag{2.90}$$

can be transformed to the ODE system (2.88), (2.89) under the Legendre transformation (Courant & Hilbert (1962)), defined implicitly by

$$\begin{aligned} a(x) + b(u) - ux &= 0 \\ u = a_x \qquad x &= b_u. \end{aligned} \tag{2.91}$$

Using (2.90), (2.91) the Hamilton-Jacobi equation may be written as

$$a_t + H(u) = 0. \tag{2.92}$$

However

$$\begin{aligned} a_t &= \dot{a} - a_x \dot{x} \\ &= \dot{a} - u\dot{x}, \end{aligned} \tag{2.93}$$

then differentiating (2.91) with respect to t,

$$\begin{aligned} \dot{a} + \dot{b} - \dot{u}x - u\dot{x} &= 0 \\ \Rightarrow \dot{a} - u\dot{x} + \dot{b} - \dot{u}x &= 0 \\ \Rightarrow a_t + \dot{b} - \dot{u}x &= 0. \end{aligned} \tag{2.94}$$

Now substitute for a_t from (2.94) into (2.92) to give

$$-(\dot{b} - \dot{u}x) + H(u) = 0 \tag{2.95}$$

and b_t may be written in a similar way to (2.93) as $b_t = \dot{b} + \dot{u}x$, hence (2.95) becomes

$$-b_t + H(u) = 0 \tag{2.96}$$

which is an ODE since it no longer contains any x derivatives. It may be rewritten as $\dot{b} = f(u)$ and $\dot{u} = 0$ (c.f. (2.88), (2.89)). It is possible to see the relationship between the variables in a diagram. See Fig. 2.18.

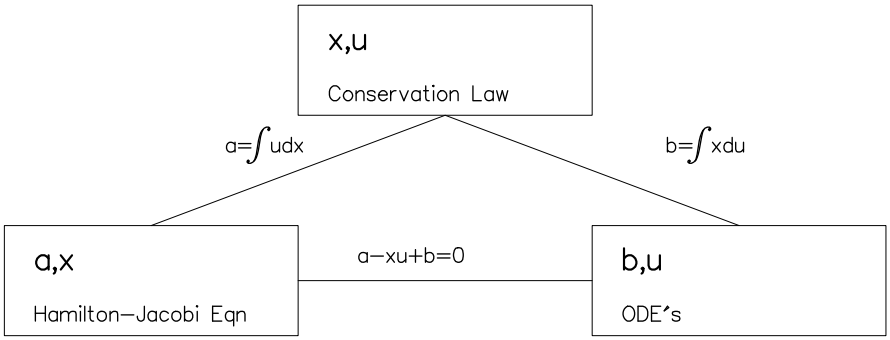

Figure 2.18: Illustration of relationship between the variables a, x, b, u.

From these transformations it can be seen that the solution of conservation laws may be carried in an equivalent way using different types of equations: equally once the solution to a conservation law has been found, we immediately have solutions to other types of equations.

2.8 A Second Legendre Transformation

Another Legendre transformation, completely separate from the one above, may be given between the variables u, x and the new variables v, m where

$$u(x) + v(m) - mx = 0 \tag{2.97}$$

$$m = u_x \qquad x = v_m. \tag{2.98}$$

The aim of using this transformation is to carry the equation to be solved into a more significant form when numerical versions of the method are given in chapters 3 and 7.

Consider the equation

$$u_t + f_x = 0 \tag{2.99}$$

which can be written as

$$\dot{u} - u_x \dot{x} + f(u)_x = 0. \tag{2.100}$$

Using (2.98) this can be transformed to

$$-\dot{v} + v_m \dot{m} + a(m v_m - v)m = 0. \tag{2.101}$$

In the case of Burgers' equation, where $a = u = mx - v$, comparing the coefficients of v_m gives

$$\dot{m} = -m^2 \tag{2.102}$$

so that

$$\dot{v} = -mv. \tag{2.103}$$

These equations can be solved exactly where m_0 and v_0 are the initial conditions, to give

$$m = \frac{1}{t + \frac{1}{m_0}} \qquad v = \frac{\frac{v_0}{m_0}}{t + \frac{1}{m_0}}. \tag{2.104}$$

It should be noted that the equations arising from this method are the same as those which arise in section 2.5 (with $m_0 = q_0$). For other equations, a projection onto a piecewise linear space is required before identifying the variables. This is described in more detail in chapter 3, section 3.14.

2.9 Calculation Of Shock Position From Multivalued Solutions

From the diagrams above (see Figs. 2.15, 2.16, 2.17) it can be seen that the problem of obtaining a discontinuous solution from the multivalued curve may be reconsidered in light of the above transformations. Consider Fig. 2.14, a

diagram of an overturned solution where the shock position has been calculated using the principle of conservation (Smoller (1983)), (Whitham (1974)). The

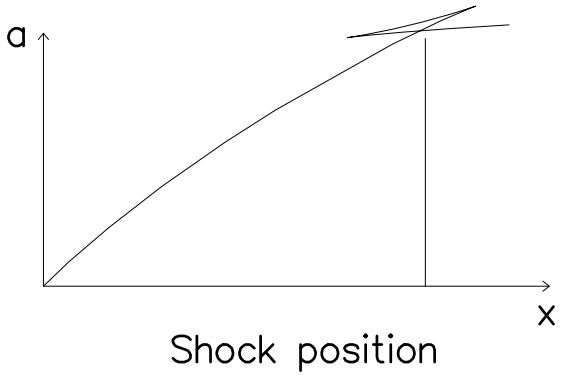

Figure 2.19: Swallow tail with shock position.

shock position calculated using the equal area argument is equivalent to the shock position calculated from the swallowtail curve using deletion (see Fig. 2.19). This occurs since the swallowtail obtained from the integration of the initial curve represents the area and hence when two areas are equivalent the curve intersects itself. It is obvious (because of the equal area argument) that the shock position found using the conservation idea is equivalent to the removal of the swallow tail, i.e. the shock position is marked in Fig. 2.19.

In a similar way the transformation to b, u variables gives a different way of calculating the shock position although it is again based on the principle of conservation. The curves shown below in Fig. 2.20 demonstrate how the curve appears before and after the shock occurring in the u, x variables has been introduced. The calculation of the shock position may be found by removing the region A, and replacing it by a straight line as shown (the convex hull). The shock position in the original variables may now be calculated using the inverse transform.

2.10 The Transport Collapse Operator Of Brenier

The aim of the method proposed by (Brenier (1984)) is to obtain a single valued approximation which satisfies the entropy condition and conservation to the mul-

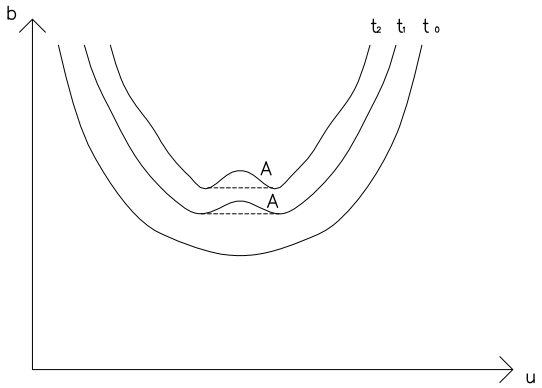

Figure 2.20: Shock position in quartic.

tivalued curve given by the classical method of characteristics (section 2.3). This method is applicable to time discretised scalar conservation laws. Although the theory of this method proposed by Brenier is given for higher dimensions in this chapter we are only concerned with one dimension.

The conservation law is given by the equation

$$u_t + f(u)_x = 0 \tag{2.105}$$

where $u = u(x,t) \in \mathbb{R}$ and $x \in \mathbb{R}$, $t > 0$. The entropy inequality given by (Lax (1972)), (2.50) can be written as follows. For each convex function V, the differential inequality

$$(V(u))_t + F(u)_x \leq 0 \tag{2.106}$$

is satisfied in the sense of distributions with

$$F(u) = \int_0^u f^{'}(w)V^{'}(w)dw \tag{2.107}$$

where $f^{'}$, $V^{'}$ denote the derivatives of f and V. (See section 2.3.7 on 'Entropy Condition - 2' for more information about $f, V,$ and F.) The introduction of (2.106) and (2.107) allow a global existence and uniqueness theorem to be proved for the initial value problem given by (2.105) (see Brenier (1984)). The following discussion leads to the introduction of the Transport Collapse operator (TC) which is an operator $T(t)$ which combines an averaging technique (which satisfies the entropy condition) with the method of characteristics.

First let us introduce the notation required for the construction of the TC operator. Equation (2.105) may be rewritten as

$$u_t + f^{'}(u)u_x = 0 \tag{2.108}$$

where $u = u(x,t) \in \mathbb{R}$, $x \in \mathbb{R}$ and $t > 0$. The characteristics are given by the ordinary differential system

$$\frac{du}{dt} = 0, \qquad \frac{dx}{dt} = f^{'}(u). \tag{2.109}$$

Let $(x,w) \in \mathbb{R} \times \mathbb{R}$, where

$$(X^t(x,w), U^t(x,w)) = F^t(x,w) \tag{2.110}$$

is defined to be the unique solution to (2.109), with (x,w) as the initial value at $t = 0$.

The graph $G(t)$ at time t of the solution of (2.108) is given by

$$G(t) = \{(x,w) \in \mathbb{R} \times \mathbb{R} : w = u(x,t)\}, \tag{2.111}$$

and from the method of characteristics, we get for any C^1 solution to (2.108),

$$G(t) = F^t G(0). \tag{2.112}$$

After finite time (see section 2.3.1), the single-valued graph $G(t)$ becomes multi-valued, hence it is no longer a solution to (2.108).

Now consider the subgraph $SG(t)$ of $G(t)$ at time t,

$$SG(t) = \{(x,w) \in \mathbb{R} \times \mathbb{R} : w \leq u(x,t)\}. \tag{2.113}$$

Provided $u(x,t)$ remains smooth, $SG(t) = F^t SG(0)$, and $F^t SG(0)$ defines the same multivalued solution as $F^t G(0)$. See Fig. 2.21, where $G(0)$, $F^t G(0)$ are the solution and $SG(0), F^t SG(0)$ are the areas under the curves. Following the method of characteristics, and integrating (2.109), we get

$$F^{'}(x,w) = (X^{'}(x,w), U^{'}(x,w)) = (x + tf^{'}(w), w) \tag{2.114}$$

and

$$F^{'}G(0) = \{(x + tf^{'}(w), w), (x,w) \in G(0)\}. \tag{2.115}$$

2.10.1 Geometrical Equal Area Construct

The equal area construction is a geometrical construction which permits the exact entropy solution to be found from the multivalued curve given by the method of characteristics. See Fig. 2.22.

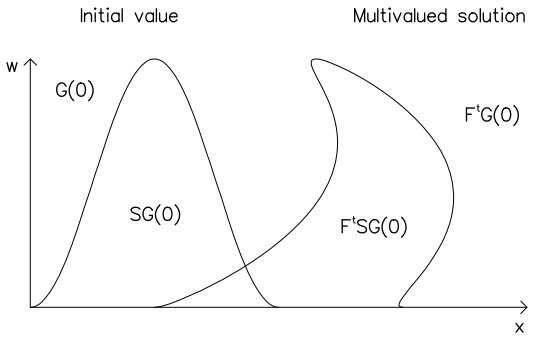

Figure 2.21: Initial and multivalued curve.

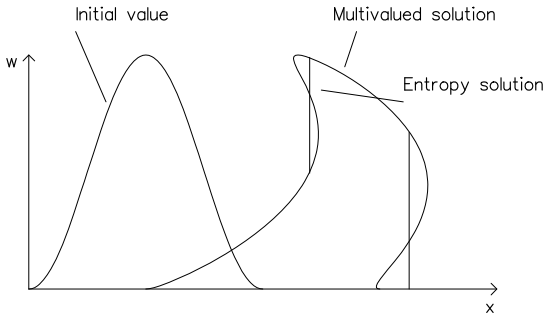

Figure 2.22: Multivalued curve with equal area construction.

Mathematically, it may be considered to be that the area between the graph $G(t)$ of the entropy solution and the multigraph $F^t(G(0))$ must be zero i.e.

$$\text{meas}(SG(t) \backslash F^t SG(0)) = \text{meas}(F^t SG(0) \backslash SG(t)) \tag{2.116}$$

where meas is defined as the Lebesgue measure. It should be noted that the conservation law described in section 2.2 is satisfied by this method. For (2.116) to be satisfied f is required to be either convex or concave.

2.10.2 Vertical Average Of Multivalued Solution

A geometrical construct (similar to the equal areas) is introduced by Brenier, which gives an approximation to the entropy condition.

For each point x, an average of the values on the multivalued curve is taken, so that they satisfy

$$\text{meas}_x(SG(t) \backslash F^t SG(0)) = \text{meas}_x(F^t SG(0) \backslash SG(t)) \tag{2.117}$$

where $\text{meas}_x(A)$ is the one-dimensional Lebesgue measure of the vertical slice $\{w \in \mathbb{R} : (x, w) \in A\}$ for any measurable subset A of $\mathbb{R} \times \mathbb{R}$. Since $\text{meas}(A) = \int_{\mathbb{R}} \text{meas}_x(A)dx$, by Fubini's theorem, (2.116) is satisfied when (2.117) holds a.e. $x \in \mathbb{R}$. A method is now required to satisfy (2.117).

2.10.3 Practical Method Of Calculation

Choose $x \in \mathbb{R}$ and let $w_0, ..., w_{2p}$ be the values of the multivalued solution at x. Let these be ordered so that

$$w_0 \leq ... \leq w_{2p}. \tag{2.118}$$

Let $\hat{w}$ be the average value, then

$$\text{meas}_x(F^t SG(0) \backslash SG(t)) = \sum_{k=1,...,p} \max(\hat{w}, w_{2k}) - \max(\hat{w}, w_{2k-1}) + \max(\hat{w}, w_0) - \hat{w} \tag{2.119}$$

$$\text{meas}_x(SG(t) \backslash F^t SG(0)) = \sum_{k=0,...,p-1} \min(\hat{w}, w_{2k+1}) - \min(\hat{w}, w_{2k}) + \hat{w} - \min(\hat{w}, w_{2p}). \tag{2.120}$$

Hence

$$\text{meas}_x(F^t SG(0) \backslash SG(t)) - \text{meas}_x(SG(t) \backslash F^t SG(0)) = \sum_{k=0,2p} (-1)^k w_k - \hat{w}. \tag{2.121}$$

Hence to satisfy (2.117)

$$\hat{w} = \sum_{k=0,2p} (-1)^k w_k. \tag{2.122}$$

and hence $\hat{w}$ satisfies $w_0 \leq \hat{w} \leq w_{2p}$. It also satisfies the convex inequality

$$V(\hat{w}) = V(\sum_{k=0,2p} (-1)^k w_k) \leq \sum_{k=0,2p} (-1)^k V(w_k) \tag{2.123}$$

which guarantees that the construction satisfies a discrete version of the entropy condition. Consequently, we have constructed from the multivalued solution to (2.105) at time t, a new single valued function which is a good approximation to the entropy solution . The operator which transforms the initial data $u(0, x)$ onto this new function is called the TC operator, (T(t)).

Geometrically, the above discussion can be seen as vertical slices through the curve. The multivalued solution at x is equivalent to the vertical slice

$$\{w \in \mathbb{R} : (x, w) \in F^t SG(0)\} =]-\infty, w_0] \cup [w_1, w_2] \cup ... \cup [w_{2p-1}, w_{2p}] \tag{2.124}$$

which is not connected and the single-valued function is obtained by replacing (2.124) by a connected curve of equivalent measure.

Note: $\hat{w} - m = (w_{2p} - w_{2p-1}) + ... + (w_2 - w_1) + (w_0 - m)$ for any $m \in \mathbb{R}$ where $m < w_0$ and $m < \hat{w}$, and this is equivalent to (2.122).

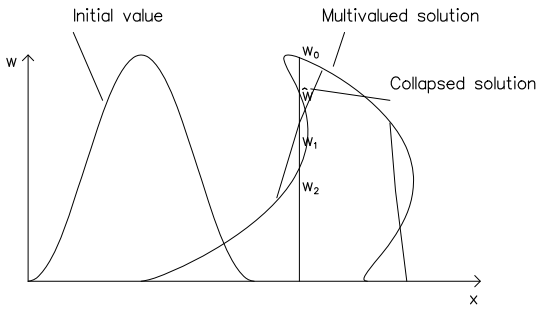

Figure 2.23: Multivalued curve with TC operator solution marked.

Using the TC operator, we have now constructed a single-valued entropy satisfying solution from the multivalued curve given by the method of characteristics. See Fig. 2.23.

2.11 Summary

In this chapter several important concepts concerning the solution of conservation laws have been introduced. Characteristics have allowed us to see how shocks and expansions form as the solution of conservation laws evolve with time. From examining both shocks and expansions the ideas of weak solutions became necessary, which in turn led to the jump and entropy conditions.

Another solution technique based upon characteristics allows the formation of multivalued solutions and the subsequent recovery of the shock position. The multivalued solutions are obtained by following the characteristics through the shock position. There are several recovery techniques discussed which are all based upon the principle of conservation.

It has been seen that the analytic calculation of the solution to conservation laws is not simple and conservation laws are the simplest nonlinear PDE's which exhibit shocks. This leads us to consider numerical methods of solution and in the

next chapter we will consider several numerical techniques for the solution of these equations. In chapter 4 a method of implementing the calculation of overturned solutions which is based on the numerical methods described in chapter 3 will be introduced. Moreover in chapter 4 the techniques for recovery of the shock position described in sections 2.9 and 2.10 will be implemented numerically.

Chapter 3

Numerical Methods In 1-D

3.1 Introduction

There are many numerical techniques available for the solution of scalar nonlinear partial differential equations each with their own advantages and disadvantages. This means that the choice of the numerical method should be made using the information known about the type of problem. In particular, solutions of nonlinear conservation laws generally form shocks or moving fronts (chapter 2). As a consequence there are regions within the solution which need high resolution and other regions which are relatively uninteresting. The resolution required may be obtained by either increasing the number of nodes or using an adaptive method.

An adaptive method is here defined as one where the grid moves with the solution, so that regions in which there are large changes in the solution have many nodes while in others the solution can be represented to the same accuracy with only a few. If a large number of uniformly spaced nodes is used instead of an adaptive grid then the grid will be expensive and wasteful, since most of the nodes will not be needed (see Fig. 3.1). Since the solutions of conservation laws generally contain moving fronts, it would therefore seem more natural to use an adaptive method as this could be more flexible and cheaper computationally (Hawken, Gottlieb & Hansen (1991)).

The numerical scheme will depend on the type of grid chosen but, disregarding our preference for adaptive methods for the moment, let us consider finite element and finite difference methods in general.

Chapter 3

Numerical Methods In 1-D

3.1 Introduction

There are many numerical techniques available for the solution of scalar nonlinear partial differential equations each with their own advantages and disadvantages. This means that the choice of the numerical method should be made using the information known about the type of problem. In particular, solutions of nonlinear conservation laws generally form shocks or moving fronts (chapter 2). As a consequence there are regions within the solution which need high resolution and other regions which are relatively uninteresting. The resolution required may be obtained by either increasing the number of nodes or using an adaptive method.

An adaptive method is here defined as one where the grid moves with the solution, so that regions in which there are large changes in the solution have many nodes while in others the solution can be represented to the same accuracy with only a few. If a large number of uniformly spaced nodes is used instead of an adaptive grid then the grid will be expensive and wasteful, since most of the nodes will not be needed (see Fig. 3.1). Since the solutions of conservation laws generally contain moving fronts, it would therefore seem more natural to use an adaptive method as this could be more flexible and cheaper computationally (Hawken, Gottlieb & Hansen (1991)).

The numerical scheme will depend on the type of grid chosen but, disregarding our preference for adaptive methods for the moment, let us consider finite element and finite difference methods in general.

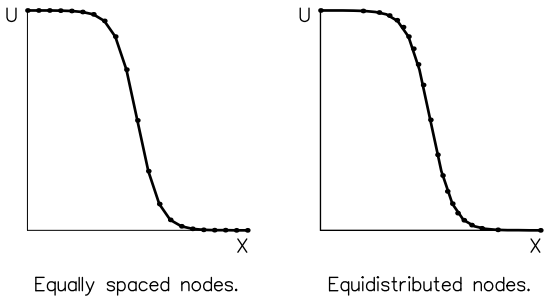

Figure 3.1: Node placement on curves.

Finite difference methods on fixed grids have been widely used for many years and a great deal is known about these schemes (see for example (Mitchell & Griffiths (1980)), (Richtmyer & Morton (1967))). Explicit and implicit methods exist using either central differences or upwind differences. Methods which combine first order central differences in space with forward differences in time are well-known to be unstable for simple linear first order differential equations and upwinding is then required. First order upwinding solves the stability problem but introduces heavy smearing. Second order schemes reduce the smearing but introduce oscillations. Recently there has been a large amount of work on second order Total Variation Diminishing (TVD) schemes, which have been very successful in reducing smearing for this type of problem while avoiding oscillations (see e.g. (Boris & Book (1973)), (Harten (1983)), (Sweby (1984)), (van Leer (1979)), (Roe (1983))).

Another important point to consider is conservation (Smoller (1983)). If we consider the equation

$$u_t + f(u)_x = 0, \tag{3.1}$$

then integrate with respect to x on a region $[A, B]$, to give

$$\int_A^B (u_t + f_x)dx = 0, \tag{3.2}$$

this may be rewritten as

$$\frac{d}{dt}\int_A^B u dx + f_B - f_A = 0 \tag{3.3}$$

where f_A, f_B are boundary terms. Conservation was defined in chapter 2 as the integral of u over the region remaining constant provided that boundary terms

are not included. If the boundary terms are ignored in equation (3.3) we get

$$\int_A^B u dx = \text{constant} \tag{3.4}$$

which shows that the equation is conservative.

Numerically, a finite difference scheme will be conservative if it is of the form

$$\frac{u_k^{n+1} - u_k^n}{\Delta t} = -\frac{h_{k+\frac{1}{2}}^n - h_{k-\frac{1}{2}}^n}{\Delta x} \tag{3.5}$$

where $h_{k+\frac{1}{2}} = h(u_{k-1}^n, ..., u_{k+r}^n)$ and h is a consistent numerical flux, i.e. $h(u, ..., u) = f(u)$. This has a similar form to (3.3) and in 1960 (Lax & Wendroff (1960)) showed that if the scheme (3.5) converges as $\Delta x \to 0$, with $\frac{\Delta t}{\Delta x}$ fixed, then it converges to a weak solution of (3.1). Since weak solutions are non-unique (see chapter 2), in order to obtain the correct physical solution an entropy condition is required. Various numerical entropy conditions have been suggested (Osher (1984)), (Lax (1972)), (Oleinik (1957)) as well as classes of schemes which are guaranteed entropy satisfying (Osher (1984)), (Tadmor (1984)), (Harten, Hyman & Lax (1976)).

Now consider finite element methods. These are less popular and less well developed than finite difference methods for the problems considered here and, on a fixed grid, oscillations occur in the neighbourhood of steep fronts when using the Galerkin approach (Herbst (1982)). More sophisticated methods such as Petrov-Galerkin (Morton (1985)) and Taylor-Galerkin (Donea (1984)) have improved the performance of finite element methods, which have the advantage of great flexibility in the mesh used. If a finite element method is applied to a conservation law, using piecewise constant or piecewise linear elements on an arbitrary grid, the integral of the initial data is conserved. This can be shown by using any Galerkin weak form of the equation. Let $\alpha_i(x)$ $(i = 1, ..., n)$ be the test functions where $\sum_{i=1}^n \alpha_i = 1$. Consider (3.1), then write it in the Galerkin weak form

$$\int_A^B (u_t + f_x)\alpha_i dx = 0 \qquad i = 1, ..., n. \tag{3.6}$$

Now summing over i gives

$$\sum_{i=1}^n \int_A^B (u_t + f_x)\alpha_i dx = 0 \tag{3.7}$$

$$\Rightarrow \int_A^B (u_t + f_x) \sum_{i=1}^{n} \alpha_i dx = 0 \tag{3.8}$$

$$\Rightarrow \int_A^B (u_t + f_x) dx = 0 \tag{3.9}$$

$$\Rightarrow \frac{d}{dt} \int_A^B u dx = f_A - f_B. \tag{3.10}$$

This shows that apart from boundary conditions the integral of the initial area is conserved.

The types of methods generally used for PDE's are based on the Galerkin finite element approach which normally has fixed grids and fixed basis functions. From the discussion of analytic methods for the solution of scalar nonlinear partial differential equations in chapter 2, it is clear that a method which permits a solution to overturn is of interest. However, in all fixed grid methods the numerical solution can never overturn or become multivalued. Since solutions of conservation laws are wavelike, often leading to overturning, it seems reasonable to introduce a moving grid and basis functions which are themselves dependent on time.

3.2 Global Moving Finite Elements

In 1981 (Miller & Miller (1981)), (Miller (1981)) introduced the Moving Finite Element (MFE) procedure to cope with problems whose solutions include steep moving fronts. This was achieved by allowing the grid to move automatically with the solution, ideally to regions where high resolution was required. The idea is now described in a general (1-D scalar) setting with t (time) as a distinguished variable.

Consider the equation

$$u_t - \mathcal{L}(u) = 0, \tag{3.11}$$

on a region $\Omega \in \mathbb{R}$, where $u = u(x,t) \in H^2$, H^2 is a Hilbert space of functions whose second derivatives are square integrable and $\mathcal{L}$ is an operator containing space derivatives u_x and u_{xx}. The problems we are considering in this thesis (see chapter 2) do not contain second derivatives of x, but the MFE method was originally developed to solve problems with near shocks in parabolic problems.

It is for this reason our description of the MFE method will include these terms. Note: we will not consider systems of PDE's in this thesis.

Let $U(t)$ be an approximate solution to (3.11) where $U(t)$ lies in a finite dimensional trial space $\subset H^2$ with linear basis functions α_i $(i = 1, ..., N)$. Note: The approximation $U(t)$ could be chosen to lie in a different linear space and the basis functions α_i could be quadratics, cubics or other functions. For descriptions of MFE using other types of basis functions (see (Jimack (1988a)), (Jimack (1988b))).

Let $U(t)$ be given by

$$U(t) = \sum_{j=1}^{N} a_j(t)\alpha_j(x, \mathbf{s}(t)) \tag{3.12}$$

where $a_j, (j = 1, ..., N)$ are the nodal amplitudes. The nodes are ordered so that $s_1 < ... < s_j < ... < s_N$, $\mathbf{s}(t) = (s_1, ..., s_N)$ and the end values are chosen so that the boundary conditions are consistent with the solution to be found. Here we assume Dirichlet boundary conditions are applied.

Differentiating (3.12) with respect to t gives

$$U_t = \sum_{j=1}^{N} \dot{a}_j(t)\alpha_j(x, \mathbf{s}(t)) + \dot{s}_j(t)\beta_j(x, a, \mathbf{s}(t)), \tag{3.13}$$

so that U_t has $2N$ unknowns (neglecting boundary conditions), $\dot{a}_j$ and $\dot{s}_j$ $(j = 1, ..., N)$ (Miller & Miller (1981)), (Miller (1981)), (Baines & Wathen (1988)), (Jimack (1988a)), (Jimack (1988b)). In (3.13) α_j and β_j are basis functions which may be defined as

$$(a) \quad \alpha_j = \frac{\partial U}{\partial a_j} \quad (b) \quad \beta_j = \frac{\partial U}{\partial s_j}. \tag{3.14}$$

Equation (3.14)(b) leads to

$$\beta_j = -m_j\alpha_j \tag{3.15}$$

which holds for linear basis functions where

$$m_j = \begin{cases} m_{j-\frac{1}{2}} & s_{j-1} < x < s_j \\ m_{j+\frac{1}{2}} & s_j < x < s_{j+1} \end{cases} \tag{3.16}$$

and

$$m_{j-\frac{1}{2}} = \frac{a_j - a_{j-1}}{s_j - s_{j-1}}. \tag{3.17}$$

However for higher order basis functions (3.15) is not always valid (Jimack (1988a)), (Lynch (1982)).

The basis functions considered here are piecewise linear and compact so that α, β are defined by

$$\alpha_j = \begin{cases} \frac{x - s_{j-1}}{s_j - s_{j-1}} & s_{j-1} \leq x \leq s_j \\ \frac{s_{j+1} - x}{s_{j+1} - s_j} & s_j \leq x \leq s_{j+1} \\ 0 & \text{elsewhere} \end{cases} \tag{3.18}$$

$$\beta_j = \begin{cases} -m_{j-\frac{1}{2}}\alpha_j & s_{j-1} < x < s_j \\ -m_{j+\frac{1}{2}}\alpha_j & s_j < x < s_{j+1} \\ 0 & \text{elsewhere} \end{cases} . \tag{3.19}$$

See Fig. 3.2. (β_j is not defined at node j). It should also be noted that the basis

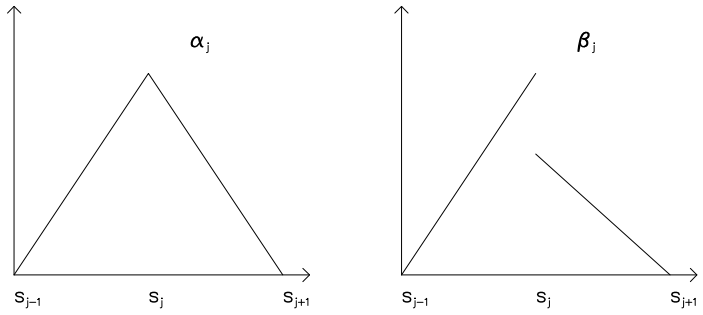

Figure 3.2: α and β basis functions.

functions β_j ($j = 1, ..., N$) are discontinuous at $x = s_j$ which implies that U_t in (3.13) is discontinuous at all nodes.

In Millers' method the equations for the solution of $\mathbf{a}$ and $\mathbf{s}$ where $\mathbf{a} = (a_1, ..., a_N)^T$, $\mathbf{s} = (s_1, ..., s_N)^T$ are obtained by minimising the L_2 residual of $U_t - \mathcal{L}(u)$. It is here that (Miller & Miller (1981)) add penalty functions in order to prevent nodes or slopes from becoming too close, and hence causing numerical problems (see section 3.5), but we will first describe the basic method without such additions.

Define the usual weighted L_2 inner product by

$$< f, g >= \int_\Omega f(x)g(x)W(x)dx \tag{3.20}$$

where f, g are integrable in the region Ω and $W(x)$ is a positive weight function. The L_2 norm squared may now be defined as

$$\|g\|^2 =< g, g > \tag{3.21}$$

where this notation will be used throughout this text.

The $\dot{a_j}$ and $\dot{s_j}$'s $(j = 1, ..., N)$ are obtained by minimising

$$\|U_t - \mathcal{L}(U)\|^2, \tag{3.22}$$

which yields, on differentiation with respect to $\dot{a}_j$ and $\dot{s}_j$,

$$\left.\begin{array}{l} < \alpha_j, U_t - \mathcal{L}(U) >= 0 \\ < \beta_j, U_t - \mathcal{L}(U) >= 0 \end{array}\right\} \quad j = 1, ..., N \tag{3.23}$$

where $< ., . >$ is defined by (3.20) . Equations (3.23) with (3.13) lead to a system of ordinary differential equations,

$$A(\mathbf{y})\dot{\mathbf{y}} = \mathbf{g}(\mathbf{y}) \tag{3.24}$$

where

$$\dot{\mathbf{y}} = (...; \dot{a}_{j-1}, \dot{s}_{j-1}; \dot{a_j}, \dot{s_j}; ...)^T, \tag{3.25}$$

$$A = \{A_{ij}\}, \qquad A_{ij} = \begin{pmatrix} < \alpha_i, \alpha_j > & < \alpha_i, \beta_j > \\ < \beta_i, \alpha_j > & < \beta_i, \beta_j > \end{pmatrix} \tag{3.26}$$

and

$$\mathbf{g} = \{g_i\}, \qquad g_i = \begin{pmatrix} < \alpha_i, \mathcal{L}(U) > \\ < \beta_i, \mathcal{L}(U) > \end{pmatrix}. \tag{3.27}$$

Since α, β straddle two intervals, A is a symmetric, 2×2 block tridiagonal, positive semi-definite matrix. It is easy to show that A is positive semi-definite since it arises from the minimisation of the term $\|U_t\|^2$ in (3.22), which by itself is

$$\|U_t\|^2 = \dot{\mathbf{y}}^T A\dot{\mathbf{y}}. \tag{3.28}$$

This shows that the quadratic form $\dot{\mathbf{y}}^T A\dot{\mathbf{y}}$ is positive semi-definite, being zero only for non-zero $\dot{\mathbf{y}}$ when A is singular (Wathen & Baines (1984)).

3.2.1 Solution Of Global MFE Equations

The system (3.24) needs to be solved for $\mathbf{y}$ and this may be done in a variety of ways. For example, a block tridiagonal solver (see e.g. Golub & Van Loan (1983)) may be used. Alternatively, other approaches include the use of the pre-conditioned conjugate gradient semi-iterative method (Golub & Van Loan (1983)) or more conventional iterative methods. Providing that A is non-singular then Jacobi, Gauss-Seidel and SOR all converge in 1-D because of the properties of A.

Eigenvalue Clustering

The eigenvalues of $D^{-1}A$, (where A is the global MFE matrix and D is a positive definite matrix comprised of the diagonal blocks of A) are $\frac{1}{2}$ and $\frac{3}{2}$ in pairs. For the Dirichlet case there are also two eigenvalues of 1. Proof of this is given in (Wathen (1987)). This result is important since it implies that the generalized conjugate gradient method will converge rapidly when applied to $D^{-1}A$.

3.3 Time-stepping

The MFE method gives rise to a system of ODE's in time, which require integration to obtain the complete solution. There are two entirely different views on how the ODE's should be integrated, dependent upon the type of approach used.

For MFE methods without penalty functions it has been suggested by (Wathen (1984)), (Johnson (1986)), (Johnson, Wathen & Baines (1988)), (Baines & Wathen (1988)) that for a wide range of problems explicit time-stepping is sufficient and that implicit methods do not give any advantage. Here time-stepping is carried out using the explicit Euler method

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \Delta t \dot{\mathbf{y}}^n \tag{3.29}$$

where

$$\mathbf{y} = (...; a_j, s_j; ...)^T. \tag{3.30}$$

Ideally we want the time-step to be as large as is consistent with good accuracy while remaining within the stability region. However to avoid node overtaking in

cases where a single-valued solution is expected, the time-step must be no larger than that which would allow it to catch up with its neighbour.

The alternative view is held by Miller, who introduced penalty functions in the original papers. On the grounds that the system of ODE's that are obtained are stiff he argues that an implicit method must be used. In recent papers the systems of ODE's obtained by this method have been solved by Miller using the implicit Euler time-stepping method,

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \Delta t \dot{\mathbf{y}}^{n+1} \tag{3.31}$$

with a Newton solver. The iteration doesn't always converge, however, and both Δt and the parameters in the penalty functions have to be tuned so that convergence takes place.

3.4 Local MFE

The local MFE approach was introduced by (Baines (1985)) who observed that U_t in (3.13) could be written in terms of local elementwise basis functions. This allows the rewriting of the original expression for U_t (3.13) as a sum over the elements in the space of piecewise linear discontinuous basis functions, instead of over the nodes. This gives

$$U_t = \sum_{j=0}^{N} (\dot{a}_j \alpha_j + \dot{s}_j \beta_j) = \sum_{k=1}^{N} (w_k^{(1)} \phi_k^{(1)} + w_k^{(2)} \phi_k^{(2)}), \tag{3.32}$$

where $\phi_k^{(i)}$ are basis functions, $w_k^{(i)}$ are coefficients related to $\dot{a}_j$, $\dot{s}_j$ and where j is a node and k is an element. The element basis functions $\phi_k^{(1)}$, $\phi_k^{(2)}$ are shown in Fig. 3.3 and are defined for piecewise linears as

$$\phi_k^{(1)} = \begin{cases} \dfrac{x - s_{j-1}}{s_j - s_{j-1}} & s_{j-1} \le x \le s_j \end{cases} \tag{3.33}$$

$$\phi_k^{(2)} = \begin{cases} \dfrac{s_j - x}{s_j - s_{j-1}} & s_{j-1} \le x \le s_j. \end{cases} \tag{3.34}$$

Let S_ϕ be the space spanned by the basis functions $\phi_k^{(i)}$ which in 1-D can be shown to be the same as the space $S_{\alpha\beta}$ space spanned by the α_j, β_j basis functions (Baines (1985)). As in the MFE method proposed by (Miller & Miller (1981)),

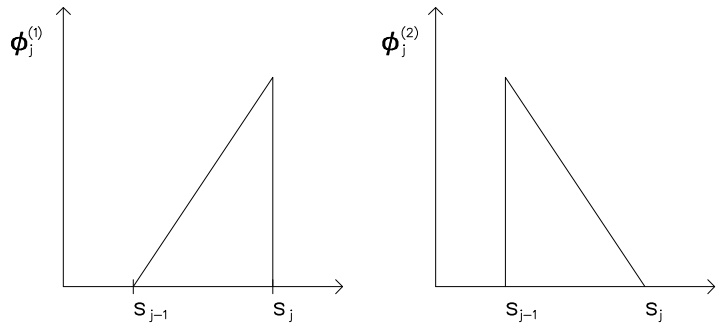

Figure 3.3: Local Basis Functions.

the residual $\|U_t - \mathcal{L}(U)\|^2$ is minimised. This is now however minimised over $w_k^{(1)}$, $w_k^{(2)}$ to obtain

$$< \phi_k^{(1)}, U_t - \mathcal{L}(U) > \;=\; 0 \tag{3.35}$$

$$< \phi_k^{(2)}, U_t - \mathcal{L}(U) > \;=\; 0 \tag{3.36}$$

for each element k. The system (3.35), (3.36) with U_t given by (3.32), decouples and may be written as N systems, each 2×2,

$$C_k \mathbf{w}_k = \mathbf{b}_k \qquad k = 1, ..., N \tag{3.37}$$

where

$$\mathbf{w}_k = \begin{pmatrix} w_k^{(1)} \\ w_k^{(2)} \end{pmatrix}, \tag{3.38}$$

$$C_k = \begin{pmatrix} < \phi_k^{(1)}, \phi_k^{(1)} > & < \phi_k^{(1)}, \phi_k^{(2)} > \\ < \phi_k^{(2)}, \phi_k^{(1)} > & < \phi_k^{(2)}, \phi_k^{(2)} > \end{pmatrix} \tag{3.39}$$

and

$$\mathbf{b}_k = \begin{pmatrix} < \phi_k^{(1)}, \mathcal{L}(U) > \\ < \phi_k^{(2)}, \mathcal{L}(U) > \end{pmatrix}. \tag{3.40}$$

From (3.18), (3.19) and (3.33), (3.34) it is easily seen that

$$\alpha_j \;=\; \phi_{j-\frac{1}{2}}^{(2)} + \phi_{j+\frac{1}{2}}^{(1)}, \tag{3.41}$$

$$\beta_j \;=\; -m_{j-\frac{1}{2}} \phi_{j-\frac{1}{2}}^{(2)} - m_{j+\frac{1}{2}} \phi_{j+\frac{1}{2}}^{(1)}, \tag{3.42}$$

where $\phi_{j-\frac{1}{2}}^{(1)}$, $\phi_{j-\frac{1}{2}}^{(2)}$ are the ϕ basis functions in the element $(j-1, j)$, i.e. the element k in the present notation. From (3.41), (3.42) and (3.32), the relationships

$$\dot{a_j} - m_{j-\frac{1}{2}} \dot{s_j} \;=\; w_{j-\frac{1}{2}}^{(2)} = w_k^{(2)} \tag{3.43}$$

$$\dot{a_j} - m_{j+\frac{1}{2}}\dot{s_j} \;=\; w^{(1)}_{j+\frac{1}{2}} = w^{(1)}_{k+1} \tag{3.44}$$

are obtained for all nodes j. Equations (3.44), (3.43) can be written as the 2×2 system

$$M_j \dot{\mathbf{y}}_j = \mathbf{w}_j \tag{3.45}$$

where

$$\dot{\mathbf{y}}_j = \begin{pmatrix} \dot{a_j} \\ \dot{s_j} \end{pmatrix}, \tag{3.46}$$

$$M_j = \begin{pmatrix} 1 & -m_{j-\frac{1}{2}} \\ 1 & -m_{j+\frac{1}{2}} \end{pmatrix} \tag{3.47}$$

and

$$\mathbf{w}_j = \begin{pmatrix} w^{(2)}_{j-\frac{1}{2}} \\ w^{(1)}_{j+\frac{1}{2}} \end{pmatrix}. \tag{3.48}$$

Since in 1-D both the Miller method and local method minimise the same residual in the same space, the MFE equations derived will be identical. Hence by writing

$$C = \begin{pmatrix} \ddots & & 0 \\ & C_k & \\ 0 & & \ddots \end{pmatrix}, \qquad M = \begin{pmatrix} \ddots & & 0 \\ & M_k & \\ 0 & & \ddots \end{pmatrix} \tag{3.49}$$

this leads to the 'Miller' global MFE matrix of (3.24) being decomposed into

$$A = M^T C M \tag{3.50}$$

where C, M are both 2×2 block diagonal (Wathen & Baines (1984)). Although both M and C are both block diagonal, the blocks are staggered with respect to each other since those in M are node based and those in C are element based.

