Diagonalization and Simultaneous Symmetrization of the Gas-Dynamic Matrices

By R. F. Warming, Richard M. Beam and B. J. Hyett

Abstract. The hyperbolicity of the unsteady, inviscid, gas-dynamic equations implies the existence of a similarity transformation which diagonalizes an arbitrary linear combination \( \sum k_j A_j \) of coefficient matrices \( A_j \). The matrix \( T \) that accomplishes this transformation is given explicitly, and the spectral norms of \( T \) and \( T^{-1} \) are computed. It is also shown that the individual matrices \( A_j \) are simultaneously symmetrized by the same similarity transformation. Applications of the transformations \( T \) and \( T^{-1} \) and their norms include the well-posedness of the Cauchy problem, linear stability theory for finite-difference approximations, construction of difference schemes based on characteristic relations, and simplification of the solution of block-tridiagonal systems that arise in implicit time-split algorithms.

1. Introduction. The inviscid gas-dynamic equations in nonconservative form in three space variables can be written as

\[
\frac{\partial u}{\partial t} + \sum_{j=1}^{3} A_j \frac{\partial u}{\partial x_j} = 0,
\]

where the vector \( u \) and the matrices \( A_j \) are

\[
u = \begin{bmatrix} \rho \\ u_1 \\ u_2 \\ u_3 \\ p \end{bmatrix}, \quad A_1 = \begin{bmatrix} \rho & 0 & 0 & 0 \\ 0 & u_1 & 0 & 1/\rho \\ 0 & 0 & u_1 & 0 \\ 0 & 0 & 0 & u_1 \end{bmatrix},
\]

\[
A_2 = \begin{bmatrix} u_2 & 0 & \rho & 0 & 0 \\ 0 & u_2 & 0 & 0 & 0 \\ 0 & 0 & u_2 & 0 & 1/\rho \\ 0 & 0 & 0 & u_2 & 0 \\ 0 & 0 & \rho c^2 & 0 & u_2 \end{bmatrix}, \quad A_3 = \begin{bmatrix} u_3 & 0 & 0 & \rho & 0 \\ 0 & u_3 & 0 & 0 & 0 \\ 0 & 0 & u_3 & 0 & 0 \\ 0 & 0 & 0 & u_3 & 1/\rho \\ 0 & 0 & 0 & \rho c^2 & u_3 \end{bmatrix}.
\]

The elements of \( u \) are density, velocity components, and pressure and \( c = c(p, \rho) \) is the local sound speed. Associated with (1.1) is a matrix \( P \) defined by

\[
P = \sum_{j=1}^{3} k_j A_j = \sum_{j=1}^{3} k_j A_j (u), \quad -\infty < k_j < \infty.
\]
The quasi-linear system (1.1) is called hyperbolic at the point $(x, t, u)$ if there exists a non-singular matrix $T(k)$ such that

$$T^{-1}PT = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix},$$

where the eigenvalues $\lambda_j$ of $P$ are real and the norms of $T$ and $T^{-1}$ are uniformly bounded in $k = (k_1, k_2, k_3)^T$; i.e.,

$$\|T\|, \|T^{-1}\| \leq K.$$ 

In Section 2 of this paper, we give explicit formulas for the matrices $T$ and $T^{-1}$. These matrices are rather simple in appearance and, by careful normalization of the columns of $T$, we have found that both $TT^T$ and its inverse are sparse ($T^T$ denotes the transpose of $T$). Consequently, the spectral norms of $T$ and $T^{-1}$ are readily computed and we give explicit formulas for them. In addition, the similarity transformation $T^{-1}A_jT$ simultaneously symmetrizes the matrices $A_j$ for arbitrary $k$ (Section 3). The importance of symmetrizing the coefficient matrices has been discussed in a recent paper by Türkei [14].

The relation between the matrices $A_j$ of the nonconservation form (1.1) and the Jacobian matrices of the conservation law form is given in Section 4. In fact, these matrices are similar [12, p. 363] and the transformation is given explicitly.

