NUMERICAL SOLUTION OF PARABOLIC EQUATIONS BY THE BOX SCHEME

Thesis by

Kirby William Fong

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1973

(Submitted September 22, 1972)

ACKNOWLEDGMENTS

The author wishes to thank Professor Herbert B. Keller for suggesting the problems considered in this thesis and for valuable discussions during the course of the research.

He wishes to thank also the National Science Foundation for fellowship support and the California Institute of Technology for graduate research and teaching assistantships.

He is also thankful to Mrs. Alrae Tingley and Mrs. Virginia Conner for their expeditious and meticulous typing of his manuscript.

ABSTRACT

The box scheme proposed by H. B. Keller is a numerical method for solving parabolic partial differential equations. We give a convergence proof of this scheme for the heat equation, for a linear parabolic system, and for a class of nonlinear parabolic equations. Von Neumann stability is shown to hold for the box scheme combined with the method of fractional steps to solve the two-dimensional heat equation. Computations were performed on Burgers' equation with three different initial conditions, and Richardson extrapolation is shown to be effective.

TABLE OF CONTENTS

Introduction		1
Chapter 1	The Box Scheme in One Space Dimension	
I. 1	A Basic Description of the Box Scheme	4
I. 2	An Energy Inequality for the Heat Equation	10
I. 3	Convergence of the Box Scheme for the Heat Equation	17
I. 4	An Energy Inequality for a Linear Parabolic System	28
I. 5	Convergence of the Box Scheme for a Linear Parabolic System	34
I. 6	Solving the Linear Finite Difference Equations	44
I. 7	Nonlinear Parabolic Equations	54
Chapter 2	The Two-Dimensional Heat Equation	
II. 1	The Method of Fractional Steps	69
II. 2	The Box Scheme and the Method of Fractional Steps	73
II. 3	Von Neumann Stability	75

TABLE OF CONTENTS (Continued)

Chapter 3	Burgers' Equation: Computational Examples	
III.1	Burgers' Equation	80
III.2	Example 1: Smoothing of a Sharp Front	84
III.3	Example 2: Formation of a Shock	94
III.4	Example 3: Interaction of Two Shocks	116
References		128

INTRODUCTION

This thesis deals with the numerical solution of parabolic equations by the box scheme. Chapter I is devoted to the analysis of problems in one space dimension. We begin with a description of the box scheme and list some situations in which it would be preferable to other methods of computation. We then give three convergence proofs for the box scheme. In each proof we use discrete analogues of energy inequalities to show that the finite difference solutions are accurate approximations of the continuous solutions. Energy inequalities are generally used to prove uniqueness of solutions of initial value problems; however, in our work we have used modified forms for initial boundary value problems with inhomogeneous terms. Section I. 2 gives the derivation of such an energy inequality for the heat equation. In Section I. 3 we show how the energy inequality can be used as a model for finite difference equations. We then generalize the convergence proof for the heat equation to a linear parabolic system. Section I. 4 gives a derivation of an energy inequality, and Section I. 5 shows how the energy inequality may be discretized with some complications to yield a convergence proof of the box scheme for parabolic systems.

The emphasis of Sections I.6 and I.7 is on the computation of the finite difference solution. In Section I.6 we prove that an upper and lower block triangular matrix factorization may be used to solve the finite difference equations for a special linear equation. Using an argument involving principal error functions, we also show how to resolve a problem about the "smoothness" of the finite difference solution that arises in Section I.5. Finally in Section I.7 we give a constructive proof that the nonlinear difference equations resulting from applying the box scheme to a particular class of nonlinear parabolic equations have a unique solution. The mean value theorem enables us to adapt the convergence proof for linear systems to this class of nonlinear equations.

In Chapter II we give an example of how the box scheme may be extended to solving the heat equation in two space dimensions by using the method of fractional steps. We show that the initial value problem with periodic data leads to a numerical scheme which is stable in the sense of von Neumann. With this type of problem our numerical scheme is consistent to second order accuracy so that the numerical solution will converge to the continuous solution with second order accuracy.

In Chapter III we present the results of computations on Burgers' equation with three different initial conditions. We sought to identify the formation of a shock numerically —that is, without looking at a graph of the solution— but have not obtained a conclusive result. We have also performed Richardson extrapolations with solutions from successively refined nets and have found empirical conditions under which the extrapolations appear to be most effective for producing more accurate solutions.

CHAPTER I

THE BOX SCHEME IN ONE SPACE DIMENSION

I. 1 A Basic Description of the Box Scheme

The box scheme for the numerical solution of parabolic equations was originally proposed by Keller [1971]. Since this scheme is of fundamental importance in this thesis, we present here a brief review of the method and indicate some of the ways in which it is superior to other numerical methods for solving parabolic problems. We consider the following special problem defined in the rectangle $0 \le x \le 1$ and $0 \le t \le T$:

$$\frac{\partial U}{\partial t} = \frac{\partial}{\partial x}\left(a(x)\,\frac{\partial U}{\partial x}\right) + c(x)U + S(x, t), \tag{1.1a}$$

$$U(x, 0) = g(x), \tag{1.1b}$$

$$\alpha_0 U(0, t) - \beta_0 a(0) U_x(0, t) = g_0(t), \tag{1.1c}$$

$$\alpha_1 U(1, t) + \beta_1 a(1) U_x(1, t) = g_1(t). \tag{1.1d}$$

An important step in the method is to reformulate the problem as a first order system of equations:

$$a(x)\frac{\partial U}{\partial x} = V, \tag{1.2a}$$

$$\frac{\partial V}{\partial x} = \frac{\partial U}{\partial t} - c(x)U - S(x,t), \tag{1.2b}$$

$$U(x,0) = g(x), \tag{1.2c}$$

$$V(x,0) = a(x)\frac{dg(x)}{dx}, \tag{1.2d}$$

$$\alpha_0 U(0,t) - \beta_0 V(0,t) = g_0(t), \tag{1.2e}$$

$$\alpha_1 U(1,t) + \beta_1 V(1,t) = g_1(t). \tag{1.2f}$$

We define a mesh over the rectangle:

$$0 = x_1 < x_2 < \cdots < x_J = 1, \tag{1.3a}$$

$$0 = t_1 < t_2 < \cdots < t_N = T. \tag{1.3b}$$

The mesh spacings are then defined by

$$h_j \equiv x_j - x_{j-1}, \tag{1.3c}$$

$$k_n \equiv t_n - t_{n-1}, \tag{1.3d}$$

for $j = 2, \cdots, J$ and $n = 2, \cdots, N$. For net functions $\{\varphi_j^n\}$ and coordinates of the net we use the following notation:

$$x_{j\pm\frac{1}{2}} \equiv \tfrac{1}{2}(x_j + x_{j\pm 1}), \tag{1.4a}$$

$$t_{n\pm\frac{1}{2}} \equiv \tfrac{1}{2}(t_n + t_{n\pm 1}), \tag{1.4b}$$

$$\varphi^n_{j\pm\frac{1}{2}} \equiv \tfrac{1}{2}\left(\varphi_j^n + \varphi^n_{j\pm 1}\right), \tag{1.4c}$$

$$\varphi_j^{n\pm\frac{1}{2}} \equiv \tfrac{1}{2}\left(\varphi_j^n + \varphi_j^{n\pm 1}\right), \tag{1.4d}$$

$$D_x^- \varphi_j^n \equiv \frac{\varphi_j^n - \varphi_{j-1}^n}{h_j}, \tag{1.4e}$$

$$D_t^- \varphi_j^n \equiv \frac{\varphi_j^n - \varphi_j^{n-1}}{k_n}. \tag{1.4f}$$

For functions $\psi(x, t)$ defined everywhere in the rectangle we use the notation

$$\psi_j^n \equiv \psi(x_j, t_n), \tag{1.4g}$$

$$\psi_{j\pm\frac{1}{2}}^n \equiv \psi(x_{j\pm\frac{1}{2}}, t_n), \tag{1.4h}$$

$$\psi_j^{n\pm\frac{1}{2}} \equiv \psi(x_j, t_{n\pm\frac{1}{2}}). \tag{1.4i}$$

The box scheme for the numerical approximation of problem (1.2) is given in terms of the net functions $\{u_j^n\}$ and $\{v_j^n\}$ with all the difference approximations centered in the middle of the box $[x_{j-1}, x_j] \times [t_{n-1}, t_n]$ or on an appropriate edge of the box when coefficients do not depend on the time variable. We have

$$a_{j-\frac{1}{2}} D_x^- u_j^n = v_{j-\frac{1}{2}}^n, \tag{1.5a}$$

$$D_x^- v_j^{n-\frac{1}{2}} = D_t^- u_{j-\frac{1}{2}}^n - c_{j-\frac{1}{2}} u_{j-\frac{1}{2}}^{n-\frac{1}{2}} - S_{j-\frac{1}{2}}^{n-\frac{1}{2}}, \tag{1.5b}$$

for $2 \le j \le J$ and $2 \le n \le N$. The initial data are taken as

$$u_j^1 = g(x_j), \tag{1.5c}$$

$$v_j^1 = a_j \frac{d\,g(x_j)}{dx}, \tag{1.5d}$$

for $1 \le j \le J$, and the boundary conditions become

$$\alpha_0\, u_1^n - \beta_0\, v_1^n = g_0^n, \tag{1.5e}$$

$$\alpha_1\, u_J^n + \beta_1\, v_J^n = g_1^n, \tag{1.5f}$$

for $n \le 2 \le N$.

We immediately see two advantages of this method over other numerical methods which have been proposed. First the mesh spacings need not be uniform so that we may use a finer net in regions where we expect rapid changes in the solution. Second, the scheme is well adapted for problems in which $a(x)$ is not continuous. For instance in a diffusion problem where $a(x)$, the diffusivity of the medium, is discontinuous, $\frac{\partial U}{\partial x}$ will also be discontinuous, but the flux $a(x)\frac{\partial U}{\partial x}$ which is one of the dependent variables in the box scheme will be continuous so that we need not make any special modifications for discontinuous coefficients other than to pick the points of discontinuity to be mesh points.

There are other desirable features which are not so apparent. Being implicit, the method will be unconditionally stable. The method has second order accuracy even with nonuniform nets. Richardson extrapolation is valid if the continuous solutions are sufficiently differentiable and yields an improvement of two orders of accuracy for each extrapolation. Both $U(x, t)$ and $\frac{\partial U(x, t)}{\partial x}$ are approximated to the same order of accuracy. Although the box scheme requires a little more computation than the Crank-Nicolson scheme, it will nevertheless be preferable for the types of problems described in the preceding paragraph.

All of the virtues cited for the box scheme are discussed by Keller [1971]; however, he does not give the complete convergence proof. Subsequently Varah [1971] presented a general stability result for difference approximations to mixed initial boundary value problems for parabolic systems and included as an example the box scheme. This result uses Fourier transforms in x and t and requires the net spacings to be uniform. Once a finite difference scheme has been shown to be stable, we need only check its consistency with a particular problem to show its convergence to the solution of that problem. That is to say, stability is a property of a difference scheme only and does not refer to any particular parabolic problem. Consistency on the other hand refers to a specific problem and tells whether or not the difference equations accurately approximate the differential equation and boundary conditions as the mesh is refined. Stability and consistency together imply convergence of the finite difference solution to the continuous solution as the mesh is refined. Obviously convergence is the behavior we seek when we compute approximate solutions to differential equations, and we would like to have an <u>a priori</u> guarantee of convergence whenever possible. Our goal is to enlarge the class of parabolic problems for which we can guarantee convergence of the box scheme. We shall give a different proof for the convergence of the box scheme for the heat equation and then show how it can be extended to a parabolic system and to a class of nonlinear parabolic equations. There will be some mild restraints on the net spacing, but basically we will be allowing nonuniform time and space steps.

Implementation of the box scheme will entail the solution of linear systems of algebraic equations. While stability or convergence will imply that the systems are nonsingular and have unique solutions, it still remains for us to choose an appropriate algorithm for obtaining these solutions. In general an algorithm for solving a linear system will require that further conditions be satisfied in addition to non-singularity before we can prove that it will produce the desired solution. We shall use the method of factorization of block tridiagonal matrices recommended by Keller [1971]; however, we shall give an alternative proof based on an analysis suggested by Varah [1972] that this algorithm will work.

I. 2 An Energy Inequality for the Heat Equation

The convergence proofs which we shall present are based on energy inequalities. Energy inequalities are often used to prove well-posedness of initial value problems or to show that a solution depends continuously on the initial data of an initial value problem. It is frequently possible to construct a discrete analogue of an energy inequality which can then be used to prove convergence of a finite difference scheme. Indeed, this is how we shall obtain our convergence proofs, and to this end we wish to study thoroughly a simple parabolic equation - namely the heat equation - beginning with an energy inequality.

We consider the following problem:

$$V = U_x, \tag{2.1a}$$

$$V_x = U_t, \tag{2.1b}$$

$$U(x, 0) = g(x), \tag{2.1c}$$

$$V(x, 0) = \frac{dg(x)}{dx}, \tag{2.1d}$$

$$\alpha_0 U(0, t) - \beta_0 V(0, t) = g_0(t), \tag{2.1e}$$

$$\alpha_1 U(1, t) + \beta_1 V(1, t) = g_1(t), \tag{2.1f}$$

$$\frac{\alpha_0}{\beta_0} \geq 0, \tag{2.1g}$$

$$\frac{\alpha_1}{\beta_1} \geq 0. \tag{2.1h}$$

The conditions (2.1g) and (2.1h) are physically natural requirements for a mixed boundary condition. If these quantities are for some reason less than zero, our energy inequality would contain integrals along the boundaries $x = 0$ and $x = 1$. With the present assumptions these integrals will have signs such that they can be dropped from the inequalities we will derive. The case of Dirichlet boundary conditions where β_0 or β_1 is zero is actually simpler than the mixed case and would require only minor modifications of the proof. We therefore consider only the mixed case.

Typically, an energy inequality argument is used when one wishes to show problem (2.1) has a unique solution. If we assume U and V are one solution pair and u and v are another solution pair and define the difference functions

$$e(x,t) \equiv U(x,t) - u(x,t)\,, \tag{2.2a}$$

$$f(x,t) \equiv V(x,t) - v(x,t)\,, \tag{2.2b}$$

we find that e and f are solutions of

$$f = e_x\,, \tag{2.3a}$$

$$f_x = e_t\,, \tag{2.3b}$$

$$e(x,0) = 0\,, \tag{2.3c}$$

$$f(x, 0) = 0 , \tag{2.3d}$$

$$\alpha_o \, e(0, t) - \beta_o f(0, t) = 0 , \tag{2.3e}$$

$$\alpha_1 \, e(1, t) + \beta_1 f(1, t) = 0 , \tag{2.3f}$$

$$\frac{\alpha_o}{\beta_o} \geq 0 , \tag{2.3g}$$

$$\frac{\alpha_1}{\beta_1} \geq 0 . \tag{2.3h}$$

One then notes that the integrals with respect to x of the squares of e and f are zero at time zero and must be non-increasing functions of time. Since these square integrals are non-negative, they must be zero; therefore, e and f are zero for all time, and the solution pairs must be identical. This, however, is not the manner in which we wish to use the energy inequality. Let us suppose instead that u and v satisfy (2. 1) but that U and V satisfy only the boundary conditions (2. 1e) and (2. 1f). Then we would have the following relations governing e and f:

$$e_x = f + \rho , \tag{2.4a}$$

$$f_x = e_t + \sigma , \tag{2.4b}$$

$$\alpha_o \, e(0, t) - \beta_o \, f(0, t) = 0, \tag{2.4c}$$

$$\alpha_1 \, e(1, t) + \beta_1 \, f(1, t) = 0, \tag{2.4d}$$

$$\frac{\alpha_o}{\beta_o} \geq 0 , \tag{2.4e}$$

$$\frac{\alpha_1}{\beta_1} \geq 0 , \tag{2.4f}$$

$$\rho \equiv U_x - V , \tag{2.4g}$$

$$\sigma \equiv V_x - U_t . \tag{2.4h}$$

The terms ρ and σ account for the fact that U and V do not necessarily satisfy the differential equations. Before proceeding with the derivation we should like to furnish some motivation by saying that in the finite difference problem u and v will be solutions of the finite difference equations while U and V will be the solutions of the continuous system we are modeling. That is, u and v are to be the computed approximations to U and V. Since the difference equations are only approximations to the differential equations, we cannot expect U and V, the solutions of the differential equations, to satisfy the difference equations exactly. The extent to which they fail to satisfy the difference equations is embodied in the truncation errors ρ and σ. The terms ρ and σ are generally small. Specifically, for the box scheme, they are $O(h^2)$. Usually e and f are zero at time zero, but we shall not assume this. With this interpretation of the various terms in (2.4), we see that the results we need are bounds on e and f at times greater than zero in terms of ρ, σ, and the initial values of e and f. This is called a convergence result because ρ and σ can be made arbitrarily small by taking a sufficiently fine mesh and because the initial error also can presumably be made arbitrarily small. Then since e and f go to zero as the mesh is refined, the finite difference solution must converge to the continuous solution.

Our plan is to derive an energy inequality from (2.4) and then try to duplicate the derivation for a discrete system. It turns out that it is convenient to take the time derivative of (2.4a):

$$e_{xt} = f_t + \rho_t. \tag{2.4i}$$

If we multiply (2.4i) by f and (2.4b) by e, add the products, integrate with respect to x from 0 to 1, and make a substitution using (2.4a), we obtain, with $(\varphi, \psi) \equiv \int_0^1 \varphi(x)\psi(x)dx$ and $\|\varphi\|^2 \equiv (\varphi, \varphi)$,

$$\begin{aligned} \tfrac{1}{2}\frac{d}{dt}\|e\|^2 + \tfrac{1}{2}\frac{d}{dt}\|f\|^2 &= -(f, \rho_t) - (e, \sigma) \\ &\quad + [f\, e_t]_0^1 - (f_x, e_t) + [e\, f]_0^1 - (f, f) - (\rho, f). \end{aligned} \tag{2.5}$$

Further substitutions, a careful examination of the boundary conditions, a time integration from 0 to t, and the Schwarz inequality lead us to

$$\begin{aligned} &\|e(t)\|^2 + \|f(t)\|^2 + 2\int_0^t (\|f_x\|^2 + \|f\|^2)ds \\ &\le C + 2\int_0^t \{ \|f\|\cdot\|\rho_t\| + \|e\|\cdot\|\sigma\| + \|f_x\|\cdot\|\sigma\| \\ &\quad + \|f\|\cdot\|\rho\| \}ds, \end{aligned} \tag{2.6}$$

where

$$C \equiv \|e(0)\|^2 + \|f(0)\|^2 + \frac{\alpha_1}{\beta_1} e^2(1, 0) + \frac{\alpha_0}{\beta_0} e^2(0, 0). \tag{2.7}$$

We recognize that on the left side of (2.6) we may change t to τ without destroying the inequality provided $0 \le \tau \le t$. Then we integrate both sides with respect to τ from 0 to t. Also the four terms on the right side of (2.6) may be separated using the generalized arithmetic-geometric inequality $ab \le \frac{1}{2}(\varepsilon a^2 + \frac{1}{\varepsilon} b^2)$ where ε is an arbitrary positive number. In two cases we take $\varepsilon = (2t)^{-1}$, and in the other two we take $\varepsilon = (4t)^{-1}$. We then find

$$\frac{1}{2}\int_0^t \|e\|^2\,ds - \frac{1}{2}\int_0^t \|f_x\|^2 ds \le Ct + 2t^2 \int_0^t \|\sigma\|^2\,ds + 4t^2 \int_0^t \|\rho\|^2 ds + 4t^2 \int_0^t \|\sigma\|^2 ds. \tag{2.8}$$

Starting again with (2.6) we find a judicious use of the generalized arithmetic-geometric inequality gives

$$\|e(t)\|^2 + \|f(t)\|^2 \le C + \left\{\frac{1}{2}\int_0^t \|e\|^2 ds - \frac{1}{2}\int_0^t \|f_x\|^2 ds\right\} + \int_0^t \|\rho_t\|^2 ds + \frac{8}{3}\int_0^t \|\sigma\|^2\,ds + \int_0^t \|\rho\|^2\,ds. \tag{2.9}$$

The final energy inequality results from substituting (2.8) into the braces in (2.9):

$$\|e(t)\|^2 + \|f(t)\|^2 \le C(1+t) + \int_0^t \|\rho_t\|^2 ds + \left(\frac{8}{3} + 6t^2\right)\int_0^t \|\sigma\|^2 ds + (1+4t^2)\int_0^t \|\rho\|^2\,ds. \tag{2.10}$$

If we restrict t to lie between 0 and T, we see that the three terms involving the truncation errors can be bounded independently of t. Referring to (2. 7) we note that C is determined by the initial errors and can be made small. Hence the energy inequality (2. 10) is in the form we need for a convergence proof of the difference equations.

The derivation given in this section was suggested by Lees [1960] and Lees [1961].

I. 3 Convergence of the Box Scheme for the Heat Equation

Let $\{u_j^n\}$ and $\{v_j^n\}$ be net functions which we shall use to approximate U and V respectively. The box scheme for the heat equation then takes the form

$$v_{j-\frac{1}{2}}^n = D_x^- u_j^n, \tag{3.1a}$$

$$D_x^- v_j^{n-\frac{1}{2}} = D_t^- u_{j-\frac{1}{2}}^n, \tag{3.1b}$$

$$u_j^1 = g(x_j), \tag{3.1c}$$

$$v_j^1 = \frac{dg(x_j)}{dx}, \tag{3.1d}$$

$$\alpha_0 u_1^n - \beta_0 v_1^n = g_0^n, \tag{3.1e}$$

$$\alpha_1 u_J^n + \beta_1 v_J^n = g_1^n, \tag{3.1f}$$

$$\frac{\alpha_0}{\beta_0} \geq 0, \tag{3.1g}$$

$$\frac{\alpha_1}{\beta_1} \geq 0. \tag{3.1h}$$

Let U and V be the solutions of (2.1) and define the error net functions

$$e_j^n \equiv U(x_j, t_n) - u_j^n, \tag{3.2a}$$

$$f_j^n \equiv V(x_j, t_n) - v_j^n. \tag{3.2b}$$

$\{e_j^n\}$ and $\{f_j^n\}$ are then solutions of

$$D_x^- e_j^n = f_{j-\frac{1}{2}}^n + \rho_{j-\frac{1}{2}}^n , \tag{3.3a}$$

$$D_x^- f_j^{n-\frac{1}{2}} = D_t^- e_{j-\frac{1}{2}}^n + \sigma_{j-\frac{1}{2}}^{n-\frac{1}{2}} , \tag{3.3b}$$

$$e_j^1 = 0 , \tag{3.3c}$$

$$f_j^1 = 0 , \tag{3.3d}$$

$$\alpha_0 e_1^n - \beta_0 f_1^n = 0 , \tag{3.3e}$$

$$\alpha_1 e_J^n + \beta_1 f_J^n = 0 , \tag{3.3f}$$

$$\frac{\alpha_0}{\beta_0} \geq 0 , \tag{3.3g}$$

$$\frac{\alpha_1}{\beta_1} \geq 0 \tag{3.3h}$$

where the local truncation errors are defined by

$$\rho_{j-\frac{1}{2}}^n \equiv \left\{ D_x^- U(x_j, t_n) - \frac{\partial U(x_{j-\frac{1}{2}}, t_n)}{\partial x} \right\} + \left\{ V(x_{j-\frac{1}{2}}, t_n) - \tfrac{1}{2}[V(x_j, t_n) + V(x_{j-1}, t_n)] \right\} , \tag{3.3i}$$

$$\sigma_{j-\frac{1}{2}}^{n-\frac{1}{2}} \equiv \left\{ \tfrac{1}{2} D_x^- [V(x_j, t_n) + V(x_j, t_{n-1})] - \frac{\partial V(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}})}{\partial x} \right\} + \left\{ \frac{\partial U(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}})}{\partial t} - \tfrac{1}{2} D_t^- [U(x_j, t_n) + U(x_{j-1}, t_n)] \right\} . \tag{3.3j}$$

If we apply the operator D_t^- to (3.3a), we obtain

$$D_t^- D_x^- e_j^n = D_t^- f_{j-\frac{1}{2}}^n + \zeta_{j-\frac{1}{2}}^{n-\frac{1}{2}}, \tag{3.3k}$$

where $\zeta_{j-\frac{1}{2}}^{n-\frac{1}{2}}$ is defined by

$$\begin{aligned} \zeta_{j-\frac{1}{2}}^{n-\frac{1}{2}} &\equiv \{V_t(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}) - D_t^- V_{j-\frac{1}{2}}^n\} \\ &\quad + \{D_t^- D_x^- U_j^n - U_{xt}(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}})\}. \end{aligned} \tag{3.3ℓ}$$

It is possible to give alternate expressions for the truncation errors if we use Taylor expansions in the above definitions. First, however, we wish to introduce some new notation. Let $h \equiv \max_j h_j$, and for some fixed $r > 0$ we assume $\max_n k_n = rh$. Let $\theta(x)$ and $\varphi(t)$ be piecewise continuous functions such that for some fixed $\delta > 0$ we have

$$\left.\begin{aligned} h_j &= \theta(x_{j-\frac{1}{2}})h, \quad 2 \le j \le J, \\ \delta &\le \theta(x) \le 1, \quad 0 \le x \le 1, \end{aligned}\right\} \tag{3.4a}$$

$$\left.\begin{aligned} k_n &= \varphi(t_{n-\frac{1}{2}})h, \quad 2 \le n \le N, \\ \delta &\le \varphi(t) \le r, \quad 0 \le t \le T. \end{aligned}\right\} \tag{3.4b}$$

Proceeding as in Keller [1971] we assume U and V have piecewise continuous derivatives of order M where any jumps must occur at the net points. Then if $2m + 2 \le M$, we can show that

$$\rho^{n}_{j-\frac{1}{2}} = \sum_{\nu=1}^{m} \left(\frac{h}{2}\right)^{2\nu} R_{\nu}\{U, V; x_{j-\frac{1}{2}}, t_n\} + O(h^{2m+2}), \tag{3.5a}$$

$$\sigma^{n-\frac{1}{2}}_{j-\frac{1}{2}} = \sum_{\nu=1}^{m} \left(\frac{h}{2}\right)^{2\nu} S_{\nu}\{U, V; x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}\} + O(h^{2m+2}), \tag{3.5b}$$

$$\zeta^{n-\frac{1}{2}}_{j-\frac{1}{2}} = \sum_{\nu=1}^{m} \left(\frac{h}{2}\right)^{2\nu} Z_{\nu}\{U, V; x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}\} + O(h^{2m+2}), \tag{3.5c}$$

where

$$R_{\nu}\{U, V; x, t\} \equiv \frac{\theta^{2\nu}(x)}{(2\nu)!} \cdot \left\{\frac{1}{2\nu+1} \cdot \frac{\partial^{2\nu+1} U(x,t)}{\partial x^{2\nu+1}} - \frac{\partial^{2\nu} V(x,t)}{\partial x^{2\nu}}\right\}, \tag{3.5d}$$

$$S_{\nu}\{U, V; x, t)\} \equiv \sum_{\mu=0}^{\nu} \frac{\varphi^{2\mu}(t)\, \theta^{2\nu-2\mu}(x)}{(2\mu)!(2\nu-2\mu)!} \cdot \left\{\frac{1}{2\nu-2\mu+1} \cdot \frac{\partial^{2\nu+1} V(x,t)}{\partial x^{2\nu-2\mu+1}\partial t^{2\mu}} - \frac{1}{2\mu+1} \cdot \frac{\partial^{2\nu+1} U(x,t)}{\partial x^{2\nu-2\mu}\partial t^{2\mu+1}}\right\}, \tag{3.5e}$$

$$Z_{\nu}\{U, V; x, t\} \equiv \sum_{\mu=0}^{\nu} \frac{\varphi^{2\nu-2\mu}(t)\, \theta^{2\mu}(x)}{(2\mu)!\,(2\nu-2\mu+1)!} \cdot \left\{\frac{1}{2\mu+1} \cdot \frac{\partial^{2\nu+2} U(x,t)}{\partial t^{2\nu-2\mu+1}\partial x^{2\mu+1}} - \frac{\partial^{2\nu+1} V(x,t)}{\partial t^{2\nu-2\mu+1}\partial x^{2\mu}}\right\}. \tag{3.5f}$$

For the purpose of this section the most important feature of the truncation errors is that all three errors are $O(h^2)$.

