A Remark on Collocation and Upwinding in First Order Hyperbolic Systems

Philip T. Keenan

November, 1992

TR92-38

A Remark on Collocation and Upwinding in First Order Hyperbolic Systems

Philip T. Keenan*

November 30, 1992

Abstract

Keenan[2] defines and analyzes a new numerical method for coupled systems of nonlinear first order hyperbolic partial differential equations with one degenerate eigenvalue. That work extends in a certain direction the collocation method described by Luskin[3], which applies to systems with all the eigenvalues uniformly bounded away from zero. Luskin's method and Keenan's method both have direct application to the study of one dimensional fluid flow through pipelines. The pressure and velocity of an isothermal fluid in a pipeline can be described by a coupled pair of nonlinear first order hyperbolic partial differential equations. When thermal effects are important a third equation for temperature is added. While Luskin's method works well for the isothermal situation he discussed, it does not apply in certain common cases when thermal effects are modeled. The analysis of the new method shows how the difficulties that come from the application of standard collocation can be overcome by using upwinded piecewise constant functions for the degenerate component of the solution. Experiments indicate that this method is a substantial improvement over standard collocation.

A number of technical details obscure the analysis presented in Keenan[2], because that work treats the general nonlinear case. The present paper describes and analyzes the method in the context of a linear, constant coefficient system of equations. In this special case the proof simplifies considerably.

Keywords: nonlinear first order hyperbolic system, collocation method, upwinding, thermal pipeline simulation. (AMS/MOS 65N35).

*Department of Computational and Applied Mathematics, Rice University. Supported in part by a National Science Foundation Postdoctoral Research Fellowship in Mathematical Sciences.

1 Introduction

Keenan[2] defines and analyzes a new numerical method for coupled systems of nonlinear first order hyperbolic partial differential equations with one degenerate eigenvalue. That work extends in a certain direction the collocation method described by Luskin[3], which applies to systems with all the eigenvalues uniformly bounded away from zero. Luskin's method and Keenan's method both have direct application to the study of one dimensional fluid flow through pipelines. The pressure and velocity of an isothermal fluid in a pipeline can be described by a coupled pair of nonlinear first order hyperbolic partial differential equations. When thermal effects are important a third equation for temperature is added. While Luskin's method works well for the isothermal situation he discussed, it does not apply in certain common cases when thermal effects are modeled. The analysis of the new method shows how the difficulties that come from the application of standard collocation can be overcome by using upwinded piecewise constant functions for the degenerate component of the solution. Experiments indicate that this method is a substantial improvement over standard collocation.

A number of technical details obscure the analysis presented in Keenan[2], because that work treats the general nonlinear case. The present paper describes and analyzes the method in the context of a linear, constant coefficient system of equations. In this special case the proof simplifies considerably.

For details on the application of the new method to pipeline simulation, as well as experimental results and convergence proofs in the general case of nonlinear coupled systems, see [2].

This paper treats a linear, constant coefficient model problem based on the thermal pipeline equations described in [2]. The model problem is presented in Section 2. Next the new numerical method is defined in Section 3. A representative theorem describing asymptotic convergence of the numerical method is presented in Section 4. Finally, the proof of the theorem is presented in Section 5.

2 The Model Problem

Consider the following system of two coupled first order constant coefficient hyperbolic partial differential equations in one space dimension:

$$\begin{aligned} p_t + v_s p_x + aT &= 0, \\ T_t + v_f T_x + bp &= 0, \end{aligned} \tag{1}$$

where $p = p(x,t)$ and $T = T(x,t)$ are sought in the region $x \in [0,1], t \in (0,1]$. Here v_s, v_f, a, b are constants,

$$v_s \gg v_f \geq 0,$$

and $p(x,0)$, $T(x,0)$, $p(0,t)$ and $T(0,t)$ are given.

The p component here is based on the pressure in the pipeline equations, and T is based on temperature. Here p is advected at a high speed v_s compared with T, which flows only slowly. In fact, v_f can equal zero, in which case T is said to stagnate. It is known that standard collocation can produce bizarre behavior in problems where stagnation can occur, such as in the thermal simulation of pipeline flow. For instance, if $v_f = 0$ and $b = 0$, a change in T at $x = 0$ would be instantly propagated as a saw tooth wave down the entire pipe.

To keep the analysis of the model problem simple, p and T are here only coupled through lower order terms. The actual thermal pipeline equations include additional coupling through x-derivative terms and in the coefficient functions. Moreover, v_f becomes an additional unknown, representing the fluid velocity. All of these features complicate the definition and analysis of the numerical method in the general case; see [2] for details. Also note that all the ideas in both papers generalize immediately to the case of systems of $n > 2$ equations for which all the eigenvalues but one are bounded uniformly away from zero.

