

Numerical Solution of a PDE System Describing a Catalytic Converter

BJÖRN ENGQUIST, BERTIL GUSTAFSSON, AND JOOP VREEBURG

Department of Computer Sciences, Uppsala University, Sweden

Received March 1, 1977

Numerical approximations are studied for a large hyperbolic system coupled to a parabolic equation and a system of algebraic equations. The equations, which all are nonlinear, describe nonviscous compressible one-dimensional gas flow in a catalytic converter. Chemical reactions within the gas are included in the model. Well-posedness of the partial differential equations is analyzed together with stability of the numerical models. In particular an investigation is made of the effect of numerical dissipation and different boundary conditions. Numerical results are presented.

1. INTRODUCTION

In this paper difference methods for solving a system of time dependent partial differential equations are studied. A thorough analysis is presented for a class of systems with application to a specific problem for a catalytic converter. The analysis and the numerical experiments include some new results, of both a practical and a theoretical nature.

The mathematical model describes nonviscous compressible gas flow in one spacial dimension. The gas consists of several components, which react chemically with each other. The interaction between the temperature of the gas and the surrounding material is contained in the model. Mathematically this is a system of nonlinear partial differential equations coupled to a system of algebraic equations. The differential equations consist of a large hyperbolic system and a parabolic equation. The upper part of the hyperbolic system (the first three equations) contains a general fluid dynamic model [1, Chap. I], and appears in a large number of applications. The lower part governs the concentrations of the different gas components and has a principal part on diagonal form. The parabolic equation describes the heat conduction in the surrounding material.

The hyperbolicity is shown in Section 2. This property is not obvious since there

* This work was sponsored by the Swedish Natural Science Research Council (NFR: F2711-014), The Swedish Board of Technical Development (STU: 72-271/U203), and The Institute for Applied Mathematics (ITM).

are multiple characteristics in the linearized system. Several ways of stating the boundary conditions to get a well-posed problem are demonstrated. From this analysis one can also exclude boundary conditions which might look physically reasonable.

The hyperbolic system is approximated by the Leap-Frog difference scheme [6, Chap. 5] with a fourth-order dissipative term [5, Chap. 9]. The Du Fort–Frankel scheme [6, Chap. 7] is used for the parabolic equation. Both schemes are explicit and of second-order accuracy. The nonlinear algebraic equations are solved by a modified version of Newton's method.

The stability analysis presented is relevant for many other applications. In particular, the importance of formulating the numerical boundary conditions is pointed out. For the Leap-Frog scheme it is necessary to give more boundary conditions than for the differential equations. A number of different conditions are analyzed using the theory in [4]. In order to get a sufficient number of conditions at the boundary we use special difference approximations of the differential equations. It is interesting to note that for the lower diagonal part of the hyperbolic equations the time derivative must be approximated by forward time differences, while for the upper part centered time differences should be used. The natural centered difference approximations at the boundaries for the parabolic equation are proved to be stable in the sense of Varah [8].

Different ways of speeding up the scheme for special applications are presented. The final version is 20 times faster than the original one. This effect is obtained mainly by using different time steps for different differential equations and by approximating the original functions, contained in the system, by simpler ones.

In certain cases the lower order terms in the differential equations make the system stiff. The difference scheme must then be modified such that these terms are treated implicitly. This technique is applied to the faster version and numerical results are also given.

In Section 4 we give results of numerical experiments for different boundary conditions, dissipation coefficients, and step sizes, e.g., the necessity of a special dissipation term at the boundary is shown.

The methods described here gives an accurate representation even of fast moving phenomena like sound waves. In some cases the solution is wanted for longer periods of time t , and then the method is too slow.

One way to increase the speed is to use a fully implicit method without any stability restriction on the time step. In this case it is not possible to describe the solution in full detail, i.e., fast varying components like sound waves cannot be well represented. However, these components can be small and the solution might be reasonable anyway.

Another way to speed up the computation is to change the mathematical model such that the fast varying components do not appear. This can be done by setting some time derivatives equal to zero. Further work along these lines is under way.

2. THE DIFFERENTIAL EQUATIONS

The model consists of a hyperbolic system with $N + 3$ equations where N is the number of different components of the gas mixture. To this system are coupled one parabolic differential equation and N nonlinear algebraic equations. (In our application the number N equals 9.)

We list here the physical significance of the independent and dependent variables:

- t = time $t \geq 0$;
- x = space variable $0 \leq x \leq L$;
- ρ = density of the gas mixture;
- w = flow speed of the gas mixture;
- T = temperature of the gas mixture;
- $Y_i = \rho_i/\rho$ where ρ_i is the density of the i th component of the gas;
- T_s = temperature of the solid;
- ρ_{is} = density of the gas components in the solid.

The mathematical model is

$$\begin{aligned}
 \epsilon \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho w) &= 0, \\
 \rho \left(\epsilon \frac{\partial w}{\partial t} + w \frac{\partial w}{\partial x} \right) &= -\epsilon^2 \frac{\partial p}{\partial x} + f_2, \\
 \rho \left(\epsilon \frac{\partial}{\partial t}(C_p T) + w \frac{\partial}{\partial x}(C_p T) \right) - \left(\epsilon \frac{\partial p}{\partial t} + w \frac{\partial p}{\partial x} \right) &= f_3, \\
 \rho \left(\epsilon \frac{\partial Y_i}{\partial t} + w \frac{\partial Y_i}{\partial x} \right) &= f_{i+3}, \quad i = 1, \dots, N, \\
 \frac{\partial T_s}{\partial t} &= a \frac{\partial^2 T_s}{\partial x^2} + f_{N+4}, \\
 g_i &= 0, \quad i = 1, \dots, N.
 \end{aligned} \tag{2.1}$$

Here $p = R\rho T \sum_{i=1}^N (Y_i/W_i)$; R and W_i are constants; $C_p = C_p(T)$, $a = a(T_s)$; ϵ is a constant; f_i and g_i are functions of the dependent variables.

The explicit form of the rather complicated functions C_p , a , f_i , and g_i which are used in our application is given in [3, Sect. 5].

