MATHEMATICS OF COMPUTATION, VOLUME 31, NUMBER 139
JULY 1977, PAGES 661–675

On the Stability of Galerkin Methods for Initial-Boundary Value Problems for Hyperbolic Systems

By Max D. Gunzburger*

Abstract. The stability of approximating the solution of mixed initial-boundary value problems for hyperbolic systems by semidiscrete Galerkin methods is studied. It is shown that a particular straightforward Galerkin method yields an unstable approximation, and that this numerical instability is caused by an improper treatment of the boundary. Stable schemes are then presented, one of which differs from the unstable scheme only insofar as the treatment of the boundary is concerned. These stable schemes make use of a particular matrix which symmetrizes the differential system. It is therefore shown that the use of this matrix is crucial to the stability of the computations as well as for obtaining a priori bounds on the energy of the continuous system. This symmetrizing matrix is also related to the diagonalizing matrix for the system of hyperbolic equations and to the Lyapunov matrix for the system of ordinary differential equations resulting from the application of Galerkin's method.

I. Introduction. In recent years there have appeared a few articles, e.g., [1], [2], [3], and [8], considering Galerkin methods for hyperbolic systems of partial differential equations. These are mostly concerned with obtaining a priori error estimates, often of optimal order, by functional analytic methods. The present work is concerned with the stability of semidiscrete Galerkin methods for initial-boundary value problems for hyperbolic systems and differs from the above references in both motivation and spirit.

Dupont [4] recently introduced the use of a symmetrizing matrix in connection with a particular hyperbolic system. This matrix was an essential ingredient in the derivation of a priori error estimates. In the present work it is shown that this symmetrizing matrix is also an essential ingredient for the computation of stable Galerkin solutions. Three Galerkin schemes are considered, only two of which are stable. The essential connection between the symmetrizing matrix and stability is clearly evident through the use of matrix theory. Connection is also made between the symmetrizing matrix and two matrices often encountered in the literature. The first is the diagonalizing matrix which transforms a given hyperbolic system into an equivalent system for characteristic variables. The second is the Lyapunov matrix encountered in the Lyapunov stability theory for the ordinary differential equations resulting from the use of

Received March 18, 1976.
AMS (MOS) subject classifications (1970). Primary 65N30; Secondary 35L50.
* This paper was prepared as a result of work performed under NASA Contract No. NAS1-14101 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, Virginia 23665.

Galerkin's method in space. In addition, an indication of the physical interpretation of the role of the symmetrizing matrix is provided.

In this work the word stable is used in two contexts. The first usage is common to both the continuous solution and its discrete approximation, and is concerned with stability in time. By definition, stability in time means that the solution is bounded as $t \longrightarrow \infty$. The second usage is in the sense of convergence of the numerical approximation, i.e., as the dimension of the approximating space becomes infinite, the approximate solution converges to the continuous one. (In finite element type approximations, the dimension of the approximating space becoming infinite is equivalent to the grid size going to zero.) The latter usage in this work is referred to simply as stability, while stability in time will always be indicated as such. Because we wish to differentiate between these two types of stability we consider below problems whose continuous solutions are bounded in time so that any unbounded growth in the approximate solution is clearly an instability of the numerical method.

II. The Continuous Problem. In this section the model r-dimensional system with constant coefficients

$$u_t = Au_x, \qquad t \geqslant 0,\ 0 \leqslant x \leqslant 1, \tag{2.1}$$

is considered. The time rate of change of the L^2 energy E_2 is given by

$$dE_2/dt = \tfrac{1}{2}(u, u)_t = (u, Au_x), \tag{2.2}$$

where the usual inner product

$$(u, v) = \int_0^1 u^T v\, dx$$

is used. In general, (2.2) gives no a priori information about how the energy behaves in time. Specifically, it cannot be deduced from (2.2) whether or not the energy is bounded by the initial energy. For example, take

$$A = \begin{pmatrix} \nu & 1 \\ 1 & \nu \end{pmatrix}, \qquad u = (u_1, u_2)^T, \tag{2.3}$$

and the boundary data

$$u_1(t, 0) = u_1(t, 1) = 0, \tag{2.4}$$

where $\nu < 1$ is a constant. Then, after integration by parts, (2.2) becomes

$$(u, u)_t = \nu[u_2^2(t, 1) - u_2^2(t, 0)]$$

which yields no a priori information since not even the sign of the right-hand side is known before the solution u is obtained.

The possibility exists of defining the energy in a different norm. Dupont [4] used such a norm in order to derive a priori error estimates for a particular hyperbolic system of two equations, and it is shown below that the use of this norm is also crucial to the stability of the actual computations.

For a given positive definite symmetric constant matrix E, the E energy is defined to be

$$E_E = \tfrac{1}{2}(u, Eu)$$

so that from (2.1)

$$dE_E/dt = (u, EAu_x). \tag{2.5}$$

If the system (2.1) is hyperbolic, then there exists a real nonsingular matrix Q such that $QAQ^{-1} = \Lambda$, where Λ is a diagonal matrix whose diagonal entries are the eigenvalues λ_j of A. Then it is well known (see Taussky [7]) that there exists a symmetric positive definite matrix E that symmetrizes A, i.e., $EA = A^TE$. We now wish to relate E to Q, and ascertain the freedom available in choosing E. Consequently we assume, without loss of generality, that Q is chosen so that the entries of Λ are ordered, i.e., $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r$. Multiple eigenvalues are then grouped together, and it is easy to verify that

$$(Q^T)^{-1}EQ^{-1} = D \quad \text{or} \quad E = Q^TDQ, \tag{2.6}$$

where D is a symmetric positive definite block diagonal matrix whose ith block has dimension equal to the multiplicity of that eigenvalue of A which appears in the corresponding position in Λ. Other than the above restrictions, the nonzero entries of D are arbitrary.

