SOLUTION ADAPTIVE MESH PROCEDURES FOR THE NUMERICAL SOLUTION OF SINGULAR PERTURBATION PROBLEMS

Thesis by
David Leslie Brown

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California
1982
(submitted April 13th, 1982)

ACKNOWLEDGEMENTS

I would like to thank my advisor Heinz-Otto Kreiss for his invaluable guidance during my tenure as a graduate student at Caltech and at the University of Uppsala, Sweden. Thanks are also due to my many friends at Caltech, in particular the members of the applied mathematics department and the participants in the Caltech chamber music program. The students and faculty of the numerical analysis group and the Stanford Exploration Project at Stanford made my stay there during the summer of 1981 a pleasant one. Special thanks are also due to Luis Reyna who worked together with me on the development of some of the numerical methods presented in this thesis.

Financial support was provided by Institute teaching assistantships and research assistantships under Office of Naval Research contract no. N00014-80-C0076, Department of Energy contract no. DE-AS03-76SF-00766 and Army Research Office contract no. DAAG29-78-C-0011. Additional computer time was provided by the Stanford Exploration Project of the Stanford University Geophysics Department and by NASA-Ames Research Center.

Finally, I would like to dedicate this thesis to the memory of my grandfather, J. Fish Smith, and to my grandmother Lillian Fountain Smith, whose many accomplishments continue to be an inspiration to me.

Solution Adaptive Mesh Procedures for the Numerical Solution of Singular Perturbation Problems

David Leslie Brown, Ph.D.
California Institute of Technology, 1982

Abstract

The accurate numerical solution of singular perturbation problems by finite difference methods is considered. (For efficient computations of this type, refinement of the finite difference mesh is important. The technique of solution-adaptive mesh refinement, in which the mesh is refined iteratively by looking at the properties of a computed solution, can be the simplest method by which to implement a mesh refinement.) The theoretical justification of solution-adaptive mesh refinement for singularly perturbed systems of first order ordinary differential equations (ODEs) is discussed. It is shown that *a posteriori* error estimates can be found for weighted one-sided difference approximations to systems of ODEs without turning points and to systems of ODEs with turning points that can be transformed to a typical normal form. These error estimates essentially depend only on the local meshwidths and on lower order divided differences of the computed solution, and so can be used in the implementation of solution-adaptive mesh refinement. It is pointed out, however, that not all systems with turning points fall into these categories, and solution-adaptive mesh refinement can sometimes be inadequate for the accurate resolution of solutions of these systems.

Numerical examples are presented in which the solutions of some model equations of fluid dynamics are resolved by transforming the problems to singularly perturbed ODEs and applying weighted one-sided difference approximations with solution-adaptive mesh refinement. In particular, well-resolved steady and moving shock solutions to Burgers' equation and to the equations of one-

dimensional isentropic gas dynamics are obtained numerically. The method is further extended to problems in two space dimensions by using the method of dimensional splitting together with careful interpolation. In particular, in this extension the mesh refinement is only used to resolve the one-dimensional problems which are solved within the splitting algorithm. Numerical examples are presented in which two-dimensional oblique shocks are resolved.

Table of Contents

I. Preliminaries

1.1 Numerical Methods for Singular Perturbation Problems (Introduction)

In this thesis we consider the accurate numerical solution of the two-point boundary value problem for a system of singularly perturbed ordinary differential equations (ODEs). An example of such a problem is given by

$$\mathbf{y}'(x) + \left\{\frac{1}{\varepsilon}\begin{bmatrix} A_{11}(x) & A_{12}(x) \\ 0 & 0 \end{bmatrix} + A_1(x)\right\}\mathbf{y}(x) = \begin{bmatrix} \frac{1}{\varepsilon}\mathbf{f}^I(x) \\ \mathbf{f}^{II}(x) \end{bmatrix} \tag{1}$$

on $0 \le x \le 1$ together with n linearly independent boundary conditions on $\mathbf{y}(0)$ and $\mathbf{y}(1)$.[1] Here $0 < \varepsilon \ll 1$ is a small parameter, $\mathbf{y}$ is a vector of length n, $\mathbf{f}^I := (f^{(1)}, f^{(2)}, \ldots, f^{(m)})^T$, $\mathbf{f}^{II} := (f^{(m+1)}, \ldots, f^{(n)})^T$, A_1 is an $n \times n$ matrix, A_{11} is an $m \times m$ matrix, A_{12} is an $m \times (n-m)$ matrix, and $m < n$. The elements of those matrices are assumed to be $\mathbf{O}(1)$. One reason this is called a *singularly* perturbed system is that the *reduced problem*, given by setting $\varepsilon = 0$ in (1):

$$\begin{aligned} &A_{11}(x)\mathbf{y}^I + A_{12}(x)\mathbf{y}^{II} = \mathbf{f}^I(x) \\ &\frac{d\mathbf{y}^{II}}{dx} + B_{22}(x)\mathbf{y}^{II} + B_{21}(x)\mathbf{y}^I = \mathbf{f}^{II}(x) \end{aligned} \tag{2}$$

is a system of ODEs of lower order than (1). The boundary conditions also have to be changed accordingly since there are only $n-m$ free parameters in the solution of (2). Here $\mathbf{y}^I := (y^{(1)}, y^{(2)}, \ldots, y^{(m)})^T$, $\mathbf{y}^{II} := (y^{(m+1)}, \ldots, y^{(n)})^T$ and

[1] In this thesis, equations, theorems and lemmas are numbered consecutively within each section. When referring to an equation in another section, the section number is prepended to the equation number, e.g. equation 1.3.5 is equation (5) in section 1.3. Also, we will usually use the following notation: Boldface (e.g. $\mathbf{y}$) is used to denote a vector, the ith component of $\mathbf{y}$ is denoted by $y^{(i)}$, $\mathbf{y}^T$ is the transpose of $\mathbf{y}$.

we have partitioned A_1 as follows:

$$A_1 = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

The problem (1) is sometimes referred to as a *stiff* boundary value problem in the numerical literature (cf. Hemker [1974], [1977]).

We are particularly interested in non-oscillatory stiff problems with turning points in the interior of [0,1]. By "non-oscillatory" we mean that those eigenvalues, $\lambda_i(x)$, of A_{11} that are not identically zero satisfy $|\mathrm{Im}\lambda_i| \leq \rho |\mathrm{Re}\lambda_i|$, where ρ is a constant of moderate size. These eigenvalues are assumed to be nonzero everywhere on [0,1] except possibly at a finite number of points x_j, $j = 1,2,\ldots,J$ where one of those eigenvalues is zero. Such a point x_j is called a *turning point* of the system (1).

Different parts of the solutions of (1) typically vary on different scales. For example, boundary layers can occur at $x = 0$ and $x = 1$, and internal layers can occur near the turning points x_j. Both of these types of phenomena have narrow widths that depend on the size of the small parameter ε. We are interested in the accurate approximate solution of problems of this type using finite-difference methods. For very small ε, the ratio of the scales on which the boundary and internal layers and the smooth parts of the solution vary can be extremely large. For an accurate representation of the solution everywhere on a finite interval, therefore, a uniform (i.e. equally-spaced) finite difference mesh would need to have a very large number of meshpoints. Since such a fine mesh is not needed in the smooth parts of the solution, this naturally leads to the idea of using *nonuniform* meshes, i.e. the local meshwidth is allowed to vary so as to be of appropriate size for the resolution of the local behavior of the solution. Simply stated, we want lots of meshpoints in the boundary and internal layers and few points in the smooth parts of the solution.

The question of the best procedure for deciding how to construct the mesh is an important one. Essentially, we seek a change of independent variable such that the solution behaves in a smooth way as a function of the new independent variable. Probably the safest approach for constructing a mesh is if we can determine *a priori* what kind of behavior to expect in the solution by looking at the coefficients of the differential equation and at the boundary conditions. If there are no turning points in the problem, then the rapidly varying parts of the solutions will be boundary layers, and correspondingly a finer mesh can be used

near the boundaries. Exactly how to stretch the mesh near the boundary in order to resolve the boundary layer is fairly well known (although perhaps not well understood) (see Keller and Cebeci [1972], p. 1198). For problems with turning points, fairly detailed information is required in order to properly refine the mesh in the turning point regions. Kreiss and Nichols [1975] and Kreiss [1976] have considered this approach to designing a mesh for problems with turning points. The procedure they discuss involves finding transformations of both the dependent and independent variables such that the resulting system is everywhere diagonally dominant (see also chapter 2 of this thesis for a discussion of diagonal dominance). The diagonal dominance insures that a difference approximation can be found that will give solutions that behave in the proper way.

One disadvantage of the approach taken by Kreiss and Nichols is that it can be computationally quite involved to implement. For this reason, the technique of *solution-adaptive mesh refinement* is much preferred (although for some problems not as reliable, as we will discuss later). This technique is one in which the mesh is determined iteratively by looking only at the computed solution to the problem. The justification for this approach is that for some problems, the error in a computed solution can be reliably estimated in terms of products of some power of the local meshwidth and some lower-order divided differences of that solution. This is essentially because the error in the computed solution depends on the truncation error of the finite-difference method which in turn depends on the local meshwidths and some derivatives of the true solution. In chapter 2 we address the question of when solution-adaptive mesh refinement can be expected to work for singular perturbation problems. We consider systems similar to (1) and find conditions on the systems such that the error can reliably be estimated in terms of the local meshwidths and some lower order divided differences of the computed solution. In particular, we find that if an appropriate difference approximation is used, solution-adaptive mesh refinement is justified for linear systems without turning points and for systems with turning points that can be transformed to a normal form discussed in section 2.4. While it is certainly possible that solution-adaptive mesh refinement will work for other types of systems, this is not true in general. In general a full investigation of the behavior of the coefficients of the differential equation such as is discussed by Kreiss and Nichols [1975] may be necessary to reliably estimate the error.

Solution-adaptive mesh refinement has been investigated both theoretically and by computational experiments by many workers, although very few people have considered its application to singular perturbation problems specifically. Denny and Landis [1971] discuss a method for solving a second order ODE on a finite interval in which the mesh is determined by attempting to drive the truncation error to zero pointwise using an iterative procedure. Pereyra and Sewell [1975] discuss theoretically the construction of "equidistributing" meshes. This is a technique in which the mesh is chosen in such a way as to distribute the local error evenly among the meshpoints. The justification for this procedure is that it solves a minimization problem for the global error in the finite difference solution. Applications of this technique are discussed by Lentini and Pereyra [1977]. More recent theoretical work can be found in Pierson and Kutler [1979] and Kautsky and Nichols [1981].

Pearson [1968] is one of the earliest workers to use solution-adaptive mesh refinement to solve singular perturbation problems with turning points numerically. He presents numerical experiments in which a centered difference scheme is used together with adaptive mesh refinement (he tests the difference between adjacent values of the computed solution) to solve second-order scalar equations of the form

$$\varepsilon y'' + a(x)y' - b(x)y = f(x).$$

Because of his use of a centered difference approximation, however, he was forced to use a very fine mesh initially, and often employed continuation in the parameter ε, presumably to avoid the rapidly oscillatory large amplitude error which is associated with centered schemes applied to such problems. Indeed, in order to avoid unnecessary mesh refinement it is important to use a difference scheme that is well-suited for solving singular perturbation problems on a coarse mesh. It has been known for some time that one-sided difference approximations give qualitatively the correct behavior for problems in which rapid transitions occur in the solutions. For example, the use of such methods for shock calculations has been discussed since the early 1950's, and is reviewed in sections 1.2 through 1.4 of this thesis as well. Dorr [1970] has discussed the use of one-sided schemes for solving singular perturbation problems in particular. He considers a system of two second-order nonlinear ODEs with a small parameter and with turning points and proves that for uniform meshwidth h, one-sided schemes give asymptotically the correct behavior for fixed h as the

small parameter ε tends towards zero. In contrast, he proves that for centered schemes, the correct behavior is only obtained for $\varepsilon \to 0$ with h/ε fixed. Similar results for systems of second-order linear ODEs without turning points and a second order scalar equation with a turning point are given by Abrahamsson et. al. [1974] and Abrahamsson [1975b], respectively. For the difference approximations considered by Dorr [1970] and by Abrahamsson [1975b], the results depend very much on the fact that the one-sided schemes have a discrete maximum priniciple (see also sections 3.1, 3.2 of this thesis).

One of the few papers in which solution-adaptive mesh refinement is used in conjunction with one-sided difference approximations for solving singular perturbation problems is by B. Kreiss and H.-O. Kreiss [1981]. They very successfully used a weighted one-sided difference approximation and mesh refinement to solve second-order linear and nonlinear ODEs with turning points. In particular, relatively few meshpoints were required to resolve the solutions, even for very small values of ε. The theoretical justification for the mesh refinement procedure employed by Kreiss and Kreiss was first outlined by Kreiss [1975]. A more detailed discussion and some extensions of that theory are given in chapter 2 of this thesis.

One field in which singular perturbation problems with turning points arises is in the study of the partial differential equations that describe shocks in fluids. Since this area has been an active one for numerical computations, much work has been done on numerical methods for computing shocks, particularly in the limiting case of vanishing viscosity. Although the emphasis in such computations is typically not on the resolution of the viscous profiles of shocks, the modern finite difference methods which are used are designed to give qualitatively correct results on a coarse mesh. Since difference schemes with this property are important for use in conjunction with solution-adaptive mesh refinement, it is of interest to review the properties of the methods used for computing solutions of conservation laws. (Indeed, Osher [1981] and Abrahamsson and Osher [1981] have applied the method of Engquist and Osher [1979], which was designed for the numerical solution of conservation laws, to the solution of singular perturbation problems). For this reason, the rest of this chapter is devoted to the discussion of numerical methods for time-dependent shock calculations.

As was mentioned above, chapter 2 of this thesis is concerned with the theoretical justification of solution-adaptive mesh refinement for singularly

perturbed ODEs. In chapter 3 we apply the theory of chapter 2 to the numerical solution of some model equations in fluid dynamics. In these computations we are interested in the use of solution adaptive mesh refinement for the resolution of steady and moving viscous shock profiles. Sections 3.3 and 3.4 give numerical examples of the application of this method to Burgers' equation. Section 3.5 is devoted to the numerical solution of the equations of isentropic gas dynamics in one space dimension. In all cases, we have reduced the problem to a singularly perturbed system of first-order ODEs and applied either the weighted one-sided difference method of Kreiss and Kreiss [1981] or a modification of that method (discussed in sections 3.1 and 3.2) together with solution-adaptive mesh refinement.

Chapter 4 describes a method developed together with Luis Reyna for resolving shock profiles in two space dimensions using the technique of dimensional splitting together with the methods developed in chapter 3 for the resolution of one-dimensional shock profiles. Of particular interest is the fact that this method only requires local refinement of the one-dimensional problems that result from the splitting procedure. The one-dimensional problems which are solved alternately in the x and y-directions are connected together by careful interpolation that does not degrade the resolution obtained in the alternating direction sweeps.

1.2 Differential Equations with Shock Solutions

Perhaps the most common example of a system of partial differential equations describing a physical situation in which shocks may occur is that of the equations of gas dynamics. In one space dimension and including all dissipative effects this system takes the form (see for example Whitham [1974], chapter 6)

$$\begin{aligned} &\rho_t + (\rho u)_x = 0 \\ &(\rho u)_t + (\rho u^2 - p_{11})_x = 0 \\ &(\rho u^2/\ 2 + \rho e)_t + [(\rho u^2/\ 2 + \rho e)u - p_{11}u + q\,]_x = 0 \end{aligned} \tag{1}$$

Here $\rho(x,t)$ is the density of the fluid, $u(x,t)$ is its velocity, $e(x,t)$ is its internal energy per unit mass, $p_{11}(x,t)$ is the stress on an element of fluid, and $q(x,t)$ is the heat conduction of the fluid. If we insert the Navier-Stokes relations for these last two quantities into (1), the resulting equations are given by

$$\rho_t + (\rho u)_x = 0$$

$$(\rho u)_t + (\rho u^2 + p)_x = \frac{4}{3}\mu u_{xx} \tag{2}$$

$$(\frac{1}{2}\rho u^2 + \rho e)_t + [(\frac{1}{2}\rho u^2 + \rho e)u + up]_x = \frac{4}{3}\mu(uu_x)_x + \lambda T_{xx}$$

where $p(x,t)$ is the pressure and $T(x,t)$ the temperature of the fluid. The parameters μ and λ are the coefficients of viscosity and heat conduction respectively and for the purposes of this discussion are assumed to be constants. When equations (2) describe the motion of a gas, these parameters can be quite small compared with unity and correspondingly the equations may admit solutions in which regions of steep gradient occur. As λ and μ tend towards zero these gradients can become infinite and the corresponding limiting solutions no longer satisfy the limiting form of equations (2) in the classical sense. This difficulty is easily resolved when we realize that equations (1) were derived from integral conservation laws by assuming that the solutions were sufficiently smooth. The limiting solutions can therefore be understood as solutions of the integral forms of equations (2) (see Whitham [1974]). Another equivalent approach which will be useful when discussing numerical methods is to say that equations (2) should always be satisfied *in the sense of distributions*. By this we mean that we multiply equations (2) by a smooth test function, $\varphi(x,t)$, and integrate by parts over the domain $t \geq 0, -\infty < x < \infty$ to obtain

$$\int\int_{t>0} (\varphi_t \rho + \varphi_x \rho u)\, dx\, dt = 0$$

$$\int\int_{t>0} (\varphi_t \rho u + \varphi_x(\rho u^2 + p))\, dx\, dt = -\int\int_{t>0} \frac{4}{3}\mu u \varphi_{xx}\, dx\, dt \tag{3}$$

$$\int\int_{t>0} ((\frac{1}{2}\rho u^2 + \rho e)\varphi_t + ((\frac{1}{2}\rho u^2 + \rho e)u + up)\varphi_x)\, dx\, dt$$

$$= -\int\int_{t>0} \frac{2}{3}\mu u^2 \varphi_{xx}\, dx\, dt - \int\int_{t>0} \lambda T \varphi_{xx}\, dx\, dt$$

(Here we have neglected the initial conditions that should appear in "by parts" terms in equations (3)). In the ("inviscid") limit of $\lambda, \mu \to 0$ the right-hand sides of (3) vanish but the equations are still valid because the physical variables ρ, u, p, and e are undifferentiated. A solution of (3) with discontinuities is called a "weak" solution of the system (2).

For $\lambda = \mu = 0$ equations (3) together with initial conditions for $t = 0$ are not sufficient to uniquely determine a solution for all $t > 0$. There is, however, a

unique physically relevant solution which is the limiting solution of equations (3) as $\lambda,\mu \to 0$. This solution can be determined without explicit reference to the limiting process if we note that by combining equation (2) we can obtain another conservation law of the form

$$(\rho S)_t + (\rho u S)_x = \frac{4\mu(u_x)^2}{3T} + \lambda\frac{T_x^2}{T^2} + \lambda\frac{\partial^2}{\partial x^2}\log T \tag{4}$$

which can also be written as

$$\int\int_{t>0} (\varphi_t \rho S + \varphi_x \rho u S) = -\int\int_{t>0} \lambda\varphi_{xx}\log T - \int\int_{t>0}\left[\frac{4\mu(u_x)^2}{T} + \lambda\frac{T_x^2}{T^2}\right]\varphi \tag{5}$$

(Here S is the entropy of the fluid. It is related to the other physical variables by the differential relation $TdS = de + pd(\frac{1}{\rho})$.) As $\lambda,\mu \to 0$, the first integral on the right-hand side of (5) goes to zero uniformly while the remaining terms are ≤ 0. Hence we conclude that

$$\lim_{\mu,\lambda\to 0} \int\int_{t>0} \varphi_t(\rho S) + \varphi_x(\rho u S) \leq 0 \tag{6}$$

Equation (6), which implies that entropy must increase or remain constant across a discontinuity in the solution, is called the "*entropy condition*". It is also sometimes called an "*integral entropy inequality*" for the system (2).

Nonuniqueness of solutions is a difficulty that is associated with nonlinear hyperbolic conservation laws in general. An approach for deriving integral inequalities that enforce uniqueness of the solutions to a given system of hyperbolic equations has been developed by Lax [1971] and others. Hopf [1969] gives a particularly lucid account of this technique for scalar conservation laws which we repeat here: We consider a scalar hyperbolic conservation law of the form

$$u_t + f(u)_x = \varepsilon u_{xx} \tag{7}$$

Here u is the conserved quantity, $f(u)$ is its flux, and $0 < \varepsilon \ll 1$ is a dissipative coefficient (viscosity, for example). We introduce an auxiliary function $U(u)$ called the "*entropy function*" (which may or may not have any relation to the physical entropy of the problem). The only requirement we make on U is that it be a convex function of u, i.e. $U''(u) \geq 0$. We define the corresponding "*entropy flux*" $F(u)$ by the relation

$$F'(u) = U'(u)f'(u) \tag{8}$$

Multiplying (7) by $U'(u)$ we obtain

$$U(u)_t + F(u)_x = \varepsilon U(u)_{xx} - \varepsilon U''(u)(u_x)^2. \tag{9}$$

We now multiply (9) by a smooth positive test function $\varphi(x,t)$ with compact support on $t \geq 0$, $-\infty < x < \infty$ and integrate the result over that region. Integration by parts gives

$$\int\int_{t>0} (U(u)\varphi_t + F(u)\varphi_x)\, dx\, dt = -\int\int_{t>0} \varepsilon U(u)\varphi_{xx}\, dx\, dt$$

$$+ \int\int_{t>0} \varepsilon U''(u)(u_x)^2\varphi\, dx\, dt \tag{10}$$

In the same way as before we conclude, therefore, that

$$\lim_{\varepsilon \to 0} \int\int_{t>0} U(u)\varphi_t + F(u)\varphi_x \geq 0. \tag{11}$$

In analogy to (6), the inequality (11) is called an "*entropy inequality*" for the conservation law

$$u_t + f(u)_x = 0 \tag{12}$$

If $u(x,t)$ is a piecewise continuous weak solution of (12) in which a single jump from a value of u^+ to u^- occurs along a line $x = x(t)$ in the $x-t$ plane, then (11) can be shown equivalent to

$$(U(u^+) - U(u^-))\dot{x}(t) - (F(u^+) - F(u^-)) \geq 0 \tag{13}$$

If we choose $U(u) = \pm u$, then $F(u) = \pm f(u)$ and (13) gives the Rankine-Hugoniot law for the speed of the discontinuity:

$$\dot{x}(t) = \frac{f(u^+) - f(u^-)}{u^+ - u^-} \tag{14}$$

If we choose first $U(u) = (u-u_0)H(u-u_0)$ and then $U(u) = (u-u_0)(H(u-u_0) - 1)$ ($H(u)$ is the Heaviside step function defined following equation (1.3.1)), then we obtain both inequalities of Oleinik's "condition E" (Oleinik [1963]):

$$\frac{f(u_0) - f(u^-)}{u_0 - u^-} \geq \dot{x}(t) \geq \frac{f(u^+) - f(u_0)}{u^+ - u_0} \tag{15}$$

Oleinik proved that inequality (15) is sufficient to guarantee uniqueness of the solutions of the initial value problem for (12). Quinn [1971] (see also Lax [1972])

showed that this uniqueness follows from the fact that the condition (15) guarantees the solution operator for (12) is an L_1-contraction.

If we consider the limiting cases $u_0 \to u^-$ and $u_0 \to u^+$ in condition (15), it is clear that condition E is equivalent to

$$f'(u^-) \le \dot{x}(t) \le f'(u^+) \tag{16}$$

which is just the geometrical condition that the characteristics must point into a shock from both sides.

Lax [1971] has shown that for strictly hyperbolic systems of conservation laws for a vector function $\mathbf{u}(x,t)$, all entropy inequalities of the form (11) are equivalent provided that for each system of equations, the entropy functions $U(\mathbf{u})$ which are considered are strictly convex functions of $\mathbf{u}$ and the weak solutions $\mathbf{u}(x,t)$ of the conservation laws have only moderately strong discontinuities. Harten [1981] has pointed out that while this result is also true for *strong* shock solutions of the equations of fluid dynamics, there exist examples showing that all entropy functions for a given system of conservation laws do not necessarily give the same unique solution for abitrarily strong shocks.

Difference approximations to systems of conservation laws of the form

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0 \tag{17}$$

can also suffer from the problem of admitting solutions that are physically unreasonable, i.e. that do not satisfy an entropy condition. For this reason, a current approach for constructing reasonable difference approximations for systems (17) is to assure that in addition to approximating weak solutions of (17) well, the solutions of these difference approximations satisfiy an entropy inequality in some appropriate sense (see e.g. Engquist and Osher [1979]). This approach is discussed further in section 1.4. In the next section, however, we first discuss some results on the behavior of difference approximations with discontinuous solutions. Section 1.4 then briefly reviews some of the more modern difference approximations used for solving hyperbolic systems of equations that admit shock solutions.

1.3 The Behavior of Difference Approximations with Discontinuous Initial Data

Some insight into the behavior of the solutions of difference approximations for problems involving shocks and contact discontinuities can be gained by considering the behavior of difference approximations to linear hyperbolic problems with discontinuous initial data. We begin by considering the problem given by

$$\frac{\partial u(x,t)}{\partial t} = \rho(x)\frac{\partial u(x,t)}{\partial x} \tag{1}$$

for $-\infty < x < \infty$, $t \geq 0$ with step-function initial data:

$$u(x,0) = 1 - H(x)$$

Here ρ is a real function of x and

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

is the "Heaviside step function", sometimes called the "characteristic function of the interval $[0,\infty)$." The solution to this problem is a step function moving to the left with local velocity $\rho(x)$. For the case ρ = const. this can be expressed as

$$u(x,t) = 1 - H(x+\rho t)$$

We approximate this problem using a finite-difference method: Establish a uniform mesh with meshpoints $x_\nu = \nu h$, $\nu = -\infty, \cdots, -1, 0, 1, \cdots \infty$, meshwidth h and timestep k with $k/h =: \lambda$ a constant and replace (1) with

$$v(x_\nu, t+k) = Qv(x_\nu,t) = \sum_{j=-\infty}^{\infty} a_j v(x_{\nu+j},t) \tag{2}$$

with initial data $v(x_\nu,0) = u(x_\nu,0)$. We assume that the method is accurate of order p, i.e.

$$Q(\xi) := \sum_{j=-\infty}^{\infty} a_j e^{ij\xi} = \exp(i\lambda\rho\xi(1+\beta\xi^p)(1+o(1))) \quad as\ \xi\to 0 \tag{3}$$

and dissipative of order $2s$, i.e. there is a constant $\gamma > 0$ such that

$$|Q(\xi)| \leq e^{-\gamma|\xi|^{2s}} \qquad \text{for } |\xi| \leq \pi \tag{4}$$

Note that necessarily by (3), $\operatorname{Im}\beta = 0$ if $2s \geq p+1$.

Various authors have found pointwise or L_2 estimates for the error $v(x_\nu,t) - u(x_\nu,t)$ in the region near the propagating discontinuity: Hedstrom

[1968] (cf. Thomee [1969]) gives pointwise estimates for the case ρ = const. Apelkrans [1968] and later Brenner and Thomee [1971] and Thomee [1971] applied a technique due to Kreiss and Lundqvist [1968] to obtain pointwise and L_2 error estimates for both the constant coefficient and variable coefficient case. We state below the L_2 result of Brenner and Thomee [1971].

Theorem 1 (Decay of Error away from a discontinuity) *Assume that* $\rho(x) > 0$ *and that the coefficient of the leading truncation error term* $\beta > 0$. *Let* δ *be the (signed) distance from the characteristic line given by* $dx/dt = \rho$. *Then to the right of this characteristic*

$$\| H(x-\delta-\rho t)v(x,nk) \|_2 \leq exp(-C_0 n(\delta/t)^{q_1}) \| v(x,0) \|_2$$

and to the left of the characteristic

$$\| H(x-\delta-\rho t)(v(x,nk)-1) \|_2 \leq exp(-C_0 n(|\delta|/t)^{q_2}) \| (v(x,0)-1) \|_2$$

where C_0 *is a positive constant and*

$$q_1 = (p+1)/p \qquad q_2 = 2s/p \qquad \text{if } p \text{ is even}$$

and

$$q_1 = q_2 = 2s/p \qquad \text{if } p \text{ is odd}$$

(If $\beta(x) < 0$ *then the exponents* q_j *reverse, i.e.* $q_1 = 2s/p$, $q_2 = (p+1)/p$ *if* p *is even.)*

Remark: $\beta < 0$ for typical even-order methods, although this is not necessarily the case. $\beta < 0$ means that the numerical phase velocity is less than the true phase velocity as $\xi \to 0$.

A proof of this theorem for the constant coefficient case is included in the appendix to this section. (A proof for the variable coefficient case can be found in Thomee [1971].)

The main conclusion we can draw from the theorem above is that the error will decay exponentially away from the propagating discontinuity if a dissipative method is used. This is an important result since a region of steep gradients or a discontinuity in a finite-difference solution is typically also a region of high error in that solution. This theorem says that this error will essentially remain localized.

For systems of hyperbolic equations, we can expect similar behavior. For example W. Gropp [1981] has considered solutions of finite difference approximations to the "telegraph equation"

$$\mathbf{u}_t + \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{u}_x + \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{u} \equiv 0 \tag{5}$$

where $\mathbf{u} = \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}$ and found similar error estimates for the region of the solution near the discontinuity by asymptotic evaluation of integral solutions of the finite difference equations. It is also expected, however, that more complicated behavior of the error will be exhibited due to coupling of the different components of the solution in a system of equations. Majda and Osher [1977] considered difference approximations for the following 2 by 2 hyperbolic system of equations:

$$\mathbf{u}_t + \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{u}_x + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \mathbf{u} = 0 \tag{6}$$

on $-\pi \leq x \leq \pi$, with initial conditions $\mathbf{u}(x,0) = \begin{bmatrix} H(x) \\ 0 \end{bmatrix}$ and proved the following

Theorem 2 (Numerical Artifact for Coupled Hyperbolic Systems) *Let* $\mathbf{v}$ *be the solution of a* p*th order accurate dissipative approximation to* (6) *which satisfies a (technical) ellipticity condition* (see Majda and Osher [1977]) *and let* R_δ *be the region in the* $x-t$ *plane between and bounded away from the characteristic lines* $|x| = |t|$*:*

$$R_\delta \equiv \left\{ (x,t) \mid \frac{|x-t|}{t} > \delta, \ \frac{|x+t|}{t} > \delta, \ |x| < t, \ \delta \leq t \leq T_0 \right\}$$

then

$$\max_{(x,t) \in R_\delta} |\mathbf{v} - \mathbf{u}| \leq C_\delta h^2$$

where C_δ *is a constant depending on* δ *and* h *is the meshwidth.*

This $O(h^2)$ error is actually present in the initial data and represents the local error at the discontinuity which results from approximating that discontinuous function on a finite grid. For a dissipative approximation to a single scalar hyperbolic equation, the effect of such an error will remain localized near the characteristic emanating from the location of the discontinuity in the data (invoke theorem 1, recalling that if we could solve the differential equation

exactly, this error would just propagate along this characteristic). For a coupled system of hyperbolic equations such as (6), however, the error in $u^{(1)}$ can be expected to influence the other component $u^{(2)}$ due to the coupling of the components by the lower order term, and we expect the region between the two characteristics emanating from the origin to be polluted by this error. The following simplified example will demonstrate this:

Consider, instead of (6), the telegraph equation (5). Clearly by a change of independent variables this can be transformed to

$$\mathbf{u}_t + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{u}_x + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \mathbf{u} = 0 \tag{7}$$

which we consider on the interval $-\pi \le x \le \pi$ with discontinuous initial data

$$\mathbf{u}(x,0) = \begin{bmatrix} S(x) \\ 0 \end{bmatrix} \tag{8}$$

where $S(x) = -x/\pi + sign(x)$ is the "sawtooth function". (Note that $S(-\pi) = S(\pi)$). We now introduce a uniform finite mesh defined by the $2N$ meshpoints $x_\nu = \nu h$, $\nu = -N+1, -N+2, \cdots, N$ with $h = \pi/N$. On such a mesh we can at best represent the function $S(x)$ with a finite Fourier series

$$\tilde{S}(x) = \sum_{l=-N+1}^{N} \tilde{s}(l) e^{ilx} \tag{9}$$

that has the interpolation property

$$\tilde{S}(x_\nu) = S(x_\nu) \quad , \nu = -N+1, \cdots, N \tag{10}$$

The continuous function $S(x)$ can be represented by the usual (infinite) Fourier series expansion given by

$$S(x) = \sum_{l=-\infty}^{\infty} \hat{s}(l) e^{ilx} \tag{11}$$

where

$$\hat{s}(l) = \begin{cases} 0 & l = 0 \\ \dfrac{(-1)^l i}{l} & l \neq 0 \end{cases}$$

By the interpolation property of $\tilde{S}(x)$, we have therefore an explicit formula for

$$\tilde{s}(l) = \sum_{\mu=-\infty}^{\infty} \hat{s}(l+2N\mu)$$

and hence for the error $e(x) \equiv \tilde{S}(x) - S(x) = \sum_{l=-N}^{N} \tilde{e}(l) e^{ilx}$ where

$$\tilde{e}(l) = \sum_{\mu \neq 0} \hat{s}(l + 2N\mu)$$

A simple computation gives for the Fourier coefficents of the error the formula

$$\tilde{e}(l) = h^2 \sum_{\mu > 0} \frac{(-1)^l 2il}{l^2 h^2 - 4\pi^2 \mu^2} \qquad \text{for } |lh| \leq \pi$$

$$=: h^2 f(l)$$

By the usual technique, in order to understand the effects of this error, we can consider the solution of (7) with the initial data (8) replaced by the error expression, i.e. put

$$\mathbf{u}(x,0) = \begin{bmatrix} h^2 f_0(x) \end{bmatrix} \tag{14}$$

where $f(x) = \sum_{|l| \leq N} f(l) e^{ilx}$ By the method of characteristics we can construct the solution to (7), (14) explicitly. It is given (in the sense of distributions) by

$$\mathbf{u}(x,t) = \begin{cases} \begin{bmatrix} h^2 f(x) \\ -h^2 f(x+t) x \end{bmatrix} & \text{for} -t \leq x \leq 0 \text{ (between characteristics)} \\ 0 & \text{elsewhere} \end{cases} \tag{15}$$

Since we have neglected the errors due to approximating (7) by a difference scheme, equation (15) indicates the best we can expect to do for such a problem.

The theorem of Majda and Osher quoted above indicates that for dissipative methods, this example gives an accurate picture of how this numerical artifact will arise and affect the solution of the difference equations.

A contact discontinuity differs from a shock in that the characteristic lines near such a discontinuity are parallel to the surface of the discontinuity in $x-t$ space, and so it is essentially a linear phenomenon. Thus we expect the theorems quoted above to give a good idea of how to expect a difference scheme to behave near such a discontinuity. In contrast, a shock surface in $x-t$ space has the property that the characteristic lines point locally into that surface. We can consider the shock surface as an internal boundary and recall from the theory for hyperbolic equations that no boundary conditions need be specified

there to determine the solution of the differential equation because the characteristics point *out* of the region of interest i.e towards the shock surface. Difference approximations, however, typically require extra "numerical" boundary conditions at such boundaries and thus unlike the continuous case, values of the solution at these boundaries can influence the solution in the interior of the region. Since the discontinuity in the solution at the shock is likely to be a region of high error in the numerical solution, we can therefore expect this error to influence the solution away from the shock. Kreiss and Lundquist [1968, Theorems 1 and 5] have shown that for the class of "contractive" difference operators (see section 1.3A) the influence of the numerical boundary conditions decays away from the boundary. The interval of influence is of length $0(|h|\log h)$. The class of contractive operators includes all dissipative approximations.

We conclude this section by remarking (as do Kriess and Lundqvist [1968, p.11] that for methods which do not require numerical boundary conditions at boundaries where the characteristics point outwards, the error at those boundaries cannot affect the interior solution. This is the motivation for the class of "upwind" difference schemes discussed in the next section. Such schemes have the (often considered desirable) property that oscillations in the numerical solution are not produced near a shock. However, they are also very dissipative and tend to produce inaccurate solutions in the region away from the shock.

1.3A Proof of Theorem 1.3.1 for the case $\rho = const.$

Since some of the details of the proofs for the estimates of the type given by Brenner and Thomee [1971],Thomee [1971] and Apelkrans [1968] are less than lucid, I am including the following proof, essentially due to Thomee, which may or may not clear up some confusion.

The proof depends on the notion of the *contractivity* of a difference approximation. (Apelkrans [1968], Brenner and Thomee [1971])

Definition 1: *The difference approximation (2) is said to be* **contractive of order** τ *if the following estimate holds for the symbol* $Q(\xi)$*:*

$$|Q(\xi - i\eta)| \leq \exp(\rho\lambda\eta)\exp(C|\eta|^{\tau}) \qquad (16)$$

where $|\eta| \le \eta_0 |$ *and* ξ *are real.*

We will first prove the following

Lemma 1 (Contractivity) *Let the difference approximation (2) be accurate of order* p *(3) and dissipative of order* $2s$. *Then*

$$\tau = \frac{2s}{2s-p} \tag{17a}$$

unless p *is even and* $\beta\rho\eta < 0$, *in which case*

$$\tau = p + 1 \tag{17b}$$

(β *is defined by (3)).*

Proof: We first recall the following variant of Holder's inequality: *For any* $a,b > 0$ *and* $\frac{1}{q} + \frac{1}{p} = 1$, *there exist constants* ε *and* δ_ε *such that the inequality*

$$a^{\frac{1}{q}} b^{\frac{1}{p}} \le \varepsilon a + \delta_\varepsilon b \tag{18}$$

holds. From this inequality it follows that for $\eta, \xi > 0$ we can always estimate $\eta^{n-j}\xi^j$ in terms of η^n and ξ^n, i.e.

$$\eta^{n-j}\xi^j \le C_j(\eta^n + \xi^n), \qquad j=1,2,\ldots,n-1$$

Now rewrite (3) as

$$Q(\xi) = \exp(i\lambda\rho\xi + \psi(\xi))$$

where $\psi(\xi) := i\lambda\rho\beta\xi^{p+1}(1 + o(1))$. Then applying (18),

$$\mathrm{Re}(\psi(\xi - i\eta) - \psi(\xi)) = (i\beta\rho\lambda(\xi - i\eta)^{p+1} - i\beta\rho\lambda\xi^{p+1})(1 + o(1))$$

$$\le C\eta(|\xi|^p + |\eta|^p) \quad \text{for } |\eta|, |\xi| \text{ small enough}$$

Using (4) we have therefore that

$$\mathrm{Re}\psi(\xi - i\eta) \le -\gamma|\xi|^{2s} + C|\eta|(|\xi|^p + |\eta|^p) \tag{19}$$

Again apply (18) to estimate $|\eta||\xi|^p$ in terms of $|\xi|^{2s}$ and $|\eta|^\tau$, where τ is determined by the condition (from Holder's inequality) $\frac{1}{\tau} + \frac{p}{2s} = 1$, i.e. $\tau = 2s/(2s-p)$. The constants in the estimate can be adjusted so that we obtain

$$\mathrm{Re}\psi(\xi - i\eta) \le C|\eta|^{2s/(2s-p)}$$

In the special case that p is even, note that we can write

$$\mathrm{Re}(i\beta\rho(\xi + i\eta)^{p+1}) \;=\; \mathrm{Re}\; i\beta\rho\sum_{j=0}^{p+1} d_j \xi^{p+1-j}(-i\eta)^j, \qquad d_j > 0$$

and so if $\mathrm{Im}\beta = 0$,

$$\mathrm{Re}(i\beta\rho(\xi + i\eta)^{p+1}) = -\beta\rho \sum_{j\ odd} i^{j+1} d_j \xi^{p+1-j}\eta^j$$

$$\leq -d_1\beta\rho\xi^p\eta + C(\eta^2|\xi|^{p-1} + \eta^{p+1})$$

Clearly if $\beta\rho\eta < 0$ we can then take $\tau = p+1$ in the estimate (16).

Proof of Theorem 1: To get the estimate for the righthand side of the discontinuity introduce the scaled variable w defined by $v = e^{\eta x}w$, $\eta > 0$. Then w satisfies

$$w(x,t+k) \;=\; \mathbf{Q}'w(x,t)$$

where $\mathbf{Q}'z(x) := e^{\eta x}\mathbf{Q}(e^{-\eta x}z(x))$. Then

$$v(x,t) \;=\; e^{-\eta x}(\mathbf{Q}')^n(e^{\eta x}v(x,0))$$

so

$$\|H(x-\delta+\rho t)v(x,t)\|_2 \;=\; \|H(x-\delta+\rho t)e^{-\eta x}(\mathbf{Q}')^n e^{\eta x}v(x,0)\|_2$$

$$\leq e^{-\eta(\delta-\rho t)}\|\mathbf{Q}^n\|_2\|e^{\eta x}v(x,0)\|_2 \qquad (20)$$

Now from lemma 1, the estimate

$$|\mathbf{Q}'(\xi)| \;=\; |\mathbf{Q}(\xi - i\eta h)| \leq e^{-\rho\eta\lambda}e^{C|\eta h|^\tau}$$

holds with τ given by formulas (17) and so using the fact that $\eta x < 0$ on the support of $v(x,0)$, the right-hand side of (20) can be estimated, giving

$$\|H(x-\delta+\rho t)v(x,t)\|_2 \;\leq\; e^{-\eta\delta}e^{Cn|\eta h|^\tau}\|v(x,0)\|_2 \qquad (21)$$

Now choose $\eta > 0$ in such a way that the exponential on the right-hand side of (21) decays: Take $Cn|\eta h|^\tau = \eta\delta/2$, in which case (21) becomes

$$\|H(x-\delta+\rho t)v(x,t)\|_2 \;\leq\; e^{-C_0 n(\delta/t)^{\tau/\tau-1}}\|v(x,0)\| \qquad (22)$$

where C_0 is a constant that depends on the mesh ration λ. To obtain the corresponding estimate on the left-hand side of the discontinuity, let

$$y(x,t) = 1 - v(x,t)$$

and note that by consistency $\sum_j a_j = 1$, so $y(x,t)$ solves the problem

$$y(x,t+k) = Qy(x,t) \tag{23}$$

with $y(x,0) = H(x)$ as initial data. Proceed with the same analysis as above but for $\delta < 0$ choose $\eta < 0$ in order to get decay in the exponential. Then replacing τ in the resulting estimate and in (22) we get the estimate of the theorem.