Finally, in the last section, several applications of the transformations are indicated.

2. Bounded Transformation Diagonalizing $P$. The eigenvalues of $P$ are easily found to be

$$\lambda_1 = \lambda_2 = \lambda_3 = k \cdot u,$$

$$\lambda_4 = k \cdot u + c(k \cdot k)^{1/2}, \quad \lambda_5 = k \cdot u - c(k \cdot k)^{1/2},$$

where

$$k \cdot k = \sum_{j=1}^{3} k_j^2, \quad k \cdot u = \sum_{j=1}^{3} k_j u_j.$$

Although the matrix $P$ has an eigenvalue of multiplicity three, it has a complete set of linearly independent eigenvectors and, consequently, the matrix $T$ is constructed using these eigenvectors as columns. Hence, we find

$$T = \begin{bmatrix} \tilde{k}_1 & \tilde{k}_2 & \tilde{k}_3 & \rho/(\sqrt{2}c) & \rho/(\sqrt{2}c) \\ 0 & -\tilde{k}_3 & \tilde{k}_2 & \tilde{k}_1/\sqrt{2} & -\tilde{k}_1/\sqrt{2} \\ \tilde{k}_3 & 0 & -\tilde{k}_1 & \tilde{k}_2/\sqrt{2} & -\tilde{k}_2/\sqrt{2} \\ -\tilde{k}_2 & \tilde{k}_1 & 0 & \tilde{k}_3/\sqrt{2} & -\tilde{k}_3/\sqrt{2} \\ 0 & 0 & 0 & \rho c/\sqrt{2} & \rho c/\sqrt{2} \end{bmatrix},$$

where $\tilde{k}_j = k_j/(k \cdot k)^{1/2}$. The $j$th column of $T$ is a right eigenvector corresponding to the eigenvalue $\lambda_j$. The inverse of $T$ was formed from the left eigenvectors of $P$.
The determinants of $T$ and $T^{-1}$ are

$$\text{det } T = (\text{det } T^{-1})^{-1} = \rho c. \tag{2.4}$$

The spectral norm of the matrix $T$ can be computed from the formula

$$\|T\| = [r(TT^t)]^{1/2}, \tag{2.5}$$

where $r(TT^t)$ is the spectral radius of $TT^t$. The eigenvectors comprising the matrix $T$ as given by (2.2) were judiciously normalized so that the product $TT^t$ led to the sparse matrix

$$TT^t = I + Q, \tag{2.6}$$

where $I$ is the identity matrix and $Q$ is a matrix whose elements are all zero except for the four corner elements:

$$q_{11} = (\rho/c)^2, \quad q_{15} = q_{51} = \rho^2, \quad q_{55} = \rho^2 c^2 - 1. \tag{2.7}$$

The singular values of $T$ (i.e., the eigenvalues $\sigma$ of $TT^t$) are readily found to be

$$\sigma = 1, 1, 1, \quad \phi \pm (\phi^2 - 4\rho^2 c^6)^{1/2}, \quad \phi = \rho^2 + c^2 + \rho^2 c^4, \tag{2.8}$$

and consequently,

$$\|T\|^2 = \max \sigma = \phi + (\phi^2 - 4\rho^2 c^6)^{1/2}/2c^2 > 1. \tag{2.9}$$

Likewise, the product $(T^{-1})^tT^{-1}$ leads to a matrix with the following simple structure:

$$(T^{-1})^tT^{-1} = I + R, \tag{2.10}$$

where $R$ is a matrix whose elements are all zero except for the three corner elements:

$$r_{15} = r_{51} = -1/c^2, \quad r_{55} = \rho^2 c^4/(2\rho^2 c^4). \tag{2.11}$$