Now, substituting (2.6) into (2.5) yields

$$dE_E/dt = \int_0^1 (w^TD\Lambda w_x)\,dx = \tfrac{1}{2}w^T(t, x)D\Lambda w(t, x)|_{x=0}^1, \tag{2.7}$$

where $w = Qu$, the vector of characteristic dependent variables for the system (2.1).

With the λ's ordered as above, they are also ordered so that $\lambda_1, \ldots, \lambda_p$ are positive and $\lambda_{p+1}, \ldots, \lambda_r$ are negative. Then Λ, D, and w can be partitioned in the following manner

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & -\Lambda_2 \end{pmatrix}, \qquad \begin{cases} \Lambda_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_p), \\ \Lambda_2 = -\operatorname{diag}(\lambda_{p+1}, \ldots, \lambda_r), \end{cases}$$

$$D = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}, \qquad w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix},$$

where w_1 and w_2 are vectors of dimension p and $q = r - p$, respectively. Then (2.7) can be written as

$$\frac{dE_E}{dt} = \tfrac{1}{2}(w_1^TD_1\Lambda_1w_1 - w_2^TD_2\Lambda_2w_2)|_{x=0}^1. \tag{2.8}$$

Furthermore, well-posed linear homogeneous boundary data can always be written as

$$w_2(t, 0) = S_2w_1(t, 0), \qquad w_1(t, 1) = S_1w_2(t, 1), \tag{2.9}$$

where S_1 and S_2 are matrices of dimension $q \times p$ and $p \times q$, respectively. The data

(2.9) is chosen to be homogeneous so that the solution of the system (2.1) is driven by the initial data. In addition, S_1 and S_2 are restricted below so that the E energy is bounded by the initial E energy. Substituting into (2.8) yields

$$2\frac{dE_E}{dt} = w_2^T(t, 1)\,[S_1^T D_1 \Lambda_1 S_1 - D_2\Lambda_2]\,w_2(t, 1) + w_1^T(t, 0)[-D_1\Lambda_1 + S_2^T D_2 \Lambda_2 S_2]\,w_1(t, 0). \tag{2.10}$$

If the E energy is to be conserved, i.e., $dE_E/dt = 0$ for any w_1 and w_2 then necessarily

$$S_1^T D_1 \Lambda_1 S_1 - D_2\Lambda_2 = 0 \quad \text{and} \quad S_2^T D_2 \Lambda_2 S_2 - D_1\Lambda_1 = 0. \tag{2.11}$$

Since $D_1^{1/2}$ and $\Lambda_1^{1/2}$ commute, as do $D_2^{1/2}$ and $\Lambda_2^{1/2}$, letting

$$\Phi_1 = (D_1\Lambda_1)^{1/2} \quad \text{and} \quad \Phi_2 = (D_2\Lambda_2)^{1/2}$$

in (2.11) yields

$$(\Phi_1 S_1 \Phi_2^{-1})^T(\Phi_1 S_1 \Phi_2^{-1}) = I \quad \text{and} \quad (\Phi_2 S_2 \Phi_1^{-1})^T(\Phi_2 S_2 \Phi_1^{-1}) = I,$$

which implies that

$$\Phi_1 S_1 \Phi_2^{-1} = V_1, \qquad \Phi_2 S_2 \Phi_1^{-1} = V_2,$$

where V_1 and V_2 are $p \times q$ and $q \times p$ matrices, respectively, both of which have orthonormal columns. Assume now that $p < q$. The orthonormality of the columns of V_1 then implies that there are q vectors of dimension $p < q$ which are linearly independent. Therefore, $p \geqslant q$. However, a similar consideration for V_2 shows that $q \geqslant p$ so that $p = q$. The matrices V_1 and V_2 are therefore square orthogonal matrices, which, along with the positivity of Φ_1 and Φ_2, imply that S_1 and S_2 are nonsingular.

To summarize, for the general homogeneous boundary data (2.9) the E energy is conserved only if: (1) A has an equal number of positive and negative eigenvalues, which trivially implies that; (2) A must be of even dimension; and (3) the matrices S_1 and S_2, which are square and of dimension $(r/2)$, must be expressible as

$$S_1 = \Phi_1^{-1} V_1 \Phi_2 \quad \text{and} \quad S_2 = \Phi_2^{-1} V_2 \Phi_1,$$

where Φ_1 and Φ_2 are arbitrary block diagonal symmetric positive definite matrices (whose block structure is that of D_1 and D_2, respectively), and V_1 and V_2 are orthogonal matrices of dimension $(r/2)$; which also implies that (4) S_1 and S_2 are nonsingular. In particular, if the eigenvalues of A are distinct, then D and therefore D_1, D_2, Φ_1 and Φ_2 are diagonal matrices.