We next introduce an inner product for net functions $\{\varphi^n_j\}$ and $\{\psi^n_j\}$:

$$(\varphi^n, \psi^n)_h \equiv \sum_{j=2}^{J} \varphi^n_{j-\frac{1}{2}} \psi^n_{j-\frac{1}{2}} h_j. \tag{3.6a}$$

If a net function is differenced, we will have

$$(\varphi^n, D_x^- \psi^n)_h \equiv \sum_{j=2}^{J} \varphi^n_{j-\frac{1}{2}} D_x^- \psi^n_j h_j. \tag{3.6b}$$

The norm associated with this inner product is

$$\|\varphi^n\|_h^2 \equiv (\varphi^n, \varphi^n)_h. \tag{3.6c}$$

We note that (3.6c) is actually a seminorm rather than a norm since a net function which oscillates along the mesh can have norm zero without itself being zero. We shall say more about this after the convergence proof. Finally, with our inner product the following identities hold:

$$(D_x^- \varphi, \psi)_h \equiv [\varphi_J \psi_J - \varphi_1 \psi_1] - (\varphi, D_x^- \psi)_h, \tag{3.7a}$$

$$(D_t^- \varphi^n, \varphi^{n-\frac{1}{2}})_h \equiv \frac{1}{2} D_t^- \|\varphi^n\|_h^2. \tag{3.7b}$$

As mentioned earlier our plan is to construct a discrete analogue of the energy inequality derived for the differential equations. The first step then is to construct the energy quantity exactly as was done for (2.5):

$$\frac{1}{2} D_t^- \|e^n\|_h^2 + \frac{1}{2} D_t^- \|f^n\|_h^2 = -(f^{n-\frac{1}{2}}, \zeta^{n-\frac{1}{2}})_h \tag{3.8}$$

$$- (e^{n-\frac{1}{2}}, \sigma^{n-\frac{1}{2}})_h + \left[f_J^{n-\frac{1}{2}} D_t^- e_J^n - f_1^{n-\frac{1}{2}} D_t^- e_1^n \right]$$

$$- (D_x^- f^{n-\frac{1}{2}}, D_t^- e^n)_h + \left[e_J^{n-\frac{1}{2}} f_J^{n-\frac{1}{2}} - e_1^{n-\frac{1}{2}} f_1^{n-\frac{1}{2}} \right]$$

$$- (D_x^- e^{n-\frac{1}{2}}, f^{n-\frac{1}{2}})_h .$$

Beyond equation (3.8) the discrete nature of the variables causes some difficulties which did not occur before. We therefore introduce new quantities which will help us notationally:

$$k_1 \equiv k_2 , \tag{3.9a}$$

$$\|f^{\frac{1}{2}}\|_h \equiv 0 , \tag{3.9b}$$

$$\|\zeta^{\frac{1}{2}}\|_h \equiv 0 , \tag{3.9c}$$

$$\|D_x^- f^{\frac{1}{2}}\|_h \equiv 0 , \tag{3.9d}$$

$$\|\sigma^{\frac{1}{2}}\|_h \equiv 0 , \tag{3.9e}$$

$$\|\rho^{\frac{1}{2}}\|_h \equiv 0 , \tag{3.9f}$$

$$s_m \equiv \frac{1}{2} \|\sigma^{m-\frac{1}{2}}\|_h + \frac{k_{m+1}}{2k_m} \|\sigma^{m+\frac{1}{2}}\|_h, \tag{3.9g}$$

$$s_1 \equiv \frac{1}{2}\|\sigma^{\frac{1}{2}}\|_h + \frac{k_2}{2k_1} \|\sigma^{\frac{3}{2}}\|_h = \frac{1}{2} \|\sigma^{\frac{3}{2}}\|_h. \tag{3.9h}$$

This notation plus additional substitutions, careful examination of the boundary conditions, a time summation from t_1 to t_n, and the Schwarz inequality lead to the analogue of (2.6):

$$\|e^n\|_h^2 + \|f^n\|_h^2 + 2\sum_{m=2}^{n} k_m \left(\|D_x^- f^{m-\frac{1}{2}}\|_h^2 + \|f^{m-\frac{1}{2}}\|_h^2\right) \le C + 2\sum_{m=1}^{n} k_m \cdot \left\{\|f^{m-\frac{1}{2}}\|_h \cdot \|\zeta^{m-\frac{1}{2}}\|_h + \|D_x^- f^{m-\frac{1}{2}}\|_h \cdot \|\sigma^{m-\frac{1}{2}}\|_h + \|f^{m-\frac{1}{2}}\|_h \cdot \|\rho^{m-\frac{1}{2}}\|_h + \|e^m\|_h \cdot s_m\right\}, \tag{3.10}$$

where

$$C \equiv \|e^1\|_h^2 + \|f^1\|_h^2 + \frac{\alpha_0}{\beta_0}(e_1^1)^2 + \frac{\alpha_1}{\beta_1}(e_J^1)^2. \tag{3.11}$$

We see that on the left side of (3.10) the index n may be changed to i where $1 \le i \le n$ yielding a set of valid inequalities. We then multiply both sides by k_i and sum from $i = 1$ to $i = n$. We again apply the arithmetic-geometric inequality to each of the four products on the right side of (3.10). With the notation $D \equiv t_n + k_1$ we now obtain

$$\tfrac{1}{2}\sum_{m=1}^{n} k_m \|e^m\|_h^2 + \sum_{m=1}^{n} k_m \|f^m\|_h^2 \tag{3.12}$$

$$- \sum_{m=1}^{n} k_m \|f^{m-\frac{1}{2}}\|_h^2 - \sum_{m=1}^{n} k_m \|D_x^- f^{m-\frac{1}{2}}\|_h^2 \le$$

$$CD + 2D^2 \sum_{m=1}^{n} k_m \|\zeta^{m-\frac{1}{2}}\|_h^2$$

$$+ D^2 \sum_{m=1}^{n} k_m \|\sigma^{m-\frac{1}{2}}\|_h^2 + 2\,D^2 \sum_{m=1}^{n} k_m \|\rho^{m-\frac{1}{2}}\|_h^2$$

$$+ \; 2\,D^2 \sum_{m=1}^{n} k_m \, s_m^2 \, .$$

(3.12) corresponds with (2.8) but has additional terms in $\|f^m\|_h^2$ and $\|f^{m-\frac{1}{2}}\|_h^2$ because a cancellation which occurred in the continuous case does not occur for discrete equations. Returning to (3.10) we use the arithmetic-geometric inequality with different parameters to deduce

$$\|e^n\|_h^2 + \|f^n\|_h^2 \le C + \Big\{ \tfrac{1}{2}\sum_{m=1}^{n} k_m \|e^m\|_h^2 \tag{3.13}$$

$$- \sum_{m=1}^{n} k_m \|f^{m-\frac{1}{2}}\|_h^2 - \sum_{m=1}^{n} k_m \|D_x^- f^{m-\frac{1}{2}}\|_h^2 \Big\}$$

$$+ 2 \sum_{m=1}^{n} k_m \|\zeta^{m-\frac{1}{2}}\|_h^2 + \sum_{m=1}^{n} k_m \|\sigma^{m-\frac{1}{2}}\|_h^2$$

$$+ 2 \sum_{m=1}^{n} k_m \|\rho^{m-\frac{1}{2}}\|_h^2 + \sum_{m=1}^{n} k_m \, s_m^2 \, .$$

Inequality (3.12) is still valid if we omit the term $\sum_{m=1}^{n} k_m \|f^m\|_h^2$ from the left side. We then substitute the remaining inequality into the braces in (3.13). With the notation

$$\tau_m^2 \equiv 2\|\zeta^{m-\frac{1}{2}}\|_h^2 + \|\sigma^{m-\frac{1}{2}}\|_h^2 + 2\|\zeta^{m-\frac{1}{2}}\|_h^2 \tag{3.14a}$$

$$+ 2\left(\tfrac{1}{2}\|\sigma^{m-\frac{1}{2}}\|_h + \frac{k_{m+1}}{2k_m}\|\sigma^{m+\frac{1}{2}}\|_h\right)^2 ,$$

$$\tau^2(n) \equiv \max_{1\le m\le n} \tau_m^2 , \tag{3.14b}$$

we conclude the result

$$\|e^n\|_h^2 + \|f^n\|_h^2 \le (1 + t_n + k_1)\Big[\|e^1\|_h^2 \tag{3.15}$$

$$+ \|f^1\|_h^2 + \frac{\alpha_0}{\beta_0}(e_1^1)^2 + \frac{\alpha_1}{\beta_1}(e_J^1)^2\Big]$$

$$+ [(t_n+k_1)^3 + t_n+k_1]\,\tau^2(n) .$$

(3.15) is the convergence result we sought. It says that the errors at a given fixed time may be made small if the initial errors are small and if the truncation errors are small. The latter error we noticed earlier was $O(h^2)$ and can be made smaller by refining the mesh. We note also that k_{m+1}/k_m is bounded by r/δ; hence, no further conditions are needed to guarantee that $\tau(n)$ is $O(h^2)$. If the initial data are approximated to $O(h^2)$ or better, then we see $\|e^n\|_h$ and $\|f^n\|_h$ are also $O(h^2)$.

There remain two points which must be clarified. The first is to show that there exists a unique solution of the finite difference problem (3.1). We shall defer this to a later section. The second is that $\|\cdot\|_h$ is a seminorm rather than a norm. This latter problem is fully discussed by Keller [1971], but we will reproduce the explanation here since possible oscillations in the finite difference solutions are of concern in practical computation.

Two net functions $\{\varphi_j\}$ and $\{\psi_j\}$ satisfy $\|\varphi - \psi\|_h = 0$ if and only if $\varphi_j = \psi_j + (-1)^j p$ for some constant p. Thus if $\|e^n - \tilde{e}^n\|_h = 0$ and $\|f^n - \tilde{f}^n\|_h = 0$, there exist p and q such that $\tilde{e}_j^n = e_j^n + (-1)^j p$ and $\tilde{f}_j^n = f_j^n + (-1)^j q$. In order that $\{e_j^n\}$, $\{f_j^n\}$, $\{\tilde{e}_j^n\}$, and $\{\tilde{f}_j^n\}$ satisfy the boundary conditions we must have

$$\left.\begin{aligned} \alpha_0 p - \beta_0 q = 0 \\ \alpha_1 p + \beta_1 q = 0 \end{aligned}\right\} . \qquad (3.16)$$

Four cases can occur. First, if $\alpha_0 \beta_1 + \alpha_1 \beta_0 \neq 0$, then $p = q = 0$, and the seminorm is actually a norm for net functions satisfying (3.3e) and (3.3f). Second, if $\beta_0 \beta_1 = 0$, then $p = 0$ so that $\{v_j^n\}$ but not $\{u_j^n\}$ may have oscillations. Third, if $\alpha_0 \alpha_1 = 0$, then $q = 0$, and $\{u_j^n\}$ may have oscillations. Finally, none of the above may happen so that both $\{u_j^n\}$ and $\{v_j^n\}$ could have oscillations. In the latter cases oscillations are eliminated by averaging neighboring values. Define

$$\bar{u}_{j-\frac{1}{2}}^n \equiv \tfrac{1}{2}(u_j^n + u_{j-1}^n), \qquad (3.17a)$$

$$\bar{v}_{j-\frac{1}{2}}^n \equiv \tfrac{1}{2}(v_j^n + v_{j-1}^n), \qquad (3.17b)$$

for $2 \le j \le J$. Then $\|\bar{u}^n\|_h = \|u^n\|_h$ and $\|\bar{v}^n\|_h = \|v^n\|_h$, but now $\|\cdot\|_h$ is a norm for net functions defined on $\{x_{j-\frac{1}{2}}: 2 \le j \le J\}$. Any oscillations are now removed, and (3. 15) therefore tells us the errors are not worse than $O(h^2)$ if U and V are piecewise four times continuously differentiable.

I. 4 An Energy Inequality for a Linear Parabolic System

We now wish to extend our analysis to a larger class of parabolic equations. Consider the following problem:

$$A^2(x)\, \underset{\sim}{U}_x(x,t) = \underset{\sim}{V}(x,t), \tag{4.1a}$$

$$\underset{\sim}{V}_x(x,t) = \underset{\sim}{U}_t(x,t) - C(x)\,\underset{\sim}{U}(x,t) - E(x)\,\underset{\sim}{V}(x,t) - \underset{\sim}{S}(x,t), \tag{4.1b}$$

$$\underset{\sim}{U}(x,0) = \underset{\sim}{g}(x), \tag{4.1c}$$

$$\underset{\sim}{V}(x,0) = A^2(x)\,\frac{d\,\underset{\sim}{g}(x)}{dx}, \tag{4.1d}$$

$$\alpha_0\,\underset{\sim}{U}(0,t) - \beta_0\,\underset{\sim}{V}(0,t) = \underset{\sim}{g}_0(t), \tag{4.1e}$$

$$\alpha_1\,\underset{\sim}{U}(1,t) + \beta_1\,\underset{\sim}{V}(1,t) = \underset{\sim}{g}_1(t), \tag{4.1f}$$

β_0^{-1} and β_1^{-1} exist, (4.1g)

$A^2(0)\beta_0^{-1}\alpha_0$ and $A^2(1)\beta_1^{-1}\alpha_1$ are positive semi-definite and symmetric, (4.1h)

$A(x)$ is symmetric and positive definite uniformly in x (4.1i)

Here $\underset{\sim}{U}$, $\underset{\sim}{V}$, $\underset{\sim}{S}(x,t)$, $\underset{\sim}{g}_0$, and $\underset{\sim}{g}_1$ are vectors of dimension p and $A(x)$, $C(x)$, $E(x)$, α_0, β_0, α_1, and β_1 are $p \times p$ matrices. The domain of the problem is $0 \le x \le 1$ and $0 \le t \le T$. Condition (4.1h) has been imposed so that boundary integrals which will arise will have signs such that they may be dropped from inequalities in which they would otherwise have to be retained. (4.1h) is a convenient assumption but not an essential one.

We now seek an energy inequality which we can use as the basis for a convergence proof. As before we suppose $\underset{\sim}{u}$ and $\underset{\sim}{v}$ are functions which satisfy the differential equations, the initial conditions, and the boundary conditions. Let $\underset{\sim}{U}$ and $\underset{\sim}{V}$ be another pair of functions which satisfy only the boundary conditions. We define

$$\underset{\sim}{e}(x, t) \equiv \underset{\sim}{U}(x, t) - \underset{\sim}{u}(x, t) , \tag{4.2a}$$

$$\underset{\sim}{f}(x, t) \equiv \underset{\sim}{V}(x, t) - \underset{\sim}{v}(x, t) . \tag{4.2b}$$

We find that $\underset{\sim}{e}$ and $\underset{\sim}{f}$ satisfy

$$A^2 \underset{\sim}{e}_x = \underset{\sim}{f} + \underset{\sim}{\rho} , \tag{4.3a}$$

$$\underset{\sim}{f}_x = \underset{\sim}{e}_t - C\underset{\sim}{e} - E\underset{\sim}{f} + \underset{\sim}{\sigma} , \tag{4.3b}$$

$$A^2 \underset{\sim}{e}_{xt} = \underset{\sim}{f}_t + \underset{\sim}{\zeta} , \tag{4.3c}$$

$$\alpha_0 \, \underset{\sim}{e}(0, t) - \beta_0 \underset{\sim}{f}(0, t) = \underset{\sim}{0} , \tag{4.3d}$$

$$\alpha_1 \, \underset{\sim}{e}(1, t) + \beta_1 \underset{\sim}{f}(1, t) = \underset{\sim}{0} , \tag{4.3e}$$

$$\underset{\sim}{\rho} \equiv A^2 \underset{\sim}{U}_x - \underset{\sim}{V} , \tag{4.3f}$$

$$\underset{\sim}{\sigma} \equiv \underset{\sim}{V}_x - \underset{\sim}{U}_t + C\underset{\sim}{U} + E\underset{\sim}{V} + \underset{\sim}{S} , \tag{4.3g}$$

$$\underset{\sim}{\zeta} \equiv \underset{\sim}{\rho}_t . \tag{4.3h}$$

$\underset{\sim}{\rho}$, $\underset{\sim}{\sigma}$, and $\underset{\sim}{\zeta}$ are error terms resulting from the fact that $\underset{\sim}{U}$ and $\underset{\sim}{V}$ do not necessarily satisfy the differential equations. If we take the dot products of $\underset{\sim}{f}$ with (4.3c) and $A^2 \underset{\sim}{e}$ with (4.3b), add the results, integrate over $0 \le x \le 1$, and integrate by parts, we arrive at

$$\frac{1}{2}\frac{d}{dt}\|A\underset{\sim}{e}\|^2 + \frac{1}{2}\frac{d}{dt}\|\underset{\sim}{f}\|^2 = -(\underset{\sim}{f}, \underset{\sim}{\zeta}) \tag{4.4}$$

$$+ (A^2\underset{\sim}{e}, C\underset{\sim}{e}) + (A^2\underset{\sim}{e}, E\underset{\sim}{f}) - (A^2\underset{\sim}{e}, \underset{\sim}{\sigma})$$

$$+ [A^2\underset{\sim}{f}\cdot\underset{\sim}{e}_t]_0^1 - (A_xA\underset{\sim}{f}, \underset{\sim}{e}_t) - (AA_x\underset{\sim}{f}, \underset{\sim}{e}_t)$$

$$- (A^2\underset{\sim}{f}_x, \underset{\sim}{e}_t) + [A^2 e\cdot f]_0^1 - (A_xA\underset{\sim}{e}, \underset{\sim}{f})$$

$$- (AA_x\underset{\sim}{e}, \underset{\sim}{f}) - (A^2\underset{\sim}{e}_x, \underset{\sim}{f}).$$

We introduce the notation $\|A\|$ for the maximum for $x \in [0, 1]$ of the Euclidean norm of the matrix $A(x)$. Let ϵ be the positive arbitrary parameter in the generalized arithmetic-geometric inequality. Define constants K, C_1 and C_2 as follows:

$$K \equiv 2\max\Big\{1 + \tfrac{1}{2}\|ACA^{-1}\|^2 + \tfrac{1}{2}\|A_x\|^2 \tag{4.5a}$$

$$+ \tfrac{1}{2}\|AA_xA^{-1}\|^2 + \tfrac{1}{2}\|CA^{-1}\|^2 + 2\|ACA^{-1}\|^2,$$

$$\|AE\|^2 + \|E\|^2 + 1 + 2\|A_x\|^2 + \frac{3}{2}\|AA_x\|^2$$

$$+ \frac{(3+\epsilon)}{2}\|A_xA\|^2\Big\},$$

$$C_1 \equiv \frac{5}{4} - \frac{1}{\epsilon}\|A^{-1}\|^2, \tag{4.5b}$$

$$C_2 \equiv 2 + 4\|A\|^2. \tag{4.5c}$$

We next make substitutions using (4. 3a) and (4. 3b), simplify (4. 4) using the boundary conditions, the Schwarz inequality, and the generalized geometric inequality, multiply both sides of (4. 4) by the integrating factor $2e^{-Kt}$, and integrate both sides from 0 to t:

$$\|A\underset{\sim}{e}(t)\|^2 + \|\underset{\sim}{f}(t)\|^2 + \int_0^t e^{K(t-s)} (2\|\underset{\sim}{f}(s)\|^2 + C_1 \|A\underset{\sim}{f}_x(s)\|^2)ds \le e^{Kt}(\|A\underset{\sim}{e}(0)\|^2 + \|\underset{\sim}{f}(0)\|^2) + \int_0^t e^{K(t-s)} \Big\{C_2 \|\underset{\sim}{\sigma}(s)\|^2 + 2\|A\underset{\sim}{e}(s)\| \cdot \|A\underset{\sim}{\sigma}(s)\| + 2\|\underset{\sim}{f}(s)\| \cdot (\|\underset{\sim}{\zeta}(s)\| + \|\underset{\sim}{\rho}(s)\|) + 2[A^2(x)\underset{\sim}{f}(x, s) \cdot \underset{\sim}{e}_t(x, s)]_0^1 \Big\} ds . \tag{4.6}$$

We have assumed ε has been chosen so that C_1 is positive. In further simplifying the boundary terms it will turn out that C_3, another constant, arises naturally:

$$C_3 \equiv A^2(0)\,\beta_0^{-1}\,\alpha_0\,\underset{\sim}{e}(0, 0) \cdot \underset{\sim}{e}(0, 0) + A^2(1)\,\beta_1^{-1}\,\alpha_0\,\underset{\sim}{e}(1, 0) \cdot \underset{\sim}{e}(1, 0) + \|A\underset{\sim}{e}(0)\|^2 + \|\underset{\sim}{f}(0)\|^2 . \tag{4.7}$$

Inequality (4. 6) is also valid if we replace t by τ on the left side and if $0 \le \tau \le t$. We integrate τ from 0 to t, simplify the boundary terms, and use the arithmetic-geometric inequality to obtain

$$\int_0^t (\|A\underset{\sim}{e}(s)\|^2 + \|\underset{\sim}{f}(s)\|^2)ds \le 2\, C_3\, t\, e^{Kt} \tag{4.8}$$

$$+ 4\, t\, e^{Kt} \int_0^t e^{-Ks}\, C_2\, \|\underset{\sim}{\sigma}(s)\|^2\, ds$$

$$+ 2\, t^2\, e^{2Kt} \int_0^t e^{-2Ks}\, [\|A\underset{\sim}{\sigma}(s)\|^2 + 2\, \|\underset{\sim}{\varsigma}(s)\|^2$$

$$+ 2\, \|\underset{\sim}{\rho}(s)\|^2\,]ds\, .$$

Inequality (4. 6) may also be reduced using the geometric inequality to

$$\|A\underset{\sim}{e}(t)\|^2 + \|\underset{\sim}{f}(t)\|^2 \le C_3\, e^{Kt} \tag{4.9}$$

$$+ e^{Kt} \int_0^t e^{-Ks} \Big\{C_2\, \|\underset{\sim}{\sigma}(s)\|^2 + e^{-Ks}\|A\underset{\sim}{\sigma}(s)\|^2$$

$$+ 2e^{-Ks}\, \|\underset{\sim}{\varsigma}(s)\|^2 + 2\, e^{-Ks}\|\, \underset{\sim}{\rho}(s)\|^2\Big\}\, ds$$

$$+ e^{Kt} \Big\{ \int_0^t (\|A\underset{\sim}{e}(s)\|^2 + \|\underset{\sim}{f}(s)\|^2)ds \Big\}\, .$$

Inequality (4. 8) is now substituted into the braces in (4. 9):

$$\|A\underset{\sim}{e}(t)\|^2 + \|\underset{\sim}{f}(t)\|^2 \le C_3\, e^{Kt}\, (1 + 2t\, e^{Kt}) \tag{4.10}$$

$$+ C_2\, e^{Kt}\, (1 + 4\, t\, e^{Kt}) \int_0^t e^{-Ks}\|\underset{\sim}{\sigma}(s)\|^2 ds$$

$$+ e^{Kt}(1 + 2t^2\, e^{2Kt}) \int_0^t e^{-2Ks}\|A\underset{\sim}{\sigma}(s)\|^2 ds$$

$$+ 2e^{Kt}(1 + 2t^2\, e^{2Kt}) \int_0^t e^{-2Ks}\, \|\underset{\sim}{\varsigma}(s)\|^2 ds$$

$$+ 2e^{Kt}(1 + 2t^2\, e^{2Kt}) \int_0^t e^{-2Ks}\, \|\underset{\sim}{\rho}(s)\|^2\, ds\, .$$

(4.10) is the desired form of energy inequality. It tells us that $\underset{\sim}{e}$ and $\underset{\sim}{f}$ can be bounded for a fixed time t in terms of $\underset{\sim}{\rho}$, $\underset{\sim}{\sigma}$, $\underset{\sim}{\zeta}$, and C_3 which depends on the initial conditions.

In conclusion we give a brief summary of the technique of energy inequalities as used in our work. One first derives a differential inequality for suitable variables such as $\|A\,\underset{\sim}{e}(t)\|^2 + \|\underset{\sim}{f}(t)\|^2$. The differential inequality is solved in the manner of Gronwall's inequality. This process can be used both for continuous and discrete equations, but since the discrete case tends to be more complicated, we first derive the continuous inequality to use as a model.

I. 5 Convergence of the Box Scheme for a Linear Parabolic System

In this section we wish to analyze the convergence of the box scheme applied to problem (4. 1) using a discrete analogue of the energy inequality derived in Section 4. Let $\{\underset{\sim}{u}_j^n\}$ and $\{\underset{\sim}{v}_j^n\}$ be net functions approximating $\underset{\sim}{U}$ and $\underset{\sim}{V}$ the solutions of (4. 1). We make the same assumptions on the matrices as before - namely (4. 1g), (4. 1h), and (4. 1i). The finite difference equations are

$$A^2_{j-\frac{1}{2}} D_x^- \underset{\sim}{u}_j^n = \underset{\sim}{v}^n_{j-\frac{1}{2}}, \tag{5. 1a}$$

$$D_x^- \underset{\sim}{v}_j^{n-\frac{1}{2}} = D_t^- \underset{\sim}{u}^n_{j-\frac{1}{2}} - C_{j-\frac{1}{2}} \underset{\sim}{u}^{n-\frac{1}{2}}_{j-\frac{1}{2}} - E_{j-\frac{1}{2}} \underset{\sim}{v}^{n-\frac{1}{2}}_{j-\frac{1}{2}} - \underset{\sim}{S}^{n-\frac{1}{2}}_{j-\frac{1}{2}}, \tag{5. 1b}$$

$$\underset{\sim}{u}_j^1 = \underset{\sim}{g}(x_j), \tag{5. 1c}$$

$$\underset{\sim}{v}_j^1 = A^2(x_j)\frac{d\,\underset{\sim}{g}(x_j)}{dx}, \tag{5. 1d}$$

$$\alpha_0 \underset{\sim}{u}_1^n - \beta_0 \underset{\sim}{v}_1^n = \underset{\sim}{g}_0(t_n), \tag{5. 1e}$$

$$\alpha_1 \underset{\sim}{u}_J^n + \beta_1 \underset{\sim}{v}_J^n = \underset{\sim}{g}_1(t_n). \tag{5. 1f}$$

We define the errors

$$\underset{\sim}{e}_j^n \equiv \underset{\sim}{U}(x_j, t_n) - \underset{\sim}{u}_j^n, \tag{5. 2a}$$

$$\underset{\sim}{f}_j^n \equiv \underset{\sim}{V}(x_j, t_n) - \underset{\sim}{v}_j^n. \tag{5. 2b}$$

The errors $\{\underset{\sim}{e}_j^n\}$ and $\{\underset{\sim}{f}_j^n\}$ satisfy

$$A^2_{j-\frac{1}{2}} D_x^- \underset{\sim}{e}_j^n = \underset{\sim}{f}^n_{j-\frac{1}{2}} + \underset{\sim}{\rho}^n_{j-\frac{1}{2}} , \tag{5.3a}$$

$$\begin{aligned} D_x^- \underset{\sim}{f}_j^{n-\frac{1}{2}} &= D_t^- \underset{\sim}{e}^n_{j-\frac{1}{2}} - C_{j-\frac{1}{2}} \underset{\sim}{e}^{n-\frac{1}{2}}_{j-\frac{1}{2}} \\ &\quad - E_{j-\frac{1}{2}} f^{n-\frac{1}{2}}_{j-\frac{1}{2}} + \sigma^{n-\frac{1}{2}}_{j-\frac{1}{2}} , \end{aligned} \tag{5.3b}$$

$$D_t^- A^2_{j-\frac{1}{2}} D_x^- \underset{\sim}{e}_j^n = D_t^- \underset{\sim}{f}^n_{j-\frac{1}{2}} + \underset{\sim}{\zeta}^{n-\frac{1}{2}}_{j-\frac{1}{2}} \tag{5.3c}$$

$$\alpha_0 \underset{\sim}{e}_1^n - \beta_0 \underset{\sim}{f}_1^n = \underset{\sim}{0} , \tag{5.3d}$$

$$\alpha_1 \underset{\sim}{e}_J^n + \beta_1 \underset{\sim}{f}_J^n = \underset{\sim}{0} , \tag{5.3e}$$

where $\left\{\underset{\sim}{\rho}^n_{j-\frac{1}{2}}\right\}$, $\left\{\underset{\sim}{\sigma}^{n-\frac{1}{2}}_{j-\frac{1}{2}}\right\}$, and $\left\{\underset{\sim}{\zeta}^{n-\frac{1}{2}}_{j-\frac{1}{2}}\right\}$ are truncation errors. As in the case of the heat equation, they will all be $O(h^2)$. In place of the function $\exp(-Kt)$ we will use its discrete analogue:

$$\left.\begin{aligned} g^1 &\equiv 1 , \\ g^n &\equiv g(t_n) \equiv \prod_{m=2}^{n} \left(\frac{2-K\,k_m}{2+K\,k_m}\right) \text{ for } n \geq 2 , \end{aligned}\right\} \tag{5.4}$$

where $k_m < 2/K$ for all m.

We begin by taking the dot products of $\underset{\sim}{f}^{n-\frac{1}{2}}_{j-\frac{1}{2}}$ with (5.3c) and $A^2_{j-\frac{1}{2}} \underset{\sim}{e}^{n-\frac{1}{2}}_{j-\frac{1}{2}}$ with (5.3b), adding the products, multiplying by h_j, and summing from $j = 2$ to $j = J$. We would like to sum by parts $(A^2 \underset{\sim}{f}^{n-\frac{1}{2}}, D_x^- D_t^- \underset{\sim}{e}^n)_h$ and $(A^2 e^{n-\frac{1}{2}}, D_x^- \underset{\sim}{f}^{n-\frac{1}{2}})_h$, but now we are faced with the problem that, for instance, $A^2 \underset{\sim}{f}^{n-\frac{1}{2}}$ means evaluating A^2

at $x_{j-\frac{1}{2}}$ and multiplying by the averaged value of $\underset{\sim}{f}_j^{n-\frac{1}{2}}$ and $\underset{\sim}{f}_{j-1}^{n-\frac{1}{2}}$ rather than averaging $A^2(x_{j-1})\underset{\sim}{f}_{j-1}^{n-\frac{1}{2}}$ and $A^2(x_j)\,\underset{\sim}{f}_j^{n-\frac{1}{2}}$. The summation by parts formula (3.7a) requires the latter quantity. In order to proceed we shall reinterpret $A^2\,\underset{\sim}{f}^{n-\frac{1}{2}}$ and $A^2\,\underset{\sim}{e}^{n-\frac{1}{2}}$ to fit the formula, but we shall then have to accept new terms $\underset{\sim}{w}^{n-\frac{1}{2}}$ and $\underset{\sim}{y}^{n-\frac{1}{2}}$ which account for the difference between the terms we actually have and the terms we need for summation by parts. An analysis of $\underset{\sim}{w}^{n-\frac{1}{2}}$ and $\underset{\sim}{y}^{n-\frac{1}{2}}$ is not needed now so that it will be postponed to a later section. At the present time we simply accept $\underset{\sim}{w}^{n-\frac{1}{2}}$ and $\underset{\sim}{y}^{n-\frac{1}{2}}$ as net functions which make the following modified summation by parts formulas true:

$$(A^2\,\underset{\sim}{f}^{n-\frac{1}{2}},\ D_x^-\,D_t^-\,\underset{\sim}{e}^n)_h = \left[A^2\,\underset{\sim}{f}^{n-\frac{1}{2}}\cdot D_t^-\,\underset{\sim}{e}^n\right]_1^J$$

$$- (D_x^-\,A^2\,\underset{\sim}{f}^{n-\frac{1}{2}},\ D_t^-\,\underset{\sim}{e}^n)_h - (\underset{\sim}{w}^{n-\frac{1}{2}},\ D_x^-\,D_t^-\,\underset{\sim}{e}^n)_h\,, \tag{5.5a}$$

$$(A^2\,\underset{\sim}{e}^{n-\frac{1}{2}},\ D_x^-\,\underset{\sim}{f}^{n-\frac{1}{2}})_h = \left[A^2\,\underset{\sim}{e}^{n-\frac{1}{2}}\cdot\underset{\sim}{f}^{n-\frac{1}{2}}\right]_1^J$$

$$- (D_x^-\,A^2\,\underset{\sim}{e}^{n-\frac{1}{2}},\ \underset{\sim}{f}^{n-\frac{1}{2}})_h - (\underset{\sim}{y}^{n-\frac{1}{2}},\ D_x^-\,\underset{\sim}{f}^{n-\frac{1}{2}})_h\,. \tag{5.5b}$$

With (5.5), the analogue of (4.4) becomes

$$\frac{1}{2} D_t^- \|A \underset{\sim}{e}^n\|_h^2 + \frac{1}{2} D_t^- \|\underset{\sim}{f}^n\|_h^2 = -\left(\underset{\sim}{f}^{n-\frac{1}{2}}, \underset{\sim}{\zeta}^{n-\frac{1}{2}}\right)_h \tag{5.6}$$

$$+ (A^2 \underset{\sim}{e}^{n-\frac{1}{2}}, C \underset{\sim}{e}^{n-\frac{1}{2}})_h + (A^2 \underset{\sim}{e}^{n-\frac{1}{2}}, E \underset{\sim}{f}^{n-\frac{1}{2}})_h$$

$$- (A^2 \underset{\sim}{e}^{n-\frac{1}{2}}, \underset{\sim}{\sigma}^{n-\frac{1}{2}})_h + [A^2 \underset{\sim}{f}^{n-\frac{1}{2}} \cdot D_t^- \underset{\sim}{e}^n]_1^J$$

$$- (D_x^- A^2 f^{n-\frac{1}{2}}, D_t^- \underset{\sim}{e}^n)_h - (\underset{\sim}{w}^{n-\frac{1}{2}}, D_x^- D_t^- \underset{\sim}{e}^n)_h$$

$$+ [A^2 \underset{\sim}{e}^{n-\frac{1}{2}} \cdot \underset{\sim}{f}^{n-\frac{1}{2}}]_1^J - ((D_x^- A^2) \underset{\sim}{e}^{n-\frac{1}{2}}, \underset{\sim}{f}^{n-\frac{1}{2}})_h$$

$$- (A^2 D_x^- \underset{\sim}{e}^{n-\frac{1}{2}}, \underset{\sim}{f}^{n-\frac{1}{2}})_h - (\underset{\sim}{y}^{n-\frac{1}{2}}, \underset{\sim}{f}^{n-\frac{1}{2}})_h .$$

As a result of the modified summation by parts formulas, the A^2 in the next to last term on the right side of (5.6) means the average of $A^2(x_j)$ and $A^2(x_{j-1})$ rather than $A^2(x_{j-\frac{1}{2}})$. Thus we cannot use (5.3a) to substitute for $A^2 D_x^- \underset{\sim}{e}^{n-\frac{1}{2}}$ in this term. As a matter of fact, we really should have a different notation for this A^2 since it has a different meaning in this term than in the other terms of (5.6). Define

$$\overline{A^2_{j-\frac{1}{2}}} \equiv \frac{1}{2}(A^2(x_j) + A^2(x_{j-1})), \tag{5.7a}$$

$$\tilde{A}_{j-\frac{1}{2}} \equiv \sqrt{\overline{A^2_{j-\frac{1}{2}}}} \; . \tag{5.7b}$$

We define $\tilde{\underset{\sim}{\rho}}^n_{j-\frac{1}{2}}$ by means of

$$\overline{A^2_{j-\frac{1}{2}}} \, D_x^- \underset{\sim}{e}^n_j = \underset{\sim}{f}^n_{j-\frac{1}{2}} + \tilde{\underset{\sim}{\rho}}^n_{j-\frac{1}{2}} . \tag{5.8}$$