3.5 Singularities Of A

In section 3.2, it was described how the MFE equations can now be solved when A is non-singular. The case when A has singularities is still to be discussed. In the local method provided that C_k and M_j are non-singular, the solution is trivial. Again, however, the singularities need to be considered. Since the introduction of the local method has provided a decomposition for A, the cases of singularities can be discussed more clearly.

3.5.1 Coincident Nodes

If we consider $A = M^T C M$, then when A is singular, this implies that either M and/or C is singular. If C is singular then this implies that at least one $\Delta s = 0$. If however C is singular, then, as we shall see, this does not necessarily imply that A is singular.

The case where two nodes merge and become coincident was originally thought to cause a singularity. However the problem was considered by (Sweby (1987)) who showed this to be generally false in the sense described below for both the local and global approach.

If two nodes become coincident (but not three) then, although C becomes singular, it can be shown that its product with M remains non-singular. If however three neighbouring nodes become coincident then $\Delta s_{j-\frac{1}{2}} = \Delta s_{j+\frac{1}{2}} = 0$ for some node j where $\Delta s_{j-\frac{1}{2}} = s_j - s_{j-1}$. This implies that two C_k's become singular, and A does become singular.

To establish these results note that C can alternatively be decomposed into 3 matrices

$$C = E^T C' E \tag{3.51}$$

where E is diagonal and C' is block diagonal such that $E = \{E_k\}$, $C' = \{C'_k\}$

$$C_k = E_k^T C'_k E_k \tag{3.52}$$

where

$$E_k = \begin{pmatrix} \Delta s_k^{\frac{1}{2}} & 0 \\ 0 & \Delta s_k^{\frac{1}{2}} \end{pmatrix}, \qquad C'_k = \frac{1}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{3.53}$$

This shows that, since C'_k is non-singular, then any singularity must be in E_k. Now, returning to the global matrix A, it can be seen that it has a five matrix decomposition

$$\begin{aligned} A &= M^T E^T C' E M \\ \Rightarrow A &= N^T C' N. \end{aligned} \tag{3.54}$$

say, where $N = EM$ so that each block of N becomes

$$N_{j-1} = \begin{pmatrix} \Delta s_{k-1}^{\frac{1}{2}} & -m_{k-1}\Delta s_{k-1}^{\frac{1}{2}} \\ \Delta s_k^{\frac{1}{2}} & -m_k \Delta s_k^{\frac{1}{2}} \end{pmatrix} \tag{3.55}$$

$$= \begin{pmatrix} \Delta s_{k-1}^{\frac{1}{2}} & -\Delta U_{k-1}\Delta s_{k-1}^{-\frac{1}{2}} \\ \Delta s_k^{\frac{1}{2}} & -\Delta U_k \Delta s_k^{-\frac{1}{2}} \end{pmatrix} \tag{3.56}$$

by the definition of m_{k-1}, m_k. Since N is block diagonal only one of the blocks N_j (3.56) need be considered.

One technique of analysis was given by (Sweby (1987)). Here we give an alternative version based upon QR decomposition. Using Gram-Schmidt orthogonalization (Golub & Van Loan (1983)) we write the N_j matrix in terms of a matrix product of Q_j and R_j. Let $N_j = (\mathbf{n_1}, \mathbf{n_2})$, $Q_j = (\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2)$ and R_j be an upper triangular with elements r_{11}, r_{12}, r_{22}. For N_j to be invertible it is necessary that r_{11} and r_{22} are non-zero.

Using the Gram-Schmidt orthogonalization we obtain

$$\hat{\mathbf{q}}_1 = \frac{\mathbf{n}_1}{n_1} \text{ then } r_{11} = n_1 \text{ where } n_1 = |\mathbf{n}_1|. \tag{3.57}$$

For r_{11} to be non-zero we require that $\sum_{i=k-1}^{k} |\Delta s_i| \neq 0$ which means that either $\Delta s_{k-1} \neq 0$ and/or $\Delta s_k \neq 0$.

The second equation becomes

$$r_{22}\hat{\mathbf{q}}_2 = \mathbf{n}_2 - (\hat{\mathbf{q}}_1.\mathbf{n}_2)\hat{\mathbf{q}}_1 \tag{3.58}$$

$$= \mathbf{n}_2 - \frac{(\mathbf{n}_1.\mathbf{n}_2)}{n_1^2}\mathbf{n}_1 \tag{3.59}$$

$$= \frac{1}{n_1^2}\mathbf{n}_1 \times (\mathbf{n}_2 \times \mathbf{n}_1). \tag{3.60}$$

This means that for r_{22} to be non-zero we require $\mathbf{n}_1 \neq 0$, $\mathbf{n}_1 \times \mathbf{n}_2 \neq 0$ and $\mathbf{n}_1$ to not be parallel to $\mathbf{n}_1 \times \mathbf{n}_2$. We already require that $\mathbf{n_1}$ is non-zero, so let us consider the case of $\mathbf{n}_1 \times \mathbf{n}_2 \neq 0$. For this to occur we again require $\mathbf{n}_1$ to be non-zero however we also require $\mathbf{n}_2$ to be non-zero. This is equivalent to $\sum_{i=k-1}^{k} |\Delta s_i| m_i^2 \neq 0$ which is evidently true. The cross product term requires $m_k \neq m_{k-1}$ which is clearly the parallelism singularity. If we assume that no parallelism occurs then N, and therefore A is non-singular. Finally it should be noted that by definition, $\mathbf{n}_1$ cannot be parallel to $\mathbf{n}_1 \times \mathbf{n}_2$.

Therefore provided at least one of the elements next to the node remains non-zero and parallelism does not occur then it is possible to pass through the singularity (although the matrix N may not have bounded elements at the singularity.)

3.5.2 A Second View Of Parallelism

Parallelism can also be analysed with the original MFE system (3.24). From (3.50) A can be decomposed into $M^T C M$. Consider the cases where two adjacent slopes in the solution are equal (see Fig. 3.4.), i.e. let $m_{j+\frac{1}{2}} = m_{j-\frac{1}{2}} = m_j$ for

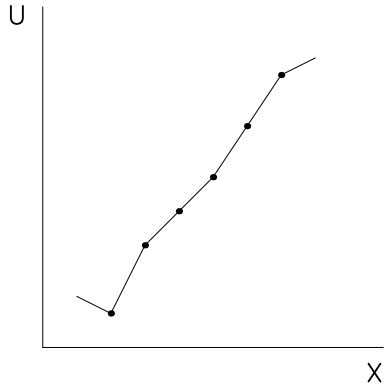

Figure 3.4: Piecewise linear curve showing parallelism.

some j. Then since $\beta_j = -m_j \alpha_j$ (3.15), it can be seen that there are two linearly dependent equations in the MFE system, consequently A is singular. However A has a simple null space spanned by $\mathbf{U}_j = (0, 0, ..., m_j, 1, ..., 0, 0)^T$, which can be used to obtain a solution despite the parallel nodes, i.e. a new system can be formed, after one of the two linearly dependent equations has been removed from the system. The new system can then be solved in the usual way and remains consistent provided $\dot{s_j}$ is set (Wathen & Baines (1984)). This is sometimes set as $\dot{s} = 0$, giving MFE with one fixed node although in general the method will depend on the specific properties of the differential equation.

3.6 Penalty Functions

Now returning to the original (Miller & Miller (1981)) method, Miller added penalty functions so as to prevent nodes moving too close to each other. Instead of minimising (3.22), he minimised

$$\|U_t - \mathcal{L}(U)\|^2 + \sum_{j=1}^{N} [\epsilon(\dot{s_j} - \dot{s}_{j-1})]^2 \tag{3.61}$$

where ϵ is a parameter, which is dependent on the problem. The penalty functions act as springs, forcing the nodes which approach each other too closely to remain

apart. According to Miller this inhibits the parallelism singularities referred to above, which otherwise may occur during the solution of (3.24). This method now requires a different solution procedure, since the simple techniques applied to (3.24) in section 3.21 will now not work (although Baines (1986) showed cases where the method of section 3.21 goes through). Moreover an implicit solver (Miller & Miller (1981)) is required because the system is now stiff.

A lot of work has been done on moving finite elements where regularisation techniques are used (e.g. Baines (1986), Juarez-Romero, Sargent & Jones (1988)). However since this type of method will not allow nodes to overtake each other it is not applicable to the present approach to overturning solutions. There is however a version of the method which is an exception to this, which is Gradient Weighted MFE (Miller (1986)), of which more later.

3.7 Variational Derivation Of MFE

The standard MFE method proposed by Miller can also be derived using a variational approach. In 1983 (Mueller & Carey (1985)) considered a time-dependent coordinate transformation in conjunction with a finite element method. We will describe this approach since the transformation is related to the Lagrangian version of the conservation law (2.11) in which the characteristics are investigated. This derivation should also be considered in conjunction with the later section 3.10 where the link between MFE and characteristics is discussed.

The equation considered is a PDE of the form

$$\frac{\partial u}{\partial t} - \mathcal{L}(u) = 0 \tag{3.62}$$

where $x \in \Omega$, $t \in (0, T)$ with T being the final time.

The coordinate transform between x, t and the new independent variables ξ, τ is of the form

$$x = \hat{x}(\xi, \tau), \qquad t = \tau \tag{3.63}$$

(c.f. section 2.3) and has Jacobian

$$J = \frac{\partial(x, t)}{\partial(\xi, \tau)} \tag{3.64}$$

where $|J| = \det(\frac{\partial x}{\partial \xi})$. The transformation is constrained to satisfy

$$|J| > 0 \quad \forall x, t; \tag{3.65}$$

which has a similar effect to the application of penalty functions in Miller's method. Similarly to (3.63), u has a coordinate transform between x, t and ξ, τ since u is dependent on x and t, u becomes

$$u = \hat{u}(\xi, \tau), \qquad t = \tau \tag{3.66}$$

(c.f. section 2.3). Using the chain rule u_t becomes

$$u_t = \hat{u}_\tau \tau_t + \hat{u}_\xi \xi_t \tag{3.67}$$

and the inverse transformation is given by

$$J^{-1} = \begin{pmatrix} \xi_x & \xi_t \\ \tau_x & \tau_t \end{pmatrix} = \begin{pmatrix} \frac{1}{x_\xi} & \frac{-x_\tau}{x_\xi} \\ 0 & 1 \end{pmatrix} \tag{3.68}$$

hence

$$\frac{\partial u}{\partial t} = \frac{\partial \hat{u}}{\partial \tau} - \frac{\partial \hat{u}}{\partial \xi} \begin{pmatrix} \frac{\partial \hat{x}}{\partial \tau} \\ \frac{\partial \hat{x}}{\partial \xi} \end{pmatrix} \tag{3.69}$$

and

$$\frac{\partial u}{\partial t} = \frac{\partial \hat{u}}{\partial \tau} - \frac{\partial u}{\partial x} \frac{\partial \hat{x}}{\partial \tau}. \tag{3.70}$$

This gives (3.62) as

$$\dot{u} - \dot{x} u_x - \mathcal{L}(u) = 0 \tag{3.71}$$

where $\dot{u} = \frac{\partial \hat{u}}{\partial \tau}$. Let R be the residual defined by (3.71) for admissible trial functions x and u so that

$$R = \dot{u} - \dot{x} u_x - \mathcal{L}(u). \tag{3.72}$$

The variational problem is now given by

$$I = \tfrac{1}{2} \int_\Omega R^2 dx \tag{3.73}$$

and is minimised over all admissible solutions u and admissible maps x. If the variations are given by $v = \delta(\frac{\partial \hat{u}}{\partial \tau})$, $z = \delta(\frac{\partial \hat{x}}{\partial \tau})$ then (3.73) gives

$$\int_\Omega (\dot{u} - \dot{x} u_x - \mathcal{L}u) v dx = 0 \tag{3.74}$$

$$\int_\Omega (\dot{u} - \dot{x} u_x - \mathcal{L}u) z \frac{\partial u}{\partial x} dx = 0 \tag{3.75}$$

for all admissible test functions (z, v), $\tau > 0$ such that all admissible transformations are invertible and (x, u) satisfies the initial and boundary conditions. We may obtain an approximate version of these equations by writing the test functions x, u as discrete approximations x_h, u_h.

$$x_h(\xi, t) = \sum_{j=1}^{N} x_j(\tau)\chi_j(\xi) \tag{3.76}$$

$$u_h(\xi, t) = \sum_{i=1}^{N} u_i(\tau)\phi_i(\xi) \tag{3.77}$$

and letting $v_h = \phi_i$ $z_h = \chi_i$ $(i = 1, ..., N)$. If we take ϕ_i and χ_i to be the basis functions α_i and β_i, then we get the MFE equations (3.23) given by Miller.

To apply the constraints (for example $|J| > 0$) the functional I is replaced by

$$I_c = I + P_\epsilon \tag{3.78}$$

where P_ϵ is a penalty term defined by

$$P_\epsilon = \int_a^b \frac{1}{2\epsilon}\left(\frac{\partial^2 \hat{x}}{\partial\xi\partial\tau} - s\frac{\partial\hat{x}}{\partial\xi}\right)dx \tag{3.79}$$

and ϵ is the strength of the penalty function and s is a parameter. A minimum value of J_{min} of the Jacobian J of the coordinate transform is specified and ϵ is chosen as $\epsilon(x) = w(\frac{J - J_{min}}{J_{min}})^2$ where w is a parameter chosen to regulate the relative size of the residual and penalty functionals.

3.8 Gradient Weighted MFE

Miller introduced (Miller (1986)) Gradient Weighted MFE (GWMFE) because of problems controlling nodes with near shocks. Where there is a very steep front, the time derivative of u was found to be almost a delta function, which is not in L_2 even in the limit. This implies that for certain types of problems, the standard L_2 norm used for the minimisation of the residual is inappropriate.

The basic MFE method remains as before, but a new norm is introduced to combat the problem of steep fronts. The L_2 norm is replaced by

$$\|.\|_N^2 = \int (.)_N^2 ds \tag{3.80}$$

$$= \int (.)^2 W(m) dx \tag{3.81}$$

where $W(m) = (1+m^2)^{-\frac{1}{2}}$, $m = |u_x|$. This means that the L_2 norm of the speed of u normal to the graph integrated with respect to arc length s has been used as the norm. This is done because $\dot{u}_N = \dot{u}(1+u_x^2)^{-\frac{1}{2}}$ is always bounded and its effect in the norm is to reduce the emphasis on the steep regions near shocks. A consequence of using the new norm is that instead of integrating the residual with respect to x, a component of the residual is now being integrated along the curve. This means that overturned solutions can be found using this method (even with penalty functions), so that it avoids problems with coincident nodes (Sweby (1987)). Note: The singularity which occurs in Global and Local MFE as one element tends to zero does not occur in this method since the integration within the minimisations is carried out along the arc length of the solution, due to the use of the weight function.

Using the basis functions and approximations for u given in section 3.2, the equations become

$$\left.\begin{aligned}(\alpha_j, U_t - \mathcal{L}(U))_N = 0 \\ (\beta_j, U_t - \mathcal{L}(U))_N = 0\end{aligned}\right\} \quad j = 1, ..., N \tag{3.82}$$

where $(.,.)_N$ is the inner product associated with the norm (3.80). This gives rise to the ODE system

$$A(\mathbf{y})\dot{\mathbf{y}} = \mathbf{g}(\mathbf{y}) \tag{3.83}$$

where

$$\dot{\mathbf{y}} = (...; \dot{a}_{j-1}, \dot{s}_{j-1}; \dot{a}_j, \dot{s}_j; ...)^T, \tag{3.84}$$

$$A = \{A_{ij}\}, \qquad A_{ij} = \begin{pmatrix} (\alpha_i, \alpha_j)_N & (\alpha_i, \beta_j)_N \\ (\beta_i, \alpha_j)_N & (\beta_i, \beta_j)_N \end{pmatrix} \tag{3.85}$$

and

$$\mathbf{g} = \{g_i\}, \qquad g_i = \begin{pmatrix} (\alpha_i, \mathcal{L}(U))_N \\ (\beta_i, \mathcal{L}(U))_N \end{pmatrix}. \tag{3.86}$$

A is symmetric, 2×2 block tridiagonal and positive semi-definite. The equations given above can be solved using for example a block tridiagonal solver or preconditioned conjugate gradient methods (Golub & Van Loan (1983)). Parallelism remains a singularity of this method.

3.9 Higher Derivatives

Up to this point we have assumed that the range of the operator $\mathcal{L}$ can be represented with linear elements. However for completeness if $\mathcal{L}$ contains second (or higher) derivatives, a way of approximating the u_{xx} terms, in terms of the piecewise linear basis functions, which only exist in the sense of distributions, must be found. In this case U_{xx} will not have a finite L_2 norm and the minimisation of (3.22) does not exist.

There are basically three techniques for dealing with this type of problem. The first was introduced by (Miller & Miller (1981)) in their original paper on MFE and involves smoothing the α basis functions so that U_{xx} terms may be defined (δ - mollification). The second method was given by (Mueller & Carey (1985)) and involves the application of Green's theorem to reduce the order of the differentiation. This method can be applied in higher dimensions. The third method (Johnson, Wathen & Baines (1988)) is the so-called recovery method, which involves fitting a polynomial (W) of sufficiently high order to the U or U_x term so that the W_{xx} term may be defined and used in place of U_{xx}. These methods are described below. Note: An alternative approach, avoiding the problem altogether (and considered by Jimack), is to replace the piecewise linear basis functions in the MFE method with quadratics or other functions so that the problem in representing the higher order terms does not occur (Jimack (1988a)), (Jimack (1988b)).

3.9.1 δ - Mollification

Consider minimising the residual

$$U_t - \mathcal{L}(\mathcal{S}(U)) \tag{3.87}$$

instead of $U_t - \mathcal{L}(U)$, where $\mathcal{S}$ is a smoothing operator defined by Miller as

$$\mathcal{S}(U)(x) = \int_{-\infty}^{\infty} \rho^{\delta}(x-y)U(y)dy \tag{3.88}$$

where ρ^{δ} is a C_0^{∞} function of unit total integral which has support within a radius δ about the origin. The α_j and β_j ($j = 1, ..., N$) basis functions are replaced by

$$\alpha_j^{\delta} = \frac{\partial \mathcal{S}(U)}{\partial a_j} = \mathcal{S}\frac{\partial U}{\partial a_j} = \mathcal{S}\alpha_j \tag{3.89}$$

$$\beta_j^\delta = \frac{\partial \mathcal{S}(U)}{\partial s_j} = \mathcal{S}\frac{\partial U}{\partial s_j} = \mathcal{S}\beta_j \tag{3.90}$$

where α_j^δ now has sufficient continuity to ensure

$$< U_t - \mathcal{L}(\mathcal{S}(U)), \beta_j > \tag{3.91}$$

exists in the limit as $\delta \to 0$. It also is assumed that the discontinuous basis function β_j takes the mean value $\frac{1}{2}(m_{j-1} + m_j)$ at node s_j and this gives

$$< u_{xx}, \beta_j > \to \tfrac{1}{2}(m_j^2 - m_{j-1}^2), \tag{3.92}$$

which is independent of the smoothing of δ as $\delta \to 0$. As a consequence, the u_{xx} terms may be evaluated and the MFE method using piecewise linear basis functions may be applied to u_{xx} terms (Miller & Miller (1981)), (Miller (1981)).

3.9.2 Mueller's Method

In the variational formulation of MFE given by (Mueller & Carey (1985)), (see section 3.7), the inner products involving 2nd order operators are evaluated using Green's theorem. If we assume that $\frac{\partial \mathcal{S}(U)}{\partial x}$ is continuous across element edges, then in 1-D integration by parts may be applied, which will reduce the order of the derivatives. This method may also be extended to higher dimensions for which Green's theorem is again required to reduce the order of the operator.

Consider the example

$$< U_{xx}, \beta_i > \tag{3.93}$$

in 1-D where $\beta_i = -\frac{\partial U}{\partial x}\alpha_i$. This becomes

$$\int_{X_{i-1}}^{X_{i+1}} U_{xx}\beta_i dx = -\int_{X_{i-1}}^{X_{i+1}} U_{xx}U_x\alpha_i dx \tag{3.94}$$

where (X_{i-1}, X_{i+1}) is the interval of support of α_i. This can now be rewritten as

$$-\int_{X_{i-1}}^{X_{i+1}} U_{xx}U_x\alpha_i dx = -\int_{X_{i-1}}^{X_{i+1}} \left(\frac{U_x^2}{2}\right)_x \alpha_i dx = \int_{X_{i-1}}^{X_{i+1}} \frac{U_x^2}{2}\frac{\partial \alpha_i}{\partial x} dx \tag{3.95}$$

which contains only U_x derivatives; consequently this may be evaluated using piecewise linear elements. Note: This method also can be applied to more general problems (Johnson, Wathen & Baines (1988)).

3.9.3 Recovery

There have been many different types of recovery implemented (Johnson (1984), Johnson, Wathen & Baines (1988)). This type of method introduced by Morton is similar to mollification. It is based on the idea that U is approximated by a smoother, recovered, function W defined as

$$W(x) = \mathcal{S}(U)(x) \tag{3.96}$$

where $\mathcal{S}(U)(x)$ is some smoothing operator (see section 3.9.1). $W(x)$ may be constructed from the piecewise linear MFE approximation U or the piecewise constant U_x so that there is enough continuity to ensure that the L_2 norm of the second derivative of

$$\mathcal{L}(\mathcal{S})U \tag{3.97}$$

exists. One method of doing this is fitting a polynomial of sufficiently high order to U or U_x. For example, consider the function $W(x)$ on element k between nodes $j-1$ and j as the Hermite cubic function satisfying

$$W(s_{j-1}) = U_{j-1} \quad W_x(s_{j-1}) = \tfrac{1}{2}(m_j + m_{j-1}) \tag{3.98}$$

$$W(s_j) = U_j \qquad W_x(s_j) = \tfrac{1}{2}(m_j + m_{j+1}) \tag{3.99}$$

where $m_j = \frac{U_j - U_{j-1}}{s_j - s_{j-1}}$. Then we approximate the operator $\mathcal{L}u$ by $\mathcal{L}W$ where W is defined by (3.98) and (3.99). Evaluating

$$< W_{xx}, \beta_i >= \int_{s_{i-1}}^{s_{i+1}} W_{xx}\beta_i dx \tag{3.100}$$

gives $-\frac{1}{2}(m_{i+1}^2 - m_i^2)$ Consequently, it can be seen that this particular recovery method is equivalent to using δ-mollification (Johnson (1986)).

However, in this thesis we shall not consider derivatives of orders higher than u_x since these inhibit the formation of shocks, which is the main focus of this work.

3.10 Lagrangian Approach To Characteristics

Before considering other adaptive finite element methods it is interesting to note that, for $\mathcal{L}(u)$ containing first derivatives only, MFE is a.e. an approximation to the method of characteristics described in chapter 2 (see e.g. Baines (1991)).

In an alternative but equivalent approach to (3.13) the MFE system is re-derived as in (Mueller & Carey (1985)). We work with the equation (c.f. (1.5)),

$$u_t + H(x, u, u_x) = 0 \tag{3.101}$$

where $u = u(x,t)$. A coordinate transform is defined (assumed non-singular) between x, t and new independent variables ξ, τ by

$$x = \hat{x}(\xi, \tau), \quad t = \tau, \quad u(x,t) = \hat{u}(\xi, \tau) \tag{3.102}$$

as in section 2.3 in chapter 2. Using the new variables (3.101) may now be written in the Lagrangian form as

$$\dot{u} - u_x \dot{x} + H(x, \hat{u}, u_x) = 0. \tag{3.103}$$

Now restricting $\hat{u}$ and $\hat{x}$ to $\hat{U}$ and $\hat{X}$, which belong to sets of admissible trial functions, (3.103) becomes

$$\dot{U} - U_X \dot{X} + H(X, \hat{U}, U_X) = R \tag{3.104}$$

where R is the residual. (This can be rewritten in the same notation describe in section 3.2 (Global MFE), with $U = a$, $X = s$ and $U_X = m$ to give (3.104) as $\dot{a} - m\dot{s} - H(X, \hat{U}, U_X) = R$.) We shall take the trial functions to be piecewise linears.

Suppose that $H(U)$ is such that the residual R equals zero, which is the case for the example of the Inviscid Burgers' Equation (2.21), and is possible with $H(U)$ in the space of piecewise linear discontinuous functions. Then (3.104) becomes

$$\dot{a} - m\dot{s} + H(s, a, m) = 0. \tag{3.105}$$

Now consider the jump in each term of the equation (3.105) across a node j (see Figs. 3.5 and 3.6), using the notation $[.]_j$ for a jump across the node.

Since $\dot{a}, \dot{s}$ are continuous at the node, taking jumps in equation (3.105) gives

$$0 - [m]_j \dot{s_j} + [H(s, a, m)]_j = 0. \tag{3.106}$$

Providing $[m]_j \neq 0$, (3.106) gives

$$\dot{s_j} = \frac{[H(s, a, m)]_j}{[m]_j}. \tag{3.107}$$

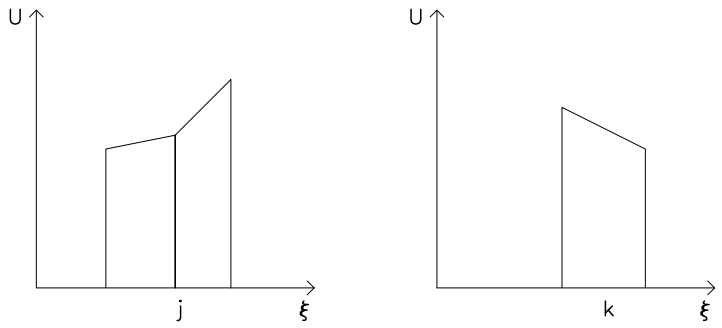

Figure 3.5: Jumps across nodes and elements.

Consider (3.105) again, and this time divide through by $m \neq 0$, to give

$$\frac{\dot{a}}{m} - \dot{s} + \frac{H}{m} = 0. \tag{3.108}$$

Now consider the jumps across node j as before, which gives

$$[m^{-1}]_j \dot{a}_j - 0 + [m^{-1}H]_j = 0. \tag{3.109}$$

This leads to

$$\dot{a}_j = -\frac{[m^{-1}H]_j}{[m^{-1}]_j}. \tag{3.110}$$

NB: $[m]_j^{-1} \neq 0$ when $[m]_j \neq 0$. (Equations (3.107) and (3.110) correspond to the solution of (3.45) in this case).

Somewhat separately, we now again consider jumps, but this time across an element k, see Figs. 3.5 and 3.6. First note that

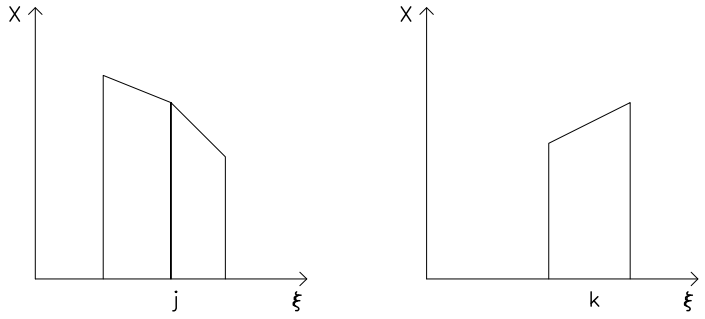

Figure 3.6: Jump across nodes and elements.

$$\dot{m}_k = \frac{\partial}{\partial \tau}\left(\frac{[a]_k}{[s]_k}\right) = \frac{([\dot{a}]_k - m[\dot{s}]_k)}{[s]_k}. \tag{3.111}$$

Then using (3.105),(3.111) becomes

$$\dot{m}_k = -\frac{[H]_k}{[s]_k}. \tag{3.112}$$

Now take limits of (3.107), (3.110) as $[s]_k \to 0$ and the number of nodes $\to \infty$, to give

$$\dot{x} = \frac{\partial H}{\partial u_x}, \quad \dot{u} = -H + u_x \frac{\partial H}{\partial u_x}. \tag{3.113}$$

Also taking the limit of (3.111), as $[s]_j \to 0$ gives

$$\dot{m} = -\frac{\partial H}{\partial x}. \tag{3.114}$$

Equations (3.113), (3.114) are the characteristic ODE's for (3.11) (see chapter 2) which shows that the characteristic equations are obtained from the MFE approximations in the limit. All this holds provided that $R = 0$. If $R \neq 0$ a projection is required (the L_2 projection of MFE, specifically that given by (3.35), (3.36)), which means that a further approximation is required.

3.11 Split Method

The split method was proposed by (Baines (1991)), (see also Edwards (1988)) and is so-called because the basic procedure is split into two sequential steps, rather than being carried out simultaneously. These may then be solved separately for $\dot{x}$ and then for $\dot{u}$, unlike the single large system obtained from the global MFE in which $\dot{x}$, $\dot{u}$ are fully coupled.

Let us again consider the equation

$$u_t + H(x, u, u_x) = 0 \tag{3.115}$$

which leads to the basic MFE equations given by (Miller & Miller (1981)) of

$$< U_t + H(X, U, U_X), \alpha_i > \;\; = \;\; 0 \tag{3.116}$$

$$< U_t + H(X, U, U_X), \beta_i > \;\; = \;\; 0 \tag{3.117}$$

where α_i, β_i $(i = 1, ..., N)$ are the usual basis functions and U_t is given by $U_t = \sum_{j=1}^{N} \dot{a}_j \alpha_j + \dot{s}_j \beta_j$ as in the global method (see section 3.2). The split method replaces (3.117), the β equation, by the weak form of the first of (3.113), namely

$$< \dot{s} - \frac{\partial H}{\partial U_X}, \alpha_i >= 0 \tag{3.118}$$

and, having found $\dot{s}$, obtains $\dot{a}$ from (3.116) with $U_t = \dot{a} - m\dot{s}$. Returning to (3.116) and (3.118), there are now two residuals to be minimised

$$\left\| \dot{s} - \frac{\partial H}{\partial U_X} \right\|^2, \tag{3.119}$$

over $\dot{s}_j$ giving $\dot{s}$ and

$$\|\dot{a} - m\dot{s} - H(X, U, U_X)\| \tag{3.120}$$

over $\dot{a}_j$ with $\dot{s}$ already prescribed. Minimising (3.119) with respect to $\dot{s}_j$ leads to a tridiagonal matrix which can easily be solved using a simple tridiagonal solver (Golub & Van Loan (1983)) to find $\dot{s}_j$ for all nodes. Then $\dot{s_j}$ is then substituted into equation (3.116), which is minimised with respect to $\dot{a_j}$, to give another tridiagonal system. This again can be solved using the same simple tridiagonal matrix solver. Consequently the initial equations (3.116), (3.118) are solved separately, and hence the name 'the split method'.

3.11.1 Singularities For The Split Method

In section 3.5.1 we have shown that the local and global MFE methods can pass through the singularity as the solution initially overturns. In the split method only one matrix occurs for the two minimisations ((3.119), (3.120)). The matrix elements are all dependent on two Δs's and from this it can be seen that if only one $\Delta s \to 0$, then the matrix is non-singular and hence the split method allows the solution to pass through the singularity.

3.12 Lagrangian Methods

Lagrangian methods use the idea of particles which move with constant 'mass' and implements them numerically. The particle idea is as follows; if a particle P is within a region R under the influence of some external force, then we follow the particle throughout the region.

In order to solve an equation of the form

$$u_t + a(u)u_x = 0 \tag{3.121}$$

or

$$\dot{u} - u_x \dot{x} + a(u)u_x = 0. \tag{3.122}$$

We set

$$\dot{u} = 0 \tag{3.123}$$

from which it follows that

$$\dot{x} = a(u). \tag{3.124}$$

For conservation laws $\dot{u} = 0$, and u is constant along certain particle paths. It should be noted that the Lagrangian method and the method of characteristics are the same analytically for conservation laws. However this no longer remains true when numerical methods are considered. The difference occurs in the discretisation of the equation which in the MFE case involves L_2 projections and in the Lagrangian case is pointwise at a node. However, the two methods come together for the inviscid Burgers' equation.

The problems normally associated with the Lagrangian method relate to tangling of the mesh because overturning is not allowed. The consequent reduction of the time-step so that no nodes overtake each other often causes problems but here we again allow the method to follow the paths past overturning so that we may replace the overturned curve by a discontinuity. This approach does not suffer from the problems normally associated with Lagrangian methods.

3.13 Boundary Conditions

The boundary required for this type of problem may be either moving or fixed. If we first consider fixed conditions, (Dirichlet or Neumann) then as time passes this may lead to regions near the boundary where there are very few nodes. This can occur because all the nodes within the region move with speed $\dot{x} = a(u)$ for conservation laws and the boundary nodes are fixed with $\dot{x} = 0$. It is possible to get round this problem by adding nodes at boundary A and removing nodes at boundary B, but the approach is rather complicated.

The moving boundaries that we will consider are where the nodes move with the characteristic speed. This causes a problem in that we are no longer looking

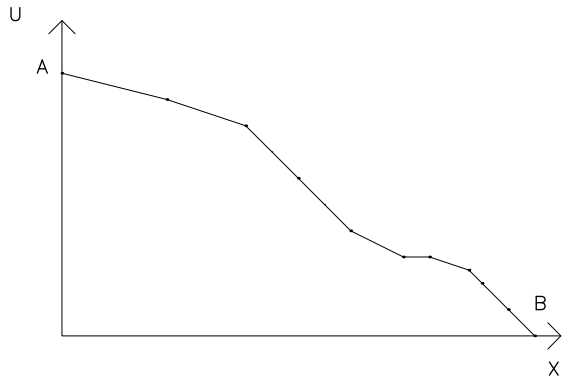

Figure 3.7: Fixed boundary.

at a fixed region, however the problem of too few nodes near one of the boundaries now disappears.

3.14 Solution In VM Space

In chapter 2 section 2.8, a Legendre transformation between the variables u, x and v, m was described. It is defined by

$$m = \frac{\partial u}{\partial x} \qquad x = \frac{\partial v}{\partial m} \tag{3.125}$$

for which

$$u(x) - mx + v(m) = 0. \tag{3.126}$$

The aim of this transformation is to simplify the equations to be solved, however it may also be significant numerically. Consider the equation

$$u_t - \mathcal{L}(u) = 0 \tag{3.127}$$

and rewrite this in Lagrangian form as

$$\dot{u} - u_x \dot{x} - \mathcal{L}(u) = 0. \tag{3.128}$$

Equation (3.126) may also be differentiated with respect to time to give

$$\dot{u} - \dot{m}x + m\dot{x} + \dot{v} = 0 \tag{3.129}$$

which can then be combined with (3.128). This results in

$$-\dot{v} + v_m \dot{m} - \mathcal{L}(mv_m - v) = 0. \tag{3.130}$$

In Burgers' Equation $\mathcal{L}(u) = -uu_x$ which, when substituted into (3.130) and the coefficients of v_m compared, gives equations

$$\dot{m} = -m^2 \quad \dot{v} = -mv. \tag{3.131}$$

Numerically, $-\dot{V}+V_M\dot{M}$ is piecewise linear discontinuous and the term $\mathcal{L}(mv_m - v)$ needs to be projected into the space spanned by the basis functions $\phi_k^{(1)}$, $\phi_k^{(2)}$. Equivalently the projection can be made into the space spanned by the set $\{1, X\}$ or $\{1, V_M\}$, using the usual L_2 minimisation. This may be carried out on each element k so that $\mathcal{L}(u)$ may be represented by the straight line $AX + B$ or $AV_M + B$. On each element A, B are found from

$$\min_{A,B} \|AV_M + B - \mathcal{L}(MV_M - V)\| \tag{3.132}$$

in the finite dimensional space V, M. Equation (3.130) now becomes

$$-\dot{V} + V_M\dot{M} + AV_M + B = 0 \tag{3.133}$$

where $V_M = X$. Comparing coefficients of V_M (the linear terms) gives the equations

$$\dot{M} = -A \quad \dot{V} = B. \tag{3.134}$$

Here there are no singularities whatsoever. These equations may be solved using an ODE solver such as Euler or Crank-Nicolson.

Once these 'simple' equations have been solved, a numerical Legendre transformation is necessary in order to obtain the solution in the original variables. Note that it is within the inverse transformation that the parallelism singularity occurs.