3 The Numerical Method

3.1 Discrete Notation for Collocation in One Space Dimension

Let N be a positive integer and let $\triangle x = 1/N$. Let

$$x_j = j\triangle x,\ j = 0, 1, \ldots, N.$$

Similarly, let

$$x_{j+1/2} = (j + 1/2)\triangle x,\ j = 0, 1, \ldots, N-1.$$

The x_j are called the knots and the $x_{j+1/2}$ the midpoints. For any function $u(x)$ let

$$\begin{aligned} u_j &= u(x_j), \\ u_{j+1/2} &= u(x_{j+1/2}), \\ u_{j,c} &= \frac{1}{2}(u_j + u_{j+1}). \end{aligned}$$

For any functions $f(x)$ and $g(x)$ let

$$\begin{aligned} (f,g)_{L^2} &= \int_0^1 f(x)g(x)\,dx, \\ < f,g >_{m^2} &= \sum_{j=0}^{N-1} f_{j+1/2}g_{j+1/2}\,\triangle x, \\ < f,g >_{l^2} &= \sum_{j=1}^{N-1} f_j g_j\,\triangle x + \frac{1}{2}(f_0 g_0 + f_N g_N)\triangle x. \end{aligned}$$

The first is the usual L^2 inner product; the latter two are discrete versions taken at the midpoints or at the knots. Also define the following L^2-like norms:

$$\begin{aligned} \|f\|_{L^2} &= \sqrt{(f,f)_{L^2}}, \\ |f|_{m^2} &= \sqrt{<f,f>_{m^2}}, \\ |f|_{l^2} &= \sqrt{<f,f>_{l^2}}. \end{aligned}$$

Next adopt the following notation for L^∞-like norms:

$$\begin{aligned} \|f\|_{L^\infty} &= \max_{x\in[0,1]} |f(x)|, \\ |f|_{m^\infty} &= \max_{j\in\{0,\ldots,N-1\}} |f_{j+1/2}|, \\ |f|_{l^\infty} &= \max_{j\in\{0,\ldots,N\}} |f_j|. \end{aligned}$$

Let M be a positive integer and let $\triangle t = 1/M$. Let

$$t^n = n\triangle t,\ n = 0, 1, \ldots, M.$$

Similarly, let

$$t^{n+\theta} = (n+\theta)\triangle t,\ n = 0, 1, \ldots, M-1,$$

for any $\theta \in (0, 1]$. For any function $u(t)$ use

$$\begin{aligned} u^n &= u(t^n), \\ u^{n+\theta} &= u(t^{n+\theta}), \\ u^{n,\theta} &= \theta u^{n+1} + (1-\theta)u^n. \end{aligned}$$

For any function $f(x,t)$, let

$$|f|_{l^\infty(L^2)} = \max_{n\in\{0,\ldots,M\}} \|f(\cdot,t^n)\|_{L^2}.$$

In general define the composition of any pair of time and space norms in the analogous way.

For any function $u(x)$ define an x-difference by

$$\partial_x u_{j+1/2} = \frac{u_{j+1} - u_j}{\triangle x},$$

and for any function $u(t)$ and any $\theta \in (0, 1]$ use the time difference

$$\partial_t u^{n+\theta} = \frac{u^{n+1} - u^n}{\triangle t}.$$

Let

$$\text{Poly}^k = \{f(x) : f \text{ is a polynomial in } x \text{ of degree at most } k\}.$$

Let

$$P_l^k = \{f(x) : f|_{[x_j, x_{j+1}]} \in \text{Poly}^k \text{ and } f \in C^l\}.$$

In particular P_0^1 is the class of continuous piecewise linear functions on the mesh defined by the x_j, while P_{-1}^0 is the class of piecewise constant functions discontinuous at the x_j.

3.1.1 Example 1: Standard Collocation

Consider the scalar, linear, first order hyperbolic partial differential equation

$$u_t + au_x = f,$$

with $a(x,t) > a_0 > 0$ for all $x \in [0,1]$ and $t \in [0,1]$, with $u(x,0) = u_0(x)$ and $u(0,t) = u_l(t)$ given. Standard collocation is a natural way to compute an approximate solution. One seeks a function $U(x,t)$ approximating $u(x,t)$ and defined as follows. At each t^n, $U(x,t^n) \in P_0^1$. Between time levels U is extended by linear interpolation in time, meaning $U(x,t^{n+\theta}) = \theta U(x,t^{n+1}) + (1-\theta)U(x,t^n)$ for all x, n and $\theta \in (0,1]$. One takes $U(x,0)$ to be an approximation to $u_0(x)$, such as the interpolant. To compute U^{n+1} from U^n requires $N+1$ equations. One is given by the boundary condition $U^{n+1}(0) = u_l(t^{n+1})$. For the rest, choose a fixed $\theta \in [1/2, 1]$ and require $U(x,t)$ to satisfy the differential equation at the N points $(x_{j+1/2}, t^{n+\theta})$, $j = 0, \ldots, N-1$. This yields the system

$$(U_t)_{j+1/2}^{n+\theta} + a_{j+1/2}^{n+\theta}(U_x)_{j+1/2}^{n+\theta} = f_{j+1/2}^{n+\theta}, \tag{2}$$

with $j = 0, \ldots, N-1$. Because U is piecewise linear in space and in time, this reduces to the system

$$\frac{U_{j+1/2}^{n+1} - U_{j+1/2}^n}{\triangle t} + a_{j+1/2}^{n+\theta}\frac{U_{j+1}^{n+\theta} - U_j^{n+\theta}}{\triangle x} = f_{j+1/2}^{n+\theta}. \tag{3}$$

Or, in the above notation for discrete derivatives,

$$\partial_t U_{j+1/2}^{n+\theta} + a_{j+1/2}^{n+\theta}\partial_x U_{j+1/2}^{n+\theta} = f_{j+1/2}^{n+\theta}. \tag{4}$$

The phrase "with $j = 0, \ldots, N-1$" will henceforth be suppressed as implied by context.