For $r = 2$, the above conditions imply that the scalars S_1 and S_2 are not zero and in fact $|S_1 S_2| = 1$. Then, choosing $\phi_2 = |S_1|\,\phi_1$ will result in the conservation of E energy. For general dimension r, the boundary matrices S_1 and S_2 and coefficient matrices A for which the above four conditions are satisfied form a very restricted set. We consider next when the E energy is bounded by the initial E energy, i.e.

$$dE_E/dt \leqslant 0. \tag{2.12}$$

Then, from (2.10) the matrices

$$(S_1^T D_1 \Lambda_1 S_1 - D_2 \Lambda_2) \quad \text{and} \quad (S_2^T D_2 \Lambda_2 S_2 - D_1 \Lambda_1) \tag{2.13}$$

must be negative semidefinite if (2.12) is to hold for all w_1 and w_2. General conditions on A, S_1 and S_2 for which D_1 and D_2 exist such that (2.13) are negative semidefinite is an open question. However, a few cases can be easily analyzed.

First, consider the pure "supersonic" case in which $p = 0$ and $q = r$. Then (2.9) and (2.10) reduce to

$$w_2(t, 0) = 0 \quad \text{and} \quad dE_E/dt = -\tfrac{1}{2} w_2^T(t, 1) D_2 \Lambda_2 w_2(t, 1),$$

respectively, since the dimension of w_1 is zero. Then for any positive diagonal matrix D_2, (2.12) holds. Of course, this is a special case of prescribing purely characteristic data, i.e., $S_1 = S_2 = 0$ for which any arbitrary positive diagonal matrices D_1 and D_2 will yield (2.12). Of more interest is when the data at $x = 1$ is characteristic but the data at $x = 0$ is not, i.e., $S_1 = 0$ but $S_2 \neq 0$. In this case only the second matrix of (2.13) need be examined and by choosing the elements of D_1 large enough and the elements of D_2 small enough, (2.12) can be satisfied.

We observe that whenever D_1 and D_2 are arbitrary, as in two of the cases above, then the role of the E matrix is purely that of a symmetrizer. Then if A itself is symmetric, Q can be chosen to be orthogonal, i.e., $QQ^T = I$ so that choosing $D = I$ yields $E = I$. A further observation is that if any characteristic data is prescribed, say at $x = 1$, then S_1 is singular; and therefore, the E energy cannot be conserved.

If the E energy can be bounded by the initial E energy,

$$(u, Eu) \leqslant (u, Eu)|_{t=0},$$

then the E matrix can be used to obtain a priori bounds on the L^2 energy since with E positive definite

$$(u, Eu)/\lambda_{\max} \leqslant (u, u) \leqslant (u, Eu)/\lambda_{\min}, \tag{2.14}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum eigenvalues of E, respectively.

In order to indicate what is the physical meaning of the E energy, consider the special case of A given by (2.3) and the boundary data (2.4). Then, with

$$E = \begin{pmatrix} 1 & -\nu \\ -\nu & 1 \end{pmatrix}$$

the E energy is conserved, i.e.,

$$\frac{1}{2}(u, Eu)_t = \frac{1}{2} \frac{\partial}{\partial t} \int_0^1 [u_1^2 - 2\nu u_1 u_2 + u_2^2]\, dx = 0 \tag{2.15}$$

which can be rewritten

$$\frac{\partial}{\partial t} \int_0^1 \frac{1}{2} [(1 - \nu^2) u_1^2 + (u_2 - \nu u_1)^2]\, dx = 0. \tag{2.16}$$

Now let

(2.17) $$u_1 = \psi_t/(1-\nu^2)^{1/2} \quad \text{and} \quad u_2 = [\psi_t + (\nu^2-1)\psi_x]/(1-\nu^2)^{1/2}.$$

Substituting (2.17) in the system, (2.1) with A given by (2.3) yields:

(2.18) $$\psi_{tt} + 2\nu\psi_{tx} + \nu^2\psi_{xx} - \psi_{xx} = 0,$$

which is a Galilean transformation of the wave equation. Then, from (2.18)

$$0 = \int_0^1 \psi_t[\psi_{tt} + 2\nu\psi_{tx} + (\nu^2-1)\psi_{xx}]\,dx$$

$$= \frac{\partial}{\partial t}\int_0^1 \frac{1}{2}[(\psi_t)^2 + (1-\nu^2)(\psi_x)^2]\,dx + \nu(\psi_t)^2|_0^1 + (\nu^2-1)\psi_t\psi_x|_0^1.$$

Using (2.4) and (2.17) to evaluate the boundary contributions yields

(2.19) $$\frac{\partial}{\partial t}\int_0^1 \frac{1}{2}[(\psi_t)^2 + (1-\nu^2)(\psi_x)^2]\,dx = 0.$$

Again using (2.17), it is obvious that (2.16) and (2.19) are identical. Therefore the E energy for the system (2.1) is the usual energy for (2.18).

III. Approximate Solution by Semidiscrete Galerkin Methods. Define the weak solution of the system (2.1) as that u which satisfies

(3.1) $$(u_t - Au_x, v) = 0$$

for all v in an appropriate test space and define the weak solution in the E inner product as that u which satisfies

(3.2a) $$(u_t - Au_x, Ev) = 0.$$

Since E is symmetric, this is equivalent to

(3.2b) $$(Eu_t - EAu_x, v) = 0.$$

Therefore, the effect of the symmetrizing matrix E may be interpreted as either changing the trial function from v to Ev or as premultiplying the system (2.1).