3.15 Summary

In this chapter the basic ideas of the moving finite element and the Lagrangian method have been introduced together with some extensions. In chapter 2 a method was proposed which involved the calculation of multivalued curves following which a second method of recovery would be applied in order to obtain the solution with shocks. Unfortunately the methods described above (excepting GWMFE and the Lagrangian method) are not able to be used to calculate

overturned solutions because of the use of the L_2 norm. The next chapter will extend the ideas of MFE described here in order to allow multivalued solutions and numerical recovery of shocks.

Chapter 4

Norms And Overturning Solutions

4.1 Introduction

In chapters 1 and 2 the analytic solution of certain nonlinear PDE'S, particularly conservation laws, was discussed. One of the main ideas involved was allowing the characteristics to cross, thus giving a multivalued solution, then recovering the shock position from this overturned manifold. In chapter 3 several adaptive finite element methods were introduced and we propose to use these to obtain overturned solutions numerically. However, for the methods based on the standard MFE approach, there is a problem with the minimisation of the L_2 norm once the solution overturns (Sweby (1990)). The problem occurs because the expression minimised in MFE does not remain positive definite and hence is not a norm once the solution has overturned.

There are three sections within this chapter. Sections 4.2-4.3 are concerned with writing the MFE method in terms of different norms which remain valid even when the solution curve becomes multivalued. In section 4.4 several moving finite element based methods are described, and the effects of the use of the different norms is discussed. Section 4.5 is a description of the methods applied to the overturned curves. Once an overturned numerical solution has been calculated then the shock position may also be found numerically. We describe several methods for this purpose which include some based upon conservation arguments and another proposed by Brenier (see chapter 2 section 2.10), which gives an approximate shock position.

4.2 A Two-Stage Procedure

In order to solve for scalar nonlinear PDE's by allowing the solution to become multivalued using the standard moving finite element method, the norm of the residual to be minimised (see chapter 3) must (by definition) remain positive definite throughout the region of solution. It has already been noted (Sweby (1990)) that once the solution becomes multivalued, this is no longer true since the integration 'reverses'. In order to continue using the norm of the residual to define the MFE method, it must be redefined in such a way that it remains positive definite.

One method of forcing the norm to remain valid (i.e. positive definite) is to integrate with respect to arc length or take note of the sign of the solution curve so that the integral will remain positive. For example, instead of simply integrating with respect to x it would be possible to integrate with respect to x with a weight function $\frac{s_x}{|s_x|}$ where s is the arc length. Using this method it would be necessary to keep track of whether the integration was over a single or multivalued curve, which would consequently make the calculations complicated.

In this section we will describe how the minimisation of the L_2 norm used in the MFE procedure may be rewritten as two separate minimisations involving norms which always remain positive definite even after overturning (Baines & Reeves (1990)). Much of what follows holds in any number of dimensions, unlike other sections of this thesis. Let us first introduce some notation.

Let u be a continuously differentiable function of the space variables x and time t, where $u = u(x,t) \in \Omega \times (0, t_1)$, where Ω is a polygonal region contained in $\mathbb{R}$, and t_1 is a fixed positive time. We shall consider the differential equation

$$u_t - \mathcal{L}(u) = 0 \tag{4.1}$$

introduced in chapter 3, where $\mathcal{L}(.)$ is a first order operator in the space variables.

Now define finite dimensional approximations to u and u_t, for all t, to be $U \in S$ and $U_t \in T$. Here U and U_t are piecewise linear functions on Π which is a partition of Ω with linear facets, for example line segments, (in higher dimensions Π will consist of triangles, tetrahedra etc.) and S, T are (generally distinct) linear spaces of piecewise linear functions.

Assuming that $\mathcal{L}(.)$ is a first order operator and that u is continuously differentiable, then $\mathcal{L}(U)$ exists and is continuous in Ω , except possibly at internal boundaries of the partition Π. If it is also assumed that U is continuously differentiable with respect to t, then

$$U_t - \mathcal{L}(U) \tag{4.2}$$

exists and is square-integrable over $\Omega, \forall t \in (0, t_1)$. Note that, although u satisfies (4.1), in general its approximation U will not. The residual R may be defined by

$$R = U_t - \mathcal{L}(U). \tag{4.3}$$

From chapter 3, it can be seen that in the MFE procedure (4.3) is minimised over all $U_t \in T$ using the L_2 norm. However, as has already been discussed, this procedure will lead to the minimisation of an invalid norm when the solution becomes multivalued. In order to redefine the norm in this minimisation in terms of norms which remain valid, some further notation must be introduced.

Consider the space S^* of piecewise linear discontinuous functions on the partition Π and let R^* be the L_2 projection of R into S^* (Miller (1988)). See Fig 4.1.

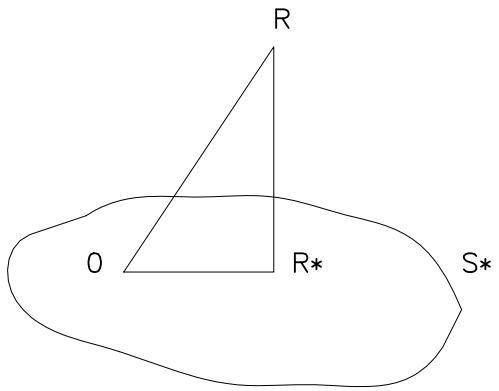

Figure 4.1: L_2 projection of R into S^*.

From the definition of the L_2 norm ((3.20)), $R^* \in S^*$ minimises

$$\|R - R^*\| \tag{4.4}$$

which implies that (see e.g. Johnson & Riess (1982))

$$< R - R^*, R^* >= 0. \tag{4.5}$$

Using (4.4),(4.5), the norm squared of the residual R can be written as the sum of two norms squared, as follows,

$$\begin{aligned} \|R\|^2 &= < R, R > \\ &= < R - R^* + R^*, R - R^* + R^* > \\ &= < R - R^*, R - R^* > + < R^*, R^* > +2 < R - R^*, R^* > \\ \Rightarrow \|R\|^2 &= \|R - R^*\|^2 + \|R^*\|^2. \end{aligned} \tag{4.6}$$

The minimisation therefore can be regarded as two successive (orthogonal) projections. Although the square of the L_2 norm has now been written as a sum of squares of two norms, defining two separate stages, the problem of the norm becoming invalid as the solution becomes multivalued still remains. The first stage presents no difficulty when the nodes overtake since the matrices in the elementwise projection are always positive definite, but the second stage remains a problem since here the integral of $\|R^*\|^2 = \int R^{*2} dx$ changes sign in the event of overturning.

Let us however consider writing $\|R^*\|$ as an l_2 norm using a coordinate system within S^*. The l_2 norm is a discrete version of the L_2 norm. If $\mathbf{u}, \mathbf{v}$ are in a finite dimensional vector space where $\mathbf{u} = (u_1, ..., u_n)$ and $\mathbf{v} = (v_1, ..., v_n)$ then the l_2 inner product is defined by

$$< \mathbf{u}, \mathbf{v} >= \sum_i u_i v_i W_i \tag{4.7}$$

where $\mathbf{W} = (W_1, ..., W_n)$ is a weight vector and the l_2 norm given by

$$\|\mathbf{u}\|_{l_2} =< \mathbf{u}, \mathbf{u} > . \tag{4.8}$$

In stage 2 the minimisation of $\|R^*\|^2$ means that we want to find $U_t \in T$ such that

$$R^* = U_t - \mathcal{L}(U)^* \tag{4.9}$$

is minimised in the L_2 norm, $\mathcal{L}(U)^*$ being the L_2 projection of $\mathcal{L}(U)$ into S^*. However since $U_t \in T$, $\mathcal{L}(U) \in S^*$, and T, S^* are both finite dimensional spaces of the same dimension, this minimisation may be rewritten as the minimisation of the l_2 norm of a vector of coordinates of (4.9). In the overturning case this

avoids the difficulty of the integral changing sign (see below) because the l_2 norm is defined algebraically with no appeal to any sense of direction.

In order to write $\mathcal{L}(U)^* \in S^*$ and $U_t \in T$ explicitly as a sum of vector coordinates, we need to introduce sets of basis functions for S^* and T. Let $\{\phi_i\}$ and $\{\delta_i\}$ be sets of basis functions which span the spaces S^* and T respectively. Let also $\{w_i\}$ and $\{q_i\}$ be the corresponding sets of coefficients for the functions $\mathcal{L}(U)^* \in S^*$ and $U_t \in T$, i.e.

$$\mathcal{L}(U)^* = \sum_i w_i \phi_i, \qquad U_t = \sum_i q_i \delta_i. \tag{4.10}$$

Since $T \subseteq S^*$, the basis functions δ_i of T may be written in terms of $\{\phi_i\}$,

$$\delta_i = \sum_j \mu_{ij} \phi_i \tag{4.11}$$

say, where μ_{ij} are coefficients.

Consequently U_t may be rewritten in terms of the basis functions of S^*,

$$\begin{aligned} U_t = \sum_i q_i \delta_i &= \sum_i q_i \sum_j \mu_{ij} \phi_j \\ &= \sum_i \sum_j q_i \mu_{ij} \phi_j \\ &= \sum_i \sum_j q_j \mu_{ji} \phi_i. \end{aligned} \tag{4.12}$$

The residual of (4.9) may now be written in terms of the basis functions of S^* so that

$$\begin{aligned} R^* &= \sum_i \sum_j q_j \mu_{ji} \phi_i - \sum_i w_i \phi_i \\ &= \sum_i (\sum_j q_j \mu_{ji} - w_i) \phi_i. \end{aligned} \tag{4.13}$$

Hence the norm squared of R^* becomes

$$\|R^*\|_{l_2}^2 = \| \sum_i (\sum_j q_j \mu_{ji} - w_i) \phi_i \|^2. \tag{4.14}$$

This norm is to be minimised over all U_t, i.e. over the coefficients of U_t. The expression (4.14) may now be rewritten using the definition of the l_2 norm as

$$\begin{aligned} \|R^*\|_{l_2}^2 &= < \sum_i (\sum_j q_j \mu_{ji} - w_i) \phi_i, \sum_k (\sum_l q_l \mu_{lk} - w_k) \phi_k > \\ &= \sum_i \sum_k (\sum_j q_j \mu_{ji} - w_i)(\sum_l q_l \mu_{lk} - w_k) < \phi_i, \phi_k > \end{aligned} \tag{4.15}$$

which is a new finite dimensional l_2 norm of the coordinates of R^* (see (4.13)) unaffected by overturning. Since μ_{ji} $(\forall ji)$ and ϕ_i $(\forall i)$ are known, and we want to minimise (4.15) over $\{q_i\}$, then (4.15) is a discrete quadratic form for the unknown q's, in terms of the w's which are obtained from the first stage, that of minimising $\|R - R^*\|$ (see (4.6)).

Returning then to the first stage of (4.6), since

$$R = U_t - \mathcal{L}(U) \tag{4.16}$$

and, because U_t is already in S^*,

$$R^* = U_t - \mathcal{L}(U)^*, \tag{4.17}$$

subtracting (4.17) from (4.16) gives

$$R - R^* = -\mathcal{L}(U) + \mathcal{L}(U)^*. \tag{4.18}$$

The first stage of the minimisation can now be seen to be to find $\mathcal{L}(U)^* \in S^*$ such that $\| - \mathcal{L}(U) + \mathcal{L}(U)^*\|$ is minimised, i.e. to find the projection $\mathcal{L}(U)^*$ of $\mathcal{L}(U)$ into S^*. (See Fig. 4.2).

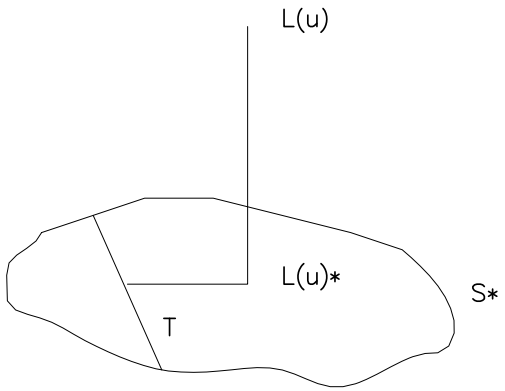

Figure 4.2: Projection of $\mathcal{L}(u)$ into the space S^*.

Now applying the notation of (4.10), the L_2 norm squared of (4.18) becomes

$$\|R - R^*\|^2 = \| - \mathcal{L}(U) + \sum_i w_i \phi_i\|^2 \tag{4.19}$$

which, when minimised, gives a set of linear equations in $\{w_i\}$ (c.f. (3.37)). When the $\{w_i\}$ from (4.19) have been found, they may be substituted into (4.15), and the $\{q_i\}$ can be found. Hence using (4.10), U_t can be calculated giving the nodal speeds.

4.3 Norms

By redefining the inner product and norm for the second stage, a new set $\{\phi_i\}$ of basis functions may be chosen so that another two-stage method may be obtained which in the non-overturning case is equivalent to the local method (Miller (1988)), (Baines & Wathen (1988)). Let $\{\phi_i\}$ be any set of linear discontinuous basis functions such that ϕ_i is zero except on element i of Π. This results in the set of linear equations for the w's (first stage) decoupling into separate element by element sets of 2 equations as in the global or local method. However the second stage is still coupled nodewise in q's, so a new norm is defined which will allow these equations to separate also.

The new inner-product $((.,.))$ is defined by

$$((\phi_i, \phi_j)) = \begin{cases} < \phi_i, \phi_j > & i = j \\ 0 & i \neq j \end{cases} \tag{4.20}$$

and the norm is defined by

$$|||\phi_i|||^2 = ((\phi_i, \phi_i)). \tag{4.21}$$

If we now replace the l_2 inner-product in (4.15) by (4.20) to give

$$\|R^*\|_d^2 = \sum_i \sum_k (\sum_j q_j \mu_{ji} - w_i)(\sum_l q_l \mu_{lk} - w_k)((\phi_i, \phi_k)) \tag{4.22}$$

we obtain

$$\|R^*\|_d^2 = \sum_i \sum_k (\sum_j q_j \mu_{ji} - w_i)^2 < \phi_i, \phi_k > \tag{4.23}$$

$$= \sum_i (\sum_j q_j \mu_{ji} - w_i) \|\phi_i\|^2. \tag{4.24}$$

This suggests the definition of a so-called local norm $|||.|||$ (Miller (1988)) to be written as

$$|||R|||^2 = \|R - R^*\|^2 + \|R^*\|_d^2 \tag{4.25}$$

which corresponds to the two stage method given in (Baines & Wathen (1988)). See Fig. 4.3.

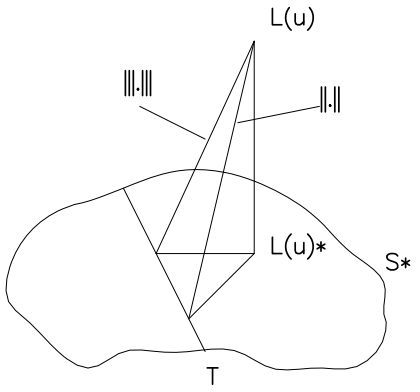

Figure 4.3: Projections for global $\|.\|$ and local $|||.|||$ MFE.

4.4 Implementation Of The Methods

Although MFE has been described in chapter 3, we consider it again from the point of view of a 1 step L_2 projection and 1 step l_2 projection. What follows is valid for both overturned and non-overturned solutions. The first stage remains the same as in chapter 3. The second stage is algebraic.

The implementation of the one-stage and two-stage methods demonstrates how various versions of MFE give rise to sets of linear equations. The methods are applied as before to the equation

$$u_t - \mathcal{L}(u) = 0 \tag{4.26}$$

where $u = u(x,t)$ and $\mathcal{L}(u)$ contains only x, u and first derivatives of u. (For convenience the methods are described here for the 1-D case although they are also valid for the 2-D case, for which see chapter 8)

4.4.1 Global Method

For this case we describe only the implementation of the two-stage method for multivalued solutions since the 1-stage method is already described in chapter 3 section 3.2 and (Miller & Miller (1981)).

From the ideas of Miller and Carlson, (see e.g. Baines (1986)), the approximation to u may be written as a linear combination of local elementwise basis functions $\phi_k^{(1)}, \phi_k^{(2)}$ in element k, which is bounded by nodes $j-1, j$. (See chapter

3, section 3.4). Let u be approximated by $\tilde{U}_k$ in element k where

$$\tilde{U}_k(\xi,\tau) = a_{j-1}(\tau)\phi_k^{(1)}(\xi) + a_j(\tau)\phi_k^{(2)}(\xi), \tag{4.27}$$

ξ being a reference variable and with $\tau = t$. Also let the x in element k be approximated by

$$\tilde{X}_k(\xi,\tau) = s_{j-1}(\tau)\phi_k^{(1)}(\xi) + s_j(\tau)\phi_k^{(2)}(\xi). \tag{4.28}$$

Using the chain rule, we get

$$\begin{aligned} \frac{\partial}{\partial\tau} &= \frac{\partial t}{\partial\tau}\frac{\partial}{\partial t} + \frac{\partial\tilde{X}}{\partial\tau}\frac{\partial}{\partial\tilde{X}} \\ &= \frac{\partial}{\partial t} + (\dot{s}_{j-1}\phi_k^{(1)} + \dot{s}_j\phi_k^{(2)})\frac{\partial}{\partial\tilde{X}} \end{aligned} \tag{4.29}$$

where the dot notation indicates differentiation with respect to τ. Hence U_t becomes

$$\begin{aligned} \frac{\partial\tilde{U}_k}{\partial t} &= \dot{a}_{j-1}\phi_k^{(1)} + \dot{a}_j\phi_k^{(2)} - \tilde{U}_X(\dot{s}_{j-1}\phi_k^{(1)} + \dot{s}_j\phi_k^{(2)}) \\ &= (\dot{a}_{j-1} - m_k\dot{s}_{j-1})\phi_k^{(1)} + (\dot{a}_j - m_k\dot{s}_j)\phi_k^{(2)} \end{aligned} \tag{4.30}$$

where m_k is the gradient U_x in the kth element.

Alternatively, writing

$$\begin{aligned} w_k^{(1)} &= \dot{a}_{j-1} - m_k\dot{s}_{j-1} \\ w_k^{(2)} &= \dot{a}_j - m_k\dot{s}_j \end{aligned} \tag{4.31}$$

we have

$$\frac{\partial\tilde{U}_k}{\partial t} = w_k^{(1)}\phi_k^{(1)} + w_k^{(2)}\phi_k^{(2)}. \tag{4.32}$$

The two forms (4.30) and (4.32) are those used by Miller and Carlson (see Baines (1986)) and by (Baines & Wathen (1988)) respectively.

Stage 1

We first minimise

$$\left\| \frac{\partial\tilde{U}_k}{\partial t} - \mathcal{L}(\tilde{U}_k) \right\| \tag{4.33}$$

over each element k, with respect to $w_k^{(1)}, w_k^{(2)}$ using (4.31), obtaining the system

$$\begin{aligned} < \phi_k^{(1)}, \frac{\partial \tilde{U}_k}{\partial t} - \mathcal{L}(\tilde{U}_k) > &= 0 \\ < \phi_k^{(2)}, \frac{\partial \tilde{U}_k}{\partial t} - \mathcal{L}(\tilde{U}_k) > &= 0 \end{aligned} \tag{4.34}$$

which gives two equations in two unknowns for each element k. The system is non-singular in general and, using (4.32), can be written in the form

$$C_k \mathbf{w}_k = \mathbf{b}_k \tag{4.35}$$

where

$$\mathbf{w}_k = \begin{pmatrix} w_k^{(1)} \\ w_k^{(2)} \end{pmatrix}, \qquad C_k = \frac{\Delta s_k}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \tag{4.36}$$

$\Delta s_k = s_j - s_{j-1}$, and

$$\mathbf{b}_k = \begin{pmatrix} < \phi_k^{(1)}, \mathcal{L}(\tilde{U}_k) > \\ < \phi_k^{(2)}, \mathcal{L}(\tilde{U}_k) > \end{pmatrix}. \tag{4.37}$$

The square of the norm (4.33) can equally be written as

$$\mathbf{w}_k^T C_k \mathbf{w}_k - 2\mathbf{w}_k^T \mathbf{b}_k + \|\mathcal{L}(\tilde{U}_k)\|^2 \tag{4.38}$$

and minimisation also leads to $C_k \mathbf{w}_k = \mathbf{b}_k$.

Alternatively, following Miller and Carlson, we may minimise (4.33) over element k with respect to $\dot{a}_k^{(1)}, \dot{a}_k^{(2)}, \dot{s}_k^{(1)}, \dot{s}_k^{(2)}$ where $\tilde{U}_k = \dot{a}_k^{(1)}\phi_k^{(1)} - m_k \dot{s}_k^{(1)}\phi_k^{(1)} + \dot{a}_k^{(2)}\phi_k^{(2)} - m_k \dot{s}_k^{(2)}\phi_k^{(2)}$, which leads to the double system

$$\begin{aligned} < \phi_k^{(1)}, \frac{\partial \tilde{U}_k}{\partial t} - \mathcal{L}(\tilde{U}_k) > &= 0 \\ < \phi_k^{(2)}, \frac{\partial \tilde{U}_k}{\partial t} - \mathcal{L}(\tilde{U}_k) > &= 0 \\ < -m_k \phi_k^{(1)}, \frac{\partial \tilde{U}_k}{\partial t} - \mathcal{L}(\tilde{U}_k) > &= 0 \\ < -m_k \phi_k^{(2)}, \frac{\partial \tilde{U}_k}{\partial t} - \mathcal{L}(\tilde{U}_k) > &= 0. \end{aligned} \tag{4.39}$$

This gives four equations in four unknowns for each element k. Since m_k is constant, the system is singular. However, considering all elements together we find that values of $\dot{a}_j, \dot{s}_j$ are defined on both sides of each node, i.e. from element $k-1$ we have $\dot{a}_{k-1}^{(2)}, \dot{s}_{k-1}^{(2)}$ and from element k we have $\dot{a}_k^{(1)}, \dot{s}_k^{(1)}$, where for continuity these need to be equal. To obtain this continuity and also to enforce boundary

conditions, constraints must be applied. The result is an assembly of the sets (4.39) to give a non-singular system.

Following (Baines (1986)), the singular system (4.39) can be written as

$$E_k \dot{\mathbf{y}}_k = \mathbf{G}_k \tag{4.40}$$

where

$$\dot{\mathbf{y}}_k = \begin{pmatrix} \dot{a}_k^{(1)} \\ \dot{s}_k^{(1)} \\ \dot{a}_k^{(2)} \\ \dot{s}_k^{(2)} \end{pmatrix} \qquad E_k = \frac{\Delta s_k}{6} \begin{pmatrix} 2\mathbf{m}_k\mathbf{m}_k^T & \mathbf{m}_k\mathbf{m}_k^T \\ \mathbf{m}_k\mathbf{m}_k^T & 2\mathbf{m}_k\mathbf{m}_k^T \end{pmatrix} \tag{4.41}$$

$$\mathbf{m}_k = \begin{pmatrix} 1 \\ -m_k \end{pmatrix} \tag{4.42}$$

and

$$\mathbf{G}_k = \begin{pmatrix} < \phi_k^{(1)}, \mathcal{L}(\tilde{U}_k) > \\ < \phi_k^{(2)}, \mathcal{L}(\tilde{U}_k) > \\ < -m_k \phi_k^{(1)}, \mathcal{L}(\tilde{U}_k) > \\ < -m_k \phi_k^{(2)}, \mathcal{L}(\tilde{U}_k) > \end{pmatrix}. \tag{4.43}$$

Note: E_k is a 4×4 matrix given by blocks of 2×2 matrices. The square of the norm (4.33) can then be written as

$$\dot{\mathbf{y}}_k^T E_k \dot{\mathbf{y}}_k - 2\dot{\mathbf{y}}_k \mathbf{G}_k + \|\mathcal{L}(\tilde{U}_k)\|^2 \tag{4.44}$$

and minimisation gives

$$E_k \dot{\mathbf{y}}_k = \mathbf{G}_k, \tag{4.45}$$

which is however a singular system unless the constraints of stage 2 are applied.

Stage 2

In stage 2 we work only with the coordinates $\mathbf{w}_k$ satisfying (4.35) or $\dot{\mathbf{y}}_k$ satisfying (4.40). In the case of the w's we need to implement the minimisation of (4.12) with the μ's given by

$$\mu_{ij} = \begin{cases} 1 & j = i \\ -m_j & j = i+1 \\ 0 & \text{otherwise} \end{cases} \tag{4.46}$$

(in the 1-D version). This gives the standard form of the MFE equations

$$M^T C M \dot{\mathbf{y}} = M^T C \mathbf{b} = \mathbf{g} \tag{4.47}$$

say, where $C = \text{diag}\{C_k\}$, $M = \text{diag}\{M_j\}$,

$$M_{j-1} = \begin{pmatrix} 1 & -m_{k-1} \\ 1 & -m_k \end{pmatrix} \tag{4.48}$$

and

$$\dot{\mathbf{y}} = (...; \dot{a}_j, \dot{s}_j; ...)^T. \tag{4.49}$$

Returning to (4.39), and applying the continuity constraints on $\dot{a}, \dot{s}$, also gives rise to the standard global MFE equations, as follows. Let

$$\dot{\mathbf{Y}}_{j-1} = \begin{pmatrix} \dot{a}_{k-1}^{(2)} \\ \dot{s}_{k-1}^{(2)} \\ \dot{a}_k^{(1)} \\ \dot{s}_k^{(1)} \end{pmatrix} = R_{j-1} \begin{pmatrix} \dot{a}_{j-1} \\ \dot{s}_{j-1} \end{pmatrix} \tag{4.50}$$

where

$$R_{j-1} = \begin{pmatrix} I_2 \\ I_2 \end{pmatrix}, \qquad I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{4.51}$$

Then over the whole system we obtain

$$\dot{\mathbf{Y}} = R \dot{\mathbf{y}} \tag{4.52}$$

where

$$R = \begin{pmatrix} R_1 & & \\ & \ddots & \\ & & R_n \end{pmatrix} \tag{4.53}$$

and

$$\dot{\mathbf{Y}} = \begin{pmatrix} \vdots \\ \dot{\mathbf{Y}}_j \\ \dot{\mathbf{Y}}_{j+1} \\ \vdots \end{pmatrix}. \tag{4.54}$$

The sum of the squares of the norm (4.33) over k may now be written

$$\dot{\mathbf{Y}}^T E \dot{\mathbf{Y}} - \dot{\mathbf{Y}}^T \mathbf{G} = \dot{\mathbf{y}}^T R^T E R \dot{\mathbf{y}} - 2 \dot{\mathbf{y}}^T R^T \mathbf{G} \tag{4.55}$$

and, minimising this expression over $\dot{\mathbf{y}}$ yields

$$R^T E R \dot{\mathbf{y}} = \mathbf{g} \tag{4.56}$$

where

$$E = \text{diag}\{E_k\}, \qquad \mathbf{g} = R^T \mathbf{G}, \qquad \mathbf{G} = \{\mathbf{G}_k\}. \tag{4.57}$$

Note that applying constraints to (4.35) is equivalent to

$$\min_{\dot{s}_{j-1}, \dot{a}_{j-1}} \left\| \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \dot{a}_{j-1} \\ \dot{s}_{j-1} \end{pmatrix} - \begin{pmatrix} \dot{a}^{(2)}_{k-1} \\ \dot{s}^{(2)}_{k-1} \\ \dot{a}^{(1)}_{k} \\ \dot{s}^{(1)}_{k} \end{pmatrix} \right\} W_{j-1} \right\| \tag{4.58}$$

where $W = (W_1, ..., W_n)$ is a weight function (matrix). By the appropriate choice of W, we can obtain either the global or local method. Global corresponds to $W = E^{\frac{1}{2}}$, local to $W = E_d^{\frac{1}{2}}$, where $E_d = \text{Diag}\{E\}$ (Diag is the diagonal of the matrix).

It is shown in (Baines (1986)) that (4.56) gives rise to the usual MFE system since

$$\begin{aligned} R^T E R &= R^T \bar{M}^T C \bar{M} R & (4.59) \\ &= M^T C M & (4.60) \\ &= A & (4.61) \end{aligned}$$

where

$$E_k = \bar{M}_k^T C_k \bar{M}_k, \quad \bar{M}_k = \begin{pmatrix} 1 & -m_k & 0 & 0 \\ 0 & 0 & 1 & -m_k \end{pmatrix} \tag{4.62}$$

$$\bar{M} = \text{diag}\{\bar{M}_k\}, \quad M = \bar{M} R, \quad E = \bar{M}^T C \bar{M}. \tag{4.63}$$

Note that the difference between the matrix decompositions (4.59) and (4.60) is essentially in the use of nodal variables (in E) rather than element variables (in C).

4.4.2 Local Method

The one-stage method has already been described in chapter 3 section 3.4. The two-stage local and global methods are identical in 1-D, whereas in higher dimensions only the first stage is the same (see chapter 8). The matrix system obtained

by applying the 2-stage method to global MFE gives the system (4.56). In 1-D $M = \bar{M}R$ is square and if it is also non-singular we may proceed as follows

$$\begin{aligned} R^T\bar{M}^T C\bar{M}R\dot{\mathbf{y}} &= \mathbf{g} \\ C\bar{M}R\dot{\mathbf{y}} &= (R^T\bar{M}^T)^{-1}\mathbf{g} \\ D\bar{M}R\dot{\mathbf{y}} &= DC^{-1}(R^T\bar{M}^T)^{-1}\mathbf{g} \\ R^T\bar{M}^T D\bar{M}R\dot{\mathbf{y}} &= R^T\bar{M}^T DC^{-1}(R^T\bar{M}^T)^{-1}\mathbf{g} \end{aligned} \tag{4.64}$$

where $D = \text{Diag}\{C\}$, so we get

$$\begin{aligned} M^T DM\dot{\mathbf{y}} &= M^T DC^{-1}M^{-1}\mathbf{g} \\ M^T DM\dot{\mathbf{y}} &= M^T D\mathbf{b} \end{aligned} \tag{4.65}$$

which is the local method. This is equivalent to letting $W = E_d^{\frac{1}{2}}$ in (4.58).

4.4.3 Split Method

The split method (Baines (1991)) is based upon the standard MFE method, consequently if the solution is allowed to overturn the expression $\|\dot{s} - \frac{\partial f}{\partial u_x}\|^2$ is again no longer positive definite. However the split method may also be written as a two stage procedure in order to overcome the problem of the norm becoming invalid. This is done in a way analogous to global MFE.

Stage 1

First minimise (4.4.3) over $\dot{s}$ in each element k,

$$\min_{\dot{s}_k^{(1)}, \dot{s}_k^{(2)}} \left\| \dot{s} - \frac{\partial f}{\partial u_x} \right\|^2 \tag{4.66}$$

where

$$\dot{s_k} = \dot{s}_k^{(1)}\phi_k^{(1)} + \dot{s}_k^{(2)}\phi_k^{(2)} \tag{4.67}$$

$\dot{s}_k^{(1)}, \dot{s}_k^{(2)}$ being the value of $\dot{s}$ at the left hand side and right hand side of the kth element. See Fig. 4.4. Note: $\dot{s_k}$ here play the role of w's elsewhere.

From (4.66) we obtain the system

$$\begin{aligned} &< \dot{s} - \tfrac{\partial f}{\partial u_x}, \phi_k^{(1)} >= 0 \\ &< \dot{s} - \tfrac{\partial f}{\partial u_x}, \phi_k^{(2)} >= 0 \end{aligned} \qquad k = 1, ..., N. \tag{4.68}$$

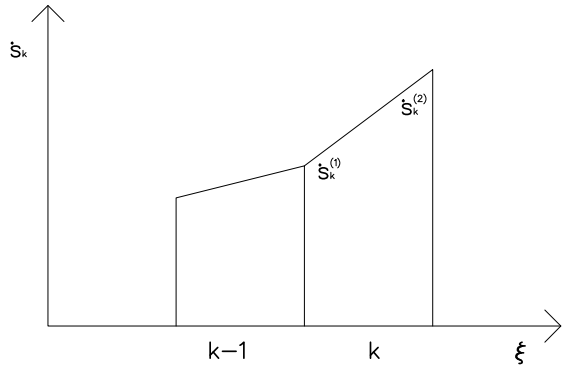

Figure 4.4: Continuous piecewise linear $\dot{s}_k(\xi)$.

This may be written as the 2×2 matrix system

$$\frac{\Delta s_k}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} \dot{s}_k^{(1)} \\ \dot{s}_k^{(2)} \end{pmatrix} = \begin{pmatrix} < \frac{\partial f}{\partial u_x}, \phi_k^{(1)} > \\ < \frac{\partial f}{\partial u_x}, \phi_k^{(2)} > \end{pmatrix} \tag{4.69}$$

whence

$$\begin{pmatrix} \dot{s}_k^{(1)} \\ \dot{s}_k^{(2)} \end{pmatrix} = \frac{2}{\Delta s_k}\begin{pmatrix} 2b_k^{(1)} - b_k^{(2)} \\ -b_k^{(1)} + 2b_k^{(2)} \end{pmatrix} \tag{4.70}$$

where $b_k^{(i)} =< \frac{\partial f}{\partial u_x}, \phi_k^{(i)} > \quad i = 1, 2$. The system (4.68) may be rewritten as

$$E_k \dot{\mathbf{y}}_\mathbf{k} = \mathbf{G}_k \tag{4.71}$$

where

$$E_k = \frac{\Delta s_k}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \tag{4.72}$$

$$\dot{\mathbf{y}}_k = \begin{pmatrix} \dot{s}_k^{(1)} \\ \dot{s}_k^{(2)} \end{pmatrix} \text{ and } \mathbf{G}_k = \begin{pmatrix} b_k^{(1)} \\ b_k^{(2)} \end{pmatrix}. \tag{4.73}$$

However this allows $\dot{s}_{k-1}^{(2)}$ and $\dot{s}_k^{(1)}$ to be unequal whereas a continuous solution $\dot{s}_{j-1}$ is required. This means that constraints must be applied as in (Baines (1986)).

Stage 2

Applying the constraints that $\dot{s}$ is continuous gives

$$\dot{\mathbf{Y}}_k = \begin{pmatrix} \dot{s}_{k-1}^{(1)} \\ \dot{s}_k^{(2)} \end{pmatrix} = R_e \dot{s}_{j-1} \tag{4.74}$$

where $R_e = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, so over the whole system we obtain

$$\dot{\mathbf{Y}} = R_0 \dot{\mathbf{y}} \tag{4.75}$$

where

$$\dot{\mathbf{y}} = \begin{pmatrix} \vdots \\ \dot{s}_{j-1} \\ \dot{s}_j \\ \vdots \end{pmatrix}, R_0 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & & 0 \\ & \ddots & & \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 1 \end{pmatrix}, \dot{\mathbf{Y}} = \begin{pmatrix} \vdots \\ \dot{\mathbf{Y}}_k \\ \vdots \end{pmatrix}. \tag{4.76}$$

Now solve the system (4.75)

$$R_0^T E R_0 \dot{\mathbf{y}} = \mathbf{g} \tag{4.77}$$

where $\mathbf{g} = R_0^T \mathbf{G}$. Using earlier notation this gives

$$R^T \bar{M}^T E \bar{M} R \dot{\mathbf{y}} = \mathbf{g} \tag{4.78}$$

or

$$M_0^T C M_0 \dot{\mathbf{y}} = \mathbf{g} \tag{4.79}$$

where $M_0 = \bar{M} R_0$.

The application of constraints to (4.71) is equivalent to

$$\min_{s_{j-1}} \left\| \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \dot{s}_{j-1} - \begin{pmatrix} \dot{s}_{k-1}^{(2)} \\ \dot{s}_k^{(1)} \end{pmatrix} \right\} W_{j-1} \right\| \tag{4.80}$$

where $W = (W_1, ..., W_N)$ is a weight function (matrix). If W is chosen to be $D^{\frac{1}{2}}$, then a local split method is obtained and if $W = C^{\frac{1}{2}}$, the comparable global method is obtained.

Let $W = D^{\frac{1}{2}}$, then (4.80) gives

$$(1 \quad 1) D_k^{\frac{1}{2}} D_k^{\frac{1}{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \dot{s}_{j-1} = (1 \quad 1) D_k^{\frac{1}{2}} D_k^{\frac{1}{2}} \begin{pmatrix} \dot{s}_{k-1}^{(2)} \\ \dot{s}_k^{(1)} \end{pmatrix} \tag{4.81}$$

where

$$D_k = \begin{pmatrix} \Delta s_{k-1} & 0 \\ 0 & \Delta s_k \end{pmatrix}. \tag{4.82}$$

Hence

$$
\begin{aligned}
(1 \quad 1)\begin{pmatrix} \Delta s_{k-1} & 0 \\ 0 & \Delta s_k \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\dot{s}_{j-1} &= (1 \quad 1)\begin{pmatrix} \Delta s_{k-1} & 0 \\ 0 & \Delta s_k \end{pmatrix}\begin{pmatrix} \dot{s}^{(2)}_{k-1} \\ \dot{s}^{(1)}_{k} \end{pmatrix} \\
\Rightarrow (\Delta s_{k-1} + \Delta s_k)\dot{s}_{j-1} &= \Delta s_{k-1}\dot{s}^{(2)}_{k-1} + \Delta s_{k-1}\dot{s}^{(1)}_{k} \\
\Rightarrow \dot{s}_{j-1} &= \frac{\Delta s_{k-1}\dot{s}^{(2)}_{k-1} + \Delta s_{k-1}\dot{s}^{(1)}_{k}}{(\Delta s_{k-1} + \Delta s_k)}. \qquad (4.83)
\end{aligned}
$$

Hence $\dot{s}_j$ $(j = 1, ..., N)$ can be found. The second equation (3.120) can be solved in a similar manner, i.e. by writing it as a two stage method.

4.5 Calculation Of Shock Position

The numerical methods described above have all been used to obtain multivalued solutions, which corresponds analytically to allowing characteristics to cross (See chapter 2). This however does not give a physical solution, for which a shock is required. The shock position must be calculated separately in order to give a valid physical solution. In order to obtain the positions of the shock throughout the period of solution, the numerical equivalent of the methods described in chapter 2 may be used. Consequently for each instant that the shock position is required, it may be calculated from the appropriate multivalued solution in the same way as in chapter 2.

4.5.1 Calculation Of The Shock Location Using Equal Area Method

One of the ideas described in chapter 2 uses the properties of conservation. If the areas in Fig. 4.5 are equal then this gives the shock position. This may be posed in terms of finding the position of the line so that the sum of the two areas (one taken to be positive, the other negative) becomes zero. In these terms, the shock position may be found using a nonlinear solver. The method chosen to be applied here is bisection. This method may be applied after any time when the solution has become multivalued.