1.4 Difference Methods for Conservation Laws

In this section we give a brief discussion of modern difference methods for solving systems of conservation laws. This is of interest in the context of numerical methods for singular perturbation problems because in both cases one of the objectives of a good method is to be able to resolve discontinuities or rapid transitions in the solutions well and at minimum expense. Also one of the predominant philosophies behind the schemes for both types of problems is the same: that "one-sided" or "upwind" type schemes tend to accomplish this objective the best. Since it is not the purpose of this thesis to discuss numerical methods for conservation laws in general, this section is more of a list rather than a review of those methods. The reader will find a much more adequate discussion of such difference methods in the review papers of Sod [1977] and of Harten, Lax and van Leer [1981].

We begin by quoting some theorems about difference schemes for conservation laws. Consider the scalar conservation law given by

$$u_t + f(u)_x = 0 \tag{1}$$

on $-\infty < x < \infty$ and specify initial data $u(x,0) = \varphi(x)$. For the numerical method we introduce a mesh $\{x_\nu\}_{-\infty}^{\infty}$ and approximate (1) on this mesh by an explicit difference scheme in *conservation form:*

$$v_\nu(t+k) = v_\nu(t) - kD_+ f_{\nu-\frac{1}{2}}(t) \tag{2}$$

$$v_\nu(0) = \varphi(x_\nu)$$

where $v_\nu(t)$ is an approximation to $u(x_\nu,t)$, k is the time step, $D_+ w_\nu := (w_{\nu+1} - w_\nu)/(x_{\nu+1} - x_\nu)$ and $f_{\nu+\frac{1}{2}} := f(v_{\nu-l+1}, v_{\nu-l+2}, \ldots, v_{\nu+l})$ where the *numerical flux* $f(\,,\cdots,\,)$ is a function of $2l$ arguments and is consistent

with the flux function $f(u)$ in the differential equation (1), i.e.

$$f(u,u,\ldots,u) = f(u).$$

Lax and Wendroff [1960] proved the following important theorem about the solutions of (2):

Theorem 1 (Lax-Wendroff): *If as k and h_ν tend to zero $v_\nu(t)$ converges boundedly almost everywhere to some function $w(x,t)$, then $w(x,t)$ is a weak solution of (1) with initial values $\varphi(x)$. Equivalently, any discontinuity of $w(x,t)$ satisfies the Rankine-Hugoniot jump conditions (1.2.14).*

It is important to the proof of this theorem that the difference approximation be written in conservation form as in (2). If a difference approximation is used that does not have this form then it is possible, and indeed often happens, that shocks that are formed in the solution will not satisfy the Rankine-Hugoniot conditions and hence will travel at the wrong speed. Conservation form is not sufficient, however, to assure that the discrete solutions will converge to the correct (physical) weak solution because the solutions of a difference equation in conservation form do not necessarily satisfy an entropy condition in the limit. A class of difference schemes that does give the correct solutions is the class of *monotone* difference schemes. Let $H(v_{\nu-l},v_{\nu-l+1},\ldots,v_{\nu+l}) := v_\nu - kD_+f_{\nu-l}$. Then

Definition: *The difference approximation (2) is said to be* **monotone** *if H is a monotone increasing function of each of its arguments.*

Harten, Hyman and Lax [1976] proved the following theorem about monotone schemes:

Theorem 2 (Harten-Hyman-Lax): *Suppose that the difference approximation (2) is monotone, i.e.*

$$\frac{\partial H}{\partial w_i}(w_{-l},\ldots,w_l) \geq 0 \quad \text{for all } -l \leq i \leq l$$

and let $w(x,t)$ be the limiting solution of (2) given in theorem 1, then $w(x,t)$ satisfies the entropy condition (2.2.15).

From the discussion in section 2.2 this result means that a monotone scheme will always give solution with the correct physical behavior in the limit. The same authors also, however, proved the following more depressing result:

Lemma 1 (Harten-Hyman-Lax [1976]): *Monotone finite-difference schemes in conservation form are at most of first order accuracy.*

From a practical point of view this means that the solutions of monotone schemes tend to have shocks and contact discontinuities that are "smeared" out, i.e. the transition from the left to the right states of the shock may occur over many meshpoints in the discrete solution. Also, if a monotone method is used everywhere on the mesh, the smooth parts of the solutions will not be computed accurately.

Monotone schemes have the sometimes desirable property that they do not produce highly oscillatory errors near discontinuities in the solutions. This is a result of the following

Theorem 3 (Harten [1977]): *A montone scheme is* **monotonicity-preserving** *i.e. if (2) is a monotone scheme, then if* $\Delta_+ v_\nu(t)$ *is of one sign, then so is* $\Delta_+ v_\nu(t+k)$.

The converse of theorem 3 is not true, however. Thus, a scheme that is monotonicity-preserving will not necessarily give limiting solutions that satisfy an entropy condition.

We now discuss some particular difference methods. Some of these methods fit into the framework developed above, others do not.

One of the earliest developed methods that is still discussed today is Godunov's method (Godunov [1959]): Consider the initial-value problem for (1) on $-\infty < x < \infty$ with piecewise constant initial data

$$u(x,0) = \begin{cases} u_L & \text{if } x < 0 \\ u_R & \text{if } x > 0 \end{cases} \tag{3}$$

This is known as the **Riemann problem** for the conservation law (1), and can often be solved exactly for systems of conservation laws of interest. Now consider the original problem (1) with general initial data $u(x,0) = \varphi(x)$. We introduce a mesh $\{x_\nu\}_{-\infty}^{\infty}$ which for convenience is assumed to be uniform ($x_{\nu+1} - x_\nu \equiv h$ = const.) and replace the initial data with a piecewise constant approximation $\tilde{u}(x,0) = \tilde{\varphi}(x)$ where

$$\tilde{\varphi}(x) = \varphi(x_\nu) \quad \text{if } x \in (x_\nu - h/2, x_\nu + h/2).$$

If we take the time step k to be small enough, this modified initial value problem for (1) can be solved exactly if the general solution of the Riemann problem is known, because in each interval $(x_\nu - h/2, x_\nu + h/2)$, we have a Riemann problem and for small enough times, the neighboring Riemann problems will not interact. Godunov's method is to do exactly that at each time step and then

average the resulting solution over each interval $(x_\nu - h/2, x_\nu + h/2)$ to get the approximate solution value $u(x_\nu, t+k)$.

There are several other methods that can be considered as modifications of the Godunov scheme. Glimm's scheme (Chorin [1976]) is similar to Godunov's scheme in that local Riemann problems are solved exactly at each time step, but in each subinterval $x_\nu - h/2, x_\nu + h/2)$ the solution at the new time level is taken to be the value of the exact solution at a random point in that subinterval. One of the most expensive parts of Godunov's or Glimm's schemes can be the exact solution of the Riemann problems. Harten and Lax [1981] have pointed out that much of the information in the exact solution of the local Riemann problems is lost when the solution is averaged or sampled and have accordingly proposed a random-choice scheme in which the local Riemann problems are only solved approximately.

The concept of "upwind" or one-sided differencing is also an old idea (see e.g. Courant, Isaacson and Rees [1952]). It has already been mentioned in section 1.3 where we pointed out that oscillatory error can be avoided near a discontinuity in the numerical solution if the differential equations are approximated in such a way that information can only travel in the true characteristic direction on the mesh. For the moment, take $f(u) = u^2/2$ in equation (1) in which case we have the inviscid form of Burgers' equation. The variable u can be identified as a fluid velocity, for example, in this equation. The upwind scheme is then given by

$$v_\nu(t+k) = \begin{cases} v_\nu(t) - kD_+ f_\nu(t) & \text{if } f'(u_\nu) < 0 \\ v_\nu(t) - kD_- f_\nu(t) & \text{if } f'(u_\nu) > 0 \end{cases} \tag{4}$$

where $f_\nu(t) := f(v_\nu(t))$. Since $f'(u) = u$, we see that the divided difference in x is always taken in the direction of larger velocity, hence the name "upwind". Equation (4) is a monotone scheme under the condition $k \max_\nu |u_\nu| / (x_{\nu+1} - x_\nu) < 1$ (the Courant condition). There are a variety of methods which have been proposed that in one way or another are variations on the upstream differencing concept. Roe's method (Roe [1981]), Steger and Warming's "flux vector splitting" (Steger and Warming [1981]), and Engquist and Osher's methods (Engquist and Osher [1979], Osher [1980]) are some examples. The differences in these methods come mostly in how they generalize to *systems* of conservation laws, and in how the differencing is done at a "sonic" or "turning" point (where $f'(u) = 0$). Engquist and Osher's first and second order methods for scalar

conservation laws are interesting from a theoretical (as well as practical) point of view because they are derived in such a way that the limiting solutions satisfy an entropy inequality (Engquist and Osher [1979]).

Since shocks and contact discontinuities are typically found only in isolated parts of a flow field, there are many methods in which an attempt is made to compute the remainder of the flow field will by using higher-order difference methods and then to use a special mechanism to get good resolution of the discontinuities in the solution without polluting the smooth part of the solution. For example, Hyman [1979], in his "method-of-lines approach to the numerical solution of conservation laws" advocates the use of higher-order centered spatial differencing together with a higher order predictor-corrector version of the "leap-frog" method in time to resolve the smooth parts of the solution and incorporates a nonlinear artificial dissipation to reduce the rapidly oscillating error associated with discontinuities in the solution. Harten's "artificial compression" method (Harten [1977],[1978]) is to use an existing scheme but to modify the flux function $f(u)$ in such a way that the characteristics point more strongly towards a shock surface. This tends to impede the propagation of error away from a discontinuity and hence gives sharper computed discontinuities. It also has the effect of turning contact discontinuities into shocks since the characteristics point towards the contact surface rather than parallel to it after artificial compression. Since the entropy condition is essentially just a mathematical statement of the fact that characteristics must point into a shock, the artificial compression method preserves the entropy producing properties of the particular difference method it is used in conjunction with.

The "flux corrected transport" method of Boris and Book [1973] is another method that makes a special effort to give good shock resolution. The "antidiffusion" step in this method is a correction step in which an attempt is made to reduce the "smearing" of the shock in the numerical solution by putting in an artificial "negative" dissipation term to try to cancel out the effect of the standard artificial dissipation in the method.

For the purposes of computing numerical solutions to singular perturbation problems the most useful concepts of those introduced above are probably the idea of one-sided difference methods and the concept of "monotonicity-preserving" schemes. We will see these ideas applied to singular perturbation problems in chapters 2 and 3 of this thesis.

II. Solution Adaptive Mesh Refinement for Singular Perturbation Problems

2.0 Introduction

In this chapter we consider the numerical solution of a system of ordinary differential equations

$$\frac{d\mathbf{y}}{dx} = A(x)\,\mathbf{y}(x) + \mathbf{f}(x), \tag{1a}$$

for $0 \le x \le L$ with n linearly independent boundary conditions

$$R_0\mathbf{y}(0) + R_1\mathbf{y}(L) = \mathbf{g}. \tag{1b}$$

Here $\mathbf{y} := (y^{(1)}, y^{(2)}, \cdots, y^{(n)})^T$ and $\mathbf{f}(x)\ \varepsilon\ C^2$ are vector functions with n components and R_0, R_1 and $A(x) := (a_{ij}(x))\ \varepsilon\ C^2[0,L]$ are $n \times n$ matrices. We solve the system (1) using a difference approximation on a nonuniform mesh $\{x_\nu\}_0^N$, $x_0 := 0$, $x_N := L$, with local meshwidth $h_\nu := x_{\nu+1} - x_\nu$, $\nu = 0,1,\ldots,N-1$. The nonuniform mesh should be chosen in such a way that the solution of the differential equation is "resolved", i.e. that the error in the computed solution will be essentially uniformly distributed over the whole interval $0 \le x \le L$. In order to do this, we therefore need to be able to estimate the error in the computed solution.

There are two approaches for estimating the error. The first is to estimate it *a priori*. This means that we use information about the functions $A(x)$, $f(x)$ and the boundary conditions for the problem to estimate where regions of rapid change and hence of potentially large computational error can occur in the solution. The computational mesh is then chosen to be finest in the regions of high predicted error. This approach has been treated for example by Kreiss and Nichols [1975] and is not the one that we will take here. Instead we will estimate

the error *a posteriori*. This means that an initial calculation is done on some initial mesh (a uniform one, for example) and then the error in this computed solution is estimated and used to construct the desired non-uniform mesh. This last procedure is usually called "*solution-adaptive mesh refinement*" since the computational mesh is adapted to resolve features in a computed solution. Such methods have been discussed and applied before, for example in the work of Kreiss [1975] and Kreiss and Kreiss [1981] on singularly perturbed ordinary differential equations and by many other investigators for other applications.

A typical procedure for solution adaptive mesh refinement is illustrated as follows for the case $n = 1$: The system (1) becomes

$$y'(x) = a_{11}(x)y(x) + f(x), \qquad 0 \le x \le L \tag{2}$$

$$y(0) = y_o$$

where we assume that $a_{11}(x) = \mathrm{Re}\, a_{11}(x) \le -\tau \le 0$. Approximate (2) with (the Backward Euler Method),[1]

$$D_- u_{\nu+1} = a_{11}(x_{\nu+1})u_{\nu+1} + f(x_{\nu+1}) \qquad \nu = 0,1,\ldots,N-1 \tag{3}$$

$$u_o = y_o$$

and determine the solution on an initial mesh $\{x_\nu\}_0^N$. Here $u_\nu := u(x_\nu)$ is an approximation to $y(x_\nu)$. The solution of (3) can be written explicitly as

$$u_\nu = S_{\nu,0}u_o + \sum_{j=1}^{\nu} S_{\nu,j}h_{j-1}f_{j-1} \tag{4}$$

where the discrete solution operator $S_{\nu,\mu}$ is defined by

$$S_{\nu,\mu} := \prod_{j=\mu}^{\nu-1} \left(\frac{1}{1 - h_j a_{11}(x_j)}\right) \tag{5}$$

In this chapter we use the following notation: undivided differences are denoted by

$$\Delta_\pm u_\nu := \pm(u_{\nu\pm1} - u_\nu).$$

Single divided differences are defined by

$$D_\pm u_\nu := (u_{\nu\pm1} - u_\nu)/(x_{\nu\pm1} - x_\nu), \quad D_0 u_\nu := (u_{\nu+1} - u_{\nu-1})/2(x_{\nu+1} - x_{\nu-1})$$

For compactness of notation higher order divided differences are denoted by (nonstandard notation):

$$D_+^j u_\nu := j!u[x_\nu, x_{\nu+1}, \ldots, x_{\nu+j}]$$

where

$u[x_\nu, x_{\nu+1}, \cdots, x_{\nu+j}] := (u[x_{\nu+1}, x_{\nu+2}, \cdots, x_{\nu+j}] - u[x_\nu, x_{\nu+1}, \cdots x_{\nu+j-1}])/(x_{\nu+j} - x_\nu)$; $u[x_\nu] := u_\nu$

is the conventional notation and definition for the jth divided difference of u_ν at the points $x_\nu, \ldots, x_{\nu+j}$. Note that on a uniform mesh, this definition reduces to the usual one for $D_+^j u_\nu$.

Note that $S_{\nu,\mu} = (h_{\mu-1}a_{11}(x_{\mu-1}))^{-1}\Delta_- S_{\nu,\mu}$ so (4) can be rewritten as

$$u_\nu = S_{\nu,0}u_\nu + \sum_{j=1}^{\nu} \frac{f_{j-1}}{a_{11}(x_{j-1})}\Delta_- S_{\nu,\mu}$$

so since $|S_{\nu,\mu}| \le 1$, it follows that the estimate

$$|u_\nu| \le |u_0| + \max_{0 \le j \le \nu-1} |f_j / a_{11}(x_j)| \tag{6}$$

holds. The error, e_ν, in the computed solution is defined by $e_\nu := u_\nu - y(x_\nu)$ and hence satisfies the equation

$$D_- e_{\nu+1} = a_{11}(x_{\nu+1})e_{\nu+1} + \delta_{\nu+1}, \qquad e_0 = 0 \tag{7}$$

where the truncation error $\delta_{\nu+1}$ satisfies the following inequality:

$$|\delta_{\nu+1}| \le h_\nu \sup_{x_\nu \le \xi_{\nu+1} \le x_{\nu+1}} |y''(\xi_{\nu+1})|. \tag{8}$$

Hence using (6) we have an estimate for the error:

$$|e_\nu| \le \max_{0 \le j \le \nu-1} \left| \frac{h_j \sup_{x_{j-1} \le \xi_j \le x_j} |y''(\xi_j)|}{a_{11}(x_j)} \right| \tag{9}$$

The standard assumption at this point in the solution adaptive mesh selection procedure is that the truncation error δ_ν, which depends in this case on the second derivative of the true solution $y''(x)$, can be estimated by replacing $y''(\xi_\nu)$ with a divided difference of the calculated solution, for example $D_- D_+ u_\nu$, possibly multiplied by a constant of moderate size, C_1. Then (9) would be replaced with

$$|e_\nu| \le \max_j \left| \frac{h_j C_1 |D_+ D_- u_j|}{a_{11}(x_j)} \right| \tag{10}$$

The new mesh is then constructed by choosing the local meshwidths h_j to be small in regions where $|D_+ D_- u_j|$ is large, and the computation of u_ν is repeated using the new mesh. Presumably, after repeating this procedure a number of times, the mesh will converge so that the error is suitably small everywhere on the mesh after the final computation.

The procedure outlined above is essentially correct, but the justification is certainly not rigorous. In particular the justification for the essential step, which is to replace the error estimate (9) with (10) is not at all clear. In the rest

of this chapter, therefore, we will consider difference approximations to the system (1) and find conditions on the system and the difference approximations such that the error in the computed solution can be estimated in terms of lower-order divided differences of the computed solution.

2.1 A Posteriori Error Estimates for Diagonally Dominant Systems

A class of problems for which the desired adaptive mesh refinement theory can be developed is the class of *diagonally dominant* systems:

Definition: *The system (2.0.1a) is said to be* **diagonally dominant** *if there are constants* $0 < \delta < 1$ *and* ρ, *with* ρ *of moderate size, such that the elements* $a_{ij}(x)$ *of* $A(x)$ *satisfy*

$$|\mathrm{Im}a_{ii}| \le \rho |\mathrm{Re}a_{ii}|, \quad i = 1,2,\ldots,n \tag{1a}$$

and

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \le (1-\delta)|\mathrm{Re}a_{ii}|, \quad i = 1,2,\ldots,n. \tag{1b}$$

Assume now that the diagonal elements can be divided into groups such that $a_{ii} \le -K < 0$ for $i = 1,2,\ldots,r$ and $a_{ii} \ge K$ for $i = r+1,\ldots,n$. Furthermore, replace the boundary conditions (2.0.1b) with

$$\mathbf{y}^I(0) = \mathbf{y}_o^I \quad \textit{and} \quad \mathbf{y}^{II}(L) = \mathbf{y}_1^{II} \tag{2}$$

where $\mathbf{y}^I := (y^{(1)},y^{(2)},\ldots,y^{(r)})^T$ and $\mathbf{y}^{II} := (y^{(r+1)},\ldots,y^{(n)})^T$. Now introduce a mesh $\{x_\nu\}_0^N$ with $x_o = 0$, $x_N = L$, local meshwidth $h_\nu := x_{\nu+1} - x_\nu$ and approximate the system (2.0.1a),(2) with the weighted one-sided difference approximation of Kreiss and Kreiss [1981]:

$$D_+\mathbf{u}_\nu = (I - \Psi_\nu)(A_{\nu+1}\mathbf{u}_{\nu+1} + \mathbf{f}_{\nu+1}) + \Psi_\nu(A_\nu\mathbf{u}_\nu + \mathbf{f}_\nu) \tag{3}$$

$$\mathbf{u}_o^I = \mathbf{y}_o^I, \quad \mathbf{u}_1^{II} = \mathbf{y}_1^{II}$$

where $A_\nu := A(x_\nu)$, $\mathbf{f}_\nu := \mathbf{f}(x_\nu)$ and Ψ_ν is an $n \times n$ matrix given by

$$\Psi_\nu := \begin{pmatrix} \alpha_\nu^{(1)} & 0 & \cdots & 0 \\ 0 & \alpha_\nu^{(2)} & & \vdots \\ \vdots & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 0 \quad \alpha_\nu^{(n)} \end{pmatrix}$$

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

I. If $h_\nu \operatorname{Re} a_{ii}(x_\nu) \le 0$ **and** $h_\nu \operatorname{Re} a_{ii}(x_{\nu+1}) \le 0$, **then**

$$\alpha_\nu^{(i)} = \begin{cases} \frac{1}{2} & \text{if} \quad |h_\nu \operatorname{Re} a_{ii}(x_\nu)| \le 1 \\ |2h_\nu \operatorname{Re} a_{ii}(x_\nu)|^{-1} & \text{if} \quad |h_\nu \operatorname{Re} a_{ii}(x_\nu)| > 1 \end{cases}$$

II. If $h_\nu \operatorname{Re} a_{ii}(x_\nu) \ge 0$ **and** $h_\nu \operatorname{Re} a_{ii}(x_{\nu+1}) \ge 0$, **then**

$$\alpha_\nu^{(i)} = \begin{cases} \frac{1}{2} & \text{if} \quad h_\nu \operatorname{Re} a_{ii}(x_{\nu+1}) \le 1 \\ 1-(2h_\nu \operatorname{Re} a_{ii}(x_{\nu+1}))^{-1} & \text{if} \quad h_\nu \operatorname{Re} a_{ii}(x_{\nu+1}) > 1 \end{cases}$$

III. If $h_\nu \operatorname{Re} a_{ii}(x_\nu) \ge 0$ **and** $h_\nu \operatorname{Re} a_{ii}(x_{\nu+1}) < 0$ **then** $\alpha_\nu^{(i)} = \frac{1}{2}$

IV. If $h_\nu \operatorname{Re} a_{ii}(x_\nu) < 0$ **and** $h_\nu \operatorname{Re} a_{ii}(x_{\nu+1}) > 0$ **then**
introduce a new mesh point x_ν^* with $x_\nu \le x_\nu^* \le x_{\nu+1}$ such that $h_\nu \operatorname{Re} a_{ii}(x_\nu^*) = 0$, and apply either I or II.

(Remark: Condition IV is necessary to assure stability of the method, and is simple to implement if mesh refinement is being done in the calculation anyway.)

Let $\Delta := \max_{[0,N]} h_\nu^2 \sum_{j=0}^{2} \max_{[-4,5-k]} |D_+^k \mathbf{u}_\nu|$ [1]. Note that Δ is a linear combination of lower order divided differences of the computed solution $\mathbf{u}_\nu$ and will certainly be small if, for example, we assure that the expression

$$\mathcal{T}_\nu := h_\nu^2 \, (D_+ D_- \mathbf{u}_\nu + D_0 \mathbf{u}_\nu + \mathbf{u}_\nu)$$

is bounded and small for all ν. Then we have the following

Main Result: *If the system (2.0.1a),(2) is diagonally dominant and* $\|A(x)\|_\infty + \|A'(x)\|_\infty + \|A''(x)\|_\infty$ *and* $\|f''(x)\|_\infty$ *are bounded then the error in the computed solution* $\mathbf{e}_\nu := \mathbf{u}_\nu - \mathbf{y}(x_\nu)$ *can be estimated in terms of* Δ.

The proof of this result can be outlined as follows: The computed solution $\mathbf{u}_\nu$ is interpolated by a piecewise polynomial function $\varphi(x) \in C^3$, and a system of

---

[1] For convenience we will use the notation $\max_{[k,l]} f_\nu := \max_{\nu-k \le j \le \nu+l} f_j$ and $\sup_{(k,l)} f(x) := \sup_{x_k \le x \le x_l} f(x)$. Also $|\mathbf{f}(x)| := \max_i |f^{(i)}(x)|$ and $\|\mathbf{f}(x)\|_\infty := \sup_x |\mathbf{f}(x)|$. If A is a matrix with elements a_{ij}, then $\|A\|_\infty := \max_i \sum_j |a_{ij}|$.

differential equations is found for the error $\mathbf{e}(x) := \varphi(x) - \mathbf{y}(x)$. This system of equations is of the same form as (2.0.1a) but the forcing terms depend on derivatives of the interpolant $\varphi(x)$. Using an interpolation result of deBoor [1975],[1981], these derivatives $\varphi^{(j)}(x)$ can be estimated in terms of the corresponding divided differences $D_+^j \mathbf{u}_\nu$ of $\mathbf{u}_\nu$. The error $\mathbf{e}(x)$ can then be estimated in terms of these divided differences by using a well-posedness estimate for the system of differential equations.

Before proceeding with the proof we remark further that if the system of equations (2.0.1) is not in diagonally dominant form, this main result is clearly still valid as long as the system can be transformed to diagonally dominant form in a bounded way. For this reason, the next sections will be concerned with finding conditions on the system (2.0.1) under which a bounded transformation to diagonally dominant form exists.

Proof of the main result: We begin with an interpolation result of deBoor [1975],[1981]:

Lemma 1 (de Boor): *Let u_ν be a discrete function on a nonuniform mesh $\{x_\nu\}_0^N$ and s be a natural number. Then there is an interpolation function $\varphi_s(x)$ ε $C^s[x_0, x_N]$ and constants $K_{s,j}$, $j = 0,1,\ldots,s$ such that*

$$\varphi_s(x_\nu) = u_\nu \qquad \nu = 0,1,\ldots,N$$

and

$$\sup_{(\nu,\nu+1)} |d^j\varphi_j / dx^s| \le \frac{1}{2} K_{s,j}(J_{\nu,j} + \tilde{h}_\nu^{s-j} J_{\nu,s}), \quad j = 0,1,\ldots,s$$

where

$$J_{\nu,s} := \max_{[-k,0]} |D_+^k u_\nu|$$

and

$$\tilde{h}_\nu := \max_{[-s,s]} h_\nu, \quad h_\nu := x_{\nu+1} - x_\nu.$$

The most important part of this result is that the constants $K_{s,j}$ are of moderate size when s and j are small. For example, $K_{1,1} = 1$, $K_{2,2} \le 3.414$, $K_{3,3} \le 6.854$, $K_{4,4} \le 11.665$, and $K_{5,5} \le 21.036$. It is this ability to estimate derivatives in terms of divided differences that makes the adaptive mesh procedure work. (On the other hand it should be remarked that $K_{6,6} \ge 11.8$, $K_{7,7} \ge 18.5$, $K_{8,8} \ge 29.1$, $K_{9,9} \ge 45.7$ and $K_{10,10} \ge 71.8$ which means that this ability to estimate

derivatives in terms of divided differences deteriorates when the order gets large.)

If we assume that the computational mesh is somewhat smooth, i.e. that there is a constant c of moderate size such that

$$c^{-1} \le h_\nu / h_{\nu+1} \le c \qquad \nu = 0,1,\ldots,N-1 \tag{5}$$

then we can estimate $\tilde{h}_\nu^{s-j} J_{\nu,s}$ in terms of $J_{\nu,j}$ in (4) and so the following estimate for the derivatives of $\varphi(x)$ can be used instead:

$$\sup_{(\nu,\nu+1)} |d^j \varphi_s / dx^j| \le \frac{1}{2} \tilde{K}_{s,j} \max_{[-s,s-j]} |D_+^j u_\nu| \tag{6}$$

Consider now the solution $\mathbf{u}_\nu$ of the difference equations (3) and interpolate it with the piecewise polynomial $\varphi(x)$ ε C^3 of deBoor's construction. By Taylor expansion we can show that the vector function $\varphi(x)$ satisfies a "nearby" system of differential equations given by

$$d\varphi(x)/dx = A(x)\varphi(x) + \mathbf{f}(x) + \mathbf{r}(x) \tag{7}$$

The error in the interpolated difference solution $\mathbf{e}(x) = \varphi(x) - \mathbf{y}(x)$ therefore satisfies

$$d\mathbf{e}(x)/dx = A(x)\mathbf{e}(x) + \mathbf{r}(x), \quad 0 \le x \le L \tag{8}$$

with

$$\mathbf{e}^I(0) = 0, \qquad \mathbf{e}^{II}(L) = 0.$$

We now want to estimate $\mathbf{e}(x)$ in terms of $\mathbf{r}(x)$. To do this we can use the following

Lemma 2 (Estimates for diagonally dominant systems): *Consider the system (2.0.1a) on the interval $0 \le x \le L$ with boundary conditions*

$$\mathbf{y}^I(0) = \mathbf{y}_0^I \quad \text{and} \quad \mathbf{y}^{II}(L) = \mathbf{y}_1^I \tag{9}$$

If this system is diagonally dominant , $a_{ii} < 0$ for $i = 1,2,\ldots,r$, and $a_{ii} > 0$ for $i = r+1,\ldots,n$ then the following estimates hold:

$$|\mathbf{y}(x)| \le \frac{1}{\delta}\left\{ \| 2(\Lambda + \Lambda^*)^{-1}\mathbf{f}(x)\|_\infty + |\mathbf{y}_0^I| s_0(x) + |\mathbf{y}_1^{II}| s_1(x) \right\} \tag{10}$$

and

$$|\mathbf{y}(x)| \le \frac{1}{\delta}\left\{L\|\mathbf{f}(x)\|_\infty + |\mathbf{y}_o^I| s_o(x) + |\mathbf{y}_1^{II}| s_1(x)\right\} \qquad (11)$$

where

$$\Lambda := \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \vdots \\ \vdots & 0 & \ddots & \ddots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix},$$

$$s_o(x) := \exp(-\int_0^x \min_{i \le r} |\mathrm{Re} a_{ii}(\xi)|\, d\xi)$$

and

$$s_1(x) := \exp(\int_1^x \min_{i > r} |\mathrm{Re} a_{ii}(\xi)|\, d\xi).$$

Proof: Consider first the case $n = 1$ with $a_{11}(x) < 0$. Then the solution of (2.0.1a),(9) is given explicitly by

$$y(x) = \int_0^x e^{\int_\eta^x a_{11}(\xi)d\xi} f(\eta)d\eta \; + \; y(0)e^{\int_0^x a_{11}(\xi)d\xi}$$

Similarly if $a_{11}(x) > 0$,

$$y(x) = \int_1^x e^{\int_\eta^x a_{11}(\xi)d\xi} f(\eta) \; + \; y(1)e^{\int_1^x a_{11}(\xi)d\xi}.$$

So the solution for $n > 1$ can be written formally as

$$y^{(i)}(x) = \int_0^x e^{\int_\eta^x a_{ii}(\xi)d\xi} f^{(i)}(\eta)d\eta + \int_0^x e^{\int_\eta^x a_{ii}(\xi)d\xi} \sum_{i \ne j} a_{ij}(\eta)y^{(j)}(\eta)d\eta$$

$$+(y_o^I)^{(i)} e^{\int_0^x a_{ii}(\xi)d\xi}, \quad \text{for } i \le r \qquad (12)$$

and

$$y^{(i)}(x) = \int_1^x e^{\int_\eta^x a_{ii}(\xi)d\xi} f^{(i)}(\eta)d\eta + \int_1^x e^{\int_\eta^x a_{ii}(\xi)d\xi} \sum_{i \ne j} a_{ij}(\eta)y^{(j)}(\eta)d\eta$$

$$+(y_1^{II})^{(i)}\, e^{\int_1^x a_{ii}(\xi)d\xi}, \quad \text{for } i > r \tag{13}$$

Consider now equation (12). Note that for $i \le r$,

$$|e^{\int_\eta^x a_{ii}(\xi)d\xi}| \le e^{-\int_\eta^x |\mathrm{Re}a_{ii}(\xi)|d\xi} = |\mathrm{Re}a_{ii}(\eta)|^{-1}\frac{\partial}{\partial\eta}e^{-\int_\eta^x |\mathrm{Re}a_{ii}(\xi)|d\xi} \tag{14}$$

Using these inequalities in the first two integrals of (12), we obtain

$$|y^{(i)}(x)| \le \|f^{(i)}(x)/\mathrm{Re}a_{ii}(x)\|_\infty + \int_0^x \frac{\sum_{i\neq j}|a_{ij}(\eta)|}{|\mathrm{Re}a_{ii}(\eta)|}|y^{(j)}(\eta)|\frac{\partial}{\partial\eta}e^{-\int_\eta^x |\mathrm{Re}a_{ii}(\xi)|d\xi}d\eta$$

$$+ |(y_0^I)^{(i)}|e^{-\int_0^x |\mathrm{Re}a_{ii}(\xi)|d\xi} \qquad \text{for } i \le r. \tag{15}$$

An inequality similar to (15) holds for $i > r$ with the last term replaced by $|(y_1^{II})^{(i)}|\, e^{-\int_1^x |\mathrm{Re}a_{ii}(\xi)|d\xi}$. Suppose now that $|\mathbf{y}(x)| = |y^{(k)}(x)|$. Then using the assumption of diagonal dominance,

$$\|\mathbf{y}(x)\|_\infty \le \|2(\Lambda+\Lambda^*)^{-1}\mathbf{f}(x)\|_\infty + (1-\delta)\|\mathbf{y}(\eta)\|_\infty$$

$$+ |\mathbf{y}_0^I|e^{-\int_0^x \min_{i\le r}|\mathrm{Re}a_{ii}(\xi)|d\xi} + |\mathbf{y}_1^{II}|e^{\int_1^x \max_{i>r}|\mathrm{Re}a_{ii}(\xi)|d\xi}$$

from which the first estimate (10) follows. To get the second estimate (11) we use inequality (14) only in the second integral of (12), using in the first integral the obvious inequality

$$|\int_0^x e^{\int_\eta^x a_{ii}(\xi)d\xi} f^{(i)}(\eta)d\eta| \le x \sup_{0\le\eta\le x}|f^{(i)}(\eta)|$$

instead.

The function $\mathbf{r}(x)$, in turn, can be estimated in terms of the derivatives of the interpolant $\varphi(x)$ and the smoothness properties of $A(x)$ and $\mathbf{f}(x)$: For convenience define

$$\mathbf{F}(x) := A(x)\mathbf{y}(x) + \mathbf{f}(x)$$

The ith equation of the system (2.0.1a) can then be written

$$dy^{(i)}/dx = F^{(i)}(x) \tag{16}$$

Correspondingly, the ith equation in the difference approximation is

$$D_+u_\nu^{(i)} = (1 - \alpha_\nu^{(i)})F_{\nu+1}^{(i)} + \alpha_\nu^{(i)}F_\nu^{(i)} \tag{17}$$

where $\mathbf{F}_\nu := \sum_{j=1}^{n} a_{ij}(x_\nu)u_\nu^{(j)} + f^{(i)}(x_\nu)$. We will first estimate the components of $\mathbf{r}(x)$ at intermediate points in each interval $[x_\nu, x_{\nu+1}]$. Since for each component $r^{(i)}(x)$ the algebra is the same, we will drop the superscript (i) for convenience of notation. Thus in the following, $r(x) := r^{(i)}(x)$, $\alpha_\nu := \alpha_\nu^{(i)}$, $F_\nu := F_\nu^{(i)}$, etc. Since $\varphi(x)$ interpolates the discrete function $\mathbf{u}_\nu$ at the points x_ν, we can replace $\mathbf{u}_\nu$ with $\varphi(x_\nu)$ in the right-hand side of (17). If we choose the intermediate point $\xi_\nu := x_{\nu+1} - \alpha_\nu h_\nu$, where $h_\nu := x_{\nu+1} - x_\nu$, then it is easy to see by Taylor expansion that the right-hand side of (17) can be written as

$$(1 - \alpha_\nu)F_{\nu+1} + \alpha_\nu F_\nu = F(\xi_\nu) + h_\nu^2 G_\nu^{(i)}$$

where

$$|G_\nu^{(i)}| \le \text{const.} \sup_{(\nu,\nu+1)} |d^2 f^{(i)}(x)/dx^2 + \frac{d^2}{dx^2}\Big(\sum_j a_{ij}(x)\varphi^{(j)}(x)\Big)|. \tag{18a}$$

Similarly, the left-hand side of (17) can be expanded as follows:

$$D_+u_\nu = \varphi_x(\xi_\nu) + \psi_\nu h_\nu B_\nu^{(i)} + h_\nu^2 H_\nu^{(i)}$$

where

$$|H_\nu^{(i)}| \le \text{const.} \sup_{(\nu,\nu+1)} |d^3\varphi^{(i)}(x)/dx^3|, \tag{18b}$$

$$|B_\nu^{(i)}| \le \text{const.} \sup_{(\nu,\nu+1)} |d^2\varphi^{(i)}(x)/dx^2|, \tag{18c}$$

and

$$\psi_\nu := \begin{cases} 0 & \text{if } \alpha_\nu^{(i)} = 1/2 \\ 1 & \text{otherwise.} \end{cases} \tag{19}$$

Therefore by the definition of $\mathbf{r}(x)$ (equation (7)) we can estimate $r(x)$ at the points $x = \xi_\nu, \nu = 0,1,\ldots,N-1$ by using the identity

$$r(\xi_\nu) = -\psi_\nu h_\nu B_\nu^{(i)} + h_\nu^2(G_\nu^{(i)} - H_\nu^{(i)}) \tag{20}$$

together with the inequalities (18). Now that we have estimates for $r(x)$ at intermediate points in each subinterval $[x_\nu, x_{\nu+1}]$, we can estimate $r(x)$ for all values of x in that subinterval by using the fact that by Taylor expansion $r(x)$ can be

constructed to within $O(h_\nu^2)$ using the three nearest adjacent known values of $r(x)$, namely $r(\xi_{\nu-1})$, $r(\xi_\nu)$, $r(\xi_{\nu+1})$. More specifically, we can find functions $\beta_j(x)$, $j = -1,0,1$, such that

$$\sup_{(\nu,\nu+1)} |r(x) - \sum_{j=-1}^{1} \beta_j(x) r(\xi_{\nu+j})| \le \bar{h}_\nu^2 \text{const.} \sup_{(\nu-1,\nu+2)} |d^2 r(x)/dx^2|$$

$$\le \bar{h}_\nu^2 \text{const.}(\bar{G}_\nu^{(i)} + \bar{H}_\nu^{(i)}) \tag{21}$$

The last part of this inequality follows from differentiating (7) to get an expression for $d^2r(x)/dx^2$. Here also $\bar{h}_\nu := \max_{[-1,1]} h_\nu$ and for convenience we have defined

$$\bar{G}_\nu^{(i)} := \sup_{(\nu-1,\nu+2)} \left| \frac{d^2}{dx^2}\left(\sum_{j=1}^{n} a_{ij}(x)\varphi^{(j)}(x) + f^{(i)}(x)\right)\right| \tag{22a}$$

and

$$\bar{H}_\nu^{(i)} := \sup_{(\nu-1,\nu+2)} |d^3\varphi^{(i)}(x)/dx^3|. \tag{22b}$$

Now

$$\left| \sum_{j=-1}^{1} \beta^j(x) r(\xi_{\nu+j})\right| \le \text{const.} \max_{[-1,1]} |r(\xi_\nu)|$$

$$\le \text{const.}\dot{\psi}_\nu \bar{h}_\nu \bar{B}_\nu^{(i)} + \text{const.}\bar{h}_\nu^2(\bar{G}_\nu^{(i)} + \bar{H}_\nu^{(i)}) \tag{23}$$

where

$$\bar{B}_\nu^{(i)} := \sup_{(\nu-1,\nu+2)} |d^2\varphi/dx^2| \tag{22c}$$

so using the triangle inequality, (21) and (23) can be combined to give an estimate for $r(x)$:

$$\sup_{(\nu,\nu+1)} |r(x)| \le \text{const.}\dot{\psi}_\nu \bar{h}_\nu \bar{B}_\nu^{(i)} + \text{const.}\bar{h}_\nu^2(\bar{G}_\nu^{(i)} + \bar{H}_\nu^{(i)}) \tag{24}$$

We can now apply lemmas 1 and 2 and estimate the error $\mathbf{e}(x)$. The results of this estimate are summarized in the following

Lemma 3 (Error Estimate for Diagonally Dominant Systems): *If* $\inf_{(0,N)} \min_i |\text{Re}a_{ii}(x)| \ge 1$ *then*

$$|\mathbf{e}(x)| \le \frac{\text{const.}}{\delta} \max_\nu \tilde{h}_\nu^2 \left\{ [1 + M_\nu]\Delta_\nu + N_\nu \right\} \tag{25}$$

where

$$M_\nu := \sup_{(\nu,\nu+1)} |2(\Lambda + \Lambda^*)^{-1}| \sum_{j=0}^{2} \sup_{(\nu-4,\nu+2+j)} |d^j A(x)/dx^j|$$

$$N_\nu := \sup_{(\nu,\nu+1)} |2(\Lambda + \Lambda^*)^{-1}| \sup_{(\nu-4,\nu+4)} |\mathbf{f}''(x)|$$

and

$$\Delta_\nu := \sum_{j=0}^{2} \max_{[-4,5-j]} |D_+^j \mathbf{u}_\nu|$$

Proof: Let k be the index such that

$$\|2(\Lambda + \Lambda^*)^{-1}\delta(x)\|_\infty = \|\delta^{(k)}(x)/\ \mathrm{Re}a_{kk}(x)\|_\infty.$$

There are two cases to consider: If $h_\nu \inf_{(\nu,\nu+1)} |\mathrm{Re}a_{kk}(x)| \le 1$ for all ν, then $\alpha_\nu^{(k)} = 1/\ 2$ and $\psi_\nu^{(k)} \equiv 0$. So applying the estimate (10) to the error equation (8) and using the inequality (24) we have

$$|\mathbf{e}(x)| \le \max_\nu \left\{ \frac{\mathrm{const.}\bar{h}_\nu^2(\bar{G}_\nu^{(k)} + \bar{H}_\nu^{(k)})}{\inf_{(\nu,\nu+1)} |\mathrm{Re}a_{kk}(x)|} \right\}$$

Now, using the definition of $\bar{G}_\nu^{(k)}$, (22a) and applying the interpolation estimate (6), we have that

$$\bar{h}_\nu^2 \bar{G}_\nu^{(k)} / \inf|\mathrm{Re}a_{kk}| \le$$

$$\le \tilde{h}_\nu^2 \left\{ \left(\sum_{j=0}^{2} \sup_{(\nu-1,\nu+2)} |d^j A(x)/dx^j| \right) \cdot \left(\sum_{j=0}^{2} \max_{[-4,5-j]} |D_+^j u_\nu^{(k)}| \right) \right.$$

$$\left. + \sup_{(\nu-1,\nu+2)} |\mathbf{f}_{xx}(x)| \right\} \cdot \sup_{(\nu,\nu+1)} |2(\Lambda + \Lambda^*)^{-1}| \qquad (26)$$

Using the definition of $\bar{H}_\nu^{(k)}$, (22b) and (6) we have also that

$$\bar{h}_\nu^2 \bar{H}_\nu^{(k)} \le \tilde{h}_\nu^2 \mathrm{const.} \max_{[-4,2]} |D_+^3 u_\nu^{(k)}|.$$

So substituting in the difference equation (3) with $\alpha_\nu^{(k)} = 1/\ 2$,

$$\bar{h}_\nu^2 \bar{H}_\nu^{(k)} \le \tilde{h}_\nu^2 \mathrm{const.} \max_{[-4,2]} |D_+^2 \Big(\sum_j (a_{jk}(x_\nu) u_\nu^{(k)} + a_{jk}(x_{\nu+1}) u_{\nu+1}^{(k)} + f_\nu + f_{\nu+1}) \Big|$$

$$\le \tilde{h}^2 \mathrm{const.} \Big(\sum_{j=0}^{2} \sup_{(\nu-4,\nu+2+j)} |d^j A(x)/dx^j| \Big) \cdot \Big(\sum_{j=0}^{2} \max_{[-4,5-j]} |D_+^j \mathbf{u}_\nu| \Big) \quad (27)$$

where the identity $D_+(a_\nu b_\nu) = a_{\nu+1}D_+b_\nu + b_\nu D_+a_\nu$ was used to obtain the last inequality. The estimate (25) then clearly follows from the inequalities (26) and (27).