An approximate solution for u is to be found by the use of Galerkin's method. The method is presented for the case of (2.1) consisting of two equations, i.e.,

(3.3) $$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \qquad u = (u_1\ u_2)^T.$$

The generalization to higher dimensions is discussed below. In addition, u is assumed to satisfy the boundary data (2.4). This particular set of boundary data is chosen so that the E energy is conserved which is accomplished by choosing

$$E = \begin{pmatrix} e_{11} & e_{12} \\ e_{12} & e_{22} \end{pmatrix}, \qquad \begin{aligned} e_{12} &= -\sigma a_{22}/a_{12}, \\ e_{22} &= \sigma, \\ e_{11} &= \sigma(a_{22}^2 - a_{22}a_{11} + a_{21}a_{12})/a_{12}^2, \end{aligned}$$

where σ is a positive constant and then

$$EA = \begin{pmatrix} \alpha & \beta \\ \beta & 0 \end{pmatrix}. \tag{3.4}$$

This clearly marks any unbounded growth in the numerical solution to be an instability of the numerical method. For more general boundary data, the system (2.1) can have exponentially growing solutions and still be well posed (in the sense of Kreiss [6]). This makes it difficult to separate growths due to instabilities of the numerical scheme from those which are due to the time instability of the continuous solution. A further reason for choosing the data (2.4) is that it will adequately display some of the pitfalls encountered in applying Galerkin's method to the system (2.1). In particular, the important role of the matrix E in the computational method will be evident.

First, it is assumed that u can be approximated by

$$u^h = \begin{pmatrix} u_1^h \\ u_2^h \end{pmatrix} = \sum_{j=0}^{n} \begin{pmatrix} b_j(t) \\ c_j(t) \end{pmatrix} \phi_j(x), \tag{3.5}$$

where the basis functions $\phi_j(x)$ are chosen (without loss of generality) so that

$$\phi_j(0) = 0, \quad j \neq 0 \quad \text{and} \quad \phi_j(1) = 0, \quad j \neq n.$$

Then the boundary data (2.4) yields that

$$b_0(t) = b_n(t) = 0;$$

and therefore, the unknowns are b_j $(j = 1, \ldots, n-1)$ and c_j $(j = 0, \ldots, n)$ which number $2n$ in total. Initial data is prescribed for u_1 and u_2. Then initial data for all b_j and c_j can be deduced by solving the interpolation problem resulting from evaluating (3.5) at $t = 0$.

The expression (3.5) is substituted into (3.1) so that u^h is required to satisfy

$$(u_t^h - Au_x^h, v^h) = 0 \tag{3.6}$$

for all v^h in a suitable trial space. An obvious choice for v^h is that it lie in that subspace of the approximating space used to approximate u (in (3.5)) for which $v_1^h(x = 0) = v_1^h(x = 1) = 0$. Such a space, of dimension $2n$, is spanned by the vectors

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \phi_l, \quad l = 1, \ldots, n-1, \qquad \begin{pmatrix} 0 \\ 1 \end{pmatrix} \phi_l, \quad l = 0, \ldots, n. \tag{3.7}$$

Choosing (3.7) for v^h and then substituting in (3.6) yields the following $2n$ ordinary differential equations for $2n$ unknowns

$$\sum_{j=0}^{n} [\dot{b}_j m_{lj} + (a_{11} b_j + a_{12} c_j) k_{lj}] = 0, \qquad l = 1, \ldots, n-1,$$

$$\sum_{j=0}^{n} [\dot{c}_j m_{lj} + (a_{21} b_j + a_{22} c_j) k_{lj}] = 0, \qquad l = 0, \ldots, n,$$

where $\cdot$ indicates differentiation with respect to time and where

$$m_{lj} = \int_0^1 \phi_j(x)\phi_l(x)\,dx \quad \text{and} \quad k_{lj} = \int_0^1 \phi_j'(x)\phi_l(x)\,dx. \tag{3.8}$$

Since b_0 and b_n are known, they may be separated from the summations to yield a square system. Let

$$U = (b_1, \ldots, b_{n-1}, c_0, \ldots, c_n)^T,$$

$$\bar{M} = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix}, \qquad \bar{K} = \begin{pmatrix} K_1 & K_3 \\ K_4 & K_2 \end{pmatrix}, \tag{3.9}$$

where

$$M_1 = (m_{lj}), \qquad K_1 = a_{11}(k_{lj}), \qquad j, l = 1, \ldots, n-1,$$

$$M_2 = (m_{lj}), \qquad K_2 = a_{22}(k_{lj}), \qquad j, l = 0, \ldots, n,$$

$$K_3 = a_{12}(k_{lj}), \qquad j = 0, \ldots, n, l = 1, \ldots, n-1,$$

$$K_4 = a_{21}(k_{lj}), \qquad j = 1, \ldots, n-1, l = 0, \ldots, n.$$

Then the system (3.7) may be written in matrix notation as

$$\bar{M}\dot{U} = \bar{K}U. \tag{3.10}$$

The matrix $\bar{M}$ is clearly symmetric since $m_{lj} = m_{jl}$. For the special case of symmetric A, $\bar{K}$ is almost skew symmetric. In fact, since by integrating by parts