To make the analysis applicable to more general systems, we rewrite the equations in the following form:

$$u_i = A_1 u_x + B y_x + F_1; \tag{2.2a}$$

$$y_i = A_2 y_x + F_2; \tag{2.2b}$$

$$v_t = av_{xx} + F_3; \quad (2.2c)$$

$$G = 0, \quad (2.2d)$$

$$u = (\rho, w, T)^T, \quad y = (Y_1, \dots, Y_N)^T, \quad v = T_s,$$

$$F_1 = (f_1, f_2, f_3)^T, \quad F_2 = (f_4, \dots, f_{N+3})^T, \quad F_3 = f_{N+4}$$

(U^T denotes the transpose of a vector U),

$$A_1 = - \begin{pmatrix} \frac{w}{\epsilon} & \frac{\rho}{\epsilon} & 0 \\ \frac{\epsilon p}{p^2} & \frac{w}{\epsilon} & \frac{\epsilon p}{T\rho} \\ 0 & \frac{p}{\epsilon\beta\rho} & \frac{w}{\epsilon} \end{pmatrix} = \begin{pmatrix} d & a_{12} & 0 \\ a_{21} & d & a_{23} \\ 0 & a_{32} & d \end{pmatrix},$$

$$B = - \begin{pmatrix} 0 & \dots & 0 \\ \frac{\epsilon RT}{W_1} & \dots & \frac{\epsilon RT}{W_N} \\ 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} 0 & \dots & 0 \\ b_1 & \dots & b_N \\ 0 & \dots & 0 \end{pmatrix},$$

$$A_2 = - \begin{pmatrix} \frac{w}{\epsilon} & & 0 \\ & \cdot & \\ & & \cdot \\ 0 & & \frac{w}{\epsilon} \end{pmatrix} = \begin{pmatrix} d & & 0 \\ & \cdot & \\ 0 & & d \end{pmatrix},$$

$$G = G(U, v, r) = (g_1, \dots, g_N)^T; \quad U = (u, y)^T; \quad r = (\rho_{1s}, \dots, \rho_{Ns})^T;$$

$$\beta = \beta(T, y) = T \frac{dC_p}{dT} + C_p - R \sum_{i=1}^N \frac{Y_i}{W_i}.$$

We will also use the following further simplified form:

$$U_t = AU_x + F;$$

$$v_t = av_{xx} + F_3; \quad (2.3)$$

$$G = 0;$$

$$U = \begin{pmatrix} u \\ y \end{pmatrix}; \quad F = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix}; \quad A = \begin{pmatrix} A_1 & B \\ 0 & A_2 \end{pmatrix}.$$

The values of U and v for $t = 0$, $0 \leq x \leq L$ are given as initial conditions.

It is not physically obvious which types of boundary conditions can be given at $x = 0$ and $x = L$ ($t > 0$). We will investigate which conditions lead to well-posed problems.

2.1. *Well-posedness of the Equations*

We mean by well-posedness that the L_2 -norm¹ of the solution of the linearized version of the equations can be estimated by the inhomogeneous terms in the differential equations and in the initial and boundary conditions. Since small perturbations of the solution, e.g., errors in the calculation, are governed by these linear equations (the variation equations), it is essential that their solution not grow too rapidly.

The unknown r can first be eliminated from the system by solving the equations $G = 0$.

We linearize the equations, neglect lower order terms, and regard A and a as constant. That is, we will study the system:

$$U_t = AU_x, \tag{2.4a}$$

$$v_t = av_{xx}. \tag{2.4b}$$

The transformations we will introduce later can also be used with the energy integral method for variable coefficients.

Let us consider Eq. (2.4a) The first question is whether the system is hyperbolic. We must check if A has real eigenvalues.

A_1 has the three eigenvalues

$$\begin{aligned} \lambda_1 &= d && (\approx -10 \text{ in our application}), \\ \lambda_2 &= d - (a_{12}a_{21} + a_{23}a_{32})^{1/2} && (\approx -400 \text{ in our application}), \\ \lambda_3 &= d + (a_{12}a_{21} + a_{23}a_{32})^{1/2} && (\approx 400 \text{ in our application}), \end{aligned}$$

and A_2 has N eigenvalues equal to d .

By definition a_{ij} all are negative, hence the eigenvalues are real. We want, however, strong hyperbolicity; i.e., that lower order terms cannot destroy the well-posedness (see [7]). In our case some eigenvalues are multiple. We need therefore to construct a symmetrizer S for the matrix A such that:

$$(SAS^{-1})^T = SAS^{-1}.$$

Let S have the partitioned form

$$\begin{aligned} S &= \begin{pmatrix} D & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} I & Q \\ 0 & I \end{pmatrix} \quad (I \text{ is the identity matrix}), \\ SAS^{-1} &= \begin{pmatrix} DA_1D^{-1} & D(B + QA_2 - A_1Q) \\ 0 & A_2 \end{pmatrix}. \end{aligned}$$

Choose the 3×3 -matrix $D = \text{diag}(d_1, 1, d_3)$ such that $DA_1D^{-1} = \tilde{A}_1$ is symmetric.

¹ The L_2 -norm of the vector valued function U is $\|U\| = (\int_0^L U(x, t)^T U(x, t) dx)^{1/2}$.

We get

$$\tilde{A}_1 = \begin{pmatrix} d & d_1 a_{12} & 0 \\ a_{21} d_1^{-1} & d & a_{23} d_3^{-1} \\ 0 & d_3 a_{32} & d_1 \end{pmatrix} = \begin{pmatrix} d & a_2 & 0 \\ a_2 & d & a_3 \\ 0 & a_3 & d \end{pmatrix}$$

where $a_2 = -(a_{12} a_{21})^{1/2}$ and $a_3 = -(a_{23} a_{32})^{1/2}$ if $d_1 = (a_{21}/a_{12})^{1/2}$ and $d_3 = (a_{23}/a_{32})^{1/2}$.

We further want $D(B + QA_2 - A_1Q) = 0$, i.e.,

$$B + QA_2 - A_1Q = B + (dI - A_1)Q = 0.$$

The matrix $(dI - A_1)$ is not invertible but despite this we can find a suitable Q since B is in the column space of A_1 ,

$$\begin{pmatrix} 0 & a_{12} & 0 \\ a_{21} & 0 & a_{23} \\ 0 & a_{32} & 0 \end{pmatrix} Q = \begin{pmatrix} 0 & \cdots & 0 \\ b_1 & \cdots & b_N \\ 0 & \cdots & 0 \end{pmatrix},$$

$$Q = \begin{pmatrix} q_{11} & \cdots & q_{1N} \\ 0 & \cdots & 0 \\ 0 & \cdots & 0 \end{pmatrix} \quad \text{where} \quad q_{1i} = \frac{b_i}{a_{21}}.$$

These calculations tell us that the system (2.4a) can be symmetrized and therefore it is strongly well posed.

Equation (2.4b) is of standard parabolic type in our application since the coefficient a is positive. Hence it is also well posed as an initial value problem.