Define the set of collocation points

$$\mathcal{CP} = \{(x_{j+1/2}, t^{n+\theta}) : j = 0, \ldots, N-1, \text{ and } n = 0, \ldots, M-1\}.$$

In equations such as (4) in which all the subscripts are $j + 1/2$ and all the superscripts are $n + \theta$, the subscripts and superscripts will henceforth be suppressed. To remind the reader of this convention the phrase "on $\mathcal{CP}$" will be appended to such equations. This convention will allow the use of subscripts for indexing component equations and variables in the case of systems of equations. With this convention, the equations for standard collocation become

$$\partial_t U + a\partial_x U = f \text{ on } \mathcal{CP}. \tag{5}$$

3.1.2 Example 2: Upwinding

Suppose now that $a(x,t)$ in Example 1 is no longer bounded away from zero. For simplicity in this example, however, assume $a(x,t) \geq 0$ for all x and t. To avoid the instabilities which arise from the application of standard collocation in this case, one can use a technique known as "upwinding".

Again one seeks a function $U(x,t)$ approximating $u(x,t)$. Now however, at each t^n, $U(x,t^n) \in P^0_{-1}$. Between time levels U is again extended by linear interpolation in time, meaning $U(x,t^{n+\theta}) = \theta U(x,t^{n+1}) + (1-\theta)U(x,t^n)$ for all x, n and $\theta \in (0,1]$. Again one takes $U(x,0)$ to be a suitable approximation to $u_0(x)$.

Requiring U to satisfy (2) no longer seems sensible, as the U_x term vanishes. Note that so far $U(x_j,t)$ is undefined for all j and t. To avoid losing information about the slope of U, one extends the definition of U by setting

$$U(x_j,t^n) = U(x_{j-1/2},t^n),$$

for $j = 1,\ldots,N$, and

$$U(0,t^n) = u_l(t^n).$$

One continues to define U at intermediate times by linear interpolation in time. The spatial asymmetry in the definition of $U(x_j,t^n)$ is due to the assumed asymmetry in the sign of the coefficient $a(x,t)$. One can interpret u as an entity being advected by a velocity a. U is defined by looking "upwind" relative to this velocity, whence the name of the technique.

To compute U^{n+1} from U^n requires N equations. One chooses a fixed $\theta \in [1/2,1]$ and requires $U(x,t)$ to *approximately* satisfy the differential equation at the N points $(x_{j+1/2},t^{n+\theta})$, $j = 0,\ldots,N-1$. That is, rather than require (2), one instead requires (3). This equation is well defined because of the extended definition of U. As in Example 1 this equation can be written compactly as (4). Using the same convention of suppressing the subscripts and superscripts, collocation using upwinded piecewise constants can be written

$$\partial_t U + a\partial_x U = f \text{ on } \mathcal{CP},$$

just as in (5) for piecewise linears.

It turns out that upwinding amounts to adding extra numerical diffusion to the numerical method, which provides the stability missing from standard collocation. Had one "downwinded" instead, the extra diffusion would have the wrong sign, making the method less stable rather than more.

Figure 1 illustrates the upwinding process in the case of velocity flowing to the right. The dashed line shows the slope used as an approximation to the x-derivative.

3.1.3 Component Notation

In both the previous examples the convention of dropping spatial subscripts and temporal superscripts was employed. This convention will be continued throughout this paper, to simplify the notation and to allow the use of subscripts denoting vector and matrix components.

In particular, if A is a matrix one writes A_{ij} for the component of A in the i'th row and j'th column. The identity matrix is represented by the Kronecker delta symbol, defined by

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

The summation convention will be used throughout this paper; it implies summations over all repeated component indices. Thus if B is another matrix of compatible dimensions, one defines $A_{ij}B_{jk}$ by

$$A_{ij}B_{jk} = \sum_j A_{ij}B_{jk}.$$

3.2 Defining the New Numerical Method

Consider now the model problem (1), written in vector form

$$u_t + Au_x + Bu = 0, \tag{6}$$

with $u = (p, T)^{tr}$. Here tr means transpose. The model problem was chosen to make the matrix A diagonal, which simplifies the analysis of the method. Notice that T_x only occurs in the T equation — in the general case the equations must be rewritten to make this happen.