The eigenvalues $\kappa$ of $(I + R)$ are

$$\kappa = 1, 1, 1, \quad \phi \pm (\phi^2 - 4\rho^2 c^6)^{1/2}/2\rho^2 c^4, \tag{2.12}$$

and hence,

$$\|T^{-1}\|^2 = \max \kappa = \phi + (\phi^2 - 4\rho^2 c^6)^{1/2}/2\rho^2 c^4 > 1. \tag{2.13}$$

Finally, we note that

$$\|T\| = (\det T)\|T^{-1}\|. \tag{2.14}$$

Since the norms of $T$ and $T^{-1}$ are independent of the real parameters $k_j$, they are uniformly bounded in $k$ as required by (1.4).
In a curvilinear coordinate system, the first-order system (1.1) is altered by the addition of a nonhomogeneous term, and the matrices $A_j$ are replaced by $A_j/h_j$ where $h_j$ is a scale factor. The transformations of this section are still valid since the $k_j$'s are arbitrary real numbers which can be redefined to absorb the scale factors.

3. Symmetrization of the Matrices $A_j$. The matrix $T$, given by (2.2), was constructed to diagonalize the matrix $P$ defined by (1.2). In addition, this same transforming matrix $T$ simultaneously symmetrizes the individual matrices $A_j$. This result can be written in partitioned form as follows:

\[(3.1)\]

\[T^{-1}A_jT = \begin{bmatrix} D_j & B_j \\ B_j^t & C_j \end{bmatrix}, \quad j = 1, 2, 3,\]

where $D_j$ and $C_j$ are the diagonal matrices

\[D_j = \begin{bmatrix} u_j & 0 & 0 \\ 0 & u_j & 0 \\ 0 & 0 & u_j \end{bmatrix}, \quad C_j = \begin{bmatrix} u_j + c\kappa_j & 0 \\ 0 & u_j - c\kappa_j \end{bmatrix},\]

and

\[B_1 = \frac{c}{\sqrt{2}} \begin{bmatrix} 0 & 0 \\ \kappa_3 & \kappa_3 \\ \kappa_2 & \kappa_2 \end{bmatrix}, \quad B_2 = \frac{c}{\sqrt{2}} \begin{bmatrix} \kappa_3 & \kappa_3 \\ 0 & 0 \\ \kappa_1 & \kappa_1 \end{bmatrix}, \quad B_3 = \frac{c}{\sqrt{2}} \begin{bmatrix} -\kappa_2 & -\kappa_2 \\ \kappa_1 & \kappa_1 \\ 0 & 0 \end{bmatrix}.\]

It may be of interest to comment on how we obtained the above symmetric form. In general, a transformation $T$ which diagonalizes $P$ will not symmetrize the individual matrices $A_j$. In fact, the original $T$, say $T_0$, which we constructed did not symmetrize the $A_j$'s. But when we made the individual calculations $T_0^{-1}A_jT_0$, the results were nearly symmetric and it was more or less obvious how to renormalize the columns of $T_0$ to achieve the symmetric form (3.1). This same renormalization also led to a sparse form for $TT^t$.

A more direct approach would be to first find some particular matrix $S$ that would simultaneously symmetrize the $A_j$'s. Then $S^{-1}PS$ would be symmetric and, consequently, this matrix could be diagonalized by an orthogonal matrix $U$. If we define $T = SU$, then this matrix would simultaneously diagonalize $P$ and symmetrize the $A_j$'s. However, we are unaware of any systematic procedure of simultaneously symmetrizing a set of noncommuting matrices such as those defined by Eq. (1.1).