2.2. Well-posedness of the Boundary Conditions

Let us first study the hyperbolic part. To analyze its boundary conditions we go a bit further and diagonalize the symmetric \tilde{A}_1 with the transformation $M\tilde{A}_1M^{-1}$:

$$\frac{a_2 a_3}{2c^2} \begin{pmatrix} 2 \frac{a_3}{a_2} & 0 & -2 \\ -\frac{c}{a_3} & \frac{c^2}{a_2 a_3} & -\frac{c}{a_2} \\ \frac{c}{a_3} & \frac{c^2}{a_2 a_3} & \frac{c}{a_2} \end{pmatrix} \begin{pmatrix} d & a_2 & 0 \\ a_2 & d & a_3 \\ 0 & a_3 & d \end{pmatrix}$$

$$= \begin{pmatrix} 1 & -\frac{a_2}{c} & \frac{a_2}{c} \\ 0 & 1 & 1 \\ -\frac{a_2}{a_3} & -\frac{a_3}{c} & \frac{a_3}{c} \end{pmatrix} = \begin{pmatrix} d & 0 & 0 \\ 0 & d - c & 0 \\ 0 & 0 & d + c \end{pmatrix}$$

where $c = (a_2^2 + a_3^2)^{1/2}$.

The characteristic quantities in the hyperbolic system A_1, \dots, A_{N+3} , i.e., the dependent variables in the diagonalized system, are related to ρ, w, T , and Y_i in the following way:

$$\begin{pmatrix} u \\ y \end{pmatrix} = S^{-1} \begin{pmatrix} M^{-1} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} A_I \\ A_{II} \end{pmatrix}$$

where

$$A_I = (A_1, A_2, A_3)^T, \quad A_{II} = (A_4, \dots, A_{N+3})^T,$$

$$\rho = p_1 A_1 + p_2 A_2 - p_2 A_3 - \sum_{i=1}^N q_{1i} A_{i+3},$$

$$w = A_2 + A_3,$$

$$T = p_3 A_1 + p_4 A_2 - p_4 A_3,$$

$$Y_i = A_{i+3}, \quad i = 1, \dots, N,$$

$$p_1 = \frac{1}{d_1}, \quad p_2 = -\frac{a_2}{d_1 c}, \quad p_3 = -\frac{a_2}{d_3 a_3}, \quad p_4 = -\frac{a_3}{d_3 c}.$$

Since the problem is subsonic, λ_3 is positive. Therefore, the boundary conditions for the diagonalized system that give well-posed problems are of the following well-known form, see, e.g., [5, p 65] or [4]. For $x = 0, t \geq 0$

$$\begin{aligned} A_1 &= s_{13} A_3 + h_1, \\ A_2 &= s_{23} A_3 + h_2, \\ A_i &= s_{i3} A_3 + h_i, \quad i = 4, \dots, N. \end{aligned} \tag{2.5}$$

For $x = L, t \geq 0$

$$A_3 = s_{31} A_1 + s_{32} A_2 + \sum_{i=4}^N s_{3i} A_i + h_3. \tag{2.6}$$

Here s_{ij} and h_i are functions of t . This means that the value of the components corresponding to characteristics that go into the region from the boundary are given. It is now easy to check whether a certain setting of boundary conditions for ρ, w, T , and Y_i after transformation have the form (2.5), (2.6).

Let us note two consequences. There must be $N + 2$ boundary conditions at $x = 0$ and one at $x = L$. For example, it is not possible to give values to all dependent variables at the inflow boundary $x = 0$. The other consequence is that $Y_i (= A_{i+3})$ should be given at $x = 0$.

At the outflow boundary we specify $p = R\rho T \sum_{i=1}^N Y_i/W_i$ and solve for T or ρ . That is

$$T = p / \left(R\rho \sum_{i=1}^N Y_i/W_i \right) \tag{2.7a}$$

or

$$\rho = p / \left(RT \sum_{i=1}^N Y_i / W_i \right). \quad (2.7b)$$

For the analysis we study the linearized form

$$T = t_0 + t_1 \rho + \sum_{i=1}^N t_{i+1} Y_i \quad (2.8a)$$

or

$$\rho = r_0 + r_1 T + \sum_{i=1}^N r_{i+1} Y_i. \quad (2.8b)$$

Here t_i and r_i are regarded as constants.

For the appropriate values of the constants in our application the following combinations give rise to well-posed problems: Two out of ρ , w , T are given at $x = 0$; T or ρ are given at $x = L$, and all Y_i at $x = 0$.

We check the conditions; ρ and T specified at $x = 0$, T specified at $x = L$. (This combination of boundary conditions is denoted $(\rho, T; T)$ in what follows. Other combinations are denoted analogously.) For $x = 0$ we have $\rho = \rho_b$, $T = T_b$, $Y_i = Y_{ib}$, $i = 1, \dots, N$; i.e.,

$$\begin{aligned} p_1 A_1 + p_2 A_2 - p_2 A_3 - \sum_{i=1}^N q_{1i} A_{i+3} &= \rho_b, \\ p_3 A_1 + p_4 A_2 - p_4 A_3 &= T_b, \\ A_{i+3} &= Y_{ib}, \quad i = 1, \dots, N. \end{aligned}$$

If the equations are rearranged they will be of the type (2.5), since $p_1 p_4 - p_2 p_3$ is nonzero:

$$\begin{aligned} A_1 &= \left(p_4 \left(\left(\sum_{i=1}^N q_{1i} Y_{ib} \right) + \rho_b \right) - p_2 T_b \right) / (p_1 p_4 - p_2 p_3), \\ A_2 &= \left(p_3 \left(\left(\sum_{i=1}^N q_{1i} Y_{ib} \right) + \rho_b \right) - p_1 T_b \right) / (p_2 p_3 - p_1 p_4) + A_3, \\ A_{i+3} &= Y_{ib}, \quad i = 1, \dots, N. \end{aligned}$$

For $x = L$ we have $T = t_0 + t_1 \rho + \sum_{i=1}^N t_{i+1} Y_i$ or

$$\begin{aligned} A_3 &= \left(t_0 + (t_1 p_1 - p_3) A_1 + (t_1 p_2 - p_4) A_2 \right. \\ &\quad \left. + \sum_{i=1}^N (t_{i+1} - t_1 q_{1i}) A_{i+3} \right) / (t_1 p_2 - p_4) \end{aligned}$$

where $t_1 p_2 - p_4 \neq 0$.

Let us finally consider the parabolic equation (2.4b). The boundary conditions for that equation are:

$$a(v(0, t)) \frac{\partial v(0, t)}{\partial x} = C_1(v(0, t) - T(0, t)),$$

$$a(v(L, t)) \frac{\partial v(L, t)}{\partial x} = C_2(v(L, t) - T(L, t)),$$

where C_1 and C_2 are constants.