If, however $h_\nu \inf_{(\nu,\nu+1)} |\mathrm{Re}a_{kk}(x)| \geq \tau \geq 1$ for some ν, then $\psi_\nu = 1$ for some values of ν. Then combining (10), (8) and (24) as above we have

$$|\mathbf{e}(x)| \leq \max_\nu \left[\frac{\text{const.}\psi_\nu \bar{h}_\nu \bar{B}_\nu^{(k)} + \text{const.}\bar{h}_\nu^2(\bar{G}_\nu^{(k)} + \bar{H}_\nu^{(k)})}{\inf_{(\nu,\nu+1)} |\mathrm{Re}a_{kk}(x)|} \right]$$

Using the interpolation estimate (6), we have that

$$\bar{H}_\nu^{(k)} \leq \text{const.} \max_{[-4,2]} |D_+^3 u_\nu^{(k)}|.$$

Since by assumption the mesh is smooth (inequality (5)), the third difference of $u_\nu^{(k)}$ can be estimated in terms of second differences, i.e.,

$$\bar{H}_\nu^{(k)} \leq \text{const.}\tilde{h}_\nu^{-1} \max_{[-4,3]} |D_+^2 u_\nu^{(k)}|$$

Also by (6),

$$\bar{B}_\nu^{(k)} \leq \text{const.} \max_{[-4,3]} |D_+^2 u_\nu^{(k)}|.$$

So we have

$$\frac{\bar{h}_\nu \bar{B}_\nu^{(k)} + \bar{h}_\nu^2 \bar{H}_\nu^{(k)}}{\inf_{(\nu,\nu+1)} |\mathrm{Re}a_{kk}(x)|} \leq \frac{\text{const.}\,\tilde{h}_\nu \max_{[-4,3]} |D_+^2 u_\nu^{(k)}|}{\inf_{(\nu,\nu+1)} |\mathrm{Re}a_{kk}(x)|}$$

$$\leq \text{const.}\frac{\tilde{h}^2}{\tau} \max_{[-4,3]} |D_+^2 u_\nu^{(k)}| \qquad (28)$$

and hence recover the $O(\tilde{h}_\nu^2)$ behavior. The desired estimate (25) now clearly follows from (26) and (28) since $\tau^{-1} \leq 1$.

The proviso in the statement of the lemma that $\inf_x \min_i |\mathrm{Re}a_{ii}(x)|$ be ≥ 1 is included as a reminder that M_ν and N_ν will become unbounded if the diagonal elements of $A(x)$ get too close to zero, and the estimate (25) will no longer be useful. If all of the elements of $A(x)$ are small enough then the following lemma gives an estimate that can be used instead:

Lemma 4: *If* $\sup_{(0,N)} \max_i |\mathrm{Re}a_{ii}(x)| \leq 1$ *then*

$$|\mathbf{e}(x)| \leq \frac{\text{const.}}{\delta} L \max_\nu \tilde{h}_\nu^2 (\tilde{M}_\nu \Delta_\nu + \tilde{N}_\nu) \qquad (29)$$

where

$$\tilde{M}_\nu := \sum_{j=0}^{2} \sup_{(\nu-4,\nu+2+j)} |d^j A(x)/dx^j|$$

and

$$\tilde{N}_\nu := \sup_{(\nu-4,\nu+4)} |\mathbf{f}''(x)|$$

Proof: Clearly $h_\nu \sup_{(\nu,\nu+1)} |\mathrm{Re} a_{ii}(x)| \le 1$ for $\nu = 0,1,\ldots,N-1$ and $i = 1,2,\ldots,n$, so $\alpha_\nu^{(i)} = 1/2$ and so using (26) and (27) together with the estimate (11) the result follows immediately.

A possibly more typical case not covered directly by either of lemmas 3 or 4 is when some of the diagonal elements a_{ii} become quite small while others stay large. This case can be taken care of by introducing an exponential scaling of some of the variables $y^{(i)}$ in the differential equation (2.0.1a). For example, if $y^{(k)}(x)$ is the offending element of $\mathbf{y}(x)$, we replace it with $\tilde{y}^{(k)}(x) = e^{-\beta x} y^{(k)}(x)$. The kth equation then becomes

$$\frac{d\tilde{y}^{(k)}}{dx} = (a_{kk}(x) - \beta)\tilde{y}^{(k)} + \sum_{j \neq k} e^{-\beta x} a_{kj}(x) y^{(j)} + e^{-\beta x} f^{(k)} \qquad (30)$$

where the constant β is chosen so that $\inf_{(0,N)} |\mathrm{Re} a_{kk}(x) - \beta| \ge 1$ is satisfied. If the constant β is of moderate size and the length of the interval L is not too large, then this scaling will allow the error to be estimated in a useful way.

In order to keep the error estimates under control we can at least conceptually divide up the interval $[0,L]$ into subintervals each of which is short enough so that any exponential scaling that might be necessary will not destroy the error estimate on that subinterval. On each subinterval we have a two-point boundary value problem for the error $\mathbf{e}(x)$ with boundary conditions given by the condition that $\mathbf{e}$ must be continuous across the boundaries of the subinterval. If these boundary values are bounded, then since on each subinterval the system of equations is diagonally dominant, we can apply lemma 2 to obtain estimates similar to (25) for the error on each subinterval.

The question of whether the boundary values for each subinterval are actually bounded depends on being able to solve all the problems on the subintervals simultaneously. Suppose for the moment that there are M such subintervals $I_i := [s_{i-1}, s_i]$ where $0 = s_0 < s_1 < \cdots < s_m = L$ are the endpoints of the

subintervals. Denote by $\mathbf{e}_i(x)$ the error function on the ith subinterval, then (8), the original problem on the whole interval $[0,L]$, can be replaced with, for $i = 1,2,\ldots,M$,

$$\frac{d\mathbf{e}_i(x)}{dx} = A(x)\mathbf{e}_i(x) + \mathbf{r}(x) \qquad \text{on } s_{i-1} \le x \le s_i \tag{31}$$

with boundary values given by continuity with the adjacent intervals, i.e.

$$\mathbf{e}_i^I(s_{i-1}) = \mathbf{e}_{i-1}^I(s_{i-1}) \qquad \mathbf{e}_i^{II}(s_i) = \mathbf{e}_{i+1}^{II}(s_i)$$

where $\mathbf{e}_0^I(x_o) = \mathbf{e}_{M+1}^{II}(s_M) = 0$. On each subinterval, once we know the boundary values $\mathbf{e}_i^I(s_{i-1})$ and $\mathbf{e}_i^{II}(s_i)$, we can determine the solution everywhere on $[s_{i-1},s_i]$, and in particular we can determine the unknown values $\mathbf{e}_i^{II}(s_{i-1})$ and $\mathbf{e}_i^I(s_i)$ at the endpoints and hence the boundary values for the adjacent intervals. In fact, since the system of equations is linear, these unknown values must be linearly related to the known boundary values and will depend also on the forcing function $\mathbf{r}(x)$ everywhere on $[s_{i-1},s_i]$. We can write this explicitly as, for $j = 1,2,\ldots,M$,

$$\begin{aligned}\mathbf{e}_j^I(s_j) &= A_j\mathbf{e}_j^I(s_{j-1}) + B_j\mathbf{e}_j^{II}(s_j) + \mathbf{g}_j^I(\mathbf{r}(x))\\ \mathbf{e}_j^{II}(s_{j-1}) &= C_j\mathbf{e}_j^I(s_{j-1}) + D_j\mathbf{e}_j^{II}(s_j) + \mathbf{g}_j^{II}(\mathbf{r}(x))\end{aligned} \tag{32}$$

where A_j, B_j, C_j, D_j are matrices and $\mathbf{g}_j^I$ and $\mathbf{g}_j^{II}$ are functionals of $\mathbf{r}(x)$ – essentially integrals of $\mathbf{r}$ over the subinterval I_j. Since the systems (31) are all diagonally dominant, we have estimates for their solutions, and in particular that means we have reasonable estimates for the size of the matrices A_j, B_j, C_j, D_j, and of $\mathbf{g}_j^I$ and $\mathbf{g}_j^{II}$. So we know that equations (32) are a linear system of equations for the unknown endstates $\mathbf{e}_j^I(s_j)$, $\mathbf{e}_j^{II}(s_{j-1})$, $j = 1,2,\ldots,M$ with coefficients and right-hand side that are of reasonable size. However, this does not tell us anything about the boundedness of the solution of the system (32) – that information and hence the global estimation of the error depends ultimately on the condition of the system (32) and hence on the well-posedness of the original problem (8) on the entire interval $[0,L]$.

Although we have not been able to say anything specific about the well-posedness of the problem (8) in the case that it cannot be written in diagonally dominant form over the whole interval, we can conclude from the above discussion that if the global problem is well-posed, then the local estimation of the error for the purposes of mesh refinement will work. In the next section we will show that if there are no turning points on the interval, then a sequence of local

transformations exist that put the system (8) into diagonally dominant form on uniformly bounded subintervals of $[0,L]$, and so the error can be estimated locally as we have outlined above.

2.2 Existence of a Transformation to Diagonally Dominant Form

As was remarked earlier, it is possible that if the system (2.0.1a) is not in diagonally dominant form, we may be able to find a bounded transformation to put it into that form. If the interval under consideration is small enough, then an exponential scaling of some of the variables may be all that is necessary. This motivates the following

Definition: *(Kreiss [1976]) Equations (2.0.1a) are said to be* **essentially diagonally dominant** *if there are constants* $0 < \delta < 1$ *and* ρ, C_1, C_2, *of moderate size such that*

$$|\mathrm{Im}a_{ii}| \le \rho|\mathrm{Re}a_{ii}| + C_1 \quad i = 1,2,\ldots,n \tag{1a}$$

and

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \le (1-\delta)|\mathrm{Re}a_{ii}| + C_2 \quad i = 1,2,\ldots,n. \tag{1b}$$

The error estimates (2.1.25) and (2.1.29) will apply to such systems if the interval length L is small enough because an essentially diagonally dominant system can be transformed to a diagonally dominant one by the moderate exponential scaling given by

$$\tilde{\mathbf{y}} := \begin{bmatrix} e^{-C_2 x}\mathbf{y}^I \\ e^{C_2 x}\mathbf{y}^{II} \end{bmatrix} \tag{2}$$

Many systems of interest can be written in the form

$$\mathbf{y}'(x) = \left[\frac{1}{\varepsilon}A_0(x) + A_1(x)\right]\mathbf{y}(x) + \mathbf{f}(x,\varepsilon) \tag{3}$$

where $A_0(x)$ and $A_1(x)$ are $n\times n$ matrices whose elements are $O(1)$, $\mathbf{y}$ and $\mathbf{f}$ are vectors of length n and $0 < \varepsilon \ll 1$ is a small parameter. If the system (3) is not essentially diagonally dominant, then under appropriate restrictions on the coefficients of the matrices A_0 and A_1 there exists a sequence of smooth transformations that will put it into essentially diagonally dominant form on

uniformly bounded subintervals. This is described in the following variant of a theorem of Kreiss [1978]:

Theorem 1 (Existence of local transformations to diagonally dominant form): *Consider the system (3) on the interval* $[0,1]$ *together with* n *linearly independent boundary conditions relating* $\mathbf{y}(0)$ *to* $\mathbf{y}(1)$. *Let* $m < n$ *be the number of non-zero eigenvalues of* $A_0(x)$ *and assume that* m *is a constant and equal to the rank of* $A_0(x)$ *everywhere on* $[0,1]$. *Assume furthermore that there is a constant* K *of moderate size and independent of* ε *such that* $\| dA_0 / dx \|_\infty \le K$ *and that* $\| A_1(x) \|_\infty \le K$. *Then the interval* $[0,1]$ *can be divided up into subintervals* $[s_j, s_{j+1}]$, $0 = s_0 < s_1 < \cdots < s_q = 1$ *with* $s_{j+1} - s_j \ge \eta$, η *independent of* ε, *on each of which there exists a smooth transformation that puts (3) into essentially diagonally dominant form.*

Proof: Since by assumption the rank of A_0 is constant and less than n, $A_0(x)$ can be written

$$A_0(x) = \begin{bmatrix} I & 0 \\ T(x) & 0 \end{bmatrix} \begin{bmatrix} \hat{A}_{11}(x) & A_{12}(x) \\ 0 & 0 \end{bmatrix}$$

where the $(n-m)\times m$ matrix $T(x)$ is clearly as smooth as $A_0(x)$. Thus we can make the smooth change of variables

$$\mathbf{w} := \begin{bmatrix} I & 0 \\ -T(x) & I \end{bmatrix} \mathbf{y}$$

and transform (3) everywhere on $[0,1]$ to the form

$$\mathbf{w}'(x) = \left[\frac{1}{\varepsilon} \begin{bmatrix} A_{11}(x) & A_{12}(x) \\ 0 & 0 \end{bmatrix} + \tilde{A}_1(x) \right] \mathbf{w}(x) + \tilde{\mathbf{f}}(x,\varepsilon) \qquad (4)$$

where $A_{11}(x) = \hat{A}_{11}(x) + T(x)A_{12}(x)$ is an $m\times m$ matrix, A_{12} is an $m\times(n-m)$ matrix and the elements of $\tilde{A}_1(x)$ are $\mathbf{O}(1)$.

We will now show that a sequence of constant transformations S_j that put $\tilde{A}_0 := \begin{bmatrix} A_{11} & A_{12} \\ 0 & 0 \end{bmatrix}$ into diagonally dominant form can be constructed. Since $\tilde{A}_1$ is $\mathbf{O}(1)$ and bounded, the entire system (4) will therefore be essentially diagonally dominant if we choose the constant C_2 appropriately. Assume that on the interval $[0,s_j]$ such a sequence of transformations $S_0, S_1, \ldots, S_{j-1}$ has been constructed. Then by a unitary transformation U_j, $\tilde{A}_0(s_j)$ can be transformed to upper triangular form, i.e.

$$U_j^{*}\tilde{A}_0(s_j)U_j = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & 0 \end{bmatrix} =: B_j(s_j)$$

where

$$\tilde{A}_{11} := \begin{bmatrix} \kappa_1 & b_{12} & \cdots & b_{1m} \\ 0 & \kappa_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \kappa_m \end{bmatrix} \quad \text{and} \quad \tilde{A}_{12} := \begin{bmatrix} b_{1,m+1} & \cdots & b_{1n} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ b_{m,m+1} & \cdots & b_{mn} \end{bmatrix}$$

with $|\kappa_1| > |\kappa_2| > \cdots > |\kappa_m|$. We assume that $\varepsilon^{-1}B_j(s_j)$ satisfies condition (1a). (If $\varepsilon^{-1}B_j(s_j)$ does not satisy condition (1a) then we can apply a stretching $x = \beta\tilde{x}$, $0 < \beta < 1$ to the independent variable. Then the system (4) becomes

$$\mathbf{w}_{\tilde{x}}(\tilde{x}) = (\frac{\beta}{\varepsilon}\tilde{A}_0 + \beta\tilde{A}_1)\mathbf{w}(\tilde{x}) + \tilde{\mathbf{f}}(\tilde{x},\varepsilon)$$

so clearly β can be chosen so that (1a) is satisfied.) The more critical requirement is that (2.1.1b) be satisfied. To accomplish this we can apply a diagonal scaling to $B_j(x_j)$, i.e. let

$$D_j = \begin{bmatrix} d_1^{(j)} & 0 & \cdots & \cdots & 0 \\ 0 & d_2^{(j)} & 0 & & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & d_m^{(j)} & \vdots \\ \vdots & & & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}$$

where $d_1^{(j)} \geq d_2^{(j)} \geq \cdots \geq d_m^{(j)} \geq 1$ and transform $B_j(x_j)$ to $C_j(x_j) = D_j^{-1}B_j(x_j)D_j$. The $d_i^{(j)}$ can be chosen large enough so that the elements c_{kl} in the first m rows of $C_j(x_j)$ satisfy condition (2.1.1b). So the net transformation $S_j := U_jD_j$ transforms $\tilde{A}_0$ to diagonally dominant form at the point x_j. We now want to show that the matrix $S_j^{-1}\tilde{A}_0(x)S_j$ will be in diagonally dominant form for all $x \in [s_j, s_{j+1}]$, where $|s_{j+1} - s_j| \geq \eta$, independent of ε. This follows in a straightforward manner from the smoothness assumption on $A_0(x)$. Clearly by construction $C_j(x)$ is as smooth as $A_0(x)$, i.e. we have $\|dC_j(x)/dx\|_\infty \leq \tilde{K}$, and in particular $|dc_{kl}(x)/dx| \leq \tilde{K}$ (where $\tilde{K}$ is a constant independent of ε and of moderate size). From these inequalities we can conclude therefore that

$$|c_{kl}(x)| \leq |c_{kl}(s_j)| + \tilde{K}|x - s_j| \qquad k = 1,\ldots,m,\ \ l = 1,\ldots,n \qquad (5a)$$

and also since $c_{kk}(s_j) \neq 0$, $k = 1,2,\ldots,m$ that

$$|c_{kk}(x)| \le |c_{kk}(s_j)|(1 + \hat{K}|x - s_j|) \le e^{\hat{K}|x - s_j|}|c_{kk}(s_j)| \qquad (5b)$$

where $\hat{K} := \tilde{K} / \min_k |c_{kk}(s_j)|$. Using (5a) we have

$$\sum_{k \neq l} |c_{kl}(x)| \le \sum_{k \neq l} |c_{kl}(s_j)| + (n-1)\tilde{K}|x - s_j|$$

$$\le (1 - \delta)|\mathrm{Re}c_{kk}(s_j)| + \hat{K}(n-1)(1 + \rho)|x - s_j||\mathrm{Re}c_{kk}(s_j)| \qquad (6)$$

where we have used both the diagonal dominance estimate at $x = s_j$ and the fact that $c_{kk}(s_j) \neq 0$ to get the result on the right-hand side of (6). Now (5b) is valid with the role of x and s_j reversed, so we can therefore estimate the right-hand side of (6) in terms of $c_{kk}(x)$: We obtain:

$$\sum_{k \neq l} |c_{kl}(x)| \le e^{\hat{K}|x - s_j|}[(1-\delta) + \hat{K}(n-1)(1 + \rho)|x - s_j|]|\mathrm{Re}c_{kk}(x)| \qquad (7)$$

Clearly we can choose $|s_{j+1} - s_j|$ small enough so that (7) can be replaced with

$$\sum_{k \neq l} |c_{kl}(x)| \le (1 - \frac{\delta}{2})|\mathrm{Re}c_{kk}(x)| \qquad \text{for } x \in [s_j, s_{j+1}]. \qquad (8)$$

2.3 Error Estimates for a Second Order Equation with a Turning Point

The existence on an interval of length $O(1)$ of a bounded transformation to diagonally dominant form depends critically on the assumption that in equation (2.2.3) the matrix $A_{11}(x)$ is always invertible, or equivalently that it has constant rank m. An isolated point x_o where $\mathrm{rank}A_{11}(x_o) < m$ is called a **turning point** of the system (2.2.3). In the interval around a turning point the error estimation procedure described in sections 2.1 and 2.2 may not be valid, and so a special investigation must be made of such cases. We begin in this section by considering a two-by-two system of first order equations and finding *a posteriori* error estimates for the numerical solution of that system near a turning point. Higher order systems are considered in section 2.4.

Consider the case of a system of the form (2.2.3) where $n = 2$ and $m = 1$. That such a system can typically be written in the form

$$\mathbf{y}'(x) + \left\{\frac{1}{\varepsilon}\begin{bmatrix} a(x) & 1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ b(x) & 0 \end{bmatrix}\right\}\mathbf{y}(x) = \begin{bmatrix} 0 \\ f(x) \end{bmatrix} \qquad (1)$$

is shown in appendix 2.3A. (Here $\mathbf{y} := (y, v)^T$). We consider (1) on an interval

$-1 \le x \le 1$ with boundary conditions $y(-1) = \alpha$, $y(1) = \beta$ given and assume that $a(0) = 0$, i.e. $x = 0$ is a turning point of the system (1). It is obvious that (1) can be rewritten as a second order scalar equation

$$\varepsilon y''(x) + a(x)y'(x) - b(x)y(x) = f(x) \tag{2}$$

where $b(x) := \tilde{b}(x) - a'(x)$. If $b(x) > 0$ for all $x \in [-1,1]$, then it is well known that there is a maximum principle for the solution $y(x)$. We have

Lemma 1 (Maximum Principle):

$$\max_{-1 \le x \le 1} |y(x)| \le \max\left[\max_{-1 \le x \le 1} \left| \frac{f(x)}{b(x)} \right|, \alpha, \beta \right]$$

Proof: Suppose that an interior maximum of y occurs at x_0. Then $y'(x_0) = 0$, and $y''(x_0) < 0$. We have therefore that $y(x_0) \le -g(x_0)/\, b(x_0)$. Similarly, if an interior minimum occurs we have the opposite inequality. If there is no interior maximum or minimum, the extremal point must occur at either $x = -1$ or $x = 1$.

Kreiss [1976],[1978] has shown that with motivation similar to that at the end of section 2.1 and in section 2.2 the system (1) can be transformed to essentially diagonally dominant form on subintervals of [-1,1]. However, near the turning point, the length of the subintervals over which the error estimates are valid approaches $O(\sqrt{\varepsilon})$ exponentially. If in the adaptive mesh refinement procedure we begin with a mesh where the local meshwidth h_ν is everywhere large compared with $\sqrt{\varepsilon}$ (i.e. $\sqrt{\varepsilon} = o(h_\nu)$), this error estimation procedure will not be useful near the turning point since the subintervals of validity are smaller than the meshwidths. (Note that we can still expect to estimate the error as before in subintervals that are away from the turning point). For this reason we will investigate the error in the region near the turning point more closely for equation (2).

We will use the method of Kreiss and Kreiss (equation (2.1.3)) to approximate (1). It is given by

$$\begin{aligned} &\varepsilon D_+ y_\nu + (1 - \alpha_\nu)(a_{\nu+1} y_{\nu+1} + v_{\nu+1}) + \alpha_\nu (a_\nu y_\nu + v_\nu) = 0 \\ &D_+ v_\nu + \frac{1}{2}(b_\nu y_\nu + b_{\nu+1} y_{\nu+1}) = \frac{1}{2}(f_\nu + f_{\nu+1}) \\ &y_{-N} = \alpha, \qquad y_N = \beta \end{aligned} \tag{3}$$

Here y_ν and v_ν are discrete approximations to $y(x_\nu)$ and $v(x_\nu)$ respectively, $a_\nu := a(x_\nu)$, $b_\nu := b(x_\nu)$ and $f_\nu := f(x_\nu)$. The computational mesh is $\{x_\nu\}_{-N}^{N}$, $x_{-N} := -1$, $x_N := 1$ and the meshwidth is $h_\nu := x_{\nu+1} - x_\nu$. The parameters α_ν are defined as in section 2.1. We will only be interested in estimating the error in this difference approximation in an interval of length $O(h_\nu)$ which contains the turning point $x = 0$.

Now let $\varphi(x) \in C^3$ and $\psi(x) \in C^3$ be the piecewise polynomial interpolants of y_ν and v_ν, respectively, given by deBoor's construction (see lemma 2.1.1). Then φ and ψ satisfy a 2×2 system of "nearby" differential equations given by

$$\varphi'(x) + \frac{1}{\varepsilon}a(x)\varphi(x) + \frac{1}{\varepsilon}\psi(x) = \delta^{(1)}(x)$$

$$\psi'(x) + \tilde{b}(x)\varphi(x) - f(x) = \delta^{(2)}(x) \tag{4}$$

In the same way as in section 2.1 $\delta^{(1)}$ and $\delta^{(2)}$ can be estimated at the intermediate points $\xi_\nu^{(1)} := x_{\nu+1} - \alpha_\nu h_\nu$ and $\xi_\nu^{(2)} := x_{\nu+1} - h_\nu/2$, respectively, in each interval $[x_\nu, x_{\nu+1}]$. Then by using a linear combination of three adjacent values $\delta^{(j)}(\xi_{\nu-1}^{(j)})$, $\delta^{(j)}(\xi_\nu^{(j)})$, $\delta^{(j)}(\xi_{\nu+1}^{(j)})$, $j = 1,2$, $\delta^{(1)}(x)$ and $\delta^{(2)}(x)$ can be estimated for all $x \in [x_\nu, x_{\nu+1}]$. In this way we obtain

Lemma 2: *The following estimates hold for $\delta^{(1)}(x)$ and $\delta^{(2)}(x)$:*

$$\varepsilon \sup_{(\nu,\nu+1)} |\delta^{(1)}(x)| \le \bar{h}_\nu^2 \text{ const.} \left\{ \sup_{(\nu-1,\nu+2)} |\varepsilon\varphi'''| + \sup_{(\nu-1,\nu+2)} |(a\varphi)''| + \sup_{(\nu-1,\nu+2)} |\psi''| \right\}$$

$$\sup_{(\nu,\nu+1)} |\delta^{(2)}(x)| \le \bar{h}_\nu^2 \text{const.} \left\{ \sup_{(\nu-1,\nu+2)} |\psi'''| + \sup_{(\nu-1,\nu+2)} |(b\varphi)''| + \sup_{(\nu-1,\nu+2)} |f''| \right\} \tag{5}$$

Here $\bar{h}_\nu := \max_{[-1,1]} h_\nu$ and differentiation with respect to x is denoted by a prime.

More useful estimates of $\delta^{(1)}$, $\delta^{(2)}$ in terms of the solution of the difference approximation (3) are given in the following

Lemma 3: *$\delta^{(1)}(x)$ and $\delta^{(2)}(x)$ satisfy the following inequalities:*

$$\varepsilon \sup_{(\nu,\nu+1)} |\delta^{(1)}(x)| \le \bar{h}_\nu^2 \text{ const.} \left\{ \max_{[-4,3]} |D_+^2(a_\nu y_\nu)| + \max_{[-4,4]} |D_+ f_\nu| \right\}$$

$$\sup_{(\nu,\nu+1)} |\delta^{(2)}(x)| \le \bar{h}_\nu^2 \text{const.} \left\{ \max_{[-4,3]} |D_+^2(b_\nu y_\nu) + \max_{[-4,3]} |D_+^2 f_\nu| \right\} \tag{6}$$

Proof: By lemma 2.1.1 the derivatives on the right-hand side of equations (5) can be estimated in terms of divided differences of the solution $(y_\nu, v_\nu)^T$ of the difference equation (3). Then v_ν can be eliminated from the estimates by substituting the second of equations (3) into the right hand sides of those estimates.

Finally we have

Lemma 4: *If $b(x) > 0$ the error in the computed solution of (2), $e(x) := \varphi(x) - y(x)$ can be estimated by*

$$|e(x)| \leq \text{const.} \max_\nu \left\{ h_\nu^2 \left(\max_{[-4,3]} |D_+^3(a_\nu y_\nu)| + \max_{[-4,4]} |D_+^2(b_\nu y_\nu)| \right.\right.$$

$$\left.\left. + \max_{[-4,4]} |D_+^2 f_\nu| \right) \right\} \qquad (7)$$

Proof: Consider the equations

$$\begin{aligned} &\varepsilon y'(x) + a(x)y(x) + v(x) = \varepsilon g(x) \\ &v'(x) + (b(x) + a'(x))y(x) = f(x) \\ &y(-1) = 0, \qquad y(1) = 0. \end{aligned} \qquad (8)$$

The variable v can be eliminated from (8) to obtain

$$\varepsilon y'' + ay' - by = \varepsilon g' + f \,, \qquad (9)$$

and hence by lemma 1, if $b(x) > 0$, we have

$$\|y(x)\|_\infty \leq \text{const. } \|\varepsilon g'(x) + f(x)\|_\infty \qquad (10)$$

Applying lemma 2.2.1, (10) can be replaced by

$$\|y(x)\|_\infty \leq \text{const. } \|\varepsilon D_+ g(x) + f(x)\|_\infty \qquad (11)$$

By equation (4), the error $e(x)$ satisfies equations of the form (8), so combining inequalities (6) and (11), the desired estimate for $e(x)$ is obtained.

Remarks: The error estimate (7) is similar to but not exactly of the same form as the error estimate for the diagonally dominant case given by lemma 2.1.3. In particular the leading term $h_\nu^2 \max|D_+^3(a_\nu y_\nu)|$ is different, and could conceivably cause trouble since it is potentially $O(h_\nu^{-1})$. If initially the right-hand side of (7) is large, then in an adaptive mesh refinement procedure, the local

meshwidth h_ν will be reduced until the error estimate is of acceptable size. The following typical example indicates that this will happen (i.e. the mesh refinement will converge).

Suppose that at the turning point ($x = 0$) $a'(0) \neq 0$. Then clearly on an interval of length $O(\sqrt{\varepsilon})$ we have that $|a(x)| \leq A\sqrt{\varepsilon}$ where $A = o(\varepsilon^{-1/2})$ is a constant. For convenience define $\tilde{a}(\tilde{x})$ by $\sqrt{\varepsilon}\tilde{a}(\tilde{x}) = a(x)$ with $\tilde{x} = \varepsilon^{-1/2}x$, and note that $\tilde{a}(\tilde{x}) = o(\varepsilon^{-1/2})$ on this interval. Suppose now that by the mesh refinement procedure using the error estimate (7) the meshwidth has been reduced to $h_\nu = O(\sqrt{\varepsilon})$ in the small interval of interest about the turning point. Let $\tilde{h}_\nu := \varepsilon^{-1/2}h_\nu$ and $w_\nu := \varepsilon^{-1/2}v_\nu$. Then the difference approximation (3) can be rewritten as

$$
\begin{aligned}
&\tilde{D}_+ y_\nu + (1-\alpha_\nu)(\tilde{a}_{\nu+1}y_{\nu+1} + w_{\nu+1}) + \alpha_\nu(\tilde{a}_\nu y_\nu + w_\nu) = 0 \\
&\tilde{D}_+ w_\nu + \frac{1}{2}(b_\nu y_\nu + b_{\nu+1}y_{\nu+1}) = \frac{1}{2}(f_\nu + f_{\nu+1}) \qquad (12) \\
&y_{-N} = \alpha\,, \quad y_N = \beta
\end{aligned}
$$

where $\tilde{D}_+ u_\nu := (u_{\nu+1} - u_\nu)/\tilde{h}_\nu$. Similarly the error estimate (7) becomes

$$
|e(x)| \leq \text{const.} \max_\nu \left\{ \tilde{h}_\nu^2 \left(\max_{[-4,3]} |\tilde{D}_+^3(\tilde{a}_\nu y_\nu)| + \max_{[-4,4]} |\tilde{D}_+^2(b_\nu y_\nu)| \right.\right.
$$

$$
\left.\left. + \max_{[-4,4]} |\tilde{D}_+^2 f_\nu| \right) \right\} \qquad (13)
$$

It follows from the following lemma that the right-hand side of (13) is bounded independently of inverse powers of ε:

Lemma 5: *If there are constants M, M_f, M_a, M_b such that* $\max_\nu |y_\nu| \leq M$, $\sum_{j=0}^{2} \max_\nu |\tilde{D}_+^j f_\nu| \leq M_f$, $\sum_{j=0}^{2} \max_\nu |\tilde{D}_+^j \tilde{a}_\nu| \leq M_a$ *and* $\sum_{j=0}^{2} \max_\nu |\tilde{D}_+^j b_\nu| \leq M_b$ *then there exist constants M_k, $k = 1,2,3$ such that the solutions of (12) satisfy the inequalities*

$$
\max_\nu |\tilde{D}_+^k y_\nu| \leq M_k,\ k = 1,2,3.
$$

Proof: Since y_ν is bounded, it follows from the difference equations (12) that $\tilde{D}_+ w_\nu$ and $\tilde{D}_+ y_\nu$ are also bounded. To obtain estimates on the higher divided differences of y_ν, take divided differences of (12) and apply the assumptions of the lemma.

If the condition $b(x) > 0$ for all $x \in [-1,1]$ does not hold, the computational error $e(x)$ can still be estimated provided that there is an estimate for the differential equation (2). Abrahamsson [1975a] has considered equation (2) and gives estimates for its solution under a variety of conditions on the coefficients. For completeness, we quote some of his results in the remainder of this section.

It is well-known that if in equation (2) there are no turning points, i.e. $a(x) \neq 0$ for $x \in [-1,1]$, then the estimate

$$\| y \|_\infty^2 \leq \text{const.} \, (\| f \|_2^2 + \alpha^2 + \beta^2) \tag{14}$$

holds for sufficiently small ε independently of the sign of $b(x)$. We therefore would expect that the estimate for (2) should depend only on the sign of $b(0)$. In fact we have

Lemma 6 (Abrahamsson [1975a],theorems 5.1 and 7.2): *Assume that $b(0) > 0$. Then there are constants δ_0, K, and ε_0 independent of f, α, β, and ε such that the solutions of (2) subject to the boundary conditions $y(-1) = \alpha$, $y(1) = \beta$ satisfy the estimate*

$$\| y \|_\infty^2 \leq K(\max_{|x|<\delta} | f |^2 + \| f \|_2^2 + \alpha^2 + \beta^2) \tag{15}$$

for all $0 < \varepsilon \leq \varepsilon_0$ and $0 < \delta \leq \delta_0$.

If $b(0) \leq 0$, then the solutions of (2) are bounded only if certain conditions of $f(x)$, α and β hold. The behavior of the solutions depends on the size and sign of the parameter $b(0)/|a'(0)|$ and the sign of $a'(0)$. In particular, for the case that $l := -b(0)/|a'(0)| = 0,1,2,\ldots$ with $a'(0) < 0$, or $l = 1,2,3,\ldots$ with $a'(0) > 0$ (the "resonance" case) the solution may be exponentially large near the turning point and therefore we cannot expect to be able to estimate the computational error near the turning point for such problems. Even for the non-resonance cases with $b(0) \leq 0$, the solutions of (2) are in general only weakly stable. For example, we have

Lemma 7 (Abrahamsson [1975a], theorem 5.2): *If $a'(0) < 0$, $b(0) < 0$ and $l \neq 0,1,2,\ldots$ then the following estimate holds for the solutions of (2) subject to $y(-1) = \alpha$, $y(1) = \beta$:*

$$\| y \|_\infty^2 \leq \text{const.} \left[\max_{|x|<\delta} | f(x) |^2 + \max_{|x|<\delta} | f^{(k)}(x) |^2 + \| f \|_2^2 + \alpha^2 + \beta^2 \right] \tag{16}$$

where k is defined to be the integer such that $l < k < l+1$.

Depending on the size of l, this estimate may depend on high-order derivatives of the right-hand side of (2) and hence cannot be used to give useful *a posteriori* error estimates for the difference approximation because of our inability to estimate reliably the higher order derivatives of the right-hand side of the error equation (see lemma 2.1.1).

2.3A Transformation of a Two-by-Two System to a Second Order Scalar Equation

Consider the two-by-two system of first order equations given by

$$\frac{d}{dx}\begin{bmatrix} y \\ w \end{bmatrix} + \begin{bmatrix} \alpha(x) & \gamma(x) \\ \beta(x) & \delta(x) \end{bmatrix}\begin{bmatrix} y \\ w \end{bmatrix} = \begin{bmatrix} f^{(1)} \\ f^{(2)} \end{bmatrix} \tag{1}$$

Introduce a new variable $\tilde{w}$, defined by $\tilde{w} := \gamma w$. Then with $\eta(x) := (\delta(x) - \gamma'(x))/\gamma(x)$, the system (1) becomes

$$\frac{d}{dx}\begin{bmatrix} y \\ \tilde{w} \end{bmatrix} + \begin{bmatrix} \alpha & 1 \\ \gamma\beta & \eta \end{bmatrix}\begin{bmatrix} y \\ \tilde{w} \end{bmatrix} = \begin{bmatrix} f^{(1)} \\ f^{(2)} \end{bmatrix} \tag{2}$$

We now exponentially scale the variable $\tilde{w}$, i.e. introduce a new variable $\tilde{v} := \tilde{w}\exp(\int \eta(x)dx)$. Then the system (2) becomes

$$\frac{d}{dx}\begin{bmatrix} y \\ \tilde{v} \end{bmatrix} + \begin{bmatrix} \alpha & e^{-\int \eta dx} \\ e^{\int \eta dx}\gamma\beta & 0 \end{bmatrix}\begin{bmatrix} y \\ \tilde{v} \end{bmatrix} = \begin{bmatrix} f^{(1)} \\ e^{\int \eta dx} f^{(2)} \end{bmatrix} \tag{3}$$

The variable $\tilde{v}$ can be eliminated from (3) by differentiation of the first equation followed by substitution of the second equation. We obtain then a single second order equation of the form

$$y''(x) + (a(x)y)' - b(x)y = \tilde{f}(x) \tag{4}$$

where

$$a(x) := \alpha(x) + \eta(x)$$

$$b(x) := \gamma(x)\beta(x) + \eta'(x) - \eta(x)\alpha(x)$$

and

$$\tilde{f}(x) := \eta(x)f^{(1)}(x) - f^{(2)}(x) + \frac{d}{dx}f^{(1)}(x).$$

Equation (4) can also be rewritten as a first order system by introducing an auxiliary variable $u(x)$ defined by $u'(x) + b(x)y(x) = f'(x)$. We then have the 2×2 system

$$\frac{d}{dx}\begin{bmatrix} y \\ u \end{bmatrix} + \begin{bmatrix} a(x) & 1 \\ b(x) & 0 \end{bmatrix}\begin{bmatrix} y \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ f'(x) \end{bmatrix} \tag{5}$$

2.4 First Order Systems with a Turning Point

In this section we consider a first order system of ordinary differential equations of the form

$$\mathbf{y}'(x) = \left[\frac{1}{\varepsilon}A_0(x) + A_1(x)\right]\mathbf{y}(x) + \mathbf{f}(x,\varepsilon) \tag{1}$$

on an interval $0 \le x \le 1$ where A_0 and A_1 are $n \times n$ matrices whose elements are smooth and $\mathbf{O}(1)$, $\mathbf{y}$ and $\mathbf{f} = \mathbf{O}(\frac{1}{\varepsilon})$ are vectors of length n, and we assume that except at a finite number of points, the number of non-zero eigenvalues of $A_0(x)$ is equal to its rank. Because of this condition on the rank of A_0, we can assume furthermore that $A_0(x)$ is in the form

$$A_0(x) = \begin{bmatrix} A_{11}(x) & A_{12}(x) \\ 0 & 0 \end{bmatrix}$$

where A_{11} is an $m \times m$ matrix.