The symmetric form (3.1) is, of course, not unique, since any orthogonal similarity transformation on (3.1) yields a symmetric matrix. For example, if $U$ is the $5 \times 5$ orthogonal matrix

\[U = \begin{bmatrix} I_3 & 0 \\ 0 & \sqrt{2} \end{bmatrix}, \quad V = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix},\]

then
where $E_j = u_j I_4$, $I_n$ is the $n \times n$ identity matrix, and

$$H_1^j = c[0, -\kappa_3, \kappa_2, \kappa_1], \quad H_2^j = c[\kappa_3, 0, -\kappa_1, \kappa_2], \quad H_3^j = c[-\kappa_2, \kappa_1, 0, \kappa_3].$$

4. Transformations for Conservation Law Matrices. The gas-dynamic equations (1.1) can also be written in conservation law form as

$$\frac{\partial \tilde{u}}{\partial t} + \sum_{j=1}^{3} \frac{\partial F_j}{\partial x_j} = 0,$$

where

$$\tilde{u} = \begin{bmatrix} \rho \\ \rho u_1 \\ \rho u_2 \\ \rho u_3 \\ e_T \end{bmatrix}, \quad F_j = \begin{bmatrix} \rho u_j \\ \rho u_1 u_j + p \delta_{1j} \\ \rho u_2 u_j + p \delta_{2j} \\ \rho u_3 u_j + p \delta_{3j} \\ u_j (e_T + p) \end{bmatrix},$$

$$e_T = \frac{\gamma p}{\gamma - 1} + \frac{\rho}{2} q^2, \quad q^2 = \sum_{j=1}^{3} u_j^2,$$

$\delta_{ij}$ is the Kronecker delta and $\gamma$ is the ratio of specific heats. Eq. (4.1) can be reexpressed as

$$\frac{\partial \tilde{u}}{\partial t} + \sum_{j=1}^{3} \tilde{A}_j \frac{\partial \tilde{u}}{\partial x_j} = 0,$$

where the $\tilde{A}_j$'s are the Jacobian matrices $\partial F_j/\partial \tilde{u}$.

The matrices $A_j$ of the nonconservative form (1.1) and the matrices $\tilde{A}_j$ of the conservative form (4.1) are related by the similarity transformation

$$A_j = M^{-1} \tilde{A}_j M,$$

where $M$ is the Jacobian matrix $\partial \tilde{u}/\partial u$. Computing the requisite derivatives, we find

$$(4.4) \quad M = \frac{\partial \tilde{u}}{\partial u} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ u_1 & \rho & 0 & 0 & 0 \\ u_2 & 0 & \rho & 0 & 0 \\ u_3 & 0 & 0 & \rho & 0 \\ q^2/2 & \rho u_1 & \rho u_2 & \rho u_3 & 1/(\gamma - 1) \end{bmatrix}$$

and

$$(4.5) \quad M^{-1} = \frac{\partial u}{\partial \tilde{u}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -u_1/\rho & 1/\rho & 0 & 0 & 0 \\ -u_2/\rho & 0 & 1/\rho & 0 & 0 \\ -u_3/\rho & 0 & 0 & 1/\rho & 0 \\ (\gamma - 1)/2 \cdot q^2 & - (\gamma - 1) u_1 & - (\gamma - 1) u_2 & - (\gamma - 1) u_3 & (\gamma - 1) \end{bmatrix}$$

Thus, for example,
\[ \mathbf{A}_1 = MA_1M^{-1} \]

\[
\begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
\frac{\gamma - 1}{2}q^2 - u_1^2 & -(\gamma - 3)u_1 & -(\gamma - 1)u_2 & -(\gamma - 1)u_3 & (\gamma - 1) \\
-u_1u_2 & u_2 & u_1 & 0 & 0 \\
-u_1u_3 & u_3 & 0 & u_1 & 0 \\
\tilde{a}_{51} & \tilde{a}_{52} & -(\gamma - 1)u_1u_2 & -(\gamma - 1)u_1u_3 & \gamma u_1 \\
\end{bmatrix},
\]

where

\[
\tilde{a}_{51} = [(\gamma - 1)u_1q^2 - \gamma e_Tu_1/\rho], \quad \tilde{a}_{52} = [\gamma e_T/\rho - (\gamma - 1)(u_1^2 + q^2/2)].
\]