Since the boundary conditions for (2.4a) do not contain v we can solve for u and consider it as known when solving for v . Hence $T(0, t)$ and $T(L, t)$ are known in (2.9). The linearized form of (2.9) is of a well-known type which gives well-posed problems.

3. THE NUMERICAL METHODS

For a problem like this, which has only one space variable but consists of a large number of nonlinear and fairly complicated equations, a difference method is the most natural choice for the numerical approximation. Since the accuracy requirements are low and some of the data have low precision, a method of higher order than two is not appropriate. A dissipative scheme is necessary because the system is nonlinear. If those effects on the solution caused by the large eigenvalues $d \pm (a_2^2 + a_3^2)^{1/2}$ of A_1 cannot be neglected, an explicit scheme should be used. The size of the time step must in any case be chosen $\approx (d + (a_2^2 + a_3^2)^{1/2})^{-1} h$, where h is the space step.

According to these considerations the Leap-Frog scheme with a fourth-order dissipation term is reasonable for the hyperbolic part of the system, and the Du Fort-Frankel scheme for the parabolic part.

3.1. Definition of the Basic Method

A mesh is defined by $x_j = jh, j = 0, 1, \dots, J, h = L/J$, and $t^n = nk, n = 0, 1, \dots; k$ being the time step. Grid functions are then defined at these grid points, e.g., $U_j^n = U(x_j, t^n)$. In order to be able to use centered difference operators when approximating (2.9), v_j^n is defined also for $j = -1, J + 1$.

Using the form (2.3) of the differential equations, the approximation is defined by

$$U_j^{n+1} = 2kA(U_j^n) D_0 D_j^n + (I - \delta_1 h^4 (D_+ D_-)^2) U_j^{n-1} + 2kF(U_j^n, v_j^n, r_j^n), \quad j = 2, 3, \dots, J - 2; \tag{3.1a}$$

$$(I + 2ka(v_j^n)/h^2) v_j^{n+1} = 2ka(v_j^n)(v_{j+1}^n - v_j^{n-1} + v_{j-1}^n)/h^2 + v_j^{n-1} + 2kF_3(U_j^n, v_j^n, r_j^n), \quad j = 0, 1, \dots, J; \tag{3.1b}$$

$$G(U_j^{n+1}, v_j^{n+1}, r_j^{n+1}) = 0, \quad j = 0, 1, \dots, J; \tag{3.1c}$$

where

$$\begin{aligned} D_+ U_j^n &= (U_{j+1}^n - U_j^n)/h, \\ D_- U_j^n &= (U_j^n - U_{j-1}^n)/h, \\ D_0 U_j^n &= (U_{j+1}^n - U_{j-1}^n)/2h. \end{aligned}$$

δ_1 is a parameter, and we will see later how it can be chosen. Equations (3.1) must be completed with boundary conditions. At the points x_1, x_{J-1} (3.1a) is changed so that a second-order dissipation term is substituted for the fourth-order one:

$$\begin{aligned} U_j^{n+1} &= 2kA(U_j^n) D_0 U_j^n + (I + \delta_2 h^2 D_+ D_-) U_j^{n-1} \\ &+ 2kF(U_j^n, v_j^n, r_j^n), \quad j = 1, J-1. \end{aligned} \quad (3.2)$$

The way of defining U_0^{n+1}, U_J^{n+1} is of course dependent on the boundary conditions for the differential equation, and as we have seen in Section 2 there are several ways in which those can be stated. Those variables, for which values are explicitly given at the boundary, are defined in the same way for the difference scheme. For the variables corresponding to the diagonal part (2.2b) of the system, one-sided difference operators are used both in space and time at the right boundary when $w > 0$

$$y_j^{n+1} = (I + kA_2(U_j^n) D_-) y_j^n + kF_2(U_j^n, v_j^n, r_j^n). \quad (3.3)$$

The missing conditions for the variables corresponding to part (2.2a) of the system are defined analogously but with centered time differences. The reason for this is discussed later.

v_{-1}^n, v_{J+1}^n are determined by

$$a(v_0^n) D_0 v_0^n = C_1(v_0^n - T_0^n), \quad (3.4a)$$

$$a(v_J^n) D_0 v_J^n = C_1(v_J^n - T_J^n). \quad (3.4b)$$

The solution $r_j^{n+1}, j = 0, 1, \dots, J$, to (3.1c) is obtained by a simplified version of Newton's method ($\lim_{k \rightarrow \infty} r^{(k)} = r^{n+1}$):

$$\begin{aligned} r_j^{(k+1)} &= r_j^{(k)} - \text{Diag} \left[\frac{\partial G}{\partial r_j} \Big|_{r_j=r_j^{(l)}} \right]^{-1} G(r_j^{(k)}), \\ r_j^{(0)} &= r_j^n, \quad l = [n/K]K; \quad k = 0, 1, \dots; \quad j = 0, 1, \dots, J. \end{aligned} \quad (3.5)$$

$\text{Diag}[A]$ denotes the matrix where the off-diagonal elements of A are set equal to zero; $[x]$ denotes "integer part of x ," i.e., the iteration matrices are updated every K th time step. The reason for ignoring the nondiagonal elements is that for our application they are small compared to the diagonal ones.

Values at the first time level $t = k$ are obtained by

$$U_j^1 = (I + kA(U_j^0) D_0) U_j^0 + kF(U_j^0, v_j^0, r_j^0), \quad j = 1, 2, \dots, J - 1; \quad (3.6a)$$

$$v_j^1 = (I + ka(v_j^0) D_+ D_-) v_j^0 + kF_3(U_j^0, v_j^0, r_j^0), \quad j = 0, 1, \dots, J. \quad (3.6b)$$

The boundary values are obtained as described above, except that one-sided time differences are used instead of centered ones.

3.2. Accuracy of the Method

The formal accuracy of the scheme (3.1a) is second order, i.e., for smooth solutions $U(x, t)$ the global error is $\mathcal{O}(k^2 + h^2)$. The consistency of the Du Fort–Frankel scheme (3.1b) requires that k tends to zero faster than h does, since the error of the principal part $\partial v/\partial t - a\partial^2 v/\partial x^2 = 0$ of the equation is

$$\frac{k^2}{6} u_{ttt} - a \frac{h^2}{12} v_{xxxx} + a \frac{k^2}{h^2} v_{tt} + \mathcal{O}\left(\frac{k^4}{h^2}\right).$$

As we will see later in this section the stability analysis of the hyperbolic part allows $k/h = \text{const}$. Formally this contradicts the consistency requirement above. However, in our application this is no limitation, since the coefficient a is very small. The spectral radius of A , $(\rho(A))$ in (3.1a) is of the order 400, and a of the order 10^{-6} , which makes the critical term in the error negligible.