We are interested in the case when the system (1) has a turning point in the interior of the interval [0,1], i.e. we assume that $A_{11}(x)$ has rank m for all $x \in [0,1]$, except at one point x_0 away from the boundaries $x = 0$ and $x = 1$ where $\text{rank}A_{11}(x_0) = r < m$. For convenience, we call this a "*rank-(m-r) turning point*". In this discussion we will limit consideration to the case of systems of type (1) with a rank-1 turning point, i.e. one eigenvalue of the matrix A_{11} is zero at the point x_0 while the remaining eigenvalues are non-zero.

The behavior of the solutions of the system (1) near a turning point of this type can often be described in terms of the two-by-two system of equations considered in section 2.3. Accordingly, we will show in this section that in the region near a rank-1 turning point the system (1) can be smoothly transformed to a normal form in which two of the equations in the system are essentially decoupled from the remaining equations and the turning-point behavior is governed by these two equations.

We first define the **turning point interval** to be a subinterval of [0,1] containing the point x_o in which one eigenvalue $\kappa(x)$ of $A_{11}(x)$ is separated from the remaining eigenvalues $\lambda_j(x)$, $j = 1,2,...,r$ of A_{11} in the following way: *There is a constant* $0 < \eta \ll 1$ *such that in the turning point interval*

$$|\kappa(x)| \le \eta\,|\lambda_j(x)|\ ,\ \ j = 1,2,...r \tag{2}$$

We then have the following

Theorem 1 (Transformation to normal form): *If the system (1) has a rank-1 turning point at some point* x_o *away from the boundaries* $x = 0$, $x = 1$, *then in the corresponding turning point interval* I_t *for almost all matrices* $\frac{1}{\varepsilon}A_0(x) + A_1(x) \in C^\infty(I_t)$ *and functions* $\mathbf{f} \in C^\infty(I_t)$ *there exists a smooth change of variables*

$$\mathbf{w}(x) := U_1(x,\varepsilon)\mathbf{y}(x) + U_2(x,\varepsilon)\mathbf{f}(x)$$

such that the system (1) is transformed to the normal form

$$\mathbf{w}'(x) = \begin{pmatrix} \frac{1}{\varepsilon}\tilde{A}_{11}(x) & \varepsilon^{p-1}\tilde{A}_{12}(x) \\ \varepsilon^{p}\tilde{A}_{21}(x) & \frac{1}{\varepsilon}\tilde{A}_{22}(x,\varepsilon) \end{pmatrix}\mathbf{w}(x) + \varepsilon^{p-1}\tilde{\mathbf{f}} \tag{3}$$

Here p *is any positive integer,* $U_1(x,\varepsilon) \in C^\infty(I_t)$, $U_2(x,\varepsilon) \in C^\infty(I_t)$ *are* $n \times n$ *matrices,* $\tilde{A}_{11}$ *is an* $r \times r$ *matrix and* $\frac{1}{\varepsilon}\tilde{A}_{22}$ *has the form*

$$\frac{1}{\varepsilon}\tilde{A}_{22}(x,\varepsilon) := \begin{pmatrix} \frac{1}{\varepsilon}a(x) & \frac{1}{\varepsilon} & 0 & \cdots & 0 \\ b(x) & 0 & * & \cdots & * \\ 0 & * & * & & \vdots \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ 0 & * & \cdots & \cdots & * \end{pmatrix}$$

where "" indicates an element of* $O(1)$. *(In this section* $C^p(I)$ *denotes the class of functions that have* p *continuous derivatives independent of* ε^{-1} *on the interval* I. *)*

Proof: The proof of this theorem follows from the following 4 lemmas:

Lemma 1: *There exists a transformation* $S_1(x) \in C^\infty(I_t)$ *such that*

$$S_1(x)\left[\frac{1}{\varepsilon}A_0(x) + A_1(x)\right]S_1^{-1}(x) = C_1(x,\varepsilon) :=$$

$$\begin{pmatrix} \frac{1}{\varepsilon}A(x) & \frac{1}{\varepsilon}\tilde{A}_{12}(x) + B_{12}(x) \\ B_{21}(x) & \frac{1}{\varepsilon}B(x) + B_{22}(x) \end{pmatrix} \tag{4}$$

where $A(x) \in C^{\infty}(I_t)$ *is an* $r \times r$ *matrix,* $\tilde{A}_{12}$ *has the form*

$$\tilde{A}_{12}(x) := \begin{pmatrix} 0 & a_{1,r+2} & \cdots & a_{1n} \\ 0 & a_{2,r+2} & \cdots & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & a_{r,r+2} & \cdots & a_{r,n} \end{pmatrix}$$

B_{12}, B_{21} *and* B_{22} *are matrices whose elements are* $O(1)$ *and* B *has the form*

$$B(x) := \begin{pmatrix} \kappa(x) & b_{12}(x) & \cdots & \cdots & b_{1,n-2}(x) \\ 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & \cdots & 0 \end{pmatrix} \tag{5}$$

Proof: Consider the submatrix $A_{11}(x)$. It is well-known that an $m \times m$ projection operator $P_{\kappa}(x)$ can be constructed that projects $\mathbb{R}^m$ into the invariant subspace corresponding to the eigenvalue $\kappa(x)$. This projection operator can for example be written explicitly in terms of an integral of the resolvent matrix of A_{11}, [Kato, 1976]:

$$P_{\kappa}(x) := -\frac{1}{2\pi i}\int [A_{11}(x) - \lambda]^{-1} d\lambda \tag{6}$$

where the integral is to be taken on a closed positively oriented path in the complex-λ plane enclosing the eigenvalue $\kappa(x)$ but not enclosing the remaining eigenvalues $\lambda_j(x)$, $j = 1,2,\ldots,r$. The condition (2) means that the eigenvalue $\kappa(x)$ is well separated from the $\lambda_j(x)$ and so this path of integration can be taken as a fixed circle centered at the origin. Then it is clear from (6) that the projection operator is as smooth as the elements of $A_{11}(x)$. The projection into the invariant subspace corresponding to the eigenvalues λ_j, $j = 1,2,\ldots r$ is given by $I - P_{\kappa}(x)$ and a similarity transformation $\tilde{S}_1$ can be constructed using these projection operators such that

$$\tilde{S}_1 A_{11}(x) \tilde{S}_1^{-1} = \begin{pmatrix} A(x) & 0 \\ 0 \cdots 0 & \kappa(x) \end{pmatrix} \tag{7}$$

where $A(x)$ is an $r \times r$ matrix with eigenvalues $\lambda_j(x)$, $j = 1,2,\ldots,r$. $S_1(x)$ is then

clearly taken to be

$$S_1(x) := \begin{bmatrix} \tilde{S}_1(x) & \\ & I \end{bmatrix}$$

The off-diagonal blocks of $C_1(x,\varepsilon)$ can be made arbitrarily small by using the change of variables described in

Lemma 2: *There exist $n\times n$ matrices $S_2(x,\varepsilon)\ \varepsilon\ C^\infty(I_t)$ and $T_2(x,\varepsilon)\ \varepsilon\ C^\infty(I_t)$ such that*

$$\mathbf{v}(x) := S_2(x,\varepsilon)\tilde{\mathbf{y}}(x) + T_2(x,\varepsilon)\mathbf{f}$$

satisfies the system of equations

$$\mathbf{v}'(x) = \begin{bmatrix} \frac{1}{\varepsilon}\tilde{A}(x) & \varepsilon^{p-1}\tilde{B}_{12}(x) \\ \varepsilon^p \tilde{B}_{21}(x) & \frac{1}{\varepsilon}B(x) + \tilde{B}_{22}(x) \end{bmatrix} \mathbf{v}(x) + \begin{bmatrix} \varepsilon^{p-1}\tilde{\mathbf{f}}^I \\ \varepsilon^{p-1}\tilde{\mathbf{f}}^{II} \end{bmatrix} \tag{8}$$

where $\tilde{A}(x)$ is an $r\times r$ matrix and $\tilde{\mathbf{y}}(x) := S_1(x)\mathbf{y}(x)$.

Proof: The proof of the lemma results essentially from the fact that by a smooth change of variables, one of the off-diagonal blocks can be made arbitrarily small without affecting the other off-diagonal block:

The function

$$\mathbf{u}(x) := \begin{bmatrix} I & S(x) \\ 0 & I \end{bmatrix} \tilde{\mathbf{y}}(x) + \begin{bmatrix} T(x)\mathbf{f}^I(x) \\ 0 \end{bmatrix}$$

where $\begin{bmatrix} \frac{1}{\varepsilon}\tilde{\mathbf{f}}^I \\ \frac{1}{\varepsilon}\tilde{\mathbf{f}}^{II} \end{bmatrix} := S_1(x)\mathbf{f}(x)$ and $\mathbf{f}^I(x) := (f^{(1)}, f^{(2)}, \ldots, f^{(r)}(x))^T$, satisfies the equation

$$\mathbf{u}'(x) = \begin{bmatrix} \frac{1}{\varepsilon}C_{11}(x) & C_{12}(x) \\ B_{21}(x) & \frac{1}{\varepsilon}C_{22}(x) \end{bmatrix} \mathbf{u}(x) + \begin{bmatrix} \mathbf{g}^I \\ \mathbf{g}^{II} \end{bmatrix} \tag{9}$$

where

$$C_{12} := \frac{1}{\varepsilon}\tilde{A}_{12} + B_{12} + \frac{1}{\varepsilon}SB + SB_{22} - \frac{1}{\varepsilon}AS - SB_{21}S + \frac{d}{dx}S$$

and

$$\mathbf{g}^I := \frac{d}{dx}(T\mathbf{f}^I) - (\frac{1}{\varepsilon}A + SB_{21})T\mathbf{f}^I + \frac{1}{\varepsilon}\mathbf{f}^I + S\mathbf{f}^{II}.$$

We can clearly choose $T(x) = A^{-1}(x)$ and $\mathbf{g}^I$ will be $\mathbf{O}(1)$. Then if we can find S satisfying

$$AS - SB = \tilde{A}_{12}$$

the matrix C_{12} will be $\mathbf{O}(1)$ as well. This is possible by the following argument: Write the matrix S in terms of its column vectors $\mathbf{s}_j$, i.e. $S = (\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{n-r})$. Then the vectors $\mathbf{s}_j$ if they exist are solutions of the system of equations

$$\begin{aligned} &(A - \kappa(x)I)\mathbf{s}_1 = \mathbf{c}_1 \\ &A\mathbf{s}_j = \mathbf{c}_j + b_{1j}I\mathbf{s}_1 \qquad j = 2,3,\ldots,n-r \end{aligned} \tag{10}$$

where the $\mathbf{c}_j$ are the column vectors of A_{12} ($\tilde{A}_{12} = (\mathbf{c}_1, \ldots, \mathbf{c}_{n-r})$) and b_{1j}, $j = 2,3,\ldots,n-r$ are defined by equation (5). It follows from the condition (2) and the fact that the eigenvalues of $A(x)$ are non-zero that equations (8) have a solution. It is also clear that the resulting transformation matrix will be as smooth as the elements of $A(x)$ and $\tilde{A}_{12}(x)$.

Note that $C_{22}(x) = B(x) + \varepsilon(B_{22} - B_{21}S) = B(x) + \mathbf{O}(\varepsilon)$, i.e. that as a result of this transformation, the lower right-hand block is perturbed only by terms which are $\mathbf{O}(\varepsilon)$. Therefore in the same way as above we can find a change of variables

$$\tilde{\mathbf{u}}(x) := \begin{bmatrix} I & \tilde{S}(x,\varepsilon) \\ 0 & I \end{bmatrix}\mathbf{u}(x) + \begin{bmatrix} \tilde{T}(x,\varepsilon)\mathbf{g}^I \\ 0 \end{bmatrix}$$

with $\tilde{S}(x,\varepsilon) := \sum_{j=1}^{p} \varepsilon^j \tilde{S}_j(x)$, $\tilde{T}(x,\varepsilon) := \sum_{j=1}^{p} \varepsilon^j \tilde{T}_j(x)$ such that $\tilde{\mathbf{u}}(x)$ satisfies

$$\tilde{\mathbf{u}}'(x) = \begin{bmatrix} \frac{1}{\varepsilon}\tilde{C}_{11}(x) & \varepsilon^{p-1}C_{12}(x) \\ B_{21}(x) & \frac{1}{\varepsilon}\tilde{C}_{22}(x) \end{bmatrix}\mathbf{u}(x) + \begin{bmatrix} \varepsilon^{p-1}\mathbf{g}^I(x) \\ \frac{1}{\varepsilon}\tilde{\mathbf{f}}^{II}(x) \end{bmatrix}$$

In making this transformation the upper right-hand block has been made arbitrarily small while the lower left-had block B_{21} has remained unaffected. It is therefore clear that in a similar way a transformation of the form

$$\mathbf{v}(x) := \begin{bmatrix} I & 0 \\ \hat{S}(x,\varepsilon) & I \end{bmatrix} \mathbf{v}(x) + \begin{bmatrix} 0 \\ \hat{T}(x,\varepsilon) \mathbf{f}^{II}(x) \end{bmatrix}$$

can be found such that $\mathbf{v}$ satisfies an equation of the form (8).

Now let $C_{22}(x) := \frac{1}{\varepsilon} B(x) + \tilde{B}_{22}(x)$. We can then show

Lemma 3: *There exists a unitary* $(n-r)\times(n-r)$ *transformation matrix* $S_3(x) \in C(I_t)$ *such that*

$$C_{\overline{22}}(x) := S_3^*(x) C_{22}^{\wedge}(x) S_3(x) = \begin{bmatrix} G_{11}(x) & G_{21}(x) \\ G_{21}(x) & G_{22}(x) \end{bmatrix}$$

where the submatrices G_{11}, G_{12} *and* G_{21} *have the form*

$$G_{11}(x) := \begin{bmatrix} \frac{1}{\varepsilon} c_{11}(x) & \frac{1}{\varepsilon} c_{12}(x) & 0 \\ c_{21}(x) & c_{22}(x) & c_{32}(x) \\ c_{31}(x) & c_{32}(x) & c_{22}(x) \end{bmatrix}, \qquad G_{12}(x) := \begin{bmatrix} 0 & \cdots & 0 \\ * & \cdots & c_{2,n-r}(x) \\ * & \cdots & c_{3,n-r}(x) \end{bmatrix}$$

$$G_{21}(x) := \begin{bmatrix} 0 & c_{42}(x) & c_{43}(x) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 0 & c_{n-r,2}(x) & c_{n-r,3}(x) \end{bmatrix}$$

and G_{22} *is an* $(n-r)\times(n-r-3)$ *matrix with elements that are* $O(1)$.

Proof: Denote the elements of $\hat{C}_{xx}(x)$ by c_{ij}. Then we can use Householder transformations to construct the matrix $S_3(x)$. Let

$$H_1(x) := \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \tilde{H}_1(x) & \\ 0 & & & \end{bmatrix}$$

with $\hat{H}_1(x) := I - \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}$ where $\mathbf{v} := \frac{1}{\varepsilon}(\tilde{c}_{12}, \hat{c}_{13}, \cdots, \hat{c}_{n,n-r})^T$ and $\tilde{c}_{12} = \hat{c}_{12} + e^{i\arg \hat{c}_{12}} \left[\sum_{j=2}^{n-r} \hat{c}_{ij}^2 \right]^{1/2}$. Then

$$H_1^* \hat{C}_{22} H_1 = \begin{bmatrix} \frac{1}{\varepsilon} c_{11}(x) & \frac{1}{\varepsilon} d_{11}(x) & 0 & \cdots & 0 \\ d_{21}(x) & * & \cdots & \cdots & * \\ \vdots & \vdots & & & \vdots \\ d_{n-r,1}(x) & * & \cdots & \cdots & * \end{bmatrix}.$$

Then take

$$H_2(x) := \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & & & \\ \vdots & \vdots & & \tilde{H}_2(x) & \\ 0 & 0 & & & \end{bmatrix}$$

with $\tilde{H}_2(x) := I - \frac{\mathbf{w}\mathbf{w}^T}{\mathbf{w}^T\mathbf{w}}$ where $\mathbf{w} := (\tilde{d}_{31}, d_{41}, \ldots, d_{n-r,1})^T$ and $\tilde{d}_{31} := d_{31} + e^{i\arg d_{31}}\left[\sum_{j=3}^{n-r} d_{j1}^2\right]^{1/2}$. Then $S_3(x) := H_2(x)H_1^*(x)$.

Finally we have

Lemma 4: *If* $c_{21}(x) \neq 0$, *then there exists a smooth transformation* $\hat{S}_4(x)$ *such that*

$$\hat{S}_4(x)\bar{C}_{22}(x)\hat{S}_4^{-1} = A_{22}(x) = \begin{bmatrix} \tilde{G}_{11}(x) & \tilde{G}_{12}(x) \\ \tilde{G}_{21}(x) & \tilde{G}_{22}(x) \end{bmatrix}$$

where $\tilde{G}_{12}$, $\tilde{G}_{21}$ *and* $\tilde{G}_{22}$ *have the same form as* G_{12}, G_{21} *and* G_{22}, *respectively, and*

$$\tilde{G}_{11} := \begin{bmatrix} \frac{1}{\varepsilon}\tilde{\kappa}(x) & \frac{1}{\varepsilon}\tilde{c}_{12}(x) & 0 \\ c_{21}(x) & \tilde{c}_{22}(x) & * \\ 0 & * & \tilde{c}_{33}(x) \end{bmatrix}$$

Proof: Obviously,

$$\hat{S}_4(x) = \begin{bmatrix} T(x) & 0 \\ 0 & I \end{bmatrix}$$

where

$$T(x) := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -c_{31}/c_{21} & 1 \end{bmatrix}.$$

The final step of the proof of theorem 1 is shown in appendix 2.3A where it is seen that provided that $\tilde{c}_{12} \neq 0$, the desired form of the equations (3) can be attained by a further change of variables.

It now remains to demostrate that the turning point behavior of the solution of (1) is governed by the equations for $w^{(r+1)}$ and $w^{(r+2)}$ and that those equations

are essentially decoupled from the rest of the system. This is clear from the following

Theorem 2 (Decoupling of the 2-by-2 system): *If the solution of (3) subject to n linearly independent boundary conditions relating* $\mathbf{w}(0)$ *and* $\mathbf{w}(1)$ *exists and is bounded independently of* ε, *then in the turning point interval* $(y,v)^T := (w^{(r+1)},w^{(r+2)})^T$ *satisfies an equation of the form*

$$\frac{d}{dx}\begin{bmatrix} y(x) \\ v(x) \end{bmatrix} = \begin{bmatrix} \frac{1}{\varepsilon}a(x) & \frac{1}{\varepsilon} \\ b(x) & 0 \end{bmatrix}\begin{bmatrix} y(x) \\ v(x) \end{bmatrix} + \begin{bmatrix} \varepsilon^{p-1}f^{(1)} \\ f^{(2)} \end{bmatrix} \tag{11}$$

where $f^{(2)} \in C^2(I_t)$ *and* $f^{(1)} = \mathbf{O}(1)$.

Proof: Away from the (possible) boundary layers at $x = 0$ and $x = 1$ the equation for $v(x)$ is of the form

$$v'(x) = b(x)y(x) + \sum_{j=r+3}^{n} b_{r+2,j}(x)w^{(j)}(x) + \mathbf{O}(\varepsilon^p), \tag{12}$$

therefore since $y(x) \in C^0(I_t)$, we have that $v(x) \in C^1(I_t)$. We can write the equations for $\mathbf{w}^{III}(x) := (w^{(r+3)}, \ldots, w^{(n)})^T$ in the form

$$\frac{d}{dx}\mathbf{w}^{III}(x) = M(x)w^{III}(x) + \mathbf{m}(x)v(x) + \mathbf{O}(\varepsilon^p) \tag{13}$$

(M is an $(n-r-3)\times(n-r-3)$ matrix and $\mathbf{m}$ is a vector of length $n-r-3$) from which we see that $\mathbf{w}^{III} \in C^2(I_t)$. Hence we can write (12) as

$$v'(x) = b(x)y(x) + f^{(2)}(x) \tag{14}$$

where $f^{(2)} \in C^2(I_t)$. A similar argument shows that the equation for $y(x)$ is in the postulated form.

2.4A Comments on Mesh Refinement Strategy and on Systems that Don't Reduce to the Normal Form

It is possible that for some systems the condition '$c_{21}(x) \neq 0$', which is a requirement for lemma 2.4.4 to be valid, may not hold. In this case the behavior of the system (2.4.1) near a turning point cannot be discussed in terms of a system of two equations; instead we may have a somewhat larger system of equations whose behavior will govern that of the larger system (2.4.1). In this appendix we will briefly consider the case when the behavior of the solutions of (1)

near a turning point are governed by a system of three coupled linear equations. In this case we find that the width of the turning point region is wider than in the previous case, and will correspondingly require less mesh refinement in order to resolve properly.

It is clear from the discussion in section 2.4 that it is the behavior of the matrix $\bar{C}_{22}(x)$ in lemma 2.4.3 that is important. If the element c_{21} of this matrix is zero then the transformation of lemma 2.4.4 will not work. In some cases, however, we can use arguments similar to those used to prove theorem 2.4.2 to conclude that the turning point behavior of the solution will be governed by the system of equations given by

$$\frac{d}{dx}\mathbf{u} = \begin{pmatrix} \frac{1}{\varepsilon}c_{11}(x) & \frac{1}{\varepsilon}c_{12}(x) & 0 \\ c_{21}(x) & c_{22}(x) & c_{23}(x) \\ c_{31}(x) & c_{32}(x) & c_{33}(x) \end{pmatrix} \mathbf{u} + \begin{pmatrix} \varepsilon^{p-1} f^{(1)} \\ \mathbf{f}^{II} \end{pmatrix} \tag{1}$$

where $\mathbf{f}^{II}\ \varepsilon C^2(I_t)$ is a vector with two elements and $\mathbf{u}$ is a vector with 3 elements. We can get an understanding for the behavior of the solutions of (1) by looking at the eigenvalues of the matrix

$$\tilde{C} := \begin{pmatrix} \frac{1}{\varepsilon}c_{11} & \frac{1}{\varepsilon}c_{12} & 0 \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{pmatrix}$$

The eigenvalues κ of $\tilde{C}$ satisfy the characteristic equation

$$\kappa^3 - (\frac{1}{\varepsilon}c_{11} + c_{22} + c_{33})\kappa^2 + (\frac{1}{\varepsilon}(c_{11}(c_{22} + c_{33}) - c_{12}c_{21}) + c_{22}c_{33} - c_{32}c_{23})\kappa + \frac{1}{\varepsilon}\Delta = 0$$

where $\frac{1}{\varepsilon}\Delta := \det|\tilde{C}| = \mathbf{O}(\frac{1}{\varepsilon})$. Using the technique of asymptotic analysis we can determine the sizes of the eigenvalues κ under different situations. We consider the following three cases:

Case I (Away from the turning point): In this case $c_{11} \neq 0$, and we expect that there will be one eigenvalue $\kappa_1 = \mathbf{O}(\frac{1}{\varepsilon})$ and two eigenvalues $\kappa_{2,3} = \mathbf{O}(1)$: Neglecting all but the leading order ($\mathbf{O}(\frac{1}{\varepsilon})$) terms in (2) we have

$$c_{11}\kappa^2 + c_{11}(c_{12} + c_{33}) - c_{12}c_{21})\kappa + \Delta = 0 \tag{3}$$

which clearly has two solutions that are $\mathbf{O}(1)$. Letting $\kappa = \varepsilon^{-1}\lambda$, $\lambda = \mathbf{O}(1)$, the

dominant balance in (2) is given by

$$\lambda^3 - c_{11}\lambda^2 = 0 \tag{4}$$

from which we conclude that there is one large eigenvalue $\kappa \sim \frac{c_{11}}{\varepsilon}$.

Case II (At the turning point, $c_{21} \neq 0$): In this case $c_{11} = 0$ by definition of the turning point. The dominant balance in (2) if we assume $\kappa = O(1)$ is then given by

$$-c_{12}c_{21}\kappa + \Delta = 0 \tag{5}$$

from which we conclude that there is only one such root. The other dominant balance comes from taking $\kappa = \varepsilon^{-1/2}\lambda$, $\lambda = O(1)$, whence

$$\lambda(\lambda^2 + c_{11}(c_{12} + c_{33}) - c_{12}c_{21}) = 0 \tag{6}$$

This equation has two non-zero solutions, and so we conclude that there are two eigenvalues $\kappa = O(1/\sqrt{\varepsilon})$. The eigensolutions of (1) are given by functions of the form $\mathbf{u}_j e^{\kappa_j x}$, $j = 1,2,3$, where $\mathbf{u}_j$ are the right eigenvectors of $\tilde{C}$, and so we conclude also that the region of rapid change near the turning point in this case has width $O(\sqrt{\varepsilon})$. This agrees with the analysis of section 2.4 which predicted that we would expect a mesh refinement to a meshwidth $h_\nu = O(\sqrt{\varepsilon})$ in this case.

Case III (At the turning point, $c_{21} = 0$): In this case there is no dominant balance for $\kappa = O(1)$ because the $O(\frac{1}{\varepsilon})$ part of the coefficient of κ^1 in (2) vanishes. The only dominant balance occurs when we take $\kappa = \varepsilon^{-1/3}\lambda$, $\lambda = O(1)$, in which case λ asymptotically satisfies

$$\lambda^3 + \Delta = 0 \tag{7}$$

Therefore we conclude that there are three eigenvalues $\kappa = O(\varepsilon^{-1/3})$. We would also conclude, therefore, that the region of rapid change near the turning point in this case has width $O(\varepsilon^{1/3})$.

In principle, we could expect to get almost any type of behavior at the turning point, depending on exactly what structure the coefficient matrix has. However (for linear problems) we can conclude that the region of rapid change in which mesh refinement will be necessary will typically be $O(\sqrt{\varepsilon})$ in width or wider. The exact behavior will depend, of course, on the behavior of the eigenvalues of $\frac{1}{\varepsilon}A_0 + A_1$ near the turning point.

It should be emphasized that in this chapter we have only discussed some specific examples of types of systems in which we can safely decide how to refine the computational mesh by looking only at lower order divided differences of the computed solution. The types of systems for which we have proved that this will work are systems with no turning points (which can always be transformed at least locally to diagonally dominant form if the coefficients for the equations are smooth enough), two-by-two systems of ordinary differential equations that can be rewritten as a second-order scalar equation, and $n \times n$ systems for which the reduction to a two-by-two system discussed in section 2.4 is possible. For the last two cases, we have shown that if the initial mesh has meshwidths which are much larger than $O(\sqrt{\varepsilon})$ in the turning point region, we should check the first *three* divided differences of the computed solution to properly estimate the error. Since this requires knowing *a priori* the location of the turning points, the realization of this method in an actual computer program will require some extra logic over problems with no turning points. In practice, however, we have found that checking only the first two divided differences of the solution everywhere is adequate to result in a proper refinement being made near a turning point. (This was done for all the numerical examples given in chapters 3 and 4). This is because if the true solution near the turning point has a region of rapid transition, the computed solution on a coarse mesh will usually exhibit that kind of behavior (or worse) near the turning point and so the first two divided differences will tend to be large there. This will cause the refinement algorithm to add points in this region. Then once the mesh has been refined such that the local meshwidth at the turning point is $O(\sqrt{\varepsilon})$, the error estimates based on the diagonal dominance assumption will again be valid (see section 2.3) and so the first two divided differences of the computed solution will give a correct estimate of the error.

For systems of ordinary differential equations that do not fall into one of the three categories mentioned above, we cannot expect solution-adaptive mesh refinement to work in general. It may be necessary to check the smoothness of the coefficients of the differential equation carefully as well. For example, Kreiss [1981] has pointed out that solution-adaptive mesh refinement can fail to resolve the solution of the problem

$$\varepsilon y'(x) + x^2 y(x) = x^2 + \varepsilon^{2/3} \qquad \text{on} -1 \le x \le 1 \tag{8}$$

with boundary condition $y(-1) = 1$ and where $0 < \varepsilon \ll 1$. It is easy to see that if

the terms containing ε can be neglected, then $y = 1$ is an approximate solution. The proper scaling near $x = 0$ is, however, $O(\varepsilon^{1/3})$, and in a region of this size about $x = 0$, the solution of (8) exhibits a small "peak" in which the solution is larger than 1. Kreiss gives a numerical example in which on a coarse mesh, the computed solution to (8) was smooth and given by $u(x) \approx 1$ on the whole interval, and so additional automatic refinement of the mesh was not made. Note that (8) is a scalar equation of the form (2.0.1a) with $a_{11}(x) = \frac{x^2}{\varepsilon}$. If we introduce an exponentially scaled variable $w = e^{-x}y$, (8) becomes

$$w'(x) + a(x)w = e^{-x}(x^2 + \varepsilon^{2/3}). \tag{9}$$

with $a(x) = \frac{x^2}{\varepsilon} + 1$. Lemma 2.1.3 gives an estimate for the error in the approximate solution of (9) assuming that the method of Kreiss and Kreiss [1981] is used to compute it. Note that the equation (2.1.25) gives a reasonable estimate only if $M_\nu := h_\nu^2 |a^{-1}(x)|(|a(x)| + |a'(x)| + a''(x)|)$ is of reasonable size everywhere. Near $x = 0$, $M_\nu \approx h_\nu^2 (x^2 + \varepsilon/3)^{-1}$, and so is only of reasonable size if $h_\nu = O(\sqrt{\varepsilon})$ in that region.

III. Numerical Methods for Problems in One Space Dimension

3.0 Introduction

All of the difference methods discussed in section 1.4 are characterized by the fact that the true physical dissipative or viscous terms in the differential equations are neglected in the numerical solution of the problem. Some of the methods are careful to introduce mechanisms (usually "artificial dissipation") by which the proper entropy production is retained. It is certainly true that for many physically interesting problems this neglect of the true viscous effects is justified and the solutions obtained using methods for the inviscid equations are valid and useful. However there are also applications in which this is not the case. A simple example is demonstrated by the following model problem for a chemically reacting flow:

$$\varepsilon u_{xx} - f(u)_x + g(x,u) = u_t \qquad + \text{ boundary conditions} \tag{1}$$

Here $g(x,u)$ is a strongly nonlinear function of u. The speed, c of a propagating front for this equation is obtained in the usual way by integrating the equation with respect to x across a region containing a front. It is given approximately by

$$c = (u_1 - u_2)^{-1} \left[f(u_1) - f(u_2) + \int_{x_1}^{x_2} g(x,u)\, dx \right] \tag{2}$$

where $u_j = u(x_j)$, $j = 1,2$. It is clear from this formula that if g is a strongly nonlinear function of u for the values of u taken on in the front region, the speed of the front and hence the global behavior of the solution can depend very critically on the proper resolution of the front profile. Since the shape and

width of this profile typically depends on the dissipative terms in the governing differential equations, it is important that a numerical method be used that models the dissipative effects properly.

Problems having differential equations that model physical processes involving a very small amount of dissipation typically fall into the class of *singular perturbation problems*. In chapter 2 we considered the theory behind some numerical methods for the accurate solution of singular perturbation problems. In particular we discussed the theoretical justification of solution-adaptive mesh refinement, a technique that allows us to get the proper resolution of boundary or internal layers in the solutions of the differential equations. This chapter gives examples in which this technique has been applied successfully. In sections 3.1 and 3.2 we first discuss in more detail the theory of difference approximations for second-order scalar ordinary differential equations with turning points, and in particular a new difference method is presented in section 3.2. Section 3.3 gives numerical examples of the application of this method to resolving steady shock profiles for Burgers' equation.

The difference method as it is implemented for Burgers' equation is implicit in time and unconditionally stable. Thus there is the temptation to use very large time steps when only the final steady profile solution is of interest. If the shock profile is stationary, then this approach will work well: the steady profile can be calculated with few time steps and it will be well-resolved. For a moving profile, however, the method tends to smear the profile unless the time step taken is very small. A loss of resolution is not just a failing of this particular method, but is typical for most methods when applied to a moving profile. A way to avoid this loss of resolution is by solving the problem in the moving reference frame in which the profile appears stationary. In section 3.4 we give numerical examples of a code in which the resolution of each moving profile is recovered by using local moving coordinate systems.

In section 3.5 we apply the theory discussed in chapter 2 to the solution of the equations for an isentropic thermally nonconducting gas in one space dimension. The differential equations are transformed to an appropriate normal form and a difference method is applied. Numerical examples are given.

3.1 Difference Approximations for Second Order Scalar Equations

In this section we discuss difference approximations for a scalar second order ordinary differential equation of the form

$$y''(x) + a(x)y'(x) - b(x)y(x) = f(x) \tag{1}$$

where $b(x) > 0$. As was pointed out in section 2.3, this equation can be transformed to a 2×2 system of first-order equations, and so we might well restrict our consideration to difference methods for first order systems. However, computational experience has shown that better numerical results can often be obtained if difference methods developed especially for equations of the form (1) are used. Accordingly, we consider the two point boundary value problem for (1), i.e. we consider (1) on the interval $0 \le x \le 1$ and specify the boundary conditions

$$y(0) = \alpha, \qquad y(1) = \beta. \tag{2}$$

We have already shown (see section 2.3) that the solution, $y(x)$, to this problem satisfies the maximum principle

$$\|y(x)\|_\infty \le \max\left\{\left\|\frac{f(x)}{b(x)}\right\|_\infty, \alpha, \beta\right\} \tag{3}$$

If certain additional conditions on $f(x)$ and $b(x)$ are satisfied, then the solution, $y(x)$, also has the property that it is monotone:

Lemma 1 (Monotonicity): *If on a subinterval* $0 \le x_1 \le x \le x_2 \le 1$, $\partial(f(x)/b(x))/\partial x \le 0$, $y(x_1) \le y(x_2)$ *and* $y'(x_1) > 0$, $y'(x_2) > 0$ *then on that interval* $y(x)$ *is monotone, i.e.* $\partial y(x)/\partial x \ge 0$. *(Clearly the result holds if all the inequalities are reversed as well.)*

Proof: Assume that for $x_1 \le \gamma \le x \le \eta \le x_2$ the result does not hold, i.e. we have $y'(x) < 0$. Divide through equation (1) by $b(x)$, differentiate with respect to x and integrate from γ to η. We have

$$\left.\frac{y''(x)}{b(x)}\right|_\gamma^\eta + \left.\frac{a(x)y'(x)}{b(x)}\right|_\gamma^\eta - \int_\gamma^\eta y'(\xi)d\xi = \int_\gamma^\eta \frac{\partial}{\partial\xi}(f(\xi)/b(\xi))d\xi$$

By construction the middle term on the left-hand side vanishes while the remaining terms are strictly positive. But by assumption $\frac{\partial}{\partial x}(f/b) \le 0$ which leads to a contradiction and hence the proof.

We are interested in solving the problem (1), (2) numerically using a solution-adaptive mesh refinement approach. in Kreiss and Kreiss [1981], the authors present a difference approximation for first order systems of ordinary differential equations (see equation (2.1.3)) that has the property that even when the solution is not adequately resolved by the computational mesh, the qualitative features of the computed solution are correct. In particular, numerical errors that oscillate rapidly on the mesh do not occur. This smoothness property of the computed solution helps make the adaptive mesh procedure an efficient process in that mesh refinement will not occur in regions where the true solution is smooth. For similar reasons we are therefore interested in difference approximations for equation (1) that satisfy a maximum principle similar to (3) and exhibit monotone solutions under conditions similar to those of lemma (1). A class of methods that has these properties are methods of *positive type: (cf. Dorr [1970] and Abrahamsson [1975b])*

We approximate (1) with a consistent three-point finite-difference scheme. We first introduce a nonuniform mesh with gridpoints $x_\nu = \sum_{j=1}^{\nu} h_j$, $\nu = 1,2,\ldots,N-1$, $x_0 = 0$, $x_N = 1$ where h_ν is the local mesh width. The function $y(x)$ is then replaced with a gridfunction y_ν at every point on the mesh. (The gridfunction y_ν is an approximation to $y(x_\nu)$.) The difference approximation then takes the form

$$\gamma_{-1}^{\nu} y_{\nu-1} + \gamma_0^{\nu} y_\nu + \gamma_1^{\nu} y_{\nu+1} = \tilde{f}_\nu \qquad \nu = 1,2,\ldots,N-1 \tag{4}$$

with boundary conditions $y_0 = \alpha$, $y_N = \beta$. Then,

Definition: *The method (4) is said to be of* **positive type** *if*

$$\gamma_1^{\nu} > 0\,,\ \gamma_{-1}^{\nu} > 0 \text{ and } \tilde{b}_\nu := \sum_{j=-1}^{1} \gamma_j^{\nu} < 0\,.$$

We then have:

Lemma 2 (Maximum principle for the difference equations): *If (4) is a method of positive type the*

$$\max_{[0,N]} |y_\nu| \le \max\left\{\max_{[0,N]} |\tilde{f}_\nu / \tilde{b}_\nu|,\ \alpha,\ \beta\right\}$$

Proof: Rewrite (4) as

$$-\tilde{b}_\nu y_\nu - \gamma_{-1}^{\nu}\Delta_- y_\nu + \gamma_1^{\nu}\Delta^+ y_\nu = \tilde{f}_\nu \tag{5}$$

(Here $\Delta_{\pm} w_\nu := \pm(w_{\nu\pm1} - w_\nu)$.) If there is an interior maximum of y_ν at x_μ then $\Delta_+ y_\mu < 0$ and $\Delta_- y_\mu > 0$. Hence $y_\mu \le -\tilde{f}_\mu / \tilde{b}_\mu$. If there is no interior extremum then the extremum must occur at the boundary.

With additional conditions on the coefficients of the scheme there is also a monotonicity result:

Lemma 3 (Monotonicity of Solutions to the Difference Equations) *If in addition to the conditions of Lemma 2, $\Delta_+(\tilde{f}_\nu / \tilde{b}_\nu) \ge 0$, $\Delta_+ y_{\nu_1} \le 0$ and $\Delta_- y_{\nu_2} \le 0$ hold on subinterval $\nu_1 \le \nu \le \nu_2$, then on that subinterval $\Delta_+ y_\nu \le 0$. (The result holds with the inequalities reversed as well.)*

Proof: Assume that on an interval $p \le \nu \le q$ this result does not hold, i.e. $\Delta_+ y_\nu \ge 0$. Divide through (5) by $\tilde{b}_\nu$ take a forward undivided difference (Δ_+) and sum from p to q. Then

$$-\sum_{\nu=p}^{q} \Delta_+ y_\nu - \left(\frac{\gamma_1^{q+1}}{\tilde{b}_{q+1}} \Delta_+ y_q - \frac{\gamma_{-1}^{p}}{\tilde{b}_p} \Delta_+ y_{p-1}\right) + \left(\frac{\gamma_1^{q+1}}{\tilde{b}_{q+1}} \Delta_+ y_{q+1} - \frac{\gamma_1^{p}}{\tilde{b}_p} \Delta_+ y_p\right) = \sum_{\nu=p}^{q} \Delta_+(\tilde{f}_\nu / \tilde{b}_\nu)$$

Because of the assumptions, we have that all of the terms on the left-hand side of this equation are negative while the right-hand side of the equation is positive. We arrive therefore at a contradiction and hence the proof.

It is clear that difference methods which satisfy the conditions of lemmas 2 and 3 will be useful for the numerical solution of (1),(2) using an adaptive mesh procedure. The property of monotonicity will assure that oscillations will not occur in the solution in undesirable places, and so the overall procedure should be relatively efficient in the sense of not refining the mesh unnecessarily. In the next several sections we discuss such a method and give some numerical examples in which it has been implemented successfully.

3.2 A Method of Positive Type for Second Order Equations on a Variable Mesh

In this section we discuss a new difference method of positive type for the linear second order ordinary differential equation

$$\varepsilon u'' + (a(x)u)' - bu = -bg(x) \qquad 0 \le x \le 1 \tag{1}$$

with $u(0) = c$, $u(1) = d$. Here $u(x)$ and $a(x)$ are real functions of x, $0<\varepsilon\ll1$ and b are constants. Furthermore the assumption is made that the inequality $b - a'(x) > 0$ is satisfied. Then by lemma 3.1.1 equation (1) has a maximum

principle:

$$\max_{0\le x\le 1}|u(x)| \le \max\left\{\max_{0\le x\le 1}|g(x)|, c, d\right\} \tag{2}$$

Furthermore, by lemma 3.1.2 if $g(x)$ is a monotone increasing (decreasing) function of x, then so is $u(x)$.