The new numerical method combines standard collocation and upwinding as follows. One seeks a vector function $U(x,t)$ approximating $u(x,t)$. At each t^n, U_1 is to be in P_0^1, but $U_2 \in P_{-1}^0$. Moreover the definition of U_2 is extended to the knots x_j by upwinding as in Example 2. In particular, one defines

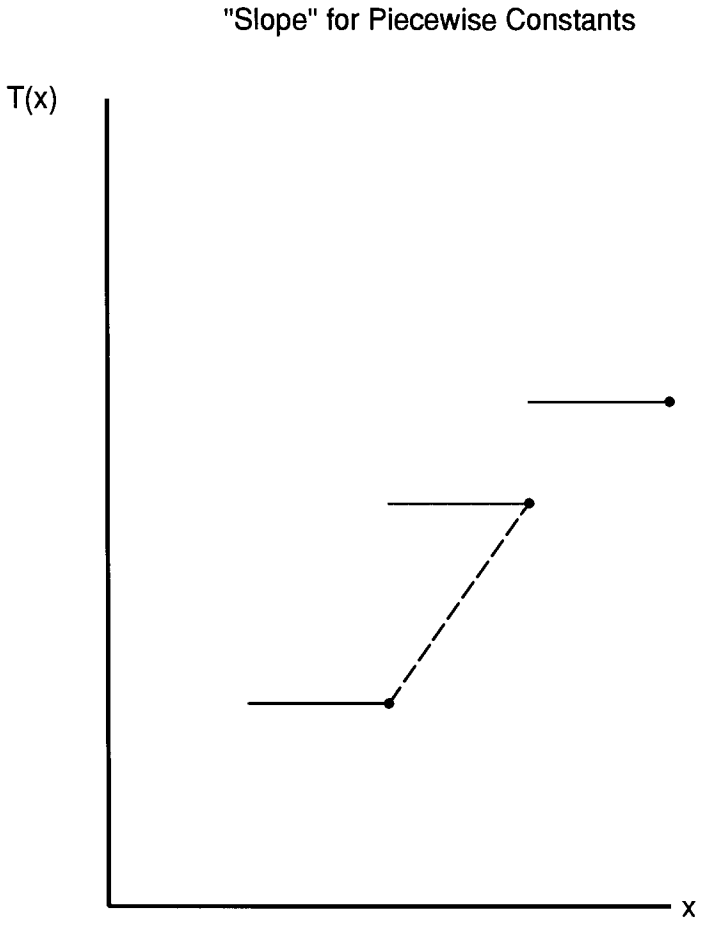

Figure 1: Upwinding Illustration

$$U_2(x_j, t^n) = U_2(x_{j-1/2}, t^n),$$

since $v_f \geq 0$. In the general case treated in [2], the velocity is one of the unknowns and can change sign, which complicates the definition of upwinding.

When $x_{j-1/2}$ falls outside the domain $[0,1]$, one inserts the corresponding boundary condition instead. Note how the upwinding technique fits naturally with the specified boundary condition:

$$U_2(x_0, t^n) = T(0, t^n).$$

Each component U_k is defined to be piecewise linear in time between the t^n's. $U(x,0)$ is taken to be a suitable approximation to $u^0(x)$; it can be the interpolant. To incorporate the remaining boundary condition one sets

$$U_1(x_0, t^n) = p(0, t^n).$$

The new numerical method determines U^{n+1} from U^n by requiring that U satisfy a certain linear system of equations at the collocation points $(x_{j+1/2}, t^{n+\theta})$, subject to the special interpretation of $U_{2,x}$ described in Example 2. Per the conventions previously described this discrete linear system can be written

$$\partial_t U + A\partial_x U + BU = 0, \tag{7}$$

where as with all following discrete equations the phrase " on $\mathcal{CP}$" is to be understood.

This discrete linear system may be solved very efficiently using a straightforward modification of the standard algorithm for tridiagonal matrices.

4 Theoretical Results

Assumption 1 *Assume $\theta \in (\frac{1}{2}, 1]$ is a given constant. Assume there is a constant K_0, independent of $\triangle x$ and $\triangle t$, such that*

$$\frac{1}{K_0} \leq \frac{\triangle x}{\triangle t} \leq K_0,$$

as both $\triangle x$ and $\triangle t$ go to zero. Moreover, assume the system (1) with given initial and boundary data has a unique solution which is smooth for all $t \in [0,1]$.

Theorem 1 *Consider the model problem (1). Let assumption 1 hold. Define a discrete solution U by (7). Then there is a constant C which depends on K_0 and on Sobolev norms for u but remains bounded even when $v_f = 0$, and which is otherwise independent of $\triangle x$ and $\triangle t$, such that for $\triangle x$ and $\triangle t$ sufficiently small,*

$$||U - u||_{l^\infty(L^2)} \leq C\triangle x.$$

This theorem is representative of those presented in [2].

As described in [2], the results in this paper generalize to nonlinear first order hyperbolic systems of any size where exactly one eigenvalue is not bounded uniformly away from zero.

4.1 Computational Results

For detailed computational results in the general nonlinear case see [2]. As an example, however, Figure 2 illustrates the convergence of the p variable, in L^∞ norm, at a particular point in time, in two particular runs of the actual thermal pipeline simulator. The two cases differ primarily in boundary conditions. In one case the system is converging smoothly toward a steady state, while in the other a pulsed boundary condition has excited a traveling sonic wave. Notice the second order convergence in the smooth steady state case and the first order convergence (because of the piecewise constants used for T) in the rapidly varying wave case.