The matrices \( \mathbf{A}_j \) are quite complicated compared with the simple form of the matrices \( A_j \) of Eq. (1.1). If one is concerned with constructing the transformation which diagonalizes \( \mathbf{P} = \sum_{j=1}^{3} k_j \mathbf{A}_j \) or if one wants to symmetrize the matrices \( \mathbf{A}_j \), it is far simpler to use the similarity transform (4.3) and work directly with the \( A_j \)'s. For example, to symmetrize \( \mathbf{A}_j \), we write simply

\[ T^{-1}M^{-1}\mathbf{A}_jMT = T^{-1}A_jT, \]

and the symmetric result is given by the previous calculation (3.1).

5. Applications.

(a) The Cauchy Problem. It is known that if a first-order system

\[ \frac{\partial \mathbf{u}}{\partial t} + \sum_j A_j \frac{\partial \mathbf{u}}{\partial x_j} = 0 \]

(with constant coefficients) is hyperbolic, then the Cauchy problem is well-posed (see, e.g., [7]). It is of interest to see how the results of Section 2 fit into a more general theory developed by Kreiss [6]. He proved the following theorem: The Cauchy problem for the system (5.1) is well-posed if and only if there exist real constants \( C > 0 \) and \( \alpha \), and a positive definite Hermitian matrix \( H(k) \) for which

\[ C^{-1}I \leq H(k) \leq CI \]

and

\[ H(k)P(ik) + \hat{P}^*(ik)H(k) \leq 2\alpha H(k), \]

where

\[ \hat{P}(ik) = i \sum k_j A_j. \]

Here Kreiss' theorem has been specialized to the particular case of the system (5.1).

For the (linearized) gas-dynamic equations, a positive definite Hermitian matrix that fulfills conditions (5.2) is the sparse matrix given by (2.10). The constant \( C \) is the maximum of the eigenvalues (2.9) and (2.13), and \( \alpha \) is zero.

The general theory of well-posedness developed by Kreiss [6] includes Eq. (5.1) with variable coefficients where the matrices \( A_j = A_j(x, t) \) are smooth functions of \( x \) and \( t \). In the nonlinear case, \( A_j = A_j(u, x, t) \) and the Cauchy problem is only well-posed in a sufficiently small interval \( 0 \leq t \leq T \) [7], [12].
(b) **Stability Theory.** Linear stability theory for finite-difference approximations to the hyperbolic system (1.1) or (4.2) is based on linearized versions where the matrices $A_j$ (or $\tilde{A}_j$) are assumed to be constant. The simplest situation occurs when the amplification matrix $G$ [12] is a polynomial in the matrix $P$ defined by (1.2). In this case, $G$ is diagonalized by a similarity transformation where the matrix that accomplishes the similarity is given by (2.2). Since $T$ and $T^{-1}$ are bounded independent of $k$, the von Neumann condition is sufficient as well as necessary for stability. Examples of two-step, Lax-Wendroff methods where $G$ is a polynomial in $P$ are the algorithms proposed by Rubin and Preiser [13], [2] and Zwas [15].

In a recent paper, Türkei [14] gave a similarity transformation that symmetrizes the matrices $\tilde{A}_j$ (see Eq. (4.2)) for the case of two spatial dimensions. (The same transformation was given earlier by MacCormack [8].) The extension to three spatial dimensions is given by the similarity transformation (4.6). Türkei used the resulting sparse symmetric forms to simplify the linear stability analysis in situations where it is necessary to compute the spectral radius of $G^*G$ to provide a sufficient stability analysis.

(c) **Difference Schemes Based on Characteristic Relations.** There are several methods of constructing difference schemes for multidimensional hyperbolic systems which utilize characteristic relations in their derivation. The eigenvalues and eigenvectors of the matrix $P$, defined by (1.2), play an essential role in these methods. The method described by Magomedov and Kholodov [10] uses the eigenvalues of the matrices $A_j$ and the similarity transformations that diagonalize each $A_j$. For the gas-dynamic equations (1.1), the requisite eigenvalues and transformations are given by Eqs. (2.1), (2.2), and (2.3) by a proper choice of $k$.