We claim now that

$$\tilde{\rho}^{n}_{j-\frac{1}{2}} = \rho^{n}_{j-\frac{1}{2}} + O(h^2), \tag{5.9}$$

although the proof will be postponed. The validity of (5.9) depends both on the smoothness of $A^2(x)$ and the (yet to be defined) "smoothness" of the $\{e^n_j\}$. We add at this time that the terms containing $w^{n-\frac{1}{2}}$ and $y^{n-\frac{1}{2}}$ will also require "smoothness" on the part of the finite difference solution. These points will be fully discussed later. Define two constants K and C_1 by

$$K \equiv \max\Big\{[2 + \|ACA^{-1}\|^2 + \|\tilde{A}CA^{-1}\|^2 + \|CA^{-1}\|^2 + \|(D_x^- A^2)A^{-1}\|^2],\ [\|\tilde{A}^{-1}D_x^- A^2\| + \|AE\|^2 + \|\tilde{A}E\| + 3\|D_x^- A^2\|^2 + 1 + \|E\|^2]\Big\}, \tag{5.10a}$$

$$C_1 \equiv 1 + \|\tilde{A}\|^2. \tag{5.10b}$$

Substitution of (5.3b) and (5.8) into (5.6) followed by applications of the Schwarz, arithmetic-geometric, and triangle inequalities and multiplication by $2\,g^{n-\frac{1}{2}}$ yield

$$D_t^-\Big\{g^n(\|A e^n\|_h^2 + \|f^n\|_h^2)\Big\} \le \tag{5.11}$$

$$g^{n-\frac{1}{2}}\Big\{2[A^2 f^{n-\frac{1}{2}} \cdot D_t^- e^n]_1^J + 2[A^2 e^{n-\frac{1}{2}} \cdot f^{n-\frac{1}{2}}]_1^J - 2(w^{n-\frac{1}{2}}, D_x^- D_t^- e^n)_h - 2(y^{n-\frac{1}{2}}, D_x^- f^{n-\frac{1}{2}})_h + 2\|f^{n-\frac{1}{2}}\|_h \cdot (\|\zeta^{n-\frac{1}{2}}\|_h + \|\tilde{\rho}^{n-\frac{1}{2}}\|_h) + C_1\|\sigma^{n-\frac{1}{2}}\|_h^2 + 2\|A e^{n-\frac{1}{2}}\|_h \cdot \|A\sigma^{n-\frac{1}{2}}\|_h\Big\}.$$

Let

$$C_2 \equiv A^2(0)\beta_0^{-1}\alpha_0\, \underset{\sim}{e}_1^1 \cdot \underset{\sim}{e}_1^1 + A^2(1)\,\beta_1^{-1}\alpha_1\, \underset{\sim}{e}_J^1 \cdot \underset{\sim}{e}_J^1 + \|A\underset{\sim}{e}^1\|_h^2 + \|\underset{\sim}{f}^1\|_h^2 . \tag{5.12}$$

Then a time summation of (5.11) and an analysis of the boundary terms lead to

$$\begin{aligned} \|A\underset{\sim}{e}^n\|_h^2 + \|\underset{\sim}{f}^n\|_h^2 \le \frac{1}{g^n} C_2 & \\ - \frac{2}{g^n}\sum_{m=2}^{n} k_m\, g^{m-\frac{1}{2}}\Big[(\underset{\sim}{w}^{m-\frac{1}{2}}, D_x^- D_t^- \underset{\sim}{e}^m)_h & \\ + (\underset{\sim}{y}^{m-\frac{1}{2}}, D_x^- \underset{\sim}{f}^{m-\frac{1}{2}})_h\Big] + \frac{1}{g^n}\sum_{m=2}^{n} k_m\, g^{m-\frac{1}{2}} \cdot & \\ \Big\{2\|A\underset{\sim}{e}^{m-\frac{1}{2}}\|_h \cdot \|A\underset{\sim}{\sigma}^{m-\frac{1}{2}}\|_h + C_1\|\underset{\sim}{\sigma}^{m-\frac{1}{2}}\|_h^2 & \\ + 2\|\underset{\sim}{f}^{m-\frac{1}{2}}\|_h \cdot (\|\underset{\sim}{\zeta}^{m-\frac{1}{2}}\|_h + \|\underset{\sim}{\tilde{\rho}}^{m-\frac{1}{2}}\|_h)\Big\} . & \end{aligned} \tag{5.13}$$

Let $k_1 \equiv k_2$ and define

$$C_3 \equiv \max_{2 \le m \le N}\left(\frac{k_m}{k_{m-1}}\right) \tag{5.14}$$

On the left side of (5.13) we change n to i then sum from $i = 2$ to $i = n$. Also we use the geometric inequality and the triangle inequality in order to get norms of time averages:

$$\sum_{m=2}^{n} k_m(\|A\underset{\sim}{e}^{m-\frac{1}{2}}\|_h^2 + \|\underset{\sim}{f}^{m-\frac{1}{2}}\|_h^2) \leq \tag{5.15}$$

$$\frac{t_n\, C_2(1+C_3)}{g^n} + C_3\, k_1\, (\|A\underset{\sim}{e}^1\|_h^2 + \|\underset{\sim}{f}^1\|_h^2)$$

$$- \frac{2\, t_n(1+C_3)}{g^n} \sum_{m=2}^{n} k_m\, g^{m-\frac{1}{2}} \Big[(\underset{\sim}{w}^{m-\frac{1}{2}}, D_x^- D_t^- \underset{\sim}{e}^m)_h$$

$$+ (\underset{\sim}{y}^{m-\frac{1}{2}}, D_x^- \underset{\sim}{f}^{m-\frac{1}{2}})_h \Big]$$

$$+ \frac{C_1(1+C_3)t_n}{g^n} \sum_{m=2}^{n} k_m\, g^{m-\frac{1}{2}} \|\underset{\sim}{\sigma}^{m-\frac{1}{2}}\|_h^2$$

$$+ \frac{t_n^2(1+C_3)^2}{(g^n)^2} \sum_{m=2}^{n} k_m (g^{m-\frac{1}{2}})^2 \|A\underset{\sim}{\sigma}^{m-\frac{1}{2}}\|_h^2$$

$$+ \frac{2\, t_n^2(1+C_3)^2}{(g^n)^2} \sum_{m=2}^{n} k_m (g^{m-\frac{1}{2}})^2 (\|\underset{\sim}{\varsigma}^{m-\frac{1}{2}}\|_h^2 + \|\underset{\sim}{\tilde{\rho}}^{m-\frac{1}{2}}\|_h^2).$$

(5.15) corresponds to (4.8). To find the discrete version of (4.9) we return to (5.13) where we estimate the terms involving truncation errors with the arithmetic-geometric inequality:

$$\|A\underset{\sim}{e}^n\|_h^2 + \|\underset{\sim}{f}^n\|_h^2 \leq \frac{1}{g^n} C_2 \tag{5.16}$$

$$- \frac{2}{g^n} \sum_{m=2}^{n} k_m\, g^{m-\frac{1}{2}} \Big[(\underset{\sim}{w}^{m-\frac{1}{2}}, D_x^- D_t^- \underset{\sim}{e}^m)_h + (\underset{\sim}{y}^{m-\frac{1}{2}}, D_x^- \underset{\sim}{f}^{m-\frac{1}{2}})_h \Big]$$

$$+ \frac{1}{g^n} \sum_{m=2}^{n} k_m\, g^{m-\frac{1}{2}} \Big\{ 2\, g^{m-\frac{1}{2}} \|\underset{\sim}{\varsigma}^{m-\frac{1}{2}}\|_h^2$$

$$+ 2\, g^{m-\frac{1}{2}} \|\underset{\sim}{\tilde{\rho}}^{m-\frac{1}{2}}\|_h^2 + C_1 \|\underset{\sim}{\sigma}^{m-\frac{1}{2}}\|_h^2 + g^{m-\frac{1}{2}} \|A\underset{\sim}{\sigma}^{m-\frac{1}{2}}\|_h^2 \Big\}$$

$$+ \frac{1}{g^n} \Big\{ \sum_{m=2}^{n} k_m (\|A\underset{\sim}{e}^{m-\frac{1}{2}}\|_h^2 + \|\underset{\sim}{f}^{m-\frac{1}{2}}\|_h^2) \Big\} .$$

The terms in the last set of braces are estimated with (5.15). Suppose we define

$$\tau_m^2 \equiv \|\underset{\sim}{\zeta}^{m-\frac{1}{2}}\|_h^2 + \|\underset{\sim}{\tilde{\rho}}^{m-\frac{1}{2}}\|_h^2 + \|\underset{\sim}{\sigma}^{m-\frac{1}{2}}\|_h^2, \tag{5.17a}$$

$$\tau^2(n) \equiv \max_{2 \le m \le n} \tau_m^2, \tag{5.17b}$$

$$C_4(t_n) \equiv \max\left\{ \frac{2}{g^n}\left(1 + \frac{t_n^2(1+C_3)^2}{(g^n)^2}\right), \right. \tag{5.17c}$$

$$\left. \frac{1}{g^n}\left[(1 + \|\tilde{A}\|^2)\left(1 + \frac{t_n(1+C_3)}{g^n}\right) + \|A\|^2\left(1 + \frac{t_n^2(1+C_3)^2}{(g^n)^2}\right)\right]\right\}.$$

Then the final inequality may be written as

$$\|A\underset{\sim}{e}^n\|_h^2 + \|\underset{\sim}{f}^n\|_h^2 \le \tag{5.18}$$

$$\frac{1}{g^n}\left(1 + \frac{t_n(1+C_3)}{g^n}\right)\left[A^2(1)\beta_1^{-1}\alpha_0\, \underset{\sim}{e}_1^1 \cdot \underset{\sim}{e}_1^1\right.$$

$$\left. + A^2(1)\beta_1^{-1}\alpha_1\, \underset{\sim}{e}_J^1 \cdot \underset{\sim}{e}_J^1\right]$$

$$+ \frac{1}{g^n}(1+C_3)\left(1 + \frac{t_n}{g^n}\right)\left(\|A\underset{\sim}{e}^1\|_h^2 + \|\underset{\sim}{f}^1\|_h^2\right)$$

$$- \frac{2}{g^n}\left(1 + \frac{t_n(1+C_3)}{g^n}\right)\sum_{m=2}^{n} k_m\, g^{m-\frac{1}{2}} \cdot$$

$$\left[(\underset{\sim}{w}^{m-\frac{1}{2}}, D_x^- D_t^- \underset{\sim}{e}^m)_h + (\underset{\sim}{v}^{m-\frac{1}{2}}, D_x^- \underset{\sim}{f}^{m-\frac{1}{2}})_h\right]$$

$$+ C_4(t_n)\tau^2(n).$$

(5.18) tells us that the error at time t_n can be bounded by an expression involving three types of quantities. The first type consists of the errors $\{\underset{\sim}{e}_j^1\}$ and $\{\underset{\sim}{f}_j^1\}$ made in the initial values. The second type consists of $\{\underset{\sim}{w}_{j-\frac{1}{2}}^{n-\frac{1}{2}}\}$ and $\{\underset{\sim}{y}_{j-\frac{1}{2}}^{n-\frac{1}{2}}\}$ combined with $\{\underset{\sim}{e}_j^n\}$ and $\{\underset{\sim}{f}_j^n\}$ which arose when we summed by parts two terms for which the summation by parts identity was not really valid. The third type involves the truncation errors $\{\underset{\sim}{\sigma}_{j-\frac{1}{2}}^{m-\frac{1}{2}}\}$, $\{\underset{\sim}{\zeta}_{j-\frac{1}{2}}^{m-\frac{1}{2}}\}$ and $\{\tilde{\underset{\sim}{\rho}}_{j-\frac{1}{2}}^{m}\}$. It is clear that the first type of term can be made $O(h^2)$ by a sufficiently accurate approximation of the initial condition. The third type will be $O(h^2)$ if the $\{\tilde{\underset{\sim}{\rho}}_{j-\frac{1}{2}}^{m}\}$ do not differ from the $\{\underset{\sim}{\rho}_{j-\frac{1}{2}}^{m}\}$ by more than $O(h^2)$ as has previously been claimed. We must investigate further the net functions $\{\tilde{\underset{\sim}{\rho}}_{j-\frac{1}{2}}^{m}\}$, $\{\underset{\sim}{w}_{j-\frac{1}{2}}^{m-\frac{1}{2}}\}$, and $\{\underset{\sim}{y}_{j-\frac{1}{2}}^{m-\frac{1}{2}}\}$.

We start with $\{\tilde{\underset{\sim}{\rho}}_{j-\frac{1}{2}}^{m}\}$. It can be shown by Taylor series expansions and (5.3a), (5.7a), (5.8), and (5.9) that

$$\tilde{\underset{\sim}{\rho}}_{j-\frac{1}{2}}^{m} = \underset{\sim}{\rho}_{j-\frac{1}{2}}^{m} + \frac{h_j^2}{8} \cdot \frac{\partial^2 A^2(\xi_{j-\frac{1}{2}})}{\partial x^2} D_x^- \underset{\sim}{e}_j^n , \tag{5.19}$$

where $\xi_{j-\frac{1}{2}}$ is some point between x_{j-1} and x_j. If the second derivative of $A^2(x)$ is uniformly bounded and if $D_x^- \underset{\sim}{e}_j^n$ is $O(1)$, then the $\{\tilde{\underset{\sim}{\rho}}_{j-\frac{1}{2}}^{m}\}$ will be $O(h^2)$. As for the $\{\underset{\sim}{w}_{j-\frac{1}{2}}^{m-\frac{1}{2}}\}$ and $\{\underset{\sim}{y}_{j-\frac{1}{2}}^{m-\frac{1}{2}}\}$, their treatments are similar so we shall consider $\{\underset{\sim}{w}_{j-\frac{1}{2}}^{m-\frac{1}{2}}\}$ alone. Again using Taylor series expansions, we find

$$A^2(x_{j-\frac{1}{2}}) \cdot \frac{\underset{\sim}{f}_j^{n-\frac{1}{2}} + \underset{\sim}{f}_{j-1}^{n-\frac{1}{2}}}{2} = \tag{5.20}$$

$$\frac{A^2(x_j)\, \underset{\sim}{f}_j^{n-\frac{1}{2}} + A^2(x_{j-\frac{1}{2}})\, \underset{\sim}{f}_{j-1}^{n-\frac{1}{2}}}{2}$$

$$+ \frac{h_j}{4} \cdot \frac{\partial A^2(x_{j-\frac{1}{2}})}{\partial x} \left[\underset{\sim}{f}_{j-1}^{n-\frac{1}{2}} - \underset{\sim}{f}_j^{n-\frac{1}{2}} \right]$$

$$+ O(h_j^2)\, \underset{\sim}{f}_j^{n-\frac{1}{2}} + O(h_j^2)\, \underset{\sim}{f}_{j-1}^{n-\frac{1}{2}} .$$

$\underset{\sim}{w}_{j-\frac{1}{2}}^{n-\frac{1}{2}}$ is defined as all the terms on the right side of (5.20) except the first one. The same definition holds for $\underset{\sim}{y}_{j-\frac{1}{2}}^{n-\frac{1}{2}}$ with $\underset{\sim}{f}_j^{n-\frac{1}{2}}$ replaced by $\underset{\sim}{e}_j^{n-\frac{1}{2}}$. If $D_x^- \underset{\sim}{e}_j^n$ and $D_x^- \underset{\sim}{f}_j^n$ are O(1), then $\underset{\sim}{w}_{j-\frac{1}{2}}^{n-\frac{1}{2}}$ and $\underset{\sim}{y}_{j-\frac{1}{2}}^{n-\frac{1}{2}}$ will be $O(h^2)$. Furthermore, the $\left\{\underset{\sim}{w}_{j-\frac{1}{2}}^{m-\frac{1}{2}}\right\}$ and $\left\{\underset{\sim}{y}_{j-\frac{1}{2}}^{m-\frac{1}{2}}\right\}$ occur in inner products and time summations with $D_x^- D_t^- \underset{\sim}{e}_j^n$ and $D_x^- \underset{\sim}{f}_j^n$. These latter must be O(1) to guarantee that our terms of the second type be $O(h^2)$. If U and V are continuously differentiable in x and the cross derivative of U is continuous, then our three conditions are that $D_x^- \underset{\sim}{u}_j^n$, $D_x^- \underset{\sim}{v}_j^n$, and $D_x^- D_t^- \underset{\sim}{u}_j^n$ be O(1) as $h \to 0$ for $\left\{\underset{\sim}{u}_j^n\right\}$ and $\left\{\underset{\sim}{v}_j^n\right\}$ solutions of the difference equations. We shall show in the next section that these conditions can be satisfied.

I.6 Solving the Linear Finite Difference Equations

Problem (5.1) may be written as a block tridiagonal linear algebraic system of equations with $2p \times 2p$ blocks. We would like to show that factorization into upper and lower block triangular matrices can be used to solve this system; however, we currently have a proof only for the case $p = 1$. In this section we restrict ourselves to an equation with constant coefficients and a net with uniform space steps h and uniform time steps k. In the next section on nonlinear parabolic equations we will generalize the proof to include variable coefficients and nonuniform meshes. After we show that the finite difference solution exists and is unique, we can show that it is "smooth" in the sense described at the end of Section 5.

In this section we consider the equation

$$u_t = a\,u_{xx} + b\,u_x + c\,u, \tag{6.1}$$

where a, b, and c are constants with $a > 0$. One can show that the matrix of the system of linear difference equations is

$$\mathscr{A} = \begin{pmatrix} B_1 & C_1 & & & 0 \\ A_2 & B_2 & C_2 & & \\ & A_3 & B_3 & C_3 & \\ & & \ddots & \ddots & C_{J-1} \\ 0 & & & A_J & B_J \end{pmatrix}, \tag{6.2a}$$

where the A_j, B_j, and C_j are given by

$$B_1 = \begin{pmatrix} \alpha_0 & -\beta_0 \\ \frac{h}{k} - \frac{ch}{2} & 1 - \frac{bh}{2} \end{pmatrix}, \tag{6.2b}$$

$$C_1 = \begin{pmatrix} 0 & 0 \\ \frac{h}{k} - \frac{ch}{2} & -1 - \frac{bh}{2} \end{pmatrix}, \tag{6.2c}$$

$$A_j = \begin{pmatrix} 1 & \frac{h}{2a} \\ 0 & 0 \end{pmatrix}, \tag{6.2d}$$

$$B_j = \begin{pmatrix} -1 & \frac{h}{2a} \\ \frac{h}{k} - \frac{ch}{2} & 1 - \frac{bh}{2} \end{pmatrix}, \tag{6.2e}$$

$$C_j = \begin{pmatrix} 0 & 0 \\ \frac{h}{k} - \frac{ch}{2} & -1 - \frac{bh}{2} \end{pmatrix}, \tag{6.2f}$$

$$A_J = \begin{pmatrix} 1 & \frac{h}{2a} \\ 0 & 0 \end{pmatrix}, \tag{6.2g}$$

$$B_J = \begin{pmatrix} -1 & \frac{h}{2a} \\ \alpha_1 & \beta_1 \end{pmatrix} \tag{6.2h}$$

for $j = 2$ to $J - 1$. We wish to show that $\mathcal{A}$ can be factored in upper and lower block triangular matrices with 2×2 blocks in the following manner:

$$\mathcal{A} = \mathcal{L}\mathcal{U}, \tag{6.3a}$$

$$\mathcal{L} = \begin{pmatrix} I & & & \\ L_2 & \ddots & & 0 \\ & \ddots & \ddots & \\ 0 & & L_J & I \end{pmatrix}, \tag{6.3b}$$

$$\mathcal{U} = \begin{pmatrix} U_1 & C_1 & & \\ & U_2 & \ddots & 0 \\ & & \ddots & C_2 \\ 0 & & & U_J \end{pmatrix}. \tag{6.3c}$$

If some right hand side vector $\underset{\sim}{f}$ is given, then we could solve $\mathcal{A}\underset{\sim}{x} = \underset{\sim}{f}$ by first solving $\mathcal{L}\underset{\sim}{w} = \underset{\sim}{f}$ for $\underset{\sim}{w}$ and then solving $\mathcal{U}\underset{\sim}{x} = \underset{\sim}{w}$. $\mathcal{L}$ is clearly nonsingular so that $\underset{\sim}{w}$ can be found by working recursively down through $\mathcal{L}$. The back substitution to find $\underset{\sim}{x}$ will require each of the U_j to be nonsingular. As a matter of fact, $\mathcal{U}$ cannot be constructed unless each U_j is nonsingular for $j = 1$ to $J - 1$. What we shall do is verify the block tridiagonal factorization and then check

U_J for invertibility. Once this is done we will know that $\mathscr{A}\underset{\sim}{x} = \underset{\sim}{f}$ has a unique solution for each $\underset{\sim}{f}$ or that the box scheme advances the solution uniquely for each time step.

If we multiply $\mathscr{L}$ and $\mathscr{U}$, we see that the following relations must hold:

$$U_1 = B_1 , \tag{6.4a}$$

$$\left.\begin{aligned} L_j &= A_j U_j^{-1} \quad (6.4b) \\ U_j &= B_j - L_j C_{j-1} \quad (6.4c) \end{aligned}\right\} \quad j = 2, \cdots, J.$$

Define

$$e^2 \equiv \frac{\det B_1}{\beta_0 + \dfrac{h\,\alpha_0}{2a}} . \tag{6.5a}$$

It can then be shown by induction, which we omit, that

$$U_i = \begin{pmatrix} \left(\dfrac{1}{e_i}\right)\left(\dfrac{h}{k} - \dfrac{ch}{2}\right) & \dfrac{h}{2a} + \left(-1 - \dfrac{bh}{2}\right)\left(\dfrac{1}{e_i}\right) \\ \dfrac{h}{k} - \dfrac{ch}{2} & 1 - \dfrac{bh}{2} \end{pmatrix}, \tag{6.5b}$$

$$e_{i+1} = - \frac{\det U_i}{-\dfrac{h}{a} + \left(\dfrac{1}{e_i}\right)\left(1 + \dfrac{bh}{2} + \dfrac{h^2}{2ak} - \dfrac{ch^2}{4a}\right)} . \tag{6.5c}$$

These recursions were suggested by Varah [1972]. The first step is to show that all of the e_i are negative. We start with e_2 which is the first of the e_i:

$$e_2 = - \frac{\left(\frac{\alpha_0}{\beta_0} + \frac{h}{k}\right) - h\left(\frac{\alpha_0}{\beta_0} \cdot \frac{b}{2} + \frac{c}{2}\right)}{1 + \frac{\alpha_0}{\beta_0} \cdot \frac{h}{2a}} \quad . \tag{6.6}$$

e_2 will be negative under our hypotheses if h is sufficiently small. Let us suppose that $e_2 \leq -M$ where M is a positive number. We can show under a mild restriction that all of the e_i will be less than or equal to $-M$. From (6.5c) we see that this would imply all of the U_j at least through $J-1$ would be nonsingular. Eliminating U_i between (6.5b) and (6.5c) results in

$$e_{i+1} = \left\{e_i - \frac{2h}{k} \cdot ch + \frac{h}{2} \cdot \left(-b + \frac{h}{ak} - \frac{ch}{2a}\right) \cdot e_i\right\} \cdot \left\{1 - \frac{h}{a} \cdot e_i + \frac{h}{2} \cdot \left(b + \frac{h}{ak} - \frac{ch}{2a}\right)\right\}^{-1} . \tag{6.7}$$

We assume $e_i \leq -M$. If we ask that e_{i+1} also be less than or equal to $-M$, then we are imposing a condition on the right side of (6.7). We find that this condition implies

$$k \leq \frac{2}{\frac{M^2}{a} + Mb + c} \quad . \tag{6.8}$$

A similar examination of $\det U_J$ shows that it is negative if we take into account the boundary conditions and if h is sufficiently small. Hence for a sufficiently fine net $\mathscr{A}$ is nonsingular.

Now that the $\{u_j^n\}$ and $\{v_j^n\}$ are known to exist uniquely, we return to the question of whether $D_x^- u_j^n$, and $D_t^- D_x^- u_j^n$ are all O(1). An examination of the difference equations shows that it would be sufficient to show $D_x^- v_j^n$ and $D_t^- v_j^n$ were O(1) for all j and n. We will do this in an inductive manner using an argument similar to one given by Strang [1960]. The essence of the argument is to interpolate the finite difference solution $\{u_j^n\}$ and $\{v_j^n\}$ at time t_n with functions $\overline{U}$ and $\overline{V}$. We insist that $\overline{U}$ and $\overline{V}$ be sufficiently smooth at time t_n and that the coefficients and boundary conditions of the differential equation be sufficiently smooth so that $\overline{U}$ and $\overline{V}$ will have five continuous derivatives at time t_{n+1}. The difference between $\{u_j^{n+1}\}$ and $\overline{U}$ at time t_{n+1} and $\{v_j^{n+1}\}$ and $\overline{V}$ at time t_{n+1} will then be equal to the first principal error terms which are $O(h^2)$ plus some residual terms. The point of the argument is to show that the residual terms are at worst $O(h^2)$. Then since $\overline{U}$ and $\overline{V}$ are smooth, $\{u_j^{n+1}\}$ and $\{v_j^{n+1}\}$ will be "smooth" also.

We begin by introducing additional notation. Let

$\begin{pmatrix} \underset{\sim}{u}^n \\ \underset{\sim}{v}^n \end{pmatrix}$ be a vector consisting of the $\{u_j^n\}$ and $\{v_j^n\}$.

Let $\begin{pmatrix} \overline{U}^n \\ \overline{V}^n \end{pmatrix}$ be a vector consisting of continuous functions $\overline{U}(x, t)$ and $\overline{V}(x, t)$ evaluated at x_j and t_n. $\begin{pmatrix} \underset{\sim}{e}^{(1,n)} \\ \underset{\sim}{f}^{(1,n)} \end{pmatrix}$ are the first principal error terms for $\overline{U}$ and $\overline{V}$ in the Richardson extrapolation of the finite difference solutions at time t_n [Keller, 1971]. $A^{n+\frac{1}{2}}$ is the matrix multiplying the vector of unknowns $\{u_j^{n+1}\}$ and $\{v_j^{n+1}\}$. $B^{n+\frac{1}{2}}$ is the matrix multiplying the vector of knowns $\{u_j^n\}$ and $\{v_j^n\}$. $\underset{\sim}{f}^{n+\frac{1}{2}}$ is a vector of inhomogeneous and boundary data. We define $\underset{\sim}{w}^n$ by

$$\begin{pmatrix} \underset{\sim}{u}^n \\ \underset{\sim}{v}^n \end{pmatrix} - \begin{pmatrix} \underset{\sim}{\overline{U}}^n \\ \underset{\sim}{\overline{V}}^n \end{pmatrix} = \begin{pmatrix} \underset{\sim}{e}^{(1,n)} \\ \underset{\sim}{f}^{(1,n)} \end{pmatrix} + \underset{\sim}{w}^n . \qquad (6.9)$$

(6.9) says that the difference between the net functions $\{u_j^n\}$ and $\{v_j^n\}$ and the functions $\overline{U}(x,t)$ and $\overline{V}(x,t)$ evaluated at the net points is equal to the first principal error terms which involve $\overline{U}$ and $\overline{V}$ plus some residual vector $\underset{\sim}{w}^n$. The system representing all of the finite difference equations and boundary conditions in the box scheme for advancing from time t_n to t_{n+1} can be written as

$$A^{n+\frac{1}{2}} \begin{pmatrix} \underset{\sim}{u}^{n+1} \\ \underset{\sim}{v}^{n+1} \end{pmatrix} = B^{n+\frac{1}{2}} \begin{pmatrix} \underset{\sim}{u}^n \\ \underset{\sim}{v}^n \end{pmatrix} + \underset{\sim}{f}^{n+\frac{1}{2}} . \qquad (6.10)$$

For the single parabolic equation considered earlier in this section $A^{n+\frac{1}{2}}$ would be $\mathcal{A}$ for all n. At time t_n we construct a smooth function of x which interpolates $\{u_j^n\}$ and has derivatives matching $\{v_j^n\}$. Let this function be an initial condition for an initial boundary value problem starting at time t_n and having the same boundary conditions as the continuous problem we are discretizing. Let $\overline{U}$ and $\overline{V}$ be the solutions of this problem, and let them have five continuous derivatives. This will in general require the initial condition, boundary data, and inhomogeneous terms in the differential equation to satisfy some differentiability conditions. Notice that $\overline{U}$ and $\overline{V}$ are not the same as U and V. The latter are solutions of an initial boundary value problem starting at time zero and which we are trying to approximate by $\{u_j^n\}$ and $\{v_j^n\}$ while the former are solutions of a problem starting at time t_n with initial data based on $\{u_j^n\}$ and $\{v_j^n\}$. The principal error terms are generally functions of U and V, but here we are substituting $\overline{U}$ and $\overline{V}$. Combining (6. 9) and (6. 10), we can show that

$$A^{n+\frac{1}{2}} \underset{\sim}{w}^{n+1} = B^{n+\frac{1}{2}} \underset{\sim}{w}^{n} + \Bigg\{ \underset{\sim}{f}^{n+\frac{1}{2}} - A^{n+\frac{1}{2}} \begin{pmatrix} \overline{\underset{\sim}{U}}^{n+1} \\ \overline{\underset{\sim}{V}}^{n+1} \end{pmatrix} + B^{n+\frac{1}{2}} \begin{pmatrix} \overline{U}^{n} \\ \overline{V}^{n} \end{pmatrix} - A^{n+\frac{1}{2}} \begin{pmatrix} \underset{\sim}{e}^{(1,n+1)} \\ \underset{\sim}{f}^{(1,n+1)} \end{pmatrix} + B^{n+\frac{1}{2}} \begin{pmatrix} \underset{\sim}{e}^{(1,n)} \\ \underset{\sim}{f}^{(1,n)} \end{pmatrix} \Bigg\} . \tag{6. 11}$$

By definition of the principal error terms, the quantities in the braces must add up to a result which is $O(h^4)$. Our choice of $\overline{U}$ and $\overline{V}$ guarantees that $\underset{\sim}{w}^n = \underset{\sim}{0}$. Therefore $\underset{\sim}{w}^{n+1}$ is equal to $(A^{n+\frac{1}{2}})^{-1}$ multiplying a vector whose terms are $O(h^4)$. The norm of the inverse of $A^{n+\frac{1}{2}}$ is at worst $O(h^{-2})$ so that $\underset{\sim}{w}^{n+1} = O(h^2)$. Since the first principal error terms are also $O(h^2)$, the left side of (6. 9) must be $O(h^2)$. In particular since $\overline{V}$ is smooth and v_j^{n+1} differs by only $O(h^2)$ at the point (x_j, t_{n+1}), $D_x^- v_j^{n+1}$ and $D_t^- v_j^{n+1}$ must be $O(1)$; hence, the desired smoothness conditions on the finite difference solutions can be satisfied, and the convergence proof is essentially complete. For the linear parabolic system where $p > 1$ and for which we do not have a proof of nonsingularity based on block factorization, we will include nonsingularity as an assumption. The proof of smoothness will then be formally identical to the one we have just given for $p = 1$.

A further remark is that while $\overline{U}$ and $\overline{V}$ may be taken to have an arbitrary number of derivatives, the magnitude of the derivatives need not be $O(1)$. In particular if at time zero there is a sharp change in the initial data over an interval of length h, then our present analysis is not adequate to show that the finite difference solutions will be smooth. On the other hand if the derivatives of the solutions U and V are small compared to the inverse of the mesh spacings, then the preceding argument when applied at each time step for $O(h^{-1})$ time steps shows that no oscillations greater than $O(h)$ can form. We recall from an earlier discussion that $\|\cdot\|_h$ is a seminorm and that for certain boundary conditions $\{u_j^n\}$ or $\{v_j^n\}$ might have oscillations. It is now clear that these oscillations will not be worse than $O(h)$ unless U and V have derivatives which are large compared to h^{-1}. Finally, it should be noted that such small amplitude oscillations are allowed under our definition of smoothness for net functions; that is, smoothness and freedom from oscillations are not equivalent.

We summarize the results of Sections 4, 5, and 6 in the following theorem.

Theorem 1: Assume (1) that the box scheme formulation (5.1) of the linear parabolic system (4.1) has a unique solution and (2) that the coefficients of (4.1) are sufficiently smooth so that an initial boundary value problem posed at any non-negative time with piecewise five times continuously differentiable initial functions will have solutions which are also five times piecewise continuously differentiable. If points of discontinuity of the derivatives are always taken to be mesh points, then the box scheme solution converges to the continuous solution of (4.1) as the mesh is refined, and the errors are $O(h^2)$.