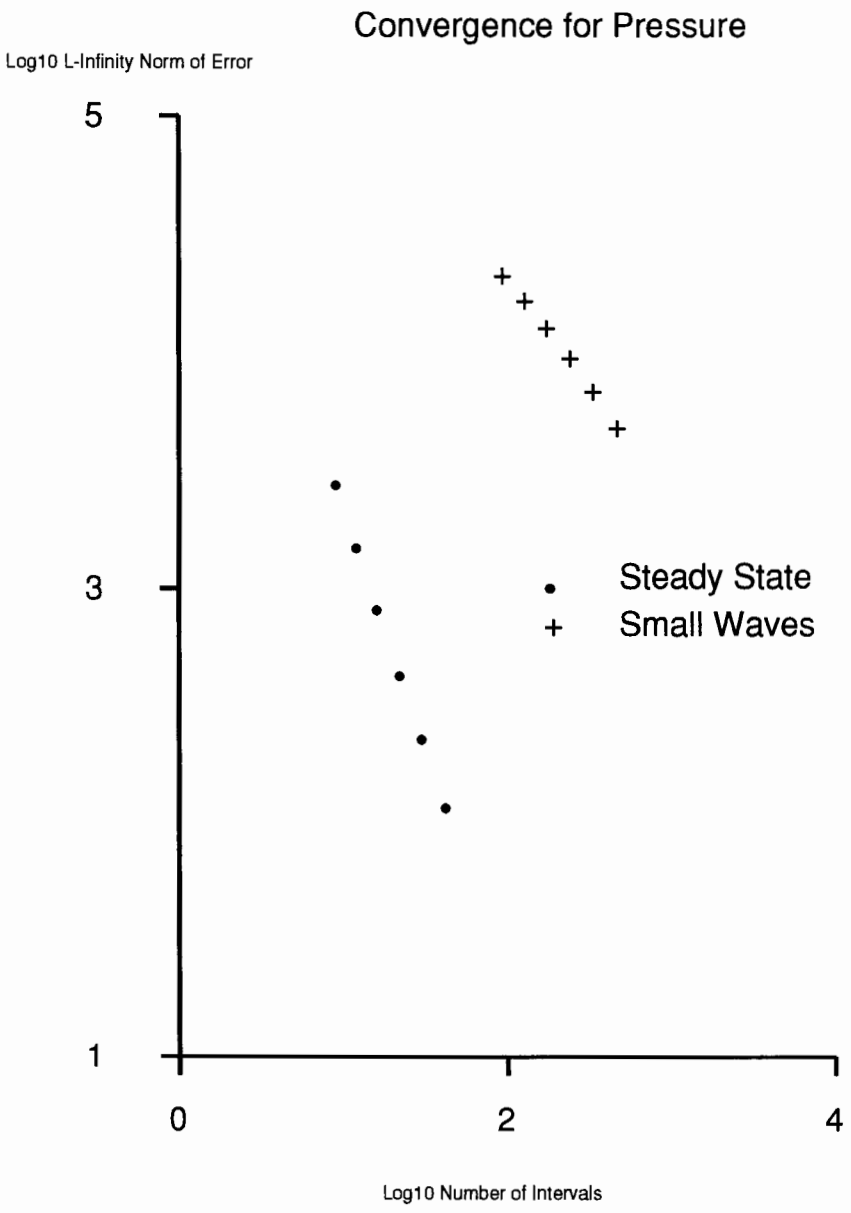

Figure 2: Convergence of the Error

5 Error Analysis

5.1 Overview

The *a-priori* error bound stated in the theorem is derived from an energy estimate based on applying the discrete scheme to $U - W$ where W is a discrete interpolant of u. The "error equation" satisfied by $U - W$ is an inhomogeneous version of the discrete scheme itself, with truncation error terms on the right hand side. In the general case, the error equation is then diagonalized by changing variables, following the earlier work of Thomeé and Luskin, though for the model problem it is already in diagonal form. Next an evolution inequality (28) is developed for certain norms of the error, using a discrete m^2 inner product of the diagonalized error equation with a certain test function. The test function is the sum of three terms representing the error, its time derivative and its space-time second derivative, each weighted with a carefully chosen power of $\triangle x$. The time derivative test function is the only one not used in Luskin[3]. In developing this evolution inequality there are many terms to estimate; these are summarized in a tableau and bounded one by one. Finally the evolution inequality is used to derive the error bounds; in the linear case presented here the usual Gronwall lemma suffices for this last step.

For a proof in the intermediate case of linear *variable* coefficient systems, see also [1].

5.2 The Error Equation

Convention 1 *In what follows let C be a generic constant whose value in any particular equation depends upon various Sobolev norms of u, and on the constants in the model problem, and the constant K_0 of Assumption 1, but which is otherwise* independent *of the discretization parameters $\triangle x$, $\triangle t$ and θ.*

Throughout the proof assume assumption 1 holds. The number 2 occurs throughout due to the special treatment of T in the model problem, but the proof generalizes to any size system with exactly one degenerate eigenvalue.