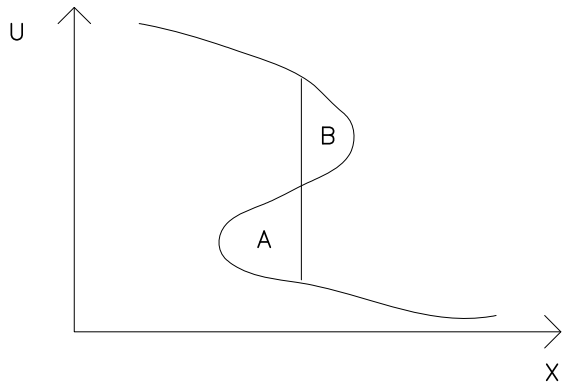

Figure 4.5: Overturned curve.

4.5.2 Calculation Of The Shock Position From The Transformed Equation

Methods which are based on the idea of conservation can be found from the transformations described in chapter 2. Once the transformation to the a, x space (see chapter 2 section 2.7.4) has been made, the shock position may be found from the self intersection of the graph (see Fig. 4.6) (Reeves (1989)). This again uses

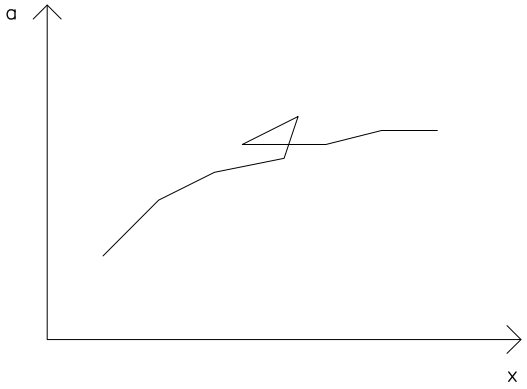

Figure 4.6: Self Intersecting curve.

the equal area principle described in section 4.5.1. Note: in the method described below, although the integral curve is made up of piecewise quadratic elements, for simplicity it is further approximated here by piecewise linears.

Consider two distinct linear segments approximating the curve given by

$$\begin{aligned} y &= p_i x + q_i \\ y &= p_j x + q_j \end{aligned} \qquad i \neq j \tag{4.84}$$

where p_i, p_j, q_i, q_j are easily calculated from the surrounding node positions. The

intersection of the segments occurs when

$$\frac{y - q_i}{p_j} = \frac{y - q_i}{p_i} \tag{4.85}$$

hence

$$y = \frac{q_j p_i - q_i p_j}{p_i - p_j} \tag{4.86}$$

$(p_i \neq p_j)$ so the shock position is given by

$$x = \frac{q_j - q_i}{p_i - p_j}. \tag{4.87}$$

There are two main problems associated with the numerical implementation of this method. The first of these refers to the numerical ill-conditioning of the problem due to segments of similar gradient. Consequently (4.87) is badly conditioned and x may be very large.

Although this problem seems to be serious, in reality the case in which two lines have similar gradients means that the intersection of these segments occurs outside the region of solution so can be ignored.

The second problem involves inappropriate intersections being found. See Fig. 4.7. This may also be avoided by the application of a simple test. If $(x - s_{i-1})(s_i -$

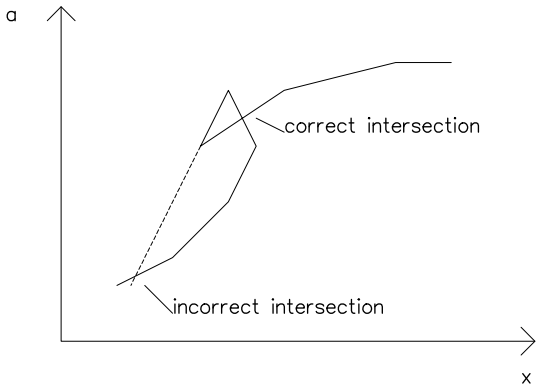

Figure 4.7: Self Intersecting curve.

$x) \geq 0$ and $(x - s_{j-1})(s_j - x) \geq 0$ the intersection is valid. i.e. although in the above calculations the piecewise quadratic curve was approximated by piecewise linears, the method of calculation remains approximately valid.

4.5.3 Note On Piecewise Linear, Piecewise Constant Elements In The Legendre Transform

Note: s is always piecewise linear for continuity in space. If the solution u of the conservation law data u is represented by piecewise constant elements, then a in the Hamilton-Jacobi equation will be represented by piecewise linear elements and b in the ODE's by piecewise constants. This causes no problem since the Legendre transformation is valid for elements of this type (see Baines (1991)).

However it would in general be better to approximate the solution u of the conservation laws by higher order elements since this would be more accurate. Suppose that u in the conservation laws is represented by piecewise linear elements U_k, and s_k (U_k is the approximation to u and s_k is the approximation to x), where

$$U_k = u_1\phi_k^{(1)}(\xi) + u_2\phi_k^{(2)}(\xi) \tag{4.88}$$

and

$$s_k = x_1\phi_k^{(1)}(\xi) + x_2\phi_k^{(2)}(\xi) \tag{4.89}$$

where $\phi_k^{(1)}(\xi)$, $\phi_k^{(2)}(\xi)$ are the usual elementwise basis functions and u_i, x_i $i = 1, 2$ are constants within k.

The transformation to (a, x) space is given by

$$\begin{aligned} a &= \int u dx \\ &= \int (u_1\phi_k^{(1)}(\xi) + u_2\phi_k^{(2)}(\xi))(x_1 d\phi_k^{(1)}(\xi) + x_2 d\phi_k^{(2)}(\xi))d\xi \qquad (4.90) \\ &= \frac{1}{\Delta\xi^2}\left((u_1(\xi_2\xi - \frac{\xi^2}{2}) + u_2(\frac{\xi^2}{2} - \xi_1\xi))(x_2 - x_1)\right) + A. \qquad (4.91) \end{aligned}$$

Similarly the transformation to (b, u) space is given by

$$\begin{aligned} b &= \int x du \qquad (4.92) \\ &= \int (x_1\phi_k^{(1)}(\xi) + x_2\phi_k^{(2)}(\xi))(u_1 d\phi_k^{(1)}(\xi) + u_2 d\phi_k^{(2)}(\xi))d\xi \qquad (4.93) \\ &= \frac{1}{\Delta\xi^2}\left((x_1(\xi_2\xi - \frac{\xi^2}{2}) + x_2(\frac{\xi^2}{2} - \xi_1\xi))(u_2 - u_1)\right) + B \qquad (4.94) \end{aligned}$$

where $\Delta\xi = \xi_2 - \xi_1$ and A, B are constants. We now need to show that the Legendre transformation holds for a, b, so calculate

$$a - ux + b = \frac{1}{\Delta\xi^2}\left((u_1(\xi_2\xi - \frac{\xi^2}{2}) + u_2(\frac{\xi^2}{2} - \xi_1\xi))(x_2 - x_1)\right) + A$$

$$
\begin{aligned}
& - \Big((u_1 \tfrac{(\xi_2 - \xi)}{\Delta\xi} + u_2 \tfrac{(\xi - \xi_1)}{\Delta\xi})(x_1 \tfrac{(\xi_2 - \xi)}{\Delta\xi} + x_2 \tfrac{(\xi - \xi_1)}{\Delta\xi}) \Big) \\
& \tfrac{1}{\Delta\xi^2} \Big((x_1(\xi_2\xi - \tfrac{\xi^2}{2}) + x_2(\tfrac{\xi^2}{2} - \xi_1\xi))(u_2 - u_1) \Big) + B \\
= \quad & A + B + \Big(u_1(\tfrac{x_2}{2} - \tfrac{x_1}{2}) + u_2(\tfrac{x_1}{2} - \tfrac{x_2}{2}) \\
& u_1(-x_1\xi_2^2 + x_2\xi_2\xi_1) + u_2(x_1\xi_1\xi_2 - \xi_1^2 x_2) \Big). \qquad (4.95)
\end{aligned}
$$

Since A and B are constants and the above expression consists of only constants then A and B may always be chosen so that

$$a - ux + b = 0. \qquad (4.96)$$

Hence the Legendre transformation holds (at least locally) when u is piecewise linear and a, b piecewise quadratic.

4.5.4 The Transport Collapse Operator Of Brenier

The description in chapter 2 of the method of calculation of the single valued approximation to the multivalued curve appears complicated (see Brenier (1984)). In practice, this method is easy to implement. This method is applied immediately after the calculation of the solution (or either after the final time-step only) after each time-step.

The algorithm is

1. Calculate initial data.
2. Calculate solution at next time-step by any method which allows multivalued solutions.
3. Calculate Brenier single-valued approximation only if the curve has overturned.
4. Goto 2.

Note: For this method it is possible to either apply the TC operator after each time-step which produces an overturned curve or apply the method only after the final time-step. The difference is that the application after each time-step produces a single-valued solution to which the MFE method may be applied, whereas the method can be applied only once after many overturned time-steps

have been calculated using, say, an MFE method. The two approaches give quantitatively different results (see chapter 5 section 5.4.3).

In order to explain how the Brenier approximation is calculated, first assume that the solution is multivalued. Now divide the curve up into five regions as shown in Fig. 4.8

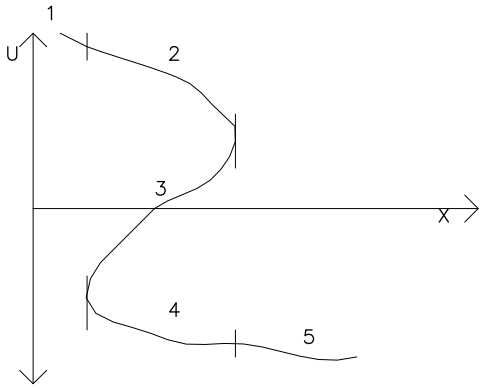

Figure 4.8: Overturned curve and 5 regions.

The Brenier approximation is the same as the solution in regions 1 and 5 where the curve is single-valued.

In region 2 the easiest way to visualise how the approximation is calculated is to add a scaling factor so that the lowest point of the multivalued curve lies above the x-axis. (See Fig. 4.9.) Consider a node j in region 2, and draw a vertical line

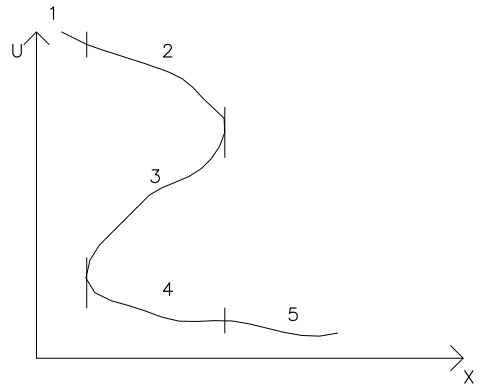

Figure 4.9: Overturned curve scaled.

down through the curve until it intersects the x-axis. The heights of two points on the curve that the line crosses can be found using linear interpolation.

In region 2 the Brenier approximation at node j is,

$$U_{Brenier}(j) = U(j) - \alpha + \beta. \tag{4.97}$$

See Fig. 4.10.

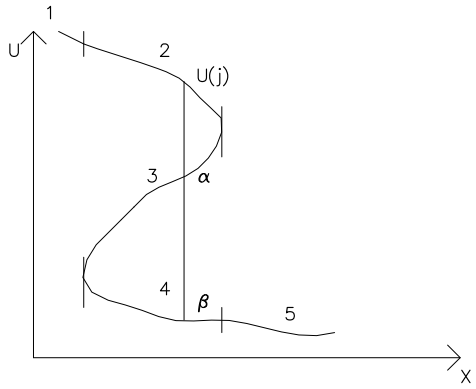

Figure 4.10: Points used in calculation of TC operator.

Following a similar method the Brenier approximation in regions 3 and 4 are

$$U_{Brenier}(j) = \alpha - U(j) + \beta \tag{4.98}$$

$$U_{Brenier}(j) = \alpha + U(j) - \beta \tag{4.99}$$

where α and β are shown in Fig. 4.11.

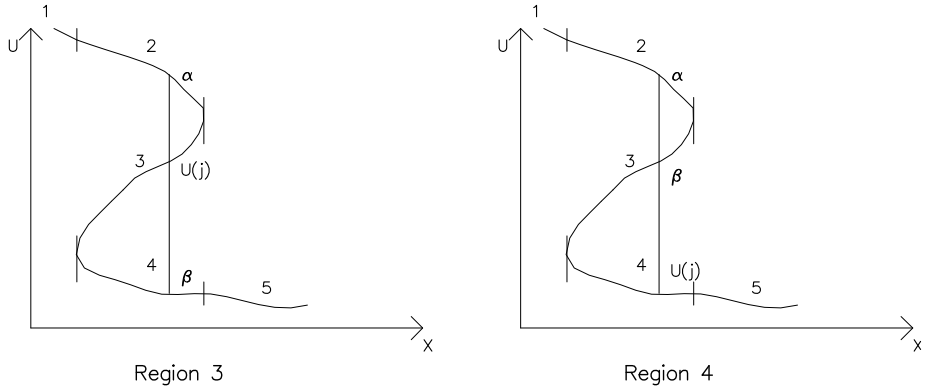

Figure 4.11: Points used in calculation of TC operator.

The next time-step is then applied to the single-valued Brenier approximation. The method is then repeated following the algorithm above until the final time-step is calculated. For more information on implementation see (Brenier (1984)) and (Böing, Werner & Jackisch (1991)).

4.6 Summary

In this chapter the numerical methods of chapter 3 have been extended in order to allow the methods of solution described in chapter 2 to be implemented. The

numerical techniques of shock recovery have also been described. The following chapter will give some numerical examples of the methods discussed here.

Chapter 5

Examples And Results In 1-D

5.1 Introduction

The examples considered here are one-dimensional scalar nonlinear partial differential equations of the form

$$u_t + f_x = 0 \tag{5.1}$$

on a region R where f is a function $f = f(u)$ and $u = u(x,t)$. The main features shown by their solutions as they evolve with time are the formation of shocks and expansions. The nonlinearity of the equations may lead to multivalued solutions (unless physical constraints are imposed). Note that it is important to have the nodes in positions which give a good definition of the initial data curve in order to improve the accuracy of the calculation of the shock position whether it is moving or stationary.

Analytic methods for the solution of (5.1) have been discussed in chapter 2, where the idea of obtaining a multivalued solution by following the characteristics (even when they have crossed) was introduced. Several methods of recovering the shock position from this multivalued curve are given in chapter 2, sections 2.9 and 2.10. The multivalued curves required by the ideas given in chapter 2 led to the use of adaptive finite element methods which are described in chapters 3 and 4. These types of methods are now used to solve a variety of problems of the form (5.1).

From chapter 3 the only direct global methods that have been applied to the examples given below are GWMFE and the Lagrangian method. The other

methods are all two-stage methods which are described in chapter 4. From chapter 4 we have a 2-stage global method, a 2-stage local method and a 2-stage split method. These methods all give very similar results so only a selection of the numerical results will be shown below.

Once the overturned solution has been calculated using a finite element procedure, the shock position may be found by using several alternative techniques. The methods for finding the shock position include two methods based upon the idea of conservation of area. The first of these is calculated using the original variables whereas the other involves a transformation (integration with respect to one variable). A second type of method of obtaining the shock position is the method of (Brenier (1984)), which provides a different kind of approximation to the shock.

The examples chosen are used to demonstrate the difficulties of the problem and show how the proposed methods cope with various aspects. The equations chosen as examples are the inviscid Burgers' equation, the equation $u_t+(u^4/4)_x = 0$ and the Buckley-Leverett Equation in one dimension (see below). These are simple examples of nonlinear conservation laws which will be applied to a variety of initial data in an attempt to show the type of problems that can occur. A Riemann problem using the conservation law $u_t + (u^4/4)_x = 0$ is also given as an example. The two sets of initial data for the Riemann problem allow the formation of a shock and the formation of an expansion. The problems are first described, then the analytic solution is given before the numerical results are shown in sections 5.3 and 5.4.

5.1.1 Problem 1 : Inviscid Burgers' Equation

The equation here is

$$u_t + uu_x = 0 \tag{5.2}$$

$$\text{or} \quad u_t + \left(\frac{u^2}{2}\right)_x = 0 \tag{5.3}$$

on the region $0 \leq x \leq 1$ with three sets of initial data given by

$$a) \quad u = \tanh(5 - 10x) \quad 0 \leq x \leq 1 \tag{5.4}$$

$$b) \quad u = \tanh(5 - 10x) + \tfrac{1}{2} \quad 0 \le x \le 1 \tag{5.5}$$

$$c) \quad u = \begin{cases} 1 & 0 \le x \le \frac{1}{4} \\ \frac{3}{2} - 2x & \frac{1}{4} \le x \le \frac{3}{4} \\ 0 & \frac{3}{4} \le x \le 1 \end{cases} \tag{5.6}$$

and with Dirichlet boundary conditions appropriate to the initial data.

Burgers' equation is given as an example since it is the simplest nonlinear partial differential equation which gives rise to the formation of a shock. Note that (5.3) is a conservation law of the form (5.1) with $f(u) = \frac{1}{2}u^2$, a convex function of u.

- In problem (1a) the central point of the initial tanh curve remains at the same position with a shock forming at $x = \frac{1}{2}$. The shock remains stationary and is formed at $t = \frac{1}{10}$. (See example 3 in chapter 2, section 2.3.8, for an analytic description.)

- Problem (1b) is similar to problem (1a) but in this case the central point of the curve is displaced. This causes a moving shock to form at $t = \frac{1}{10}$, $x = \frac{1}{2}$, which moves with speed $\frac{1}{2}$.

- Problem (1c) has initial data in the form of a ramp, which steepens to form a shock at $t = \frac{1}{2}$, $x = \frac{3}{4}$ (similar to example 1 in chapter 2 sections 2.3.2 and 2.3.4).

5.1.2 Problem 2 : $u_t + (u^4/4)_x = 0$

This problem is given by the equation

$$u_t + u^3 u_x = 0 \tag{5.7}$$

$$\text{or} \quad u_t + \left(\frac{u^4}{4}\right)_x = 0 \tag{5.8}$$

on the region $[0, 1]$ with initial data given by

$$u = \tanh(5 - 10x) \tag{5.9}$$

and with Dirichlet conditions given on the boundaries. Equation (5.8) is given as an example because it is a conservation law which does not admit an exact

solution with piecewise linear moving finite elements, unlike Burgers' equation. The flux function is also convex providing another example to be used with the Brenier shock recovery method.

- Problem (2) forms shocks initially just to the right and to the left of $x = \frac{1}{2}$; however almost immediately they join to form a shock at $x = \frac{1}{2}$.

5.1.3 Problem 3 : Buckley-Leverett Equation

This equation is given by

$$u_t + \left(\frac{u^2}{u^2 + \frac{1}{2}(1-u)^2} \right)_x = 0 \qquad 0 \leq x \leq 2 \tag{5.10}$$

with initial condition

$$u = \frac{1}{1+10x} \qquad 0 \leq x \leq 2 \tag{5.11}$$

and Dirichlet conditions are imposed at the boundaries. The equation provides a model for the flow of oil in porous media and was first described in (Buckley & Leverett (1942)). This equation is interesting because its solution is a combination of a shock and an expansion, due to the flux function being non-convex. This can lead to difficulties in calculating the shock speed (Concus & Proskurowski (1979)). We will not give any solution for this equation because of these difficulties, however a brief description of the problems that occur is given below.

Let us consider the simplest case of using initial Riemann data. See Fig. 5.1. First let us examine the initial data where u_L and u_R are shown. Now consider the diagram showing the flux function where again u_L and u_R are marked. The construction required (see Concus & Proskurowski (1979)) is carried out by drawing a line from u_R to a point u_Q where it is tangent to the flux function. This point u_Q denotes the split between the formation of a shock and the formation of an expansion as the solution evolves. See Fig. 5.2. It should be noted that this becomes much more complicated for non-Riemann initial data.

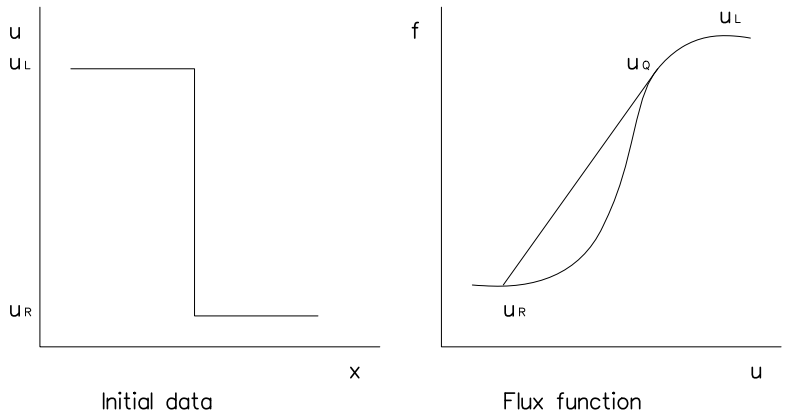

Figure 5.1: Initial data and flux function.

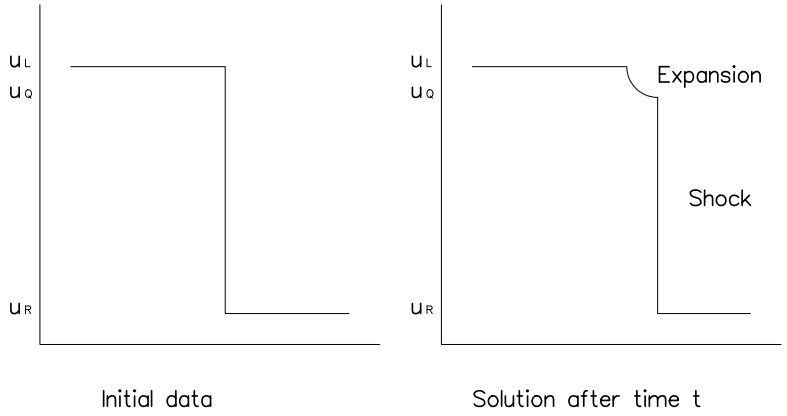

Figure 5.2: Initial data and solution after time t.

5.1.4 Problem 4 : Riemann Problem

The Riemann problem is given by two different levels of constant initial data in conjunction with a conservation law. Here the conservation law is taken as

$$u_t + \left(\frac{u^4}{4}\right)_x = 0 \tag{5.12}$$

on the region $[0, 1]$. The two sets of initial data to be tested are given by

$$a) \quad u = \begin{cases} 1 & 0 \leq x \leq \frac{1}{2} \\ 0 & \frac{1}{2} \leq x \leq 1 \end{cases} \tag{5.13}$$

$$b) \quad u = \begin{cases} 0 & 0 \leq x \leq \frac{1}{2} \\ 1 & \frac{1}{2} \leq x \leq 1 \end{cases} . \tag{5.14}$$

This will enable us to demonstrate the problem that with a poor initial representation, i.e. badly placed nodes, the final solution will be poor also. The two sets of initial data give very different results, (4a) forms a shock and (4b) forms an expansion.

5.2 Representation Of Initial Data

Three methods of initial data representation are used here. Each involves sampling the function on an initial grid. The first grid is obtained by equi-spacing the nodes. The second grid involves placing the nodes in way influenced by the type of solution to be formed. The final grid comes from a form of equidistribution and requires that the function is twice differentiable. The nodes are equidistributed using a weight function $(U_{0_{xx}})^{\frac{2}{5}}$, where U_0 is the initial data (Carey & Dinh (1985)). This means that the nodes are distributed according to

$$\frac{\int_{s_0}^{s_N} (U_{0_{xx}}(x))^{\frac{2}{5}}}{N+1} = \int_{s_j}^{s_{j+1}} (U_{0_{xx}})(x)^{\frac{2}{5}} dx \tag{5.15}$$

where the s_j $(j = 0, ..., N)$ are the node positions. This initially places more nodes at regions of high curvature and fewer elsewhere.

5.3 Overturning Solutions

There are several methods described in chapters 3 and 4 which can be used to calculate overturned solutions. These methods are applied to some of the problems above to in order to demonstrate their ability to solve different types of equation. In this section the solutions given will be the multivalued solutions (i.e. the type of solution which would be obtained by following the characteristics through their intersection). Methods of shock recovery can be applied to to these overturned solutions in order to obtain the shock position. This is done in section 5.4.

The first method we will consider is the global or local MFE method modified so that it is written in a 2-stage form (see chapter 4 section 4.4.1) so as to permit overturning.

5.3.1 MFE - 2 Stage

This method is described in chapter 4 section 4.4.1 and permits an overturned solution to form. The method is applied to all the problems given above to show how multivalued solutions form.

Problem 1a : Inviscid Burgers' Equation

We will first apply this method to problem 1, with initial data (a) which is a tanh curve. This will produce a stationary shock (when a jump condition is applied) or an overturned curve which evolves with time always cutting the x-axis at $x = \frac{1}{2}$. The nodes are initially placed using the equidistribution routine described above. The results are given in Fig. 5.3. Since the MFE method is equivalent to the

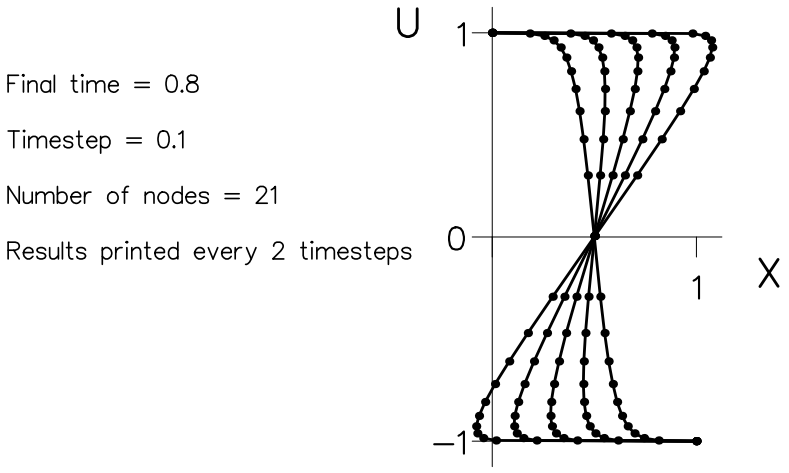

Figure 5.3: Problem (1a) - MFE 2 stage method.

method of characteristics in this case, the equations of characteristics being linear, the same final result could have been produced by using a single time-step of size 0.8. This occurs because $\frac{dx}{dt}$ is constant ($= u_0$) along the characteristics, so that Euler's method is exact for arbitrary time-steps. From Fig. 5.3 the nodes can be

seen to stay with the regions of high curvature as the solution evolves with time. This enables the curved solution to be well represented.

Problem 1b : Inviscid Burgers' Equation

This is similar to the above problem but the initial tanh data is shifted vertically, which gives a moving shock or an overturning curve which changes both shape and position as it evolves with time. The nodes are placed initially by using the equidistribution routine described above. The results are given in Fig 5.4 shown below. Again, these results could have also been obtained by using a single time-

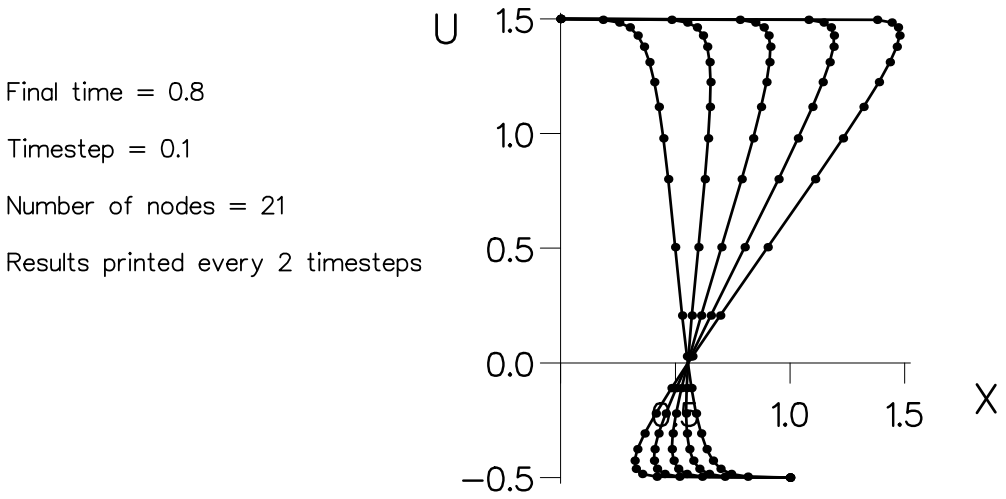

Figure 5.4: Problem (1b) - MFE 2 stage method.

step, since we are solving Burgers' equation (see problem 1(a)). The nodes can again be seen to cluster around regions where high resolution is needed, with less where the curve is straight.

Problem 1c : Inviscid Burgers' Equation

The initial data here is given by a ramp, i.e. a piecewise linear function. This function does not have a second derivative, therefore the equidistribution routine described above cannot be used. Initially the nodes are equi-spaced over the region. The results for this problem are given in Fig. 5.5, from which it can be seen that if the initial ramp data is not well represented then this poor representation is carried throughout the calculations. Note: From other numerical experiments

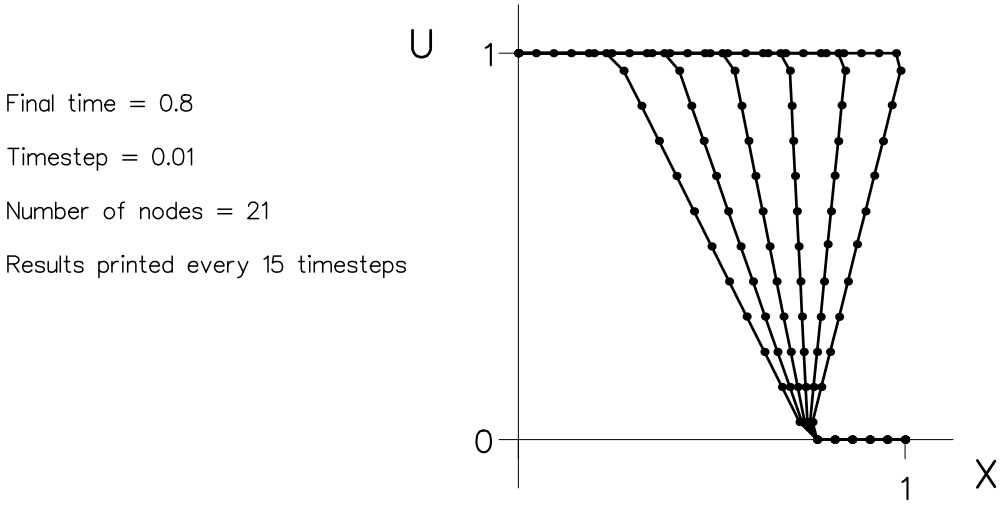

Figure 5.5: Problem (1c) - MFE 2 stage method.

it has been found that if the initial data given here is represented by only 4 nodes (chosen to be at 0.0, 0.25, 0.75 and 1.0) then a more accurate solution than that given above is obtained. This shows that the accuracy of the solution is not dependent on just the number of nodes, but on their representation of the initial data.

Problem 2 : $u_t + (u^4/4)_x = 0$

The equation above is applied to the same tanh initial data that was used in problem (1a). The nodes are again initially placed using the equidistribution routine. This allows us to compare how the two different equations affect the initial data. The results are shown in Fig. 5.6 below. It can be seen that unlike Burgers' equation, this equation requires a small controlled time-step for solution. The nodes can be seen to move horizontally by the motion of the characteristics but a slight vertical motion can also be seen. This vertical motion is caused by the projection step of the MFE method (a step that is not needed when considering Burgers' equation). The nodes can also be seen to move to regions of high curvature leaving fewer to represent the straighter regions.

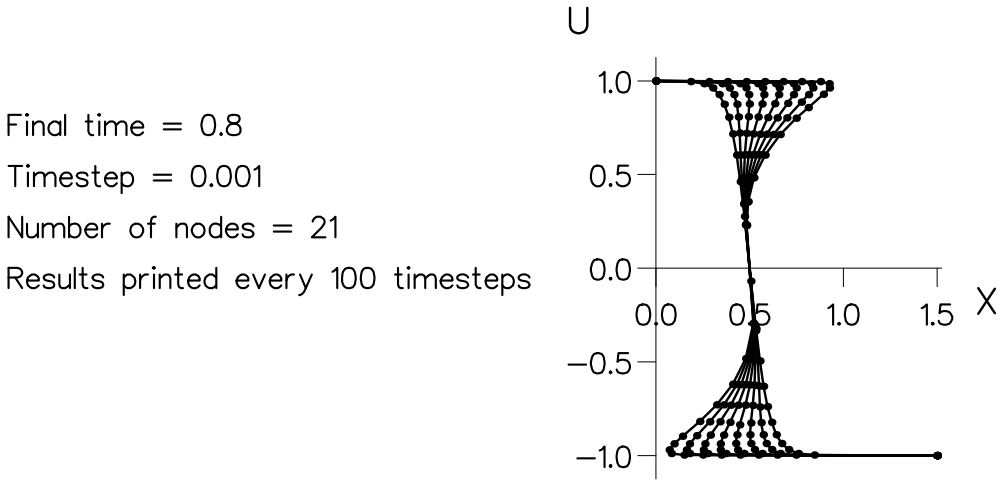

Figure 5.6: Problem (2) - MFE 2 stage method.

Problem 3 : Buckley-Leverett Equation

The initial data is given by $1/(1+10x)$ and the nodes are initially place using the equidistribution routine described in section 5.2. The results are given for this problem in Fig. 5.7 below. Although this problem gives rise to a combined

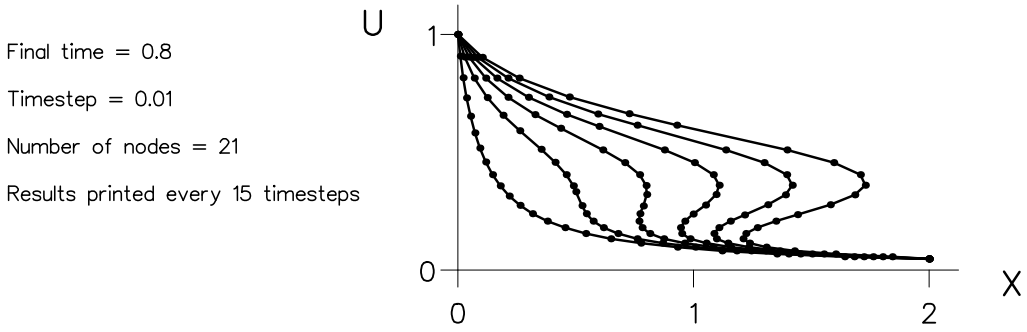

Figure 5.7: Problem (3) - MFE 2 stage method.

shock and expansion, it is not possible to distinguish it from the overturned manifolds formed by shocks in the above problems. Apart from this the results to this problem are very similar to those of problem 2. Reminder: there is a special procedure for the calculation of the shock speed for this example (with

non-convex or non-concave flux function), (see (Concus & Proskurowski (1979))).

Problem 4 : Riemann Problem

For this problem we will show the results for two sets of initial data. Each problem has initial data given by piecewise constant regions. In both cases the 'jump' between the regions is approximated by a function representing a very steep slope.

For problem (4a) the initial data is given by

$$u = \begin{cases} 1 & 0 \leq x \leq 0.495 \\ -100x + 50.5 & 0.495 \leq x \leq 0.505 \\ 0 & 0.505 \leq x \leq 1 \end{cases} . \tag{5.16}$$

We will space the nodes equally throughout the region (Note: no second derivatives available for equidistribution). As a consequence the representation of the very steep slope is dependent upon the number of nodes used. In problem (4a), a shock is formed almost immediately which then moves across the region. In this section an overturned manifold is generated, the shock position being recovered in a later section. The results are shown in Fig. 5.8. The extra number of nodes

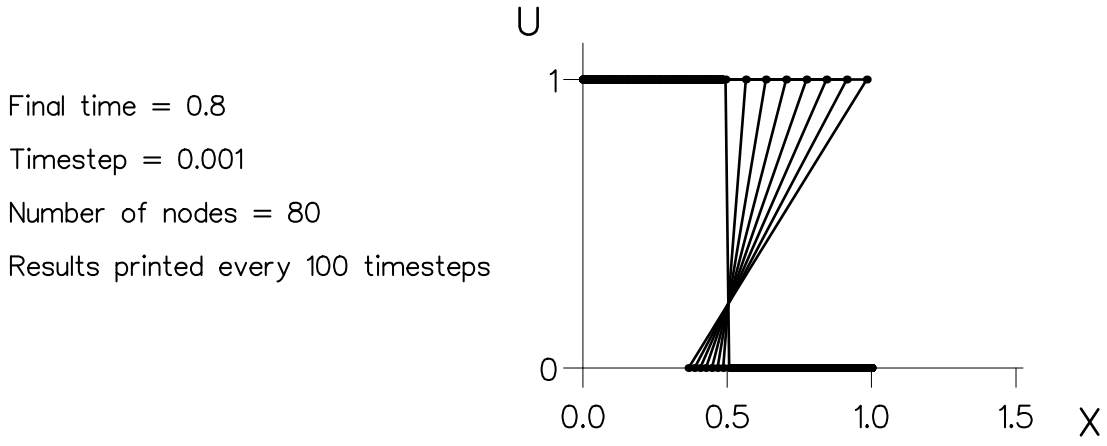

Figure 5.8: Problem (4a) - MFE 2 stage method.

used in this example are wasted in representing the two constant regions whilst there are no nodes in the very steep region. Clearly equi-spacing is not always

the best method of placing nodes, and the initial placement of nodes requires us to respect both the initial data and the subsequent motion.

This is not an economic method of placing the nodes and negates the adaptive grid strategy, so in this problem we have to consider the type of solution to be formed in choosing the initial node positions. For this problem we shall place several nodes within the very steep region so that as the expansion forms there will be enough nodes to represent the curve. The results are shown in Fig. 5.9 below.

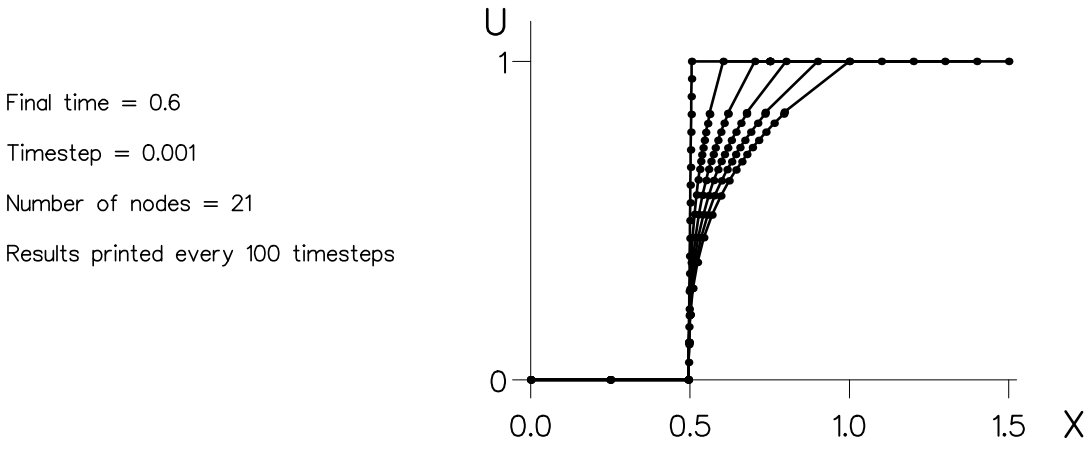

Figure 5.9: Problem (4b) - MFE 2 stage method.