As mentioned above, the modified Newton’s method (3.6) is motivated by the strong diagonal dominance of $(\partial G/\partial r_j)^{-1}$. For our application the magnitude is of an order 10^4 times larger for the diagonal elements than for the off-diagonal ones. Comparisons made with the full Newton method show a difference only in the fifth digit. Two iterations with (3.5) gave a sufficient accuracy.

3.3. Stability of the Method

The stability analysis of the approximation is carried out for the linearized system with constant coefficients (as for the differential equations in Section 2). A straightforward calculation shows that the von Neumann condition for (3.1a) is fulfilled if

$$\delta_1 < \frac{1}{8} \quad (3.7a)$$

and

$$\frac{k}{h} \rho(A) \leq \min_{0 \leq \theta \leq 1} \frac{1 - 8\delta_1 \Theta^4}{2\Theta(1 - \Theta^2)^{1/2}}. \quad (3.7b)$$

Equation (3.7b) is satisfied if, for example,

$$\frac{k}{h} \rho(A) \leq 1 - 8\delta_1. \quad (3.7c)$$

From (3.7a,b) we can also derive the sufficient conditions

$$\delta_1 < \frac{1}{16}, \tag{3.7d}$$

$$\frac{k^2}{h^2} [\rho(A)]^2 \leq \frac{1 + (1 - 16\delta_1)^{1/2}}{2}. \tag{3.7e}$$

See also [5].

The Du Fort–Frankel scheme (3.1b) is unconditionally stable for the Cauchy problem which is shown, e.g., in [6, Section 7.5].

When including also the boundary conditions into the stability analysis, we make use of the theory in [4] and [8]. We first prove that the parabolic part is stable. Dropping lower order terms and considering the quarter-plane problem $0 \leq x < \infty$, $t \geq 0$, we study the solutions to $v_t = av_{xx}$ with boundary condition $v_x(0, t) = g(t)$. The resolvent equation corresponding to the Du Fort–Frankel scheme is

$$z^2 \hat{v}_j - \frac{\sigma z}{1 + \sigma} (\hat{v}_{j+1} + \hat{v}_{j-1}) - \frac{1 - \sigma}{1 + \sigma} \hat{v}_j = 0, \quad j = 0, 1, \dots \tag{3.8a}$$

where $\sigma = 2ak/h^2$. The boundary condition is

$$\hat{v}_1 - \hat{v}_{-1} = 2hg. \tag{3.8b}$$

The solution to (3.8a) which lies in $l_2(0, \infty)$ for $|z| > 1$, can be written $\hat{v}_j = \lambda \kappa^j$ where κ satisfies

$$z^2 \kappa - \frac{\sigma z}{1 + \sigma} (\kappa^2 + 1) - \frac{1 - \sigma}{1 + \sigma} \kappa = 0, \quad |\kappa| < 1 \text{ for } |z| > 1. \tag{3.9}$$

Looking for eigenvalues, we put $g = 0$ in (3.8b) and get the condition $\kappa^2 - 1 = 0$, i.e., $\kappa = \pm 1$ for the existence of a nontrivial solution. Under the condition $\kappa = 1$, $|z| \geq 1$, (3.9) implies $z = 1$, and for this z value $\kappa = 1$ is a double root of (3.9). Therefore

$$\kappa = 1 - c(z - 1)^{1/2} + \mathcal{O}(|z - 1|), \quad c \neq 0,$$

and we get from Eq. (3.8b):

$$\lambda(\kappa^2 - 1) = 2\kappa hg$$

which implies

$$|\lambda| \leq \text{const} \frac{h |g|}{|z - 1|^{1/2}}, \quad |z| > 1. \tag{3.10}$$

This is the stability condition of Varah [8].

The other critical point is $\kappa = -1, z = -1$. This is not covered by Varah's theory. However, stability proofs can be carried out also in this case if the condition (3.10) is changed to

$$|\lambda| \leq \frac{h|g|}{|z+1|^{1/2}} \quad \text{for all } z \rightarrow -1, \quad |z| > 1. \quad (3.11)$$

This estimate is obtained in precisely the same way as (3.10) was obtained for $\kappa = 1$; $\kappa = -1$ is, therefore, a double root of (3.9) for $z = -1$.

The h factor in (3.10), (3.11) is lost at the error estimate, but since the truncation error (corresponding to g) is of the order h^2 , we obtain a second-order convergence rate.

A stability analysis has not been carried out for the upper part (2.2a) of the hyperbolic system. The motivation for choosing the boundary conditions as described in Section 3.1 is the relation between the system (2.4a) and

$$U_t = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} U_x. \quad (3.12)$$

This relation is seen when carrying out the diagonalization of A in Section 2. The boundary conditions using centered time differences can be shown to be stable for (3.12), but not the ones using forward time differences. Numerical results confirm these arguments.

For the lower diagonal part (2.2b) we have made an investigation with regard to generalized eigenvalues, and that analysis will be presented here since it might be of interest also for other applications.

We consider the system $y_t = A_2 y_x$ where $A_2 = dI, d < 0$. The resolvent equation corresponding to the difference approximation for each scalar equation is

$$(z^2 - 1) \hat{y}_j + z\tau(\hat{y}_{j+1} - \hat{y}_{j-1}) + \delta_1(\hat{y}_{j+2} - 4\hat{y}_{j+1} + 6\hat{y}_j - 4\hat{y}_{j-1} + \hat{y}_{j-2}) = 0, \quad j = 2, 3, \dots \quad (3.13a)$$

where $\tau = -kd/h > 0, \delta_1 \geq 0$.