Consider the model problem in vector form (6), with the numerical method given by (7). Recall that U is piecewise linear in time, U_1^n is piecewise linear in space, and U_2^n is discontinuous piecewise constant, upwinded by v_f. It will now be useful to introduce a discrete interpolant W of u. Such a function is defined in the same discrete space as U. Therefore $U - W$ is also in the discrete space, and thus is easier to analyze than $U - u$. Define W by $W_1^n(x_j) = u_1^n(x_j)$ and $W_2^n(x_{j+1/2}) = u_2^n(x_{j+1/2})$, with W_2^n at the knots upwinded by v_f, so $W_2^n(x_j) = W_2^n(x_{j-1/2})$, for $j = 1, \ldots, N$.

Define the total error

$$\Psi = u - U,$$

the discrete error

$$\zeta = W - U,$$

and the approximation error

$$e = u - W.$$

Under reasonable conditions it is standard to show that e is small; since $\Psi = e + \zeta$, it will suffice to show that ζ is small. In particular, the interpolant W satisfies the equation

$$\partial_t W + A\partial_x W + BW = TE, \tag{8}$$

where the *local truncation error* TE is given by

$$TE = (\partial_t W - u_t) + A(\partial_x W - u_x) + B(W - u).$$

By standard calculations involving Taylor series expansions, one can show that for some C independent of $\triangle x$, $\triangle t$ and θ,

$$|TE^{n+\theta}|_{m^2} \leq C(\triangle x + \triangle t^2 + (\theta - \frac{1}{2})\triangle t), \quad \text{for all } n. \tag{9}$$

Similarly one may show

$$|(TE_1^{n+1+\theta} - TE_1^{n+\theta})|_{m^2} \leq C(\triangle x \triangle t + \triangle t^2). \tag{10}$$

Note that this equation does not apply to the second component of the truncation error, which in general is one order less accurate due to the use of piecewise constants.

Subtracting (8) and (7) shows that the discrete error ζ satisfies

$$\underset{\partial_t \zeta}{\boxed{A}} + \underset{A\partial_x \zeta}{\boxed{B}} + \underset{B\zeta}{\boxed{D}} = \underset{TE}{\boxed{C}}. \tag{11}$$

where each of the four terms is labeled with a letter for future convenience.

Thus the discrete error ζ satisfies an inhomogeneous version of the same equation satisfied by U.

In the general case the next step in the proof is to diagonalize the matrix A by changing variables. This requires introducing extra notation. Fortunately in the model problem A is already diagonal. The nonlinear version of (11) is further complicated by the appearance of the usual "shower of terms" from differentiating A and B.

Let

$$P = \left\{ \begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array} \right\}.$$

Now define the test function to be used in the energy analysis:

$$\varphi = \overset{\boxed{1}}{\zeta} + \alpha \triangle x \overset{\boxed{2}}{P} \partial_t \zeta + \beta \triangle x \overset{\boxed{3}}{P A} \partial_t \partial_x \zeta, \tag{12}$$

where α and β are two unspecified but non-negative parameters which will be determined later and which will be independent of $\triangle x$ and $\triangle t$. In the general case an extra factor appears in some terms of φ to handle the boundary conditions, but in the constant coefficient model problem is it not needed.

5.3 The 12 Product Terms

Now form the vector inner product of both sides of (11) with φ, producing another equation involving 12 product terms which must be considered individually. The following chart summarizes the situation.

		A	B	C	D
$-$	$+$	$-$	$-$	$-$	$-$
1	|	L	L	R	R
2	|	L	R	R	R
3	|	L	L	R	R

The five terms marked with an "L" are primarily "helping" or "left hand side terms"; the other 7 are right hand side terms. The analysis of each product term is conducted as follows. The product equation holds at every point of $\mathcal{CP}$; for each time level $t^{n+\theta}$, multiply by $\triangle x$ and sum over all $x_{j+1/2}$'s, thus forming the discrete spatial midpoint-based m^2 norm.

For each right hand side term an upper bound will be given for the sum over the $x_{j+1/2}$'s. For the terms marked L, a positive lower will be given instead. The bounds may not be obvious at first, but they follow in straightforward ways from the properties of the objects involved, in particular from knowing that ζ is piecewise linear in time and either piecewise linear or piecewise constant in space.

For non-constant coefficients additional lower order terms would appear in the following bounds, as in the nonlinear case. However, the constant coefficient analysis does include the analysis of all the highest order terms.

Later, some right hand side terms will be hidden by direct subtraction, and Gronwall's inequality will be used to handle the rest.

Define

$$\hat{\mathcal{P}}_x = \{x_{j+1/2} : j = 0, \ldots, N-1\}.$$

For any function $f(t)$ let

$$(f^{n+\theta})^+ = f^{n+1},$$

and

$$(f^{n+\theta})^- = f^n.$$

First consider the three main "helping terms" $A-1$, $A-2$, and $B-3$.

5.3.1 A-1

The steps leading to the bound are spelled out in some detail for this first left-hand side term:

$$\partial_t\zeta\cdot\zeta \geq \frac{1}{2}\partial_t(\zeta^2) + (\theta-\frac{1}{2})\triangle t(\partial_t\zeta)^2, \tag{13}$$

The above equation holds at each point of $\mathcal{CP}$, hence for each t^n,

$$\sum_{\hat{\mathcal{P}}_x}\partial_t\zeta\cdot\zeta\triangle x \geq \frac{1}{2}\partial_t|\zeta|^2_{m^2} + (\theta-\frac{1}{2})\triangle t|\partial_t\zeta|^2_{m^2}.$$

Here C is a generic constant independent of $\triangle x$, $\triangle t$; as always it can depend on norms of u.