From these results it can be seen that if there are not enough nodes in the vertical part of the initial data then as the expansion forms there will not be enough nodes to represent the solution: for example if there were none in the vertical part, the expansion would come out linear, which is not correct in this case.

It should be noted that representation of a curved expansion by only one element would be a very poor solution (see Fig. 5.9). If the number of nodes in the initial vertical region is increased then more nodes move into the region of the curved expansion. Now, if the solution (i.e. the function representing the curved region) is given by f, then the error between the function f and the piecewise linear representation satisfies

$$| \text{error} | \leq Ch^2 |f''|_{max} \tag{5.17}$$

where f is the maximum distance between the nodes $|.|_{max}$ is the maximum norm and C is a constant. If we want the error to be less than some tolerance and an approximation to $\max|f''|$ can be estimated then the maximum required node spacing can be found. This allows the number of nodes to be placed in the vertical region to be chosen.

5.3.2 Split MFE, GWMFE, Split GWMFE, etc

These methods have all been applied to the above problems but give extremely similar results, therefore no results will be shown for these methods.

5.3.3 VM Method

This method involves calculating the initial data in the original variables then transforming to v, m coordinates using the Legendre transformation described in chapter 2 section 2.8. The transformation is given by

$$u(x) + v(m) - mx = 0 \tag{5.18}$$

$$m = u_x \qquad x = v_m. \tag{5.19}$$

The solution is calculated using the transformed coordinates before returning back to the original variable to display the results. A picture of the solution evolving with time in the v, m coordinates is not usually informative. The results in x, u coordinates are again similar to those already given above.

5.3.4 Solution Via Integrated Form

In this section we consider results obtained by using a different transformation. The initial data is given in the original variables, then transformed to the a, x space where $a = \int u dx$ by integrating (see chapter 2 section 2.7.4). The solution is calculated by the local 2-stage (or any other method which allows an overturned curve to form) in the a, x space, then the solution is transformed back to x, u where the results are required. Here the solutions are shown in transformed coordinates only, since the solutions in x, u space are very similar to those already shown above.

Problem 1a : Inviscid Burgers' Equation

This first example shows problem (1a) transformed to a, x variables as the shock forms. See chapter 4 section 4.5.2. See Fig. 5.10. There are less nodes used in this

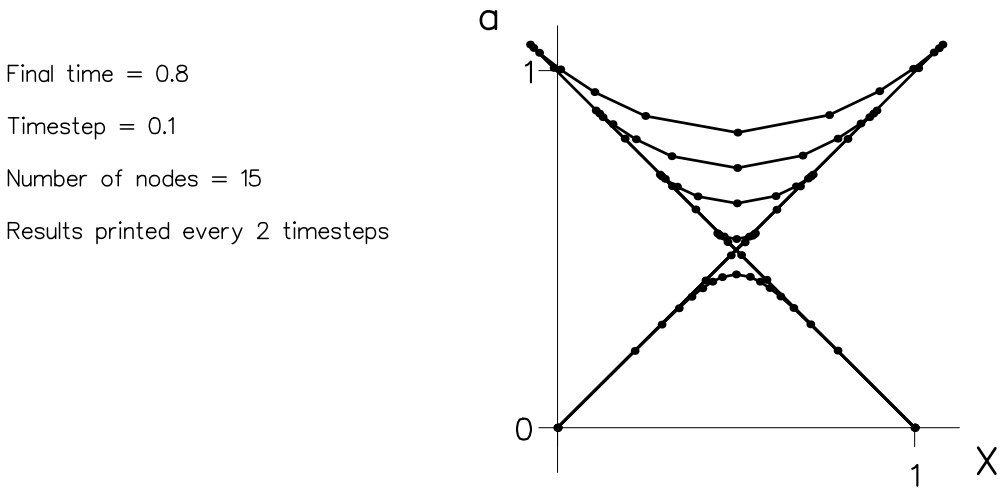

Figure 5.10: Problem (1a) - Transformed to a, x variables.

calculation so that the results are easier to see. The initial data transformed looks approximately like a $\sin^2$ curve and as time increases a characteristic swallow-tail is formed. The a, x variables can then be transformed back into u, x space to give the same solution as Fig. 5.3.

Problem 1b : Inviscid Burgers' Equation

The second example here shows the moving shock formed in example (1b). See Fig. 5.11. After the calculation the a, x curve is then transformed back to the original variables to give the same solution to Fig. 5.4. It can be seen that the shock moves to the right and that the speed must be correct owing to conservation.

5.3.5 Summary Of Overturned Results

The results show that both expansions and overturned manifolds (caused by the formation of shocks) can be calculated for a variety of problems. It has also been evident that care is needed in choosing the representation of the initial data so that a good solution can be found.

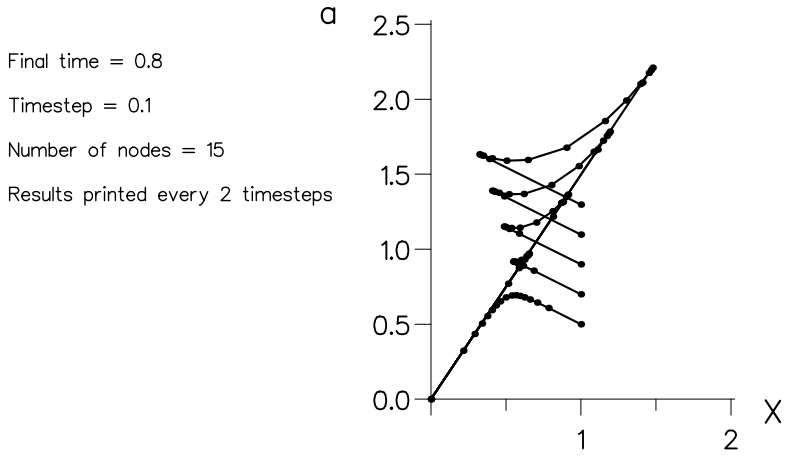

Figure 5.11: Problem (1b) - Transformed to a, x variables.

5.4 Recovery Of Shock Position From An Overturned Curve

All the methods used above can give a multivalued solution for conservation laws and consequently may have one of the recovery methods based on conservation (given in chapter 4 section 4.5) applied to them. Note: the transport collapse operator of Brenier is different in that the solution obtained by applying it to a multivalued curve then affects the solution at the next time-step. The other two methods are only applied when the shock position is required and this has no effect on solutions at subsequent time-steps.

The methods of recovery given in the examples below are all applied to the 2 stage local method since the results will apply equally to any overturned curve. There are three methods of shock recovery discussed below;

(i) equal area,

(ii) via an integral and

(iii) the transport collapse operator of Brenier

described in section 2.9. It should be noted that for convex or concave f in (5.1) all three methods may be used, but for non-convex (or non-concave) f the third method (Brenier) may not be used.

5.4.1 Equal Area Method - Bisection

Problem 1a : Inviscid Burgers' Equation

The corresponding overturned curve for this initial data is shown in Fig 5.3. The shock position location can be seen in Fig. 5.12. The results shown with

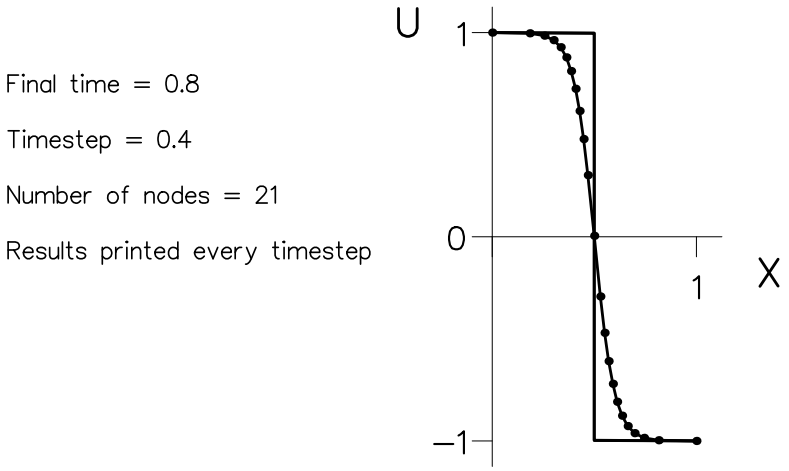

Figure 5.12: Problem (1a) - Equal area method.

node positions displayed show that the curve has not yet overturned and no recovery technique has been applied. The shock positions calculated from the overturned curves are shown without any node positions since they are then no longer applicable. This is because the node positions are then located in the overturned solution (Fig. 5.3) which is used to give the solution at later time. The equal area method calculation simply gives a representation of the shock position and does not affect the overturning solution given earlier.

This example shows that the method may be used to calculate a stationary shock position. Similarly the next example uses the same method to calculate a moving shock.

Problem 1b : Inviscid Burgers' Equation

The overturning curve associated with this problem is shown in Fig. 5.4. Here the results are given in Fig. 5.13 showing the location of the shock position. It can be seen that the shock moves to the right.

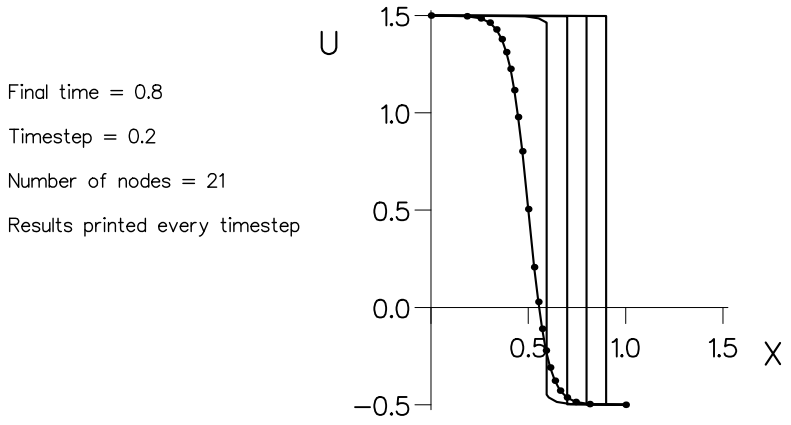

Figure 5.13: Problem (1b) - Equal area method.

5.4.2 Via The Integral Transformation

This method is also based upon the principle of conservation It gives a representation of the shock position at any instant but does not affect the overturned solution calculated in the above section. These methods have been run for all the test problems and give similar results. This method is different to the one described in section 5.3.5 because the solution is calculated in x, u variables and only integrated to a, x variables when the shock position is required. The transformation is given by

$$a(x) + b(u) - ux = 0 \tag{5.20}$$

$$u = a_x \qquad x = b_u. \tag{5.21}$$

The shock position is calculated from the self-intersection of the curve, then given in terms of the original variables.

Problem 3 : Buckley-Leverett Equation

The results together with the integrated curve are shown in Fig 5.14. The self-intersection point on the integrated curve marks the shock position. Owing to conservation the method of calculating the shock position using equal area is valid although both a shock and an expansion form in this problem. The only difficulty with such non-convex flux functions arises in calculating the shock speed which

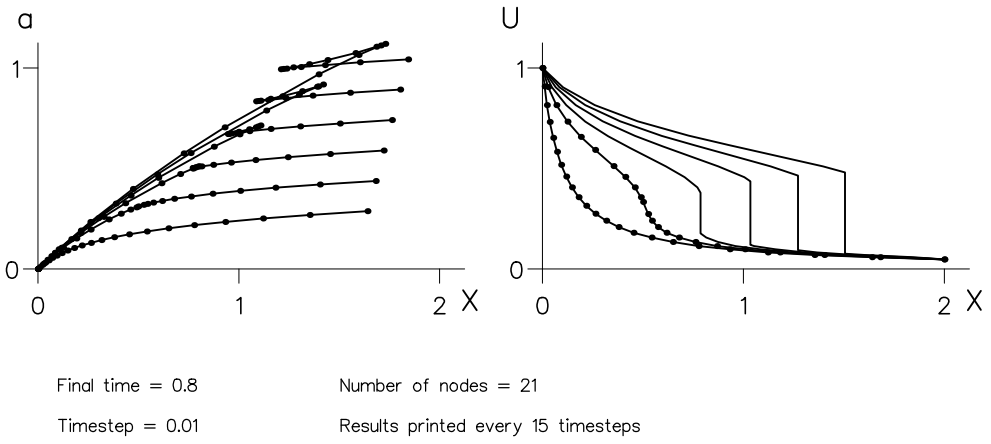

Figure 5.14: Problem (3) - Shock position calculated using Integrated curve.

in this case requires a special construction. See (Concus & Proskurowski (1979)).

Problem 4 : Riemann Problem

The nodes are initially placed at 0.0, 0.25, 0.495, 0.505, 0.75, 1.0 with the remainder equally spaced between 0.495 and 0.505 in order to give a good representation of the shock. The piecewise constant data in the statement of the Riemann problem is again approximated as in (section 5.3.1) problem 4.

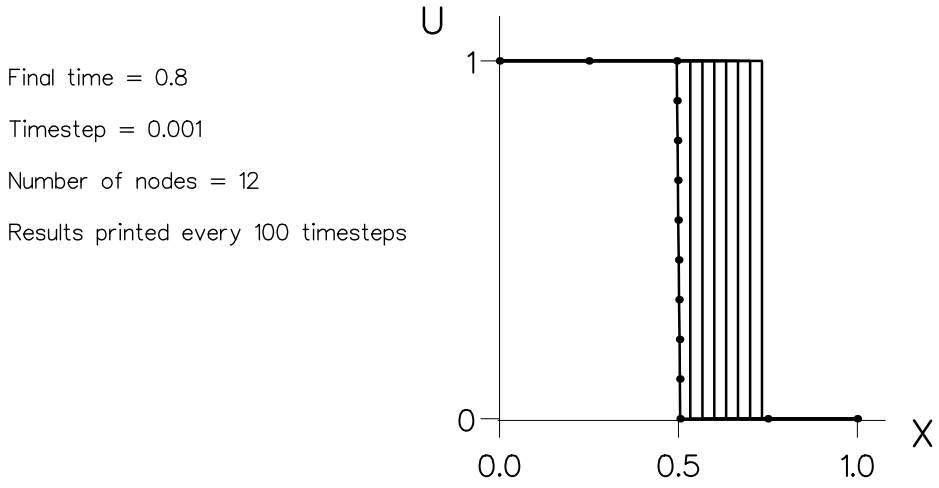

Figure 5.15: Problem (1c) - Shock position using integral method.

5.4.3 Brenier

This method is different to those above in that this method reconnects the nodes. It can therefore be applied after every time-step or only after the final time-step. The results are given below for the two cases.

Problem 1a : Inviscid Burgers' Equation

In this case the TC operator is applied after the final time-step only. The nodes are initially placed using the equidistribution routine described earlier. The results are given in Fig. 5.16. The result of solving problem (1a) with the TC

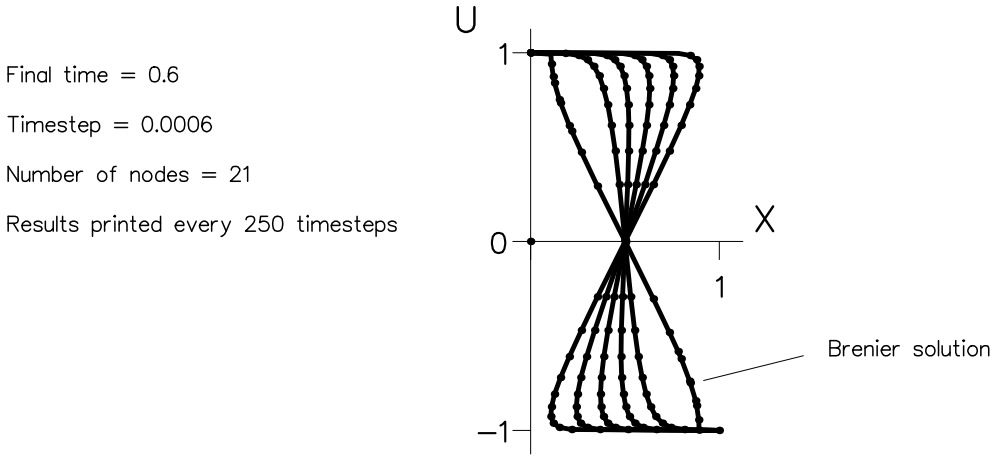

Figure 5.16: Problem (1a) - Shock position using TC operator.

operator applied after each time-step gives a result similar to that of the equal area method. i.e. a vertical shock position.

Problem 1b : Inviscid Burgers' Equation

In this case the TC operator is applied after every time-step for which there is an overturned curve. The initial data nodes are placed using an equidistribution routine. The results are shown below in Fig. 5.17. If the TC operator is applied after every time-step then the results are the same as the other two shock recovery methods.

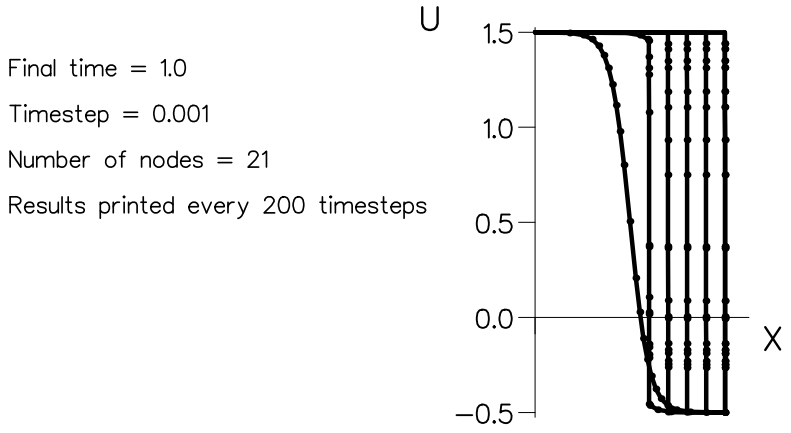

Figure 5.17: Problem (2) - Shock position using TC operator.

5.4.4 Summary On Shock Position Calculations

It has been seen that the shock position can be calculated using the three methods proposed in chapters 2 and 4. The results show that the equal area method, calculation via an integral and the TC operator applied after every time-step give good results.

In the case where Brenier (TC operator) is only applied after the final time-step, the results are very different. This approximation to the shock is significantly worse than the other methods (since if the equal area method or integral methods were applied only after the final time-step, then a vertical non-smeared shock would result.)

5.5 Summary

In this chapter we have considered various numerical methods and several test cases to illustrate the work described in chapters 2-5. From the results it can be seen that overturned solutions may be calculated for a variety of initial data and for various conservation laws. The results also show that the shock position may be recovered from the overturned curves.

The results also show that the initial data representation (which includes both the number and position of the nodes) is very important for this type of method.

This can be seen most clearly in both the formation of the ramp (see Fig. 5.5) and the Riemann problems (see Figs. 5.8 and 5.9), however it applies equally to the calculation of all initial data.

It should also be noted that the method cannot cope with the case when two shocks merge. Another important point to consider is that some examples require a very small time-step so that no instabilities occur.

In the following chapters we will extend the work of chapters 2-5 to higher dimensions.

Chapter 6

Analytic Methods For Conservation Laws In Higher Dimensions

6.1 Introduction

In this chapter we will consider possible analytic solutions to first order scalar nonlinear partial differential equations of the form

$$F(\mathbf{x}, u, \mathbf{m}) = 0 \tag{6.1}$$

given on the region $\Omega \in \mathbb{R}^n$, where $\mathbf{x} = (x_1, ..., x_n)$, $\mathbf{m} = (m_1, ..., m_n)$ and $m_i = \frac{\partial u}{\partial x_i}$ and n is the number of dimensions. This class of equations is generally too complicated to fully analyse theoretically. However the types of behaviour exhibited by such equations are of interest since this knowledge is of considerable use when numerical methods are developed. The formation of shocks and expansions are the main areas of interest here. We shall therefore concentrate on the conservation laws because they are the simplest class of equations which form shocks and expansions. In particular we will consider Riemann problems for conservation laws since they provide a basis on which to develop solutions for more general problems. For general initial data, it can be shown that discontinuities (shocks) occur in finite time. There are also some existence and uniqueness theorems known for this class of problems. For a few specific problems an analytic solution may be found. Such cases are discussed below.

6.2 Characteristics

For a general first order PDE in higher dimensions, characteristics can be found in exactly the same way as in 1-D (see chapter 2). The characteristics of (6.1) are given by

$$\frac{dx_i}{dt} = F_{m_i} \qquad i = 1, ..., n \tag{6.2}$$

$$\frac{du}{dt} = \sum_{i=1}^{n} m_i F_{m_i} \tag{6.3}$$

$$\frac{dm_i}{dt} = -F_{x_i} - m_i F_u \qquad i = 1, ..., n. \tag{6.4}$$

From this we have a system of $2n + 1$ ODE's and one algebraic equation $F = 0$ for the $2n + 1$ functions $x_1(t), ..., x_n(t), u(t), m_1(t), m_n(t)$. It can be shown that F is an integral of the characteristics, hence $F = 0$ must be satisfied at some initial point $t = 0$ of (6.1) for it to be satisfied for all t. For more information on this equation see (John (1971)). Although families of characteristics can be found, it is difficult to use them in a similar way to chapter 2 since the characteristics now form surfaces. Let us now reduce the class of equations we are considering to the conservation laws.

6.3 Conservation Laws

As in 1-D conservation laws have solutions which contain both shocks and expansions. Since the aim of this chapter is to provide an analytic background for numerical methods to be developed in later chapters to approximately locate shocks then the conservation laws provide a good class of equations to investigate. These equations have been extensively studied in 1-D but there has been comparatively little done in two or higher dimensions.

6.3.1 Derivation

Let us first consider a derivation of the conservation law in 2-D. This is carried out in a similar way to that in 1-D (chapter 2) which may be readily extended to higher dimensions. Let $u(\mathbf{x}, t)$ (e.g. density) be defined on a region Ω, then the

$$\text{Rate of Change of integral of } u \text{ in } \Omega \quad = \quad \text{Total flux of } u \text{ through } \partial\Omega \text{ (into } \Omega) \tag{6.5}$$

$$\Rightarrow \frac{d}{dt}\int_\Omega u d\Omega \quad = \quad -\int_{\partial\Omega} \mathbf{F}(u).\mathbf{ds} \tag{6.6}$$

where $\mathbf{F}$ is the flux function

$$\mathbf{F} = \begin{pmatrix} f(u) \\ g(u) \end{pmatrix}. \tag{6.7}$$

Equation (6.6) can now be written as

$$\int_\Omega u_t d\Omega = -\int_\Omega \boldsymbol{\nabla} .\mathbf{F}(u) d\Omega, \tag{6.8}$$

using Gauss' Theorem. This gives

$$\int_\Omega (u_t + f_x + g_y) d\Omega = 0, \tag{6.9}$$

and since this holds for all Ω, however small, then

$$u_t + f_x + g_y = 0. \tag{6.10}$$

The problems we shall consider here are given by the conservation law (6.10), with various f, g and initial data and boundary conditions applied where appropriate.

The conservation law (6.10) may be written in many forms, some of which are helpful when considering certain methods of solution. Since f, g are functions of u then (6.10) can be written as

$$u_t + f^{'}(u)u_x + g^{'}(u)u_y = 0. \tag{6.11}$$

In a similar way to 1-D, let $f^{'}(u) = a(u)$, $g^{'}(u) = b(u)$ to give

$$u_t + a(u)u_x + b(u)u_y = 0 \tag{6.12}$$

which is a 2-D quasi-linear equation where a and b are known as the wave speeds.

6.3.2 Characteristics

In a similar manner to 1-D, characteristic equations for conservation laws can be found. Using the general equations for characteristics in section 6.2, and applying them to (6.12) gives characteristic equations

$$\frac{du}{dt} = 0, \quad \frac{dx}{dt} = a(u), \quad \frac{dy}{dt} = b(u). \tag{6.13}$$

These equations can be solved, by integration with respect to t to give

$$u = \text{constant} = u_0 \tag{6.14}$$

$$x = a(u_0)t + x_0 \tag{6.15}$$

$$y = b(u_0)t + y_0 \tag{6.16}$$

where u_0, x_0 and y_0 are constants, which are straight lines.

Let us now consider the theory for conservation laws in 2-D.

6.3.3 Theory

There are several existence and uniqueness theorems for the general conservation law. Using general initial data on (6.12) the existence of a weak solution has been given by (Conway & Smoller (1966)). Vol'pert and Kružkov proved existence and uniqueness of a weak solution satisfying the entropy condition in the class of bounded measurable spaces (Kružkov (1969)), (Vol'pert (1967)). An important point to note is that Conway shows that for smooth initial data that the solution will form discontinuities in finite time (Conway (1977)). These existence theorems do not however provide a method of solution to (6.12).

In order to consider the occurrence of discontinuities in the solution of the conservation laws, we will first write (6.10) in Lagrangian form.

6.3.4 Lagrangian Form Of Conservation Laws

The Lagrangian form in 2-D may be obtained by considering a coordinate transformation between x, y, t and independent variables ξ, η, τ where

$$x = \hat{x}(\xi, \eta, \tau), \quad y = \hat{y}(\xi, \eta, \tau), \quad t = \tau, \quad u = \hat{u}(\xi, \eta, \tau). \tag{6.17}$$

Using the new variables (6.12) may be written as

$$\frac{\partial \hat{u}}{\partial \tau} - \frac{\partial u}{\partial x}\frac{\partial \hat{x}}{\partial \tau} - \frac{\partial u}{\partial y}\frac{\partial \hat{y}}{\partial \tau} + a(u)u_x + b(u)u_y = 0. \tag{6.18}$$

Now using similar notation to chapter 2, section 2.3, we let

$$\dot{u} = \frac{\partial \hat{u}}{\partial \tau}, \quad \dot{x} = \frac{\partial \hat{x}}{\partial \tau}, \quad \dot{y} = \frac{\partial \hat{y}}{\partial \tau}. \tag{6.19}$$

Substituting these into (6.18) gives

$$\dot{u} - u_x\dot{x} - u_y\dot{y} + a(u)u_x + b(u)u_y = 0. \tag{6.20}$$

If we now compare coefficients of u_x and u_y in (6.20), we get

$$\dot{u} = 0 \tag{6.21}$$

$$\dot{x} = a(u) \tag{6.22}$$

$$\dot{y} = b(u) \tag{6.23}$$

which are the same as the characteristics in section 6.3.2.

6.4 Blow-up

In this section we will describe one method (Conway (1977)), (Majda (1984)) of showing that the solution of (6.10) forms a discontinuity in finite time. We will apply this technique to conservation laws in two dimensions of the form

$$u_t + a(u)u_x + b(u)u_y = 0. \tag{6.24}$$

Differentiate (6.24) with respect to x to give

$$u_{xt} + a'(u)u_x^2 + a(u)u_{xx} + b'u_xu_y + b(u)u_{yx} = 0 \tag{6.25}$$

and with respect to y to give

$$u_{yt} + a'u_yu_x + a(u)u_{xy} + b'(u)u_y^2 + b(u)u_{yy} = 0. \tag{6.26}$$

If we write (6.24) in Lagrangian form it becomes

$$\dot{u} - u_x\dot{x} - u_y\dot{y} + a(u)u_x + b(u)u_y = 0. \tag{6.27}$$

Following the Lagrangian method we choose $\dot{x} = a(u)$ and $\dot{y} = b(u) \Rightarrow \dot{u} = 0$. Similarly (6.25) and (6.26) become

$$\dot{u}_x - u_{xx}\dot{x} - u_{yx}\dot{y} + a^{'}(u)u_x^2 + a(u)u_{xx} + b^{'}(u)u_x u_y + b(u)u_{yx} = 0 \tag{6.28}$$

and

$$\dot{u}_y - u_{yx}\dot{x} - u_{yy}\dot{y} + a^{'}(u)u_x u_y + a(u)u_{xy} + b^{'}(u)u_x^2 + b(u)u_{yy} = 0. \tag{6.29}$$

With $\dot{x} = a(u)$ and $\dot{y} = b(u)$, (6.28) and (6.29) reduce to

$$\dot{u}_x = -a^{'}(u)u_x^2 - b^{'}(u)u_x u_y \tag{6.30}$$

and

$$\dot{u}_y = -a^{'}(u)u_x u_y - b^{'}(u)u_x^2. \tag{6.31}$$

Now let $q = a(u)_x + b(u)_y = div(a, b) = a^{'}(u)u_x + b^{'}(u)u_y$. Differentiate q with respect to t to give

$$\begin{aligned} \dot{q} &= a^{''}(u)\dot{u}u_x + a^{'}(u)\dot{u}_x + b^{''}\dot{u}u_y + b^{'}(u)\dot{u}_y & (6.32)\\ &= a^{'}(u)\dot{u}_x + b^{'}(u)\dot{u}_y. & (6.33) \end{aligned}$$

Substitute (6.30) and (6.31) into (6.33) to give

$$\begin{aligned} \dot{q} &= a^{'}(u)(-a^{'}(u)u_x^2 - b^{'}(u)u_x u_y) + b^{'}(u)(-a^{'}(u)u_x u_y - b^{'}(u)u_y^2) & (6.34)\\ &= -(a^{'}(u)u_x + b^{'}(u)u_y)^2 & (6.35)\\ &= -q^2. & (6.36) \end{aligned}$$

Writing this as

$$\frac{dq}{d\tau} = -q^2 \tag{6.37}$$

and integrating gives

$$\begin{aligned} \int \frac{dq}{q^2} &= -\int d\tau & (6.38)\\ \Rightarrow -\frac{1}{q} &= \tau + C & (6.39)\\ \Rightarrow q &= \frac{1}{\tau - C} & (6.40) \end{aligned}$$

where C is a constant. Let $q = q_0(\xi)$ at $\tau = 0$ which gives

$$q(\xi) = \frac{q_0(\xi)}{1 + q_0(\xi)\tau}. \tag{6.41}$$

From this it can be seen that

$$q(\xi) \to \infty \text{ as } \tau \to -\frac{1}{q_0(\xi)}. \tag{6.42}$$

which indicates that the solution blows up at this time for some ξ. To find the time that the solution first blows-up, we want the smallest positive value of $q_0(\xi)(\forall\xi)$ for which $q(\xi) \to 0$. This is equivalent to the discontinuity that was said to occur by Vol'pert.

It can also be seen from (6.30), (6.31) that

$$\frac{\dot{u}_x}{\dot{u}_y} = \frac{u_x}{u_y} \tag{6.43}$$

from which

$$\frac{\partial}{\partial\tau}\ln u_x = \frac{\partial}{\partial\tau}\ln u_y \tag{6.44}$$

$$\ln u_x = \ln u_y + \text{function of } \xi \tag{6.45}$$

$$\frac{u_x}{u_y} = A(\xi). \tag{6.46}$$

Using this result and the definition of q, allows u_x, u_y to be found as functions of ξ and hence the solution of the equation to be found.

From blow-up, and also from (Vol'pert (1967)), we can see that a weak formulation of the equation is required since discontinuities will occur in the solution.

6.5 Weak Solutions

From the above section it can be seen that weak solutions should be considered when solving conservation laws. The need for weak solutions has been noted by (Vol'pert (1967)), (Kružkov (1969)) and (Conway & Smoller (1966)). The version proposed by (Conway & Smoller (1966)) is given below for (6.10).

A bounded measurable function $u : \mathbb{R}^+ \times \mathbb{R}^2 \to \mathbb{R}$ is said to be a weak solution to (6.10) with initial data $u = u_0(x, y)$ if

$$\int_{\mathbb{R}^+}\int_{\mathbb{R}^2}\left[u\frac{\partial}{\partial t}\phi + f(u)\frac{\partial}{\partial x}\phi + g(u)\frac{\partial}{\partial y}\phi\right]dxdydt = 0 \tag{6.47}$$

for every test function $\phi \in C_0^\infty(\mathbb{R}^+ \times \mathbb{R}^2)$ and if $u(t,.,.) \to u_0$ in L^1_{loc} as $t \to 0$.

6.6 Riemann Problems

A large area of interest concerns Riemann problems (Wagner (1983)), (Klingenberg (1986)), (Lindquist (1986)). This type of problem is given by conservation laws which have initial data given by piecewise constant values in adjacent regions. The solution leads to the formation of shocks and expansions at the initial time t_0. This is a separate problem to the case where smooth initial data is given.

6.6.1 The Problem

In 2-D a general type of Riemann problem is given by

$$\frac{\partial u}{\partial t}(x,y,t) + \frac{\partial}{\partial x} f(u(x,y,t)) + \frac{\partial}{\partial y} g(u(x,y,t)) = 0 \tag{6.48}$$

with initial data

$$u(0,x,y) = \begin{cases} u_1 & x > 0 \quad y > 0 \\ u_2 & x < 0 \quad y > 0 \\ u_3 & x < 0 \quad y < 0 \\ u_4 & x > 0 \quad y < 0 \end{cases} \tag{6.49}$$

where $u_1, ..., u_4$ are constants. A possible set of initial data is shown in Fig. 6.1 below. From this we can see that a shock and/or expansion (or both) will form at the initial time t_0.

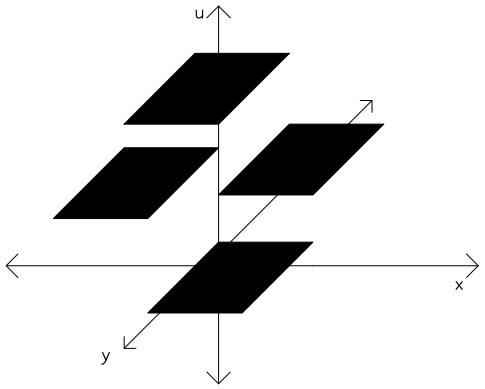

Figure 6.1: Initial data for a Riemann Problem.

6.6.2 Analytic Solution Of The Riemann Problem.

Several methods of solution have been suggested (Wagner (1983)), (Klingenberg (1986)), (Lindquist (1986)), (Guckenheimer (1975)), for the solution of this Rie-

mann problem subject to different restrictions on the functions f and g.

Case 1: $f = g$.

This is the main case considered in the literature (Wagner (1983), Lindquist (1986), Klingenberg (1986)). Analytic solutions have been given when

1. $f'' > 0$; the construction of the solution gives a well defined function which satisfies the entropy condition (see Wagner (1983)).
2. f has at most one inflection point; the solution can be characterized in terms of non-linear waves (shocks and rarefactions) analogously to the 1-D Riemann problem. (See (Lindquist (1986)).
3. no restriction on f. (See Chang & Klingenberg (1986).)

Case 2: f close to g.

This only holds for certain orderings of $u1, u2, u3, u4$. For every function h such that $h'' > 0$, $\exists \epsilon > 0$ such that $\|f - h\|_{C_2} < \epsilon$ and $\|g - h\|_{C_2} < \epsilon$ which implies that Wagner's construction produces a well defined function which satisfies the entropy condition.

Case 3: f, g polynomials and $f \neq g$.

Klingenberg also considered the case where $f \neq g$ and f, g are both polynomials. See (Hsaio & Klingenberg (1984)) for details. Klingenberg considered the cases where $f = u^2, g = u^3$ as conjectures that this is the generic case. This would mean that there would be little more complication in considering other polynomials (see also Hsaio & Klingenberg (1984)).

Conservation laws with piecewise constant data have also been studied by (Guckenheimer (1975)), but the initial data is of a different form than (6.49). For example Guckenheimer has only three regions for his calculations but shows that the solution to the conservation law remains very complicated.

6.6.3 Examples

The examples given here all have analytic solutions which can be calculated. We will only consider the problems where either

$$u_1 < u_2 < u_4 < u_3 \tag{6.50}$$

or

$$u_1 < u_4 < u_2 < u_3 \tag{6.51}$$

since these give straight line shocks (Wagner (1983)) without expansions. This allows the solution to the problem to be calculated relatively simply.

Problem 1

Let us consider the inviscid Burgers' equation in 2-D

$$u_t + (\tfrac{1}{2}u^2)_x + (\tfrac{1}{2}u^2)_y = 0 \tag{6.52}$$

with initial data given by

$$u_0(x,y) = \begin{cases} u_1 = -1 & x, y > 0 \\ u_2 = \frac{1}{2} & y > 0, x < 0 \\ u_3 = 1 & x, y < 0 \\ u_4 = 0 & y < 0, x > 0 \end{cases} \tag{6.53}$$

on region $[-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}]$. The characteristics for (6.52) are given by

$$\frac{\partial u}{\partial t} = 0 \quad \frac{\partial x}{\partial t} = u \quad \frac{\partial y}{\partial t} = u \tag{6.54}$$

whose solution is the straight lines given by

$$u = u_0, \quad x = u_0 t + x_0, \quad y = u_0 t + y_0. \tag{6.55}$$

This problem has an analytic solution, and the solution at time t is given in Fig. 6.2. The points A and B given in the diagram above are given by

$$A = (\frac{-t}{4}, \frac{3t}{4}), \qquad B = (\frac{t}{2}, \frac{-t}{2}) \tag{6.56}$$

and P is at the origin. These points are calculated from the shock speeds found at the jumps in the initial data.

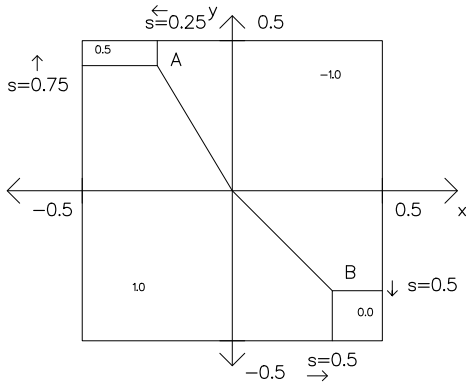

Figure 6.2: Solution for a 2-D Riemann Problem.

Problem 2

We will again consider the inviscid Burgers' equation in 2-D with initial data given by

$$u_0(x,y) = \begin{cases} u_1 = -\frac{1}{2} & x, y > 0 \\ u_2 = \frac{1}{2} & y > 0, x < 0 \\ u_3 = 1 & x, y < 0 \\ u_4 = 0 & y < 0, x > 0 \end{cases} \tag{6.57}$$

on the region $[-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}]$. The characteristics are the straight lines

$$u = u_0, \quad x = u_0 t + x_0, \quad y = u_0 t + y_0. \tag{6.58}$$

The solution can be found by applying the jump condition along the shock interfaces to find the speed with which the shocks are moving. The solution is shown in Fig. 6.3 below, where A and B denote the intersection of the shocks. A and

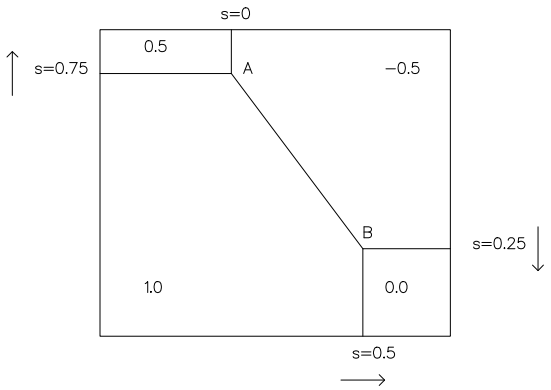

Figure 6.3: Solution for a 2-D Riemann Problem.