Now approximate (1) with a finite difference method. If possible, the method should have a maximum principle and have a monotonicity property similar to those for the differential equation. To accomplish this, write (1) as a first order system of equations by introducing a new variable $v(x)$:

$$\begin{aligned} \varepsilon u' + au + v &= 0 \\ v' + b(u-g) &= 0 \end{aligned} \tag{3}$$

These two equations are clearly equivalent to (1): Differentiate the first of equations (3) with respect to x and then eliminate v using the second equation. Now approximate (3) by a general two-point scheme:

$$\begin{aligned} \frac{\varepsilon}{h_\nu}\Delta_+ u_\nu + (1-\alpha_\nu)(a_{\nu+1}u_{\nu+1} + v_{\nu+1}) + \alpha_\nu(a_\nu u_\nu + v_\nu) &= 0 \\ \frac{1}{h_\nu}\Delta_+ v_\nu + (1-\beta_\nu)b(u_{\nu+1} - g_{\nu+1}) + \beta_\nu b(u_\nu - g_\nu) &= 0 \end{aligned} \tag{4}$$

Here u_ν and v_ν are approximations to $u(x_\nu)$ and $v(x_\nu)$ respectively, $a_\nu = a(x_\nu)$, and $g_\nu = g(x_\nu)$. The parameters $0 \le \alpha_\nu \le 1$ and $0 \le \beta_\nu \le 1$ are arbitrary and will be chosen later so as to assure that the approximation has a maximum principle and is as accurate as possible consistent with the maximum principle. (A similar idea was used by both Abrahamsson [1975b] and Hemker [1974] to find methods of positive type for equation (1).) The difference scheme (4) is at least first-order accurate and hence consistent with the differential equation (3). We see this by calculating the truncation error:

$$T_\nu \equiv \begin{bmatrix} \frac{\varepsilon}{h_\nu}\Delta_+ u(x_\nu) + (1-\alpha_\nu)(a(x_{\nu+1})u(x_{\nu+1}) + v(x_{\nu+1})) + \alpha_\nu(a(x_\nu)u(x_\nu) + v(x_\nu)) \\ \frac{1}{h_\nu}\Delta_+ v(x_\nu) + (1-\beta_\nu)b(u(x_{\nu+1}) - g(x_{\nu+1})) + \beta_\nu b(u(x_\nu) - g(x_\nu)) \end{bmatrix}$$

$$= h_\nu \begin{bmatrix} \varepsilon(\alpha_\nu - \frac{1}{2})u''(\xi) \\ (\beta_\nu - \frac{1}{2})v''(\xi) \end{bmatrix} + h_\nu^2 \begin{bmatrix} \varepsilon(\frac{3\alpha_\nu^2}{2} + \frac{\alpha_\nu}{2} - \frac{1}{6})u'''(\xi) \\ (-\frac{\alpha_\nu^2}{2} + 2\alpha_\nu\beta_\nu - \frac{\alpha_\nu}{2} + \beta_\nu)v'''(\xi) \end{bmatrix} + O(h_\nu^3) \tag{5}$$

Note that if $\alpha_\nu = \beta_\nu = 1/2$ the scheme (4) reduces to the Trapezoidal rule and is second-order accurate ($T_\nu = 0(h_\nu^2)$). In general it will not be possible to have both second-order accuracy everywhere on the mesh and satisfy the conditions for a maximum principle. Note, however, that if $\varepsilon \leq const.h_\nu$, the scheme will be second-order accurate if the single condition $\beta_\nu = 1/2$ is satisfied. We will use the results of lemma 3 to choose the parameters α_ν and β_ν so that a maximum principle holds for the difference scheme. To do this we must first rewrite (4) as a single second-order finite difference equation. This is easily done by eliminating v_ν and $v_{\nu+1}$ from equations (4). We obtain

$$\varepsilon\left(\frac{1}{h_\nu}\Delta_+ - \frac{1}{h_{\nu-1}}\Delta_-\right)u_\nu + \left((1-\alpha_\nu)\Delta_+ + \alpha_{\nu-1}\Delta_-\right)(a_\nu u_\nu)$$

$$-b\left[(1-\alpha_\nu)(1-\beta_\nu)h_\nu(u_{\nu+1}-g_{\nu+1}) + \left((1-\alpha_\nu)\beta_\nu h_\nu + \alpha_{\nu-1}(1-\beta_{\nu-1})h_{\nu-1}\right)(u_\nu - g_\nu)\right.$$

$$\left. + \alpha_{\nu-1}\beta_{\nu-1}h_{\nu-1}(u_{\nu-1}-g_{\nu-1})\right] = 0 \qquad (6)$$

In the notation of equation (3.1.4), we have for this scheme that

$$\gamma_1^\nu = h_\nu^{-1} + (1-\alpha_\nu)a_{\nu+1} - (1-\alpha_\nu)(1-\beta_\nu)h_\nu b$$

$$\gamma_0^\nu = -\left[h_\nu^{-1} + h_{\nu-1}^{-1} + (1-\alpha_\nu-\alpha_{\nu-1})a_\nu + \left((1-\alpha_\nu)\beta_\nu h_\nu + \alpha_{\nu-1}(1-\beta_{\nu-1})h_{\nu-1}\right)b\right]$$

$$\gamma_{-1}^\nu = h_{\nu-1}^{-1} - \alpha_{\nu-1}a_{\nu-1} - \alpha_{\nu-1}\beta_{\nu-1}h_{\nu-1}b$$

$$\tilde{g}_\nu = -b\left[(1-\alpha_\nu)(1-\beta_\nu)h_\nu g_{\nu+1} + \left((1-\alpha_\nu)\beta_\nu h_\nu + \alpha_{\nu-1}(1-\beta_{\nu-1})h_{\nu-1}\right)g_\nu + \alpha_{\nu-1}\beta_{\nu-1}h_{\nu-1}g_{\nu-1}\right]$$

We have therefore that

$$\tilde{b}_\nu \equiv -\sum_j \gamma_j^\nu = (1-\alpha_\nu)(h_\nu b - \Delta_+ a_n) + \alpha_{\nu-1}(h_{\nu-1}b - \Delta_- a_\nu) > 0$$

as long as not both $\alpha_\nu = 1$ and $\alpha_{\nu-1} = 0$. So by the conditions of lemma 3.1.3, the maximum principle

$$\max_\nu |u_\nu| \leq \max\left(|c|, |d|, \left|\frac{\tilde{g}_\nu}{\tilde{b}_\nu}\right| = \left|\frac{g(x_\nu)}{b - a'(x_\nu)} + 0(h_\nu)\right|\right) \qquad (7)$$

holds if in addition, α_ν and β_ν are chosen so that $\gamma_1^\nu \geq 0$ and $\gamma_{-1}^\nu \geq 0$ for $\nu = 1,2,\ldots,N-1$. Conditions on α_ν and β_ν which assure that these inequalities hold are given in the following

Proposition 1: *The conditions* $\gamma_{-1}^\nu \geq 0$ *and* $\gamma_1^\nu \geq 0$ *will hold for every* ν *if the following conditions on* α_ν *and* β_ν *are met:*

1) If $a_\nu \geq 0$ and $a_{\nu+1} \geq 0$, require that

$\alpha_\nu \leq (h_\nu(a_\nu + \beta_\nu h_\nu b))^{-1}$ and

either $\beta_\nu \geq 1 - a_{\nu+1}/(h_\nu b)$

or $\alpha_\nu \geq 1 + (h_\nu(a_{\nu+1} - (1-\beta_\nu)h_\nu b))^{-1}$

2) If $a_\nu \leq 0$ and $a_{\nu+1} \leq 0$, require that

$\alpha_\nu \geq 1 + (h_\nu(a_{\nu+1} + (1-\beta_\nu)h_\nu b))^{-1}$ and

either $\beta_\nu \leq a_\nu/(h_\nu b)$

or $\alpha_\nu \leq (h_\nu(a_\nu + \beta_\nu h_\nu b))^{-1}$

3) If $a_\nu \geq 0$ and $a_{\nu+1} \leq 0$, require that

$(h_\nu(a_\nu + \beta_\nu h_\nu b))^{-1} \leq \alpha_\nu \leq 1 + ((h_\nu(a_{\nu+1} - (1-\beta_\nu)h_\nu b))^{-1}$

4) If $a_\nu \leq 0$ and $a_{\nu+1} \geq 0$, require that

either $\beta_\nu \geq 1 - a_{\nu+1}/(h_\nu b)$

or $\alpha_\nu \geq 1 + (h_\nu(a_{\nu+1} - (1-\beta_\nu)h_\nu b))^{-1}$ and

either $\beta_\nu \leq -a_\nu/(h_\nu b)$

or $\alpha_\nu \leq (h_\nu(a_\nu + \beta_\nu h_\nu b))^{-1}$

Proof: These conditions are verified by tedious algebraic manipulation of the inequalities $\gamma_1^{\nu+1} \geq 0$ and $\gamma_1^\nu \geq 0$.

It is unfortunately not so easy to obtain conditions which will assure that the scheme will have monotone solutions. The scheme (6) can clearly be written in the form

$$\sum_j \tilde{\gamma}_j^\nu u_{\nu+j} = \sum_j \delta_j^\nu g_{\nu+j} \tag{8}$$

where $\tilde{\gamma}_j^\nu = \gamma_j^\nu / \tilde{b}_\nu$ and $\sum_j \delta_j^\nu g_{\nu+j} = \tilde{g}_\nu / \tilde{b}_\nu$. Assume now that $g(x)$ is a monotone increasing function of x, i.e. $\Delta_+ g(x) > 0$ for $0 \leq x \leq 1$. The conditions for monotone increasing solutions u_ν of (8) are, by lemma 4, that

$$\Delta_+ \frac{\tilde{g}_\nu}{\tilde{b}_\nu} = \Delta_+ \sum_j \delta_j^\nu g_{\nu+j} > 0 \tag{9}$$

Note that in the special case $\delta_j^\nu = \delta_j = const.$, $j = -1,0,1$ we have that

$$\Delta_+ \sum_j \delta_j g_{\nu+j} = \sum_j \delta_j \Delta_+ g_{\nu+j} > 0$$

by the assumptions on the function $g(x)$, and hence u_ν will be a monotone increasing function. In general, however, the coefficients δ_j^ν are not independent of ν and so the inequality (9) will be difficult to satisfy. For this scheme, the coefficients δ_j^ν are given by

$$\begin{aligned}
\delta_1^\nu &= \frac{(1-\alpha_\nu)(1-\beta_\nu)h_\nu b}{(1-\alpha_\nu)(h_\nu b - \Delta_+ a_\nu) + \alpha_{\nu-1}(h_{\nu-1} - \Delta_- a_\nu)} \\
\delta_0^\nu &= \frac{((1-\alpha_\nu)\beta_\nu h_\nu + \alpha_{\nu-1}(1-\beta_{\nu-1})h_{\nu-1})b}{(1-\alpha_\nu)(h_\nu b - \Delta_+ a_\nu) + \alpha_{\nu-1}(h_{\nu-1} - \Delta_- a_\nu)} \qquad (10) \\
\delta_{-1}^\nu &= \frac{\alpha_{\nu-1}\beta_{\nu-1}h_{\nu-1}b}{(1-\alpha_\nu)(h_\nu b - \Delta_+ a_\nu) + \alpha_{\nu-1}(h_{\nu-1} - \Delta_- a_\nu)}
\end{aligned}$$

The inequality (9) can be rewritten as

$$\sum_j \delta_j^\nu \Delta_+ g_{\nu+j} + \sum_j g_{\nu+j+1}\Delta_+ \delta_j^\nu \geq 0 \qquad (11)$$

It is clear from (10) that if the parameters α_ν and β_ν and the function $a(x_\nu)$ are smooth enough, the first term in (11) will dominate and the monotonicity property will hold. Because of the complexity of (10) it is difficult to get precise conditions that this be so. In practice, however, we have found that the solutions of difference equations similar to (4) are almost everywhere monotone if the coefficients are chosen so as to satisfy the conditions of Lemma 3.1.2 and to be smooth functions of $a(x)$ and b. We give computational examples of this in section 3.3.

A difference approximation that satisfies the conditions of Proposition 1 is given by the following choices for the parameters α_ν and β_ν:

I. If $a_\nu \geq 0$ and $a_{\nu+1} \geq 0$:

If $a_{\nu+1} \geq h_\nu b$, set $\beta_\nu = 1/2$, otherwise $\beta_\nu = 1 - a_{\nu+1}/2h_\nu b$

then if $a_\nu + \beta_\nu h_\nu b \leq \varepsilon / h_\nu$ set $\alpha_\nu = 1/2$, otherwise $\alpha_\nu = \varepsilon / (2h_\nu(a_\nu + \beta_\nu h_\nu b))$

II. If $a_\nu \leq 0$ and $a_{\nu+1} \leq 0$:

If $a_\nu \leq -h_\nu b$ set $\beta_\nu = 1/2$,

otherwise $\beta_\nu = -a_\nu / 2h_\nu b$

then if $a_{\nu+1} - (1-\beta_\nu)h_\nu b \geq -\varepsilon / h_\nu$ set $\alpha_\nu = 1/2$, otherwise $\alpha_\nu = 1 + \varepsilon / (2h_\nu(a_{\nu+1} - (1-\beta_\nu)h_\nu b))$

IIIa If $a_\nu < 0$ and $a_{\nu+1} > 0$ and $-a_\nu > a_{\nu+1}$,

If $a_\nu > -h_\nu b$ and $a_{\nu+1} < h_\nu b/2$ set $\beta_\nu = -a_\nu/2h_\nu b$
otherwise $\beta_\nu = 1/2$
then if $|a_{\nu+1} - (1-\beta_\nu)h_\nu b| \le \varepsilon b$ or $a_{\nu+1} \ge h_\nu b/2$ set $\alpha_\nu = 1/2$
otherwise $\alpha_\nu = 1 - \varepsilon/(2h_\nu|a_{\nu+1} - (1-\beta_\nu)h_\nu b|)$

IIIb If $a_\nu < 0$ and $a_{\nu+1} > 0$ and $-a_\nu < a_{\nu+1}$
if $a_{\nu+1} < h_\nu b$ and $a_\nu > -h_\nu b/2$ set $\beta_\nu = 1 - a_{\nu+1}/2h_\nu b$
otherwise $\beta_\nu = 1/2$
then if $|a_\nu + \beta_\nu h_\nu b| \ge -\varepsilon/h_\nu$ or $a_\nu \le -h_\nu b/2$ set $\alpha_\nu = 1/2$
otherwise $\alpha_\nu = \varepsilon/(2h_\nu|a_\nu + \beta_\nu h_\nu b|)$

IV If $a_\nu > 0$ and $a_{\nu+1} < 0$
if $a_\nu + h_\nu b/2 \le \varepsilon/h_\nu$ and $a_{\nu+1} - h_\nu b/2 \ge -\varepsilon/h_\nu$ then $\beta_\nu = 1/2$ and

$$\alpha_\nu = \frac{1}{2}\left[1 + \frac{\varepsilon}{h_\nu(a_{\nu+1} - h_\nu b/2)} + \frac{\varepsilon}{h_\nu(a_\nu + h_\nu b/2)}\right]$$

otherwise add a point x_ν^* to the mesh with $x_\nu \le x_\nu^* \le x_{\nu+1}$ where $a(x_\nu^*) = 0$.
Then either I or II is applicable.

Since we will usually be using this method together with an adaptive mesh refinement procedure, the special case under IV will not be difficult to implement. If the first condition of IV is not satisfied, then there is no choice of α_ν and β_ν that will guarantee that the difference approximation (6) satisfies a maximum principle. Numerical tests have shown that if an extra mesh point is not added, the solution may explode near the interval in which $a(x)$ changes sign from positive to negative.

This method looks rather formidable as presented above, but actually is fairly simple to implement. When $|a_\nu|$ is large relative to $|h_\nu b|$ it is essentially a one-sided scheme, and when $|a_\nu|$ is small relative to $|h_\nu b|$, it is essentially a centered scheme. However, the coefficients of the difference scheme are taken to be continuous functions of the coefficients of the differential equation. This is done in order to avoid convergence problems that might arise when using the scheme for a nonlinear problem.

3.3 Numerical Example: Stationary Shocks for Burgers' Equation

Burgers' equation is given by

$$\varepsilon U_{xx} + f(U)_x = U_t \tag{1}$$

with $f(U) = U^2/2$ and where $U = U(x,t)$. In this section we discuss a difference approximation for (1) on the interval $-1 \le x \le 1$ with initial data $U(x,0) = u_0(x)$ and boundary data $U(-1,t) = \alpha(t)$, $U(1,t) = \beta(t)$. We first approximate the time derivative using the "Backward Euler" method to get

$$\varepsilon u_{xx} + f(u)_x - u/k = u^*/k \tag{2}$$

Here $u = u(x)$ is an approximation to $U(x,t)$, $u^* = u^*(x)$ is an approximation to $U(x,t-k)$, and k is the time step. The same boundary and initial data are used. Since (2) is both nonlinear and implicit, we need an iteration procedure to solve for u. We use Newton's method. For solving a differential equation, Newton's method amounts to successively linearizing the differential equation about the current best guess to the solution. Denote by u^n the nth iterate in the Newton procedure, and let $\tilde{u} = u^{n+1} - u^n$, then the iteration for (2) is described by

$$\varepsilon \tilde{u}_{xx} + (f(u^n)\tilde{u})_x - \tilde{u}/k = (u^* - u^n)/k - f(u^n)_x - \varepsilon u^n_{xx} \qquad n = 0,1,2,\ldots \tag{3}$$

As an initial guess we take $u^0 = u^*$. At the $(n+1)$st iteration, the right-hand side of (3) is known, so we see that this equation is in the form of equation (3.2.1). We therefore approximate (3) with the method given by equation (3.2.4) with the parameters α_ν and β_ν chosen as described at the end of section 3.2.

Since, depending on the initial and boundary data specified, we expect boundary and internal layers to form, we use a solution adaptive mesh strategy to construct the mesh in order to resolve the solution. The procedure is given schematically at the end of this section. Note that the Newton iteration is in the inner loop, i.e. we solve the nonlinear problem completely on each mesh in the mesh iteration. It should be remarked that the theoretical presentation of chapter 2 would suggest the more conservative procedure of resolving each linearized problem completely, i.e. the mesh refinement should be the inner loop in the procedure. This is because in chapter 2 we have only treated the problem of how to solve a *linear* system of equations by mesh refinement. The estimates (more commonly available) for the solutions of linear differential equations rather than nonlinear equations allow us to estimate the computational error so that mesh refinement can be make in regions of high error. We

then expect the mesh to converge as it resolves the solution to the differential equation provided that the first few derivatives of the true solution are not infinite.

Doing the mesh refinement in the inner loop would, however, be much more expensive than the first method, and in practice for the *time-dependent* problems I have tested it has never been a problem that the Newton iteration is the innermost loop. (That this works results essentially from the fact that the time-step k is small. For non-time-dependent singular perturbation problems, such as those considered by Kreiss and Kreiss [1981] this procedure is less likely to work). Another feature of the implementation of the procedure to note is that points are only *added* to the mesh. This was done because the initial mesh $\mathbf{M}_0$ is always taken to be a uniform coarse mesh at every time step, and so it wasn't expected that overresolution of the solution would be a problem. This was only done for convenience, and in general for problems in which stationary shocks form, it would be more efficient to use the final mesh at the previous time step for the initial guess $\mathbf{M}_0$ and allow for mesh points to be taken out if the solution becomes too smooth.

In the first two numerical examples (figs. 1 and 2) the difference method of section 3.2 was applied to a uniform mesh to demonstrate the monotonicity-preserving property of the method. Recall from section 3.2 that the method is not actually monotonicity-preserving for all data, but is only "nearly monotonicity-preserving". In the examples I have run, however, the solutions have always been monotone in the sense that no numerical overshoots or oscillations are produced in the solution. This is an important property for a method to have when used in conjunction with a solution-adaptive mesh selection procedure because in this way unnecessary refinement is avoided. In both figures $\varepsilon = 1/100$ and $k = 1/10$. In fig. 1 the minimum value of the initial data is 0 and the maximum is 1. A left-moving shock is formed on the left side of the solution and a rarefaction wave on the right. Figure 2 is similar except that the minimum value of the initial data is -1 and so initially the shock is stationary. Since the characteristic points out of the calculation interval on the right, a boundary layer begins to form there.

Figure 3 shows an example using mesh refinement. For this example, $\varepsilon = 10^{-4}$, $k = 1/20$ and the end states of the initial data are $u = \pm 1$.

Figures 4 and 5 demonstrate that for calculating a steady profile, a large time step can be taken: $k = 1/20$ in figure 4 and $k = 1/4$ in figure 5. Although

the intermediate states in figure 5 do not agree well with those of figure 4, the final state does. This is to be expected because the equation for steady state solutions of the differential equation (1) is the same as that for the time-differenced equation(2): Setting $u_t = 0$ in (1) or $u = u^*$ in (2) yield the same steady state equation:

$$\varepsilon u_{xx} + f(u)_x = 0 \tag{4}$$

Unfortunately, the same is not true for travelling wave solutions of (1) and (2), i.e. moving profiles. A travelling wave solution of (1) or (2) moving at speed c will have the form $u(x,t) = w(x-ct)$. Substituting this expression into (1) and (2) we see that such solutions of the differential equation satisfy the equation

$$\varepsilon w'' + (f(w))' + cw' = 0 \tag{5}$$

while those for the difference equation satisfy

$$(\tfrac{1}{2}kc^2 + \varepsilon)w'' + (f(w))' + cw' = c^3O(k^2) \tag{6}$$

The effective dissipation in the last equation is increased by a factor depending on the size of the time step and the velocity of the profile. Thus we expect moving profiles to be smeared by this method even if the mesh is refined enough so that equation (2) is resolved.

This smearing is already apparent in the example of figure 6 (same as figure 2 but with $\varepsilon = .0025$ and mesh refinement). The shock is initially stationary and resolved well, but then the rarefaction wave begins to interact with the shock causing it to begin moving to the right, and by time $t = 1$ the profile has been smeared out slightly. This effect is much more evident in figure 7 where the endstates of the initial profile are $u = -.5$ and 1.5.

By inspection of equation (5) we can see that there are two obvious ways to overcome this difficulty. One is to reduce the stepsize k. This was done in the example shown in figure 8 where k has been reduced by a factor of 4. The profile is sharper at $t = 1$ in figure 8 than in figure 7, although the shock is still not as sharp as in figure 4. Indeed, we would expect to have to take a time step $k = o(\varepsilon)$ in order for the smearing effect to be avoided.

The other method by which the smearing effect can be eliminated is to solve the problem (2) in a moving coordinate frame. If the speed of the coordinate frame is taken to be that of the travelling wave then the steady state equations again both become (4) and we can expect sharp profiles if mesh refinement

is used. In the next section we describe and give examples of a code in which locally moving mesh segments are embedded in the finite difference mesh in order to resolve the moving features of the solution.

/* MESH REFINEMENT ALGORITHM FOR BURGERS' EQUATION [1]

u_n^m is the nth iterate on the mth mesh

$\mathbf{M}_m$ is the mth mesh

δ_n and δ_m are predetermined constants

$$\Delta(u) := \frac{h^2}{h_\nu + h_{\nu-1}}\left(\frac{|u_{\nu+1} - u_\nu|}{h_\nu} - \frac{|u_\nu - u_{\nu-1}|}{h_{\nu-1}}\right)$$

where $h_\nu = x_{\nu+1} - x_\nu$ and $h = \max(h_\nu, h_{\nu-1})$

*/

```
{
    /* Take one time step: */
    u^∞_{-1} = u*; /* Initial guess for solution
        is final solution at previous time step */
    m = 0;
    Construct M_0;    /* Initial guess for the mesh */
    while( M_m ≠ M_{m-1})
    {
        Interpolate u^∞_{m-1} onto M_m
            (call this u^0_m);
        n = 0;
        while(‖u^{n+1}_m − u^n_m‖ > δ_n)
        {
            Solve Newton equations (3);
            n = n+1;
        }
        u^∞_m ≡ u^n_m;
        Add mesh points until Δ(u^∞_m) < δ_m everwhere;
        Call this mesh M_{m+1};
        m = m+1;
    }
    u* = u^∞_m;
}
```

[1] /* This is a comment */
" **while**(condition) {procedure}" means to repeat the {procedure} until the (condition) is no longer true.

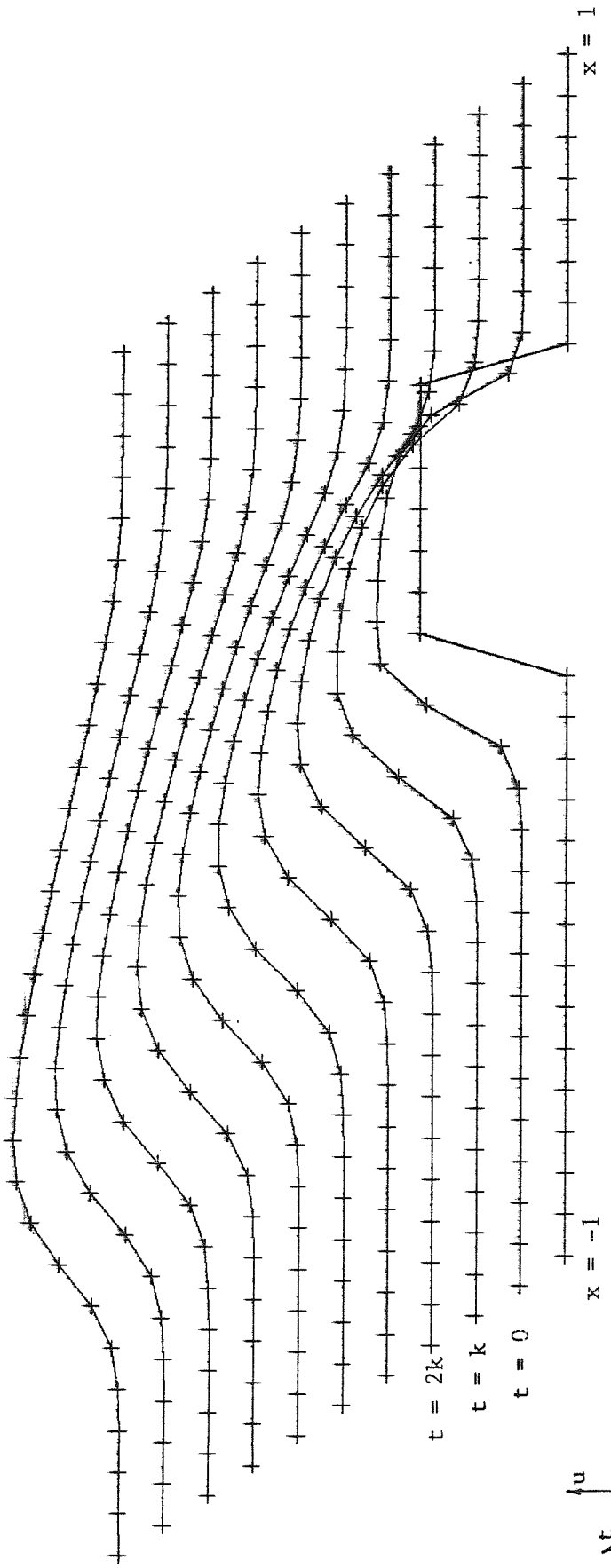


Figure 3.3.1

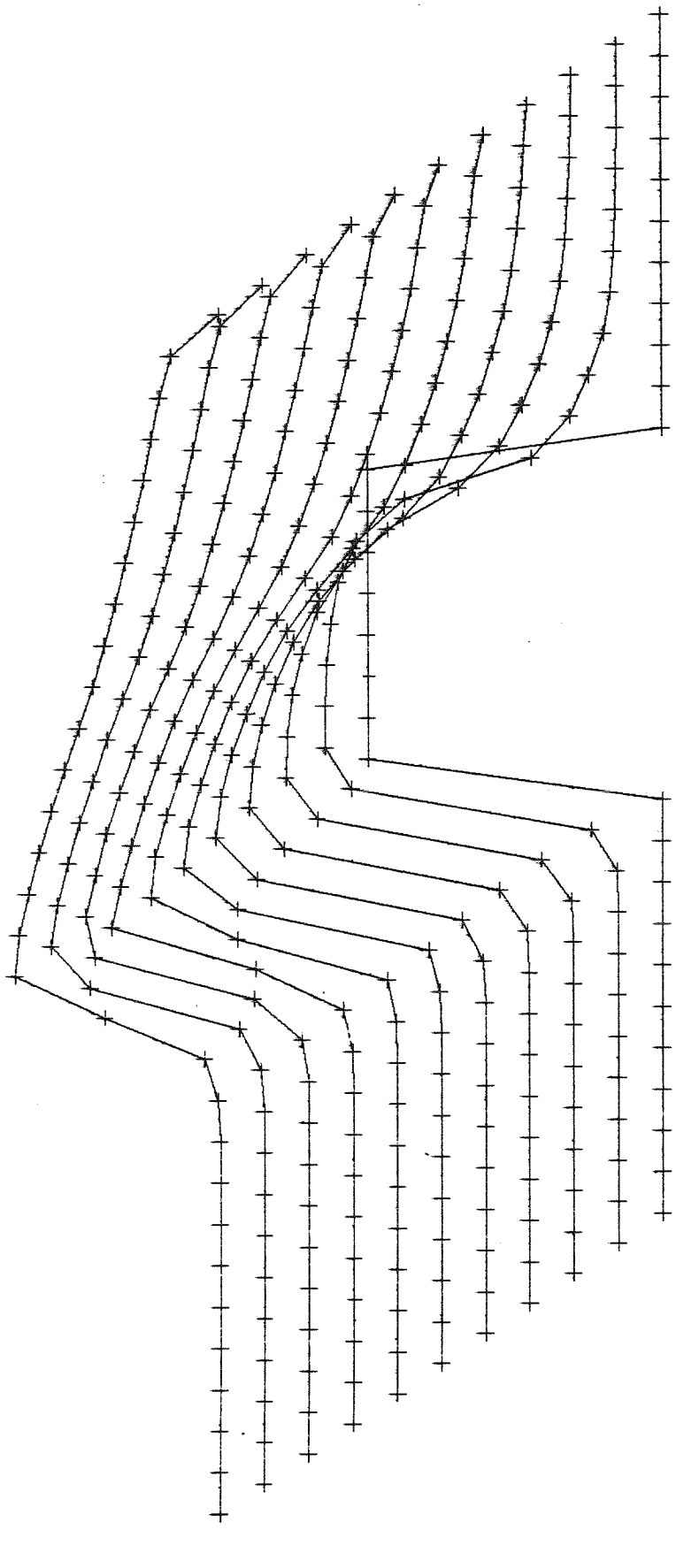
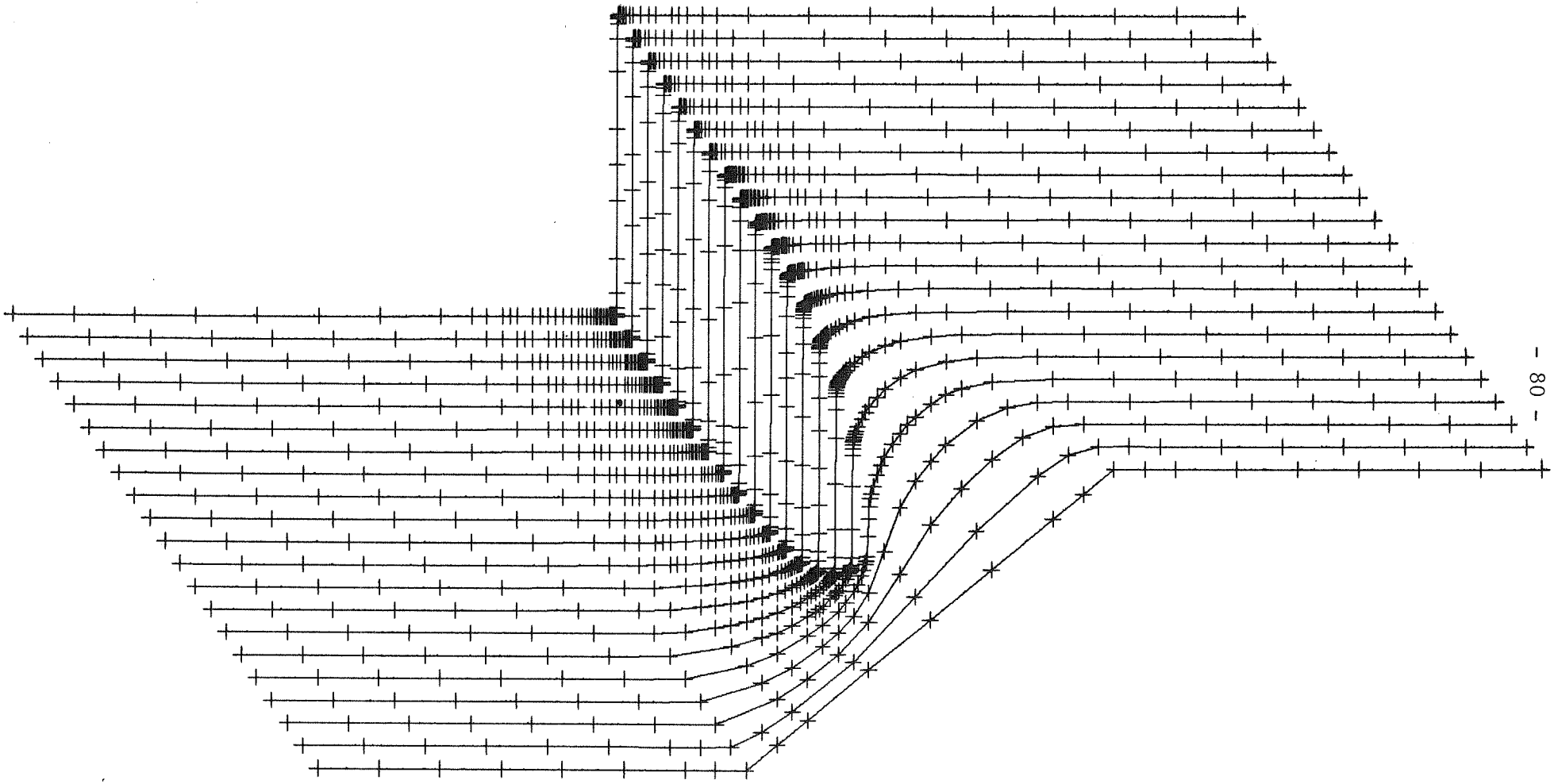
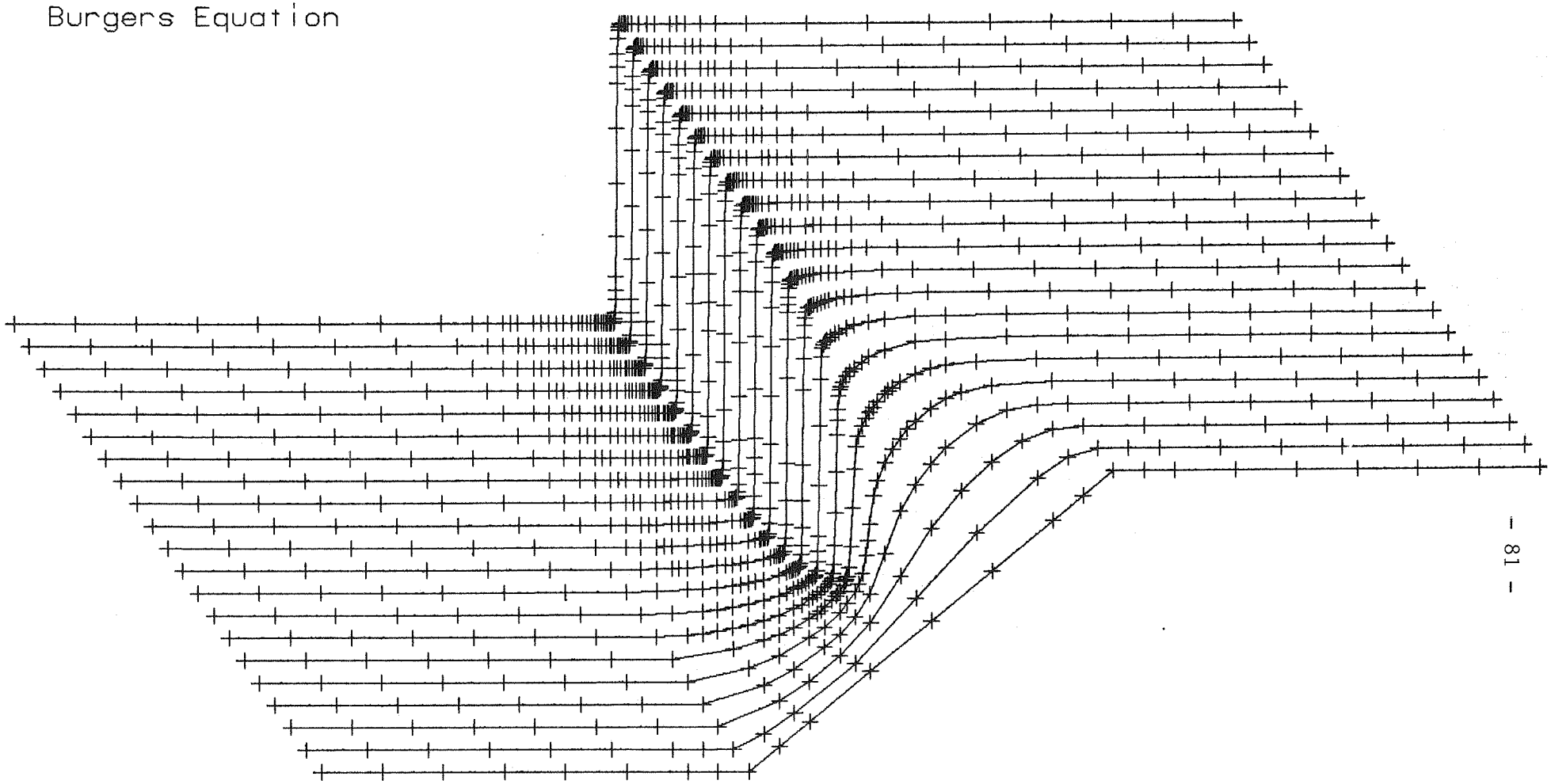