I. 7 Nonlinear Parabolic Equations

In this section we wish to study the problem

$$V = a(x, U)U_x , \tag{7.1a}$$

$$V_x = U_t - S(x, t, U, V) , \tag{7.1b}$$

$$U(x, 0) = g(x) , \tag{7.1c}$$

$$V(x, 0) = a(x, g(x)) \frac{dg(x)}{dx} , \tag{7.1d}$$

$$\alpha_0 \rho(0, t) - \beta_0 V(0, t) = g_0(t) , \tag{7.1e}$$

$$\alpha_1 U(1, t) + \beta_1 V(1, t) = g_1(t) , \tag{7.1f}$$

$$\frac{\alpha_0}{\beta_0} \geq 0 , \tag{7.1g}$$

$$\frac{\alpha_1}{\beta_1} \geq 0 , \tag{7.1h}$$

$$0 < a_* \leq a(x, U) \leq a^* < \infty , \tag{7.1i}$$

$$|a_u(x, U)| \leq a^{**} < \infty , \tag{7.1j}$$

$$|S(x, t, U, V)| \leq s^* < \infty , \tag{7.1k}$$

$$|S_U(x, t, U, V)| \leq S^* < \infty , \tag{7.1\ell}$$

$$|S_V(x, t, U, V)| \leq S^* , \tag{7.1m}$$

where in (7. 1i) through (7. 1m) the inequalities hold uniformly in x, t, U, and V. The box scheme applied to this problem yields the following finite difference equations:

$$a(x_{j-\frac{1}{2}}, u_{j-\frac{1}{2}}^{n-\frac{1}{2}}) D_x^- u_j^{n-\frac{1}{2}} = v_{j-\frac{1}{2}}^{n-\frac{1}{2}} , \tag{7.2a}$$

$$D_x^- v_j^{n-\frac{1}{2}} = D_t^- u_{j-\frac{1}{2}}^n - S(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}, u_{j-\frac{1}{2}}^{n-\frac{1}{2}}, v_{j-\frac{1}{2}}^{n-\frac{1}{2}}) , \tag{7.2b}$$

$$u_j^1 = g(x_j) , \tag{7.2c}$$

$$v_j^1 = a(x_j, g(x_j)) \frac{dg(x_j)}{dx} , \tag{7.2d}$$

$$\alpha_0 u_1^n - \beta_0 v_1^n = g_0(t_n) , \tag{7.2e}$$

$$\alpha_1 u_J^n + \beta_1 v_J^n = g_1(t_n) . \tag{7.2f}$$

The domain of this problem is $0 \leq x \leq 1$ and $0 \leq t \leq T$. The net is the same as that used earlier.

We would first like to discover under what conditions (7.2) will have a unique solution. Furthermore we would like to know how to construct the solution. Let us then consider the matter of advancing the finite difference solution from time t_{n-1} to time t_n. Basically we have a nonlinear system of equations in the form $\underset{\sim}{F}(\underset{\sim}{y}) = \underset{\sim}{0}$ where $\underset{\sim}{y}$ is a vector consisting of the unknown $\{u_j^n\}$ and $\{v_j^n\}$ arranged in the order $(u_1^n, v_1^n, u_2^n, v_2^n, \ldots, u_J^n, v_J^n)^t$. The equations are ordered in the following way:

$$F_1(\underset{\sim}{y}) \equiv \alpha_0 u_1^n - \beta_0 v_1^n - g_0^n , \tag{7.3a}$$

$$F_{2j-2}(\underset{\sim}{y}) \equiv -2\,h_j\Big[D_x^- v_j^{n-\frac{1}{2}} - D_t^- u_{j-\frac{1}{2}}^n$$

$$+ S(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}, u_{j-\frac{1}{2}}^{n-\frac{1}{2}}, v_{j-\frac{1}{2}}^{n-\frac{1}{2}})\Big], \tag{7.3b}$$

$$F_{2j-1}(\underset{\sim}{y}) \equiv -2\,h_j\Big[a(x_{j-\frac{1}{2}}, u_{j-\frac{1}{2}}^{n-\frac{1}{2}})D_x^- u_j^{n-\frac{1}{2}} - v_{j-\frac{1}{2}}^{n-\frac{1}{2}}\Big], \tag{7.3c}$$

$$F_{2J}(\underset{\sim}{y}) \equiv \alpha_1 u_J^n + \beta_1 v_J^n - g_1^n, \tag{7.3d}$$

where j ranges from 2 to J. We wish to solve this system iteratively using the chord method so we next calculate $\mathcal{J}$, the Jacobian of $\underset{\sim}{F}$. The order of the unknowns and the equations was chosen so that $\mathcal{J}$ would be block tridiagonal:

$$\mathcal{J} = \begin{pmatrix} B_1 & C_1 & & & & \\ A_2 & B_2 & C_2 & & 0 & \\ & A_3 & B_3 & C_3 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & & & C_{J-1} \\ & & & & A_J & B_J \end{pmatrix} \tag{7.4a}$$

Elements of the blocks are labeled in the following manner:

$$A_j \equiv \begin{pmatrix} A_j^1 & A_j^2 \\ 0 & 0 \end{pmatrix}, \tag{7.4b}$$

$$B_j = \begin{pmatrix} B_j^1 & B_j^2 \\ B_j^3 & B_j^4 \end{pmatrix} , \tag{7.4c}$$

$$C_j \equiv \begin{pmatrix} 0 & 0 \\ C_j^3 & C_j^4 \end{pmatrix} . \tag{7.4d}$$

If we use the notation

$$a_{j-\frac{1}{2}}^{n-\frac{1}{2}} \equiv a(x_{j-\frac{1}{2}}, u_{j-\frac{1}{2}}^{n-\frac{1}{2}}) , \tag{7.4e}$$

$$a_{u_{j-\frac{1}{2}}}^{n-\frac{1}{2}} \equiv \frac{\partial}{\partial u} a(x_{j-\frac{1}{2}}, u_{j-\frac{1}{2}}^{n-\frac{1}{2}}) , \tag{7.4f}$$

$$S_{u_{j-\frac{1}{2}}}^{n-\frac{1}{2}} \equiv \frac{\partial}{\partial u} S(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}, u_{j-\frac{1}{2}}^{n-\frac{1}{2}}, v_{j-\frac{1}{2}}^{n-\frac{1}{2}}) , \tag{7.4g}$$

$$S_{v_{j-\frac{1}{2}}}^{n-\frac{1}{2}} \equiv \frac{\partial}{\partial v} S(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}, u_{j-\frac{1}{2}}^{n-\frac{1}{2}}, v_{j-\frac{1}{2}}^{n-\frac{1}{2}}) , \tag{7.4h}$$

the matrix elements take the form

$$B_1^1 \equiv \alpha_0 , \tag{7.4i}$$

$$B_1^2 \equiv -\beta_0 , \tag{7.4j}$$

$$B_J^3 \equiv \alpha_1 , \tag{7.4k}$$

$$B_J^4 \equiv \beta_1 , \tag{7.4\ell}$$

$$A_j^1 \equiv -\frac{h_j}{2} a_{u_{j-\frac{1}{2}}}^{n-\frac{1}{2}} D_x^- u_j^{n-\frac{1}{2}} + a_{j-\frac{1}{2}}^{n-\frac{1}{2}} , \tag{7.4m}$$

$$A_j^2 \equiv \frac{h_j}{2} , \qquad (7.4n)$$

$$B_j^1 \equiv -\frac{h_j}{2} a_{u_{j-\frac{1}{2}}}^{n-\frac{1}{2}} D_x^- u_j^n - a_{j-\frac{1}{2}}^{n-\frac{1}{2}} , \qquad (7.4o)$$

$$B_j^2 \equiv \frac{h_j}{2} , \qquad (7.4p)$$

$$B_j^3 \equiv -\frac{h_{j+1}}{2} S_{u_{j+1-\frac{1}{2}}}^{n-\frac{1}{2}} + \frac{h_{j+1}}{k_n} , \qquad (7.4q)$$

$$B_j^4 \equiv -\frac{h_{j+1}}{2} S_{v_{j+1-\frac{1}{2}}}^{n-\frac{1}{2}} + 1 , \qquad (7.4r)$$

$$C_j^3 \equiv B_j^3 , \qquad (7.4s)$$

$$C_j^4 \equiv -\frac{h_{j+1}}{2} S_{v_{j+1-\frac{1}{2}}}^{n-\frac{1}{2}} - 1 . \qquad (7.4t)$$

Since the matrix $\mathcal{J}$ comes from linearizing a system of equations, it has the same general form as the matrix for the box scheme solution of a linear parabolic equation with variable coefficients. We have already examined the special case of a linear equation with constant coefficients and uniform net spacing, and we shall use the previous study as a model for the current nonlinear problem. As before we wish to show that $\mathcal{J}$ can be factored as $\mathcal{L}\mathcal{U}$ where $\mathcal{L}$ and $\mathcal{U}$ are lower and upper block triangular (in fact bidiagonal) matrices. We shall use the same notation as in (6.3) and (6.4).

As before we define a sequence $\{e_j : j = 2, \cdots, J\}$ by

$$e_2 \equiv - \frac{\det B_1}{\beta_0 A_2^1 + \alpha_0 A_2^2} , \tag{7.5a}$$

$$U_j = \begin{pmatrix} B_j^1 + \frac{1}{e_j} C_{j-1}^3 & B_j^2 + \frac{1}{e_j} C_{j-1}^4 \\ B_j^3 & B_j^4 \end{pmatrix}, \tag{7.5b}$$

$$e_{j+1} = - \frac{\det U_j}{A_{j+1}^1\left(- B_j^2 - \frac{1}{e_j} C_{j-1}^4\right) + A_{j+1}^2\left(B_j^1 + \frac{1}{e_j} C_{j-1}^3\right)} . \tag{7.5c}$$

We wish to show that all the e_j will be negative. This is equivalent to showing U_j is nonsingular for $j = 1, \cdots , J-1$. We start with e_2. In the case where $\beta_0 = 0$, e_2 will be negative for h_2 sufficiently small. If $\beta_0 \neq 0$, it will in general also be required that

$$- \tfrac{1}{2} a_{u_{2-\frac{1}{2}}}^{n-\frac{1}{2}} (u_2^{n-\frac{1}{2}} - u_1^{n-\frac{1}{2}}) + a_{2-\frac{1}{2}}^{n-\frac{1}{2}} > 0 \tag{7.6}$$

as $h_2 \to 0$ so that $e_2 < 0$ for sufficiently small h_2. We shall say more about this requirement later. Let us assume now that $e_2 \leq -M$ where $M > 0$. We wish to show that $e_j \leq -M$ for $j = 2, \cdots, J$. Let us then assume it is true for j and see what condition is necessary to insure it for $j + 1$. Multiply the numerator and denominator on the right side of (7.5c) by $(A_{j+1}^1 C_{j-1}^4 / e_j)^{-1}$. We introduce new notation for several important groupings of terms:

$$D_{j+1} \equiv B_j^1 \; B_j^4 \, (A_{j+1}^1 \, C_{j-1}^4)^{-1} , \tag{7.7a}$$

$$E_{j+1} \equiv B_j^2 \; B_j^3 (A_{j+1}^1 \, C_{j-1}^4)^{-1} \; h_j^{-1} , \tag{7.7b}$$

$$G_{j+1} \equiv \frac{A_{j+1}^1 \, B_j^2 - B_j^1 \, A_{j+1}^2}{A_{j+1}^1 \, C_{j-1}^4 \, h_j} , \tag{7.7c}$$

$$H_{j+1} \equiv \frac{k_n}{h_j} \cdot \frac{C_{j-1}^3 \, B_j^4 - B_j^3 \, C_{j-1}^4}{A_{j+1}^1 \, C_{j-1}^4} , \tag{7.7d}$$

$$I_{j+1} \equiv \frac{A_{j+1}^2 \; C_{j-1}^3}{A_{j+1}^1 C_{j-1}^4 \, h_j} . \tag{7.7e}$$

If $u_j^{n-\frac{1}{2}} - u_{j-1}^{n-\frac{1}{2}} = O(h)$ as $h \to 0$, then $D_{j+1} \geq 0$ and $E_{j+1} \leq 0$ for h_j and h_{j+1} sufficiently small. If we assume this is the case, we can estimate e_{j+1} by

$$\begin{aligned} e_{j+1} \leq & \left\{ -D_{j+1} M + h_j \, E_{j+1} \, M \right. \\ & \left. + C_{j-1}^3 (A_{j+1}^1 \, C_{j-1}^4)^{-1} \; B_j^4 - B_j^3 (A_{j+1}^1)^{-1} \right\} \cdot \\ & \left\{ 1 + h_j \, G_{j+1} \, e_j - A_{j+1}^2 \, C_{j-1}^3 (A_{j+1}^1 \, C_{j-1}^4)^{-1} \right\} . \end{aligned} \tag{7.8}$$

Under our current assumptions $G_{j+1} \leq 0$ for h_j and h_{j+1} sufficiently small. If we also assume $H_{j+1} < 0$ for h_j, h_{j+1}, and k_n sufficiently small and if we ask for the right side of (7.8) to be less than or equal to $- M$, we arrive at the condition

$$\frac{1}{k_n} \geq \left\{(D_{j+1} - 1 - h_j E_{j+1} + h_j I_{j+1}) M + h_j G_{j+1} M^2\right\} \cdot \left\{h_j H_{j+1}\right\}^{-1} . \tag{7.9}$$

E_{j+1} and I_{j+1} remain bounded as $k_n \to 0$ if and only if h_{j+1}/k_n is bounded which we have already assumed. G_{j+1} is independent of k_n. H_{j+1} is $O(1)$ as $k_n \to 0$. We are thus left with showing that $D_{j+1} - 1$ is $O(h_j)$ as $k_n \to 0$ in order for the right side of (7.9) to be bounded as $k_n \to 0$. If we write out $D_{j+1} - 1$, we will discover that it is $O(h_j)$ if $u_{j+1}^{n-\frac{1}{2}} - u_j^{n-\frac{1}{2}} = O(h)$. We find therefore that if the net function $\{u_j^n\}$ is smooth, then for sufficiently fine net spacing all of the e_j will be less than or equal to $-M$. The requirement that $u_{j+1}^{n-\frac{1}{2}} - u_j^{n-\frac{1}{2}} = O(h)$ is a result of the discretization of $U(x, t)$. It arises only in conjunction with the a_u terms and corresponds to a difference approximation of the derivative. This discrete condition is analogous to asking $U(x, t)$ to have a continuous x derivative. At any rate we now know that the factorization of $\mathcal{J}$ is possible. It further turns out that $\det U_J < 0$ without any additional assumptions so that $\mathcal{J}$ is nonsingular.

We wish to find the solution of $\underset{\sim}{F}(\underset{\sim}{y}) = \underset{\sim}{0}$ by the chord method. This is an iterative method of the form

$$\underset{\sim}{y}^{\nu+1} = \underset{\sim}{y}^{\nu} - A^{-1} \underset{\sim}{F}(\underset{\sim}{y}^{\nu}) . \tag{7.10}$$

We take $\underset{\sim}{y}^{\circ}$, the initial guess for the solution at time t_n, to be the same as the solution already known at time t_{n-1}. In the chord method A is chosen to be the Jacobian of $\underset{\sim}{F}$ evaluated at the initial guess

$\underset{\sim}{y}^{\circ}$. If $\{u_j^{n-1}\}$ is smooth, then A will be nonsingular for a sufficiently fine net. Let

$$\underset{\sim}{g}(\underset{\sim}{y}) \equiv \underset{\sim}{y} - A^{-1} \underset{\sim}{F}(\underset{\sim}{y}) , \tag{7.11}$$

where $\underset{\sim}{y}$ is any vector of length 2J. If we can show that $\underset{\sim}{g}$ is a contracting map in some neighborhood containing $\underset{\sim}{y}^{\circ}$ and that $\underset{\sim}{g}$ maps the neighborhood into itself, then we will know there exists a unique solution of the nonlinear difference equations in that neighborhood. Let $\underset{\sim}{s}$ and $\underset{\sim}{t}$ be vectors.

$$\underset{\sim}{g}(\underset{\sim}{s}) - \underset{\sim}{g}(\underset{\sim}{t}) = \underset{\sim}{s} - \underset{\sim}{t} - A^{-1}[\underset{\sim}{F}(\underset{\sim}{s}) - \underset{\sim}{F}(\underset{\sim}{t})]. \tag{7.12}$$

We apply the mean value theorem with vector $\underset{\sim}{r}$ lying between $\underset{\sim}{s}$ and $\underset{\sim}{t}$:

$$g(s) - g(t) = A^{-1}\left[A - \frac{\partial \underset{\sim}{F}(\underset{\sim}{r})}{\partial \underset{\sim}{y}}\right](\underset{\sim}{s} - \underset{\sim}{t}) . \tag{7.13}$$

We define the matrix M to be the matrix in the brackets in (7.13). Let the vector $\underset{\sim}{r}$ have the components $\{ru_j\}$ and $\{rv_j\}$. We proceed to evaluate the matrix elements $M_{i,j}$:

$$\left.\begin{aligned} M_{1,m} &= 0 , \\ M_{2J,m} &= 0 , \end{aligned}\right\} \quad m = 1, \cdots, 2J , \tag{7.14a}$$

$$\begin{aligned} M_{2j-2,\,2j-3} = \frac{h_j}{2}\Big[&S_u(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}, u_{j-\frac{1}{2}}^{n-\frac{1}{2}}, v_{j-\frac{1}{2}}^{n-\frac{1}{2}}) \\ &- S_u(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}, ru_{j-\frac{1}{2}}, rv_{j-\frac{1}{2}})\Big], \end{aligned} \tag{7.14b}$$

$$M_{2j-2,\,2j-2} = -\frac{h_j}{2}\Big[S_v(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}, u_{j-\frac{1}{2}}^{n-\frac{1}{2}}, v_{j-\frac{1}{2}}^{n-\frac{1}{2}}) - S_v(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}, ru_{j-\frac{1}{2}}, rv_{j-\frac{1}{2}})\Big], \tag{7.14c}$$

$$M_{2j-2,\,2j-1} = -M_{2j-2,\,2j-3}, \tag{7.14d}$$

$$M_{2j-2,\,2j} = M_{2j-2,\,2j-2} \tag{7.14e}$$

for $j = 2, \cdots, J$. All other $M_{2j-2,\,i}$ are zero.

$$M_{2j-1,\,2j-3} = -\frac{h_j}{2}\Big[a_u(x_{j-\frac{1}{2}}, u_{j-\frac{1}{2}}^{n-\frac{1}{2}}) D_x^- u_j^{n-1} - a_u(x_{j-\frac{1}{2}}, ru_{j-\frac{1}{2}}) D_x^- ru_j\Big] + \Big[a(x_{j-\frac{1}{2}}, u_{j-\frac{1}{2}}^{n-\frac{1}{2}}) - a(x_{j-\frac{1}{2}}, ru_{j-\frac{1}{2}})\Big], \tag{7.14f}$$

$$M_{2j-1,\,2j-1} = M_{2j-1,\,2j-3} \tag{7.14g}$$

for $j = 2, \cdots, J$. All other $M_{2j-1,\,i}$ are zero. Thus the matrix M is block tridiagonal with 2×2 blocks. The terms in the even numbered rows can be made arbitrarily small by taking h small enough. The odd numbered rows require in addition that $\underset{\sim}{y}^{\circ}$ and $\underset{\sim}{r}$ be smooth. If $\underset{\sim}{s}$ and $\underset{\sim}{t}$ are restricted to be near $\underset{\sim}{y}^{\circ}$, then $\underset{\sim}{r}$ will smooth, and if h is sufficiently small, we find that $\underset{\sim}{g}$ is contracting in a neighborhood of $\underset{\sim}{y}^{\circ}$. The fact that M is block tridiagonal guarantees that its maximum absolute row sum can be made small even as h goes to zero. Our previous investigation of the linear algebraic system shows that A^{-1} exists as h goes to zero so that the norm of A^{-1} (not necessarily the maximum absolute row sum) must be bounded in the limit. The size of the neighborhood in which contraction occurs for any pair of vectors $\underset{\sim}{s}$ and $\underset{\sim}{t}$ ie essentially proportional to $(a^{**})^{-1}$. However, we wish to avoid smoothness assumptions on $\underset{\sim}{s}$ and $\underset{\sim}{t}$ so we shall restrict them to a smaller spherical neighborhood of radius R about $\underset{\sim}{y}^{\circ}$ where R is some fixed multiple r_1 of h. We will specify r_1 shortly. Let $d \equiv \| \underset{\sim}{y}^1 - \underset{\sim}{y}^{\circ} \|$. We note that d is O(h) provided g_0 and g_1 are continuous. Thus for sufficiently small h we can choose r_1 so that $R > d$. We then take an even smaller h so that we also have $\| A^{-1} \| \cdot \| M \| < 1 - (d/R)$. Then

$$\| \underset{\sim}{g}(\underset{\sim}{s}) - \underset{\sim}{g}(\underset{\sim}{t}) \| \leq (1 - \frac{d}{R}) \cdot \| \underset{\sim}{s} - \underset{\sim}{t} \| \tag{7.15}$$

for $\underset{\sim}{s}$ and $\underset{\sim}{t}$ in the sphere. If $\underset{\sim}{s}$ is taken to be $\underset{\sim}{y}^{\circ}$, we find $\| \underset{\sim}{y}^1 - \underset{\sim}{g}(\underset{\sim}{t}) \|$ is less than or equal to R-d. Hence if $\underset{\sim}{t}$ is any vector within a distance R of $\underset{\sim}{y}^{\circ}$, $\underset{\sim}{g}(\underset{\sim}{t})$ will also be within R of $\underset{\sim}{y}^{\circ}$. In other words there is a neighborhood of $\underset{\sim}{y}^{\circ}$ which is mapped into itself by $\underset{\sim}{g}$ and in which $\underset{\sim}{g}$ contracts. Therefore $\underset{\sim}{g}$ has a unique fixed point in this neighborhood, and the nonlinear difference equations have a solution.

We further note that as with the linear parabolic equation we may ask the linearized initial boundary value problem to have very smooth solutions given sufficiently smooth initial data when posed at times greater than or equal to zero. As before, we can then show that the iterates are smooth by studying a series of linear problems - one for each iterate. Indeed this study is necessary for an inductive proof of the existence of a solution of the nonlinear difference equations for succeeding time steps, but it has the further implication that the matrix A might be re-evaluated at each iteration rather than at each time step as in the chord method. In fact the Jacobian of $\underset{\sim}{F}$ can be shown to be nonsingular in a neighborhood of $\underset{\sim}{y}^{o}$ under certain conditions of smallness on h and smoothness of the "point" of evaluation. The previous methods can again be used to show contraction of the series of maps $\underset{\sim}{g}^{\nu}$ and of the mapping into themselves of successive neighborhoods. Of course we may have to start with a smaller initial neighborhood or equivalently a finer net spacing to insure that successive Jacobians remain nonsingular, but otherwise the use of Newton's method is justified.

Returning now to the question of convergence, we find that nearly all of the analysis we exhibited for linear systems can also be adapted to the present nonlinear problem with the use of the mean value theorem. As before u is the finite difference solution and U is the continuous solution. The mean value theorem must be applied several times, but since it is not necessary to keep track of each application, we will use $\bar{u}$ as a generic symbol to indicate some function value intermediate to u and U. The three basic equations involving the truncation errors may then be written as

$$a(x_{j-\frac{1}{2}}, \bar{u}_{j-\frac{1}{2}}^{n-\frac{1}{2}}) D_x^- e_j^{n-\frac{1}{2}} = f_{j-\frac{1}{2}}^{n-\frac{1}{2}} + \rho_{j-\frac{1}{2}}^{n-\frac{1}{2}} - \left[a_u(x_{j-\frac{1}{2}}, \bar{u}_{j-\frac{1}{2}}^{n-\frac{1}{2}}) D_x^- \bar{u}_j^{n-\frac{1}{2}} \right] e_{j-\frac{1}{2}}^{n-\frac{1}{2}} , \qquad (7.16a)$$

$$D_x^- f_j^{n-\frac{1}{2}} = D_t^- e_{j-\frac{1}{2}}^n + \sigma_{j-\frac{1}{2}}^{n-\frac{1}{2}} \tag{7.16b}$$

$$- S_u(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}, \bar{u}_{j-\frac{1}{2}}^{n-\frac{1}{2}}, \bar{v}_{j-\frac{1}{2}}^{n-\frac{1}{2}}) e_{j-\frac{1}{2}}^{n-\frac{1}{2}}$$

$$- S_v(x_{j-\frac{1}{2}}, t_{n-\frac{1}{2}}, \bar{u}_{j-\frac{1}{2}}^{n-\frac{1}{2}}, \bar{v}_{j-\frac{1}{2}}^{n-\frac{1}{2}}) f_{j-\frac{1}{2}}^{n-\frac{1}{2}},$$

$$a(x_{j-\frac{1}{2}}, \bar{u}_{j-\frac{1}{2}}^{n-\frac{1}{2}}) D_t^- D_x^- e_j^n = D_t^- f_{j-\frac{1}{2}}^n + \zeta_{j-\frac{1}{2}}^{n-\frac{1}{2}} \tag{7.16c}$$

$$- \left[D_t^- (a_u(x_{j-\frac{1}{2}}, \bar{u}_{j-\frac{1}{2}}^{n-\frac{1}{2}}) D_x^- \bar{u}_j^n)\right] e_{j-\frac{1}{2}}^{n-\frac{1}{2}}$$

$$- \left[a_u(x_{j-\frac{1}{2}}, \bar{u}_{j-\frac{1}{2}}^{n-\frac{1}{2}}) D_x^- \bar{u}_j^{n-\frac{1}{2}}\right] D_t^- e_{j-\frac{1}{2}}^n$$

$$- \left[D_t^- a(x_{j-\frac{1}{2}}, \bar{u}_{j-\frac{1}{2}}^n)\right] D_x^- e_j^{n-\frac{1}{2}},$$

where

$$e_j^n \equiv U(x_j, t_n) - u_j^n, \tag{7.17a}$$

$$f_j^n \equiv V(x_j, t_n) - v_j^n. \tag{7.17b}$$

We multiply (7.16c) by $f_{j-\frac{1}{2}}^{n-\frac{1}{2}} h_j$, multiply (7.16b) by $a(x_{j-\frac{1}{2}}, \bar{u}_{j-\frac{1}{2}}^{n-\frac{1}{2}}) \cdot e_{j-\frac{1}{2}}^{n-\frac{1}{2}} h_j$, add the results, and sum from $j = 2$ to J:

$$
\begin{aligned}
&(a(\bar{u}^{n-\frac{1}{2}})e^{n-\frac{1}{2}}, D_t^- e^n)_h + (f^{n-\frac{1}{2}}, D_t^- f^n)_h = \\
&\quad (f^{n-\frac{1}{2}}, a(\bar{u}^{n-\frac{1}{2}})D_t^- D_x^- e^n)_h \qquad (7.18)\\
&\quad + (f^{n-\frac{1}{2}}, (D_t^-(a_u(\bar{u}^n)D_x^- \bar{u}^n)\, e^{n-\frac{1}{2}})_h \\
&\quad + (f^{n-\frac{1}{2}}, (a_u(\bar{u}^{n-\frac{1}{2}})D_x^- \bar{u}^{n-\frac{1}{2}})D_t^- e^n)_h \\
&\quad + (f^{n-\frac{1}{2}}, (D_t^- a(\bar{u}^n))D_x^- e^{n-\frac{1}{2}})_h - (f^{n-\frac{1}{2}}, \zeta^{n-\frac{1}{2}})_h \\
&\quad + (a(\bar{u}^{n-\frac{1}{2}})e^{n-\frac{1}{2}}, D_x^- f^{n-\frac{1}{2}})_h \\
&\quad + (a(\bar{u}^{n-\frac{1}{2}})e^{n-\frac{1}{2}}, S_u(\bar{u}^{n-\frac{1}{2}}, \bar{v}^{n-\frac{1}{2}})e^{n-\frac{1}{2}})_h \\
&\quad + (a(\bar{u}^{n-\frac{1}{2}})e^{n-\frac{1}{2}}, S_v(\bar{u}^{n-\frac{1}{2}}, \bar{v}^{n-\frac{1}{2}})f^{n-\frac{1}{2}})_h \\
&\quad - (a(\bar{u}^{n-\frac{1}{2}})e^{n-\frac{1}{2}}, \sigma^{n-\frac{1}{2}})_h .
\end{aligned}
$$

Equation (7.18) has the same form as (5.6) except for the variable coefficients depending on the solution. This means additional applications of the mean value theorem will be necessary in order to carry out the analysis of Section 5. As before, the requirement of smoothness on the finite difference solution arises and is handled as in the linear case by an examination of the system of linear algebraic equations obtained from the Jacobian of the nonlinear system and by requiring U and V and solutions of the linearized initial boundary value problem to have x derivatives which are small compared to the inverse of the net spacing. The results of this section are summarized in the following theorem.

Theorem 2: Assume (1) that the nonlinear problem (7.1) has a unique solution and (2) that the linearization of problem (7.1) at any piecewise five times continuously differentiable U and V results in a differential equation with coefficients sufficiently smooth so that an initial boundary value problem posed at any non-negative time with initial functions both piecewise five times continuously differentiable will have solutions which are also piecewise five times continuously differentiable. If the points of discontinuity of the derivatives are always taken to be mesh points, then for sufficiently small h the box scheme formulation (7.2) of the nonlinear parabolic problem (7.1) has a unique solution which approximates the solution of (7.1) to $O(h^2)$.

CHAPTER II

THE TWO DIMENSIONAL HEAT EQUATION

II.1 The Method of Fractional Steps

The numerical solution of multidimensional parabolic problems is of concern to scientists and engineers because there are many physical processes in which boundary conditions or properties of materials prevent realistic modeling by one-dimensional equations. In particular we are interested in two space dimensions. We could of course write various sets of difference equations coupling net points in both space directions, but this means we would have to solve a large algebraic system for all the net points in the domain at each time step. Instead we restrict ourselves to a rectangular domain and ask if the two-dimensional problem can be reduced to a series of one-dimensional problems. Just such a reduction is accomplished by the method of fractional steps. This method has not yet received a complete theoretical justification for all the problems in which we would like to use it and is still undergoing active investigation. The reader is referred to Yanenko [1971] for an exposition of the techniques and applications of the method of fractional steps. For our present purposes we shall present a simple example using this method. In this case the method will be theoretically justified.

Consider the equation

$$U_t = U_{xx} + U_{yy} \tag{1.1}$$

on the domain $D \equiv [0,1] \times [0,1]$ in the (x,y) plane. We require U to be zero on the boundary of D for all time $t \geq 0$. At time $t = 0$, $U(x,y,t)$ is equal to some given function $g(x,y)$. As is well-known, this problem can be solved by separation of variables and Fourier series. The solution is then represented as a sum of terms of the form

$$e^{-(m^2+n^2)\pi^2 t} \sin m\pi x \sin n\pi y \ , \tag{1.2}$$

where m and n are positive integers. Now pick two different points in time t_1 and t_2 such that $t_2 > t_1 \geq 0$. t_1 and t_2 need not be close together. We pose two more problems. First

$$V_t = V_{xx} \quad \text{on} \quad D \times [t_1, t_2] \ , \tag{1.3a}$$

$$V(x,y,t) = 0 \quad \text{on} \quad \partial D \ , \tag{1.3b}$$

$$V(x,y,t_1) = U(x,y,t_1) \ . \tag{1.3c}$$

This is a heat equation in one space dimension with y as a parameter. V is chosen to coincide with U at time t_1. Now consider the second problem:

$$W_t = W_{yy} \quad \text{on} \quad D \times [t_1, t_2] , \tag{1.4a}$$

$$W(x,y,t) = 0 \quad \text{on} \quad \partial D , \tag{1.4b}$$

$$W(x,y,t_1) = V(x,y,t_2) . \tag{1.4c}$$

This is another one-dimensional heat equation but with x as a parameter. Notice that the initial value of W is the value of V at t_2. We now assert that

$$W(x,y,t_2) = U(x,y,t_2) . \tag{1.5}$$

This is an example of how the method of fractional steps reduces a two-dimensional problem to two one-dimensional problems. We can easily verify (1.5) for this simple case when separation of variables is legitimate. A component of the form (1.2) when used as an initial condition for the V problem evolves to

$$e^{-m^2\pi^2 t_2 - n^2\pi^2 t_1} \sin m\pi x \sin n\pi y , \tag{1.6}$$

where the factor $\exp(-n^2\pi^2 t_1)$ acts as though it were a multiplicative constant. When (1.6) is used as an initial condition for the W problem, the other factor, $\exp(-m^2\pi^2 t_2)$, acts as a multiplicative constant so that at time t_2 we get

$$e^{-m^2\pi^2 t_2 - n^2\pi^2 t_2} \sin m\pi x \sin n\pi y\ . \tag{1.7}$$

Thus we see that Fourier components of the two-dimensional problem evolve to the same extent as Fourier components of two successive one-dimensional problems.

II.2 The Box Scheme and the Method of Fractional Steps

Having reduced the two-dimensional heat equation to two one-dimensional heat equations with parameters, we easily see how to apply the box scheme. We place a rectangular grid over the unit square D. For each horizontal grid line we solve an equation in x, then we change directions to solve an equation in y for each vertical grid line. The procedure is then repeated for the next time step.