$$k_{lj} = -k_{jl} + \phi_j(1)\phi_l(1) - \phi_j(0)\phi_l(0),$$

it is clear that

$$\begin{aligned} k_{lj} &= -k_{jl}, \qquad l \neq j, l, j = 0, \ldots, n, \\ k_{jj} &= 0, \qquad j = 1, \ldots, n-1, \end{aligned} \tag{3.11}$$

so that (with $a_{12} = a_{21}$) $K_4 = -K_3^T$, K_1 is skew symmetric and K_4 is skew symmetric except for the two nonzero diagonal entries $(a_{22}k_{00})$ and $(a_{22}k_{nn})$. Then $\bar{K}$ itself is skew symmetric except for the two nonzero diagonal elements of K_4.

For the system (3.10), the stability in time is determined by the eigenvalues λ of

$$\lambda\bar{M}y = \bar{K}y. \tag{3.12}$$

The eigenvalues of (3.12) were computed for the special case of A given by (2.3) where the basis functions $\phi_j(x)$ were chosen to be cubic B-splines on a uniform mesh. Computations were performed for various values of n ranging from 5 to 100. The computational results show that:

1. The eigenvalues of $\bar{K}$ have zero real part.

2. $(2n-4)$ eigenvalues of (3.12) have zero real part, 2 have negative real part, and 2 have positive real part.

3. If h is the mesh size, then the real parts of the eigenvalues with positive real part grow as $1/h$ as h tends to zero.

From this it can be concluded that the solution of (3.10) will grow as $\exp\{\kappa t/h\}$ as t increases, where κ is a positive constant. This growth is clearly an instability of the numerical scheme since the continuous solution is bounded in time due to its conserved E energy and (2.14). Also, note that as h tends to zero, the instability in time is intensified.

Since these computational results indicate that for a representative approximation space the obvious Galerkin scheme considered above yields an unstable (in time) approximation to the solution of (2.1), this scheme is now abandoned. However, before proceeding to a different scheme, there are a few observations to be made.

First, the reason that the above scheme is unstable is the manner in which the boundary is being treated. This is made clear by observing that the right-hand side of

$$\text{(3.13)} \qquad \tfrac{1}{2}(U^T\bar{M}U)_t = U^T(\bar{K} + \bar{K}^T)U = a_{22}[k_{00}c_0^2(t) + k_{nn}c_n^2(t)]$$

involves terms which are affected by the boundary. In fact, from (3.5) one immediately deduces that

$$c_0(t) \approx u_2(t, 0)/\phi_0(0) \quad \text{and} \quad c_n(t) \approx u_2(t, 1)/\phi_n(1)$$

so that c_0 and c_n are clearly boundary terms.

A second observation is to note that the scheme considered in this section is stable in time for the wave equation. In fact, whenever $a_{12} = a_{21}$ and $a_{22} = 0$, $K_4 = 0$ so that $\bar{K}$ is skew symmetric. Then from (3.13) the $\bar{M}$ energy is bounded by the initial M energy, i.e. $(U^T\bar{M}U)_t = 0$. Therefore, so long as $\|\bar{M}\|$ and $\|\bar{M}^{-1}\|$ are bounded as h tends to zero (which is shown in the Appendix), the method discussed in this section is stable for $a_{22} = 0$ and the boundary data (2.4). (The norm $\|\cdot\|$ is assumed to be the Euclidean norm.)

IV. The Galerkin Scheme Using the E Matrix. As was pointed out in Section III, the role of the E matrix in the search for weak solutions of (2.1) may be interpreted in two ways. Of course, this dual role extends to the role of E in the Galerkin approximation. Therefore, after approximating u by (3.5), one may choose v^h as in Section III and substitute into (3.2b) or alternately one may choose $\{Ev^h\}$ to be the trial space and substitute into (3.2a). The trial space $\{Ev^h\}$ is spanned by

$$\text{(4.1)} \qquad \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix}\phi_l(x), \quad l = 1, \ldots, n-1, \quad \begin{pmatrix} e_{21} \\ e_{22} \end{pmatrix}\phi_l(x), \quad l = 0, \ldots, n.$$

Choosing either approach again results in a system of ordinary differential equations which may be written in matrix notation as

$$\text{(4.2)} \qquad M\dot{U} = KU,$$

where

$$M = \begin{pmatrix} M_1 & M_3 \\ M_3^T & M_2 \end{pmatrix}, \qquad K = \begin{pmatrix} K_1 & K_3 \\ -K_3^T & 0 \end{pmatrix},$$

$$M_1 = e_{11}(m_{lj}), \qquad K_1 = \alpha(k_{lj}), \qquad j, l = 1, \ldots, n-1,$$

$$M_2 = e_{22}(m_{lj}), \qquad j, l = 0, \ldots, n,$$

$$M_3 = e_{12}(m_{lj}), \qquad K_3 = \beta(k_{lj}), \qquad j = 0, \ldots, n, l = 1, \ldots, n-1,$$

where α and β are defined in (3.4). Clearly M is symmetric and, using (3.11), K is skew symmetric. Note that the symmetry of E is necessary for M to be symmetric, and that K is skew symmetric because EA is symmetric and has a zero element in the lower right-hand corner (see (3.4)). This shows how the properties of E enter into the matrices appearing in the differential system (4.2). The fact that E is positive definite is used in the Appendix to show that M is also positive definite.