Figure 3.3.2



Burgers Equation
eps = 0.00010
k = 0.05000

Figure 3.3.3



eps = 0.00100
k = 0.050

Figure 3.3.4

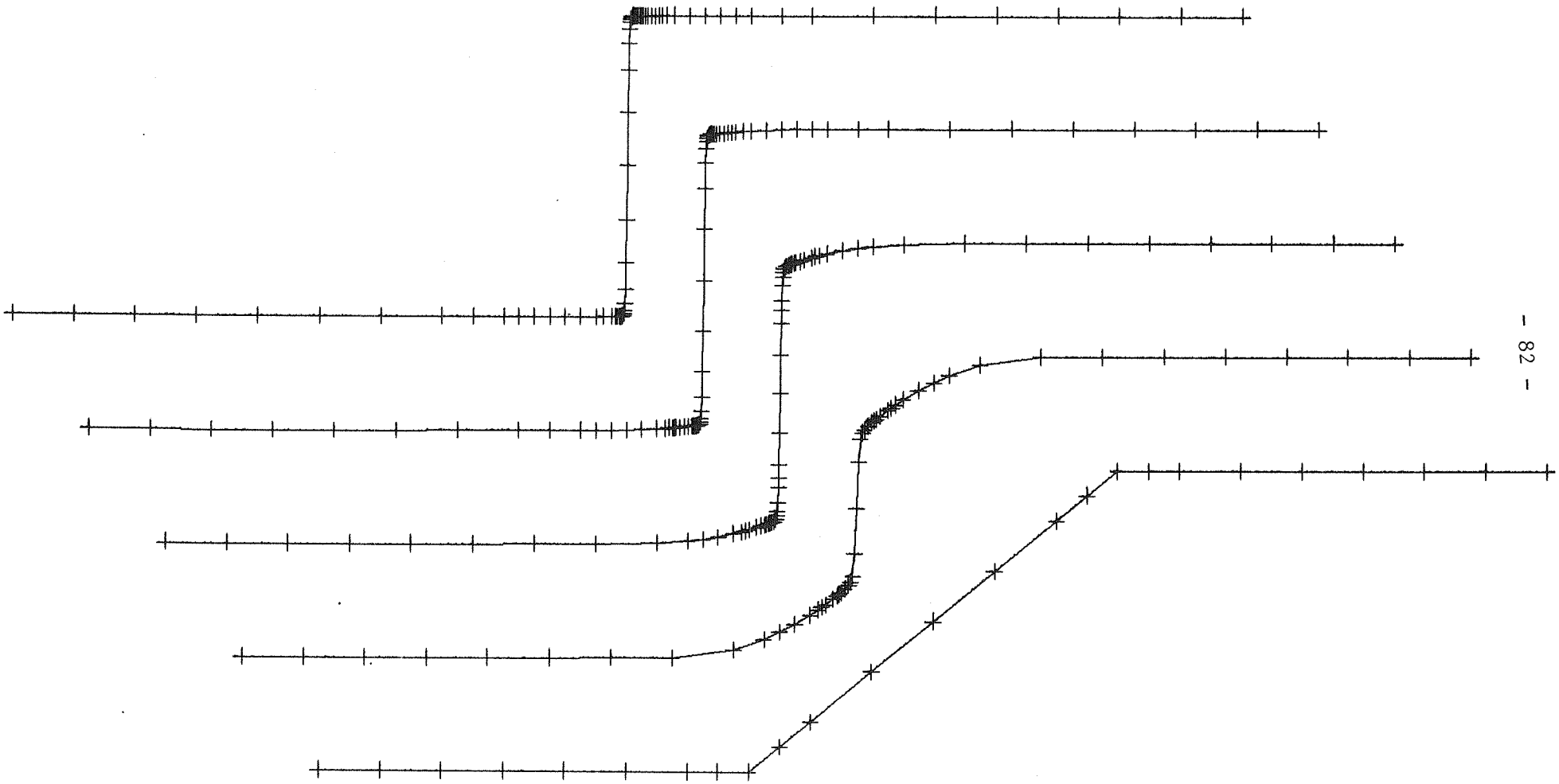


Figure 3.3.5

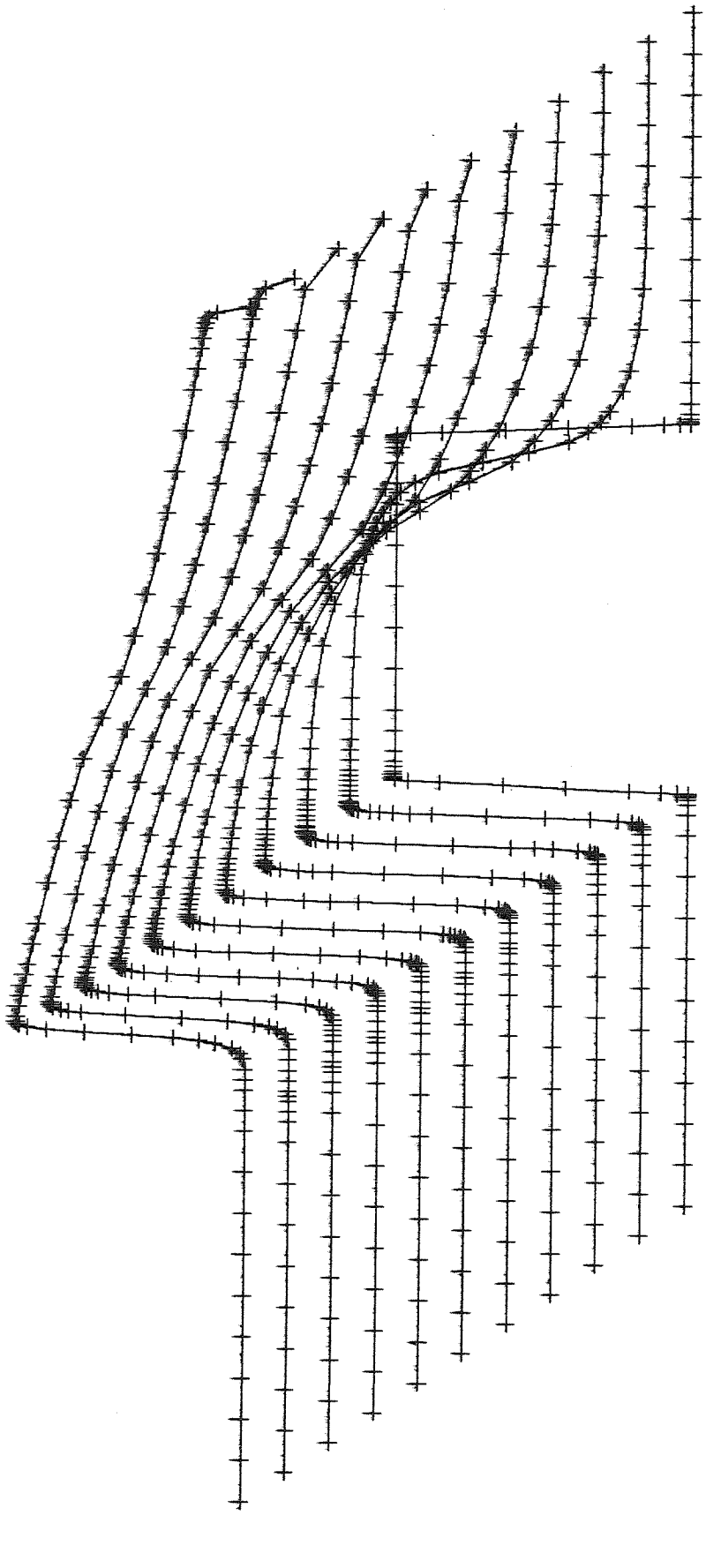
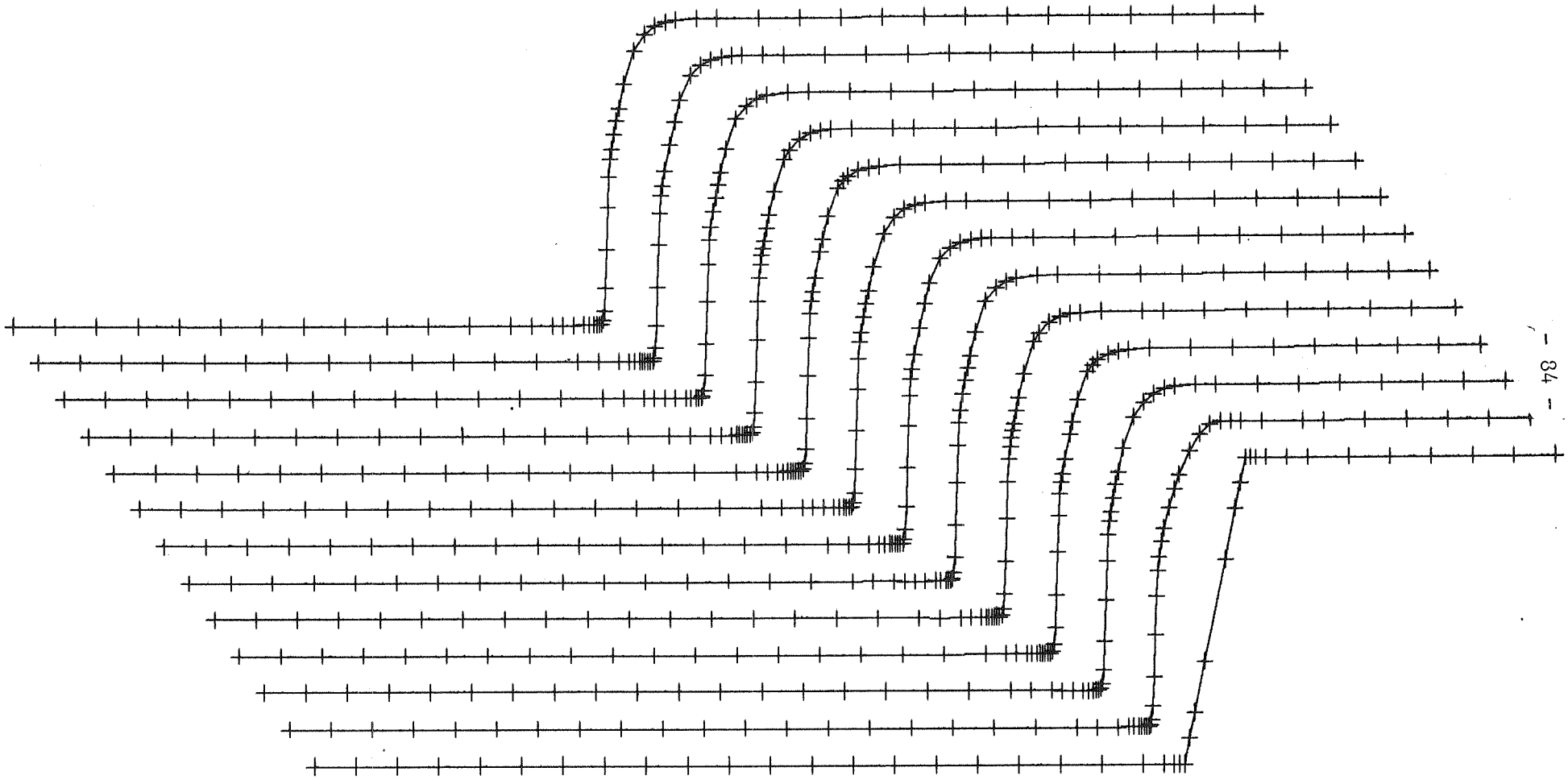
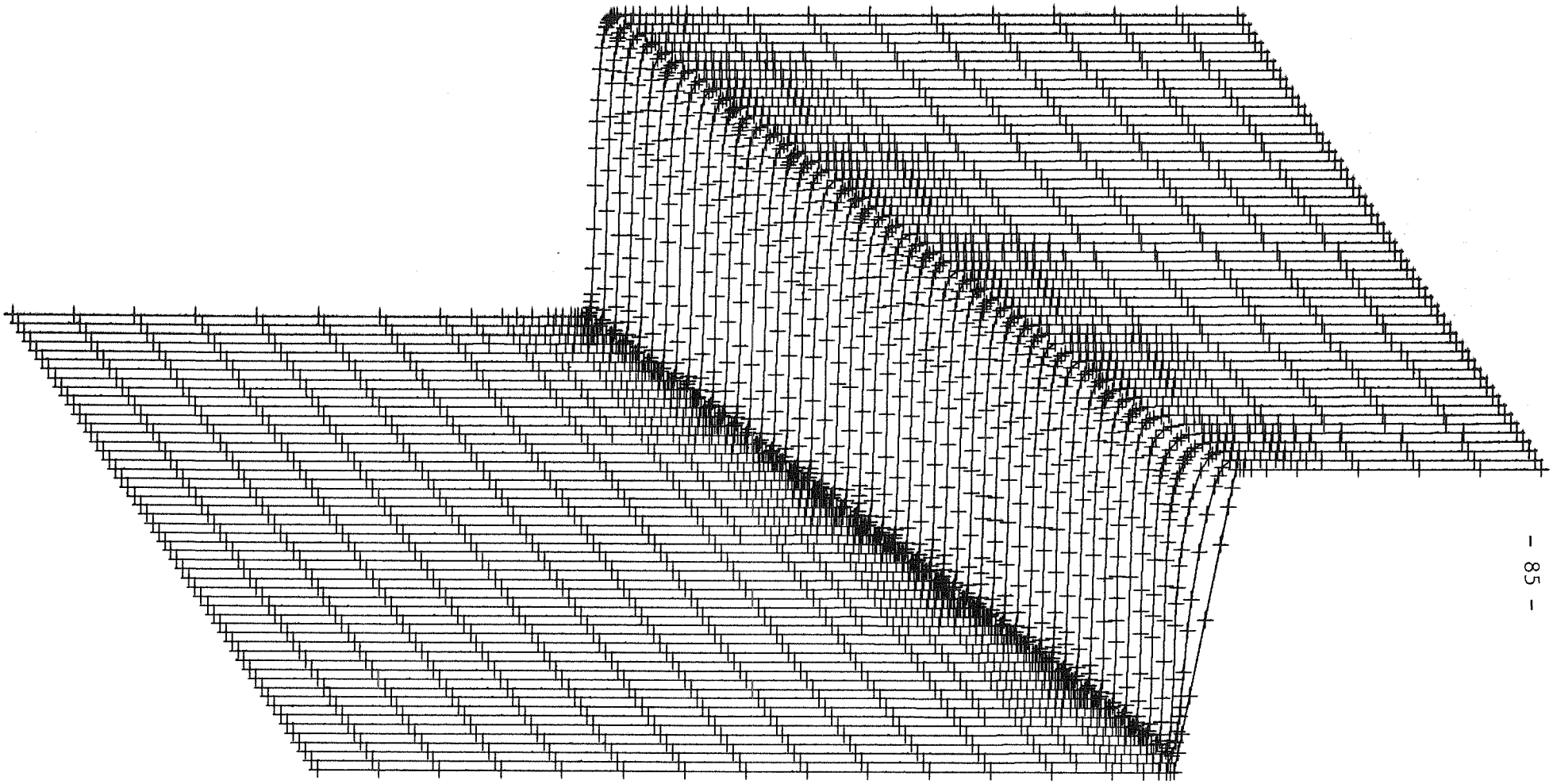


Figure 3.3.6



Burgers Equation
eps = 0.00100
k = 0.08000

Figure 3.3.7



Burgers Equation
eps = 0.00100
k = 0.02000

Figure 3.3.8

3.4 Numerical Example: Resolution of Moving Shocks for Burgers' Equation

In this section a procedure is discussed by which moving profile solutions of (3.3.1) can be resolved. As was suggested in the last section, if each travelling wave in the solution is computed in a local moving coordinate system in which the travelling wave appears stationary, then it will be possible to resolve the travelling waves because we will be computing solutions to the correct steady-state equation in the limit of spatial mesh-width going to zero. Even before a steady travelling wave has formed in the solution, it is a good idea for reasons of accuracy to use a moving coordinate system in regions of high computational error. The speed of the moving coordinate system in such a case can be taken to be one such that in the resulting coordinate frame, some measure of the change in the solution with time is minimized. By minimizing the change in the solution we will also tend to reduce the truncation error and hence the computational error associated with the time stepping procedure.

Briefly, the procedure for taking one time step can be described as follows: An initial calculation is done on a coarse mesh on the entire interval. Then the error is estimated and subintervals are defined on which that error is unacceptably large. On each of the subintervals thus defined we recompute the solution using a combination of mesh refinement and a moving coordinate system to reduce the error to an acceptable size. We will first discuss some of the ideas behind this procedure in a general way, and then will give the specific algorithm.

The local regions in the solution that might be improved by using a locally moving coordinate system can be chosen to be regions in which the estimated error after an initial computation is higher than some acceptable value. Thus, the same method used to define regions where the mesh should be refined can be used to decide where a moving coordinate system should be employed to improve the computed solution. We proceed as before, estimating the error and determining regions where the solution needs to be improved but with the difference that we attempt to improve the solution by simultaneously refining the mesh and adjusting the local speed of the coordinate system.

At each time step, rather than using the relatively expensive implicit method over the whole interval as we did in the last section, we will first take a time step using an explicit method on a relatively coarse mesh. This coarse mesh is chosen so that the smooth parts of the solution will be well-resolved on it (while the regions of rapid change will not necessarily be resolved). After taking a time step from t to $t + k$, for example, we estimate the error in the

computed solution on the coarse mesh at time $t + k$ and choose regions where this error is higher than is acceptable. Since the method used to compute this preliminary solution was explicit, we know that there is a finite signal speed associated with it, and thus we can determine the domain of dependence at time t corresponding to each region at time $t + k$ in which the estimated error is too large, i.e. we can isolate the region at time t that produced the part of the solution at time $t + k$ with unacceptably high error. We will then take each such subregion at time t separately, and use the solution at time t on that region to recompute the unacceptable parts of the solution at time $t + k$. This time, however, we will use the implicit method on each local region and will also use mesh refinement and a moving coordinate system to reduce the error.

In order to recompute the solution using the implicit method, we will need boundary conditions for each (local moving) subinterval. These can be safely gotten from the initial solution in which we used the explicit scheme on the coarse mesh, because those boundary conditions come from parts of that initial solution where we judged the error to be of acceptable size for the final solution.

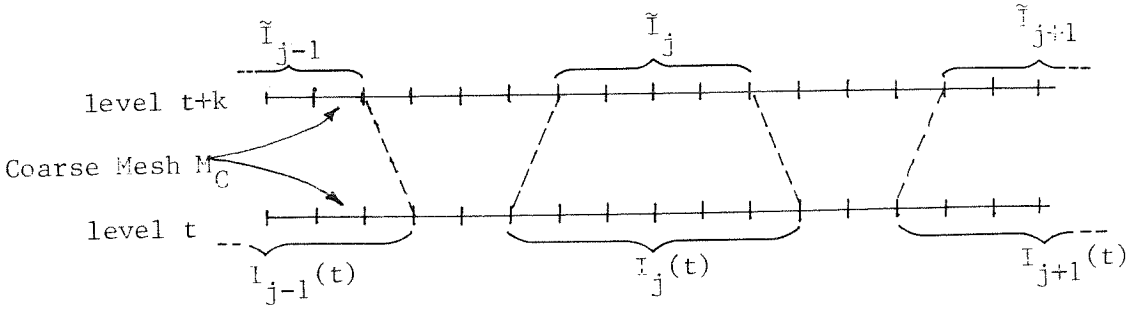
In determining the final refined mesh on each subinterval we recompute the solution several times, re-estimating the error after each computation and then adding or deleting points from the mesh as appropriate so that the error will be reduced when the solution is recomputed. We have chosen to estimate the speed of the subinterval within this mesh iteration procedure. After each intermediate solution $u(x,t+k)$ is computed from the solution at time t, $u(x,t)$, we choose the speed for the moving coordinate system by minimizing the discrete L_2-norm of the change in the solution with respect to the speed c, i.e. we seek an approximate solution c to

$$\min_c \| u(x,t+k) - u(x-ck,t) \|_{L_2}$$

(This is similar to an idea suggested by Hyman [1981].) This speed is then used for the next solution computation in the mesh iteration.

Let us formalize this procedure: For the duration of the calculation, define an underlying nonmoving coarse mesh $\mathbf{M}_C$ on the total interval of interest I_C. This mesh is chosen in such a way that the smooth features of the solution can be well resolved on it. Assume the solution of the differential equation (3.3.1 for example) has been calculated accurately up to time t. By this we mean that the solution has been well-resolved at time t. (If we need values of the solution at time t which are not on the computational mesh, then these can be found

accurately by interpolation.) Call this resolved solution $u_\infty(x,t)$. Take $u_C(x,t)$ to be the restriction of this solution on the coarse mesh $\mathbf{M}_C$. Now use an explicit difference approximation (we call this the "coarse mesh scheme") and take a time step of length k to get $u_C(x,t+k)$. (For a differential equation such as (3.3.1), ε is a small parameter and so an explicit method with a time-step restriction that is reasonable can be found.) Now estimate the error at each point of $u_C(x,t+k)$ and define intervals $\tilde{I}_j$, $j = 1,2,\ldots$, where the error is larger than is acceptable. The domain of dependence $I_j(t)$ at time level t can be found for each $\tilde{I}_j$ since the coarse mesh method is explicit. (See diagram below.) For each j we define $I_j(t+k)$ to be is same interval shifted to the right or left a distance depending on the speed which we will subsequently estimate for the interval. (See second diagram below.)

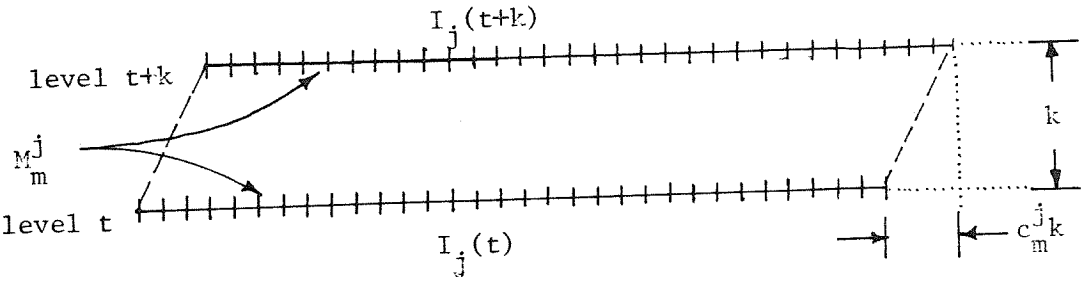


Let $\mathbf{M}_0^j$ be the initial fine mesh at time $t+k$ in the jth interval. It is defined to be the part of the coarse mesh $\mathbf{M}_C$ included in $I_j(t+k)$. Let $u_0^j(x,t)$ and $u_0^j(x,t+k)$ be some representation on this "fine" mesh of $u_C(x,t)$ and $u_C(x,t+k)$ respectively. Now begin the fine mesh iteration: Denote by m the number of times we have refined the mesh at this time step. Then starting with $m = 0$, we determine a speed c_m^j for each interval I_j by approximately minimizing the expression

$$\| u_m^j(x,t+k) - u_m^j(x - c_m^j k, t) \|_{L_2}$$

(i.e. we find a local moving coordinate system in which the solution changes the least measured in the L_2 norm). We then define a new mesh (at $t+k$) $\mathbf{M}_{m+1}^j$ by estimating the error in $u_m^j(x,t+k)$ and adding or deleting points from $\mathbf{M}_m^j$. Recall that the error is estimated by looking at the lower order divided differences of the computed solution as described in chapter 2. We then take a step with the implicit fine mesh scheme in the moving coordinate system of I_j.

The boundary conditions for the interval are interpolated from the coarse mesh solution $u_C(x,t+k)$.



We then repeat this fine mesh iteration until the mesh converges (i.e. until $\mathbf{M}^j_{m+1} \equiv \mathbf{M}^j_m$).

Once this procedure has been done for every I_j, we define the discrete function $u_\infty(x,t+k)$ to be made up of the final values on each I_j combined with the values of $u_C(x,t+k)$ on $\overline{I_C \cap (\bigcup_j I_j)}$ (the part of I_C not included on any fine mesh interval I_j). We can now take another time step using the same procedure. Note that at each time step, the local moving subintervals I_j are completely redetermined. Thus the intervals at a given time step may not correspond to the intervals at the preceding or subsequent time step, although in actual computations they will tend to do so. This allows for the subintervals of refinement to appear and disappear as the local smoothness of the solution changes.

This procedure was coded for Burgers' equation and some examples are presented below. I used Lax-Wendroff for the coarse mesh scheme and the method presented in section 3.3 for the fine mesh scheme. In the Lax-Wendroff part of the solution no attempt was made to approximate the dissipative term εu_{xx} of Burgers' equation. Since Lax-Wendroff is a dissipative scheme and ε is very small, the artificial dissipation in the method would swamp any accurate approximation of this term. Lax-Wendroff can be written as a two step scheme as follows: (Richtmyer and Morton [1967]) We introduce a uniform coarse mesh with grid points x_ν, $\nu = 0,1,\ldots,N$ and a uniform meshwidth h. The solution $u_\nu(t+k)$ after a single time step is given by

$$\tilde{u}_\nu(t+k) = \frac{1}{2}(u_\nu(t) + u_{\nu-1}(t)) + \lambda(f(u_\nu) - f(u_{\nu-1})) \qquad \nu = 1,2,\ldots,N$$

$$u_\nu(t+k) = u_\nu(t) + \lambda(f(\tilde{u}_{\nu+1} - f(\tilde{u}_\nu)) \qquad \nu = 1,\ldots,N-1$$

Here $\lambda = k/h$, and the differential equation being approximated is

$$u_t = f(u)_x$$

All interpolation was done linearly when values of the solution were needed at level t that had not been calculated by the difference method. The minimization process in which the local speed of the solution is estimated was done by a "Golden Section search" technique. (This is certainly not the most efficient approach). At each time step, only the difference in speed from that at the previous time step was estimated.

In the first example, initial data with constant endstates and a "ramp" connecting them were given:

$$u(x,0) = \begin{cases} 0 & \text{if } -1 \le x \le .4 \\ 20.(x-.4) & \text{if } .4 < x < .5 \\ 2 & \text{if } .5 \le x \le 1 \end{cases}$$

The boundary conditions were $u(-1,t) = -1.$, $u(1,t) = 1$. The coarse mesh width was $h = 1/10$, the time step was $k = 1/20$ and $\varepsilon = .0025$. In figure 1, the solution was computed using the method of section 3.3, i.e. no moving coordinate system was used. Note that although the transition region is moving at the correct speed of -1, the width of the transition region is much greater than the expected $O(\varepsilon)$. Figure 2 shows the same problem with the moving mesh code implemented. The circles indicate the ends of the fine mesh interval. The transition is much better resolved in this calculation. The correct speed for the shock is -1 for this problem. The estimated speeds for the moving coordinate system were -1.031, -.992, -1.005, -1.000, -1.000, -1.001, -1.000, -1.000, -1.000, -1.000, -1.000, -1.001, -1.001, -.999, -.999, -1.000, -1.000, -.999, -1.000, and -.999 for time steps 1 through 20 respectively. In the speed estimation procedure the golden section search was considered to have converged if two successive iterates agreed to within 10^{-3}, so these results indicate that once a steady moving profile has formed, the speed estimated is the correct shock speed.

For the next example the following initial data were used:

$$u(x,0) = \begin{cases} -.1 & \text{if } -1.0 \le x \le -.1 \\ 20.(x+.1) - .1 & \text{if } -.1 < x \le 0. \\ 1.9(1-x) & \text{if } 0. \le x \le 1. \end{cases}$$

In the solution of this problem the a shock forms on the left and a rarefaction wave forms on the right. I took $k = 1/20$, $h = 2/21$ and $\varepsilon = 5\times10^{-4}$. These initial data are interesting because the speed of the resulting shock transition changes as the rarefaction wave interacts with the shock, so the method has to be able to catch this. By inspection of the computed profiles in figure 3 we see that the shock profiles are sharp and so the shock speed has been estimated well.

Figure 4 shows an example in which time-varying boundary conditions were used so that the "smooth" part of the solution would not be a constant or linear function. Except at one time step, the error estimation procedure has decided automatically that the Lax-Wendroff solution is accurate enough to be used in the region away from the shock.

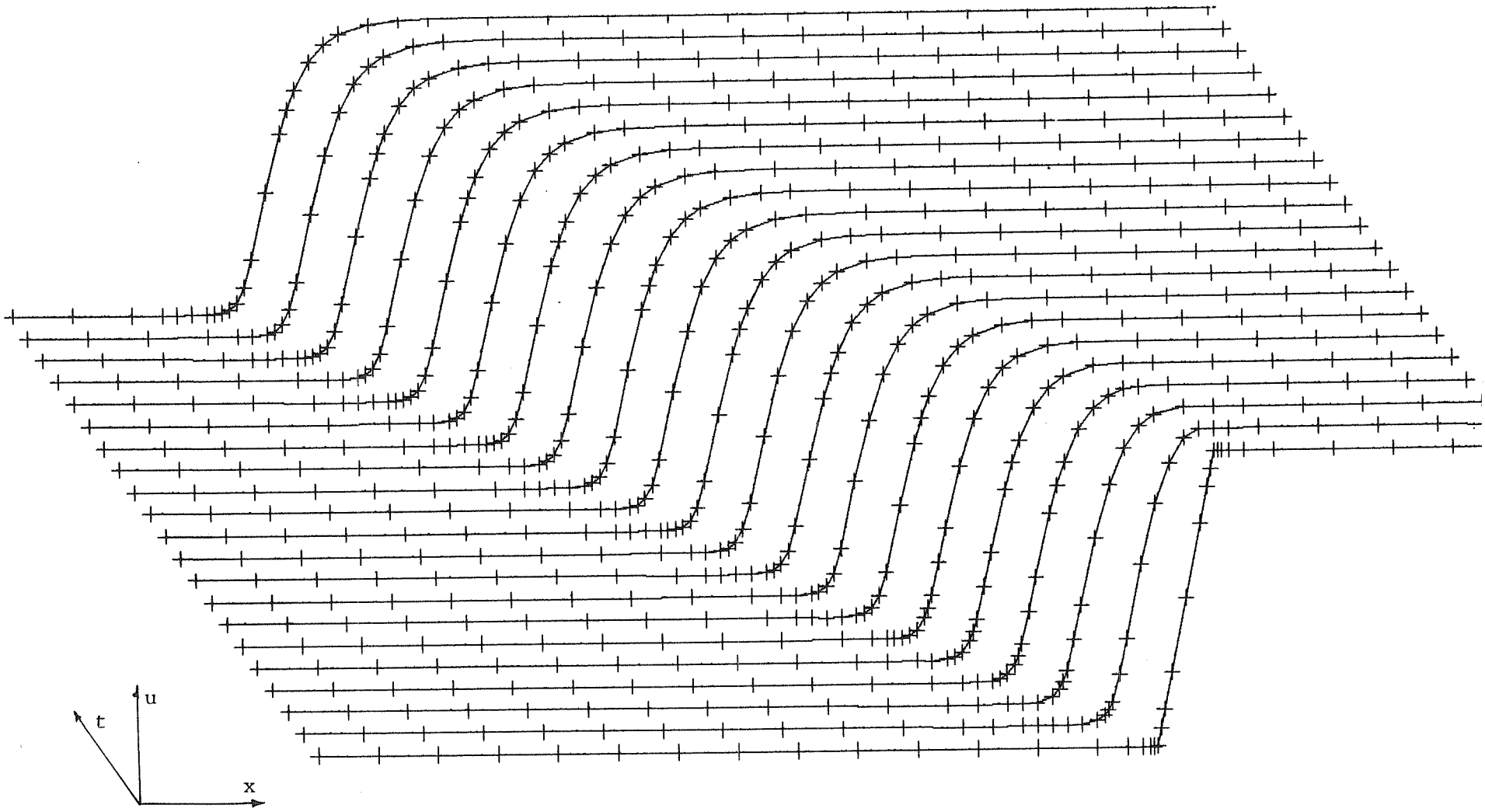
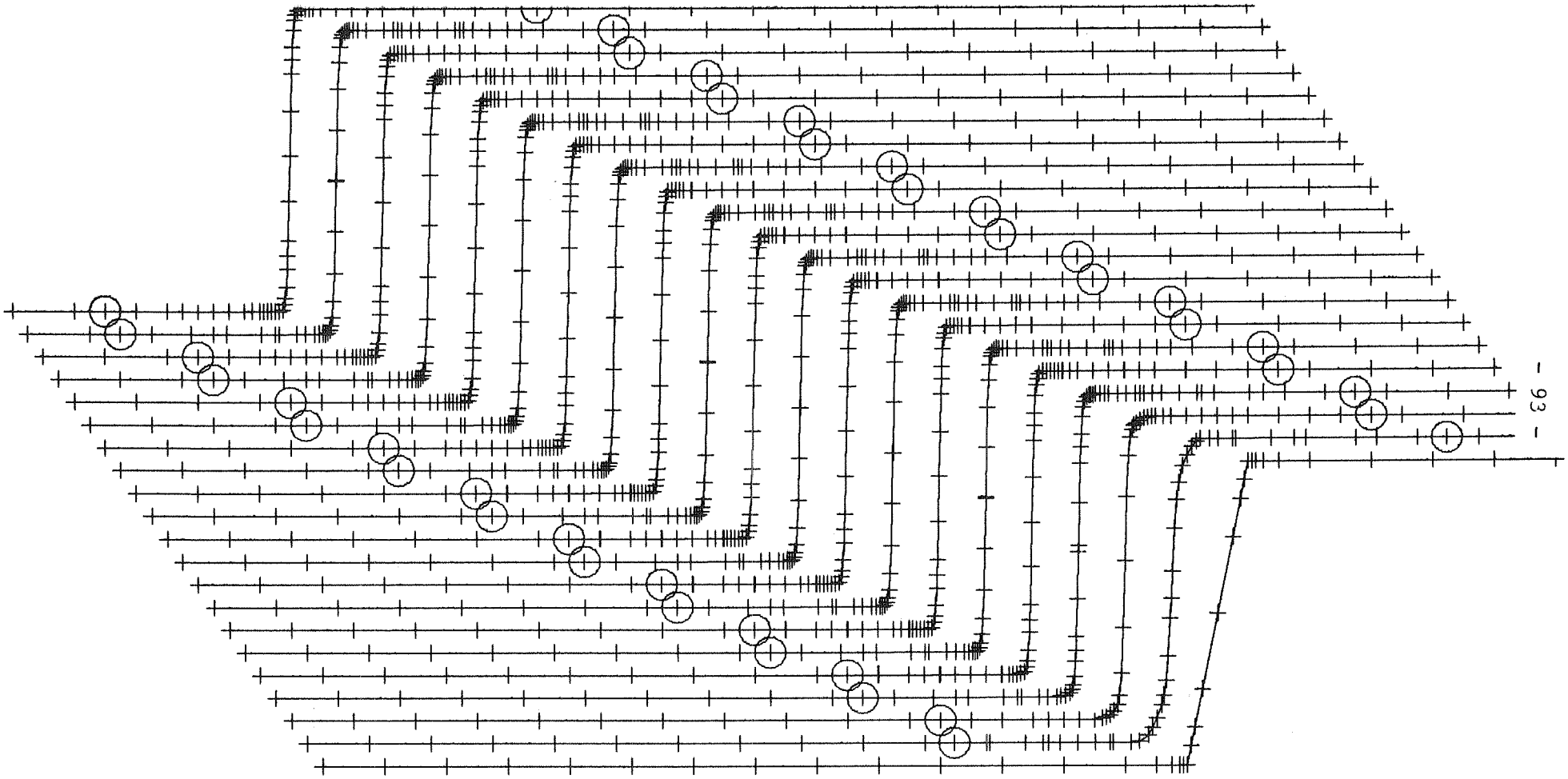


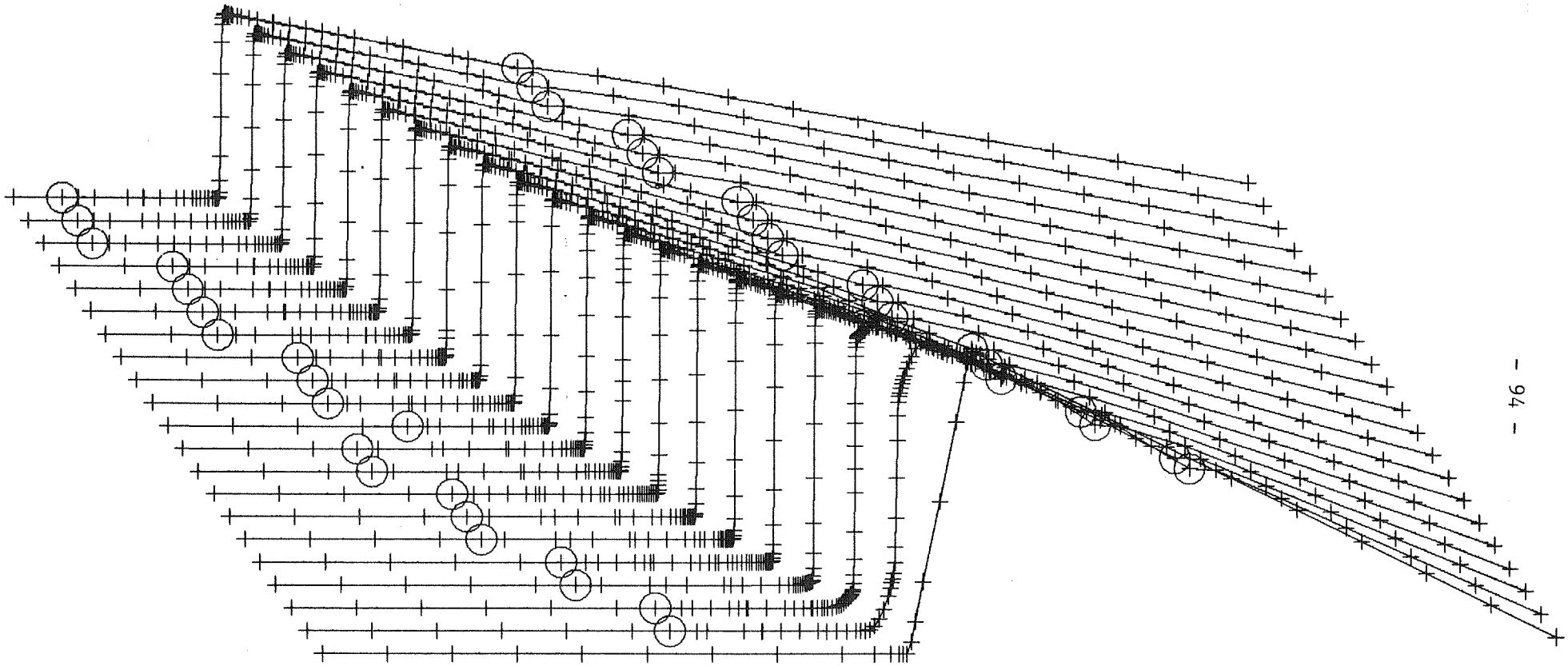
Figure 3.4.1



k = 0.05000
eps = 0.00250
lambda = 0.50000

Figure 3.4.2

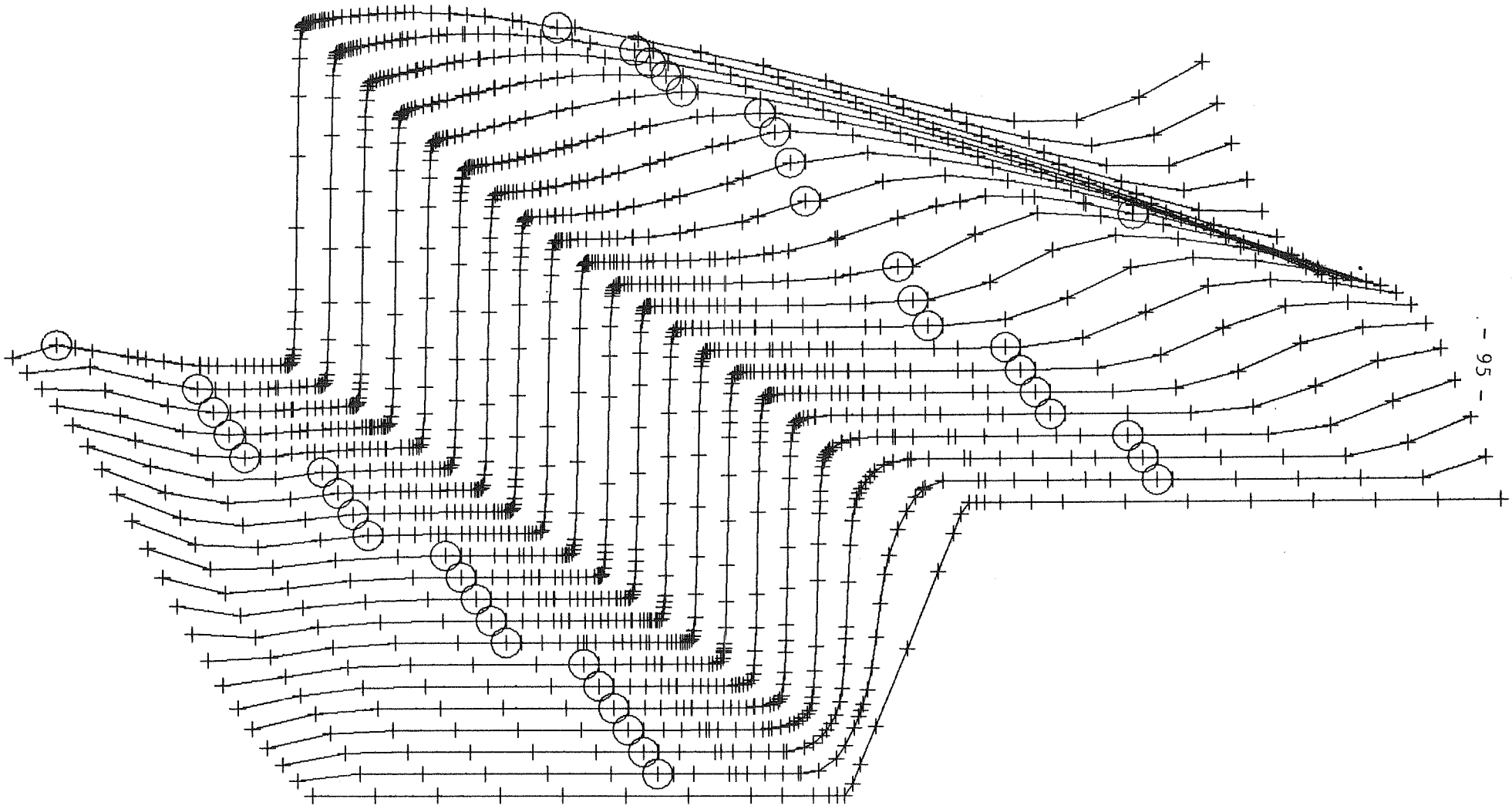
Moving Coordinates



k = 0.05000
eps = 0.00050
lambda = 0.47500

Figure 3.4.3

Moving Coordinates



k = 0.05000
eps = 0.00250
lambda = 0.47500

Moving Coordinates

Figure 3.4.4

3.5 Example: Stationary Shocks in Isentropic Gas Dynamics

The equations of gas dynamics for a polytropic gas can be written as follows in conservation form:

$$\begin{aligned} &\rho_t + (\rho u)_x = 0 \\ &(\rho u)_t + (\rho u^2 + (\gamma - 1)e)_x = \varepsilon u_{xx} \\ &(\tfrac{1}{2}\rho u^2 + \rho e)_t + [(\tfrac{1}{2}u^2 + e)\rho u + (\gamma - 1)e\rho u]_x = \varepsilon(uu_x)_x + O(\lambda) \end{aligned} \tag{1}$$

where $\rho = \rho(x,t)$ is the density, $u = u(x,t)$ the velocity and $e = e(x,t)$ the internal energy of the fluid. The ratio of specific heats γ is a constant for a given gas and is of order unity. The small parameters $0 \le \varepsilon \ll 1$ and $0 \le \lambda \ll 1$ are the coefficients of viscosity and thermal conductivity of the gas. (ε is here 4/3 times the usual coefficient of viscosity μ.) By linear combination of the equations (1), the third equation can be reduced to the form

$$s_t + us_x = \frac{\varepsilon}{\rho^\gamma}(u_x)^2 + O(\lambda) \tag{2}$$

where $s := e/\rho^{\gamma-1}$ is proportional to e^S, S being the entropy of the fluid. If an initially uniform state and weak shocks are assumed, then to a good approximation, equation (2) can be replaced by

$$s \equiv s_0 = \text{const.} \tag{3}$$

This is the isentropic assumption. Equations (1) then become

$$\begin{aligned} &\rho_t + (\rho u)_x = 0 \\ &(\rho u)_t + (\rho u^2 + \vartheta\rho^\gamma)_x = \varepsilon u_{xx} \end{aligned} \tag{4}$$

where $\vartheta := (\gamma - 1)s_0$.

Equation (4) can now be treated as a singular perturbation problem. In order to apply the theory of chapter 2, we rewrite (4) as a first-order system of ordinary differential equations by replacing the time derivative with the Backward Euler approximation and introducing an auxiliary variable $y(x)$:

$$\begin{aligned} &(\hat{\rho}\hat{u})_x + (\hat{\rho} - \rho^*)/k = 0 \\ &y_x + (\hat{\rho}\hat{u} - \rho^*u^*)/k = 0 \\ &\varepsilon\hat{u}_x - \hat{\rho}\hat{u}^2 + y - \vartheta\hat{\rho}^\gamma = 0 \end{aligned} \tag{5}$$

Here $\hat{\rho} = \hat{\rho}(x)$ and $\hat{u} = \hat{u}(x)$ are approximations to $\rho(x,t)$ and $u(x,t)$ respectively, while $\rho^* = \rho^*(x)$ and $u^* = u^*(x)$ are approximations to $\rho(x,t-k)$ and $u(x,t-k)$. The parameter k is the time step.