B are found to be

$$A = (0, 0.75t) \quad \text{and} \quad B = (0.5t, -0.25t) \tag{6.59}$$

at time t. This solution is given in (Wagner (1983)) for various cases of u. For more detail and explanation of calculations of expansions see (Wagner (1983)).

6.7 Legendre Transformation

Consider now the Legendre Transformation between the variables x, y and m, n with dual functions $u(x, y)$, $v(m, n)$ which satisfy

$$u(x, y) - mx - ny + v(m, n) = 0 \tag{6.60}$$

with $m = \frac{\partial u}{\partial x}$, $n = \frac{\partial u}{\partial y}$, $x = \frac{\partial v}{\partial m}$ and $y = \frac{\partial v}{\partial n}$. Define the coordinate transformation

$$x, y, t \to \xi, \eta, \tau \quad \text{and} \quad m, n, t \to \mu, \nu, \tau \tag{6.61}$$

by

$$x = \hat{x}(\xi, \eta, \tau) \tag{6.62}$$

$$y = \hat{y}(\xi, \eta, \tau) \quad t = \tau, \quad \hat{u}(\xi, \eta, \tau) = u(x, y, t) \tag{6.63}$$

$$m = \hat{m}(\mu, \nu, \tau) \tag{6.64}$$

$$n = \hat{n}(\mu, \nu, \tau) \quad t = \tau, \quad \hat{v}(\mu, \nu, \tau) = v(x, y, t). \tag{6.65}$$

Now differentiate (6.60) with respect τ to give

$$\dot{u} - \dot{m}\hat{x} - \dot{n}\hat{y} - \dot{x}\hat{m} - \dot{y}\hat{n} + \dot{v} = 0. \tag{6.66}$$

Let us consider the equation

$$u_t + H(x, y, u, u_x, u_y) = 0 \tag{6.67}$$

which includes the conservation laws. Write this in the dual variables

$$-\dot{v} + x\dot{m} + y\dot{n} + H(x, y, m, n) = 0 \tag{6.68}$$

where $x = v_m$ and $y = v_n$.

We are using this transformation partly to see if a simplified set of equations can be found and partly to link with the exact solution procedure in section 6.4. It is also a useful viewpoint when we come to consider the MFE approximation in chapter 7. Initially let us consider the special case of the inviscid Burgers' equation

$$u_t + uu_x + uu_y = 0 \tag{6.69}$$

which can be written in the dual variables (c.f. (6.68)) as

$$-\dot{v} + x\dot{m} + y\dot{n} + um + un = 0. \tag{6.70}$$

The u can be substituted using the transformation (6.60) to give

$$-\dot{v} + x\dot{m} + y\dot{n} + (mx + ny - v)m + (mx + ny - v)n = 0. \tag{6.71}$$

We can now compare the coefficients of x, y and the constant term to give

$$\dot{m} + m^2 + mn = 0 \tag{6.72}$$

$$\dot{n} + n^2 + mn = 0 \tag{6.73}$$

$$\dot{v} + vm + vn = 0. \tag{6.74}$$

Equations (6.72) and (6.73) can be written in the same form as the equations describing blow-up in section 6.4. Add equations (6.72) and (6.73) together to give

$$\dot{m} + \dot{n} = -m^2 - 2mn - n^2 \tag{6.75}$$

$$\Rightarrow \frac{d}{d\tau}(m + n) = -(m + n)^2. \tag{6.76}$$

Now let $q = m + n$ which gives the equation

$$\dot{q} = -q^2 \tag{6.77}$$

which is the equation used to describe blow-up in section 6.4. Equation (6.74) can also be written in terms of q as

$$\dot{v} = -vq. \tag{6.78}$$

These equations can be solved to give the solution in terms of q and v which may then be transformed back into u, x variables. Equation (6.77) can be solved to

give

$$q = \frac{q_0(\xi)}{1 + q_0(\xi)\tau} \tag{6.79}$$

$$\Rightarrow m + n = \frac{q_0(\xi)}{1 + q_0(\xi)\tau} \tag{6.80}$$

where q_0 is the initial value of q. Equation (6.78) can now be solved using (6.79) to give

$$\frac{\dot{v}}{v} = \frac{q_0(\xi)}{1 + q_0(\xi)\tau} \tag{6.81}$$

$$\Rightarrow \ln v = \ln(1 + q_0\tau) + \ln C(\xi) \tag{6.82}$$

$$\Rightarrow v = C(\xi)(1 + q_0\tau). \tag{6.83}$$

v can also be written in terms of m and n as $v = \frac{C(\xi)q_0}{m+n}$. The solution can now be found for m and n, by using equations (6.72) and (6.73) to give m as

$$\frac{\dot{m}}{\dot{n}} = \frac{m}{n} \Rightarrow m = A(\xi)n \tag{6.84}$$

(c.f. (6.46)). Now using this and (6.80), we can solve for m and n.

For general conservation laws the above method may also be carried out numerically either by means of a projection, or pointwise, in order that a piecewise linear representation can be found. A similar set of equations to the blow-up equations is found after the projection is applied. This will be considered in more detail in chapter 7. Note: if u is absent from $H(x, u, u_x)$ the characteristic equations are the Hamiltons equations (Courant & Hilbert (1962)),

$$\dot{x} = \frac{\partial H}{\partial u_x} \qquad \dot{y} = \frac{\partial H}{\partial u_y} \tag{6.85}$$

$$\dot{u_x} = -\frac{\partial H}{\partial x} \qquad \dot{u_y} = -\frac{\partial H}{\partial y}. \tag{6.86}$$

If equation (6.85) and the Legendre transformation given by (6.60) are combined then (6.86) can be obtained. Similarly (6.86) and (6.60) can be combined to give (6.85).

6.8 Summary

This chapter has discussed the analytic solution of conservation laws in 2-D. In summary we have found that, although for a general conservation law a solution

may exist, there is not necessarily a method of finding it. From the above discussion it can be seen that both shocks and expansions form in 2-D and that it is in general very complicated to calculate solutions to these problems. This leads us to consider numerical techniques of solution in the following two chapters.

Blow-up and the existence of discontinuities within finite time are considered. The Legendre transformation into v, m space is also investigated within the special case of the Inviscid Burgers' equation; links are found to exist between the equations obtained and those found from the earlier discussion of blow-up. The v, m Legendre transformation will again be considered in chapter 7 when the projection of a general equations into a piecewise linear space is introduced.

Chapter 7

Moving Finite Element Methods In Higher Dimensions

7.1 Introduction

In two and higher dimensions we concentrate only on moving finite element methods. These methods are chosen because we wish to develop the ideas described in 1-D (chapters 2-5). For this it is again necessary to have an adaptive method which will also allow solutions to overturn. The standard MFE method generalizes to higher dimensions and mostly follows the 1-D structure and ideas described in chapter 3. However, in 2-D and higher dimensions the global and local MFE methods differ due to the structure of the matrices formed during the implementation of the method. Moreover, the simple problems in 1-D which involved node overtaking now incur triangle folding in 2-D, and the case of parallelism increases in complexity.

We will first discuss the basic types of MFE in both two and higher dimensions, highlighting the differences between these and the 1-D methods. In particular we will examine the problems of implementation of MFE in higher dimensions.

In 2-D MFE was introduced by Alexander, Manselli & Miller (1979), continued by Djomeri, Doss, Gelinas, & Miller (1985) and developed by Carlson & Miller (1986). In 1-D there were several variations on the MFE method described in chapter 3. These have all been extended to two and higher dimensions and include Local MFE, Global MFE, GWMFE and Split MFE.

7.2 Introduction Of Problem

Analogously to the 1-D case the equation we are solving is of the form

$$u_t - \mathcal{L}(u) = 0 \tag{7.1}$$

in n space dimensions $u = u(\mathbf{x}, t)$ on the region $\Omega \in \mathbb{R}^n$, with boundary conditions given on $\partial\Omega$ and initial data given at $t = 0$, where $\mathbf{x} = (x_1, ..., x_n)$ and the possibly nonlinear operator $\mathcal{L}$ contains all first derivatives of only first order $u_{x_1}, ..., u_{x_n}$. In the 2-D case $u = u(x, y, t)$ and $\mathcal{L}$ contains derivatives u_x, u_y. We will not consider problems containing second order terms in two or higher dimensions here. The problems in 2-D of representing u_{xx} and u_{yy} in terms of piecewise linear basis functions have been discussed in (Johnson, Wathen & Baines (1988)). Nor will we consider systems here, for discussion of the issues involved see (Edwards (1988)).

Equation (7.1) is a scalar, nonlinear PDE whose solution evolves with time. This type of equation may form solutions which contain discontinuities or expansions (see chapter 6). Although we will discuss the MFE method for equations of the general form of (7.1), we are really interested in overturning solutions which are provided in the simplest manner by conservation laws. We will therefore consider equations of the form

$$u_t + \text{div}(\mathbf{f}(u)) = 0 \tag{7.2}$$

where $\mathbf{f}$ is a function of u.

For such conservation laws, many finite element methods imitate the conservation. To show this, let us consider a Galerkin weak form of equation (7.2)

$$\int_\Omega (u_t + \text{div}(\mathbf{f}(u)))\alpha_i d\Omega = 0 \quad i = 1, ..., N \tag{7.3}$$

where α_i is a trial function with $\sum_{i=1}^N \alpha_i = 1$. Now summing over all elements gives

$$\sum_{i=1}^N \int_\Omega (u_t + \text{div}(\mathbf{f}(u)))\alpha_i d\Omega = 0 \tag{7.4}$$

$$\Rightarrow \int_\Omega (u_t + \text{div}(\mathbf{f}(u))) \sum_{i=1}^N \alpha_i d\Omega = 0 \tag{7.5}$$

$$\Rightarrow \int_\Omega (u_t + \text{div}(\mathbf{f}(u)))d\Omega = 0 \tag{7.6}$$

$$\Rightarrow \frac{d}{dt}\int_\Omega u d\Omega + \oint \mathbf{f}(u).d\mathbf{S} = 0. \tag{7.7}$$

This shows that the initial data is conserved for all the trial functions which satisfy $\sum_{i=1}^{N} \alpha_i = 1$ (apart as usual from the boundary conditions) (c.f. chapter 6 section 6.3.1).

7.3 Global MFE

The 1-D Global MFE method generalizes to higher dimensions straight forwardly so that the ideas and structures remain similar although the implementation increases in complexity. We seek a continuous piecewise linear approximation U to u, on simplex elements (so in 2-D the elements are triangles, in 3-D the elements are tetrahedra, etc.). Let U be an approximation of the form

$$U = \sum_{j=1}^{N} a_j(t)\alpha_j(\mathbf{r}, \mathbf{s}(t)) \tag{7.8}$$

where a_j $(j = 1, ..., N)$ are the nodal amplitudes, and $\alpha_j(\mathbf{r}, \mathbf{s}(t))$ are basis functions. Here $\mathbf{r} = (x_1, ..., x_n)$ is the position vector of a point, N is the number of nodes, and $\mathbf{s}$ contains the nodal position vectors $\mathbf{s} = (\mathbf{s_1}, ..., \mathbf{s_N})$, where each element of $\mathbf{s}$ is given in coordinate form by $\mathbf{s}_j = (X_{1j}, ..., X_{nj})$. Also in 2-D $\mathbf{r} = (x, y)$ and each element of $\mathbf{s}$ is given by $\mathbf{s}_j = (X_j, Y_j)$. In 2-D α_j is the pyramidal piecewise linear finite element basis function taking the value 1 at node j and 0 at surrounding nodes (see Fig. 7.1). Differentiating (7.8) with respect to t gives

$$U_t = \sum_{j=1}^{N} \dot{a}_j(t)\frac{\partial U}{\partial a_j} + \dot{\mathbf{s}}_j(t).\nabla_{\mathbf{s}_j} U \tag{7.9}$$

which in 2-D becomes

$$U_t = \sum_{j=1}^{N} (\dot{a}_j\alpha_j - \dot{X}_j\frac{\partial U}{\partial X} - \dot{Y}_j\frac{\partial U}{\partial Y}). \tag{7.10}$$

This can be shown to reduce to (Miller & Miller (1981), Baines & Wathen (1988))

$$U_t = \sum_{j=1}^{N} (\dot{a}_j\alpha_j + \dot{X}_j\beta_j + \dot{Y}_j\gamma_j) \tag{7.11}$$

where β_j and γ_j are piecewise linear basis functions defined within each element by

$$\beta_j = -\frac{\partial U}{\partial X}\alpha_j \qquad \gamma_j = -\frac{\partial U}{\partial Y}\alpha_j. \tag{7.12}$$

Note: $\frac{\partial U}{\partial X}$ and $\frac{\partial U}{\partial Y}$ are the components of the gradient of U on each element in the X and Y directions respectively. The β_j and γ_j basis functions have the same support as α_j, but are discontinuous across all the element edges through node j. See Fig. 7.1.

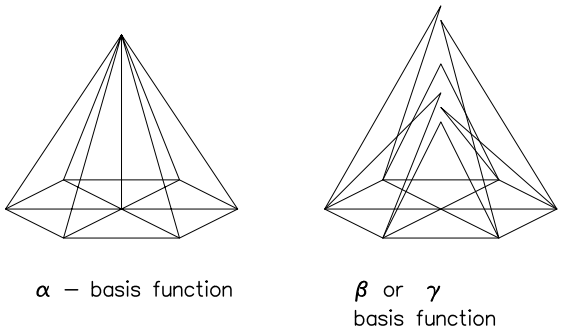

Figure 7.1: Basis functions in 2-D.

Again, following the 1-D approach, we minimise the residual

$$\|U_t - \mathcal{L}(U)\|^2 \tag{7.13}$$

with respect to $\dot{a}_j$, $\dot{s}_{jm}$ $(m = 1, ..., n\ j = 1, ..., N)$ in n space dimensions using N nodes, giving rise to the $N(n+1)$ equations

$$< U_t - \mathcal{L}(U), \alpha_j > \ = \ 0 \tag{7.14}$$

$$< U_t - \mathcal{L}(U), \beta_{jm} > \ = \ 0 \tag{7.15}$$

where β_{jm} $(j = 1, ..., N\, m = 1, ..., n)$ are the n piecewise linear discontinuous basis functions (c.f. β_j, γ_j in 2-D) defined by $\beta_{jm} = -\frac{\partial U}{\partial x_m}\alpha_j$. In 2-D, after minimising (7.13) with respect to a_j, X_j, Y_j $(j = 1, ..., N)$, we have $3N$ equations

$$\left.\begin{array}{l} < U_t - \mathcal{L}(U), \alpha_j >= 0 \\ < U_t - \mathcal{L}(U), \beta_j >= 0 \\ < U_t - \mathcal{L}(U), \gamma_j >= 0 \end{array}\right\} j = 1, ..., N. \tag{7.16}$$

The sets of equations (7.14), (7.15) or (7.16) can again be written in the form of a system

$$A(\mathbf{y})\dot{\mathbf{y}} = \mathbf{g}(\mathbf{y}) \tag{7.17}$$

where $\mathbf{y} = (a_1, \mathbf{s}_1; ...; a_N, \mathbf{s}_N)^T$ in n-D and $\mathbf{y} = (a_1, X_1, Y_1; ...; a_N, X_N, Y_N)^T$ in 2-D.

In both cases A is square and symmetric, and consists of inner products of basis functions in blocks. A is also positive semi-definite because it arises from the minimisation of the form $\|U_t\|^2$ in (7.13), which is the quadratic term $\dot{\mathbf{y}}^T A\dot{\mathbf{y}}$. This is only zero when $\dot{\mathbf{y}} \neq 0$ exists such that $\dot{\mathbf{y}}^T A\dot{\mathbf{y}} = 0$ (i.e. when A is singular). In n-dimensions, the (i,j)th block contains $(n+1)^2$ elements and is given by

$$A_{ij} = \begin{pmatrix} <\alpha_i,\alpha_j> & <\beta_{i1},\alpha_j> & ..... & <\beta_{1n},\alpha_j> \\ <\alpha_i,\beta_{j1}> & <\beta_{i1},\beta_{j1}> & ..... & <\beta_{1n},\beta_{j1}> \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ <\alpha_i,\beta_{jn}> & <\beta_{in},\beta_{jn}> & ..... & <\beta_{1n},\beta_{jn}> \end{pmatrix}. \tag{7.18}$$

The righthand side vector is given by $\mathbf{g}(\mathbf{y}) = (\mathbf{g}_1, ..., \mathbf{g}_N)^T$ where

$$\mathbf{g}_j = \begin{pmatrix} <\mathcal{L}(U),\alpha_j> \\ <\mathcal{L}(U),\beta_{j1}> \\ \vdots \\ <\mathcal{L}(U),\beta_{jn}> \end{pmatrix}. \tag{7.19}$$

Before considering the structure of the matrices given by the MFE method and subsequently the method of solution (see section 7.5), let us discuss an elementwise construction.

7.4 Local Basis Functions

In order to introduce an elementwise construction, it is first necessary to define elementwise basis functions (Baines & Wathen (1988)). Let $\phi_k^{(\nu)}$ be a linear elementwise basis function, with support only on element k, with the values 1 at corner ν and 0 at the other corners. In 2-D element k with local node numbering $\nu = 1, 2, 3$ has the local basis functions shown in Fig. 7.2. Discontinuous approximations to u and u_t may now be written in terms of these basis functions and hence a discontinuous version of the method above can be written in terms of ϕ's. Later we apply constraints to give continuity of U.

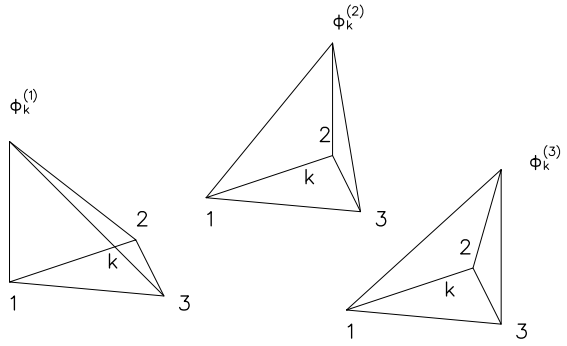

Figure 7.2: Local element basis functions in 2-D.

The approximation U_t is then rewritten in terms of the local basis functions

$$U_t = \sum_{k=1}^{E} \sum_{\nu=1}^{n+1} w_k^{(\nu)}(t) \phi_k^{(\nu)}(\mathbf{r}, \mathbf{s}(t)) \tag{7.20}$$

where E is the number of elements, n is the dimension, $\mathbf{r}$ is the position vector of a point and $\mathbf{s}$ contains the positions of the nodes.

If we now minimise (7.13) using the new approximation to U_t, over $w_k^{(\nu)}$ ($\nu = 1, 2, 3, k = 1, ..., E$) we have

$$< \phi_k^{(1)}, U_t - \mathcal{L}(U) > \quad = 0 \tag{7.21}$$

$$< \phi_k^{(2)}, U_t - \mathcal{L}(U) > \quad = 0 \tag{7.22}$$

$$< \phi_k^{(3)}, U_t - \mathcal{L}(U) > \quad = 0. \tag{7.23}$$

In n-D this generalizes to

$$< \phi_k^{(i)}, U_t - \mathcal{L}(U) >= 0 \qquad i = 1, ..., n. \tag{7.24}$$

In both cases (7.21), (7.22), (7.23) or (7.24), the equations can be written as the system

$$C_k \mathbf{w}_k = \mathbf{b}_k, \tag{7.25}$$

where C_k is an $(n+1) \times (n+1)$ square element mass matrix with inner-products as entries (except possibly at the boundaries where the size may be reduced). If the basis functions are ordered in a corresponding way to the w's then

$$\mathbf{w}_k = (w_1^{(1)}, ..., w_1^{(n+1)}; ...; w_E^{(1)}, ..., w_E^{(n+1)})^T \tag{7.26}$$

and $\mathbf{b} = \{\mathbf{b}_k\}$ where

$$\mathbf{b}_k = \begin{pmatrix} < \phi_k^{(1)}, \mathcal{L}(U) > \\ \vdots \\ < \phi_k^{(n+1)}, \mathcal{L}(U) > \end{pmatrix}. \tag{7.27}$$

Since the $\phi_k^{(\nu)}$'s have local support the inner products in C are zero unless the ϕ's belong to the same element and C becomes a block diagonal matrix. This gives C as $C = \{C_k\}$ where $C_k = \{C_{ki}\}$ and

$$C_{ki} = \begin{pmatrix} < \phi_k^{(1)}, \phi_i^{(1)} > & < \phi_k^{(2)}, \phi_i^{(1)} > & .... & .... & < \phi_k^{(n)}, \phi_i^{(1)} > \\ < \phi_k^{(1)}, \phi_i^{(2)} > & & & & \vdots \\ \vdots & & & & \vdots \\ < \phi_k^{(1)}, \phi_i^{(n+1)} > & .... & .... & .... & < \phi_k^{(n+1)}, \phi_i^{(n+1)} > \end{pmatrix}. \tag{7.28}$$

In 2-D the C matrix is block diagonal with blocks

$$C_k = \frac{\Delta_k}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \tag{7.29}$$

where Δ_k is the area of the element k.

Now, as we shall see, enforcing the constraint that (7.20) is the same as (7.9) gives

$$M_j \dot{\mathbf{y}}_j = \mathbf{w}_j \tag{7.30}$$

where M_j is a rectangular matrix obtained by writing $\alpha_j, \boldsymbol{\beta}_j$ where $\boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jn})$ in terms of $\phi_k^{(\nu)}$ $(\nu = 1, ..., n+1)$. The M matrix is made up from the rectangular blocks M_j as 'diagonal' entries, using nodal numbering.

From (7.9)

$$U_t = \sum_{j=1}^{N} (\dot{a}_j \alpha_j + \boldsymbol{\beta}_j . \dot{\mathbf{s}}_j) \tag{7.31}$$

$$= \dot{\mathbf{y}}^T \boldsymbol{\alpha} \tag{7.32}$$

where $\boldsymbol{\alpha} = (\alpha_1, \boldsymbol{\beta}_1^T; ...; \alpha_N, \boldsymbol{\beta}_N^T)$. Using the definition of the $\boldsymbol{\beta}_j$ basis functions (7.31) becomes

$$U_t = \sum_{j=1}^{N} (\dot{a}_j - \nabla U . \dot{s}_j) \alpha_j. \tag{7.33}$$

Similarly, from (7.20)

$$U_t = \mathbf{w}^T \boldsymbol{\phi} \tag{7.34}$$

where $\boldsymbol{\phi} = \{\phi_k^{(\nu)}(\mathbf{r}, \mathbf{s}(t))\}$ Since each α_j is the sum of some sets of $\phi_k^{(\nu)}$ we can write

$$\boldsymbol{\alpha} = M^T \boldsymbol{\phi}, \tag{7.35}$$

where M is a rectangular matrix of 1's, 0's and components of $\boldsymbol{\nabla} U$. M is block diagonal (using nodal numbering) with blocks M_j. Since, further, $A =< \boldsymbol{\alpha} , \boldsymbol{\alpha} >$ it follows that

$$A = M^T C M \tag{7.36}$$

also holds in two and higher dimensions.

The local MFE method is obtained by using the local basis functions and will be described in section 7.9. We will now use the decompostion of the global MFE matrix A to examine the solution of the equations obtained from that system.

7.5 Solution Of MFE Equations

We now return to the global MFE equations (7.14), (7.15). In 1-D M and C are block 2×2 diagonal matrices. From this form it was noted in chapter 3 that the equations decouple into trivial 2×2 matrix systems in general. The solution is therefore straightforward, although there do arise some situations in which the 2×2 matrices cannot be inverted, namely, when either M or C are singular due to the configuration of the solution.

Since it has been shown that the decomposition $A = M^T C M$ occurs for higher dimensions, although the matrix M does not remain square, it is possible to exploit this form in the solution of the system $A\dot{\mathbf{y}} = \mathbf{g}$.

7.5.1 Non-singular A

If A is non-singular then the system

$$A\dot{\mathbf{y}} = \mathbf{g} \tag{7.37}$$

can be solved for $\dot{\mathbf{y}}$. However for $n \geq 2$ there is no longer the decoupling and simple inversion technique which exists in 1-D. The size of the matrix A will

usually be very large (if any physical problem is tackled) and hence an efficient solver is required. The best solver is found to be the conjugate gradient method with diagonal preconditioning, using the property of eigenvalue clustering given by (Wathen (1987)) which is solution and mesh geometry independent.

If we consider the matrix D, which contains all the diagonal blocks of A, then the eigenvalue clustering results (given in chapter 3 section 3.2.1) (Wathen (1984)) extended to higher dimensions, prove that the eigenvalue spectral radius ρ of $D^{-1}A$ always satisfies

$$\rho(D^{-1}A) \in [\tfrac{1}{2}, 1 + \frac{n}{2}] \tag{7.38}$$

where n is the number of space dimensions. This result is independent of the number of nodes and the mesh configuration and is valid providing there is no element folding or parallelism. This result indicates that the conjugate gradient method (Golub & Van Loan (1983)) with D^{-1} preconditioning should converge very rapidly, which is borne out in practice (Johnson, Wathen & Baines (1988)).

7.5.2 A Is Singular

Now consider the case where A is singular (see Wathen (1984)). Returning to the decomposition

$$A = M^T C M, \tag{7.39}$$

if A is singular, then

$$\exists \dot{\mathbf{y}} \neq 0 \text{ such that } A\dot{\mathbf{y}} = 0 \tag{7.40}$$

$$\Rightarrow \quad \dot{\mathbf{y}}^T A \dot{\mathbf{y}} = 0 \tag{7.41}$$

$$\Rightarrow \quad \dot{\mathbf{y}}^T M^T C M \dot{\mathbf{y}} = 0 \tag{7.42}$$

$$\Rightarrow \exists \mathbf{w} \text{ such that } \quad \mathbf{w}^T C \mathbf{w} = 0 \tag{7.43}$$

where

$$\mathbf{w} = M\dot{\mathbf{y}}. \tag{7.44}$$

From (7.43) and (7.44) it can be seen that either C is singular or $\mathbf{w}$ is zero (with $\dot{\mathbf{y}} \neq 0$), i.e. M is column rank deficient (However the converse is not always true as we have already seen in 1-D, chapter 3 section 3.5.)

M is column rank deficient

To examine this problem, it is necessary to first find the form of M. Since M is given by writing the α_j, $\boldsymbol{\beta}_j$ functions in terms of the local basis functions $\phi_k^{(\nu)}$, consider the identity

$$U_t = \sum_{j=1}^{N}(\dot{a}_j\alpha_j + \sum_{m=1}^{n}\dot{s}_{jm}\beta_{jm}) = \sum_{k=1}^{E}\sum_{\nu=1}^{n} w_k^{(\nu)}\phi_k^{(\nu)}. \tag{7.45}$$

This gives

$$\sum_{\nu=1}^{N} w_l^{(\mu)}\phi_l^{(\mu)} = \dot{a}_i\alpha_i + \sum_{m=1}^{n}\dot{s}_{im}\beta_{im} \tag{7.46}$$

$$= \left(\dot{a}_i - \sum_{m=1}^{n}\frac{\partial U}{\partial x_m}\dot{s}_{im}\right)\alpha_i. \tag{7.47}$$

By the definitions of the α_i and $\phi_l^{(\mu)}$ functions

$$\phi_l^{(\mu)} = \alpha_i \tag{7.48}$$

on element l. Then (7.47) becomes

$$w_l^{(\mu)} = \dot{a}_i - \sum_{m=1}^{n}\frac{\partial U}{\partial x_m}\dot{s}_{im}. \tag{7.49}$$

In the matrix M, therefore, the row corresponding to $w_l^{(\mu)}$ contains

$$(1, -\frac{\partial U}{\partial x_1}, ..., -\frac{\partial U}{\partial x_n}) = \mathbf{p}_l^T, \tag{7.50}$$

say. If we assume this row lies in the ith column block, then for each element k surrounding node i a vector similar to $\mathbf{p}_l^T$ will also appear in the ith column block. The matrix M can therefore be reordered into a block diagonal matrix by reordering the elements of $\mathbf{w}$ into a nodewise list. This means that there exists a permutation matrix Q such that $N = QM$ where N has rectangular blocks. The block ik has the same number of rows as elements surrounding node i and $n+1$ columns corresponding to $\mathbf{p}_k^T$.

We now consider the case when M or N becomes column rank deficient. This is equivalent to considering the row rank deficiency of $\{\mathbf{p}_l^T\}$, since M is column rank deficient if there exists a non zero linear combination of the components of $\mathbf{p}_l^T$. This gives rise to two levels of parallelism, which is different to the simple case which occurred in 1-D. Following Wathen (1984) we will give an example in

2-D for clarity and in order to emphasize the differences between one and higher dimensions.

Let us consider the system

$$\mathbf{w} = M\dot{\mathbf{y}} \tag{7.51}$$

where the blocks of M are

$$\begin{pmatrix} \vdots & \vdots & \vdots \\ 1 & -m_k & -n_k \\ \vdots & \vdots & \vdots \end{pmatrix}. \tag{7.52}$$

Then M is column rank deficient if, for all pairs of elements k and l adjacent to node j, (i) $m_k = m_l$ and $n_k = n_l$ or (ii) $\exists\ \lambda, \mu \in \mathbb{R}$ such that λ, μ are not both zero such that $\lambda m_k + \mu n_k = \lambda m_l + \mu n_l$. As in 1-D this type of degeneracy may again be dealt with by considering a block of A

$$A_{ij} = \begin{pmatrix} <\alpha_i, \alpha_j> & <\alpha_i, \beta_j> & <\alpha_i, \gamma_j> \\ <\beta_i, \alpha_j> & <\beta_i, \beta_j> & <\beta_i, \gamma_j> \\ <\gamma_i, \alpha_j> & <\gamma_i, \beta_j> & <\gamma_i, \gamma_j> \end{pmatrix}. \tag{7.53}$$

There are two cases to consider.

Case (i)

When $m_k = m_l$ and $n_k = n_l$ then β_j, γ_j are parallel to α_j in all elements k, l surrounding node j. In this case there is a unique m and n $(\forall k, l)$ and the solution in the whole patch of elements surrounding node j is coplanar.

Hence the MFE equations

$$< U_t - \mathcal{L}(U), \alpha_j >= 0 \tag{7.54}$$

$$< U_t - \mathcal{L}(U), \beta_j >= 0 \tag{7.55}$$

$$< U_t - \mathcal{L}(U), \gamma_j >= 0 \tag{7.56}$$

are linearly dependent. The β and γ equations may be omitted, together with the corresponding columns of A. This gives a non-singular matrix and the solution is consistent with any values of $\dot{X}$ and $\dot{Y}$, which may then be chosen arbitarily.

A reduced system

$$A^*(\dot{\mathbf{y}}^*)\mathbf{y}^* = \mathbf{g}^*(\mathbf{y}^*) \tag{7.57}$$

is therefore solved (as in 1-D), to give a complete solution

$$\dot{\mathbf{y}} = \dot{\mathbf{y}}^* + \mathbf{c}^T \mathbf{u}_j \tag{7.58}$$

where $\mathbf{u}_j = (\mathbf{u}_j^1, \mathbf{u}_j^2)^T$ is the null vector of the full system, which has components of the form

$$\mathbf{u}_j^1 = (0,0,0;...;m_j,1,0;...;0,0,0)^T \tag{7.59}$$

$$\mathbf{u}_j^2 = (0,0,0;...;n_j,0,1;...;0,0,0)^T \tag{7.60}$$

and $\mathbf{c} = (c_1, c_2)^T$ can be chosen to satisfy some external criterion, e.g. averaging.

Case (ii)

In this case $\exists\ \lambda, \mu$, where λ and μ are not both zero such that

$$\lambda m_k + \mu n_k = \lambda m_l + \mu n_l. \tag{7.61}$$

It follows that $\lambda\beta_j + \mu\gamma_j$ is parallel to α_j in all elements k, l surrounding node j. The vectors $p_k = (1, -m_k, -n_k)^T\ \forall k$ surrounding node j span a 2-D space and the null space is the orthogonal space. The null space may be spanned by

$$\mathbf{n} = [m_k(n_k - n_l) + (m_k - m_l)n_k, n_k - n_l, m_k - m_l] \tag{7.62}$$

where k, l are chosen such that $\mathbf{p}_k \neq \mathbf{p}_l$. The set can then be written in the form

$$\dot{\mathbf{y}} = \dot{\mathbf{y}}^* + c\mathbf{u}_j \tag{7.63}$$

where $\mathbf{y}^*$ is the solution of the reduced system and

$$\mathbf{u}_j = (0,0,0;...;\mathbf{n};...;0,0,0) \tag{7.64}$$

and c is a constant chosen to satisfy external criteria.

7.5.3 C Is Singular

We have shown that if A is singular, then $M^T C M$ is singular which implies that either C or M is singular. If C is singular then $\Delta_k = 0$ for some k. i.e. one of the triangles has zero area.

On the other hand all this depends on M and C being independent. However, in the overturning case, M, C are obviously dependent so we will reconsider the singularities of A.

If we consider the global matrix A, then it has been shown that it has the decomposition $A = M^T C M$ where

$$A_{ij} = \begin{pmatrix} 1 & -m_{j1} & -n_{j1} \\ \vdots & \vdots & \vdots \\ 1 & -m_{jp} & -n_{jp} \\ \vdots & \vdots & \vdots \\ 1 & -m_{jt} & -n_{jt} \end{pmatrix}^T \frac{\Delta_{ji}}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & -m_{j1} & -n_{j1} \\ \vdots & \vdots & \vdots \\ 1 & -m_{jp} & -n_{jp} \\ \vdots & \vdots & \vdots \\ 1 & -m_{jt} & -n_{jt} \end{pmatrix} \tag{7.65}$$

are the blocks of the M and C matrices and t is the number of elements surrounding a node. A can be rewritten as

$$A = M^T (D^T E D) M \tag{7.66}$$

so that the blocks are

$$A_k = M_k^T \begin{pmatrix} \Delta_k^{\frac{1}{2}} & 0 & 0 \\ 0 & \Delta_k^{\frac{1}{2}} & 0 \\ 0 & 0 & \Delta_k^{\frac{1}{2}} \end{pmatrix} \frac{1}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} \Delta_k^{\frac{1}{2}} & 0 & 0 \\ 0 & \Delta_k^{\frac{1}{2}} & 0 \\ 0 & 0 & \Delta_k^{\frac{1}{2}} \end{pmatrix} M_k \tag{7.67}$$

which is

$$A_k = M_k^T D_k^T E_k D_k M_k \tag{7.68}$$

where $D = \text{diag}\{D_k\}$ and $E = \text{diag}\{E_k\}$.

Let $N = DM$ and collect the terms of N from the global A matrix in a nodewise manner to give

$$N_j = \begin{pmatrix} \Delta_{j1}^{\frac{1}{2}} & -\Delta_{j1}^{\frac{1}{2}} m_{j1} & -\Delta_{j1}^{\frac{1}{2}} n_{j1} \\ \vdots & \vdots & \vdots \\ \Delta_{jl}^{\frac{1}{2}} & -\Delta_{jl}^{\frac{1}{2}} m_{jl} & -\Delta_{jl}^{\frac{1}{2}} n_{jl} \\ \vdots & \vdots & \vdots \\ \Delta_{jt}^{\frac{1}{2}} & -\Delta_{jt}^{\frac{1}{2}} m_{jt} & -\Delta_{jt}^{\frac{1}{2}} n_{jt} \end{pmatrix}. \tag{7.69}$$

This gives

$$\dot{\mathbf{y}}^T M^T C M \dot{\mathbf{y}} = 0 \tag{7.70}$$

$$\Rightarrow \qquad \dot{\mathbf{y}}^T N^T E N \dot{\mathbf{y}} = 0 \tag{7.71}$$

$$\Rightarrow \quad \begin{cases} \mathbf{z}^T E \mathbf{z} = 0 \\ \quad \text{where} \;\; \mathbf{z} = N^T \dot{\mathbf{y}} \end{cases} \tag{7.72}$$

However it can be seen that E is non-singular which implies that $\mathbf{z} = 0$ with $\dot{\mathbf{y}} \neq 0$, which implies that N^T is rank deficient.

Since the singularity occurs when the areas tend to zero then it is useful to separate these from the other terms so let $m_{jl} = p_{jl}/\Delta_{jl}$ and $n_{jl} = q_{jl}/\Delta_{jl}$ with p and q non-zero in general. The gradients m_{jl} and n_{jl} can be chosen to give N with linearly independent columns.

In a similar manner to 1-D we will find the QR decomposition of the N matrix in order to show that there are conditions under which it is non-singular. Let us write N as a vector of three column vector $\mathbf{c}_1$, $\mathbf{c}_2$ and $\mathbf{c}_3$. The QR decomposition will be given by

$$N = (\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3) = QR = (\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2, \hat{\mathbf{q}}_3) \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{pmatrix}. \tag{7.73}$$

Now using Gram-Schmidt orthogonalization (Golub & Van Loan (1983)) to give the QR decomposition, we get

$$\hat{\mathbf{q}}_1 = \frac{\mathbf{c}_1}{c_1} \Rightarrow r_{11} = c_1 \text{ where } c_1 = |\mathbf{c}_1|. \tag{7.74}$$

The second vector becomes

$$r_{22}\hat{\mathbf{q}}_2 = \mathbf{c}_2 - (\hat{\mathbf{q}}_1.\mathbf{c}_2)\hat{\mathbf{q}}_1 \tag{7.75}$$

$$= \mathbf{c}_2 - \frac{(\mathbf{c}_1.\mathbf{c}_2)}{c_1^2}\mathbf{c}_1 \tag{7.76}$$

$$= \mathbf{C}. \tag{7.77}$$

Then $\hat{\mathbf{q}}_2 = \frac{\mathbf{C}}{C}$ which implies that $r_{22} = C$ where $C = |\mathbf{C}|$. The third vector is slightly more complicated and is

$$r_{33}\hat{\mathbf{q}}_3 = \mathbf{c}_3 - (\hat{\mathbf{q}}_1.\mathbf{c}_3)\hat{\mathbf{q}}_1 - (\hat{\mathbf{q}}_2.\mathbf{c}_3)\hat{\mathbf{q}}_2 \tag{7.78}$$

$$= \mathbf{c}_3 - \frac{(\mathbf{c}_1.\mathbf{c}_3)}{c_1^2}\mathbf{c}_1 - \frac{(\mathbf{C}.\mathbf{c}_3)}{C^2}\mathbf{C}. \tag{7.79}$$

To avoid the singularity (occuring as the solution becomes multivalued) we require $r_{11} \neq 0$, $r_{22} \neq 0$ and $r_{33} \neq 0$.