The box scheme requires however that we give the derivatives of the initial data as well as the data themselves. For a sweep in the x direction we deal with u and u_x, but when we wish to perform a sweep in the y direction, we must give u_y as part of the initial condition. What we must do is ignore u_x after an x sweep and construct u_y using the computed values of u. We have chosen Lagrange interpolation as a way to do this. At any net point take the value of u and combine it with the values of u at the next two nearest net points on the grid line to form the Lagrange quadratic polynomial which interpolates all three function values. We then use the derivative of the quadratic. Furthermore we do not change directions at every half step since it is not really necessary. Suppose for instance that in solving for u at t_2 we start at t_1 and perform an x sweep followed by a y sweep. It will be necessary to fabricate y derivatives when changing directions; however, in moving from t_2 to t_3, we perform the y sweep first followed by the x sweep. Proceeding in this way, we need create derivatives only once per time step rather than twice.

We have performed computations on the two-dimensional heat heat equation with the following initial condition:

$$g(x,y) = \sum_{m,n=1}^{3} \frac{5}{2m+3n} \sin m\pi x \sin n\pi y \ . \qquad (2.1)$$

The continuous solution is

$$U(x,y,t) =$$

$$\sum_{m,n=1}^{3} \frac{5}{2m+3n} e^{-(m^2+n^2)\pi^2 t} \sin m\pi x \sin n\pi y \ . \qquad (2.2)$$

Using several different mesh spacings we have found that the error of the computed solution u is $O(h^2)$ where h was the size of the steps in both space directions and in time.

In the present problem there is no doubt about the consistency of the numerical scheme with the continuous problem since the fractional steps are theoretically correct and since the box scheme is consistent with one-dimensional heat equations to second order; that is, the truncation errors are $O(h^2)$. Convergence however is yet to be shown. For this problem we choose to show that the numerical scheme is stable in the sense of von Neumann. This stability analysis is appropriate for difference schemes with constant coefficients and which correspond to pure initial value problems with periodic initial data [Isaacson & Keller, 1966].

II.3 Von Neumann Stability

The basic idea of a von Neumann stability analysis is to decompose the finite difference solution into Fourier components and then to show that none of the components can grow in amplitude as time increases. In this section we will use complex harmonics and coefficients rather than sines and cosines with real coefficients. We shall study the evolution of the general harmonic $e^{i\alpha x}e^{i\beta y}$ where $i = \sqrt{-1}$ and α and β are fixed, arbitrary real numbers. We assume that all steps in the x direction are of size hx, all steps in the y direction are of size hy, and all time steps are k. Let u^n and v^n be the net functions at time t_n which approximate U and U_x. Let $\tilde{u}^{n+1}$ and $\tilde{w}^{n+1}$ be the net functions after a half time step which approximate V and V_y. Finally let u^{n+1} and w^{n+1} be the net functios at time t_{n+1} which approximate U and U_y. These net functions must each be some multiple of $e^{i\alpha x}e^{i\beta y}$ evaluated at the net points:

$$u^n = c_1^n e^{i\alpha x}e^{i\beta y}, \tag{3.1a}$$

$$v^n = c_2^n e^{i\alpha x}e^{i\beta y}, \tag{3.1b}$$

$$\tilde{u}^{n+1} = \tilde{c}_1^{n+1} e^{i\alpha x}e^{i\beta y}, \tag{3.1c}$$

$$\tilde{w}^{n+1} = \tilde{c}_3^{n+1} e^{i\alpha x}e^{i\beta y}, \tag{3.1d}$$

$$u^{n+1} = c_1^{n+1} e^{i\alpha x}e^{i\beta y}, \tag{3.1e}$$

$$w^{n+1} = c_3^{n+1} e^{i\alpha x}e^{i\beta y}. \tag{3.1f}$$

The above net functions are involved in going from time t_n to t_{n+1} with an x sweep followed by a y sweep. We in turn proceed from time t_{n+1} to t_{n+2} with a y sweep followed by an x sweep. That is, one complete cycle covers two time steps. Now suppose c_1^n and c_2^n are given. The difference scheme will determine c_1^{n+2} and c_2^{n+2}. What we must show is that c_1^{n+2} does not exceed c_1^n in amplitude and that c_2^{n+2} does not exceed c_2^n in amplitude. This would then imply that no harmonic can grow in amplitude; hence, the numerical scheme is stable.

We introduce two symbols which we shall use in simplifying notation:

$$\theta \equiv \frac{1}{2}\,\alpha hx \ , \tag{3.2a}$$

$$\varphi \equiv \frac{1}{2}\,\beta hy \ . \tag{3.2b}$$

Since y derivatives are constructed from the most recent values of u and do not involve any information about U_x or its approximation, it is clear that $\tilde{c}_3^{n+1}$ can depend only on $\tilde{c}_1^{n+1}$. In fact

$$\tilde{c}_3^{n+1} = \frac{\tilde{c}_1^{n+1}}{hy}\, i \sin \beta hy \ . \tag{3.3}$$

(3.3) and the substitution of (3.1a) through (3.1d) into the finite difference equations for an x sweep yield the following relations between the coefficients:

$$\begin{pmatrix} \tilde{c}_1^{n+1} \\ \tilde{c}_3^{n+1} \end{pmatrix} = \frac{1}{a} A \begin{pmatrix} c_1^n \\ c_2^n \end{pmatrix} , \tag{3.4a}$$

where

$$a = \frac{2}{hx^2} \sin^2 \theta + \frac{1}{k} \cos^2 \theta , \tag{3.4b}$$

$$\mathbf{A} \equiv \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix} , \tag{3.4c}$$

$$A_1 = \frac{1}{k} \cos^2 \theta , \tag{3.4d}$$

$$A_2 = \frac{i}{hx} \sin \theta \cos \theta , \tag{3.4e}$$

$$A_3 = \frac{i}{hy\,k} \sin \beta hy \cos^2 \theta , \tag{3.4f}$$

$$A_4 = - \frac{1}{hy\,hx} \sin \beta hy \sin \theta \cos \theta . \tag{3.4g}$$

Similarly the substitution of (3.1c) through (3.1f) into the difference equations for a y sweep yields

$$\begin{pmatrix} c_1^{n+1} \\ c_3^{n+1} \end{pmatrix} = \frac{1}{b} B \begin{pmatrix} \tilde{c}_1^{n+1} \\ \tilde{c}_3^{n+1} \end{pmatrix} , \tag{3.5a}$$

where

$$b = \frac{2}{hy^2} \sin^2 \varphi + \frac{1}{k} \cos^2 \varphi , \tag{3.5b}$$

$$B \equiv \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix}, \tag{3.5c}$$

$$B_1 = \frac{1}{k} \cos^2 \varphi , \tag{3.5d}$$

$$B_2 = \frac{i}{hy} \sin \varphi \cos \varphi , \tag{3.5e}$$

$$B_3 = \frac{2i}{hy\ k} \sin \varphi \cos \varphi , \tag{3.5f}$$

$$B_4 = -\frac{2}{hy^2} \sin^2 \varphi . \tag{3.5g}$$

These two transformations must be composed and followed with two more similar transformations representing the progression from time t_{n+1} to t_{n+2}. We shall not give the details, but we write the composition of the four transformations symbolically as

$$\begin{pmatrix} c_1^{n+2} \\ c_2^{n+2} \end{pmatrix} = G \begin{pmatrix} c_1^{n} \\ c_2^{n} \end{pmatrix}, \tag{3.6}$$

where G is a 2×2 matrix. One of the eigenvalues of G is zero. The other is

$$\left\{ \frac{\left(\frac{1}{k} \cos^2\theta - \frac{2}{hx^2} \sin^2\theta\right)}{\left(\frac{2}{hx^2} \sin^2\theta + \frac{1}{k_n} \cos^2\theta\right)} \right\} \cdot \left\{ \frac{\left(\frac{1}{k} \cos^2\varphi - \frac{2}{hy^2} \sin^2\varphi\right)}{\left(\frac{2}{hy^2} \sin^2\varphi + \frac{1}{k} \cos^2\varphi\right)} \right\} .$$

$$\left\{\frac{\left(\frac{1}{k}\cos^2\theta - \frac{1}{hx^2}\sin\theta\cos\theta\sin\alpha hx\right)}{\left(\frac{2}{hx^2}\sin^2\theta + \frac{1}{k}\cos^2\theta\right)}\right\} \cdot$$

$$\left\{\frac{\left(\frac{1}{k}\cos^2\varphi - \frac{1}{hy^2}\sin\varphi\cos\varphi\sin\beta hy\right)}{\left(\frac{2}{hy^2}\sin^2\varphi + \frac{1}{k}\cos^2\varphi\right)}\right\} \quad . \qquad (3.7)$$

It will be found upon study that each of the four terms in braces must lie between -1 and $+1$; hence, the eigenvalues of G are real and are less than or equal to one in magnitude. Since α and β were arbitrary, we have shown that no harmonic can increase in amplitude, and the numerical scheme is stable.

CHAPTER III

BURGERS' EQUATION: COMPUTATIONAL EXAMPLES

III.1 Burgers' Equation

In this chapter we shall describe the results of computing on Burgers' equation with three different initial conditions. Burgers' equation is

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} = \nu \frac{\partial^2 U}{\partial x^2} , \qquad (1.1)$$

where $U = U(x,t)$ and ν is a positive number. This equation is discussed by Cole [1951] who describes its applications and its general solution. His result is that if $\theta(x,t)$ is any solution to the heat equation

$$\frac{\partial \theta}{\partial t} = \nu \frac{\partial^2 \theta}{\partial x^2} , \qquad (1.2a)$$

then

$$U(x,t) = -2\nu \frac{\theta_x}{\theta} \qquad (1.2b)$$

is a solution of (1.1). We shall use this result in our third computational example.

We are particularly interested in the situation when ν is small compared to unity. In fact let us first examine the case when $\nu = 0$:

$$\frac{\partial U}{\partial t} + U \frac{\partial U}{\partial x} = 0 \quad . \tag{1.3}$$

The characteristic ordinary differential equations are

$$\frac{dx}{d\xi} = U \, , \tag{1.4a}$$

$$\frac{dt}{d\xi} = 1 \, , \tag{1.4b}$$

$$\frac{dU}{d\xi} = 0 \, . \tag{1.4c}$$

If we take the initial condition $x = \eta$ at time $t = 0$ and if $U(x,0) = U_0(x)$, we find

$$x = \eta + \xi \, U_0(\eta) \, , \tag{1.5a}$$

$$t = \xi \, , \tag{1.5b}$$

$$U = U_0(\eta) \, . \tag{1.5c}$$

ξ is a parameter along the characteristic curve in the (x,t) plane on which U is a constant. An interesting feature of (1.5) is that characteristics originating at two different points η_1 and η_2 on the x axis at time $t = 0$ can intersect at $(\bar{x}, \bar{t})$ where $\bar{x}$ and $\bar{t}$ are given by

$$\bar{x} = \frac{\eta_1 U_0(\eta_2) - \eta_2 U_0(\eta_1)}{U_0(\eta_2) - U_0(\eta_1)} , \tag{1.6a}$$

$$\bar{t} = - \frac{(\eta_2 - \eta_1)}{U_0(\eta_2) - U_0(\eta_1)} . \tag{1.6b}$$

Since we are interested only in positive time, we restrict our attention to initial functions $U_0(\eta)$ such that intersection times $\bar{t}$ are positive. This means U_0 should be a decreasing function of η in some range of η. The first instant in time at which an intersection occurs is determined by the greatest negative slope of U_0. We may see this by allowing η_1 and η_2 to approach each other in (1.6b). At this point we say a shock is forming in the solution; that is, the solution will become discontinuous. The initial function U_0 determines which characteristics will lead into the shock and what the magnitude of the discontinuity will be. We note that the value of U must always lie in the range between the minimum and maximum values of U_0 so that if $U_0(\eta)$ has a bounded range, the jump in the solution must also be bounded. The speed with which the shock propagates is determined by the values of U just ahead of and just behind the shock. We rewrite (1.3) as

$$\frac{\partial U}{\partial t} + \frac{1}{2}\frac{\partial}{\partial x}[U^2] = 0 . \qquad (1.7)$$

If we seek a steadily propagating solution of the form $U = U(\zeta)$ where $\zeta = x - vt$, (1.7) becomes

$$-v\frac{\partial}{\partial \zeta}U + \frac{1}{2}\frac{\partial}{\partial \zeta}U^2 = 0 \qquad (1.8a)$$

which we integrate from ζ_1 to ζ_2:

$$-v[U(\zeta_2) - U(\zeta_1)] + \frac{1}{2}[U^2(\zeta_2) - U^2(\zeta_1)] = 0. \qquad (1.8b)$$

(1.8b) may be solved for v:

$$v = \frac{1}{2}[U(\zeta_2) + U(\zeta_1)] . \qquad (1.9)$$

If ζ_1 and ζ_2 are chosen on opposite sides of the shock, then (1.9) tells us that the velocity of a steadily moving shock is equal to the average of the function values just ahead of and just behind the shock.

If $\nu > 0$, we no longer have shocks in the sense of intersecting characteristics. Nevertheless, if ν is small, we should be able to see shocks trying to form. The small amount of diffusion will prevent the complete formation of a shock. On the other hand if U_0 is a step function, the diffusion will smooth the step into a continuous function. We shall demonstrate both of these situations in the computational examples.

III. 2 Example 1: Smoothing of a Sharp Front

The domain for all three examples will be $0 \leq x \leq 1$. In this example the initial function U_0 is 1 for $0 \leq x \leq 0.48$ and 0 for $0.52 \leq x \leq 1$. In the interval $0.48 \leq x \leq 0.52$ U_o is the unique cubic polynomial with value 1 at the left end, 0 at the right end, and zero derivatives at both ends. The boundary conditions are $U(0,t) = 1$ and $U(1,t) = 0$. We take $\nu = 3 \times 10^{-3}$. With the exception of the initial condition, this example is the same as that given by Swartz & Wendroff [1969]. Their initial condition was a step function. We use a cubic transition function instead because we need to specify x derivatives as initial data for the box scheme.

We have performed the computation using both uniform and nonuniform net spacings in the x direction. The time spacing was always taken to be uniform. Newton's method was used for solving the nonlinear difference equations in all of the examples. All computations were performed in double precision on an IBM 370/155. The first four figures[1] show the solution for uniform meshes. The curves were plotted at intervals of 0.1 time units so that the last curve corresponds to time $t = 0.5$. The oscillations which appear most prominently in Figure 1 are actually a part of the numerical solution since according to our boundary conditions we cannot have any oscillations of constant amplitude over the entire net which must be averaged out.

---

[1] Tables and figures are at the ends of the sections in which they are discussed with tables (if any) preceding the figures.

In Figures 5 through 8 we show a nonuniform mesh which is successively refined. These curves are also given at time intervals of 0.1.

In Figure 8 we see most clearly the behavior of the solution. The "corners" of the initial function are quickly rounded off, and a shock-like profile is moving to the right with speed one half. For an economical computation, one should probably change the space net as the shock propagates by deleting net points behind the shock and interpolating additional points in the neighborhood of the shock.

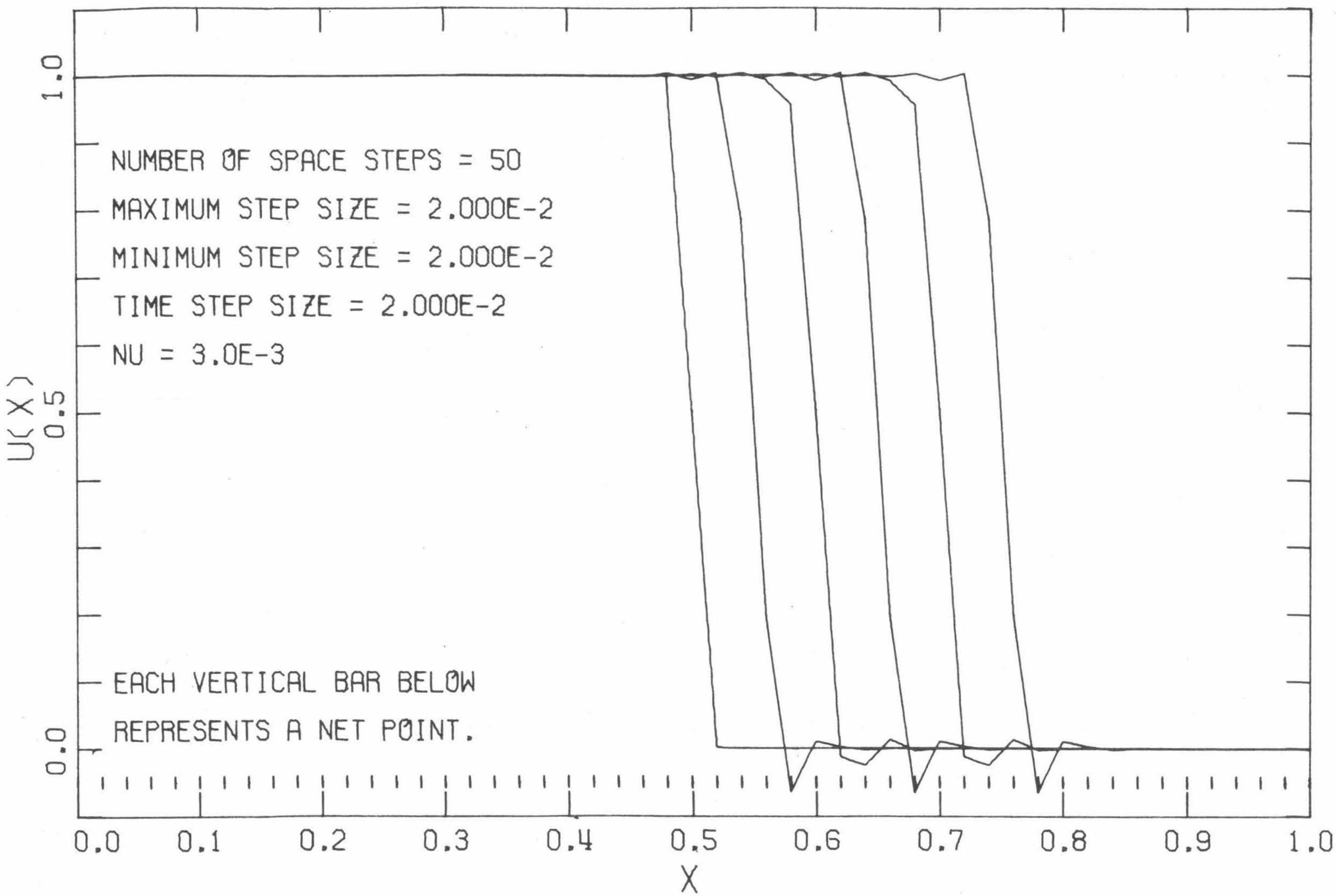

NUMBER ØF SPACE STEPS = 50
MAXIMUM STEP SIZE = 2.000E-2
MINIMUM STEP SIZE = 2.000E-2
TIME STEP SIZE = 2.000E-2
NU = 3.0E-3
EACH VERTICAL BAR BELOW
REPRESENTS A NET PØINT.
U(X)
1.0
0.5
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
X

Figure 1

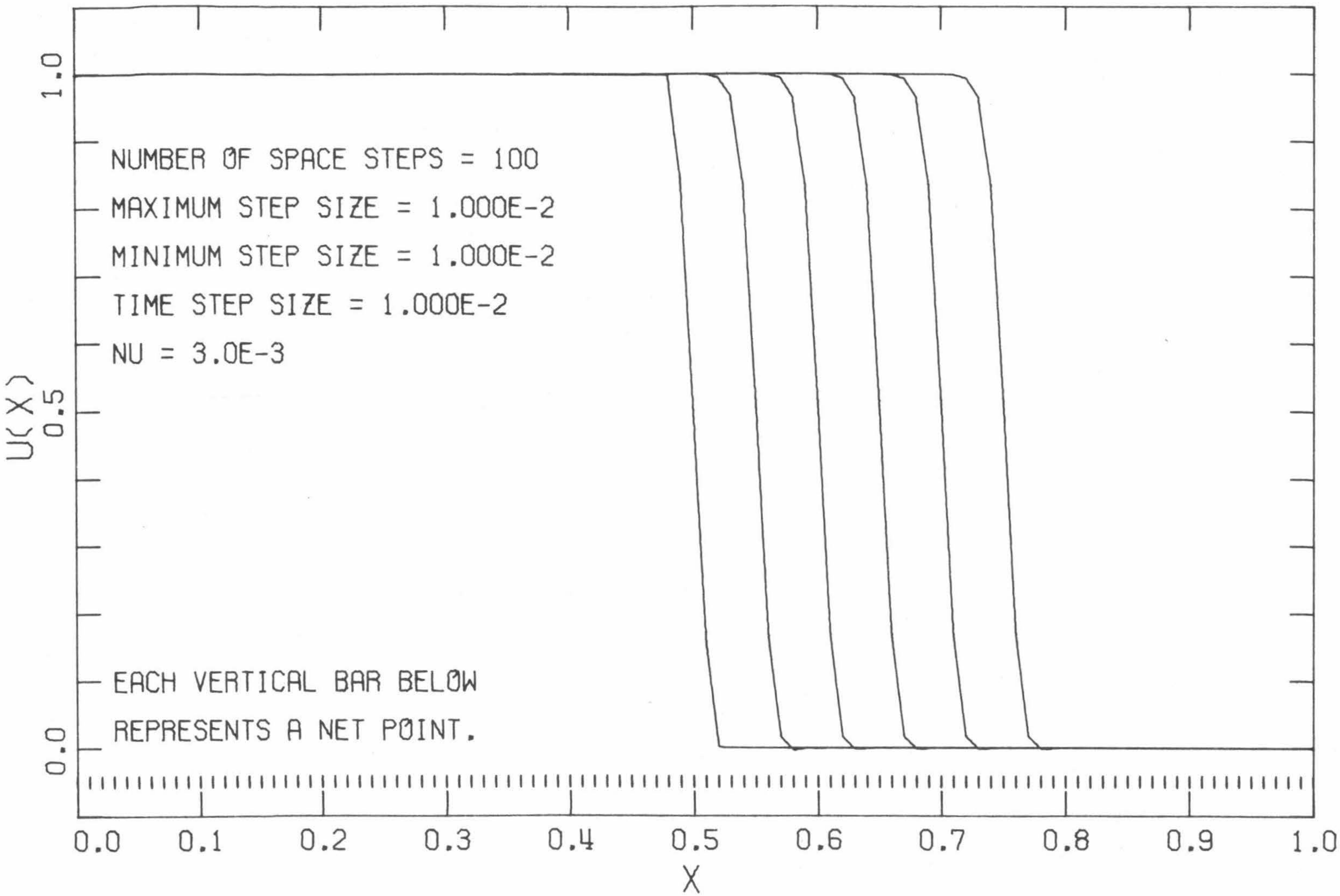

NUMBER OF SPACE STEPS = 100
MAXIMUM STEP SIZE = 1.000E-2
MINIMUM STEP SIZE = 1.000E-2
TIME STEP SIZE = 1.000E-2
NU = 3.0E-3
EACH VERTICAL BAR BELOW
REPRESENTS A NET POINT.
U(X)
1.0
0.5
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
X

Figure 2

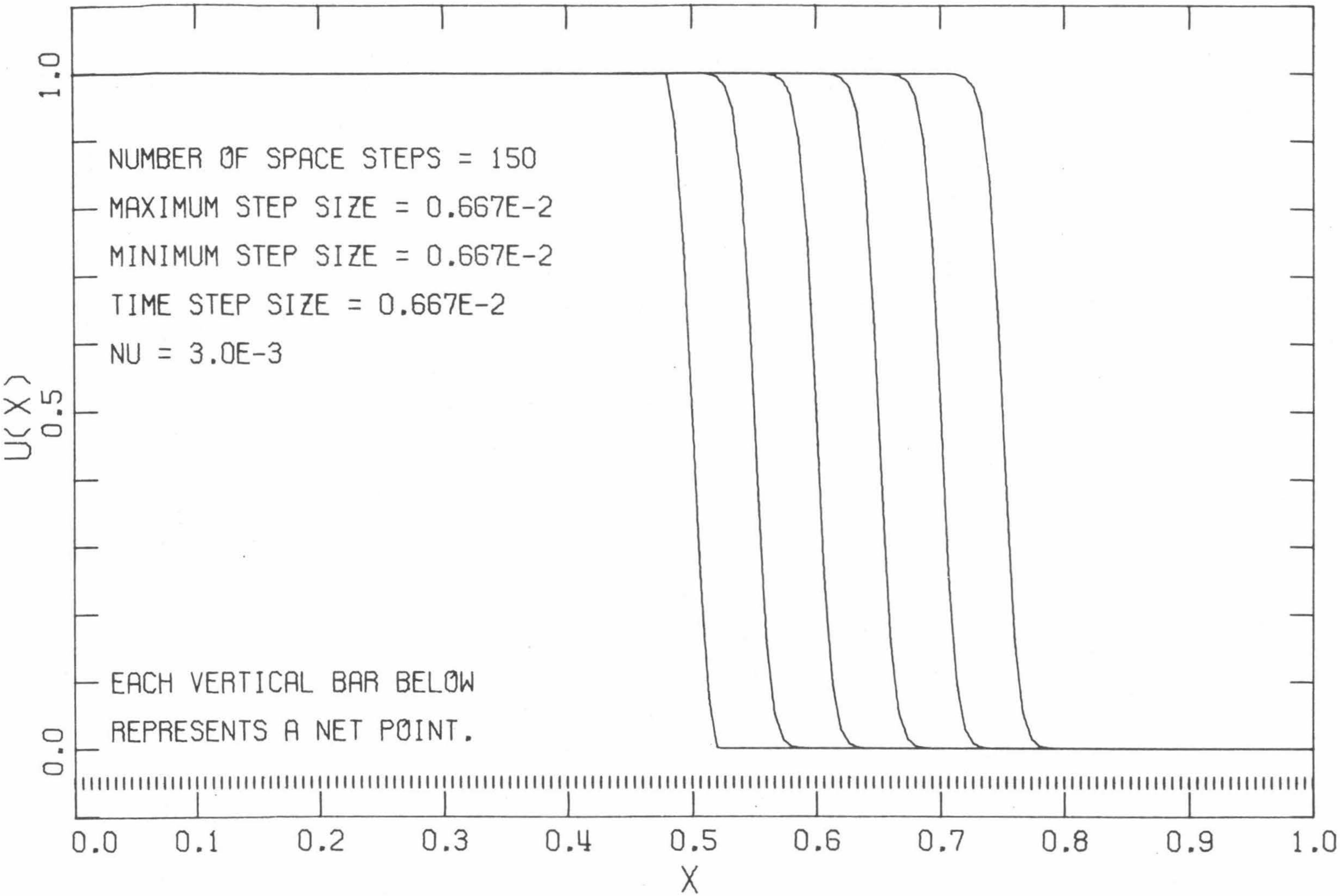

1.0
0.5
0.0
U(X)
NUMBER ØF SPACE STEPS = 150
MAXIMUM STEP SIZE = 0.667E-2
MINIMUM STEP SIZE = 0.667E-2
TIME STEP SIZE = 0.667E-2
NU = 3.0E-3
EACH VERTICAL BAR BELØW
REPRESENTS A NET PØINT.
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
X

Figure 3

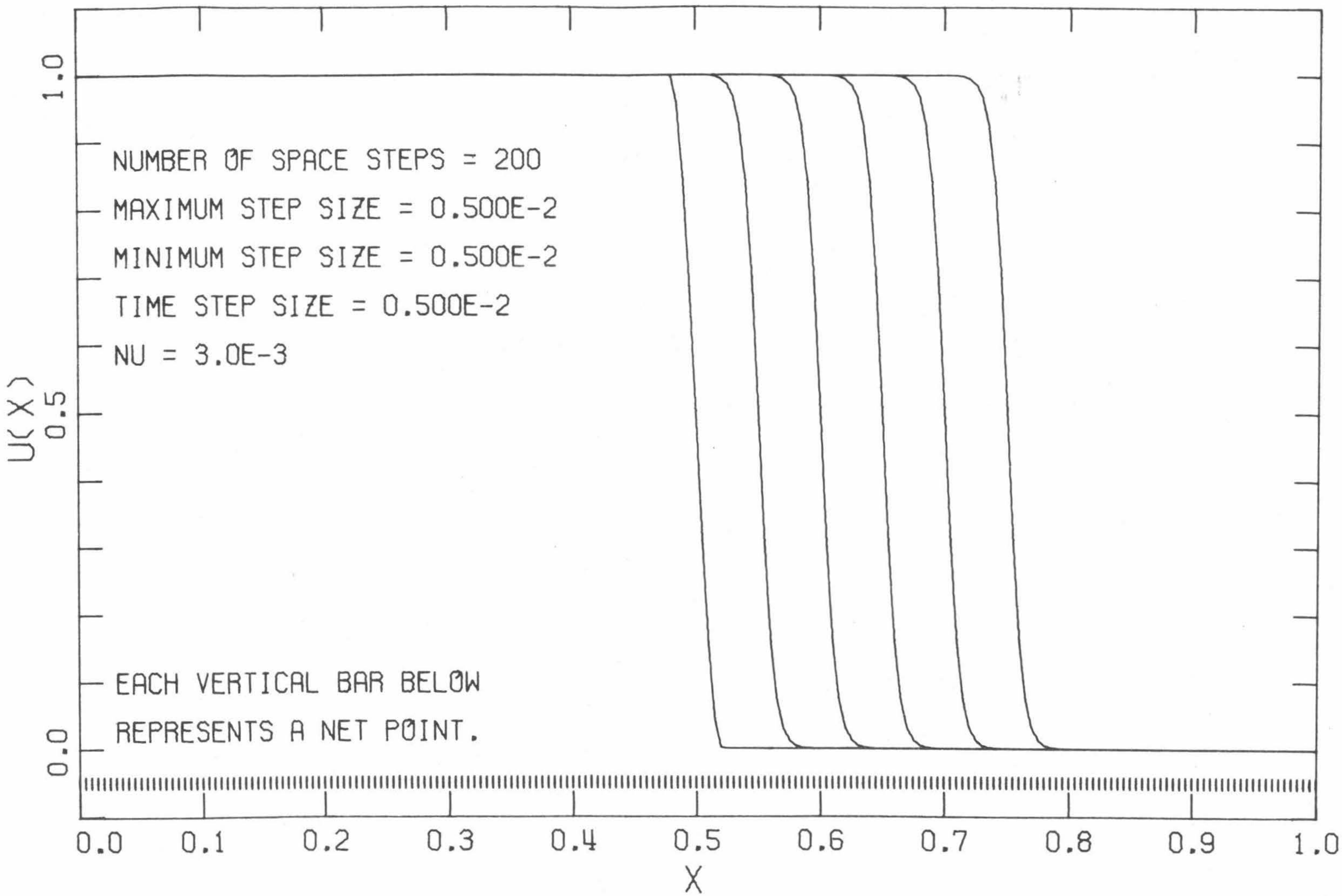

NUMBER ØF SPACE STEPS = 200
MAXIMUM STEP SIZE = 0.500E-2
MINIMUM STEP SIZE = 0.500E-2
TIME STEP SIZE = 0.500E-2
NU = 3.0E-3
EACH VERTICAL BAR BELØW
REPRESENTS A NET PØINT.
U(X)
1.0
0.5
0.0
X
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 4