Johnston and Pal [5] have derived a difference method which makes use of the bicharacteristics of the system. In this technique, one needs the eigenvalues of the matrix $P$. Appropriate derivatives of these eigenvalues yield the bicharacteristics of the system. The difference scheme then follows by an approximate integration along a bicharacteristic. The resulting difference schemes [5] are such that the amplification matrix is a linear function of $P$. Hence, the von Neumann condition is necessary and sufficient for stability, as described in the first paragraph of Section 5(b).

(d) **Implicit Split Algorithms.** Efficient explicit algorithms for the hyperbolic system (1.1) or (4.1) can be constructed with the aid of the method of fractional steps or time splitting (see, e.g., [9], [3]). There is some motivation for considering implicit methods because of the time-step limitations inherent in explicit algorithms. Although multidimensional implicit methods require the solution of large sparse matrices, the method of fractional steps provides an efficient solution procedure. For example, a time-split, implicit, (Crank-Nicolson) algorithm for the system (1.1) is the following:

\[(5.3a) \quad (L_+^1)u^{n+1/3} = (L_-^1)u^n,\]

\[(5.3b) \quad (L_+^2)u^{n+2/3} = (L_-^2)u^{n+1/3},\]

\[(5.3c) \quad (L_+^3)u^{n+1} = (L_-^3)u^{n+2/3},\]
where
\[(L^j) = [I \pm \Delta t A_j (D_0)_{x_j}]\]

and
\[(D_0)_{x_1} u = \frac{u_{j+1/2, i, j} - u_{j-1/2, i, j}}{2 \Delta x_1} ,\]

with similar definitions for \((D_0)_{x_2} , (D_0)_{x_3} .\)

In order that the above algorithm be uniformly second order in the time and spatial increments \(\Delta t, \Delta x_j ,\) the matrix coefficients \(A_j\) must be evaluated within some predictor-corrector sequence and the order in which the operators \(L^1, L^2, L^3 ,\) appear must be reversed when the solution is advanced in the next time step, i.e., \(n + 1 \rightarrow n + 2 .\)

We will not go into these details which are discussed elsewhere [11].

Each fractional step of the algorithm (5.3) requires the solution of a block-tridiagonal system whose submatrices are each of order 5. There is a well-known direct method [4] for solving tridiagonal matrices and a generalization for block tridiagonals. For the present application, \(5 \times 5\) matrices must be inverted in the sequence of steps required for the direct solution. However, we can take advantage of the fact that the matrices \(A_j\) are each diagonalized by the similarity transformation \(T^{-1} A_j T\) by appropriate choice of the parameters \(k_j .\) If the \(A_j\)'s were spatially constant, we could change variables by introducing \(w = T^{-1} u\) and obtain an uncoupled set of equations for the components of \(w .\) Otherwise, we can still effect a considerable simplification if we properly take into account the dependence of the \(A_j\)'s upon the vector of dependent variables \(u .\) For example, if we multiply Eq. (5.3a) by \(T^{-1}(k)\) with \(k^f = (1, 0, 0)\) and write out the resulting equations, we find we can invert the operator on the left of (5.3a) by solving one block-tridiagonal system with \(2 \times 2\) submatrices (which are easily inverted analytically) followed by three scalar tridiagonal inversions. An equivalent reduction can be achieved by reordering (multiplying by a permutation matrix) the elements of the vector \(u\) in each fractional step of algorithm (5.3). A more complete description of the algorithm and numerical computations for unsteady aerodynamic flows will be reported elsewhere [1].

Computational Fluid Dynamics Branch
Ames Research Center, NASA
Moffett Field, California 94035

1. R. BEAM & R. F. WARMING, "Implicit, time-split difference schemes for the transient solution of hyperbolic equations." (In preparation.)