The discrete M energy of the solution of the system (4.2) is conserved since

$$(U^T M U)_t = U^T (K + K^T) U = 0,$$

due to the skew symmetry of K. Therefore, if $\|M\|$ and $\|M^{-1}\|$ are bounded as the dimension of the approximating space becomes infinite, then the approximation U to u is a stable one. The question of bounding $\|M\|$ and $\|M^{-1}\|$ is discussed in the Appendix.

It is interesting to note that when M is symmetric and positive definite and K is skew symmetric, then the eigenvalues λ of

$$\lambda M y = K y \tag{4.3}$$

are pure imaginary since λ is also an eigenvalue of $M^{-1/2} K M^{-1/2}$ which is a skew symmetric matrix.

The Galerkin method presented in this section therefore yields a stable approximation U to the solution u of (2.1). It is important to note that the E matrix plays a crucial role in this method, as it did in obtaining bounds on the continuous energy. From the results concerning the method presented in Section III, it is seen that it is essential that the E matrix be used in the computations.

The matrix M, and therefore E, is intimately related to the Lyapunov stability theory for the system of ordinary differential equations (4.2). In the Lyapunov theory, one seeks a positive definite symmetric matrix H such that in the H norm, the energy can be bounded. In fact, for the system (4.2), which results from imposing the homogeneous boundary data (2.4), one actually seeks an H such that the H energy is conserved. This leads to the matrix problem

$$H(M^{-1}K) + (K^T M^{-1})H = 0,$$

which obviously has the solution $H = M$. Therefore, the matrix M is precisely the Lyapunov matrix for the system (4.2).

The Galerkin method presented above suffers from the serious drawback that M

is considerably less sparse than $\bar{M}$. Even if the unknowns are reordered so that the c_j's and b_j's interleave, the bandwidth of M will be double that of $\bar{M}$. This drawback is greatly alleviated by the method considered below.

The Galerkin Scheme Using E at the Boundary. Consider the trial space spanned by

$$\begin{pmatrix}1\\0\end{pmatrix}\phi_l(x), \qquad \begin{pmatrix}0\\1\end{pmatrix}\phi_l(x), \qquad l = 1, \ldots, n-1,$$

$$\begin{pmatrix}e_{21}\\ \\e_{22}\end{pmatrix}\phi_l(x), \qquad l = 0 \text{ and } n. \tag{4.4}$$

Clearly, the vectors (4.4) also span $\{Ev^h\}$ and therefore may be used instead of (4.1) in (3.2a). However, note that for $l \neq 0$ and $l \neq n$, the basis vectors (4.4) are identical to those used in Section III.

Choosing (4.4) as the trial space leads to the system

$$\hat{M}\dot{U} = \hat{K}U, \tag{4.5}$$

where

$$\hat{M} = \begin{pmatrix}M_1 & 0\\ \\M_3 & M_2\end{pmatrix}, \qquad \hat{K} = \begin{pmatrix}K_1 & K_3\\ \\K_4 & K_2\end{pmatrix},$$

$$M_1 = (m_{lj}), \qquad K_1 = a_{11}(k_{lj}), \qquad l, j = 1, \ldots, n-1,$$

$$K_3 = a_{12}(k_{lj}), \qquad l = 1, \ldots, n-1, j = 0, \ldots, n,$$

$$M_2 = (m_{lj}), \qquad \hat{K}_2 = a_{22}(k_{lj}), \qquad l, j = 0, \ldots, n,$$

$$K_2 = \hat{K}_2 - \begin{pmatrix}\epsilon_1\\ \hline 0\\ \hline \epsilon_{n+1}\end{pmatrix}\hat{K}_2, \qquad \epsilon_1 = (1, 0, \ldots, 0),\ \epsilon_{n+1} = (0, \ldots, 0, 1),$$

$$M_3 = \frac{e_{12}}{e_{22}}\begin{pmatrix}\xi\\ \hline 0\\ \hline \eta\end{pmatrix}, \qquad \xi = (\xi_1, \ldots, \xi_{n-1}),\ \eta = (\eta_1, \ldots, \eta_{n-1}),\ \xi_j = m_{0j},\ \eta_j = m_{nj},\ j = 1, \ldots, n-1, \tag{4.6}$$

$$\hat{K}_4 = a_{21}(k_{lj}), \qquad l = 0, \ldots, n, j = 1, \ldots, n-1,$$

$$K_4 = \hat{K}_4 + (\beta/e_{22} - a_{21})\begin{pmatrix}\bar{\xi}\\ \hline 0\\ \hline \bar{\eta}\end{pmatrix}, \qquad \bar{\xi} = (\bar{\xi}_1, \ldots, \bar{\xi}_{n-1}),\ \bar{\eta} = (\bar{\eta}_1, \ldots, \bar{\eta}_{n-1}),\ \bar{\xi}_j = k_{0j},\ \bar{\eta}_j = k_{nj},\ j = 1, \ldots, n-1.$$

The associated generalized eigenvalue problem is

$$\lambda \hat{M} y = \hat{K} y, \tag{4.7}$$

where now $\hat{M}$ is not symmetric and $\hat{K}$ is not skew symmetric.