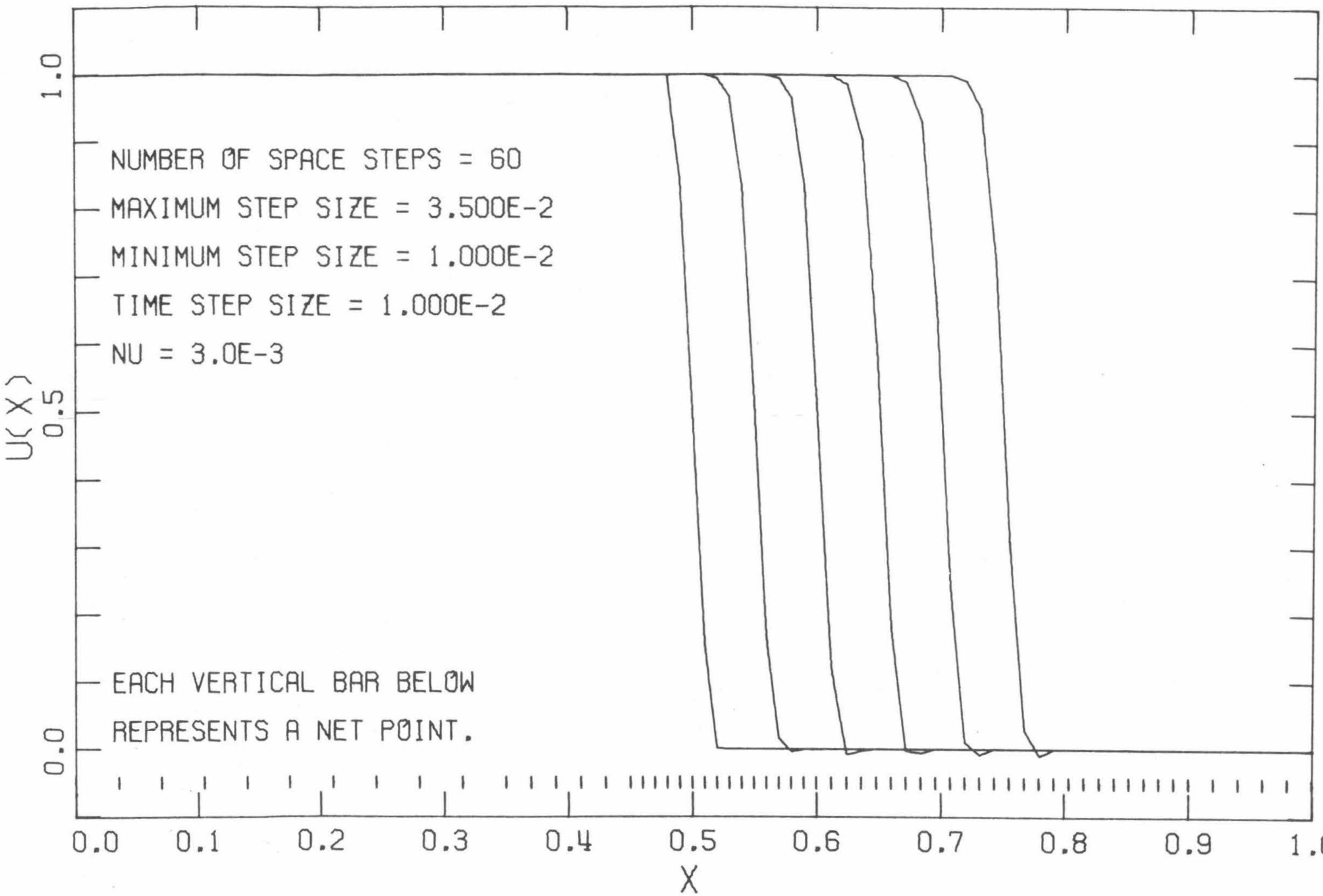

NUMBER ØF SPACE STEPS = 60
MAXIMUM STEP SIZE = 3.500E-2
MINIMUM STEP SIZE = 1.000E-2
TIME STEP SIZE = 1.000E-2
NU = 3.0E-3
EACH VERTICAL BAR BELØW
REPRESENTS A NET PØINT.
U(X)
1.0
0.5
0.0
X
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 5

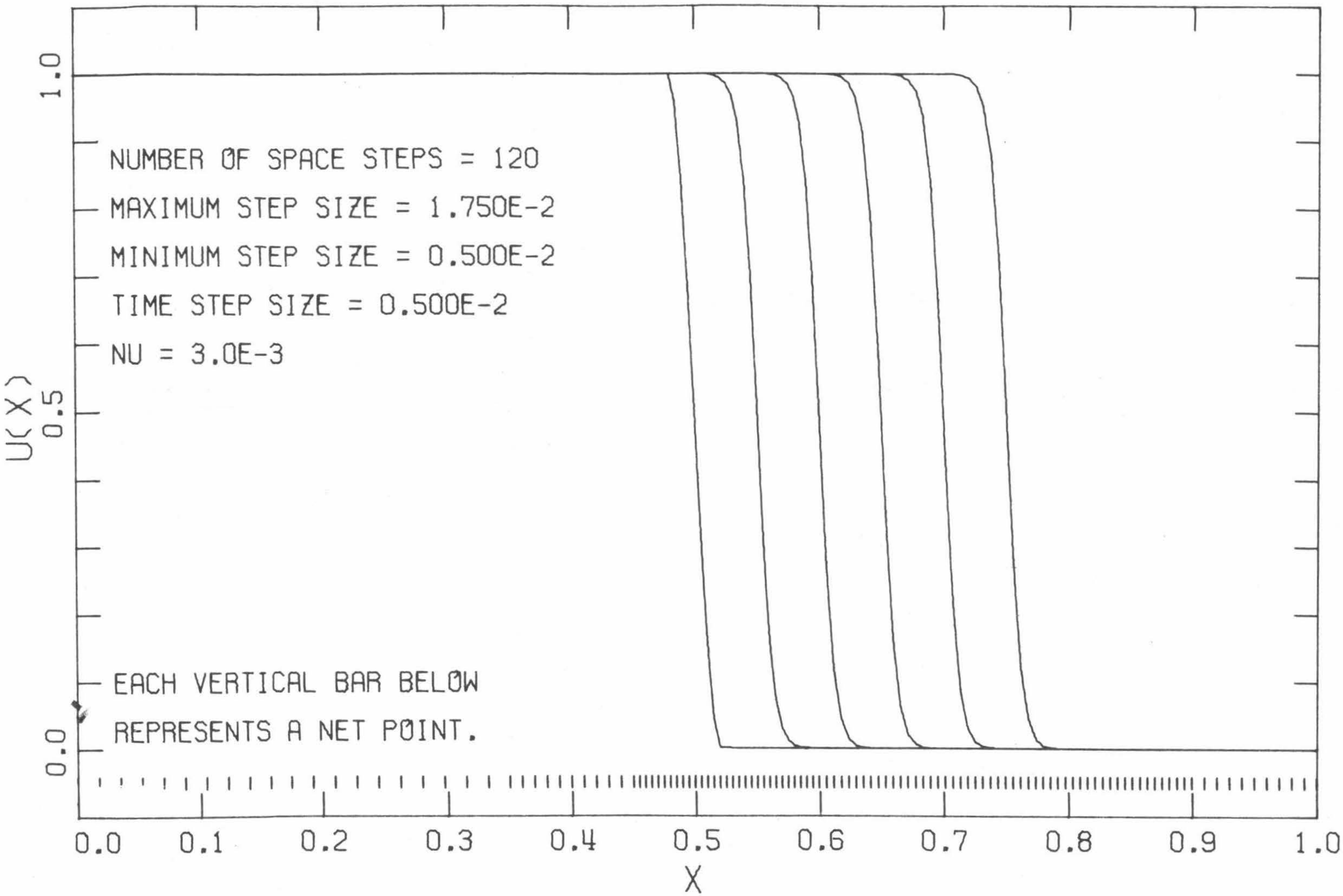

1.0
0.5
0.0
U(X)
NUMBER ØF SPACE STEPS = 120
MAXIMUM STEP SIZE = 1.750E-2
MINIMUM STEP SIZE = 0.500E-2
TIME STEP SIZE = 0.500E-2
NU = 3.0E-3
EACH VERTICAL BAR BELØW
REPRESENTS A NET PØINT.
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
X

Figure 6

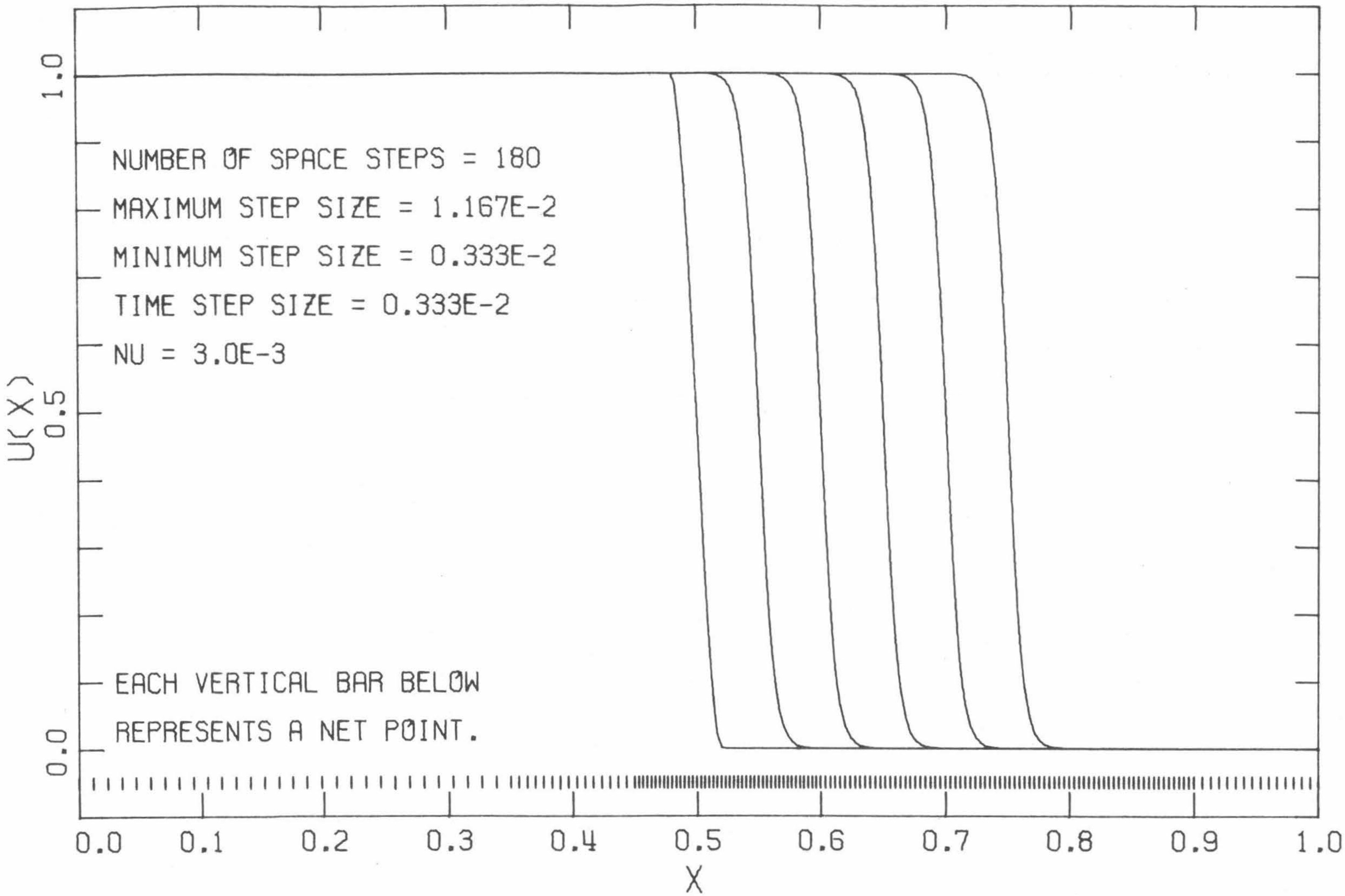
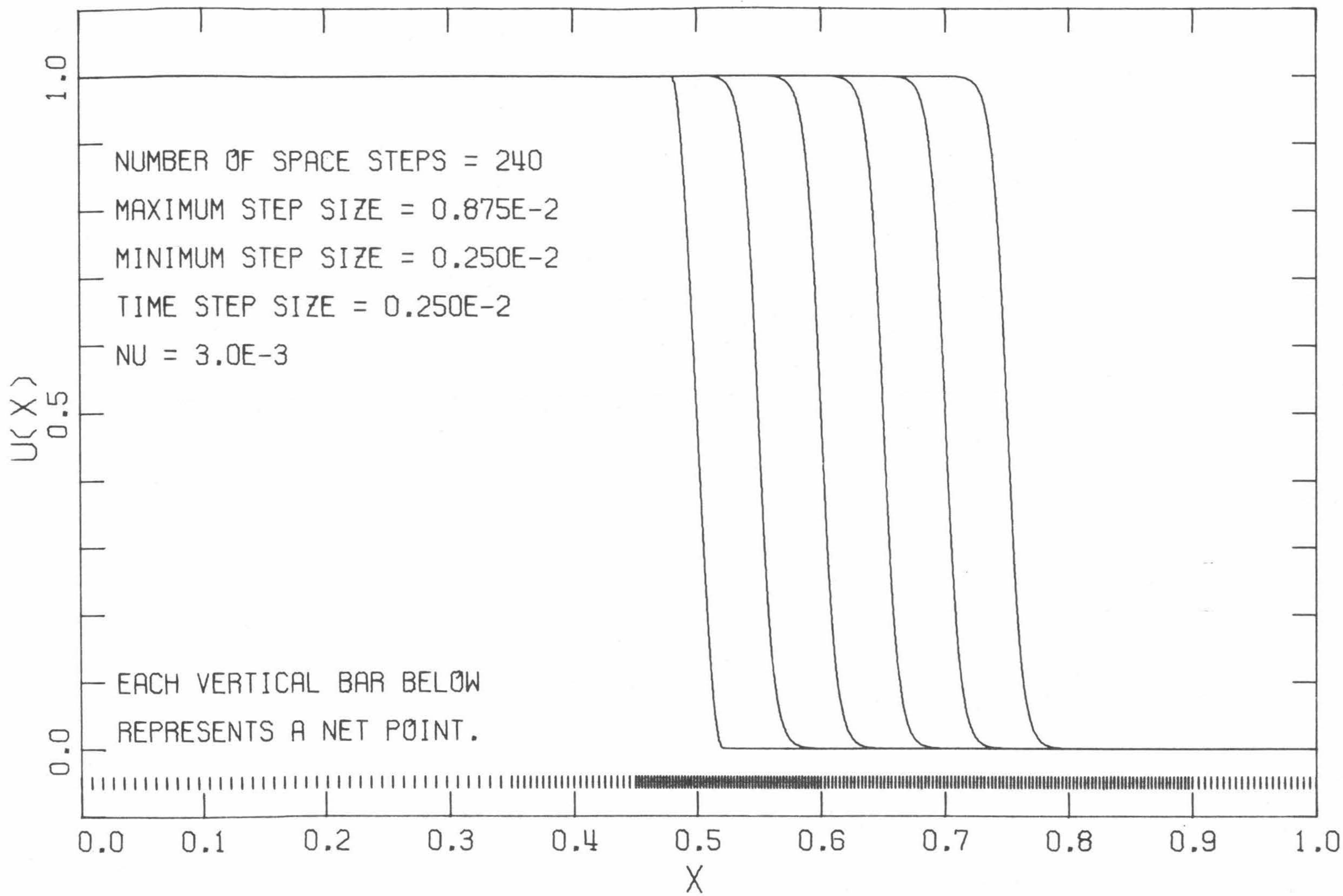

NUMBER ØF SPACE STEPS = 180
MAXIMUM STEP SIZE = 1.167E-2
MINIMUM STEP SIZE = 0.333E-2
TIME STEP SIZE = 0.333E-2
NU = 3.0E-3
EACH VERTICAL BAR BELØW
REPRESENTS A NET PØINT.
U(X)
1.0
0.5
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
X

Figure 7



1.0
0.5
0.0
U(X)
NUMBER ØF SPACE STEPS = 240
MAXIMUM STEP SIZE = 0.875E-2
MINIMUM STEP SIZE = 0.250E-2
TIME STEP SIZE = 0.250E-2
NU = 3.0E-3
EACH VERTICAL BAR BELØW
REPRESENTS A NET PØINT.
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
X

Figure 8

III.3 Example 2: Formation of a Shock

In this example the initial function U_0 is

$$U_0(x) = \begin{cases} 1 & \text{for } 0.0 \leq x \leq 0.1 \\ \frac{1}{2}\left(1 + \cos\left(\frac{x-0.1}{0.4}\pi\right)\right) & \text{for } 0.1 \leq x \leq 0.5 \\ 0 & \text{for } 0.5 \leq x \leq 1.0 \end{cases}. \qquad (3.1)$$

The boundary conditions are $U(0,t) = 1$ and $U(1,t) = 0$. We again take $\nu = 3 \times 10^{-3}$. If we examine the characteristics when $\nu = 0$, we learn that a shock will start to form at $(x,t) = (0.4273, 0.2546)$ and will be fully developed at $(x,t) = (0.5, 0.4)$.

For this example we have used uniform nets consisting of 100, 200, 300, and 400 space steps. The time step k was always equal to the space step h. Figures 9 through 12 show plots of the solution for the four meshes. The curves were plotted at time intervals of 0.1 so that the rightmost curve corresponds to time $t = 0.8$. Visually it appears that by time $t = 0.4$ the solution has reached its final shape and is traveling to the right at speed one half without further change.

This leads us to ask whether or not it is possible to detect numerically when a shock has formed; that is, can we tell when the solution has reached a steady profile. We decided to recompute the case $h = k = 0.01$, but now at each time we performed an inverse Neville interpolation on the solution for five different values of U:

0.10, 0.25, 0.50, 0.75, and 0.90. In other words we ask for what values of x does the solution take the five prescribed values. We then took time differences over one time step to approximate the speed at which each of the five given values of U moves to the right. We point out that when $\nu = 0$, a point of the solution will propagate to the right at a speed equal to its amplitude. For example, the point $U = 0.6$ will appear to move at speed 0.6 to the right until it is absorbed into the shock after which it moves at the speed of the shock. When $\nu \neq 0$, we would hope to see the speed decrease gradually (or increase gradually when $U < \frac{1}{2}$) to speed one half, and we would want it to reach this ultimate velocity at or before time $t = 0.4$ when the shock for $\nu = 0$ would be fully formed. In Table 1 we give the results of these computations. In the first column are the time values midway between the time mesh points. In the succeeding columns are the velocities found by inverse interpolation and time differencing. These numbers are presented graphically in Figures 13 and 14. We have plotted also a line segment from the first velocity point to the value on the velocity axis to which it corresponds. We add that the values $U = 0.25$ and $U = 0.75$ should be absorbed into the shock at $(x,t) = (0.4333, 0.2667)$. Since it appeared that the velocities were oscillating, we decided to average them over two time steps or equivalently to difference the locations over two time steps. Table 2 gives these results. The time in column one is the midpoint of two time steps. We see that on the average these five points of the solution do approach speed one half though it is difficult to say when. One further computation was done with $\nu = 3 \times 10^{-2}$ and three prescribed values of U. These results are presented in Table 3 and Figure 17. We see that with greater diffusivity the velocities no longer oscillate.

We are also interested in achieving more accurate solutions through Richardson extrapolation. Suppose u_1 is a finite difference solution and is related to the continuous solution U by

$$u_1(x,t) = U(x,t) + h^2 f_2(x,t) + h^4 f_4(x,t) + h^6 f_6(x,t) + O(h^8) \ . \tag{3.2}$$

u_1 of course is defined only for (x,t) in the mesh. The principal errors f_2, f_4, and f_6 are defined for all x and t and are independent of h, but in (3.2) they are used only at mesh points. We are assuming $k = h$. If u_2 denotes the finite difference solution for the same problem but with the mesh refined by cutting all intervals in half, then u_2 must satisfy

$$u_2(x,t) = U(x,t) + \frac{h^2}{4} f_2(x,t) + \frac{h^4}{16} f_4(x,t) + \frac{h^6}{64} f_6(x,t) + O(h^8) \ . \tag{3.3}$$

We can now take a combination of u_1 and u_2 at each of the net points of u_1 in such a way as to eliminate the $O(h^2)$ term leaving us with an $O(h^4)$ approximation to U at those mesh points. The combination is $(4u_2 - u_1)/3$. If we divide the basic mesh into thirds and quarters, we can compute u_3 and u_4 which will have expansions similar to (3.2) but with h replaced by $h/3$ or $h/4$. With four finite difference solutions we have six possible pairs which we can extrapolate to $O(h^4)$, or we might take three solutions and extrapolate to $O(h^6)$. Even though we do not know what U is, we can perform the extrapolations and look for an agreement to a greater number of significant figures in the extrapolated solutions than in the original finite difference solutions. Let us introduce a notation for extrapolated solutions where subscripts indicate the u_i used. For example, u_{14} is extrapolated from u_1 and u_4. u_{124} is extrapolated from u_1, u_2, and u_4.

In this example we have selected the solution at $(x,t) = (0.35, 0.1)$ for extrapolation. The results are given in Table 4. In Table 5 we give the equivalent extrapolations for the flux V. We see that pairwise extrapolations give very good agreement, but for our particular data an extrapolation of three solutions appears not to yield much further improvement. Iterations in Newton's method were stopped when the relative change in two succeeding iterates was less than 2×10^{-7} except we did not compute the relative change for components less than 8×10^{-9} in magnitude. Since the maximum value of the solution is very nearly unity, the maximum absolute error in any component should be 2×10^{-7}, and for most components it should be even less. This assumes that round-off error can be neglected.

We also performed extrapolations in Example 1, but the increase in agreement was not as great as in Example 2. Another computation in which the transition from 1 to 0 was spread out over a wider interval yielded better extrapolations. On the basis of a limited number of computations involving different initial functions and different values of ν, we believe, although we cannot prove, that the deciding factor in whether or not Richardson extrapolation will yield a significant improvement is the ratio of the magnitudes of the derivatives of the solutions (including their initial values) to the inverse of the mesh spacing h. Examination of seven computations indicates that this ratio should not exceed $1/10$. We therefore propose that in the numerical computations of shocks in the presence of a small amount of diffusion, an examination of the solution should be made every few time steps to detect regions of rapid change and that new net points should be interpolated in accordance with the empirical ratio given above.

Table 1, Part 1

PROPAGATION VELOCITIES OF POINTS OF THE SOLUTION TO BURGERS' EQUATION OBTAINED BY DIFFERENCING OVER ONE TIME STEP

T	U = .10	U = .25	U = .50	U = .75	U = .90
0.005	0.131896	0.263950	0.500229	0.735932	0.867148
0.015	0.133366	0.264944	0.500385	0.735190	0.865906
0.025	0.134943	0.266091	0.500418	0.734038	0.864276
0.035	0.136955	0.267448	0.500454	0.732675	0.862196
0.045	0.139206	0.268947	0.500494	0.731170	0.859946
0.055	0.141434	0.270197	0.500539	0.729910	0.857525
0.065	0.144154	0.272047	0.500589	0.728056	0.854938
0.075	0.146622	0.274142	0.500644	0.725949	0.852194
0.085	0.149018	0.276230	0.500704	0.723863	0.850064
0.095	0.150915	0.278279	0.500771	0.721798	0.847822
0.105	0.154249	0.281238	0.500844	0.718828	0.844465
0.115	0.157861	0.284647	0.500927	0.715464	0.840790
0.125	0.161544	0.287128	0.501012	0.712926	0.836811
0.135	0.165826	0.291021	0.501101	0.709069	0.832552
0.145	0.169377	0.296058	0.501209	0.704125	0.828908
0.155	0.170938	0.300379	0.501313	0.699780	0.826832
0.165	0.175938	0.304443	0.501428	0.695773	0.821377
0.175	0.181671	0.312055	0.501507	0.688387	0.815165
0.185	0.188251	0.319649	0.501691	0.680944	0.808207
0.195	0.195237	0.322264	0.501701	0.678189	0.800558
0.205	0.197095	0.334374	0.501983	0.666758	0.799516
0.215	0.202681	0.345836	0.501764	0.655675	0.791408
0.225	0.213199	0.344235	0.502387	0.656417	0.779761
0.235	0.226338	0.366175	0.501641	0.637038	0.765761
0.245	0.240473	0.371462	0.502931	0.629535	0.757152
0.255	0.232511	0.373649	0.501104	0.628673	0.756301
0.265	0.251871	0.414196	0.503774	0.594255	0.733844
0.275	0.284354	0.374014	0.500032	0.620296	0.703383
0.285	0.295886	0.432928	0.504988	0.579894	0.714271
0.295	0.265001	0.405740	0.498274	0.583154	0.699956
0.305	0.352190	0.433709	0.506678	0.583511	0.637966
0.315	0.348102	0.464705	0.495878	0.520129	0.679412
0.325	0.276867	0.403774	0.508778	0.615186	0.648747
0.335	0.514069	0.533543	0.492992	0.486143	0.538052
0.345	0.182232	0.360312	0.511175	0.620607	0.720932
0.355	0.543654	0.564241	0.489895	0.464828	0.514237
0.365	0.273807	0.351342	0.513646	0.617004	0.660511
0.375	0.477959	0.591057	0.486864	0.449309	0.534815
0.385	0.529732	0.342818	0.516005	0.611680	0.550754
0.395	0.248122	0.613707	0.484124	0.438972	0.606499

Table 1, Part 2

PROPAGATION VELOCITIES OF POINTS OF THE SOLUTION TO BURGERS' EQUATION OBTAINED BY DIFFERENCING OVER ONE TIME STEP

T	U = .10	U = .25	U = .50	U = .75	U = .90
0.405	0.877383	0.335297	0.518095	0.605468	0.398020
0.415	-0.072516	0.632246	0.481803	0.432635	0.724722
0.425	0.991424	0.328991	0.519848	0.599165	0.289208
0.435	-0.159028	0.647034	0.479933	0.429077	0.804172
0.445	1.096824	0.323949	0.521249	0.593422	0.255958
0.455	-0.236592	0.658496	0.478493	0.427235	0.813541
0.465	1.189696	0.320022	0.522329	0.588570	0.231211
0.475	-0.303239	0.667245	0.477415	0.426374	0.819653
0.485	1.268497	0.317076	0.523140	0.584678	0.213138
0.495	-0.357689	0.673741	0.476632	0.426016	0.823654
0.505	1.332260	0.314878	0.523732	0.581711	0.200089
0.515	-0.400985	0.678557	0.476071	0.425894	0.826185
0.525	1.382645	0.313301	0.524162	0.579480	0.190796
0.535	-0.433887	0.682015	0.475679	0.425874	0.827933
0.545	1.420732	0.312150	0.524465	0.577886	0.184145
0.555	-0.458746	0.684542	0.475406	0.425880	0.828987
0.565	1.449442	0.311352	0.524681	0.576715	0.179499
0.575	-0.476679	0.686304	0.475220	0.425900	0.829816
0.585	1.470086	0.310772	0.524830	0.575922	0.176159
0.595	-0.489877	0.687591	0.475092	0.425907	0.830251
0.605	1.485286	0.310384	0.524936	0.575334	0.173886
0.615	-0.499025	0.688460	0.475007	0.425919	0.830683
0.625	1.495790	0.310097	0.525006	0.574961	0.172218
0.635	-0.505721	0.689104	0.474949	0.425916	0.830845
0.645	1.503500	0.309915	0.525057	0.574670	0.171133
0.655	-0.510187	0.689521	0.474911	0.425920	0.831093
0.665	1.508623	0.309775	0.525089	0.574504	0.170299
0.675	-0.513504	0.689841	0.474885	0.425912	0.831132
0.685	1.512446	0.309692	0.525114	0.574359	0.169798
0.695	-0.515613	0.690036	0.474869	0.425914	0.831228
0.705	1.514862	0.309623	0.525129	0.574291	0.169376
0.715	-0.517240	0.690195	0.474857	0.425906	0.831273
0.725	1.516745	0.309587	0.525141	0.574216	0.169157
0.735	-0.518203	0.690282	0.474851	0.425908	0.831379
0.745	1.517845	0.309553	0.525147	0.574192	0.168937
0.755	-0.519004	0.690363	0.474846	0.425901	0.831343
0.765	1.518776	0.309538	0.525153	0.574151	0.168851
0.775	-0.519426	0.690399	0.474843	0.425903	0.831421
0.785	1.519256	0.309520	0.525155	0.574147	0.168731
0.795	-0.519826	0.690441	0.474841	0.425898	0.831380

Table 2, Part 1

PROPAGATION VELOCITIES OF POINTS OF THE SOLUTION TO BURGERS' EQUATION OBTAINED BY DIFFERENCING OVER TWO TIME STEPS

T	U = .10	U = .25	U = .50	U = .75	U = .90
0.010	0.132631	0.264447	0.500307	0.735561	0.866527
0.020	0.134154	0.265517	0.500401	0.734614	0.865091
0.030	0.135949	0.266769	0.500436	0.733356	0.863236
0.040	0.138080	0.268197	0.500473	0.731922	0.861071
0.050	0.140320	0.269572	0.500516	0.730540	0.858735
0.060	0.142794	0.271122	0.500564	0.728983	0.856231
0.070	0.145388	0.273094	0.500616	0.727002	0.853566
0.080	0.147820	0.275186	0.500674	0.724906	0.851129
0.090	0.149966	0.277254	0.500737	0.722830	0.848943
0.100	0.152582	0.279758	0.500807	0.720313	0.846143
0.110	0.156055	0.282942	0.500885	0.717146	0.842627
0.120	0.159702	0.285887	0.500969	0.714195	0.838800
0.130	0.163685	0.289074	0.501056	0.710997	0.834681
0.140	0.167601	0.293539	0.501155	0.706597	0.830730
0.150	0.170157	0.298218	0.501261	0.701952	0.827870
0.160	0.173438	0.302411	0.501370	0.697776	0.824104
0.170	0.178804	0.308249	0.501467	0.692080	0.818271
0.180	0.184961	0.315852	0.501599	0.684665	0.811686
0.190	0.191744	0.320956	0.501696	0.679566	0.804382
0.200	0.196166	0.328319	0.501842	0.672473	0.800037
0.210	0.199888	0.340105	0.501873	0.661216	0.795462
0.220	0.207940	0.345035	0.502075	0.656046	0.785584
0.230	0.219768	0.355205	0.502014	0.646727	0.772761
0.240	0.233405	0.368818	0.502286	0.633286	0.761456
0.250	0.236492	0.372555	0.502017	0.629104	0.756726
0.260	0.242191	0.393922	0.502439	0.611464	0.745072
0.270	0.268112	0.394105	0.501903	0.607275	0.718613
0.280	0.290120	0.403471	0.502510	0.600095	0.708827
0.290	0.280443	0.419334	0.501631	0.581524	0.707113
0.300	0.308595	0.419724	0.502476	0.583332	0.668961
0.310	0.350146	0.449207	0.501278	0.551820	0.658689
0.320	0.312484	0.434239	0.502328	0.567657	0.664079
0.330	0.395468	0.468658	0.500885	0.550664	0.593399
0.340	0.348150	0.446927	0.502083	0.553375	0.629492
0.350	0.362943	0.462276	0.500535	0.542717	0.617584
0.360	0.408730	0.457791	0.501770	0.540916	0.587374
0.370	0.375883	0.471199	0.500255	0.533156	0.597663
0.380	0.503845	0.466937	0.501434	0.530494	0.542784
0.390	0.388927	0.478262	0.500064	0.525326	0.578626
0.400	0.562752	0.474502	0.501109	0.522220	0.502259

Table 2, Part 2

PROPAGATION VELOCITIES OF POINTS OF THE SOLUTION TO BURGERS' EQUATION OBTAINED BY DIFFERENCING OVER TWO TIME STEPS

T	U = .10	U = .25	U = .50	U = .75	U = .90
0.410	0.402434	0.483771	0.499949	0.519051	0.561371
0.420	0.459454	0.480618	0.500825	0.515900	0.506965
0.430	0.416198	0.488012	0.499890	0.514121	0.546690
0.440	0.468898	0.485491	0.500591	0.511249	0.530065
0.450	0.430116	0.491222	0.499871	0.510328	0.534749
0.460	0.476552	0.489259	0.500411	0.507902	0.522376
0.470	0.443229	0.493633	0.499872	0.507472	0.525432
0.480	0.482629	0.492160	0.500277	0.505526	0.516395
0.490	0.455404	0.495408	0.499886	0.505347	0.518396
0.500	0.487286	0.494309	0.500182	0.503863	0.511871
0.510	0.465638	0.496717	0.499901	0.503802	0.513137
0.520	0.490830	0.495929	0.500116	0.502687	0.508490
0.530	0.474379	0.497658	0.499920	0.502677	0.509364
0.540	0.493422	0.497082	0.500072	0.501880	0.506039
0.550	0.480993	0.498346	0.499935	0.501883	0.506566
0.560	0.495348	0.497947	0.500043	0.501297	0.504243
0.570	0.486381	0.498828	0.499950	0.501307	0.504657
0.580	0.496704	0.498538	0.500025	0.500911	0.502987
0.590	0.490105	0.499181	0.499961	0.500914	0.503205
0.600	0.497704	0.498987	0.500014	0.500620	0.502068
0.610	0.493130	0.499422	0.499971	0.500626	0.502284
0.620	0.498383	0.499278	0.500006	0.500440	0.501450
0.630	0.495035	0.499600	0.499977	0.500438	0.501531
0.640	0.498890	0.499509	0.500003	0.500293	0.500989
0.650	0.496656	0.499718	0.499984	0.500295	0.501113
0.660	0.499218	0.499648	0.500000	0.500212	0.500696
0.670	0.497560	0.499808	0.499987	0.500208	0.500715
0.680	0.499471	0.499766	0.499999	0.500135	0.500465
0.690	0.498417	0.499864	0.499991	0.500136	0.500513
0.700	0.499625	0.499829	0.499999	0.500102	0.500302
0.710	0.498811	0.499909	0.499993	0.500098	0.500324
0.720	0.499752	0.499891	0.499999	0.500061	0.500215
0.730	0.499271	0.499934	0.499996	0.500062	0.500268
0.740	0.499821	0.499917	0.499999	0.500050	0.500158
0.750	0.499421	0.499958	0.499996	0.500046	0.500140
0.760	0.499886	0.499950	0.499999	0.500026	0.500097
0.770	0.499675	0.499968	0.499998	0.500027	0.500136
0.780	0.499915	0.499959	0.499999	0.500025	0.500076
0.790	0.499715	0.499980	0.499998	0.500022	0.500055