For $r_{11} \neq 0$ we require that $c_1 \neq 0$, which implies that $\sum_i |\Delta_i| \neq 0$ which means that at least one of the triangles surrounding the node must have non-zero area.

The other two cases are much more complicated and involve values of u: hence they are based upon the ideas of parallelism described in section 7.5.2. For some values of the gradients with only one non-zero area the solution can overturn without causing singularities.

The second case where $r_{22} = C$ is required to be non-zero has to be expanded. For C to be non-zero then

$$\mathbf{C} = \mathbf{c}_2 - \frac{(\mathbf{c}_1.\mathbf{c}_2)}{c_1^2}\mathbf{c_1} \tag{7.80}$$

which can be rewritten in terms of vector products as

$$\mathbf{C} = \frac{1}{c_1^2}\mathbf{c}_1 \times (\mathbf{c}_2 \times \mathbf{c}_1). \tag{7.81}$$

For $r_{22} \neq 0$ we require

a) $\mathbf{c}_1 \neq 0$

b) $\mathbf{c}_2 \times \mathbf{c}_1 \neq 0$

c) $\mathbf{c}_1$ not parallel to $\mathbf{c}_2$.

Now if $r_{11} \neq 0$ then (a) holds. If we consider (b) then we now require that $\mathbf{c}_1$, $\mathbf{c}_2$ to be non-zero and $\mathbf{c}_1$ not to be parallel to $\mathbf{c}_2$. we already require $\mathbf{c}_1$ to be non-zero, so now we also require $\mathbf{c}_2$ to be non-zero. Finally for (b) to hold we also require $\mathbf{c}_1 \neq \mathbf{c}_2$, which means that not all p_i/A_i can be equal. This corresponds to parallelism. (c) holds since by definition of the cross product $\mathbf{c}_1$ cannot be parallel to $\mathbf{c}_2 \times \mathbf{c}_1$.

The final case to consider is even more complicated and yet more involved with the cases of parallelism. After some calculation it can be seen that if $\exists \Delta_i \neq 0$ and provided that no parallelism occurs then we can pass through the singularity.

7.6 Time-stepping

In two and higher dimensions the MFE method gives rise to a system of ODE's which must be integrated to give the solution. The method of time-stepping to

be applied is dependent upon the MFE approach. If penalty functions are used then an implicit solver must be used (Miller & Miller (1981)) but otherwise then the time-stepping may be carried out using any convenient method. e.g. Euler explicit

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \Delta t \dot{\mathbf{y}}^n \tag{7.82}$$

where $\mathbf{y}^n = (...; a_j, \mathbf{s}_j; ...)$ (c.f. Johnson, Wathen & Baines (1988)).

7.7 Regularization

There have been many regularization techniques, each with different strengths and weaknesses, depending upon the problem under consideration. The aim of these methods is similar to that of penalty functions (see chapter 3), i.e. to constrain the nodes so that no element folding occurs.

The regularization is carried out by adding a variety of terms dependent on the type of regularization under consideration. These terms can include such terms as tangential velocity (see Baines (1986), Sweby (1987)) or tangential acceleration.

We will not consider this further since it is not appropriate because we wish to allow the solution to overturn (following the 1-D approach). This means that elements must cross thus eliminating the need to apply constraints to prevent this.

Besides the basic global and local approach, there are other variations of the MFE method, analogous to those described in 1-D.

7.8 Gradient Weighted MFE

The GWMFE method introduced by Miller can also be extended to higher dimensions (Carlson & Miller (1986)). We again consider the PDE, $u_t - \mathcal{L}(u) = 0$. The same approximation to u is made and U_t becomes

$$U_t = \sum_{j=1}^{N} \dot{a}_j \alpha_j + \dot{X}_j \beta_j + \dot{Y}_j \gamma_j \tag{7.83}$$

or

$$U_t = \sum_{j=1}^{N} \dot{a}_j \frac{\partial U}{\partial a_j} + \dot{X}_j \frac{\partial U}{\partial X_i} + \dot{Y}_j \frac{\partial U}{\partial Y_i} \tag{7.84}$$

in 2-D and either

$$U_t = \sum_{j=1}^{N}(\dot{a}_j(t)\alpha_j + \sum_{m=1}^{N} \dot{\mathbf{s}}_j(t)\beta_{jm}) \tag{7.85}$$

or

$$U_t = \sum_{j=1}^{N} \dot{a}_j(t)\frac{\partial U}{\partial a_j} + \dot{\mathbf{s}}_j(t).\nabla_{\mathbf{s}_j} U \tag{7.86}$$

generally.

According to Miller the problem with basic MFE is that the L_2 norm is inappropriate for problems with moving fronts in that all the nodes move to the steepest part of the curve. In 2-D it is argued that the term $\dot{u}(1 + u_x^2 + u_y^2)^{-\frac{1}{2}}$ is a bounded L_2 function of the surface area, which is independent of the steepness of the front, and hence should produce a more satisfactory nodal placement. A similar argument can be applied to higher dimensions in order to find a satisfactory weight function.

Consider therefore the minimisation of the residual

$$U_t - \mathcal{L}(U) \tag{7.87}$$

using the gradient weighted norm

$$\|\dot{U} - \mathcal{L}(U)\|_N^2 = \int [\dot{U} - \mathcal{L}(U)]_N^2 ds \tag{7.88}$$

$$= \int [\dot{U} - \mathcal{L}(U)]^2 w dx dy. \tag{7.89}$$

The weighting function $w = (1 + u_x^2 + u_y^2)^{-\frac{1}{2}}$ in 2-D helps to space the nodes out, so that they move to regions of high curvature.

This method now gives rise to the MFE equations

$$(U_t - \mathcal{L}(U), \alpha_i)_N = 0 \tag{7.90}$$

$$(U_t - \mathcal{L}(U), \beta_{im})_N = 0 \tag{7.91}$$

where $i = 1, ..., N$ and $m = 1, ..., n$. In 2-D we obtain the system

$$(U_t - \mathcal{L}(U), \alpha_i)_N = 0 \tag{7.92}$$

$$(U_t - \mathcal{L}(U), \beta_i)_N = 0 \tag{7.93}$$

$$(U_t - \mathcal{L}(U), \gamma_i)_N = 0 \tag{7.94}$$

where $(.,.)_N$ is the inner product associated with the gradient weighted norm. The equations obtained are then solved in a similar manner to those generated using the global MFE method.

7.9 Local MFE

The local MFE method in two or more dimensions uses the same ideas as the 1-D method. U_t is approximated discontinuously using the local basis functions $\phi_k^{(\nu)}$ so that

$$U_t = \sum_{k=1}^{E} \sum_{\nu=1}^{n+1} w_k^{(\nu)}(t) \phi_k^{(\nu)}(\mathbf{r}, \mathbf{s}(t)) \tag{7.95}$$

where E is the number of elements, n is the dimension, $\mathbf{r}$ is the position vector of a point and $\mathbf{s}$ denotes the position of the node. In 2-D this simplifies to

$$U_t = \sum_{k=1}^{E} (w_k^{(1)} \phi_k^{(1)} + w_k^{(2)} \phi_k^{(2)} + w_k^{(3)} \phi_k^{(3)}) \tag{7.96}$$

where $\phi_k^{(i)}$ are the local element basis functions. There now follows the use of the local basis functions in obtaining a decomposition of the global MFE matrix. We again minimise

$$\|U_t - \mathcal{L}(U)\|^2 \tag{7.97}$$

over $w_k^{(i)}$ to give

$$< \phi_k^{(i)}, U_t - \mathcal{L}(U) >= 0 \quad i = 1, .., n+1 \tag{7.98}$$

which is

$$< \phi_k^{(i)}, U_t - \mathcal{L}(U) >= 0 \quad i = 1, 2, 3 \tag{7.99}$$

in 2-D. For each element this is written as

$$C_k \mathbf{w}_k = \mathbf{b}_k \tag{7.100}$$

which is a 3×3 system in 2-D, where C_k is

$$C_k = \frac{\Delta_k}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}, \tag{7.101}$$

Δ_k is the area of element k and

$$\mathbf{b}_k = \begin{pmatrix} < \phi_k^{(1)}, \mathcal{L}(U) > \\ < \phi_k^{(2)}, \mathcal{L}(U) > \\ < \phi_k^{(3)}, \mathcal{L}(U) > \end{pmatrix}. \tag{7.102}$$

In n-D we obtain the same equations as in the section 7.4 on local basis functions (see equations (7.26), (7.27), (7.28)).

As in section 7.4 we obtain the system

$$M\dot{\mathbf{y}}_j = \mathbf{w}_j \tag{7.103}$$

for each node j. In two and higher dimensions the M matrix is rectangular so that (7.103) cannot be multiplied by its inverse. Instead we solve the system

$$M_j^T D_j M_j \dot{\mathbf{y}} = M_j^T D_j \mathbf{w}_j \tag{7.104}$$

where D_j is the diagonal of C_j. The local and global methods in two and higher dimensions are not equivalent because D is used and not C.

The GWMFE method above (see section 7.8) has an analogous local method which can be found by using the $(.,.)_N$ inner product in the local method.

7.10 Local And Global MFE Methods

In 1-D the Local and Global methods are identical but in two and higher dimensions this does not remain true. Local methods can also be used in 2-D, but here we are more interested in considering them as two stage procedures (see chapter 4 and chapter 8) in order that overturning solutions may be calculated.

There are several varieties of two stage methods which give rise to both local and global methods depending upon the type of weighting function used in the minimisation. In 1-D these methods would only give rise to one method because of the equivalence of local and global MFE.

7.11 Legendre Transformation In 2-D

Now consider the transformation (described in chapter 6 section 6.7) between the variables x, y and m, n with dual functions $u(x, y)$ and $v(m, n)$ which satisfy

$$u(x, y) - mx - ny + v(m, n) = 0 \tag{7.105}$$

with $m = \frac{\partial u}{\partial x}$, $n = \frac{\partial u}{\partial y}$, $x = \frac{\partial v}{\partial m}$ and $y = \frac{\partial v}{\partial n}$. Let us consider the equation

$$u_t + H(x, y, u, u_x, u_y) = 0 \tag{7.106}$$

and write this in terms of the dual variables as (see chapter 6 section 6.7)

$$-\dot{v} + x\dot{m} + y\dot{n} + H(x, y, u, m, n) = 0. \tag{7.107}$$

Now approximate $\hat{u}, \hat{x}, \hat{y}$ by $\hat{U}, \hat{X}, \hat{Y}$ in a finite dimensional space where $\hat{U}, \hat{X}, \hat{Y}$ are piecewise linear. We now project H into $\tilde{H}$ into the space of piecewise linear functions on element k to give

$$R = -\dot{V}_k + X\dot{M}_k + Y\dot{N}_k + \breve{H}(X, Y, V_k, M_k, N_k) \tag{7.108}$$

where $\breve{H}(X, Y, V_k, M_k, N_k) = \tilde{H}(X, Y, M_k X + N_k Y - V_k, M_k, N_k)$ is linear in X, Y.

Since $\breve{H}$ is linear in X, Y it can be written as

$$\breve{H} = \breve{H}_{k0}(V_k, M_k, N_k) + X\breve{H}_{k1}(V_k, M_K, N_k) + Y\breve{H}_{k2}(V_k, M_k, N_k). \tag{7.109}$$

Now the residual R will vanish for all X, Y if

$$\dot{M}_k = -\breve{H}_{k1}(V_k, M_k, N_k) \tag{7.110}$$

$$\dot{N}_k = -\breve{H}_{k2}(V_k, M_k, N_k) \tag{7.111}$$

$$\dot{V}_k = \breve{H}_{k0}(V_k, M_k, N_k) \tag{7.112}$$

(c.f. the equations given analytically for Burgers' equation in chapter 6 section 6.7), which in this case are given by

$$\breve{H}_{k0} = (M + N)V \tag{7.113}$$

$$\breve{H}_{k1} = (M + N)M \tag{7.114}$$

$$\breve{H}_{k2} = (M + N)N \tag{7.115}$$

which implies that

$$\dot{M}_k = -(M + N)M \tag{7.116}$$

$$\dot{N}_k = -(M + N)N \tag{7.117}$$

$$\dot{V}_k = (M + N)V \tag{7.118}$$

which are numerical versions of (6.72), (6.73) and (6.74). in chapter 6.

It should also be noted that the projection into piecewise linear space can be used in developing the following method. Since (7.109) is linear in X, Y then $\breve{H}_{k1}$

is the gradient of U in the X direction and $\breve{H}_{k2}$ is the gradient of U in the Y direction. As the area of element $k \to 0$ then

$$\dot{M} = \breve{H}_{k1} \to \dot{u}_x = -\frac{\partial H}{\partial x} \tag{7.119}$$

$$\dot{N} = \breve{H}_{k2} \to \dot{u}_y = -\frac{\partial H}{\partial y} \tag{7.120}$$

as in the theory of characteristics. Thus in 2-D the MFE method in v, m, n space is again a discretisation of the method of characteristics. The inverse transformation in the analytic case leads back to

$$\dot{X} = \frac{\partial H}{\partial m} \quad \text{and} \quad \dot{Y} = \frac{\partial H}{\partial n}. \tag{7.121}$$

In the numerical case the transformation back to U, X, Y is however a further approximation to maintain the topology of the grid.

7.12 Split Method

This method is based upon MFE but the basic procedure is divided into two sequential steps. In 2-D this allows $\dot{x}, \dot{y}$ to be solved separately from $\dot{u}$ since the equations simplify to give a decoupled system.

Consider the equation

$$u_t - \mathcal{L}(u) = 0 \tag{7.122}$$

where $u = u(\mathbf{x}, t)$, $\mathbf{x} = (x_1, ..., x_n)$ and $\mathcal{L}(u)$ is a function of x, y, u and its first derivatives. The usual MFE method leads to the global MFE equations

$$< U_t - \mathcal{L}(U), \alpha_i >= 0 \qquad i = 1, ..., N \tag{7.123}$$

$$< U_t - \mathcal{L}(U), \beta_{im} >= 0 \qquad m = 1, ..., n \tag{7.124}$$

where α_i, β_{im} are the usual basis functions n is the number of space dimensions and N is the number of nodes. Now consider (7.106) for which $\mathcal{L}(u) = -H(x, y, u, u_x, u_y)$. The split method replaces the β equations (7.124) by

$$< \dot{X}_j - \frac{\partial H}{\partial u_{X_j}}, \alpha_i >= 0 \quad i = 1, ..., N, \quad j = 1, ...n. \tag{7.125}$$

The idea is justified from the equations above (section 7.11) by extending the characteristic theory. In the limit as the number of nodes increases we have

(7.119), (7.120) which leads to

$$\dot{x_j} = \frac{\partial H}{\partial u_{x_j}} = \frac{\partial H}{\partial m_j} \tag{7.126}$$

where $m_j = u_{x_j}$. The equations are solved by minimising

$$\left\| \dot{X}_j - \frac{\partial H}{\partial M_j} \right\| \tag{7.127}$$

over $\dot{X}_j$ and similarly for $\dot{Y}_j$

$$\|\dot{U} - U_x\dot{X} - U_y\dot{Y} - \mathcal{L}(U)\| \tag{7.128}$$

over $\dot{U}$.

7.13 Lagrangian Methods

From chapter 6, the equations of interest here are the conservation laws. Analytically the method of characteristics and the Lagrangian method are the same (see chapter 6 section 6.3.4) for conservation laws, although this is not true for more general equations. The Lagrangian method is identical to the method of characteristics since if we consider the equation

$$u_t + a(u)u_x + b(u)u_y = 0 \tag{7.129}$$

then the Lagrangian method becomes

$$\dot{u} = 0, \quad \dot{x} = a(u) \quad \dot{y} = b(u). \tag{7.130}$$

This is same as the characteristic method described in chapter 6 where u_t is written as $\dot{u} - \dot{x}u_x - \dot{y}u_y$ substituted into (7.129) then the coefficients of u_x, u_y are compared to give (7.130). The difference between the two methods arises in the discretisation of the equations. This can be more clearly seen by considering the equation

$$u_t + H(x, y, u, u_x, u_y) = 0. \tag{7.131}$$

The Lagrangian method is given by

$$\dot{u} = 0 \tag{7.132}$$

and any values of $\dot{x}$ and $\dot{y}$ satisfying

$$\dot{x}u_x + \dot{y}u_y = H. \tag{7.133}$$

The method of characteristics solution is

$$\dot{u} = -H + u_x \frac{\partial H}{\partial u_x} + u_y \frac{\partial H}{\partial u_y} \tag{7.134}$$

$$\dot{x} = \frac{\partial H}{\partial u_x} \tag{7.135}$$

$$\dot{y} = \frac{\partial H}{\partial u_y}. \tag{7.136}$$

Note: $\dot{u} \neq 0$ and only if $H = a(u)u_x + b(u)u_y$ does it become so. If we now discretise these equations using finite elements or finite differences, it can be seen that different sets of equations can be found. In particular $\dot{u} \neq 0$ although it may be small. Note: The MFE methods described (above) are all slightly different, but they are all approximations to the method of characteristics or the Lagrangian methods, so $\dot{u}$ will always be small. Note: The Lagrangian method is approximate when applied numerically. This occurs because the linear elements between the nodes will not necessarily remain linear when the Lagrangian method is applied to a general conservation law.

7.14 Boundary Conditions

There are several types of conditions that may be applied to the boundaries but we do not consider them in detail here. If we consider fixed boundary conditions, then we can apply either Dirichlet, Neumann or a combination of these conditions at the boundary. If boundaries are fixed this may cause large elements near the boundary since the first non-fixed node may want to move into the region while the boundary nodes remain fixed. An alternative to this is moving boundaries. This eliminates the large elements around the boundary but the region will now not remain fixed. e.g. it will change shape and size.

7.15 Summary

In this chapter moving finite element methods have been considered in two and higher dimensions. The problems of their implementation have been discussed,

including the singularities of the matrices caused by element folding or parallelism.

There is a problem in the minimisation of the norm when the solution becomes multivalued because the norm is no longer well defined. This problem also occured in 1-D and was solved by minimising a different norm which remained valid when multivalued solutions formed. This idea is extended to two and higher dimensions in chapter 8.

Chapter 8

Overturning In Higher Dimensions

8.1 Introduction

The method of obtaining shocked solutions discussed here involves the calculation of multivalued solutions, followed by the recovery of the shock position from the multivalued curve. The calculations are made using both Lagrangian methods and MFE based methods (see chapter 7). The basic MFE method is invalid when used for overturned solutions since the L_2 norm of the residual does not remain positive definite, but may be rewritten as a two stage procedure where the norms in both stages remain positive definite (see chapter 4). These methods are valid in two and higher dimensions, however we will only describe the implementation of the various methods in 2-D. This is because as the number of dimensions increases, then so do the size and complexity of the matrices.

We follow the pattern of chapter 4, with the work split into three separate sections. In section 8.2 the theory describing the replacement of the L_2 norm by a sum of two other norms is given. In section 8.3 the numerical implementations of the new methods are described followed by section 8.4 which contains the algorithm for fitting the shock.

8.2 Summary Of Description Of Overturned Norms

Here we will give a summary of the work described in chapter 4. Chapter 4 was restricted to 1-D whereas here the work is described relating to two and higher dimensions (see also Baines & Reeves (1990)). The usual L_2 norm becomes invalid as the solution becomes multivalued. As a consequence, the L_2 minimisation used in the the MFE methods is no longer appropriate. In the 1-D theory of chapter 4, the L_2 norm is replaced by the sum of two different norms. Here we extend the argument to 2-D but first we summarize the work of chapter 4 sections 4.1-4.3.

Let us consider the equation

$$u_t - \mathcal{L}(u) = 0 \tag{8.1}$$

then the piecewise linear approximation is given by

$$U_t - \mathcal{L}(U) = R \tag{8.2}$$

where R is the residual, $U \in S$ and $U_t \in T$. S and T are generally distinct spaces of piecewise linear functions. Let S^* be a space of piecewise discontinuous functions and R^* is the L_2 projection of R into S^*. The minimisation of R can be rewritten as

$$\|R\|^2 = \|R - R^*\|^2 + \|R^*\| \tag{8.3}$$

(since $< R - R^*, R >= 0$, see chapter 4, equation (4.5)). This means that the minimisation can be split into two sequential projections. It is found that there still remains a problem when the solution overturns (because $\|R^*\|$ is not valid once the solution becomes multivalued).

Let us now rewrite $\|R^*\|$ as an l_2 norm (a discrete version of the L_2 norm as defined in chapter 4, equation (4.8)) using a coordinate system in S^*. To do this it is necessary to introduce sets of basis functions $\{\phi_i\}$ in S^*, $\{\delta_i\}$ in T where $\{w_i\}$ and $\{q_i\}$ are the corresponding sets of coefficients. This gives

$$\mathcal{L}(U)^* = \sum_i w_i \phi_i, \qquad U_t = \sum_i q_i \delta_i. \tag{8.4}$$

After some manipulation we find that

$$\|R^*\|^2 = \|\sum_i (\sum_j q_j \mu_{ji} - w_i)\phi_i\|^2 \tag{8.5}$$

where

$$\delta_i = \sum_j \mu_{ij}\phi_i. \tag{8.6}$$

Now $\|R^*\|^2$ becomes

$$\|R^*\|^2 = \sum_i \sum_k (\sum_j q_j \mu_{ji} - w_i)(\sum_l q_l \mu_{lk} - w_k) < \phi_i, \phi_k > \tag{8.7}$$

which is a new finite dimensional l_2 norm of the coordinates of R^* (moreover it is unaffected by overturning).

We may also replace $\|R^*\|^2$ by

$$\|R^*\|_d^2 = \sum_i \sum_k (\sum_j q_j \mu_{ji} - w_i)(\sum_l q_l \mu_{lk} - w_k)((\phi_i, \phi_k)) \tag{8.8}$$

where $((.,.))$ is defined by

$$((\phi_i, \phi_j)) = \begin{cases} < \phi_i, \phi_j > & i = j \\ 0 & i \neq j \end{cases}, \tag{8.9}$$

to give a new norm (Miller (1988))

$$|||R|||^2 = \|R - R^*\|^2 + \|R^*\|_d^2 \tag{8.10}$$

which corresponds to a 2-stage or a local method. These norms decouples the equations so that they are separable element by element. This norm remains valid when the solution overturns and allow us to write the MFE methods in two stages.

8.3 Implementation Of The Methods In 2-D

We shall now consider the implementation of a variety of MFE methods (in 2-D). The descriptions which follow are applicable to both overturning and non-overturning solutions but are only necessary if overturned solutions are to be calculated.

8.3.1 ϕ Basis Functions (2 Stage Method)

This is a two stage method using ϕ basis functions which can give either a global or local method. The first stage is common to both methods and the second stage differs only in the choice of weighting function.

Stage 1

Consider the equation

$$u_t - \mathcal{L}(u) = 0 \tag{8.11}$$

where $u = u(x, y, t)$ and $\mathcal{L}$ contains derivatives u_x, u_y only on a region R, with Dirichlet boundary conditions. Using ϕ basis functions, u is approximated (discontinuously) by $\tilde{U}_k$ in element k where

$$\begin{aligned}
\tilde{U}_k &= a_k^{(1)}\phi_k^{(1)} + a_k^{(2)}\phi_k^{(2)} + a_k^{(3)}\phi_k^{(3)} && (8.12)\\
\tilde{X}_k &= X_k^{(1)}\phi_k^{(1)} + X_k^{(2)}\phi_k^{(2)} + X_k^{(3)}\phi_k^{(3)} && (8.13)\\
\tilde{Y}_k &= Y_k^{(1)}\phi_k^{(1)} + Y_k^{(2)}\phi_k^{(2)} + Y_k^{(3)}\phi_k^{(3)} && (8.14)
\end{aligned}$$

where $a_k^{(i)}, X_k^{(i)}, Y_k^{(i)}$ are the values of a, X, Y at a node of element k corresponding to the local node numbering of the basis functions. Now $\frac{\partial \tilde{U}_k}{\partial t}$ is given (also discontinuously) by

$$\begin{aligned}
\frac{\partial \tilde{U}_k}{\partial t} &= \dot{a}_k^{(1)}\phi_k^{(1)} + \dot{a}_k^{(2)}\phi_k^{(2)} + \dot{a}_k^{(3)}\phi_k^{(3)}\\
&\quad -m_k(\dot{X}_k^{(1)}\phi_k^{(1)} + \dot{X}_k^{(2)}\phi_k^{(2)} + \dot{X}_k^{(3)}\phi_k^{(3)})\\
&\quad -n_k(\dot{Y}_k^{(1)}\phi_k^{(1)} + \dot{Y}_k^{(2)}\phi_k^{(2)} + \dot{Y}_k^{(3)}\phi_k^{(3)}) \qquad (8.15)
\end{aligned}$$

where m_k and n_k are the U_x and U_y gradients in element k. This can also be written as

$$\begin{aligned}
w_k^{(1)} &= \dot{a}_k^{(1)} - m_k\dot{X}_k^{(1)} - n_k\dot{Y}_k^{(1)} && (8.16)\\
w_k^{(2)} &= \dot{a}_k^{(2)} - m_k\dot{X}_k^{(2)} - n_k\dot{Y}_k^{(2)} && (8.17)\\
w_k^{(3)} &= \dot{a}_k^{(3)} - m_k\dot{X}_k^{(3)} - n_k\dot{Y}_k^{(3)} && (8.18)
\end{aligned}$$

so that

$$\frac{\partial \tilde{U}_k}{\partial t} = w_k^{(1)}\phi_k^{(1)} + w_k^{(2)}\phi_k^{(2)} + w_k^{(3)}\phi_k^{(3)}. \tag{8.19}$$

We first minimise

$$\left\| \frac{\partial U_k}{\partial t} - \mathcal{L}(U_k) \right\|^2 \tag{8.20}$$

over each element k with respect to $w_k^{(i)}$ $(i = 1, 2, 3)$ and E is the number of elements. This gives the system

$$< U_t - \mathcal{L}(U_k), \phi_k^{(1)} >= 0 \tag{8.21}$$

$$< U_t - \mathcal{L}(U_k), \phi_k^{(2)} >= 0 \tag{8.22}$$

$$< U_t - \mathcal{L}(U_k), \phi_k^{(3)} >= 0 \tag{8.23}$$

which can be written as

$$C_k \mathbf{w}_k = \mathbf{b}_k \tag{8.24}$$

for each element k, where

$$C_k = \frac{\Delta_k}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}, \qquad \mathbf{b}_k = \begin{pmatrix} < \phi_k^{(1)}, \mathcal{L}(U) > \\ < \phi_k^{(2)}, \mathcal{L}(U) > \\ < \phi_k^{(3)}, \mathcal{L}(U) > \end{pmatrix} \tag{8.25}$$

and Δ_k = area of element k,

$$\mathbf{w}_k = \begin{pmatrix} w_k^{(1)} \\ w_k^{(2)} \\ w_k^{(3)} \end{pmatrix}. \tag{8.26}$$

Taken over all the elements, the C matrix becomes block diagonal, with block entries C_k. The w's may now be calculated over the whole system.

This can be rewritten as the minimisation of (8.20) with respect $a_k^{(i)}, X_k^{(i)}, Y_k^{(i)}$ $(i = 1, 2, 3)$. This leads to a square system of 9 equations in 9 unknowns for each element k, but the system is singular. The situation is more complicated than in 1-D where $\dot{a}, \dot{s}$ were defined on each side of the node, because here $\dot{a}, \dot{X}, \dot{Y}$ are defined at each corner of each element which meets the node. For continuity all values of $\dot{a}, \dot{X}, \dot{Y}$ surrounding the node j need to be equal. This may be obtained by applying a set of constraints to these values, which returns the system to the smaller node-based numbering and the familiar non-singular systems are obtained.

The singular system can be written as

$$E_k \dot{\mathbf{y}}_k = \mathbf{G}_k \tag{8.27}$$

where $\dot{\mathbf{y}}_k = (\dot{a}_k^{(1)}, \dot{X}_k^{(1)}, \dot{Y}_k^{(1)}; \dot{a}_k^{(2)}, \dot{X}_k^{(2)}, \dot{Y}_k^{(2)}; \dot{a}_k^{(3)}, \dot{X}_k^{(3)}, \dot{Y}_k^{(3)})$

$$E_k = \frac{\Delta_k}{12} \begin{pmatrix} 2\mathbf{m}_k\mathbf{m}_k^T & \mathbf{m}_k\mathbf{m}_k^T & \mathbf{m}_k\mathbf{m}_k^T \\ \mathbf{m}_k\mathbf{m}_k^T & 2\mathbf{m}_k\mathbf{m}_k^T & \mathbf{m}_k\mathbf{m}_k^T \\ \mathbf{m}_k\mathbf{m}_k^T & \mathbf{m}_k\mathbf{m}_k^T & 2\mathbf{m}_k\mathbf{m}_k^T \end{pmatrix} \tag{8.28}$$

$$\mathbf{G}_k = \begin{pmatrix} < \phi_k^{(1)}, \mathcal{L}(U) > \\ < \phi_k^{(2)}, \mathcal{L}(U) > \\ < \phi_k^{(3)}, \mathcal{L}(U) > \\ < -m_k\phi_k^{(1)}, \mathcal{L}(U) > \\ < -m_k\phi_k^{(2)}, \mathcal{L}(U) > \\ < -m_k\phi_k^{(3)}, \mathcal{L}(U) > \\ < -n_k\phi_k^{(1)}, \mathcal{L}(U) > \\ < -n_k\phi_k^{(2)}, \mathcal{L}(U) > \\ < -n_k\phi_k^{(3)}, \mathcal{L}(U) > \end{pmatrix} \tag{8.29}$$

and $\mathbf{m}_k = (1 \quad -m_k \quad -n_k)^T$. It should be noted that E_k is a 9×9 matrix.

Stage 2

For both the global and local method, we use the minimisation of (8.20) which leads to either $C\mathbf{w} = \mathbf{b}$ or $E\dot{\mathbf{y}} = \mathbf{G}$. In the Miller and Carlson approach we can define E_{dk} to be

$$E_{dk} = \frac{\Delta_k}{12} \begin{pmatrix} 2\mathbf{m}\mathbf{m}^T & 0 & 0 \\ 0 & 2\mathbf{m}\mathbf{m}^T & 0 \\ 0 & 0 & 2\mathbf{m}\mathbf{m}^T \end{pmatrix} = \tilde{M}_k^T D_k \tilde{M}_k \tag{8.30}$$

where $D_k = \text{Diag}\{C_k\}$ so that $E_k\dot{\mathbf{y}}_k = \mathbf{G}_k$ may be written as

$$E_{dk}\dot{\mathbf{y}}_k = E_{dk}E_k^{-1}\mathbf{G}_k. \tag{8.31}$$

We now apply the constraints as before. To apply constraints to the system $E_k\mathbf{y}_k = \mathbf{G}_k$, let us consider node j, where j is surrounded by p elements. Let $\dot{\mathbf{Y}}_j$

be given by the values of $(\dot{a}_\nu^{(l)}, \dot{X}_\nu^{(l)}, \dot{Y}_\nu^{(l)})$ ν, being the numbers of the p elements surrounding node j, and where l is the local element numbering $(l = 1, 2, 3)$ referring to node j.

$$\dot{\mathbf{Y}}_j = \begin{pmatrix} \vdots \\ \dot{a}_j^\nu \\ \dot{x}_j^\nu \\ \dot{y}_j^\nu \\ \vdots \end{pmatrix}_{(\forall \nu)} = R_j \begin{pmatrix} \dot{a}_j \\ \dot{X}_j \\ \dot{Y}_j \end{pmatrix} \tag{8.32}$$

where

$$R_j = \begin{pmatrix} I_3 \\ I_3 \\ \vdots \\ I_3 \end{pmatrix}, \qquad I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{8.33}$$

Then over the whole system this gives

$$\dot{\mathbf{Y}} = R\dot{\mathbf{y}} \tag{8.34}$$

where $R = \text{diag}\{R_j\}$ where R is a non-square matrix and

$$\dot{\mathbf{Y}} = \begin{pmatrix} \vdots \\ \dot{\mathbf{Y}}_j \\ \vdots \end{pmatrix}. \tag{8.35}$$

Thus to minimise this over $\dot{\mathbf{y}}$ gives

$$\min_{\dot{\mathbf{y}}} \|(R\dot{\mathbf{y}} - \dot{\mathbf{Y}})W\| \tag{8.36}$$

where W is a weight function. For the global method $W = E^{\frac{1}{2}}$ and for the local $W = E_d^{\frac{1}{2}}$ where $E_d = \text{Diag}\{E\}$. This can be shown to be equivalent to the usual MFE system.

8.3.2 Introduction To $|||.|||$ And $\tilde{\phi}$ Basis Functions

Using the $|||.|||$ norm and the usual element basis functions ϕ (see chapter 7) (Baines & Wathen (1988)), (Miller (1988)) showed that there exists an equivalence between the local MFE method and a new set of basis functions $\tilde{\phi}$ with the usual norm.

Consider the elementwise basis functions ϕ where the new basis functions $\tilde{\phi}$ are defined by

$$\tilde{\phi}_k^{(1)} = p_k^{(1)}\phi_k^{(1)} + p_k^{(2)}\phi_k^{(2)} + p_k^{(3)}\phi_k^{(3)} \tag{8.37}$$

where

$$< \phi_k^{(1)}, \tilde{\phi}_k^{(1)} > \;=\; < \phi_k^{(1)}, \phi_k^{(1)} > \tag{8.38}$$

$$< \phi_k^{(2)}, \tilde{\phi}_k^{(1)} > \;=\; 0 \tag{8.39}$$

$$< \phi_k^{(3)}, \tilde{\phi}_k^{(1)} > \;=\; 0. \tag{8.40}$$

Since $\tilde{\phi}$ is linear it may be defined as

$$\tilde{\phi}_k^{(1)} = \sum_{i=1}^{N} p_k^{(i)}\phi_k^{(i)}, \tag{8.41}$$

$$< \phi_k^{(1)}, \tilde{\phi}_k^{(1)} > = < \phi_k^{(1)}, \phi_k^{(1)} > \tag{8.42}$$

and

$$< \phi_k^{(1)}, \tilde{\phi}_k^{(i)} > = 0 \quad i = 2, ..., N. \tag{8.43}$$

This gives

$$p_k^{(1)} = 1 + \frac{n}{n+2}, \quad p_k^{(2)} = ... = p_k^{(n)} = \frac{-2}{n+1} \quad p_k^{(n+1)} = \frac{-2}{n+2} \tag{8.44}$$

where n is the space dimension, so in 2-D the basis functions for element k are

$$\tilde{\phi}_k^{(1)} \;=\; \frac{3}{2}\phi_k^{(1)} - \frac{1}{2}\phi_k^{(2)} - \frac{1}{2}\phi_k^{(3)} \tag{8.45}$$

$$\tilde{\phi}_k^{(2)} \;=\; -\frac{1}{2}\phi_k^{(1)} + \frac{3}{2}\phi_k^{(2)} - \frac{1}{2}\phi_k^{(3)} \tag{8.46}$$

$$\tilde{\phi}_k^{(3)} \;=\; -\frac{1}{2}\phi_k^{(1)} - \frac{1}{2}\phi_k^{(2)} + \frac{3}{2}\phi_k^{(3)} \tag{8.47}$$

and the $\tilde{\alpha}$ basis functions are defined as

$$\tilde{\alpha}_k = \tilde{\phi}_k^{(1)} + \tilde{\phi}_k^{(2)} + \tilde{\phi}_k^{(3)} \tag{8.48}$$

with $\tilde{\beta}_k = -m_k\tilde{\alpha}_k$, $\tilde{\gamma}_k = -n_k\tilde{\alpha}_k$.

The methods described below are all variations on the standard MFE methods described in chapter 7, however either the basis functions or the norms are replaced by by those described in this chapter.

8.3.3 $\tilde{\phi}$ Basis Functions (2 Stage Method)

This is the same method as chapter 7 section 7.4 where we have ϕ basis functions. Again either a local or global method can be found, dependent upon the weighting function chosen.

8.3.4 $\tilde{\alpha}$ Basis Function (1 Stage Method)

Note: The use of the $\tilde{\alpha}$ basis functions and the global method gives a one stage method. i.e. we use the global method described in chapter 7 section 7.3 but replace the α basis functions with $\tilde{\alpha}$ basis functions.

8.3.5 ϕ Or α Basis Functions And $|||.|||$ Norm

There are two methods which arise from the use of ϕ's with $|||.|||$ and two forms with the α and $|||.|||$ with a global or local method being determined from the weighting function.

8.4 Calculation Of Shock Position From Overturned Curve

If we recall the methods of shock recovery in 1-D, there were two separate approaches. One method was based on the TC operator of Brenier and the second was based on conservation of area. We will now consider how these methods may be extended to 2-D.

Let us first examine the TC operator of Brenier. If we consider an overturned region, see Fig. 8.1, where A is higher than B, it can be seen that for a smooth solution the heights should reduce from A to B. If we applied the TC operator, the connections of the nodes would prevent this occurring without the application of either a regridding or reconnection of the nodes. Regridding or reconnection is expensive to carry out, especially since this method gives its best solution when applied after every time-step (see chapter 5 section 5.4.3). This means that regridding would have to be carried out after every time-step. For these reasons

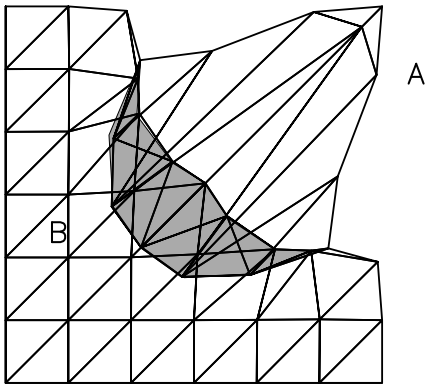

Figure 8.1: Overturned region in 2-D.

we do not pursue this approach further.

Let us now consider the equal area method and its extensions to 2-D. This method now relies on equal volumes, conditions for which can be found. Unfortunately, there is now not enough information known to find a unique position of the shock, since the extra degree of freedom is not controlled. This means that this method is not satisfactory by itself although the 1-D version of the method works well.

We nevertheless consider how we can apply the 1-D equal area method to our 2-D problem. One approach to a shock recovery scheme is to take a 1-D 'slice' through the region (in the direction of the wave speed) and apply a 1-D recovery method to this slice. The reason that the shock normal is used is that we are using a locally 1-D approach normal to the shock. This is consistent with tangential continuity. The algorithm below describes in more detail how this method may be used in the calculation.

8.4.1 Algorithm

1) Find the region where the elements are overturned (i.e. the elements have negative area). Note: the shock position will occur within this region R. See Fig. 8.2.