Since equations (5) will give rise to an implicit set of equations to solve when the x-derivatives are replaced by a difference approximation, we first linearize the equations. The resulting equations are then iterated to convergence. Let the nth iterate be denoted by $\mathbf{u}^n$ and let $\tilde{\mathbf{u}} := \mathbf{u}^{n+1} - \mathbf{u}^n$. Also for notational convenience let $\bar{\mathbf{u}} := \mathbf{u}^n$. Then the iteration scheme is given by

$$\left\{\begin{bmatrix}\bar{u} & 0 & \bar{\rho}\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\tilde{\mathbf{u}}\right\}_x + \left\{\frac{1}{\varepsilon}\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ -(\bar{u}^2+\bar{c}^2) & 1 & -2\bar{\rho}\bar{u}\end{bmatrix} + \begin{bmatrix}1/k & 0 & 0\\ \bar{u}/k & 0 & \bar{\rho}/k\\ 0 & 0 & 0\end{bmatrix}\right\}\tilde{\mathbf{u}}$$

$$= \begin{bmatrix}(\rho^* - \bar{\rho})/k\\ (\rho^* u^* - \bar{\rho}\bar{u})/k\\ \bar{\rho}\bar{u}^2 + \vartheta\bar{\rho}^\gamma - \bar{y}\end{bmatrix} - \begin{bmatrix}\bar{\rho}\bar{u}\\ \bar{y}\\ \varepsilon\bar{u}\end{bmatrix}_x \tag{6}$$

where $\tilde{\mathbf{u}} := (\tilde{\rho}, \tilde{y}, \tilde{u})^T$, $\bar{\mathbf{u}} := (\bar{\rho}, \bar{y}, \bar{u})^T$ and $\bar{c}^2 := \gamma\vartheta\bar{\rho}^{\gamma-1}$. With the obvious definitions, (6) is of the form

$$(H(\bar{\mathbf{u}})\tilde{\mathbf{u}})_x + (\frac{1}{\varepsilon}A_0 + A_1)\tilde{\mathbf{u}} = \mathbf{g}(x) + \mathbf{f}(x)_x \tag{7}$$

In terms of the dependent variables defined by $\tilde{\mathbf{w}} := H(\bar{\mathbf{u}})\tilde{\mathbf{u}}$ and defining $\tilde{A}_j := A_j H^{-1}$, $j = 0,1$, (7) can be rewritten as

$$\mathbf{w}_x + (\frac{1}{\varepsilon}\tilde{A}_0 + \tilde{A}_1)\mathbf{w} = \mathbf{g}(x) + \mathbf{f}(x)_x \tag{8}$$

where both $\tilde{A}_0$ and $\tilde{A}_1$ are $O(1)$. Explicitly,

$$\tilde{A}_0 := \begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ -(\bar{u}^2+\bar{c}^2)/\bar{u} & 1 & \rho(\bar{c}^2 - \bar{u}^2)\end{bmatrix}$$

so since equation (8) is in the form of equation (2.4.1) the turning points of the system are characterized as points where the one non-zero eigenvalue $\kappa := -\bar{\rho}(\bar{u} - \bar{c})(\bar{u} + \bar{c})$ of A_0 vanishes. Thus we will use the difference method

$$\Delta_+(\mathbf{w}_\nu - \mathbf{f}_\nu) + h_\nu(I - \Psi_\nu)\tilde{\mathbf{F}}_{\nu+1} + h_\nu\Psi_\nu\tilde{\mathbf{F}}_\nu = 0 \tag{9}$$

where

$$\tilde{\mathbf{F}}_\nu := \tilde{\mathbf{F}}(x_\nu) := (\frac{1}{\varepsilon}\tilde{A}_0(x_\nu) + \tilde{A}_1(x_\nu))\mathbf{w}_\nu - \mathbf{g}(x_\nu)\,,$$

$\mathbf{f}_\nu := \mathbf{f}(x_\nu)$, h_ν is the local meshwidth, $\mathbf{w}_\nu$ is an approximation to $\mathbf{w}(x_\nu)$ and $\Psi_\nu := \begin{bmatrix} 1/2 & & \\ & 1/2 & \\ & & \alpha_\nu \end{bmatrix}$. The parameter $\alpha_\nu = \alpha_\nu(\kappa)$ is to be chosen according to the sign and size of κ. We use the choice of Kreiss and Kreiss [1981] which is described in section 2.1 of this thesis. To get the appropriate values for α_ν, replace the expression "Re$a_{ii}(x_\nu)$" with $-\kappa(x_\nu)$ in the formulas given following equation 2.1.3.

This difference method can then be applied together with a mesh refinement procedure to resolve the solutions of (4). As with the example of Burgers' equation (sections 3.3 and 3.4) we expect to be able to resolve steady shocks well, while moving shocks will be smeared. This effect can again be remedied by introducing locally moving coordinate systems.

Before giving numerical examples, we make the following remarks:

Remark 1: There is no need to compute with the variables $\mathbf{w}$ in equation (8). It is clear by inspection of (9) that the variables $\mathbf{u}$ can be used with no problem, i.e. we compute using

$$\Delta_+(H_\nu \mathbf{u}_\nu - \mathbf{f}_\nu) + h_\nu(I - \Psi_\nu)\mathbf{F}_{\nu+1} + h_\nu \Psi_\nu \mathbf{F}_\nu = 0 \tag{10}$$

where

$$\mathbf{F}_\nu := (\frac{1}{\varepsilon}A_0(x_\nu) + A_1(x_\nu))\mathbf{u}_\nu - \mathbf{g}(x_\nu) .$$

The construction of equation (8) was only necessary in order to determine the eigenvalue $\kappa(x_\nu)$.

Remark 2: The variable $\tilde{y}_\nu$ can be eliminated from equations (10) thus reducing the number of equations from $3N$ to $2N$, where N is the number of computational mesh points. This also has the effect of reducing the bandwidth of the system of equation to be solved from 7 to 5. The last two equations of (10) can then be replaced with the single equation

$$\Delta_+ \frac{\varepsilon}{h_\nu}\Delta_- u_\nu + (1 - \alpha_\nu)(\Delta_+ D_\nu - \frac{h_\nu}{2}(G_\nu + G_{\nu+1}))$$

$$+ \alpha_{\nu-1}(\Delta_- D_\nu - \frac{h_{\nu-1}}{2}(G_\nu + G_{\nu-1})) = 0 \tag{11}$$

where $u_\nu := \bar{u}_\nu + \tilde{u}_\nu$,

$$D_\nu := -((\bar{u}^2 + \bar{c}^2)\tilde{\rho} + 2\bar{\rho}\bar{u}\tilde{u} + \bar{\rho}\bar{u}^2 + \vartheta\tilde{p}')\Big|_{x = x_\nu}$$

and

$$G_\nu := (\bar{u}\tilde{\rho} + \bar{\rho}\tilde{u} + \bar{\rho}\bar{u} - \rho^* u^*)/\ k\Big|_{x = x_\nu}$$

Remark 3: The turning points of the system (6) are the places where the eigenvalue κ becomes zero, which are clearly at the (linearized) sonic points of the flow, $\bar{u} = \pm\, \bar{c}$. Since a steady shock in gas dynamics is characterized by having supersonic flow on one side and subsonic flow on the other, the scheme described above can be thought of as essentially a weighted upwinding scheme.

The scheme of Kreiss and Kreiss that we use is designed to work well in the regions of the solution away from the turning points, but from the discussion of Kreiss and Nichols [1975] would not be expected to perform well at the turning points unless an appropriate scaling of the independent variable x, (i.e. a mesh refinement) is made there. In practice, if a mesh refinement is not made, the solution typically exhibits overshoot at the transition region but elsewhere remains smooth, so we expect the method to perform well as the basic difference method in a solution adaptive mesh refinement procedure. Figures 1 show the initial conditions and the computed solution after 21 time steps of length k=.02 for the density ρ on a uniform mesh with 21 points. The endstates of the initial conditions were ρ =.4,.6 and u = 2.1,1.4, and so a steady shock is produced. The viscosity was $\varepsilon = 10^{-2}$ for this calculation. Note that, as expected for a one-sided scheme, the solution stays smooth except near the transition region.

Figures 2 show the initial conditions and computed solution at t = .92 for ρ with the solution adaptive mesh refinement implemented. In this example, the endstates are again ρ = .4,.6 and u = 2.1,1.4, $\varepsilon = 5\times10^{-4}$ and k = .02. The number of meshpoints used to define the initial data was 33 and the number of meshpoints used at t = .92 was 82, most of which are in the shock transition region.

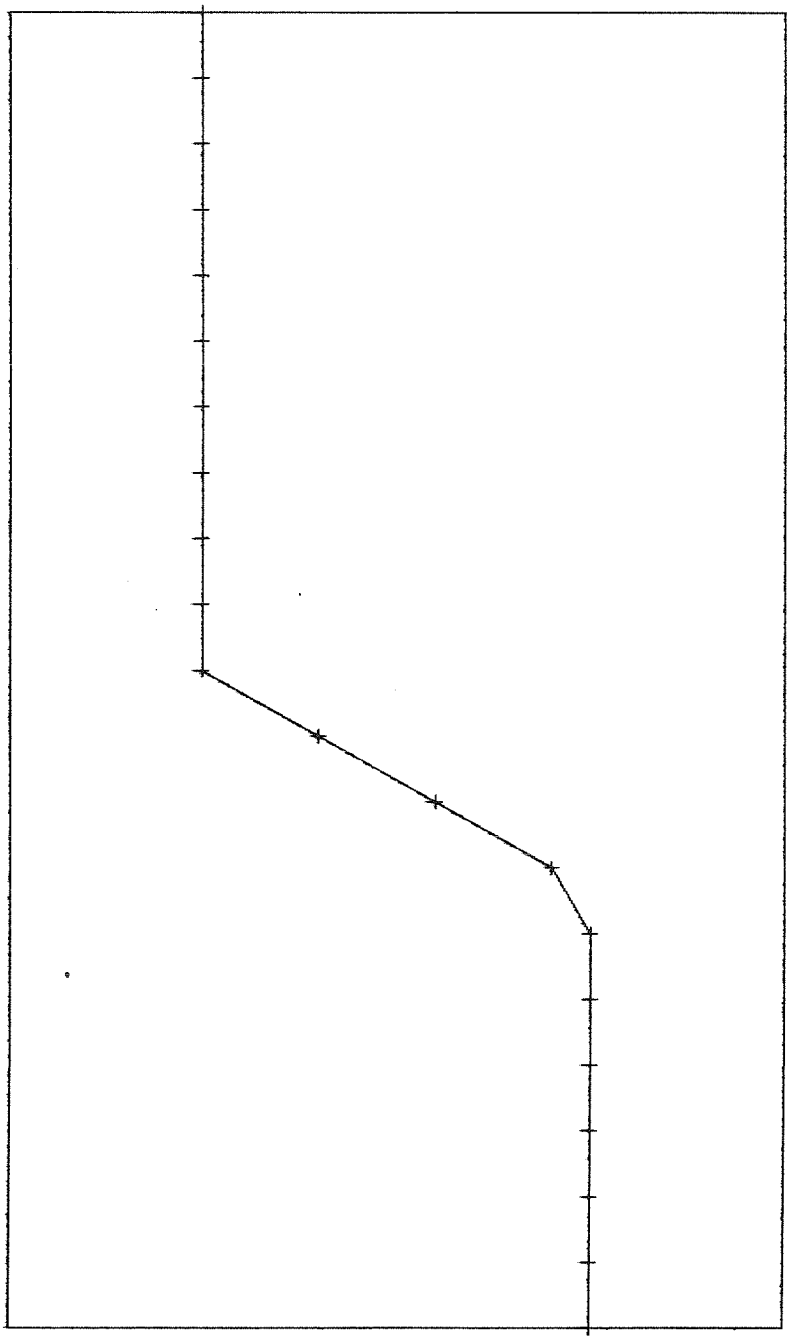


Figure 3.5.1a

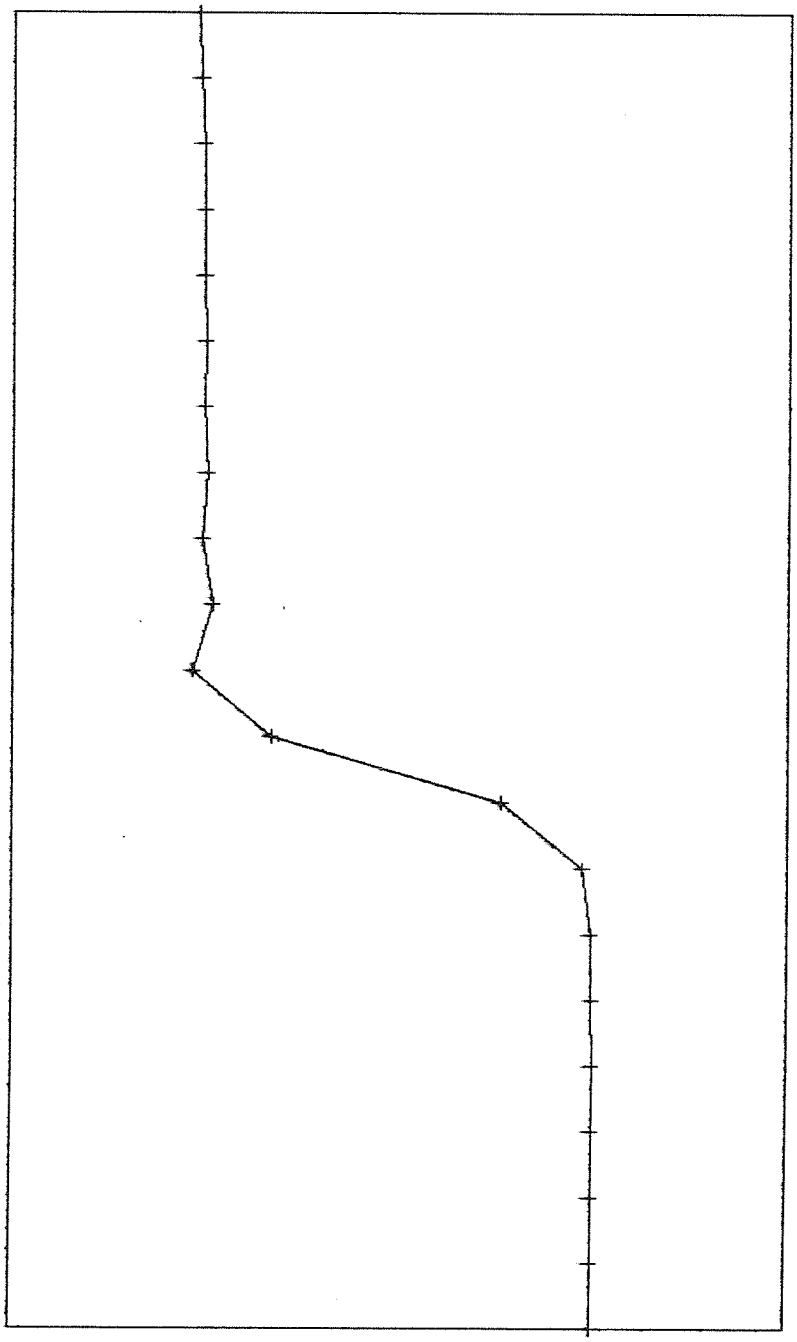


Figure 3.5.1b

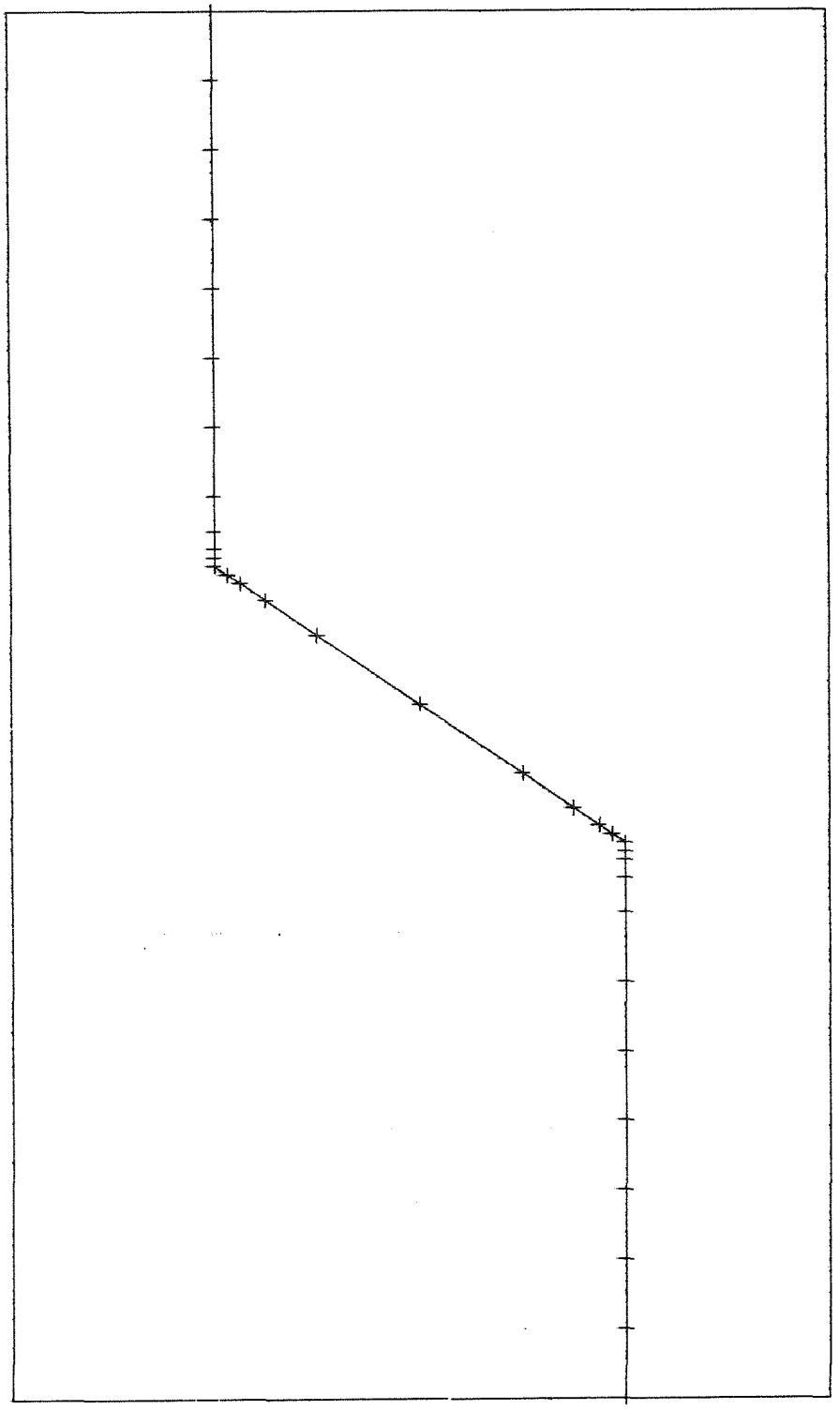


Figure 3.5.2a

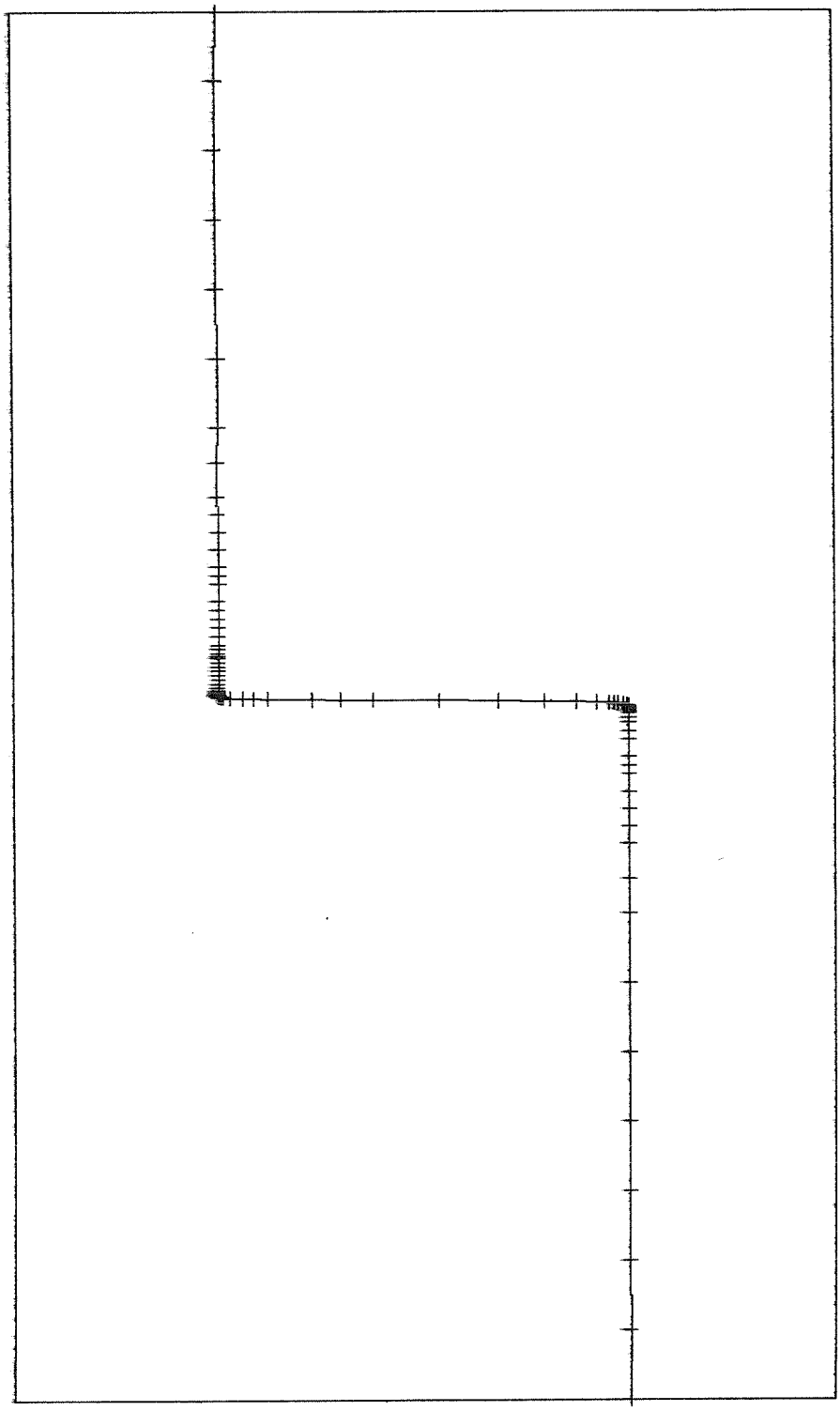
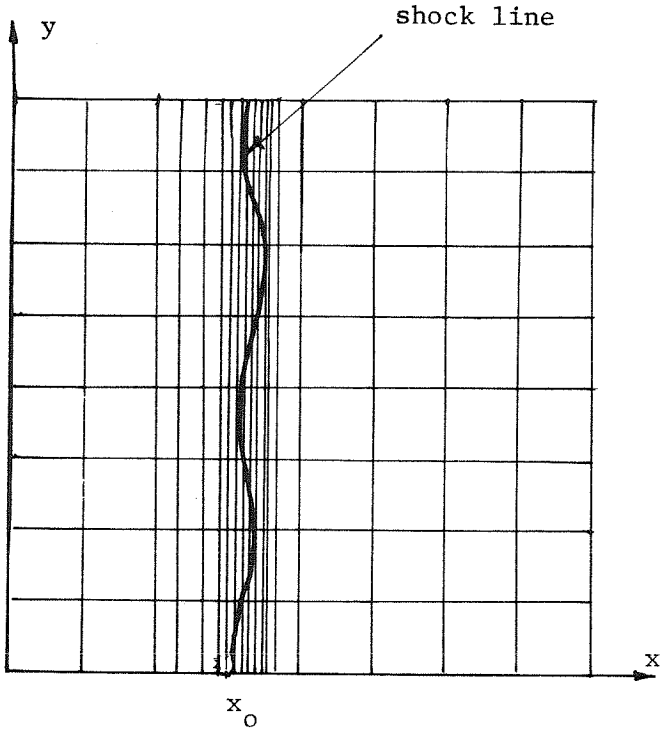


Figure 3.5.2b

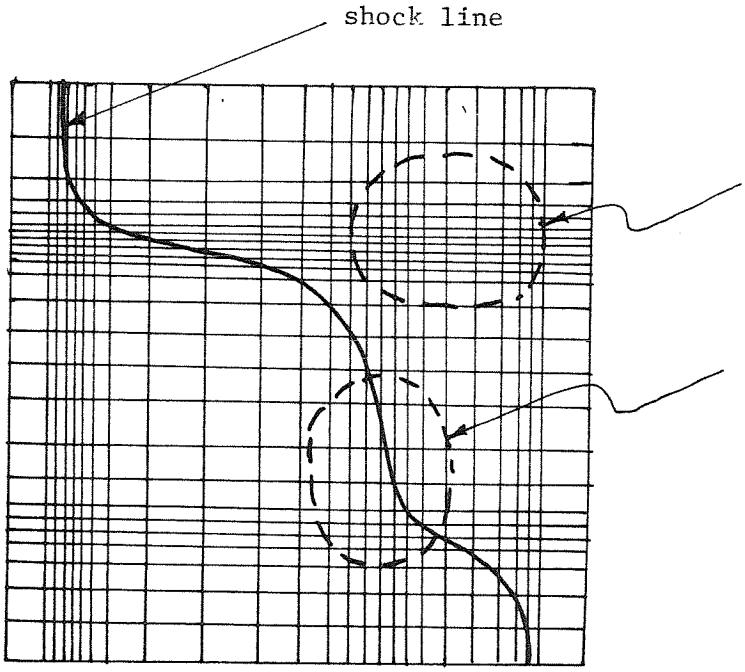
IV. Numerical Methods for Problems in More Than One Space Dimension

4.1 An Unconventional Approach to Splitting

In this section we will discuss an extension of the mesh refinement technique of chapter 3 to problems in two space dimensions. We are particularly concerned with computing accurate solutions of problems whose solutions exhibit rapid transitions that are essentially one-dimensional in nature such as shocks. For a problem in two dimensions in which the shock line is very nearly linear and oriented so as to be parallel to one set of coordinate lines, it is clear how to implement a mesh refinement. If, for example, the shock lies essentially parallel to a line $x = x_0$, a refinement in the direction normal to the shock (the x-direction) could be made (see figure 4.1.1a). No refinement would probably be necessary in the y-direction in this case. It is clear, however, that this will not always be true. We will not always have the freedom to choose the orientation of the computational mesh in such a way as to have "one-dimensional" rapid transitions oriented with the mesh. A nice feature of the mesh refinement indicated in figure 4.1.1a is that if on the coarse mesh there are N meshpoints in each direction (for a total of N^2 meshpoints), the number of points added in the mesh refinement is only $\mathbf{O}(N/\delta)$ where δ is a measure of the width of the rapid transition (see Swartz [1981]). If the shock line was not oriented with the mesh, then adding lines to refine the mesh would result in lines being added in both the x and y directions. The number of additional mesh points would then tend to be $\mathbf{O}(N^2/\delta)$. In particular, we would also be refining the mesh in regions where the solution is smooth (see figure 4.1.1b). This large number of added points in the mesh can clearly be reduced if we truncate the added lines so that they do not extend into smooth parts of the solution (see figure 4.1.2a). The reduction in the

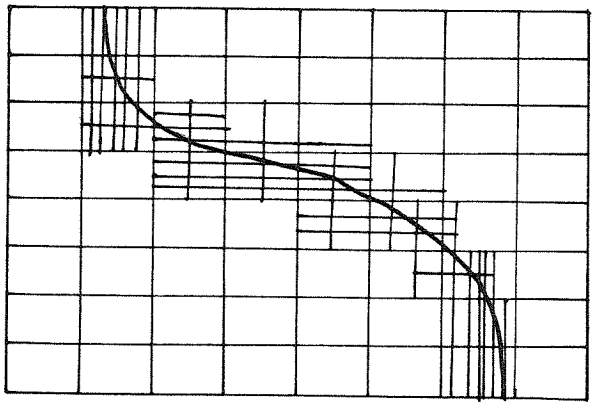


(a)

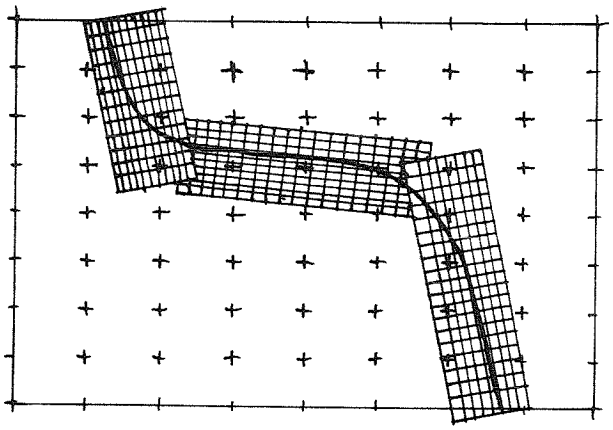


(b)

Figure 4.1.1



(a) Truncated mesh lines



(b) Oliger, Berger and Gropp method of

(c) method of B. Kreiss

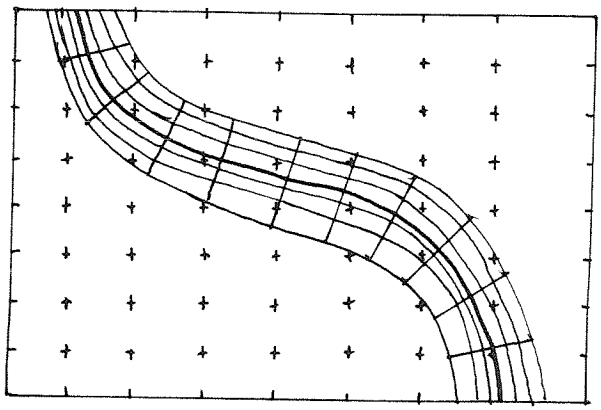


Figure 4.1.2

number of meshpoints will, however, be at the expense of programming complexity. A more rational approach to local mesh refinement is to embed local oriented grids in the coarse grid and interpolate between the different grids when solving the differential equations. Oliger [1981], Berger [1982] and Gropp [1980] have used local oriented rectangular moving grids to accomplish this. Figure 4.1.2b illustrates the basic idea. With a simple extension of the grid generation approach of B. Kreiss [1981], curvilinear grids could also be embedded in the coarse rectangular grid in such a way as to resolve a shock (see figure 4.1.2c). Again interpolation would be used to connect the solutions on the two grids together. One feature that would be common to solution adaptive mesh procedures that refined in two dimensions using one of these last three techniques is that in order to do the mesh refinement at each time step, we would have to look at the global solution at that time step. In this section we will propose and discuss a method in which only local information about the solution is used for purposes of mesh refinement.

Let us consider the numerical solution of Burgers' equation in two space dimensions:

$$U_t + (\tfrac{1}{2}U^2)_x + (\tfrac{1}{2}U^2)_y = \varepsilon(U_{xx} + U_{yy}) \quad for\ -\infty < x < \infty,\ t > 0 \tag{1}$$

with $U(x,y,0) = u_o(x,y)$ given. The difference method we used for the one-dimensional case (see section 2.3) is implicit in the time direction. A convenient way to implement an implicit difference method in two dimensions is to use operator splitting: As in section 3.3, we begin by approximating the time derivative in (1) with the Backward Euler method, and in addition use operator splitting to reduce the computational problem to a sequence of one dimensional problems: We introduce an "underlying" coarse mesh $\{x_i,y_j\}_{0,0}^{N,M}$ and solve approximately the equations

$$\varepsilon\tilde{u}_{xx}(x,y_j) + f(\tilde{u}(x,y_j))_x - \frac{1}{k}\tilde{u}(x,y_j) = -\frac{1}{k}u(x,y_j,t-k) \quad for\ j = 1,2,\ldots,M-1$$

and (3)

$$\varepsilon u_{yy}(x_i,y,t) + f(u(x_i,y,t))_y - \frac{1}{k}u(x_i,y,t) = -\frac{1}{k}\tilde{u}(x_i,y) \quad for\ i = 1,2,\ldots,N-1$$

with initial conditions $u(x,y,0) = u_o(x,y)$. The method we use for for the approximate solution of each of equations (3) is the "method of positive type" with solution-adaptive mesh refinement described and used in sections 3.2 and

3.3.

Let us suppose that in the initial conditions $u_o(x,y)$ there is a region of rapid transition oriented obliquely to the mesh. We would begin by solving the first of equations (3) on each of the lines $y = y_j$, $j = 1,2,...,M-1$. Because of the rapid transition region, we would expect automatic refinement to occur so that the solution of each of those one-dimensional problems would be resolved. We then solve the second of equations (3) on each of the lines $x = x_i$, $i = 1,2,..,N-1$. We expect again that automatic mesh refinement will occur in the region near the rapid transition. Note, however, that the right-hand side of the equation for $u(x_i,y,t)$ depends on values of the computed solution $\tilde{u}(x_i,y)$ at the previous step. If points are added to the one-dimensional mesh between the coarse mesh lines $y = y_j$, this means that we will need values of $\tilde{u}$ at points where they have not been computed (see figure 4.1.3).

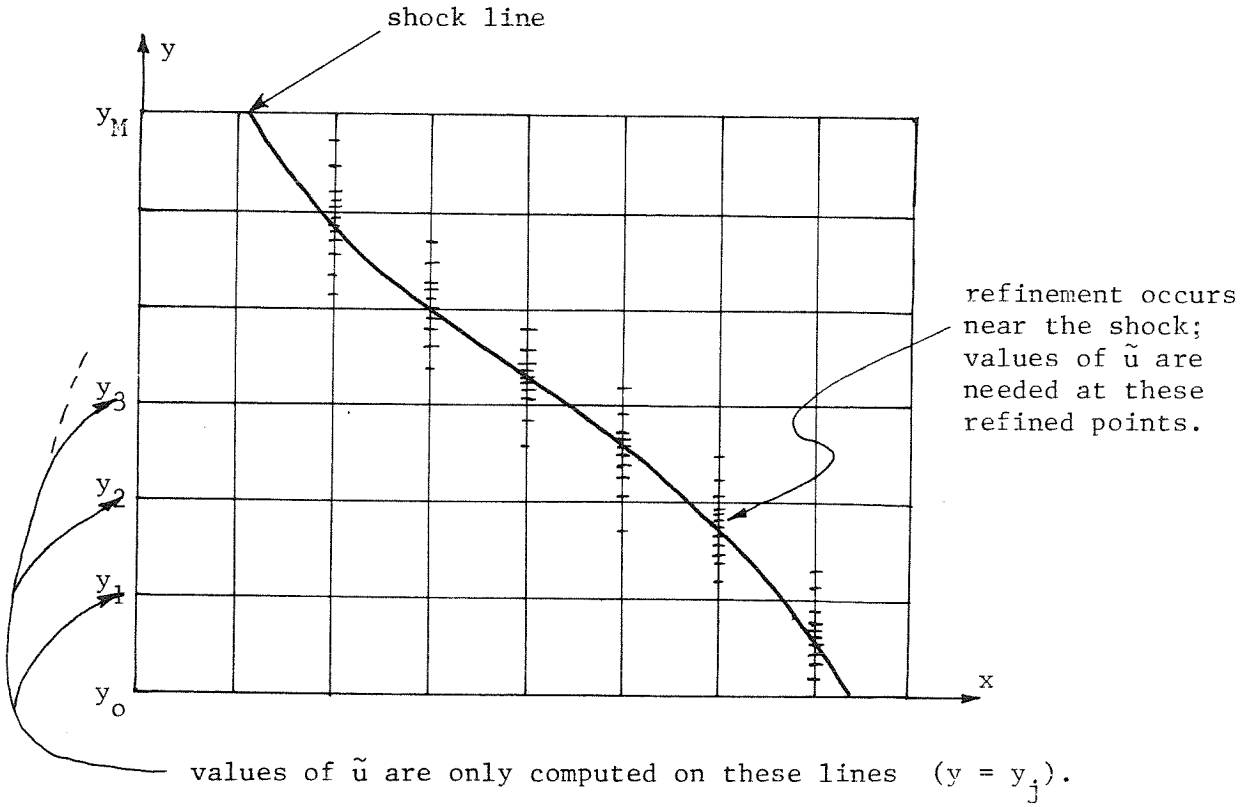


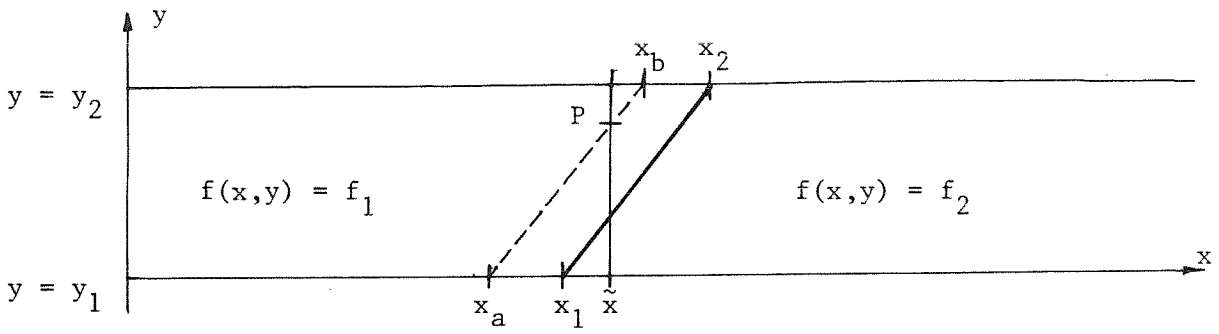
Figure 4.1.3

Recall that for the one-dimensional computation presented in chapter 3, we used linear interpolation with good results. Clearly for the two-dimensional case, we will also need to use some form of interpolation to get solution values at

the previous splitting step at the refined points. Simple linear interpolation will not work in general, however, because the solution being interpolated is not sufficiently smooth. L. Reyna [1982] has developed and implemented an interpolation procedure that is designed with the idea in mind of interpolating two-dimensional functions in which one-dimensional regions of rapid transition are present. Suppose that we know the values of a function on two sufficiently close parallel lines and we wish to find values of that function at some points between those lines. In addition, suppose that the function is known to exhibit one-dimensional regions of rapid transition. The need for a special interpolation method and an appropriate solution to that need can best be described using the following example. Let

$$f(x,y) = \begin{cases} f_1 & \text{if } x < 2(y - y_1) + x_2 \\ f_2 & \text{otherwise} \end{cases}$$

where $f_1 \neq f_2$, for $y_1 < y < y_2$.



We are interested in determining values $f(\tilde{x},y)$ for $y_1 < y < y_2$ when we only know the functions $f(x,y_1)$ and $f(x,y_2)$. (Here $x_1 < \tilde{x} < x_2$). Linear interpolation along the line $x = \tilde{x}$ will clearly give an inappropriate answer if $|f_1 - f_2|$ is large. Using simple linear interpolation along $x = \tilde{x}$ we will always get a value between f_1 and f_2, while the correct value should be *either* f_1 or f_2. To remedy this problem, we can eliminate the restriction that interpolation always be made along lines of constant x. We can get a reasonable value for f at the point P if, for example, we interpolate linearly along the straight line between (x_a,y_1) and (x_b,y_2). Reyna's procedure for finding a value of $f(x,y)$ between the lines $y = y_1$ and $y = y_2$ is to always interpolate along lines that do not intersect a

region of rapid transition in the function. The details of this procedure can be found in Reyna [1982] and Brown and Reyna [1982]. It is this method that we will use for interpolation in the two-dimensional problems.

Reyna's interpolation procedure works well as long as the regions of rapid transition of the function $f(x,y)$ do not lie nearly parallel to lines $y = const.$ In that case the interpolation procedure would have to use values of $f(x,y)$ that are separated by a large distance, most likely giving an unacceptably inaccurate interpolated value. In particular, if the region of rapid transition lies *parallel* to lines $y = const.$, it would be impossible to interpolate in this way. In the context of solving a differential equation in two space dimensions numerically, however, this case should not be a problem, because it is the case of a shock oriented parallel to the mesh. We explained at the beginning of this section that that problem is the "easiest" one for which mesh refinement could be implemented in two space dimensions.

In the rest of this section, we will discuss and give numerical examples of the use of the method outlined above to resolve stationary rapid transitions oriented obliquely to a mesh in solutions of the two-dimensional Burgers' equation (equation (1)).

Figures 4.1.4 show the initial data and solution at time $t = 1$ for a computation using this method. The initial data (figure 4.1.4a) is a ramp oriented obliquely with respect to the mesh connecting the constant values $u = \pm 1$. Figure 4.1.4b shows the solution after the last sweep in x and figure 4.1.4c shows the solution after the last sweep in y. The plus signs '+' indicate the locations of the mesh points in the final refined mesh. Lines are also drawn in the direction perpendicular to the sweep direction (e.g. in the y-direction in figure 4.1.4a) to indicate the location of the underlying coarse mesh. (Note that figure 4.1.4c is reversed in orientation from the other two plots in this series.)