Considering first the boundary $x = 0$, we have the boundary conditions

$$\hat{y}_0 = g_0, \quad (3.13b)$$

$$(z^2 - 1) \hat{y}_1 + z\tau(\hat{y}_2 - \hat{y}_0) - \delta_2(\hat{y}_2 - 2\hat{y}_1 + \hat{y}_0) = g_1, \quad \delta_2 \geq 0. \quad (3.13c)$$

The solution to (3.13a) can be written $\hat{y}_j = \lambda_1 \kappa_1^j + \lambda_2 \kappa_2^j$ for $\kappa_1 \neq \kappa_2$, where κ_1, κ_2 are roots to

$$(z^2 - 1) \kappa^2 + z\tau\kappa(\kappa^2 - 1) + \delta_1(\kappa - 1)^4 = 0, \quad |\kappa| < 1 \text{ for } |z| > 1. \quad (3.14)$$

Again we put $g_0 = g_1 = 0$ and obtain the boundary conditions

$$\begin{aligned} \lambda_1 + \lambda_2 &= 0, \\ \{(z^2 - 1)\kappa_1 + z\tau(\kappa_1^2 - 1) - \delta_2(\kappa_1 - 1)^2\}\lambda_1 + \{(z^2 - 1)\kappa_2 + z\tau(\kappa_2^2 - 1) \\ &\quad - \delta_2(\kappa_2 - 1)^2\}\lambda_2 = 0. \end{aligned} \quad (3.15)$$

Looking for generalized eigenvalues, i.e., $|\kappa_1| = 1$, $|z| = 1$, it is sufficient to investigate the case $\kappa_1 = 1$ since the scheme is dissipative.

Corresponding to $\kappa_1 = 1$ there are two z values; $z = \pm 1$. Under the condition $|\kappa_1(z)| < 1$ for $|z| > 1$ the only z value is $z = -1$, and therefore the condition for a nontrivial solution is

$$\tau(\kappa_2^2 - 1) + \delta_2(\kappa_2 - 1)^2 = 0, \quad |\kappa_2| < 1 \quad \text{for } |z| > 1,$$

i.e.,

$$\kappa_2 = \frac{\delta_2 - \tau}{\delta_2 + \tau}.$$

We put $z = -1$ in (3.14), divide by $(\kappa - 1)$, and put $\kappa = \kappa_2$. This gives

$$-\tau\kappa_2(\kappa_2 + 1) + \delta_1(\kappa_2 - 1)^3 \equiv \frac{-2\tau\delta_2(\delta_2 - \tau)}{(\delta_2 + \tau)^2} - \frac{8\delta_1\tau^3}{(\delta_2 + \tau)^3} = 0$$

which is equivalent to the condition

$$\begin{aligned} \tau^2 &= \frac{\delta_2^3}{\delta_2 - 4\delta_1} \quad \text{for } \delta_2 \neq 4\delta_1, \\ \delta_1 &= \delta_2 = 0 \quad \text{if } \delta_2 = 4\delta_1. \end{aligned} \quad (3.16)$$

Since τ is arbitrary in the interval $0 < \tau < 1$ we must choose δ_2 such that one of the conditions

$$\delta_2 \leq 4\delta_1, \quad (3.17a)$$

$$\delta_2^3 \geq \delta_2 - 4\delta_1 \quad (3.17b)$$

is fulfilled.

($\delta_1 = \delta_2 = 0$ gives the usual Leap-Frog scheme which is stable, [4].)

We now check for multiple roots κ_1 , i.e., a solution with the form $y_j = (\lambda_1 + j\lambda_2)\kappa_1^j$. The boundary conditions now imply

$$\begin{aligned} \lambda_1 &= 0, \\ \{(z^2 - 1)\kappa_1 + z\tau(\kappa_1^2 - 1) - \delta_2(\kappa_1 - 1)^2\}\lambda_1 + \{(z^2 - 1)\kappa_1 + z\tau \cdot 2\kappa_1^2, \\ &\quad - \delta_2(2\kappa_1^2 - 2\kappa_1)\}\lambda_2 = 0, \end{aligned} \quad (3.18)$$

and the condition for a nontrivial solution with $z = -1, \kappa_1 = 1$ is $\tau = 0$, which is a contradiction.

We have not checked the possibility of an eigenvalue $z = z_0, |z_0| = 1, |\kappa_1| < 1, |\kappa_2| < 1$ which is more difficult to do. However, in that case we permit an estimate $|\lambda_i| \leq |z - z_0|^{-1} (|g_0| + |g_1|)$ (cf. Theorem 10.3 in [4]). To violate this condition, there must be a z_0 such that the determinants in (3.15), (3.18) vanish for $z = z_0$, and furthermore the terms of order $|z - z_0|^{1/2}$ and $|z - z_0|$ must cancel. This is very unlikely but has not been checked. The case $|z_0| > 1$ has not been treated theoretically. However, this type of eigenvalue shows up as an instability after a few time steps and this has never occurred.

To simplify the notation when analyzing the boundary conditions at $x = L$, we study the same quarter-plane problem as above, but with $d > 0$. No boundary conditions are then given for the differential equation at $x = 0$. The condition for the $\lambda_i : s$ now will be

$$\begin{aligned} & \{z - 1 + \tau(\kappa_1 - 1)\} \lambda_1 + \{z - 1 + \tau(\kappa_2 - 1)\} \lambda_2 = 0, \\ & \{(z^2 - 1) \kappa_1 + z\tau(\kappa_1^2 - 1) - \delta_2(\kappa_1 - 1)^2\} \lambda_1 + \{(z^2 - 1) \kappa_2 + z\tau(\kappa_2^2 - 1) \\ & \quad - \delta_2(\kappa_2 - 1)^2\} \lambda_2 = 0. \end{aligned} \tag{3.19}$$

Since $z = -1$ for $\kappa_1 = 1$ we arrive at exactly the same condition (3.16) as in the previous case, τ now lying in the interval $-1 < \tau < 0$. This is also true for the case $\kappa_1 = \kappa_2$, and we get no new restriction.

It should be noted here that with the boundary condition

$$y_N^{n+1} = 2kd D_- y_N^n - y_N^{n-1} \tag{3.20}$$

the coefficient for λ_1 in the first equation of (3.19) will be $z^2 - 1 + 2\tau z(\kappa_1 - 1)$, which vanishes for $z = -1, \kappa_1 = 1$. Therefore we will have an instability for all $\delta_1 \geq 0, \delta_2 \geq 0$.

3.4. The Use of Different Time Steps for Different Equations

Due to the complexity of the system, in particular the lower order terms, a large number of arithmetic operations must be carried out for each point and every time step. For stability reasons the time step must be chosen very small, since $\rho(A)$ in Eq. (3.7b) is large. This causes the computing time to be very long. In this section we will describe a faster version of the program, where different time steps are used for different equations in the system. We have also approximated parts of the functions f_i by piecewise linear functions.

The matrix A of the hyperbolic part of the linearized system (2.3) has the form

$$A = \begin{pmatrix} A_1 & B \\ 0 & A_2 \end{pmatrix}.$$

In our application the spectral radius of A_2 is less than 3% of that of A_1 . Since the lower hyperbolic part (2.2b) does not contain u_x terms, it is therefore possible to use longer time steps for advancing y .