Table 3, Part 1

PROPAGATION VELOCITIES OF POINTS OF THE SOLUTION TO BURGERS' EQUATION OBTAINED BY DIFFERENCING OVER ONE TIME STEP

T	U = .25	U = .50	U = .75
0.005	0.392306	0.499836	0.607234
0.015	0.405520	0.499951	0.594442
0.025	0.419353	0.499956	0.580939
0.035	0.432555	0.499989	0.568132
0.045	0.444864	0.500059	0.555963
0.055	0.455285	0.500159	0.545359
0.065	0.464216	0.500256	0.536250
0.075	0.471848	0.500321	0.528484
0.085	0.478082	0.500354	0.522167
0.095	0.483595	0.500363	0.516598
0.105	0.487853	0.500357	0.512309
0.115	0.491822	0.500342	0.508316
0.125	0.494684	0.500322	0.505455
0.135	0.497534	0.500299	0.502577
0.145	0.499445	0.500277	0.500703
0.155	0.501464	0.500255	0.498617
0.165	0.502738	0.500234	0.497421
0.175	0.504203	0.500214	0.495904
0.185	0.504805	0.500197	0.495173
0.195	0.506253	0.500180	0.494071
0.205	0.506590	0.500167	0.493655
0.215	0.506812	0.500151	0.492861
0.225	0.507502	0.500143	0.492657
0.235	0.508472	0.500127	0.492096
0.245	0.507854	0.500123	0.492032
0.255	0.508197	0.500116	0.491649
0.265	0.508744	0.500091	0.491676
0.275	0.508641	0.500127	0.491432
0.285	0.508358	0.500114	0.491512
0.295	0.508976	0.499986	0.491379
0.305	0.508747	0.500102	0.491488
0.315	0.508325	0.500266	0.491444
0.325	0.508544	0.499980	0.491563
0.335	0.508810	0.499854	0.491593
0.345	0.508348	0.500254	0.491708
0.355	0.508091	0.500238	0.491802
0.365	0.508277	0.499854	0.491908
0.375	0.508178	0.500008	0.492027
0.385	0.507919	0.500234	0.492153
0.395	0.507722	0.499996	0.492357

Table 3, Part 2

PROPAGATION VELOCITIES OF POINTS OF THE SOLUTION TO BURGERS' EQUATION OBTAINED BY DIFFERENCING OVER ONE TIME STEP

T	U = .25	U = .50	U = .75
0.405	0.507758	0.499991	0.492325
0.415	0.507514	0.500209	0.492563
0.425	0.507436	0.500019	0.492791
0.435	0.507186	0.499881	0.492921
0.445	0.507187	0.500178	0.492731
0.455	0.506891	0.500212	0.493249
0.465	0.506900	0.499904	0.493423
0.475	0.506590	0.499943	0.493392
0.485	0.506626	0.500198	0.493308
0.495	0.506292	0.500119	0.493952
0.505	0.506353	0.499942	0.493831
0.515	0.506002	0.500020	0.493943
0.525	0.506085	0.500127	0.494048
0.535	0.505718	0.500621	0.494456
0.545	0.505825	0.500007	0.494156
0.555	0.505444	0.500047	0.494633
0.565	0.505571	0.500072	0.494691
0.575	0.505178	0.500047	0.494834
0.585	0.505326	0.500039	0.494599
0.595	0.504921	0.500048	0.495273
0.605	0.505089	0.500051	0.495140
0.615	0.504674	0.500044	0.495264
0.625	0.504859	0.500044	0.495124
0.635	0.504434	0.500044	0.495752
0.645	0.504635	0.500044	0.495498
0.655	0.504201	0.500041	0.495754
0.665	0.504416	0.500041	0.495611
0.675	0.503972	0.500038	0.496131
0.685	0.504200	0.500037	0.495851
0.695	0.503745	0.500033	0.496222
0.705	0.503982	0.500031	0.496021
0.715	0.503514	0.500027	0.496481
0.725	0.503756	0.500023	0.496202
0.735	0.503274	0.500017	0.496627
0.745	0.503516	0.500011	0.496372
0.755	0.503015	0.500002	0.496817
0.765	0.503251	0.499994	0.496529
0.775	0.502725	0.499980	0.496972
0.785	0.502943	0.499967	0.496680
0.795	0.502384	0.499948	0.497127

Table 4

Richardson Extrapolations

of Example 2 at

$(x, t) = (0.35, 0.1)$

u_1	=	0.50032235050
u_2	=	0.50008107259
u_3	=	0.50003613622
u_4	=	0.50002035943
u_{12}	=	0.50000064662
u_{13}	=	0.50000035944
u_{14}	=	0.50000022669
u_{23}	=	0.50000018712
u_{24}	=	0.50000012171
u_{34}	=	0.50000007499
u_{123}	=	0.50000012969
u_{124}	=	0.50000008672
u_{134}	=	0.50000005602
u_{234}	=	0.50000003761
u_{1234}	=	0.50000003147

Table 5

Richardson Extrapolations

of Example 2 at

$(x, t) = (0.35, 0.1)$

v_1	=	-0.018540499230
v_2	=	-0.018499265130
v_3	=	-0.018491656850
v_4	=	-0.018488994690
v_{12}	=	-0.018485520430
v_{13}	=	-0.018485551553
v_{14}	=	-0.018485561054
v_{23}	=	-0.018485570226
v_{24}	=	-0.018485571210
v_{34}	=	-0.018485571913
v_{123}	=	-0.018485576451
v_{124}	=	-0.018485574595
v_{134}	=	-0.018485573270
v_{234}	=	-0.018485572475
v_{1234}	=	-0.018485572210