5.3.2 A-2

$$\sum_{\hat{\mathcal{P}}_x}\partial_t\zeta\cdot\alpha\triangle x P\partial_t\zeta\triangle x \geq \alpha\triangle x|P\partial_t\zeta|^2_{m^2}. \tag{14}$$

5.3.3 B-3

$$\sum_{\hat{\mathcal{P}}_x}\partial_x(A\zeta)\cdot\beta\triangle x P\partial_t\partial_x(A\zeta)\triangle x$$
$$\geq \beta\triangle x\left(\frac{1}{2}\partial_t|P\partial_x(A\zeta)|^2_{m^2} + (\theta-\frac{1}{2})\triangle t|P\partial_t\partial_x(A\zeta)|^2_{m^2}\right). \tag{15}$$

Next upper bounds for right hand side terms are derived, beginning with the easiest ones.

5.3.4 D-1

$$|\sum_{\hat{\mathcal{P}}_x}B\zeta\cdot\zeta\triangle x| \leq C|\zeta|^2_{m^2}. \tag{16}$$

5.3.5 B-2

$$|\sum_{\hat{\mathcal{P}}_x}\partial_x(A\zeta)\cdot\alpha\triangle x P\partial_t\zeta\triangle x| \leq \frac{\alpha\triangle x}{64}|P\partial_t\zeta|^2_{m^2} + \alpha\triangle x C|P\partial_x(A\zeta)|^2_{m^2}. \tag{17}$$

5.3.6 D-2

$$|\sum_{\hat{\mathcal{P}}_x} B\zeta \cdot \alpha\triangle x P\partial_t\zeta \triangle x| \leq \frac{\alpha\triangle x}{64}|P\partial_t\zeta|^2_{m^2} + \alpha\triangle x C|\zeta|^2_{m^2}. \tag{18}$$

5.3.7 C-1

$$\sum_{\hat{\mathcal{P}}_x} TE \cdot \zeta\triangle x \leq C(|\zeta|^2_{m^2} + |TE|^2_{m^2}). \tag{19}$$

5.3.8 C-2

$$\sum_{\hat{\mathcal{P}}_x} TE \cdot \alpha\triangle x P\partial_t\zeta\triangle x \leq \frac{\alpha\triangle x}{64}|P\partial_t\zeta|^2_{m^2} + \alpha\triangle x C|TE|^2_{m^2}. \tag{20}$$

5.3.9 B-1

In the constant coefficient case this term contains only helping terms, so a lower bound is derived.

$$\sum_{\hat{\mathcal{P}}_x} \partial_x(A\zeta) \cdot \zeta\triangle x \geq \frac{1}{2}((v_s l_1 \zeta_1^2) + (v_f \zeta_2^2))|_{x=0}^{x=1}. \tag{21}$$

5.3.10 A-3

Note that ζ_2 does not appear in this term, so only piecewise linear functions need be considered.

$$\beta\triangle x \sum_{\hat{\mathcal{P}}_x} \partial_t\zeta \cdot P\partial_x(A\partial_t\zeta)\triangle x \geq \frac{\beta\triangle x}{2}(v_s l_1(\partial_t\zeta_1)^2)|_{x=0}^{x=1}. \tag{22}$$

The spatial boundary terms in (21) and (22) turn out to give non-negative helping terms. In the general nonlinear case this requires a slightly more general test function with a parameter to be chosen sufficiently small relative to certain $O(1)$ constants depending only on u. This is based on a trick used by Luskin and pioneered by Thomeé. In the present case, however, it is clear by inspection. The boundary term in (21) is

$$\frac{1}{2}(-v_s\zeta_1^2(0) - v_f\zeta_2^2(0) + v_s\zeta_1^2(1) + v_f\zeta_2^2(1)).$$

Now $\zeta(0) = 0$ by choice of boundary conditions, and the remaining terms are non-negative because of the sign of the velocities. Similar arguments apply to the $\partial_t\zeta$ terms from (22).

5.3.11 C-3

Use the following formula for summation by parts in time at $t = t^{n+\theta}$, in which for clarity time superscripts are not suppressed:

$$\begin{aligned} a^{n+\theta}\partial_t b^{n+\theta} &= \frac{1}{\triangle t} a^{n+\theta}(b^{n+1} - b^{n-1}) \\ &= \frac{1}{\triangle t}(a^{n+\theta}b^{n+1} - a^{n-1+\theta}b^{n-1}) - \frac{1}{\triangle t}(a^{n+\theta} - a^{n-1+\theta})b^{n+1}. \end{aligned} \tag{23}$$

Thus

$$\begin{aligned} &\sum_{\hat{\mathcal{P}}_x} TE \cdot \beta \triangle x P \partial_t \partial_x (A\zeta) \triangle x = \\ &\quad \frac{\beta \triangle x}{\triangle t} \sum_{\hat{\mathcal{P}}_x} (TE^{n+\theta} \cdot P\partial_x(A\zeta^{n+1}) - TE^{n-1+\theta} \cdot P\partial_x(A\zeta^{n-1})) \triangle x \\ &\quad - \beta \frac{\triangle x}{\triangle t} \sum_{\hat{\mathcal{P}}_x} (TE^{n+\theta} - TE^{n+\theta-1}) \cdot P\partial_x(A\zeta^{n+1}) \triangle x. \end{aligned} \tag{24}$$