Let k_1 be the time step for u and $k_2 = Mk_1$ the time step for y and v , where M is a positive integer, With n denoting the time index in the coarser grid and $Mn + \nu$ the time index in the finer one, the numerical scheme takes the following form:

$$u^{Mn+\nu+1} = 2k_1A_1(u^{Mn+\nu}, \hat{y}^{Mn+\nu}) D_0u^{Mn+\nu} + B(u^{Mn+\nu}, \hat{y}^{Mn+\nu}) D_0\hat{y}^{Mn+\nu} \tag{3.21a}$$

$$+ (I - \delta_1h^4(D_+D_-)^2) u^{Mn+\nu-1} + 2k_1\tilde{F}_1^{Mn+\nu}, \quad \nu = 0, 1, \dots, M - 1;$$

$$y^{n+1} = 2k_2A_2(u^{Mn}, y^n) D_0y^n + (I - \delta_1h^4(D_+D_-)^2) y^{n-1} + 2k_2F_2^n; \tag{3.21b}$$

$$(I + 2k_2a(v_j^n)/h^2) v_j^{n+1} = 2k_2a(v_j^n)(v_{j+1}^n - v_j^{n-1} + v_{j-1}^n)/h^2 + v_j^{n-1} + 2k_2F_{3j}^n; \tag{3.21c}$$

$$G(u^{M(n+1)}, y^{n+1}, v^{n+1}, r^{n+1}) = 0, \quad n = 1, 2, \dots \tag{3.21d}$$

(The space index j is deleted in (3.21a, b, d)). The vector $\hat{y}^{Mn+\nu}$ denotes the interpolated value

$$\hat{y}^{Mn+\nu} = y^n + \nu(y^{n+1} - y^n)/M.$$

The vector valued function $\tilde{F}_1^{Mn+\nu}$ denotes the extrapolated value

$$\tilde{F}_1^{Mn+\nu} = \left(1 + \frac{\nu}{M}\right) F_1(u^{Mn}, y^n, v^n, r^n) - \frac{\nu}{M} F_1(u^{M(n-1)}, y^{n-1}, v^{n-1}, r^{n-1}).$$

Note that F_1 itself is extrapolated, not its arguments. This is of great importance for saving computing time, since now F_1 need not be evaluated for every small time step.

Equations (3.21) have all the same space step h , and the space index j runs over interior points as in (3.1). The necessary modifications in the boundary conditions are exactly analogous to the changes in the difference equations for the interior points.

The starting procedure is identical to (3.6) but with the use of different time steps. The first time steps in the inner loop for u ($n = 0, \nu = 1, 2, \dots, M - 1$) are treated as in (3.21a), but with \tilde{F}_1 defined by

$$\tilde{F}_1^\nu = F_1(u^0, y^0, v^0, r^0). \tag{3.22}$$

If M is regarded as a fixed number independent of h and k_1 , the scheme is still of second-order accuracy. The approximation (3.22) which is only first-order accurate, is used only $M - 1$ times and hence, it does not destroy the global accuracy.

As for the original method we disregard the lower order terms in the stability analysis. Then the lower part (3.21b, c) of the difference equations does not depend on

the vector u , and can therefore be treated separately. Equation (3.21c) is unconditionally stable, and the stability condition for (3.21b) will be

$$k_2^2(\rho(A_2))^2 \leq h^2(1 + (1 - 16\delta_1)^{1/2}). \quad (3.23)$$

We can now consider y and v as known functions in (3.21a). The additional stability criterion therefore is

$$k_1^2(\rho(A_1))^2 \leq h^2(1 + (1 - 16\delta_1)^{1/2}). \quad (3.24)$$

Of course the condition (3.7a) must also be fulfilled together with the conditions (3.17).

The truncation error for (3.21b, c) naturally grows with increasing M . For the practical choice $M \approx \rho(A_1)/\rho(A_2)$ this change in accuracy is negligible according to numerical experiments over short time intervals.

We have also approximated parts of the functions f_i by piecewise linear functions of T . Table 3.1 shows the significant gain in computing time with this technique. We cannot hope to get a more efficient scheme by a further optimization since the computing time is only twice as long as for the reduced system obtained when y and v are deleted.

3.5. Treatment of Stiff Equations

In some applications $\partial f_2/\partial v$ and $\partial f_3/\partial T$ are negative and large in magnitude. This makes the system stiff. Some components in the solution of the differential equations decay very fast with time. Our scheme with centered time differences is badly suited to deal with stiff equations (see [2]). The solution contains error components of the form $(-1)^n e^{ct}$, $c > 0$. These strong oscillations with time are shown in Fig. 3.1 ($L - F$). To overcome this difficulty the scheme can be changed so that the stiff terms are treated

TABLE 3.1
Computing Time in Seconds for 0.1 sec of Physical Time t^a

A. Basic method	600 [s]
B. A with simplified functions	280
C. B with different time steps	40
D. C with f_2 and f_3 extrapolated	26
E. C with implicit treatment of f_2, f_3	140
F. Reduced system (first three equations only)	14

^a $h = 2.2 \times 10^{-2}$, $k_1 = 4.0 \times 10^{-5}$, $M = 21$.

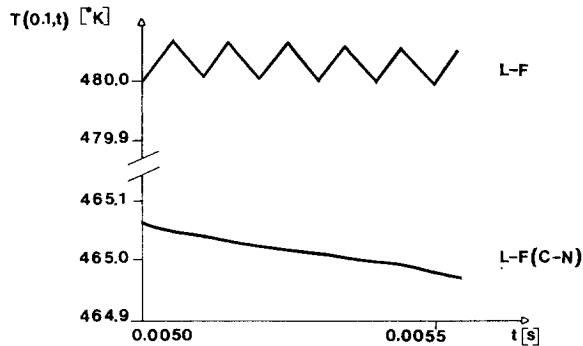


FIG. 3.1. The temperature T at $x = 0.1$ as a function of time for the two methods. $L - F$ denotes the scheme from Section 3.1, $L - F(C - N)$ the scheme described in Section 3.5. Boundary conditions: $(w, T; T)$ (for notation see Section 2.2) $h = 0.028$, $k_1 = 0.49 \times 10^{-5}$, $k_2 = 0.1 \times 10^{-8}$, $\delta_1 = 0.04$, $\delta_2 = 0.16$.

as in the Crank–Nicholson scheme [6, Chap. 6], i.e., in (3.21a) $2\tilde{F}_1^{Mn+\nu}$ is substituted by

$$F_1(u^{Mn+\nu-1}, \tilde{y}^{Mn+\nu-1}, \tilde{v}^{Mn+\nu-1}) + F_1(u^{Mn+\nu+1}, \tilde{y}^{Mn+\nu+1}, \tilde{v}^{Mn+\nu+1})$$

where

$$\tilde{v}^{Mn+\nu} = (1 - \nu/M)v^n + \nu v^{n+1}/M.$$

The nonlinear equations obtained are solved by Newton's method. In Fig. 3.1 it is seen that the oscillating error components mentioned above vanish with this implicit technique ($L - F(C - N)$).