The basis vectors (4.4) are a linear combination of the basis vectors (4.1) and the system (4.5) is a linear combination of the system (4.2). This observation makes it easy to define the symmetric matrix

$$R = \begin{pmatrix} R_1 & R_3^T \\ R_3 & R_2 \end{pmatrix}, \tag{4.8}$$

where

$$R_1 = e_{11} I_{n-1}, \qquad R_2 = e_{22} I_{n+1}, \qquad R_3 = e_{12} \begin{pmatrix} 0_{n-1} \\ \hline I_{n-1} \\ \hline 0_{n-1} \end{pmatrix},$$

and where I_j and 0_j are the identity matrix and zero vector of dimension j, respectively. Then

$$M = R\hat{M} \quad \text{and} \quad K = R\hat{K} \tag{4.9}$$

so that R simultaneously symmetrizes $\hat{M}$ and skew symmetrizes $\hat{K}$. By multiplying (4.7) on the left by R, (4.3) is recovered so that the eigenvalues of (4.7) are identical to those of (4.3). Therefore, the eigenvalues of (4.7) have zero real part.

The question of stability can be answered in the affirmative if a norm can be found in which the energy is conserved. That norm exists and is the M norm introduced above since

$$\begin{aligned} (U^T M U)_t &= U^T (M\hat{M}^{-1}\hat{K} + \hat{K}^T \hat{M}^{T-1} M) U \\ &= U^T (R\hat{M}\hat{M}^{-1}\hat{K} + \hat{K}^T \hat{M}^{T-1} \hat{M}^T R^T) U = U^T (K + K^T) U = 0. \end{aligned}$$

This also shows that the Lyapunov matrix for the system (4.5) is again M.

The role of E is, therefore, crucial to the computation only at the boundary; and E need be applied only there. This last Galerkin method is identical to the unstable method presented in Section III except for the treatment of the boundary, so once again it is evident that the stability of the Galerkin scheme is dependent on a correct treatment of the boundary. This is also clear from observing that the basis vectors (3.7) are not linear combinations of (4.1) as are the basis vectors (4.4). A further observation is that the matrix $\hat{M}$ is only slightly less sparse than $\bar{M}$. In fact, M and $\bar{M}$ are identical except for the nth and $2n$th rows. On the other hand, M has approximately twice as many nonzero entries as does $\bar{M}$.

The final step in proving the stability of either of the Galerkin schemes considered in this section is the bounding of $\|M\|$ and $\|M^{-1}\|$. This question is considered in the Appendix.

The mechanical construction of the above numerical methods extend in a straightforward manner from two to higher dimensions. In addition, the generalization to the general boundary data (2.9) is also easily accomplished, especially if the

system (2.1) is transformed to the diagonal (or characteristic) form $w_t = \Lambda w_x$ before the Galerkin methods are applied. However, the above stability analysis does not generalize so easily, especially when the E energy is not conserved but is merely bounded by the initial E energy. Part of the difficulty is finding an M norm for which, as the case may be, the discrete M energy is conserved or is bounded by the initial M energy. This difficulty is analogous to the search for the corresponding E matrix in the continuous problem. An additional difficulty in the case of the E energy being merely bounded is that since the search for the M matrix is aimed at bounding the M energy by its initial value, the Lyapunov matrix H is now required to make $H(M^{-1}K) + (K^TM^{-1})H$ negative semidefinite. This, of course, is not a trivial extension of the analysis of Sections III and IV. However, the rather specialized example treated in those sections clearly shows how essential the proper treatment of the boundary is to the stability of Galerkin solutions of hyperbolic systems.

Appendix—Properties of the Matrices M, $\bar{M}$ and $\hat{M}$. In Section IV it was required that $\|M\|$ and $\|M^{-1}\|$ be bounded. This problem for M can be reduced to the equivalent problem for $\bar{M}$ as follows. From (4.9), $M = R\hat{M}$. Furthermore, by comparing (3.9) and (4.6) it is clear that $\hat{M}$ and $\bar{M}$ have the same eigenvalues; and therefore, $\hat{M}$ is positive definite. In fact

$$\hat{M} = P\bar{M}P^{-1},$$

where

$$P = \begin{pmatrix} I_{n-1} & 0 \\ W & I_{n+1} \end{pmatrix}, \qquad W = -\frac{e_{12}}{e_{22}} \begin{pmatrix} 0_{n-1} \\ \hline I_{n-1} \\ \hline 0_{n-1} \end{pmatrix}.$$

Then

$$P^{-1} = \begin{pmatrix} I_{n-1} & 0 \\ -W & I_{n+1} \end{pmatrix},$$

and

$$M = RP\bar{M}P^{-1} \quad \text{and} \quad M^{-1} = P\bar{M}^{-1}P^{-1}R^{-1}$$

so that

$$\text{(A.1)} \qquad \|M\| \leqslant \|R\|\,\|P\|\,\|\bar{M}\|\,\|P^{-1}\| \quad \text{and} \quad \|M^{-1}\| \leqslant \|P\|\,\|\bar{M}^{-1}\|\,\|P^{-1}\|\,\|R^{-1}\|.$$

First, consider the symmetric matrix

$$PP^T = \begin{pmatrix} I_{n-1} & W^T \\ W & WW^T + I_{n+1} \end{pmatrix},$$

whose eigenvalues λ_p are

$$\lambda_p = \begin{cases} 1, & \text{twice}, \\ [2 + s^2 \pm s\sqrt{s^2 + 4}]/2, & (n-1) \text{ times}, \end{cases}$$

where $s = -e_{11}/e_{22}$. These eigenvalues are independent of n and are bounded from above and below by positive constants which are independent of n. Hence,

$$\|P\| \, \|P^{-1}\| \leqslant C_1, \tag{A.2}$$

where C_1 does not depend on n.