The first term will telescope when summed over n. One can bound the second sum by

$$C\beta \triangle x |P\partial_x(A\zeta^+)|^2_{m^2} + \beta \frac{\triangle x}{\triangle t^2} |TE^{n+1+\theta} - TE^{n+\theta}|^2_{m^2}.$$

5.3.12 D-3

As in $C-3$, sum by parts:

$$\begin{aligned} &\sum_{\hat{\mathcal{P}}_x} B\zeta \cdot \beta \triangle x P \partial_t \partial_x (A\zeta) \triangle x = \\ &\quad \frac{\beta \triangle x}{\triangle t} \sum_{\hat{\mathcal{P}}_x} (B\zeta \cdot P\partial_x(A\zeta^+) - (B\zeta)^{n-1+\theta} \cdot P\partial_x(A\zeta^-)) \triangle x \\ &\quad - \beta \triangle x \sum_{\hat{\mathcal{P}}_x} BP(\zeta^{n+\theta} - \zeta^{n+\theta-1}) \cdot P\partial_x(A\zeta^+) \triangle x. \end{aligned} \tag{25}$$

One can bound the second sum by

$$C\beta \triangle x |P\partial_x(A\zeta^+)|^2_{m^2} + \frac{\alpha \triangle x}{64}(|P\partial_t\zeta|^2_{m^2} + |P\partial_t\zeta^{n+\theta-1}|^2_{m^2}).$$

5.4 The Evolution Inequality

Now collect all 12 terms forming an evolution inequality. In the general nonlinear case one must carefully formulate induction hypotheses in order to analyze this inequality. In the present case, however, the ordinary discrete Gronwall lemma will suffice.

In particular, take α to be small based on some other order one constants. Next take β small relative to α again based on order one constants, and finally require $\triangle x$ and $\triangle t$ to be sufficiently small with respect to these other constants.

A number of right hand side terms now can be directly subtracted off from left hand side terms. These are terms with a small multiplier on them, usually written as $\frac{1}{64}$ above. This results in the inequality

$$\partial_t |\zeta|^2_{m^2} + \triangle x |P\partial_t \zeta|^2_{m^2} + \triangle x \partial_t |AP\partial_x \zeta|^2_{m^2} + TS - \frac{\triangle x}{64} |P\partial_t \zeta^{n+\theta-1}|^2_{m^2} \leq$$
$$C\left(|\zeta^+|^2_{m^2} + |\zeta^-|^2_{m^2} + \triangle x |AP\partial_x \zeta^+|^2_{m^2} + \triangle x |AP\partial_x \zeta^-|^2_{m^2} + \triangle x^2\right). \quad (26)$$

where use was made of Assumption 1 and equations (9) and (10). Here TS stands for the telescoping terms in $C-3$ and $D-3$. Multiplying by $\triangle t$, summing on n and using the fact that the initial error ζ^0 is zero by construction, one obtains

$$|\zeta^N|^2_{m^2} + \triangle x |P\partial_t \zeta|^2_{m^2(m^2)} + \triangle x |AP\partial_x \zeta^N|^2_{m^2} \leq$$
$$C\left(|\zeta|^2_{l^2(m^2)} + \triangle x |AP\partial_x \zeta|^2_{l^2(m^2)} + \triangle x^2\right). \quad (27)$$

Gronwall's lemma then implies

$$|\zeta|^2_{l^\infty(m^2)} + |AP\partial_x \zeta|^2_{l^\infty(m^2)} \leq C\triangle x^2. \quad (28)$$

Careful reading of the proof shows that one can prove a stronger theorem than claimed, in the constant coefficient case, but the intent here has been to present the theorem and proof of the nonlinear case in a simpler context. See also [1] for details of the theorems one can prove in the linear variable coefficient case.

References

[1] P. T. KEENAN, *An error estimate for a new scheme for the general variable coefficient linearized thermal pipeline equations*, Tech. Report 90--20, Department of Computer Science, University of Chicago, 1990.

[2] ——, *Thermal simulation of pipeline flow*, Tech. Report 91–19, Department of Computer Science, University of Chicago, 1991.

[3] M. LUSKIN, *An approximation procedure for nonsymmetric, nonlinear hyperbolic systems with integral boundary conditions*, SIAM J. Numer. Anal., 16 (1979), pp. 145–164.

[4] M. LUSKIN AND T. BLAKE, *The existence of a global weak solution to the nonlinear waterhammer problem*, Comm. Pure Appl. Math, 35 (1982), pp. 697–735.

[5] V. THOMÉE, *A stable difference scheme for the mixed boundary problem for a hyperbolic, first order system in two dimensions*, J. Soc. Indust. Appl. Math., 10 (1962), pp. 229–245.

[6] E. B. WYLIE AND V. STREETER, *Fluid Transients*, McGraw Hill, New York, 1978.