4. NUMERICAL EXPERIMENTS

In this section we will present results obtained from calculations based on the method given in Section 3.4. We will concentrate on the effect of dissipation. Initial and boundary values, parameters, and functions are specified in [3].

As mentioned in Section 3 the dissipation term was added to the ordinary Leap-Frog scheme to prevent oscillations in the solution. The necessity of this term is clearly illustrated by Fig. 4.1. (The notations from Sections 2 and 3 are used.) In the nondissipative case presented there, the oscillations increased rapidly with each step and the solution degenerated completely.

For this short time the boundary dissipation does not make any essential difference.

In Fig. 4.2, however, we see how the lack of boundary dissipation causes oscillations. The solution T develops a discontinuity near the boundary $x = 0$, and at a later time this effects the velocity w . In the case $\delta_2 = 4\delta_1$ the solution is smooth over long time intervals.

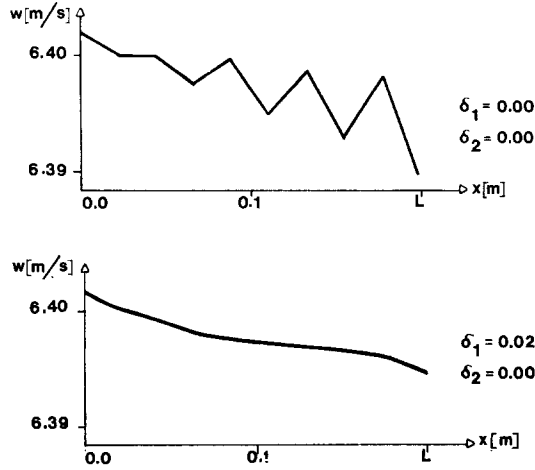


FIG. 4.1. The velocity w as a function of x for different dissipation coefficients. $h = 0.022$, $k_2 = 0.2 \times 10^{-4}$, $n = 350$, $t = 0.007$. Boundary conditions: $(w, T; T)$.

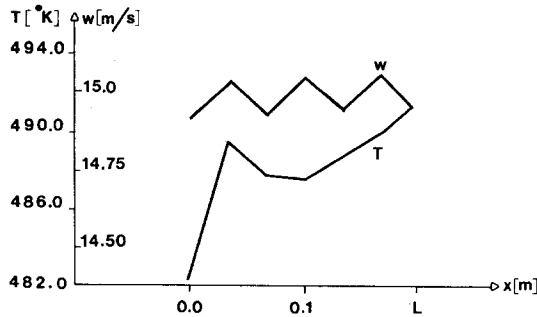


FIG. 4.2. The velocity w and temperature T as a function of x with dissipation coefficients $\delta_1 = 0.02$, $\delta_2 = 0.0$. $h = 0.032$, $k = 0.6 \times 10^{-4}$, $t = 0.090$. Boundary conditions: $(\rho, T; T)$.

The instability introduced by the wrong extra boundary condition (3.20) for the vector y is very weak. The use of wrong time differencing in the boundary conditions for u is more severe.

This stronger instability for Eqs. (2.2a) than for (2.2b) can depend on the fact that A_1 has both positive and negative eigenvalues. The error is in this case reflected back and forth between the boundaries and amplified each time.

Another possible error source could be the numerical solution of the nonlinear equations (2.2d) as described in Section 3. However, the special technique using a diagonal iteration matrix two times enters no errors of practical importance. Test runs with the full Newton method and many iterations effected only the fourth digit in the solution. Finally, Fig. 4.3 shows the solutions for different types of well-posed boundary conditions.

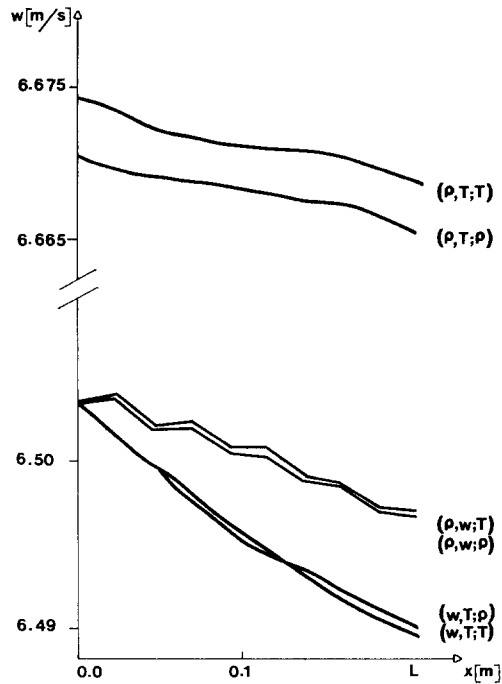


FIG. 4.3. The velocity w as a function of x for different boundary conditions: $h = 0.022$, $k = 0.4 \times 10^{-4}$, $n = 390$, $t = 0.015$, $\delta_1 = 0.02$, $\delta_2 = 0.08$.

REFERENCES

1. R. COURANT AND K. O. FRIEDRICHS, "Supersonic Flow and Shock Waves," Interscience, New York, 1948.
2. G. DAHLQUIST, *Trans. Roy. Inst. Tech. Stockholm* No. 130 (1959).
3. B. ENGQUIST, B. GUSTAFSSON, AND J. VREEBURG, "A Difference Method for a Non-linear Mixed Hyperbolic-Parabolic Problem I," Uppsala University, Dept. Comp. Sciences, Report No. 52, 1974.
4. B. GUSTAFSSON, H.-O. KREISS, AND A. SUNDSTRÖM, *Math. Comp.* **26** (1972), 649-686.
5. H.-O. KREISS AND J. OLIGER, "Methods for the Approximate Solution of Time Dependent Problems," GARP Publication Series, No. 10, 1973.
6. R. D. RICHTMYER AND K. W. MORTON, "Difference Methods for Initial-Value Problems," Interscience, New York, 1967.
7. G. STRANG, *J. Math. Kyoto Univ.* **6** (1967), 397-417.
8. J. M. VARAH, *SIAM J. Numer. Anal.* **8** (1971), 598-615.