2) Find the edge of the region R comprised of the sides of some of the triangles in the region. These edges are now ordered in such a way as to give a continuous connected line Γ which defines the edge of the region R. See Fig. 8.3.

3) Calculate a 'slice' through the region from each edge node on Γ in the direction

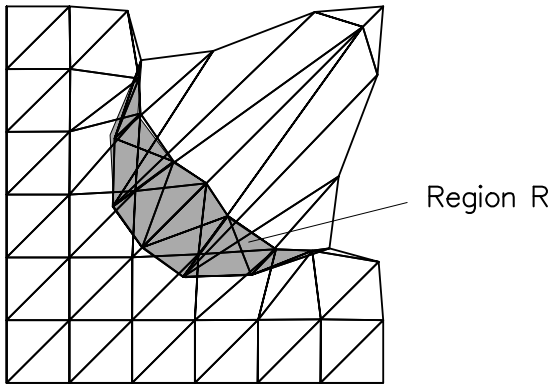

Figure 8.2: Overturned region in 2-D.

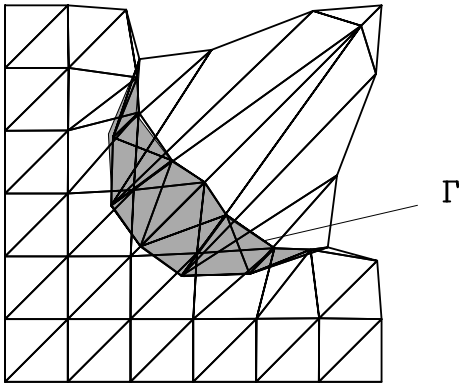

Figure 8.3: Shock region in 2-D.

$(\frac{f}{|f|}, \frac{g}{|g|})$. At each entry to an element whilst crossing the region, the height of the elements is found. Note: There will usually be more that one layer of elements so several heights need to be found. This information is used to construct a piecewise linear curve (c.f. 1-D). See Fig. 8.4.

4) To each 'slice' a 1-D shock recovery method is applied. The shock position is then calculated from this.

5) Join up the 1-D shock positions in order to obtain a 2-D shock curve S. Note that it is possible to calculate a new direction for each slice so that it will be normal to the shock. This can be done by iterating around steps 3, 4 and 5 but replacing the direction of the slice in step 3 by the new direction calculated in step 5.

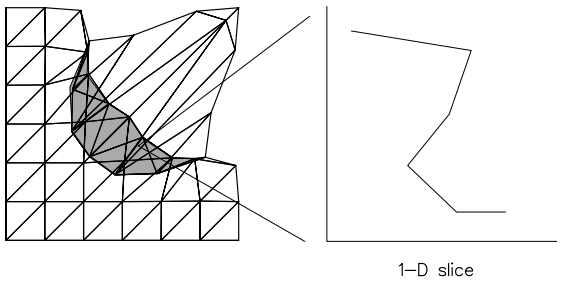

Figure 8.4: 'Slice' through overturned region.

8.5 Summary

In this chapter the modifications needed to the MFE methods in 2-D (see chapter 7) in order that overturned solutions can be calculated are described. The numerical implementation of these methods is discussed with special reference to the types of norms and basis functions which can be used. Finally, a description of an algorithm which may be used to recover the shock position in 2-D is given. In the following chapter, numerical examples are given to demonstrate the methods described here and in chapter 7.

Chapter 9

Numerical Results And Examples In 2-Dimensions

9.1 Introduction

In this chapter we consider equations of the form

$$u_t + f_x + g_y = 0 \tag{9.1}$$

where f, g are functions of $u = u(x, y)$ and initial data is given on the region Ω. This type of equation forms shocks and/or expansions. The nonlinearity of these equations leads to the formation of multivalued solutions. The type of analytic solution and the behaviour of these equations was discussed more fully in chapter 6.

The methods of solution applied to this equation are the Lagrangian method (see chapter 7) and the 2-stage local MFE method (see chapters 7 and 8). These methods both allow the formation of multivalued solutions which arise from following the characteristics (see chapter 7). These multivalued curves can then be used to obtain the shock position by applying a recovery technique. The shock position is recovered from the overturned solution using the '1-D' slicing technique described in chapter 8 section 8.4.

9.2 Description Of Test Problems

We will consider a variety of problems in order to illustrate how the solution technique copes with different cases. The problems to be considered include the formation of both shocks and expansions, thus illustrating the capabilities of the method. Examples of Riemann problems are given, including problems where $f \neq g$. Figures of the initial data are also given with the test problems in order that they may be compared with the final multivalued solution obtained. Let us first describe the test problems.

9.2.1 Problem 1

The equation we are solving here is similar to inviscid Burgers' equation in two dimensions, which is the simplest equation that gives rise to the formation of shocks and expansions. We will change the sign in the inviscid Burgers' equation since this will move the data towards the origin and make the results in the figures below easier to see. The equation we consider can be written in conservation form as

$$u_t - \left(\frac{u^2}{2}\right)_x - \left(\frac{u^2}{2}\right)_y = 0 \tag{9.2}$$

or as

$$u_t - uu_x - uu_y = 0 \tag{9.3}$$

on the region $[0,1] \times [0,1]$. The initial data is given by the equation

$$u = \frac{(\tanh(9x + 9y - 9) + 1)}{9} \tag{9.4}$$

and is shown in Fig. 9.1. The boundary conditions used are Lagrangian (see section 9.4), so that the region moves as the solution evolves. The initial data moves under the influence of the equation to form a straight line shock. This problem is used to demonstrate that the method can cope with the simple case where the shock position can be easily verified. The second problem we will consider involves the formation of a curved shock.

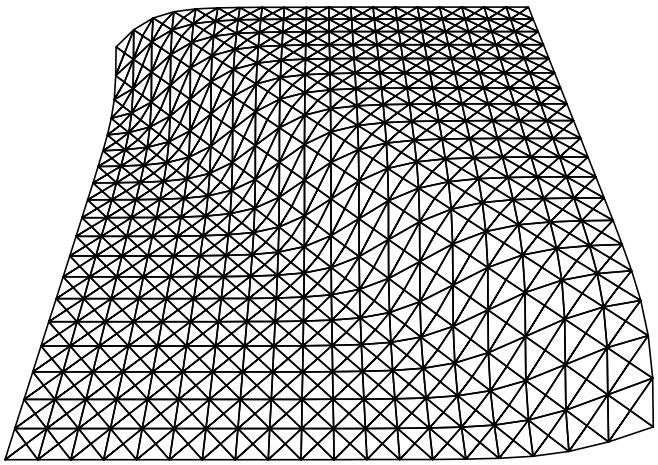

Figure 9.1: Problem 1 - Initial data given by a tanh curve.

9.2.2 Problem 2

We will again use the 'negative' inviscid Burgers' equation but now the initial data is given so that as the solution evolves a curved shock will be formed. The initial data is given in terms of r where

$$r = \sqrt{\left(\left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2\right)} \tag{9.5}$$

and is given by

$$u = \begin{cases} 0 & r > 0.5 \\ \cos^2(\pi r) & \text{otherwise} \end{cases} \tag{9.6}$$

on the region $[0,1] \times [0,1]$. The initial data is shown in Fig. 9.2 below. The

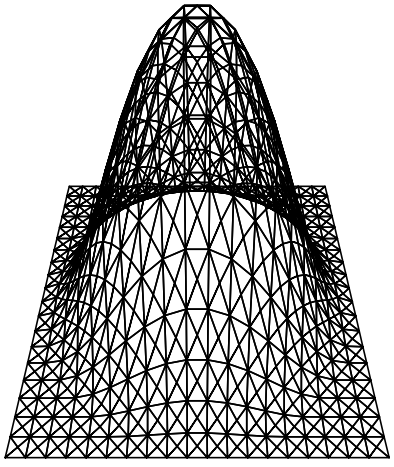

Figure 9.2: Problem 2 - Initial data given by a plane and a cosine curve.

boundary conditions applied are Lagrangian (see section 9.4), but in this example there will initially be little interaction between the cosine curve and the boundary.

9.2.3 Problem 3

In this example we will examine the formation of an expansion. We use the equation

$$u_t + 5\left(\frac{u^3}{3}\right)_x + 5\left(\frac{u^3}{3}\right)_y = 0 \tag{9.7}$$

with initial data given by

$$u = \frac{(\tanh(9x^2 + 9y^2 + 9) - 0.5)}{9} \tag{9.8}$$

on the region $[0,1] \times [0,1]$. The initial data can be seen in Fig. 9.3 below. In this

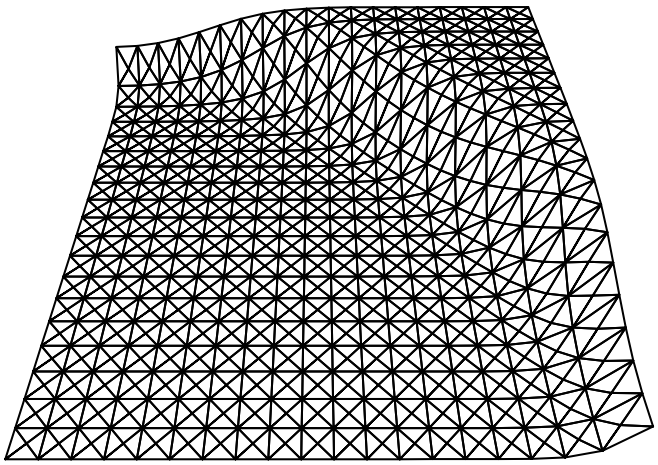

Figure 9.3: Problem 3 - Initial data given by a tanh curve.

problem the slope becomes less steep as time evolves and the solution expands. The boundary conditions given are Lagrangian (see section 9.4).

9.2.4 Problem 4

Let us now consider the 'negative Buckley-Leverett' equation in 2-D which is given by

$$u_t - \left(\frac{u^2}{u^2 + \frac{1}{2}(1-u)^2}\right)_x - \left(\frac{u^2}{u^2 + \frac{1}{2}(1-u)^2}\right)_y = 0. \tag{9.9}$$

The initial data is given in terms of r where

$$r = \sqrt{((x-1)^2 + (y-1)^2)} \tag{9.10}$$

and u is given by

$$u = \begin{cases} 0 & r > 0.5 \\ \cos^2(\pi r) & \text{otherwise} \end{cases} \tag{9.11}$$

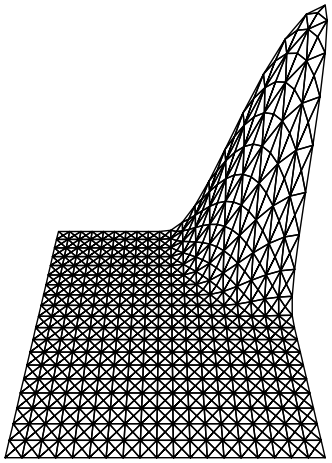

Figure 9.4: Problem 4 - Initial data given by a plane and a cosine curve.

on the region $[0,1] \times [0,1]$. The initial data is shown in Fig. 9.4 above. This is very similar to problem 3 in 1-D (see chapter 5 section 5.3.1). A curved shock is formed as time evolves, however it is a combination of a shock and an expansion since f and g are not convex (or concave). See Concus & Proskurowski (1979).

9.2.5 Problem 5

In this problem we show the effect of a non-symmetric equation on symmetric data. The equation to be solved is non-symmetric because $f \neq g$ and is given by

$$u_t - 5\left(\frac{u^2}{2}\right)_x - 5\left(\frac{u^3}{3}\right)_y = 0 \tag{9.12}$$

and the initial data is given by

$$u = \frac{(\tanh(9x^2 + 9y^2 + 9) - 0.5)}{9} \tag{9.13}$$

on the region $[0,1] \times [0,1]$ (see Fig. 9.3). The initial data is smooth and curved so that a curved shock will be formed. This problem can be used to demonstrate that the method can be applied where $f \neq g$. The boundary conditions are Lagrangian; for more information see section 9.4.

9.2.6 Problem 6

We now consider a series of Riemann problems. The equation used in these examples is again the 'negative' inviscid Burgers' equation. The first set of initial

data is given by

$$u = \begin{cases} 1 & x > 0.5 \quad y > 0.5 \\ 0.5 & x < 0.5 \quad y > 0.5 \\ -1 & x < 0.5 \quad y < 0.5 \\ 0 & x > 0.5 \quad y < 0.5 \end{cases} \tag{9.14}$$

and is given on the region $[0,1] \times [0,1]$. Note: a similar problem is solved in chapter 6, section 6.6.3. The solution involves the formation of four shocks and their interaction. The initial data is shown in Fig. 9.5 below. The boundary

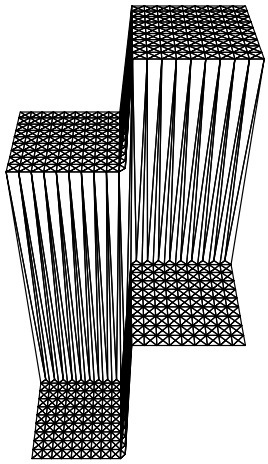

Figure 9.5: Problem 6 - Initial data given for a Riemann problem.

conditions applied are Lagrangian (see section 9.4).

9.2.7 Problem 7

The second example of initial data gives rise to four expansions. See Wagner (1983). The equation we are considering is the 'positive' Burgers' equation

$$u_t + uu_x + uu_y = 0 \tag{9.15}$$

and the initial data is given by

$$u = \begin{cases} 1 & x > 0.5 \quad y > 0.5 \\ 0.5 & x < 0.5 \quad y > 0.5 \\ -1 & x < 0.5 \quad y < 0.5 \\ 0 & x > 0.5 \quad y < 0.5 \end{cases} \tag{9.16}$$

on the region $[0,1] \times [0,1]$. See Fig. 9.5. The boundary conditions are Lagrangian and move with the expansion (see section 9.4).

9.2.8 Problem 8

We will again use the 'negative' inviscid Burgers' equation but now the initial data is in a form which combines a Riemann Problem with a solution which evolves in such a way that a curved shock will be formed. The initial data is given in terms of r where

$$r = \sqrt{\left(\left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2\right)} \tag{9.17}$$

and is given by

$$u = \begin{cases} 0 & r > 0.25 \\ \cos^2(r) & \text{otherwise} \end{cases} \tag{9.18}$$

on the region $[0,1] \times [0,1]$. The initial data is shown in Fig. 9.6 below. The

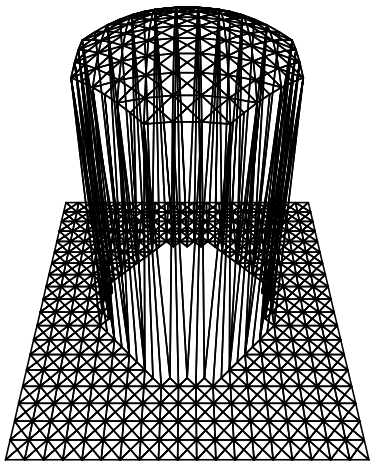

Figure 9.6: Problem 7 - Initial data given by a plane and a cosine curve.

boundary conditions applied are Lagrangian (see section 9.4), but in this example there will again initially be little interaction between the cosine curve and the boundary.

9.3 Initial Data Representation

We use three basic grids upon which to represent the initial data. See Fig. 9.7. These are very basic and do not really demonstrate MFE to its full potential but we use them to allow the full movement of the nodes by the method to be clearly seen.

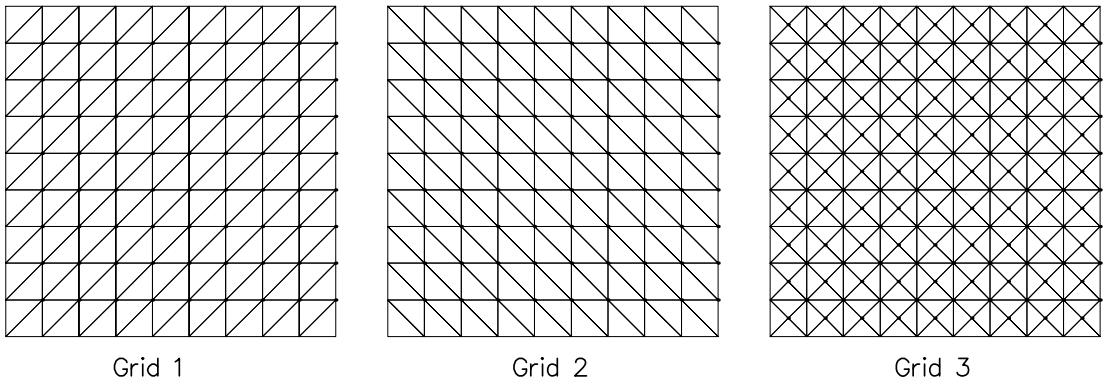

Figure 9.7: Initial grids.

The grids do not necessarily give a good representation of the initial data. A better (lower L_2 error) representation of the initial data can be found by using non-regular grids (see Malcolm (1991) and Baines (1990)).

9.4 Boundary Conditions

Boundary conditions are a separate issue so where possible we try to give examples where this will not play a major part in affecting the results. The Lagrangian boundary conditions applied in the above problems are obtained by letting $\dot{X} = f'(U)$, $\dot{Y} = g'(U)$ and $\dot{U} = 0$. This allows the region to move without distortion from the finite element technique and hence reduces the affect of the boundaries on the method.

9.5 Numerical Results For MFE 2-stage Method

Here the results are given for the problems described above. The method applied to the above problems is a 2-stage local MFE method. The shock position is obtained by allowing the solution to become multivalued by the applying the recovery technique described in chapter 8 section 8.4.1. The results are given for

varying numbers of nodes, different time-steps and different initial grids in order to demonstrate the difficulties of the problems.

9.5.1 Problem 1

This problem forms a straight line shock. The results shown below have been calculated using 100 nodes, a time-step of 0.1 and are given after 10 time-steps. The results show that a straight line shock can be formed and remain straight as it traverses the region (see Fig. 9.8). The shock position is marked in red on the grid. The picture shows the overturned curve at the same time. This can be compared with the initial data (see Fig. 9.1).

We can also use this example to show how the three trial grids appear after several time-steps. The results shown in Fig. 9.9 show that ignoring the boundaries the results are very similar for the three grids. The results are given using 100 nodes for the first two grids and 181 nodes for the third. The time-step is 0.01 and the pictures are given after 60 time-steps. As it can be see from these pictures, different triangles will overturn at different times, consequently the shock position formed will differ. i.e. the method is grid dependent.

Figure 9.8: Problem 1 - Solved using MFE method.

Figure 9.9: Problem 1 - Comparison of three grids for the initial data.

9.5.2 Problem 2

The initial data is given on grid 1 using 100 nodes and the time-step used is 0.001. The results are given after 40 time-steps. The results are shown in Fig. 9.10. below.

9.5.3 Problem 3

In this example we see the formation of an expansion to show that the MFE method can solve this type of problem. The results are shown in Fig. 9.11. The problem was solved using a time-step of 0.001 and the results are given after 460 time-steps and the initial data is given on grid 1. The resulting grid shows how the triangles are stretched and the how the boundary moves as the expansion forms.

Figure 9.10: Problem 2 - Solved using MFE method.

Figure 9.11: Problem 3 - Solved using MFE method.

9.5.4 Problem 4

The results are given using 100 nodes, a time-step of 0.001 and are given after 180 time-steps. This equation requires the small time-step to prevent oscillations forming owing to stability restrictions. The region can be seen to move as the solution forms. See Fig. 9.12. After a greater number of time-steps, the shock position becomes more jagged. This possibly occurs because the overturned region is large and the jump direction is not clear.

9.5.5 Problem 5

In this problem we applied a non-symmetric equation to symmetric data. From the results shown in Fig. 9.13 it can be seen that the MFE method can move the nodes successfully when $f \neq g$. It can also be seen that the shock has been formed unsymmetrically as expected for this type of equation. It can also be seen that spurious shocks can be formed when the triangles overturn due to poor movement of the nodes near the boundary. The results are calculated with 181 nodes on grid 3 and a time-step of 0.001. The results are given at after 220 time-steps.

Figure 9.12: Problem 4 - Solved using MFE method.

Figure 9.13: Problem 5 - Solved using MFE method.

9.5.6 Problem 6

This initial data is only approximate Riemann data because the planes are joined together by steep linear slopes and not by vertical planes. This means that it takes several time-steps before the four initial shocks are formed. The top diagram is shown after 80 time-steps and the lower diagram is shown after 120 time-steps. In Fig 9.14a, given immediately after all shocks have formed, it can be seen that the centre node where the four heights meet has moved to produce a spike. In Fig 9.14b the results show that the shocks move correctly where there is no interaction between them, however in the centre where they should interact the method cannot cope with this situation. It can also be seen that the triangles do not overturn so as to give a single line shock. For a large time-step or few nodes the MFE method fails at the centre where the 4 different heights meet.

9.5.7 Problem 7

This problem uses initial Riemann data and applies the inviscid Burgers' equation to this data. This causes each of the initial discontinuities to form expansions. The time-step used is 0.001 and the results are given after 220 time-steps using grid 3 with 181 nodes to represent the initial data. See Fig. 9.15.

Figure 9.14: Problem 6 - Solved using MFE method.

Figure 9.15: Problem 7 - Solved using MFE method.

9.5.8 Problem 8

The results are shown using grid 3 with 181 nodes and are calculated using a time-step of 0.001, and are given after 100 time-steps. The results are shown in Fig 9.16. It can be seen from the results how the initial grid moves to the shape of the data. This problem is difficult because of the steep slope joining the plane and the cosine curve. The shock position is marked in red and can be seen to be curved. This problem is very similar to the Riemann problems given earlier. The multivalued curve can be seen to be very similar to the initial data (see Fig. 9.6), however it is 'tilted' so as to form an expansion at about $(0.75, 0.75)$ and a shock at about $(0.25, 0.25)$.

9.6 Numerical Results Using Lagrangian Method

We also consider problems 1, 4 and 6 being solved using a Lagrangian method. The initial data and functions remain the same as before. The method works in a similar way to the examples above but the nodes now move differently according to the equations $\dot{U} = 0$, $\dot{X} = f^{'}(U)$ and $\dot{Y} = g^{'}(U)$. See chapter 7 section 7.13.

9.6.1 Problem 1

The results are given using grid 3 to represent the initial data and the calculations are carried out with a time-step 0.001 to give the results after 60 time-steps. For this example the nodes move in a very similar way to the 2-stage MFE method. The results obtained are very similar and both methods work equally well. See Fig. 9.17.

Figure 9.16: Problem 8 - Solved using MFE method.

Figure 9.17: Problem 1 - Solved using Lagrangian method.

9.6.2 Problem 4

Problem 4 uses grid 3 for the representation of the initial data. The time-step is 0.001 and the results are given after 8 time-steps. For this example neither grid 1 nor 2 was satisfactory and the MFE method works better that the Lagrangian method for this example. The overturned curve appears similar to that obtained using the MFE method, even though the shock position calculations differ.

9.6.3 Problem 6

The solution to problem 3 is calculated using 761 nodes and using grid 3 to represent the initial data. The time-step is 0.001 and the results are given after 5 time-steps. For this example the Lagrangian method works better that the MFE method. The nodes move to better positions and consequently the shock position is better defined than when using the MFE method.

Figure 9.18: Problem 4 - Solved using Lagrangian method.

Figure 9.19: Problem 6 - Solved using Lagrangian method.

9.6.4 Comparison Between MFE And Lagrangian Methods

The results shown in the diagram overleaf are given at the time of initial formation of the shock. The results are given using 100 nodes and a time-step of 0.001. For the first example, problem 4 the MFE method of solution in the topleft is given after 30 time-steps and the Lagrangian method is given after 32 time-steps. For problem 2 the results are given for the MFE method after 68 time-steps and for the Lagrangian method after 64 time-steps.

Comparing the MFE and the Lagrangian method, it can be seen that the shocks initially form at approximately the same time and position. This allows us to verify that the program was working since it is known analytically that the movement of nodes is along the approximate characteristics. See Fig. 9.20.

Figure 9.20: Comparison between MFE and Lagrangian methods.

9.7 Summary

It has been seen that for a variety of nonlinear problems an approximate solution can be found using moving grid methods.

The results are grid dependent since if the triangles are orientated differently they will then overturn at different times. This means that the initial representation is important.

There are problems with intersecting shocks as the shock constructions program has not been designed to cope with this type of problem. This is particularly apparent in problem 6, the Riemann problem, where the four shocks try to interact.

Chapter 10

Conclusions And Further Work

10.1 Conclusion

In this thesis we have examined the numerical solution of nonlinear scalar PDE's, particularly conservation laws, in one and higher dimensions by moving element methods, with emphasis on the formation of shocks and expansions. The methods considered are finite element in nature, including the classical MFE method, its derivatives and Lagrangian methods. The MFE methods have been rewritten in a form which allows the calculation of multivalued solutions. These multivalued solutions have been calculated using moving finite element techniques which as has been demonstrated, approximately follow the characteristics of the equation. Special attention has been paid to the recovery of shock positions from a multivalued solution. Numerical examples in both one and two dimensions have been given to demonstrate these solution techniques.

In chapter 2, we introduced some of the analytical techniques available to describe the solution and properties of the conservation laws. This included the ideas of weak solutions in order to allow discontinuities and the need for entropy conditions to impose uniqueness. The method of characteristics and consequently the notion of overturning solutions were also introduced. Following the ideas of multivalued solutions, methods of finding the shock position based on 'conservation' were described.

Chapter 3 introduced the basic ideas of MFE and Lagrangian methods with particular reference to the solution of the conservation laws. These methods are

adaptive and were chosen because they permitted the formation of overturned curves which is required to follow the ideas of chapter 2.

In chapter 4, the consequences of applying MFE methods to the formation of multivalued curves were discussed. In this situation it was found that the MFE method became invalid, and as a consequence it was rewritten in terms of a two stage procedure which remained valid once overturning had occurred.

Numerical examples were given in chapter 5 to demonstrate the methods given in chapters 2-4. From these results it has been seen that extensions to the method are required for the interaction of shocks. It has also been seen that the method is dependent upon the initial data representation.

In chapter 6 the analytical methods in 2-D were considered and we found that analytical solutions have only been found for a very small class of problems. The main class of problems, for which analytic solutions have been found are Riemann problems. The ideas of characteristics and blow-up generalize from 1-D to higher dimensions and supported the extension of the numerical methods from 1-D to 2-D.

Chapter 7 discussed the background to the basic MFE and Lagrangian methods considered in this thesis and chapter 8 described the modifications needed in order to calculate multivalued solutions. In chapter 8, a method of shock recovery in 2-D was proposed. This led to the numerical experiments being carried out in chapter 9.

The results of chapter 9 showed that the method was able to calculate the shock position for a variety of problems. The method was also found to be initial grid dependent and in its present form is unable to cope with the calculation of shock interactions.

It is clear that approximate solutions can be calculated using the techniques described, but we have also seen that there are problems in the recovery of the shock position in two dimensions. Another problem that may occur in both one and two dimensions is the inability of the method to cope with the interaction of shocks.

10.2 Further Work

Following the work in this thesis we can see that there are several paths available for future work. Another development would be to generalise the method to equations other than scalar conservation laws. Although both the MFE methods and Lagrangian methods can deal with the solution of the conservation laws, for other non-linear PDE's only the MFE method is available. The generalisation is not difficult and much of the work of the thesis goes over immediately to such equations.

The method has so far only been applied to scalar conservation laws. The method for finding the position of the shock in 2-D is only an extension of the 1-D method.

The first extension would be to find a proven 2-D method for recovering the shock position from the overturned manifold. One way for doing this would be to return to the ideas of conservation of area and tangential continuity, seeing whether a condition could be proved in order to generate a unique shock position.

It would also be useful to generalize the shock fitting algorithm so that it can cope with the interaction of shocks.

A final path would be an extension of the use of the Lagrangian method in conjunction with the shock fitting technique. The idea is that this method could then be applied to systems of conservation laws. Since systems of conservation laws such as the Euler equations can by solved using Lagrangian methods, then it would remain to choose a 'Monitor' variable (e.g. density) to which our shock fitting technique could be applied. This should then give the position of the shock formed for all variables of the system.

References

Alexander, R., Manselli, P. & Miller, K. (1979) *Moving Finite Elements for the Stephan problem in two-dimensions.* Accadamia Nazionale dei Lincei. Serie VIII Vol LXVII Fasc. 1-2 pp 57-61.

Baines, M.J. (1985) *Local Moving Finite Elements.* Numerical Analysis Report 9/85, University of Reading, U.K.

Baines, M.J. (1986) *Moving Finite Elements: Regularisation Techniques.* Numerical Analysis Report 19/86, University of Reading, U.K.

Baines, M.J. (1990) *On best piecewise linear L_2 fits with adjustable nodes: the 2-D case.* Numerical Analysis Report 15/90, University of Reading, U.K.

Baines, M.J. (1991) *An analysis of the moving finite element procedure.* SIAM J. Numer. Anal. To appear.

Baines, M.J. & Reeves, C.P. (1990) *Moving Finite Element Procedures and Overturning Solutions.* Numerical Analysis Report 11/90, University of Reading, U.K.

Baines, M.J. & Wathen, A.J. (1988) *Moving Finite Element Methods for Evolutionary Problems (I) Theory.* J. Comput. Phys. **79,** pp 245-269.

Böing, H., Werner, K. & Jackisch, H. (1991) *Construction of the Entropy Solution of Hyperbolic Conservation Laws by a Geometrical Interpretation of the Conservation Principle.* J. Comput. Phys. **95** pp 40-58.

Boris, J.P. & Book, D.L. (1973) *Flux corrected Transport. I SHASTA, A Fluid Transport Algorithm that works.* J. Comput. Phys. **11** pp 38-69.

Brenier, Y. (1984) *Averaged Multivalued Solutions for Scalar Conservation Laws.* SIAM J. Numer. Anal. **21** pp 1031-1037.

Buckley, S.E. & Leverett, M.C. (1942) *Mechanism of Fluid Displacement in Sands.* Trans. AIME **146** pp 107-116.

Carey, G.F. & Dinh, H.T. (1985) *Grading Functions and Mesh Redistribution.* SIAM J. Numer. Anal. **22** pp 1028-1040.

Carlson, N. & Miller, K. (1986) *Gradient Weighted Moving Finite Elements in two Dimensions.* PAM-347. Center for Pure and Applied Mathematics, University of California, Berkley, USA.

Chang, T. & Klingenberg, C. (1986) *The Riemann problem of a scalar hyperbolic conservation law in two space dimensions.* preprint - Maths. Institute, University of Heidelberg, Germany.

Concus, P. & Proskurowski, W. (1979) *Numerical Solution of a Nonlinear Hyperbolic Equation by the Random Choice Method.* J. Comput. Phys. **30** pp 153-166.

Conway, E. (1977) *The formation and decay of shocks for Conservation laws in Several Dimensions.* Arch. Rat. Mech. Anal. **64** pp 135-151.

Conway, E. & Smoller, J. (1966) *Global solutions of the Cauchy problem for quasi-linear 1st order equations in several space variables.* Comm. Pure Appl. Math. **19** pp 95-105.

Courant, R. & Hilbert, D. (1962) Methods of Mathematical Physics. Vol 2. Wiley, New York.

Djomeri, J., Doss, S., Gelinas, R., & Miller, K. (1985) *Applications of the moving finite element for systems.* J. Comput. Phys. pp 1-41.

Donea, J. (1984) *A Taylor-Galerkin Method for Convective Transport Problems.* Int. Num. Meth. Eng. **20** pp 101-119.

Edwards, M.G. (1988) *The Mobile Element Method for Systems of Conservation Laws.* Proc. of the VIIth GAMM conference. M.Deville (ed). pp 72-79. Viewig.

Golub, G.H. & Van Loan, C.F. (1983) *Matrix Computations.* North Oxford Academic. 1st ed.

Guckenheimer, J. (1975) *Shocks and rarefactions in two-space dimensions.* Arch. Rational Mech. Anal. **59** pp 281-291.

Harten, A. (1983) *High Resolution Schemes for Conservation Laws.* J. Comput. Phys. **49** pp 357-393.

Harten, A., Hyman, J.M. & Lax, P.D. (1976) *On finite-difference approximations and entropy conditions for shocks.* Comm. Pure & Appl. Math. **29** pp 297-322.

Hawken, D.F., Gottlieb, J.J. & Hansen, J.S. (1991) *Review of Some Adaptive Node-Movement Techniques in Finite-Element and Finite-Difference Solutions of Partial Differential Equations.* J. Comput. Phys. **95** pp 254-302.

Herbst, B.M. (1982) *Moving Finite Element Methods For The Solution Of Evolutionary Problems.* Ph.D. Thesis. University of the Orange Free State, South Africa.

Hsaio, & Klingenberg, C. (1984) *The construction of the solution to a non-convex two-dimensional Riemann problem.* Maths. Institute, University of Heidelberg, Germany.

Jimack, P. (1988a) *High order moving finite elements I.* Report number:AM-88-03, School of Mathematics, University of Bristol, U.K.

Jimack, P. (1988b) *High order moving finite elements II.* Report number:AM-88-11, School of Mathematics, University of Bristol, U.K.

John, F. (1971) *Partial Differential Equations.* Springer-Verlag.

Johnson, I.W. (1984) *The Moving Finite Element for the viscous Burger's Equation.* Numerical Analysis Report 3/84, University of Reading, U.K.

Johnson, I.W. (1986) *Moving Finite Elements For Diffusion Problems.* PhD. Thesis. University of Reading, U.K.

Johnson, I.W., Wathen, A., & Baines, M.J. (1988) *Moving Finite Elements for Evolutionary Problems (II) Applications.* J. Comput. Phys. **79** pp 270-297.

Johnson, L.W. & Dean Riess, R. (1982) *Numerical Analysis.* Addison-Wesley.

Juarez-Romero, D., Sargent, R.H.W. & Jones, W.P. (1988) *Improving the robustness of the moving finite element method.* Comp. Chem. Engineering. **12** pp 433-442.

Klingenberg, C. (1986) *Hyperbolic conservation laws in two dimensions: some numerical and theoretical results.* Report 2: (1986) Institut Mittag-Leffler.

Kružkov, S.M. (1969) *Generalized solutions of the Cauchy problem in the large for nonlinear equations of first order.* Soviet Math. Dokl. **10** pp 785-788.

Kružkov, S.M. (1970) *First Order Quasilinear Equations with Several Independent Variables.* Mat. Sb. **81** (123) pp 217-243.

Lax, P.D. (1972) *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves.* SIAM Regional Conference Series. Lectures in Applied Mathematics II.

Lax, P.D. & Wendroff, B. (1960) *Systems of Conservation Laws.* Comm. Pure Appl. Math., **13** pp 217-237.

van Leer, B. (1979) *Towards the Ultimate Finite Difference Scheme V. A Second Order Scheme to Godunov's Method.* J. Comput. Phys. **32** pp 101-136.

Lindquist, W.B. (1986) *Construction of solutions for two-dimensional Riemann problems.* Comp. and Math. with Appl. **129** Nos 4/5 pp 615-630.

Lynch, D.R. (1982) *Unified approach to simulations on deforming elements with applications to phase change problems.* J. Comput. Phys. **47** pp 347-411.

Majda, A. (1984) *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables.* Springer-Verlag.

Malcolm, A.J. (1991) *Data dependent triangular grid generation.* PhD. Thesis, University of Reading, U.K. (In preparation.)

Miller, K. (1981) *Moving Finite Elements, Part II.* SIAM J. Numer. Anal. **18.** pp 1033-1057.

Miller, K. (1986) *Recent results on Finite Element Methods with moving nodes.* Accuracy Estimates and Adaptive Refinement in Finite Element Calculations. Babuška, I., Zienkiewicz, O.C. et als. eds. Wiley, New York. pp 325-338.

Miller, K. (1988) *On the mass matrix spectrum bounds of Wathen and the local moving finite elements of Baines.* PAM-430. Center for Pure and Applied Mathematics. University of California, Berkley, USA.

Miller, K. & Miller, R.N. (1981) *Moving Finite Elements, Part I.* SIAM J. Numer. Anal. **18** pp 1019-1032.

Mitchell, A.R. & Griffiths, D.F. (1980) *The Finite Difference Method in Partial Differential Equations.* Wiley-Interscience.

Morton, K.W. (1985) *Generalized Galerkin methods for hyperbolic problems.* Comp. Meth. in Appl. & Engng. **52** pp 847-871.

Mueller, A.C. & Carey, G.F. (1985) *Continuously deforming finite elements for transport problems.* Int. J. for Numer. Meth. in Engng. **21** pp 2099-2126.

Oleinik, O.A. (1957) *Discontinuous solutions of Non-Linear Differential Equations.* Amer. Math. Soc. Transl. Ser. 2 **26** pp 95-171.

Osher, S. (1984) *Riemann solvers, the entropy condition, and difference approximations.* SIAM J. Numer. Anal. **21** pp 217-235.

Reeves, C.P. (1989) *A Comparison Between Two Moving Finite Element Methods.* Numerical Analysis Report 17/89. University of Reading, U.K.

Richtmyer, R. & Morton, K.W. (1967) *Difference Methods for Initial Value Problems.* Wiley-Interscience.

Roe, P.L. (1983) *An Introduction to Numerical Methods Suitable for the Euler Equations.* Lecture notes for the von Karman Institute for Fluid Dynamics. Lecture Series 'Introduction to Computational Fluid Dynamics'. January 24-28th.

Sewell, M.J. (1987) *Maximum and minimum principles.* Cambridge.

Smoller, J. (1983) *Shock Waves and Reaction-Diffusion Equations.* Springer-Verlag, Berlin.

Sweby, P.K. (1984) *High Resolution Schemes using Flux Limiters for Hyperbolic Conservation Laws.* SIAM J. Numer. Anal. **21** pp 995-1011.

Sweby, P.K. (1987) *Some Observations on the Moving Finite Element Method and its Implications.* Numerical Analysis Report. 13/87. University of Reading, U.K.

Sweby, P.K. (1990) Private Communication.

Tadmor, E. (1984) *Numerical viscocity and the entropy condition for conservative difference schemes.* Math. Comp. **43** pp 369–381.

Vol'pert, A.I. (1967) *The spaces BV and quasilinear equations.* Math USSR-Sb **2** pp 225-267.

Wagner, D.H. (1983) *The Riemann problem in two space dimensions for a single conservation law.* SIAM. J. Math. Anal. **14** pp 534-559.

Wathen, A.J. (1984) *Moving Finite Elements And Oil Reservoir Modelling.* Ph.D. Thesis, University of Reading, U.K.

Wathen, A.J. (1987) *Realistic Eigenvalue Bounds for the Galerkin Mass Matrix.* I.M.A. J. Numer. Anal. **7** pp 449-457.

Wathen A.J. & Baines, M.J. (1984) *On the Structure of the Moving Finite Element Equations.* IMA J. Num. Anal. **5** pp 161-182.

Whitham, G.B. (1974) *Linear and Nonlinear Waves.* Wiley-Interscience.