Next consider the symmetric matrix R which is defined by (4.8). The $2n$ eigenvalues λ_R of R can easily be shown to be

$$\lambda_R = \begin{cases} 1, & \text{twice,} \\ \text{eigenvalues of } E, & (n-1) \text{ times.} \end{cases}$$

Since E is positive definite, so is R. Since both R and $\hat{M}$ are positive definite, then so is M. Furthermore, the eigenvalues of R are bounded from above and below by positive constants which are independent of n. Since R is symmetric, this implies that

$$\|R\| \leqslant C_2 \quad \text{and} \quad \|R^{-1}\| \leqslant C_3, \tag{A.3}$$

where C_3 and C_4 do not depend on n. Note that the positive definiteness of E is necessary for R, and therefore M, to be positive definite.

Substituting (A.2) and (A.3) into (A.1) yields

$$\|M\| \leqslant C_4 \|\bar{M}\| \quad \text{and} \quad \|M^{-1}\| \leqslant C_5 \|\bar{M}^{-1}\|, \tag{A.4}$$

where $C_4 = C_1 C_2$ and $C_5 = C_1 C_3$ are positive constants independent of n. Therefore, the problem of bounding $\|M\|$ and $\|M^{-1}\|$ has been reduced to bounding $\|\bar{M}\|$ and $\|\bar{M}^{-1}\|$, respectively.

The matrix $\bar{M}$ is defined by (3.9); and therefore, $\|\bar{M}\|$ will depend on the choice of interpolating space. If the ϕ_j's are orthonormal, then $\bar{M} = I$ so that $\|\bar{M}\| = \|\bar{M}^{-1}\| = I$; and trivially these norms are bounded by constants independent of n.

The Galerkin approximation can be accomplished by the use of finite elements. Then typically the basis functions $\phi_j(x)$ are polynomials with compact support which in turn results in the Gram matrices M_1 and M_2 being banded. Furthermore, M_1 and M_2 can be assembled from element mass matrices. Following Fried [5], it can be shown that

$$\lambda_{\max}(\bar{M}) \leqslant s\mu_{\max} \quad \text{and} \quad \lambda_{\min}(\bar{M}) \geqslant \mu_{\min},$$

where $\mu_{\max}$ and $\mu_{\min}$ are the largest and smallest eigenvalues of the element mass matrices, respectively, and s is the maximum number of nonzero elements in a row of $\bar{M}$. In general, s is independent of n. The elements of $\bar{M}$ are proportional to $h \sim 1/n$, the measure of the grid size. However, this h scales the whole matrix $\bar{M}$ (and also $\hat{M}$ and M) and dividing equations such as (3.10), (4.2), and (4.5) by h scales the K matrices by $1/h$. This does not affect any of the results of Sections III and IV, but does make $\mu_{\max}$ and $\mu_{\min}$ independent of n, and therefore $\lambda_{\max}$ and $\lambda_{\min}$ can be bounded from above and below, respectively, by positive constants which are independent of n. Then since $\bar{M}$ is symmetric, $\|\bar{M}\|$ and $\|\bar{M}^{-1}\|$ are bounded by positive constants which are independent of n, and then by (A.4), the same is true for $\|M\|$ and $\|M^{-1}\|$.

Acknowledgement. The author gratefully acknowledges the aid, both in substance and in spirit, of Dr. David Gottlieb and Dr. James Ortega.

Mathematics Department
The University of Tennessee
Knoxville, Tennessee 37916

1. G. A. BAKER, "A finite element method for first order hyperbolic equations," *Math. Comp.*, v. 29, 1975, pp. 995–1006.

2. J. E. DENDY, "Two methods of Galerkin type achieving optimum L^2 rates of convergence for first order hyperbolics," *SIAM J. Numer. Anal.*, v. 11, 1974, pp. 637–653. MR **50** # 6178.

3. T. DUPONT, "Galerkin methods for first order hyperbolics: an example," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 890–899. MR **50** # 1540.

4. T. DUPONT, "A Galerkin method for liquid pipelines." (To appear.)

5. I. FRIED, "Condition of finite element matrices generated from nonuniform meshes," *AIAA J.*, v. 10, 1972, pp. 219–221.

6. H.-O. KREISS & J. OLIGER, *Methods for the Approximate Solution of Time Dependent Problems*, Global Atmospheric Research Programme, Publications Series, no. 10, Geneva, 1973.

7. O. TAUSSKY, "Positive-definite matrices and their role in the study of the characteristic roots of general matrices," *Advances in Math.*, v. 2, 1968, pp. 175–186. MR **37** # 2785.

8. L. WAHLBIN, "A modified Galerkin procedure with Hermite cubics for hyperbolic problems," *Math. Comp.*, v. 29, 1975, pp. 978–984. MR **52** # 9643.