Figures 4.1.5 show the initial data and computed solutions at time $t = 0.2$ and time $t = 1$ for another example using this method. The coarse mesh in this case was not a uniform one, but was finer near the center of the domain where the corner of the "wedge" occurs. This was done in an attempt to resolve that corner. The initial data also consist of ramps connecting the two constant states $u = \pm 1$. The two ramps are oriented in such a way that the one on the left evolves into a shock while the one on the right forms a contact discontinuity. Because of the dissipative terms in Burgers' equation, of course, the shock has finite width, and the contact discontinuity becomes wider with time. In this

series of plots, the orientation is the same for all sweeps shown. The meshpoints are indicated with small squares and plus signs. The squares denote meshpoints that lie on the underlying coarse mesh. Figures 4.1.5b and 4.1.5d are the solutions after the x-sweep at $t = 0.2$ and $t = 1.0$ respectively; figures 4.1.5c and 4.1.5e show the solutions after the corresponding y-sweeps. In all the computations presented in this section, $\varepsilon = 1/400$, and the time step was $k = 1/20$. Note that in particular, the intended objective of this method, to resolve steady two-dimensional rapid transitions, has been realized.

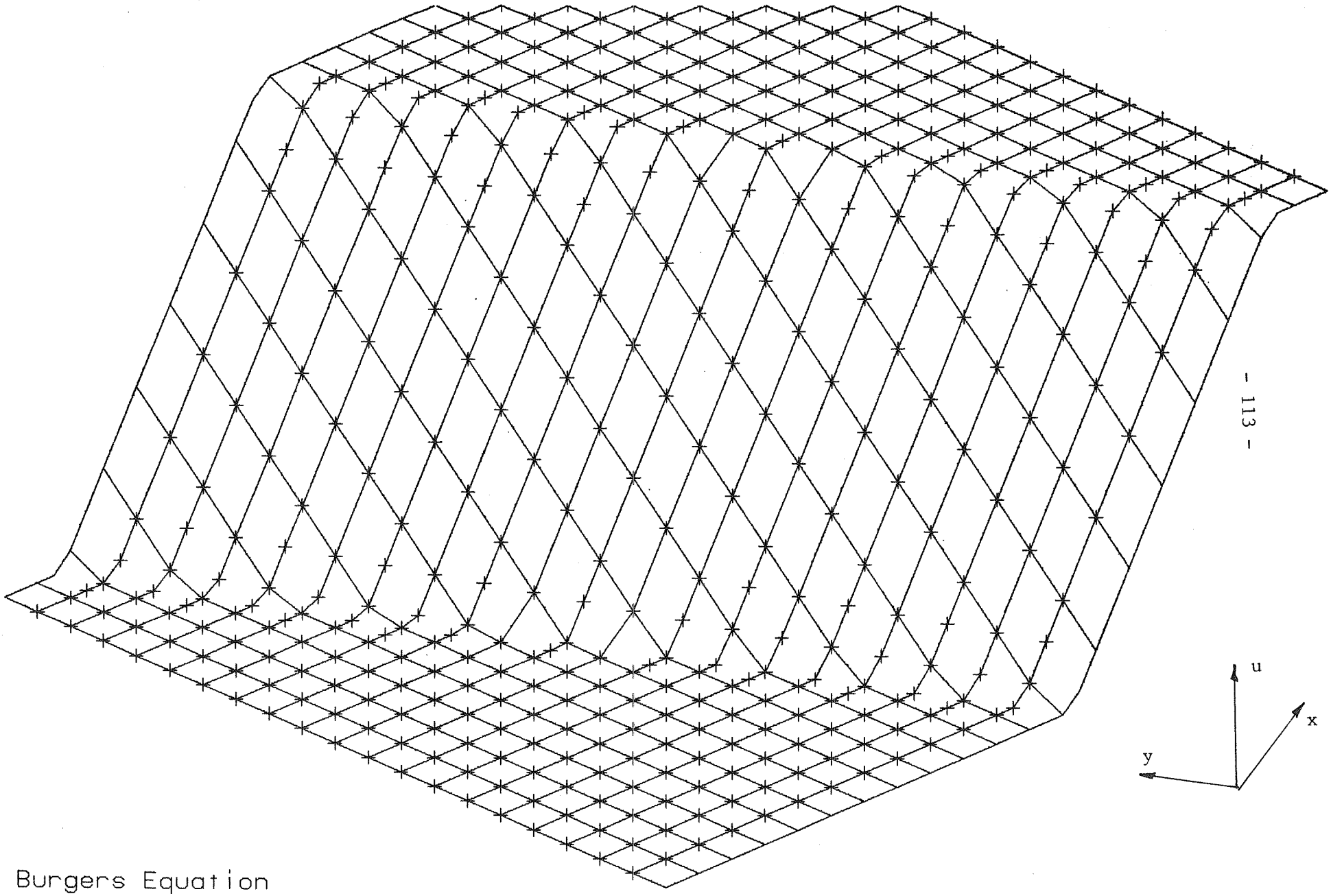


Figure 4.1.3a

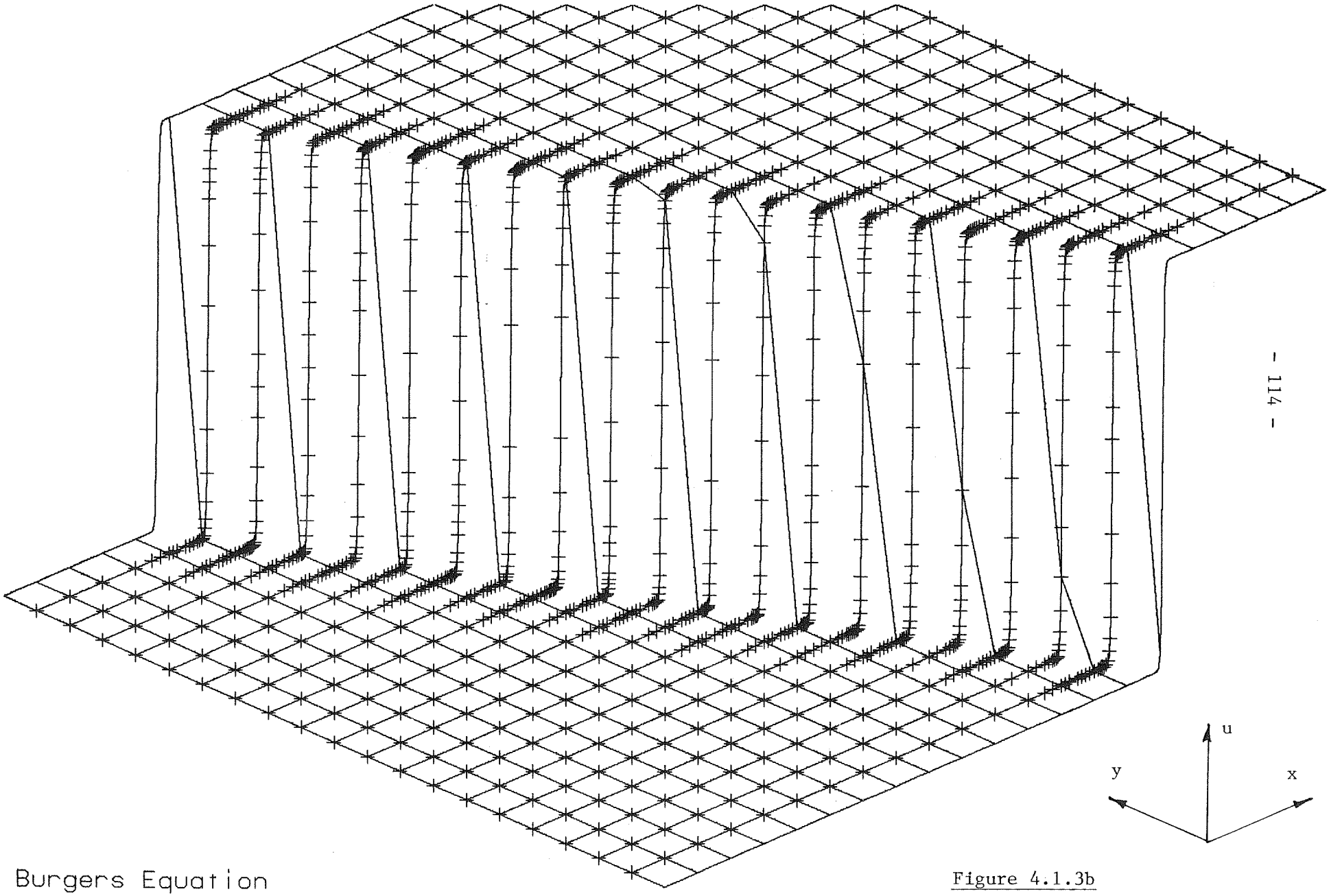


Figure 4.1.3b

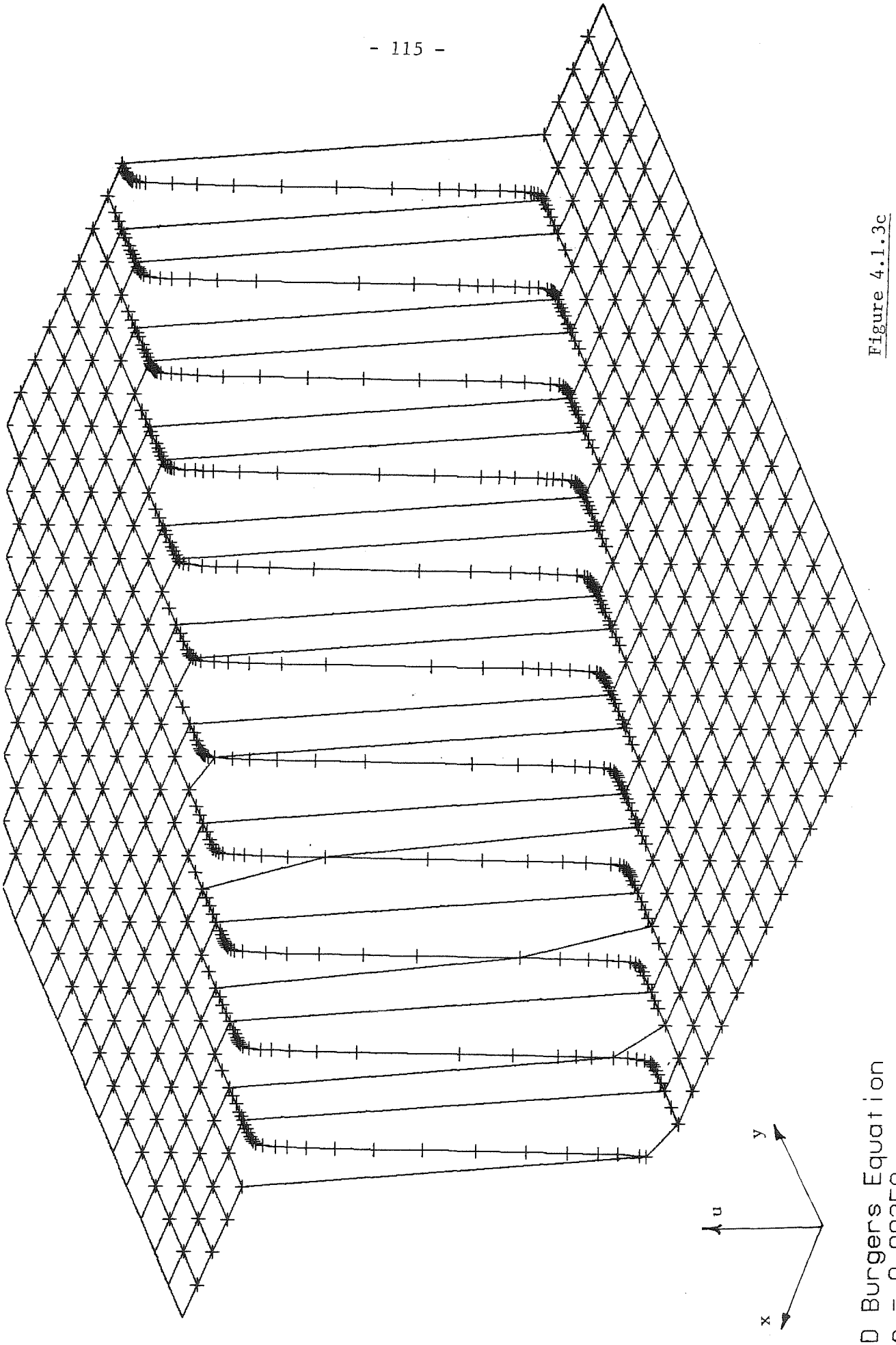


Figure 4.1.3c

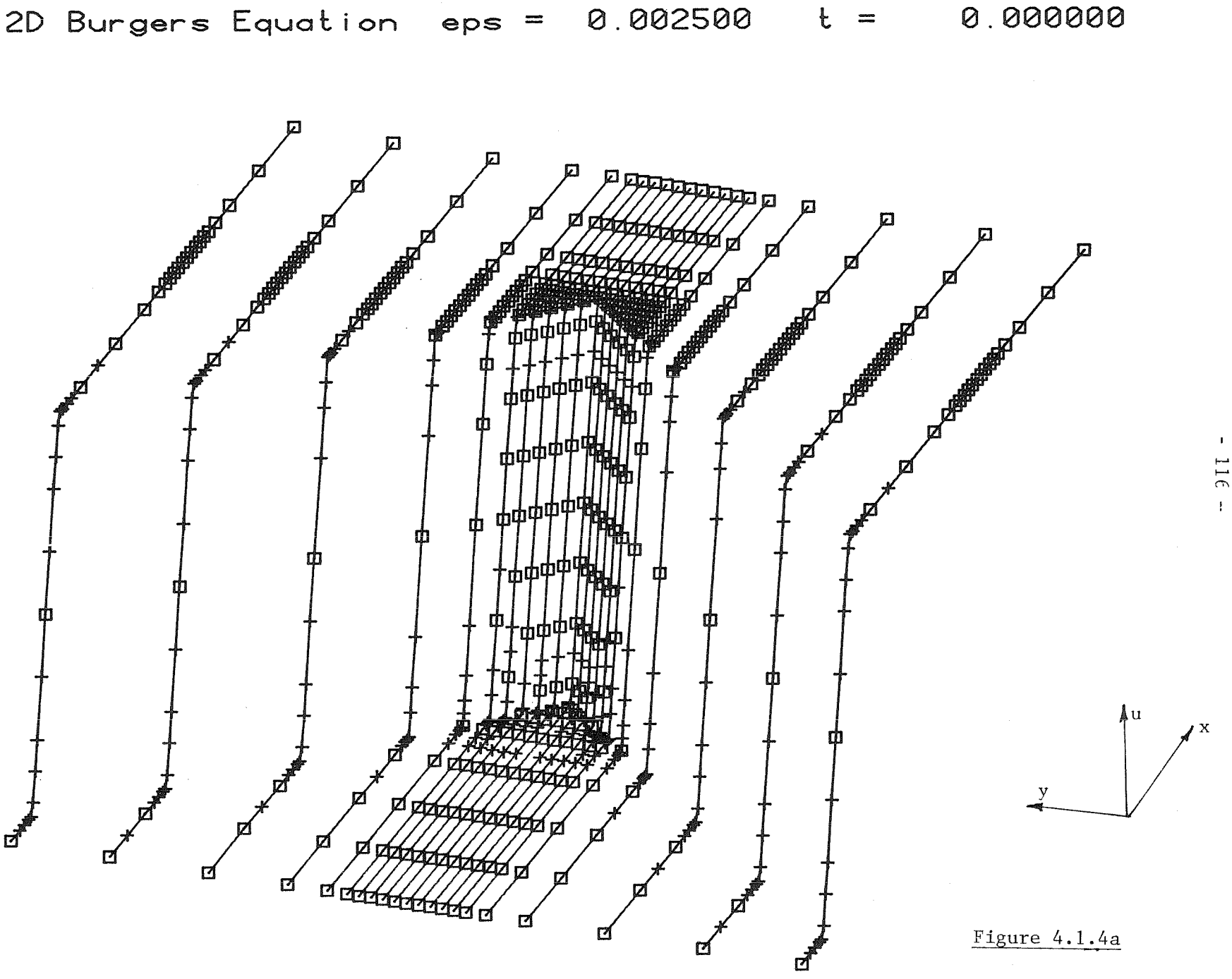


Figure 4.1.4a

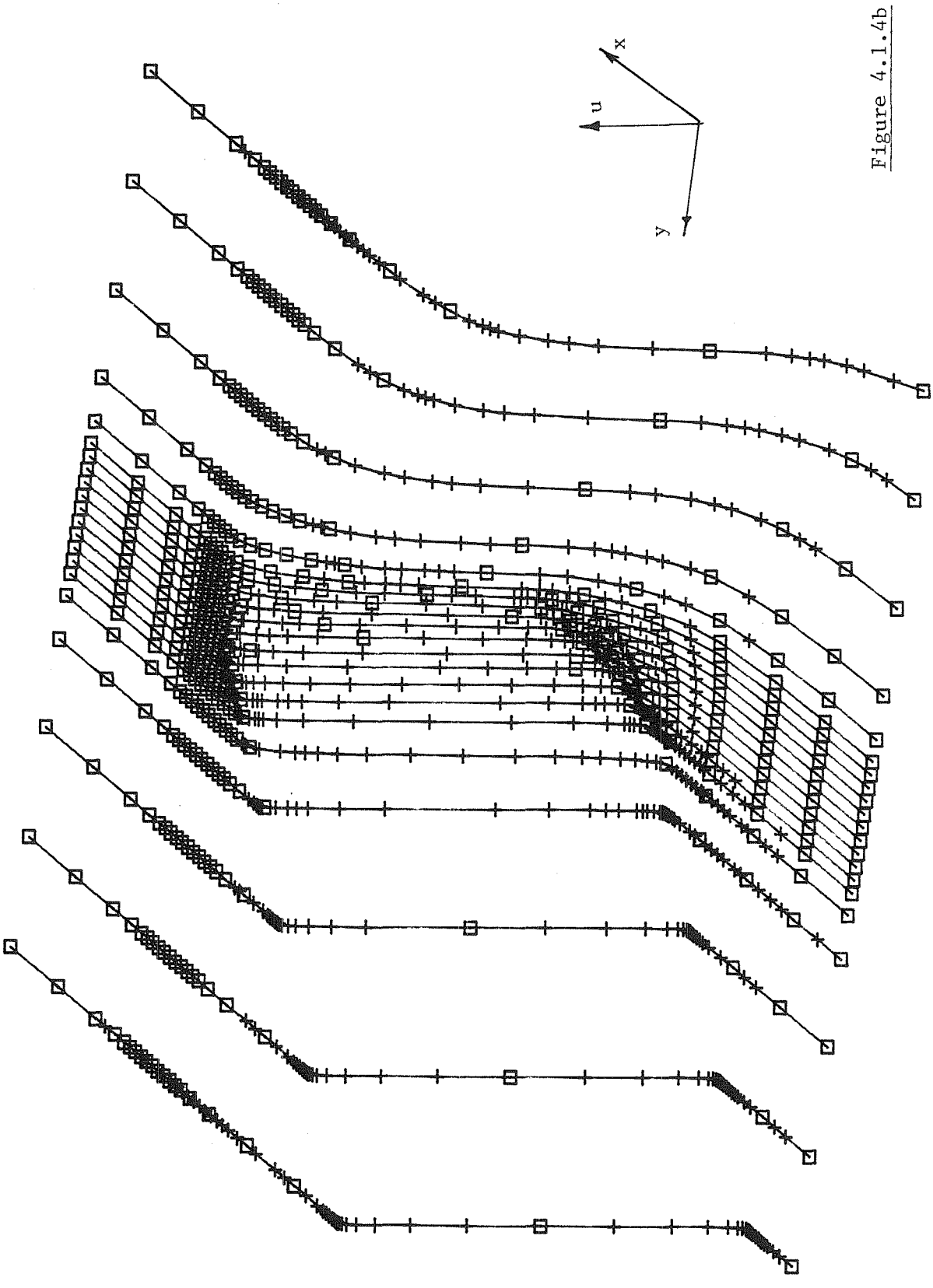


Figure 4.1.4b

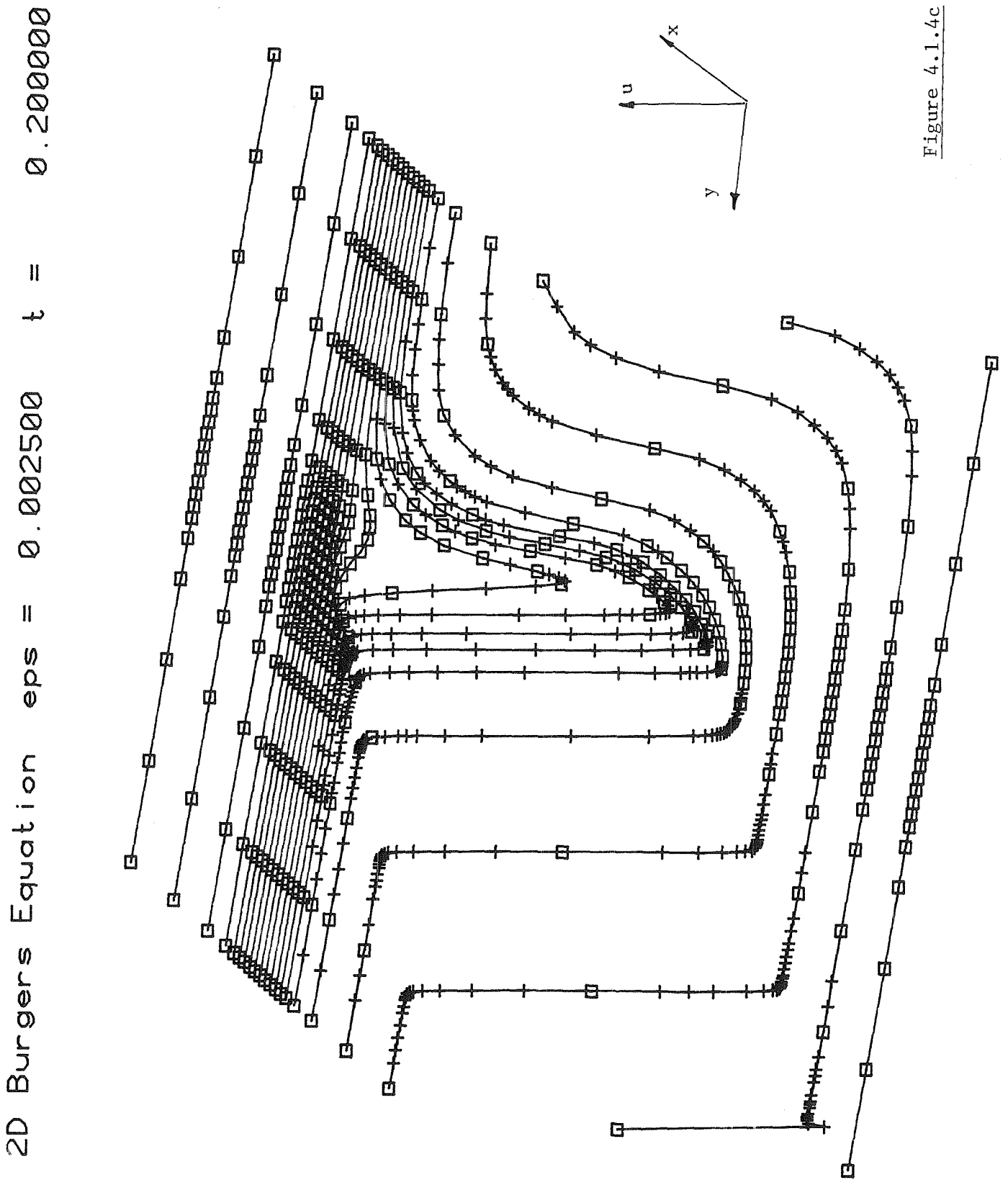


Figure 4.1.4c

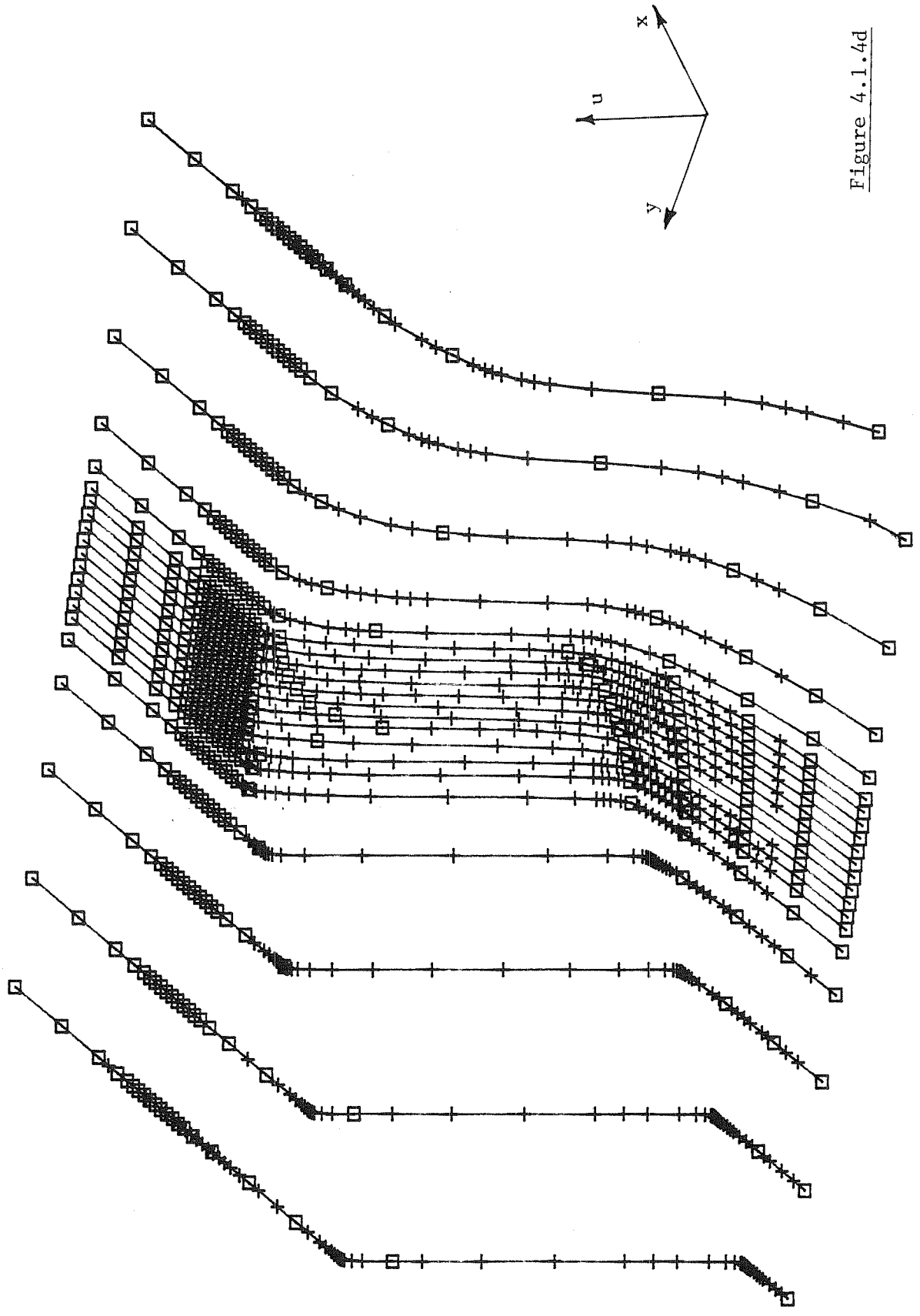


Figure 4.1.4d

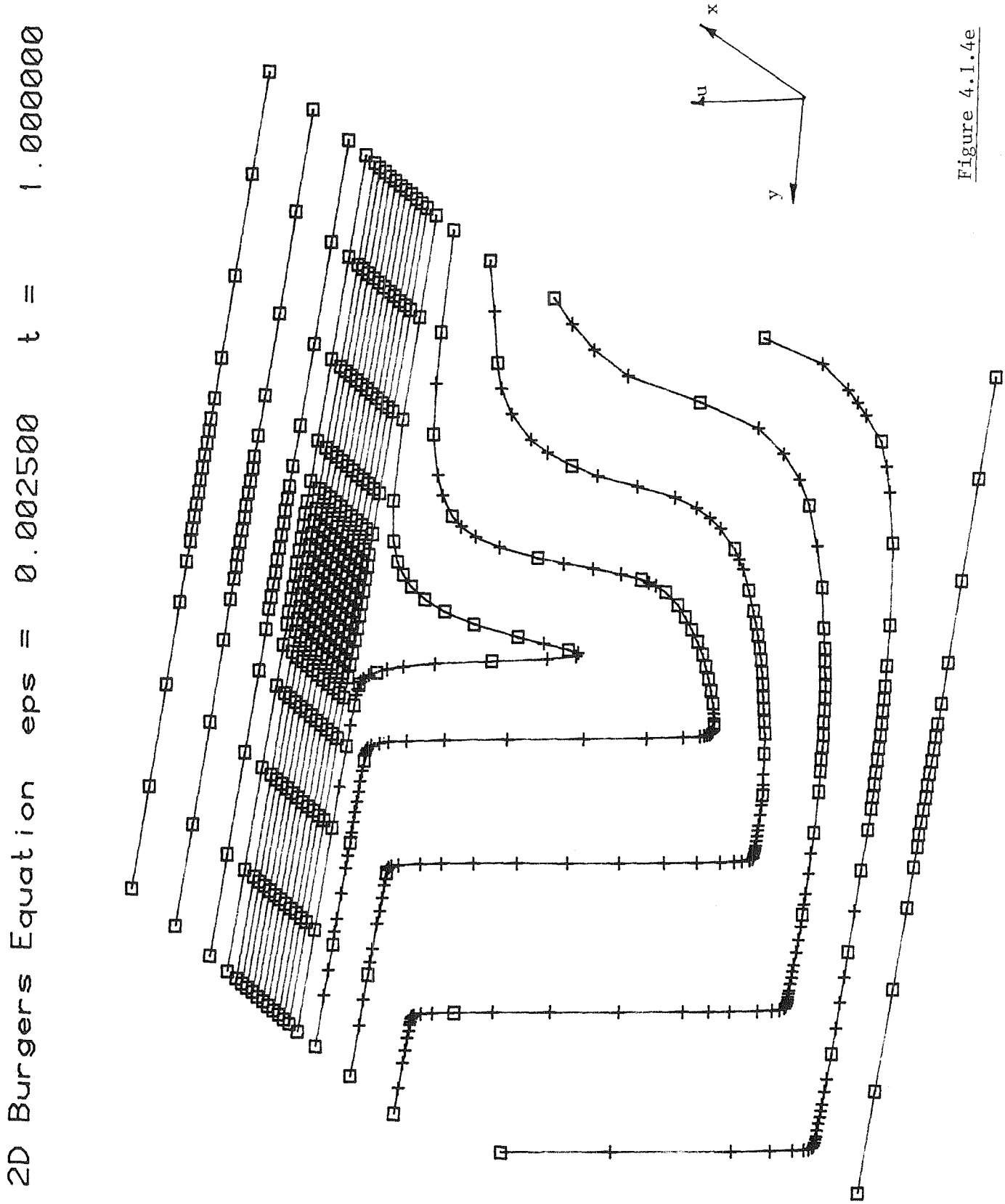


Figure 4.1.4e

REFERENCES

Abrahamsson, L.R. [1975a], A priori estimates for solutions of singular perturbation problems with a turning point, *Uppsala University Dept. of Comp. Sci. Rept. no. 56*

Abrahamsson, L.R. [1975b], Difference approximations for singular perturbation problems with a turning point, *Uppsala University Dept. of Comp. Sci. Rept. no. 57*

Abrahamsson, L.R., H.B. Keller and H.O. Kreiss [1974], Difference approximations for singular perturbations of systems of ordinary differential equations, *Numer. Math.*, 22, pp. 367-391.

Abrahamsson, L.R. and S. Osher [1981], Monotone difference schemes for singular perturbation problems, *UCLA preprint.*

Apelkrans, M.R. [1968], On difference schemes for hyperbolic equations with discontinuous initial values, *Math. Comp.*, 22, pp. 525-539

Berger, M. [1982], Ph.D. Thesis, Stanford University Dept. of Computer Science.

Boris, J. and D. Book [1973], Flux-corrected transport, I. SHASTA, A fluid transport algorithm that works, *J. Comp. Phys.*, 11, pp. 38-69.

Brenner, Ph. and Vidar Thomee [1971], Estimates near discontinuities for some difference schemes, *Math. Scand.*, 28, pp.329-340.

Brown, D.L. and L.G.M. Reyna [1982], to appear.

Chorin, A.J. [1976], Random choice solution of hyperbolic systems, *J. Comp. Phys.*, 22, pp. 517-533.

Courant, R., E. Isaacson and M. Rees [1952], *Comm. Pure and App. Math.*, 5, p. 243.

deBoor, C. [1975], A smooth local interpolant with "small" k-th derivative, in: A.K. Aziz, ed., **Numerical Solutions of Boundary Value Problems for Ordinary Differential Equations,** New York: Academic Press, pp. 177-197

deBoor, C. [1981], Smooth and rough interpolation, *Eidegenoessische Technische Hochschule Seminar fuer Angewandte Mathematik (Zurich) Research Report no. 81-03.*

Denny, V.E. and R.I. Landis [1971], A new method for solving two-point boundary value problems using optimal node distribution, *J. Comp. Phys.*, 9, pp. 120-137.

Dorr, F.W. [1970], The numerical solution of singular perturbations of boundary value problems, *SIAM J. Num. Anal.*, 7, pp., 281-313.

Engquist, B. and S. Osher [1979], One-sided difference approximations for non-linear conservation laws, preprint.

Engquist, B. and S. Osher [1980a], One sided difference schemes and transonic flow, *Proc. Natl. Acad. Sci. USA*, 77, pp. 3071-3074.

Engquist, B. and S. Osher [1980b], Stable and entropy satisfying approximations for transonic flow, *Math. Comp.*, 34, pp. 45-75.

Godunov, S.K. [1959], Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics, *Matematicheski Sbornik*, 47, p. 271 (in Russian) (translation by I. Bohachevsky).

Gropp, W.D. [1980], A test of moving mesh refinement for 2-D scalar hyperbolic problems, *SIAM J. Sci. Stat. Comput.*, 1, pp. 191-197.

Gropp, W.D. [1981], Ph.D. Thesis, Stanford University Dept. of Comp. Sci.

Harten, A. [1977], The artificial compression method for computation of shocks and contact discontinuities, I. Single Conservation Laws, *Comm. Pure and Appl. Math.* 30, pp. 611-638.

Harten, A. [1978], The artificial compression method for computation of shocks and contact discontinuities: III, Self-adjusting hybrid schemes, *Math. Comp.*, 32, pp. 363-389.

Harten, A. [1981], Lectures at Stanford University, July 27,30, 1981.

Harten, A., J.M. Hyman and P.D. Lax [1976],On finite difference approximations and entropy conditions for shocks, *Comm. Pure and Appl. Math.*, 29, pp. 297-322.

Harten, A. and P.D. Lax [1981], A random choice finite difference scheme for hyperbolic conservation laws, *SIAM J. Num. Anal.*, 18, pp. 289-315.

Hedstrom, G. [1968], The rate of convergence of some difference schemes, *SIAM J. Num. Anal.*, 5, pp. 366-406.

Hemker, P.W. [1974], A method of weighted one-sided differences for stiff boundary value problems with turning points, *Mathematisch Centrum (Amsterdam), Afdeling Numerieke Wiskunde Report NW 9/74.*

Hemker, P.W. [1977], A numerical study of stiff two-point boundary problems,

Ph.D. Thesis, Mathematisch Centrum, Amsterdam.

Hopf, E. [1969], On the right weak solution of the Cauchy problem for a quasilinear equation of first order, *J. of Math. and Mech.*, 19, pp. 483-487.

Hyman, J.M. [1979], A method of lines approach to the numerical solution of conservation laws, *Los Alamos Preprint, LA-UR-79-837.*

Hyman, J.M. [1981], personal communication.

Kato, T. [1976], **Perturbation Theory for Linear Operators,** Berlin: Springer-Verlag, 619 pp.

Kautsky, J. and N.K. Nichols [1981], Equidistributing meshes with constraints, *SIAM J. Sci. and Stat. Comput.*, 1, pp. 449-511.

Keller, H.B. and T. Cebeci [1972], Accurate numerical methods for boundary-layer flows, II: Two-dimensional turbulent flows, *AIAA Journal*, 10, pp. 1193-1199.

Kreiss, B. [1981], Construction of curvilinear grids, *Uppsala University Dept. of Comp. Sci. Rept. no. 89.*

Kreiss, B. and H.O. Kreiss [1981], Numerical methods for singular perturbation problems, *SIAM J. Num. Anal*, 18 pp. 262-276.

Kreiss, H.O. [1974], Numerical solution of singular perturbation problems, in: Willoughby, R.A., ed., **Stiff Differential Systems,** New York: Plenum Press, pp. 165-170.

Kreiss, H.O. [1975], Difference approximations for singular perturbation problems, in: A.K. Aziz, ed., **Numerical Solution of Boundary Value Problems for Ordinary Differential Equations,** New York: Academic Press, pp. 199-211.

Kreiss, H.O. [1976], Numerical methods for singular perturbation problems, *SIAM-AMS Proceedings*, vol. 10, pp. 73-86.

Kreiss, H.O. [1978], Difference methods for stiff ordinary differential equations, *SIAM J. Num. Anal.*, 15, pp. 21-58.

Kreiss, H.O. [1981], A posteriori error estimates for scalar equations, unpublished.

Kreiss, H.O. and E. Lundqvist [1968], On difference approximations with the wrong boundary values: *Math. Comp.*, 22, pp. 1-12.

Kreiss, H.O. and N. Nichols [1975], Numerical methods for singular perturbation problems, *Uppsala University Dept. of Comp. Sci. Rept. no. 57.*

Lax, P.D. [1971], Shock waves and entropy, *Proc. Symp. at Univ. of Wisc.*, E.H. Zarantonello, ed., pp. 603-634.

Lax, P.D. [1972], **Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves,** Philadelphia: SIAM, 48 pp.

Lax, P. and B. Wendroff [1960], Systems of conservation laws, *Comm. Pure and Appl. Math.*, 13, pp. 217-237.

Lentini, M. and V. Pereyra [1977], An adaptive finite difference solver for nonlinear two-point boundary value problems with mild boundary layers, *SIAM J. Numer. Anal.*, 14, pp. 91-111.

Majda, A. and S. Osher [1977], Propagation of error into regions of smoothness for accurate difference approximations to hyperbolic equations: *Comm. Pure and Appl. Math.*, 30, pp. 671-705.

Oleinik, O.A. [1963], Uniqueness and stability of the generalized solution of the Cauchy problem for a quasilinear equation, *Amer. Math. Soc. Trans., Ser. 2*, 33, pp. 285-290.

Oliger, J. [1981], Adaptive composite grid methods for time dependent problems, presented at the Adaptive Mesh Workshop, Center for Nonlinear Studies, Los Alamos Natl. Lab., August 2, 1981.

Osher, S. [1980], Numerical solution of singular perturbation problems and hyperbolic systems of conservation laws, *UCLA preprint.* (also in North-Holland Mathematics Studies #47, O. Axelsson, L.S. Frank, A. van der Sluis, eds., [1981], pp. 179-205).

Osher, S. [1981], Nonlinear singular perturbation problems and one-sided difference schemes, *SIAM J. Num. Anal.*, 18, pp. 129-144.

Pearson, C.E. [1968], On a differential equation of boundary layer type, *J. Math. Physics*, 47, pp. 134-154.

Pereyra, V. and E.G. Sewell [1975], Mesh selection for discrete solution of boundary value problems in ordinary differential equations, *Numer. Math.*, 23, pp. 261-268.

Pierson, B.L. and P. Kutler [1979], Optimal nodal distribution for improved accuracy in computational fluid dynamics, *Proceedings of the 17th Aerospace Sciences Meeting (AIAA)*, New Orleans, January 15-17, 1979.

Quinn, B.K. [1971], Solutions with shocks: an example of an L_1-contractive semigroup, *Comm. Pure and Appl. Math.*, 24, pp. 125-132.

Reyna, L.G.M. [1982], Ph.D. Thesis, California Institute of Technology Dept. of Applied Math.

Richtmyer, R.W. and K.W. Morton [1967], **Difference Methods for Initial Value Problems,** 2nd ed., New York: Interscience Publisheres, Inc., 405 pp.

Sod, G. [1977], A survey of numerical methods for compressible fluids, *Courant Institute Report COO-3077-145.*

Steger, J.L. and R.F. Warming [1981], Flux vector splitting of the inviscid gas-dynamic equations with applications to finite difference methods, *J. Comp. Phys.*, 40, pp. 263-293.

Swartz, B. [1981], Courant-like conditions limit reasonable mesh refinement to order h^2, *Los Alamos Report LA-UR-81-2037*, (submitted to SIAM J. on Sci. and Stat. Comp.)

Thomee, V. [1969], Stability theory for partial difference operators, *SIAM Review*, 11, pp. 152-195.

Thomee, V. [1971], On the rate of convergence of difference schemes for hyperbolic equations, in: Hubbard, B., ed., *Numerical Solution of Partial Differential Equations-II SYNSPADE 1970*, New York: Academic Press, pp. 586-622.

van Leer, B. [1979], Towards the ultimate conservative difference scheme, V. A second order sequel to Godunov's method, *J. Comp. Phys.*, 32, pp. 101-136.

Whitham, G. [1974], **Linear and Nonlinear Waves**, New York: Wiley-Interscience Publishers, 636 pp.