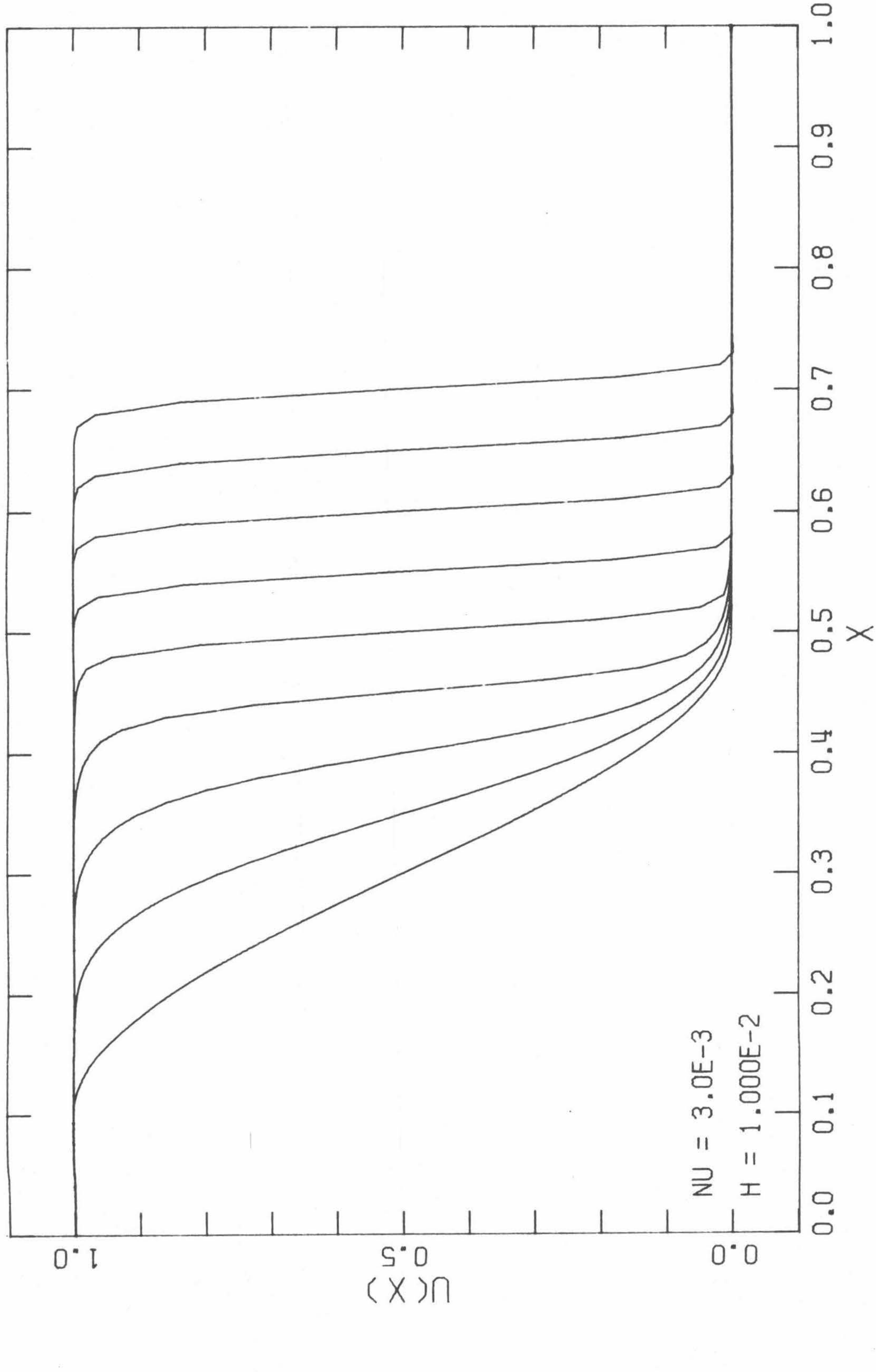


Figure 9

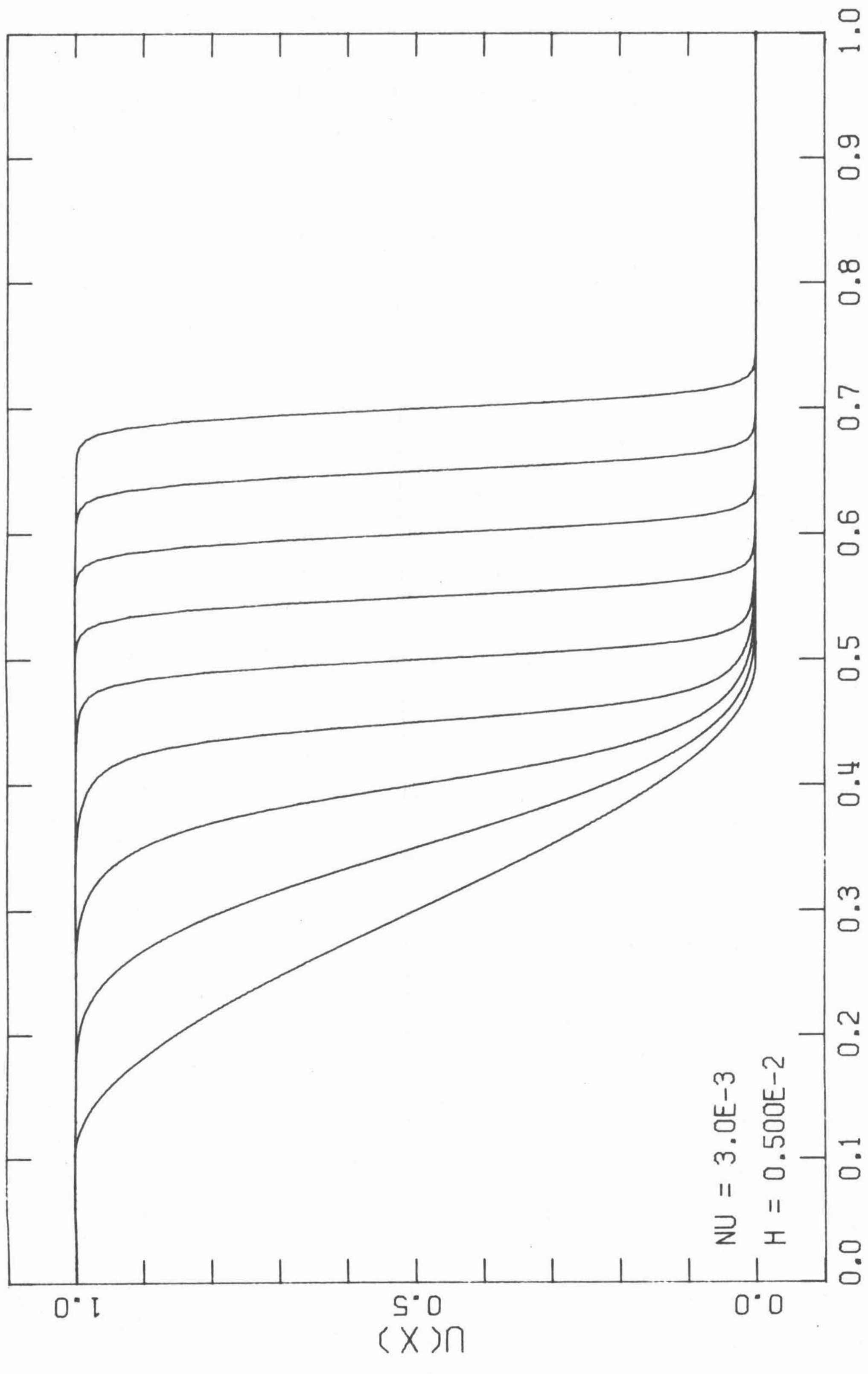


Figure 10

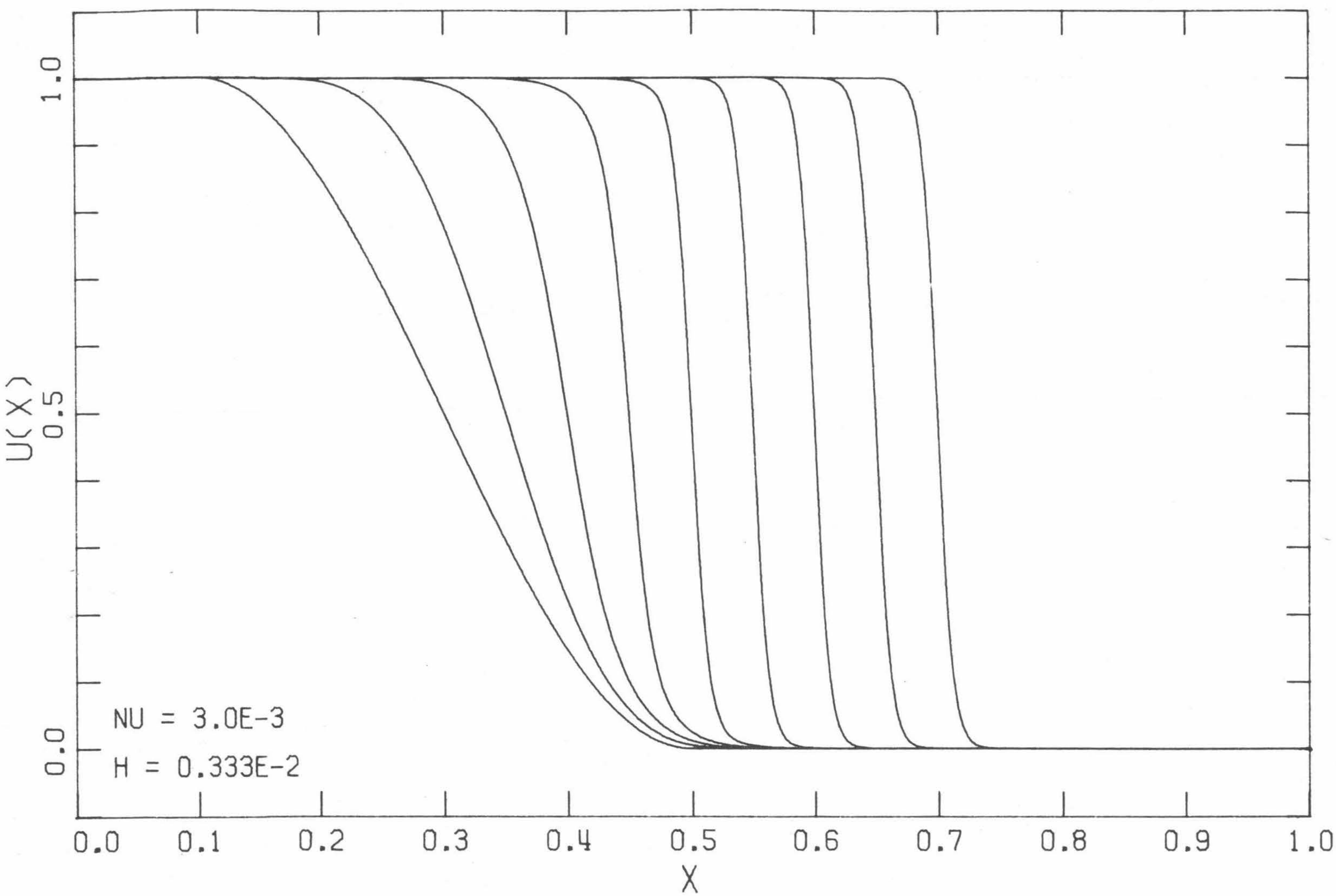


Figure 11

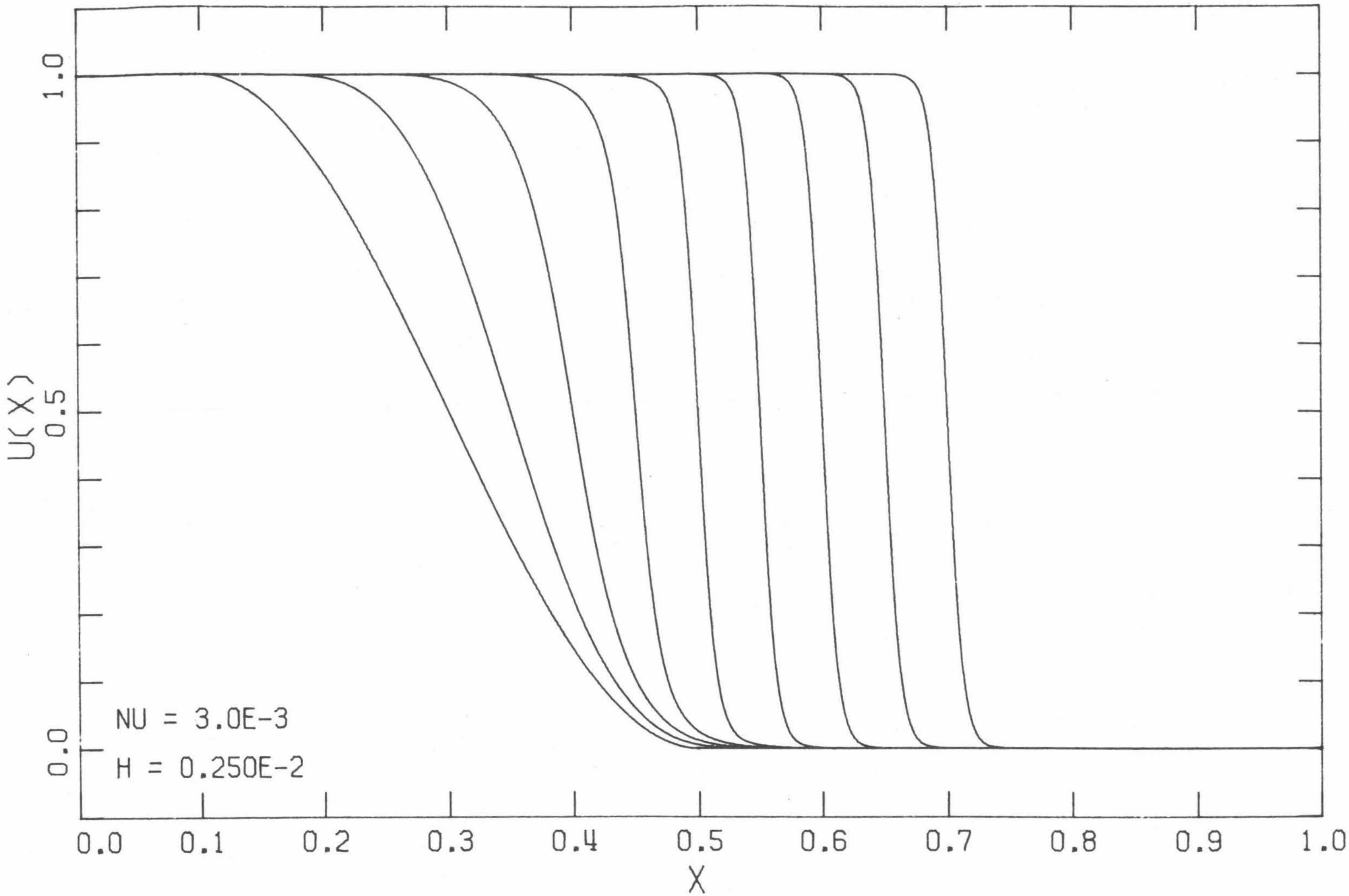


Figure 12

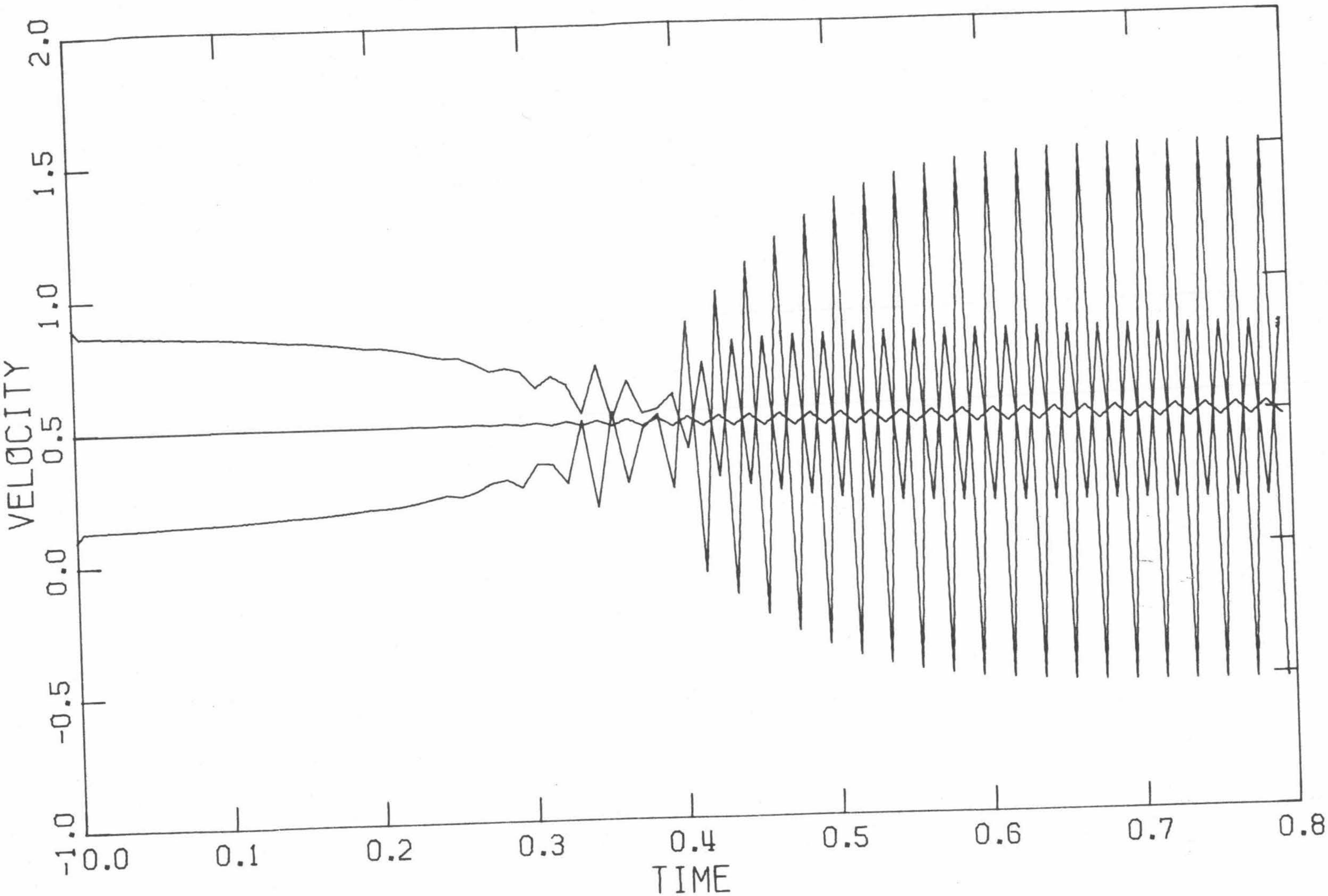


Figure 13

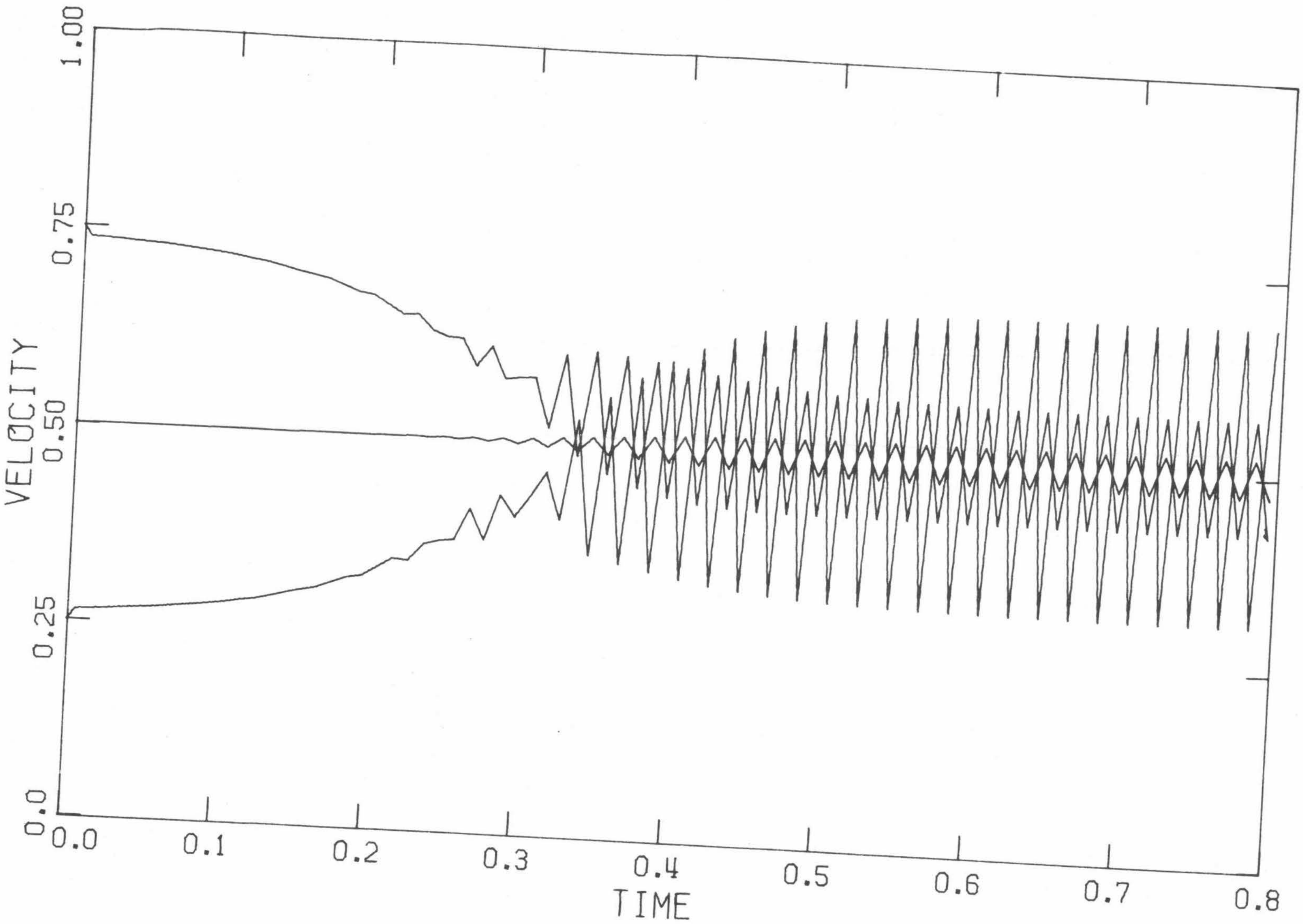


Figure 14

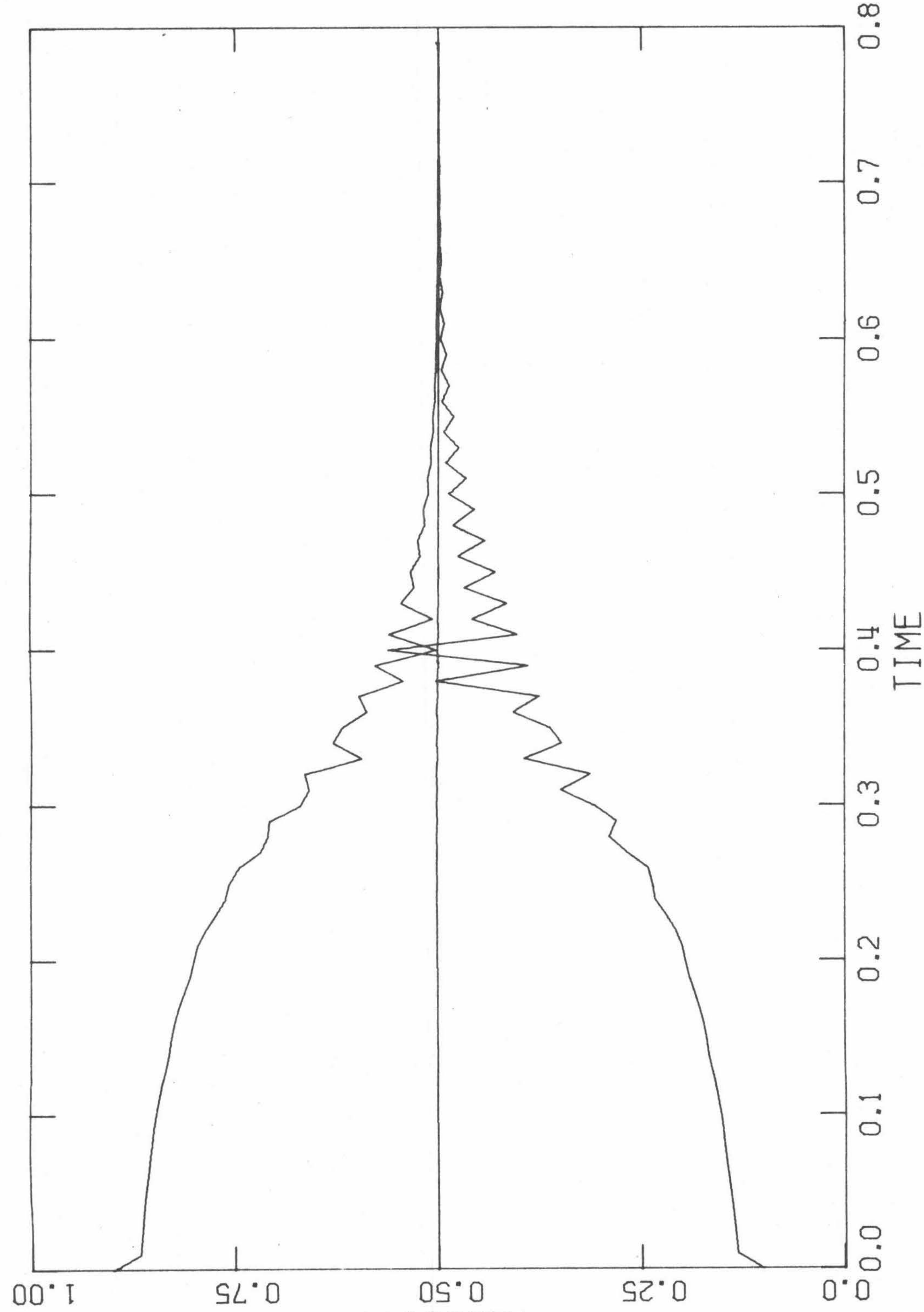


Figure 15

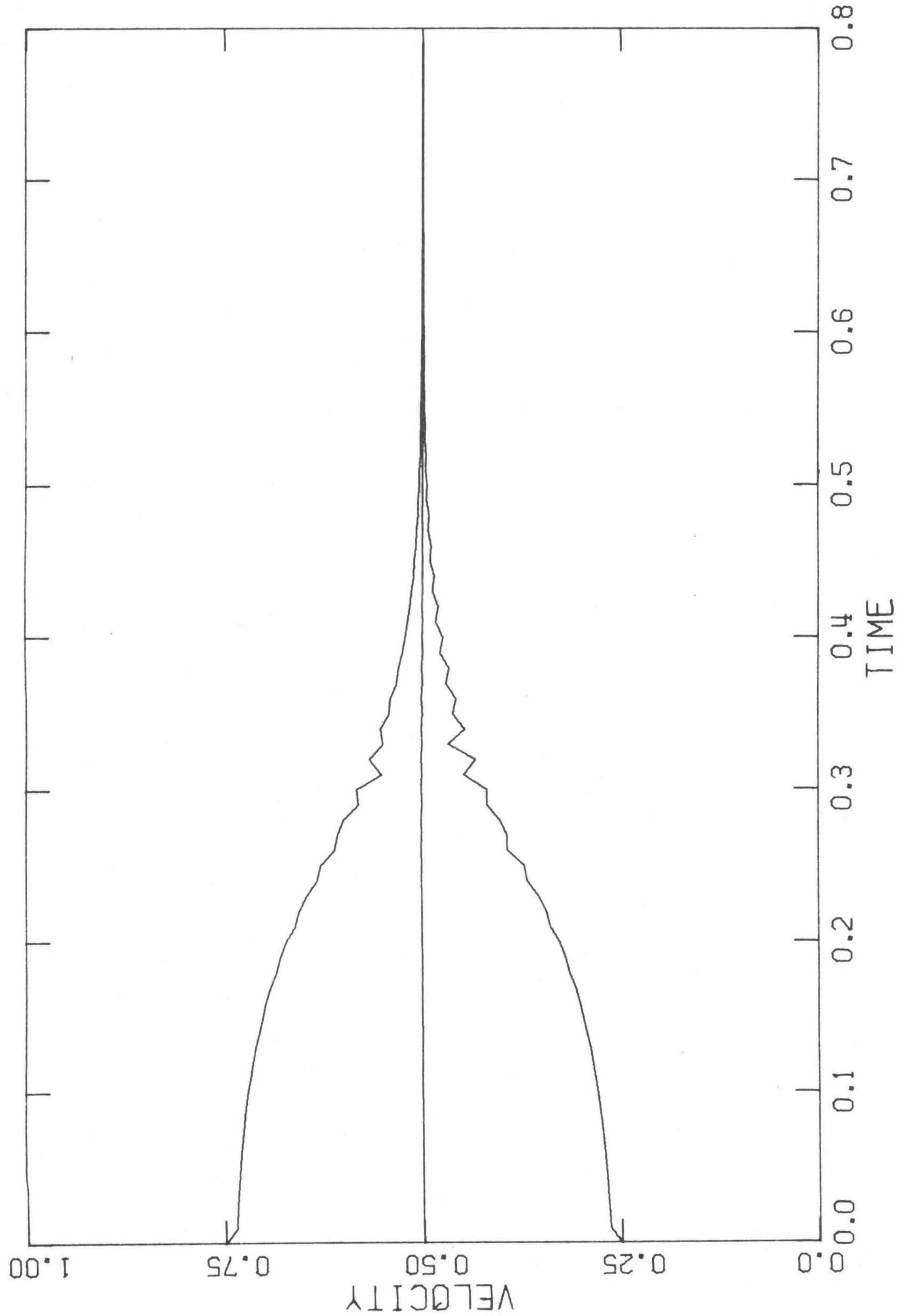


Figure 16

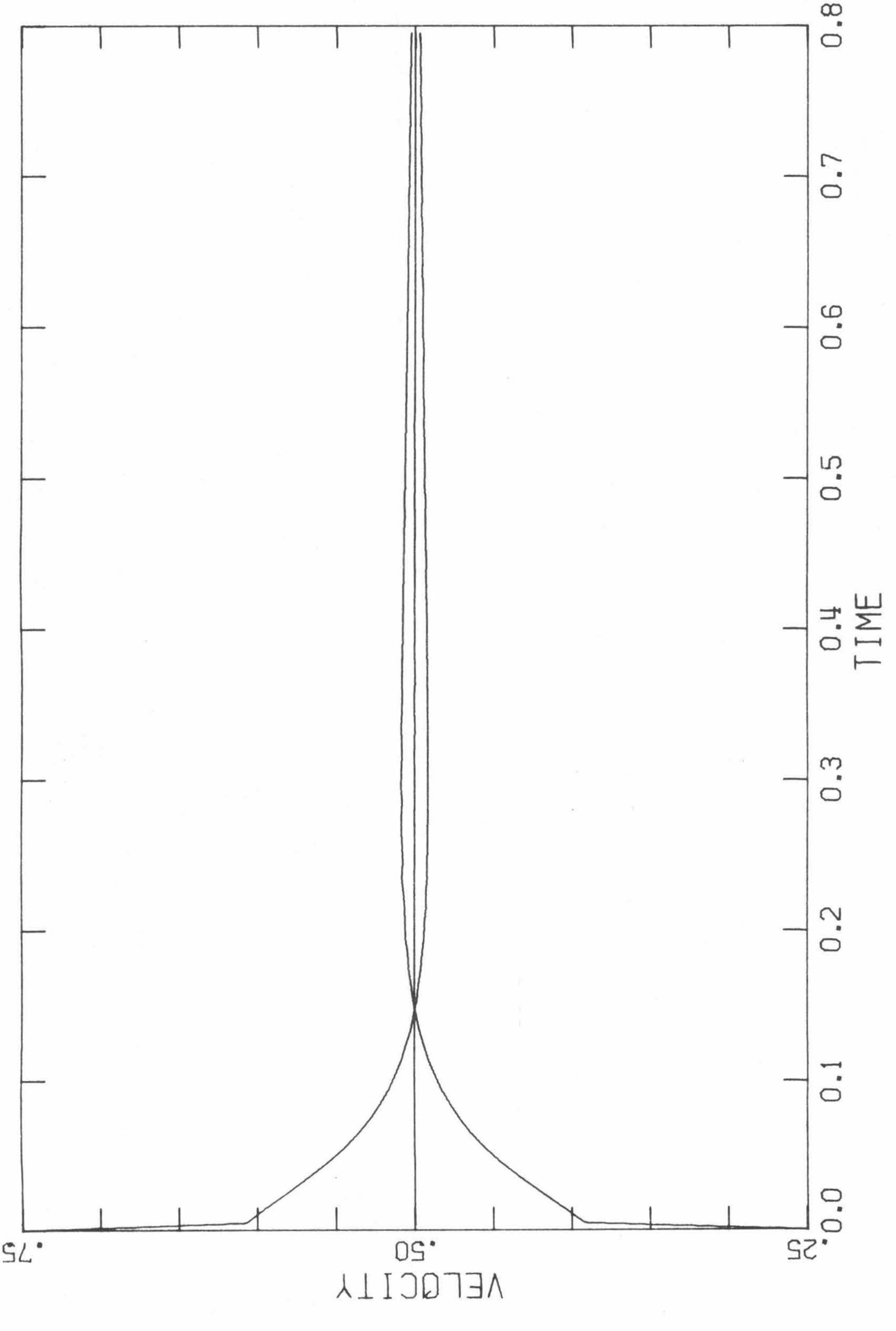


Figure 17

III.4 Example 3: Interaction of Two Shocks

In (1.2) we gave a transformation which converts solutions of the heat equation into solutions of Burgers' equation. This transformation has been used to advantage by Whitham [1972] to construct an exact solution of Burgers' equation representing the overtaking of one shock by another. This solution is

$$
\begin{aligned}
W(x,t) = & \left\{ \frac{1}{10} e^{-\frac{(x-\frac{1}{2})}{20\nu} - \frac{99t}{400\nu}} + \frac{1}{2} e^{-\frac{(x-\frac{1}{2})}{4\nu} - \frac{3t}{16\nu}} \right. \\
& \left. + e^{-\frac{(x-\frac{3}{8})}{2\nu}} \right\} \cdot \left\{ e^{-\frac{(x-\frac{1}{2})}{20\nu} - \frac{99t}{400\nu}} \right. \\
& \left. + e^{-\frac{(x-\frac{1}{2})}{4\nu} - \frac{3t}{16\nu}} + e^{-\frac{(x-\frac{3}{8})}{2\nu}} \right\}^{-1} .
\end{aligned}
\tag{4.1}
$$

In the limit as ν goes to zero we have the initial condition

$$
W_0(x) = \left\{ \begin{array}{ll} 1 & \text{for } x < 0.25 \\ \frac{1}{2} & \text{for } 0.25 < x < 0.5 \\ \frac{1}{10} & \text{for } 0.5 < x \end{array} \right\} .
\tag{4.2}
$$

For $\nu = 0$ the solution would be a shock moving at speed 0.75 and starting at $x = 0.25$ overtaking a shock moving at speed 0.3 and starting at $x = 0.5$. The shocks would merge at $(x,t) = (2/3,\ 5/9)$ and continue as a single shock of strength 0.9 and speed 0.55. For a small, positive ν the initial condition would resemble a staircase function with the corners rounded.

In Example 3 we take $\nu = 3 \times 10^{-3}$. The initial and boundary conditions are

$$U(x,0) = W(x,0), \tag{4.3a}$$

$$U(0,t) = W(0,t), \tag{4.3b}$$

$$U(1,t) = W(1,t). \tag{4.3c}$$

We use uniform meshes consisting of 100, 200, 300, and 400 space steps over the interval $0 \leq x \leq 1$. We always take $k = h$. Graphs of the solutions for each of the four nets are given in Figures 18 through 21. The curves are plotted at time intervals of 0.1 so that the right-most curve, which goes off the right edge at about $U = 0.96$, corresponds to $t = 1.2$.

As in Example 2 we have selected a point at which to perform all of the possible Richardson extrapolations with four solutions u_1, u_2, u_3, and u_4. These are given in Tables 6 and 7. We notice that u_{12} is more accurate than u_4. In order to find if this might be true in general, we select twelve different points, and at each tenth of a time unit, for comparison. In Table 8, Part 1 we give u_1, u_2 and u_3. In Part 2 we list u_4, U, and u_{12}. We see that u_4 has at best three digit accuracy as is the case with u_{12}. We conclude that

u_{12} gives roughly the same number of correct digits as u_4. This is a significant result when we recall that the amount of computation required by various meshes is proportional to the square of the ratio of steps. In particular u_2 requires four times as much computing as u_1, and u_4 requires sixteen times as much as u_1. Thus if u_{12} is comparable to u_4 in accuracy but requires only 5/16 as much computing as u_4, it will clearly be preferable to compute u_1 and u_2 and to extrapolate rather than to compute u_4 with a very fine mesh. We note incidentally that u_1, u_2, u_3 and u_4 become progressively more accurate. This means we have not yet refined the mesh to such an extent that the increased amount of computation causes significant round-off errors. For reference, Part 3 gives two more extrapolations involving u_3.

Table 6

Richardson Extrapolations

of Example 3 at

$(x, t) = (0.56, 0.2)$

u_1	=	0.30225365244
u_2	=	0.30059271979
u_3	=	0.30026632734
u_4	=	0.30015060111
u_{12}	=	0.30003907557
u_{13}	=	0.30001791170
u_{14}	=	0.30001039769
u_{23}	=	0.30000521338
u_{24}	=	0.30000322822
u_{34}	=	0.30000181024
u_{123}	=	0.30000098061
u_{124}	=	0.30000083839
u_{134}	=	0.30000073681
u_{234}	=	0.30000067586
u_{1234}	=	0.30000065555

Exact Solution: $U = 0.30000056686$

Table 7

Richardson Extrapolations

of Example 3 at

(x, t) = (0.56, 0.2)

v_1	=	-0.020013464931
v_2	=	-0.020008700290
v_3	=	-0.020004435706
v_4	=	-0.020002664316
v_{12}	=	-0.020007112076
v_{13}	=	-0.020003307053
v_{14}	=	-0.020001944275
v_{23}	=	-0.020001024039
v_{24}	=	-0.020000652325
v_{34}	=	-0.020000386815
v_{123}	=	-0.020000263034
v_{124}	=	-0.020000221675
v_{134}	=	-0.020000192132
v_{234}	=	-0.020000174406
v_{1234}	=	-0.020000168498

Exact Solution: V = -0.020000182204

Table 8, Part 1

Example 3

t	x	u_1	u_2	u_3
0.1	0.50	0.45162932809	0.45213909765	0.45223813643
0.2	0.50	0.49260003939	0.49284055156	0.49288720790
0.3	0.55	0.47374691848	0.47464422617	0.47481860040
0.4	0.60	0.42107292706	0.42301439573	0.42335725720
0.5	0.65	0.33964540345	0.34073709968	0.34090349587
0.6	0.70	0.33473215370	0.32641934827	0.32523826859
0.7	0.75	0.44725589532	0.43114939356	0.42808375484
0.8	0.80	0.60225117856	0.59134149643	0.58882496729
0.9	0.85	0.74645749560	0.74395011947	0.74323437130
1.0	0.90	0.85541977691	0.856508357749	0.85674654929
1.1	0.97	0.44512538419	0.42816511566	0.42491123354
1.2	0.99	0.99248461873	0.98966518105	0.98983381566

Table 8, Part 2

Example 3

t	x	u_4	U	u_{12}
0.1	0.50	0.45227329626	0.45231905503	0.45230902084
0.2	0.50	0.49290374042	0.49292520906	0.49292072228
0.3	0.55	0.47488025263	0.47496004739	0.47494332873
0.4	0.60	0.42347622629	0.42362844757	0.42366155195
0.5	0.65	0.34095825381	0.34102590553	0.34110099842
0.6	0.70	0.32485439735	0.32438289934	0.32364841313
0.7	0.75	0.42702022924	0.42566448866	0.42578055964
0.8	0.80	0.58790276994	0.58668558549	0.58770493572
0.9	0.85	0.74296136413	0.74259164202	0.74311432743
1.0	0.90	0.85683280384	0.85694563756	0.85687121768
1.1	0.97	0.42378437932	0.42235376627	0.42251169282
1.2	0.99	0.99015457747	0.99161484779	0.98872536849

Table 8, Part 3

Example 3

t	x	u_{23}	u_{123}
0.1	0.50	0.452317367 45	0.45231841078
0.2	0.50	0.49292453297	0.49292500931
0.3	0.55	0.47495809978	0.47495994616
0.4	0.60	0.42363154638	0.42362779568
0.5	0.65	0.34103661282	0.34102856462
0.6	0.70	0.32429340485	0.32437402881
0.7	0.75	0.42563124386	0.42561257939
0.8	0.80	0.58681174398	0.58670009501
0.9	0.85	0.74266177276	0.74260520343
1.0	0.90	0.85693710273	0.85694533836
1.1	0.97	0.42230812784	0.42228268222
1.2	0.99	0.98996872335	0.99012414271

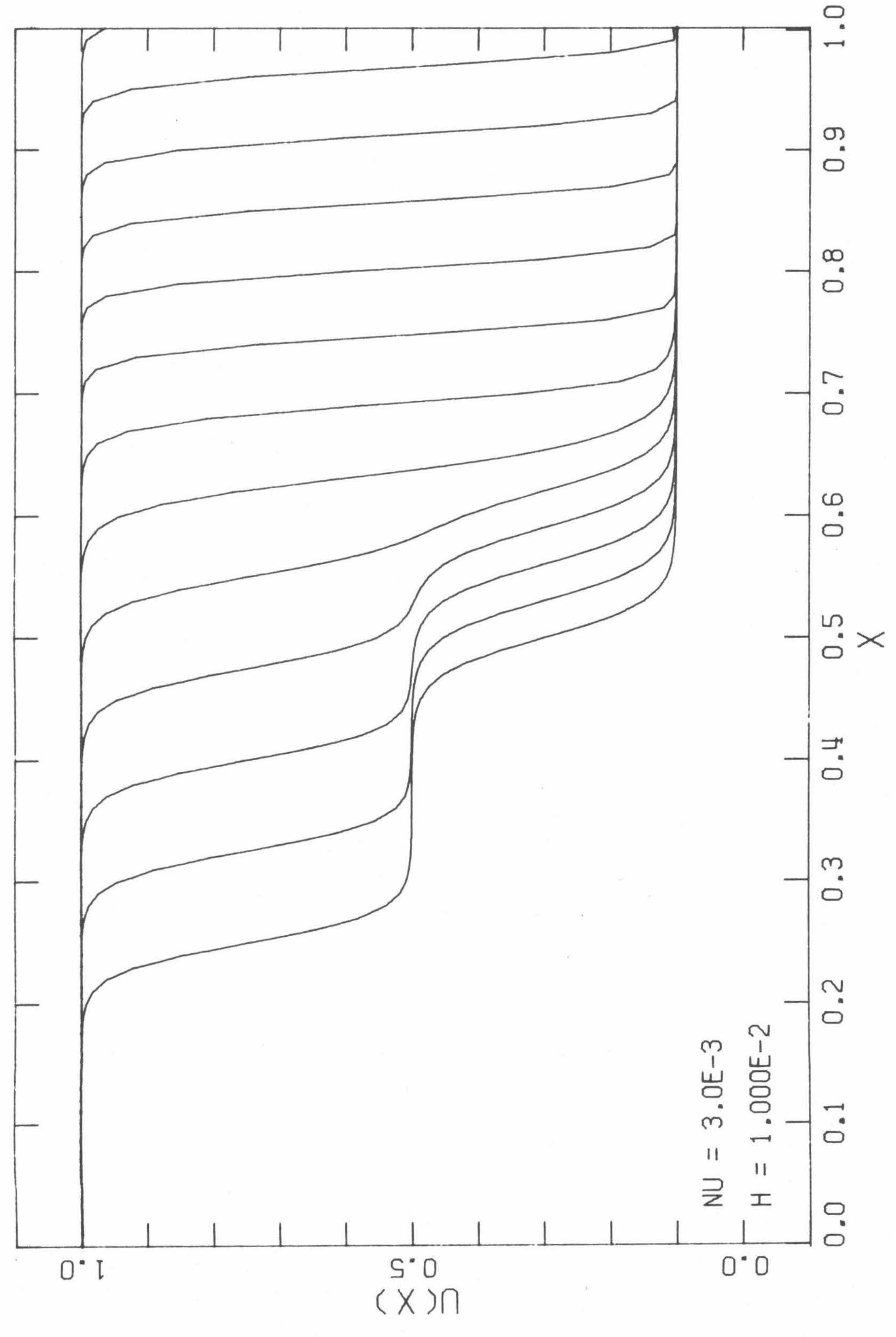


Figure 18

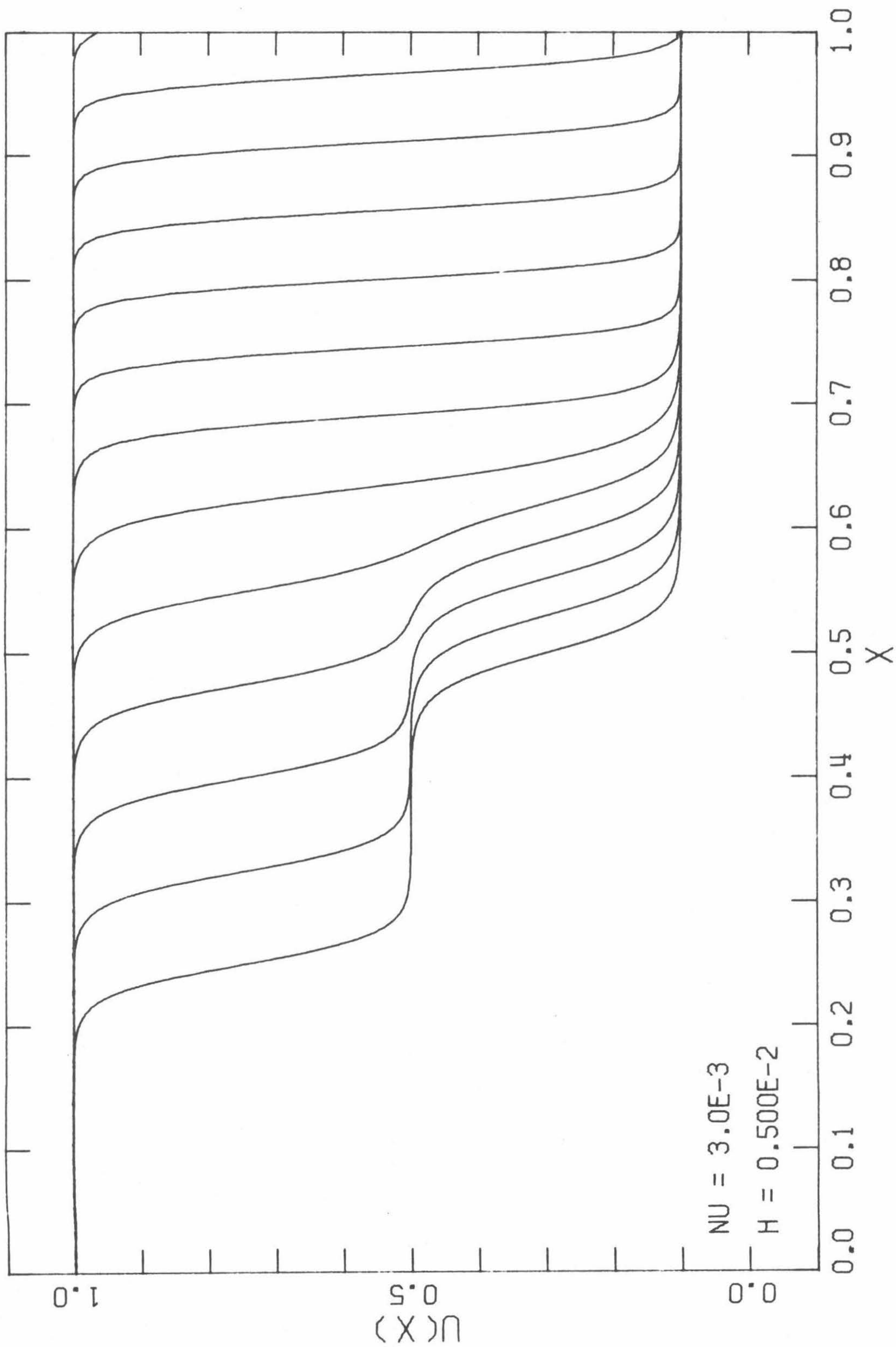


Figure 19

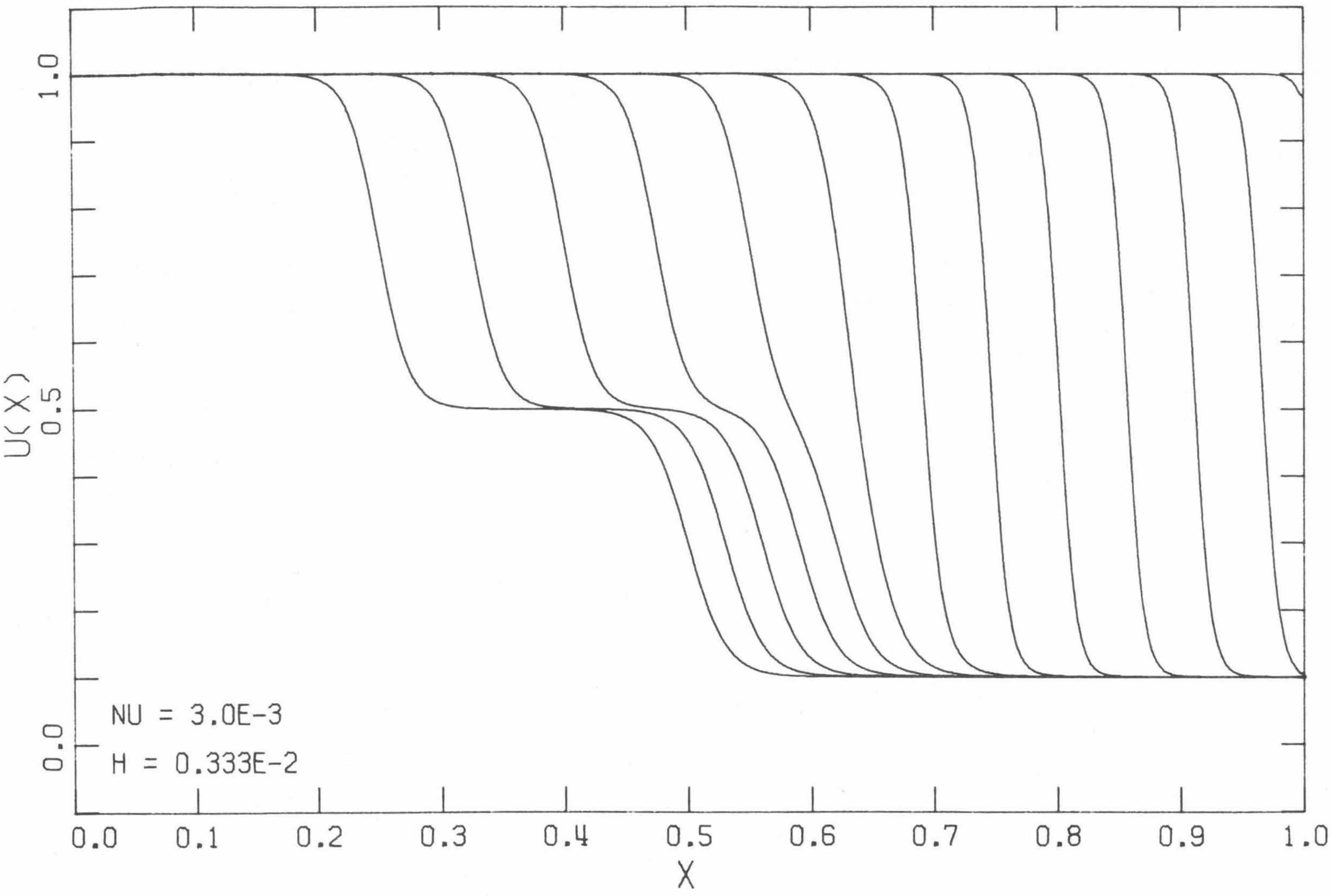


Figure 20

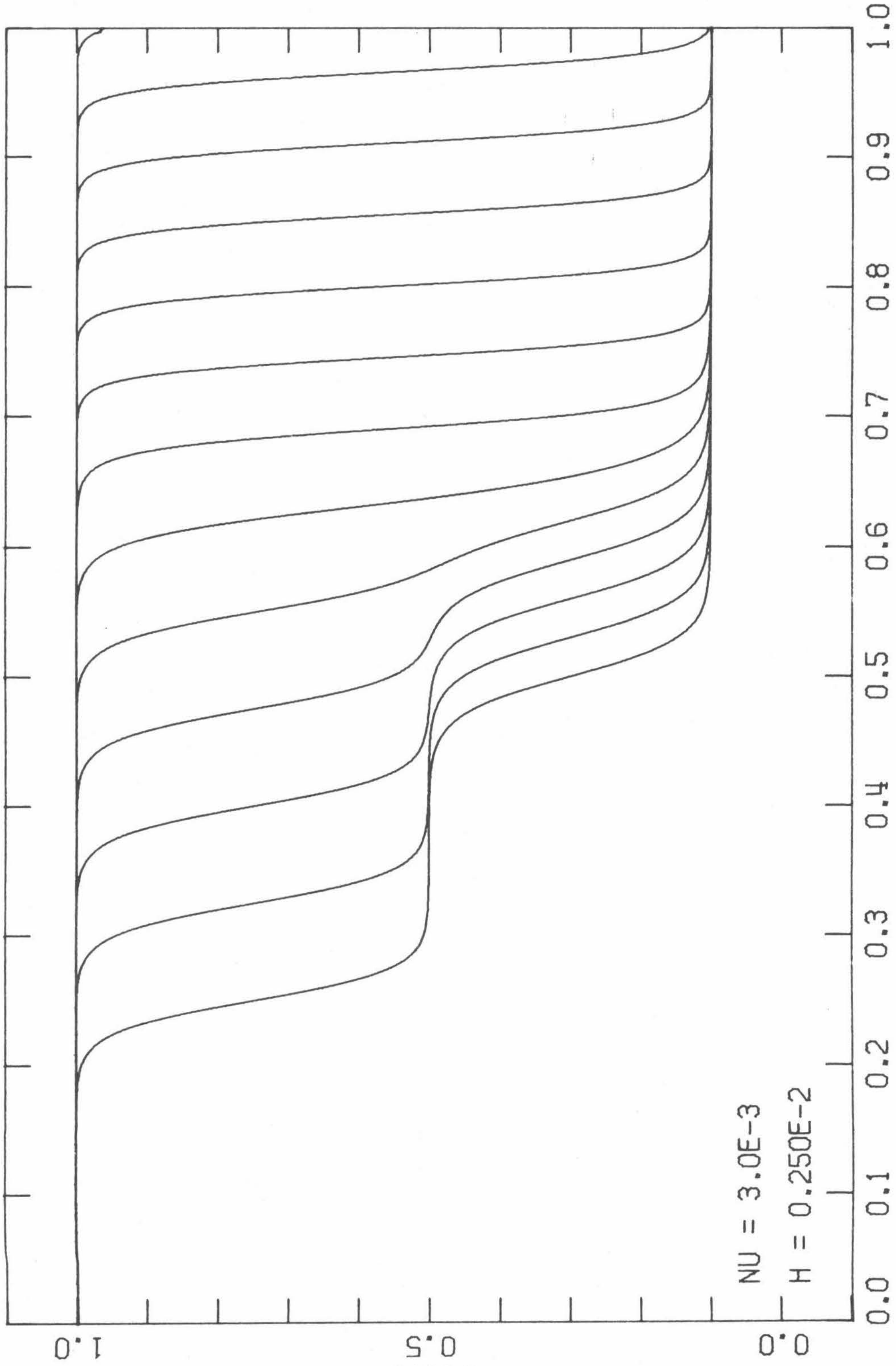


Figure 21

REFERENCES

Cole, Julian D. [1951]: "On a Quasi-Linear Parabolic Equation Occurring in Aerodynamics." Quarterly of Applied Mathematics, Vol. 9, No. 3, pp. 225-236.

Isaacson, Eugene, and Herbert Bishop Keller [1966]: Analysis of Numerical Methods. John Wiley & Sons, Inc., New York.

Keller, Herbert B. [1971]: "A New Difference Scheme for Parabolic Problems." Numerical Solution of Partial Differential Equations -- II: Synspade 1970. Edited by Bert Hubbard. Academic Press, New York, pp. 327-350.

Lees, Milton [1960]: "Energy Inequalities for the Solution of Differential Equations." American Mathematical Society Transactions, Vol. 94, pp. 58-73.

Lees, Milton [1961]: "Energy Inequalities for Partial Difference Equations." Lecture notes, Mathematics Department, California Institute of Technology, Pasadena. (Mimeographed.)

Strang, W. Gilbert [1960]: "Difference Methods for Mixed Boundary-Value Problems." Duke Mathematical Journal, Vol. 27, No. 2, pp. 221-231.

Swartz, Blair, and Burton Wendroff [1969]: "Generalized Finite-Difference Schemes." Mathematics of Computation, Vol. 23, No. 105, pp. 37-49.

Varah, James M. [1971]: "Stability of Difference Approximations to the Mixed Initial Boundary Value Problems for Parabolic Systems." SIAM Journal of Numerical Analysis, Vol. 8, No. 3, pp. 598-615.

Varah, James M. [1972]: On the Solution of Block Tridiagonal Systems Arising from Certain Finite-Difference Equations. Technical Report No. 72-02, Department of Computer Science, University of British Columbia, Vancouver.

Whitham, Gerald B. [1972]: private communication.

Yanenko, N. N. [1971]: The Method of Fractional Steps: The Solution of Problems of Mathematical Physics in Several Variables. Translated by Northrop Corporate Laboratories. Translation edited by Maurice Holt. Springer-Verlag, New York.