NASA Contractor Report 191436

ICASE Report No. 93-9

ICASE

TIME-STABLE BOUNDARY CONDITIONS FOR FINITE-DIFFERENCE SCHEMES SOLVING HYPERBOLIC SYSTEMS: METHODOLOGY AND APPLICATION TO HIGH-ORDER COMPACT SCHEMES

Mark H. Carpenter
David Gottlieb
Saul Abarbanel

NASA Contract Nos. NAS1-19480 and NAS1-18605
March 1993

Institute for Computer Applications in Science and Engineering
NASA Langley Research Center
Hampton, Virginia 23681-0001

Operated by the Universities Space Research Association

National Aeronautics and
Space Administration
Langley Research Center
Hampton, Virginia 23681-0001

TIME-STABLE BOUNDARY CONDITIONS FOR FINITE-DIFFERENCE SCHEMES SOLVING HYPERBOLIC SYSTEMS: METHODOLOGY AND APPLICATION TO HIGH-ORDER COMPACT SCHEMES

Mark H. Carpenter
Aerospace Engineer, Theoretical Flow
Physics Branch, Fluid Mechanics Division
NASA Langley Research Center
Hampton, VA. 23681-0001

David Gottlieb[1]
Division of Applied Mathematics
Brown University
Providence, RI 02912

Saul Abarbanel[1]
Department of Mathematical Sciences
Division of Applied Mathematics
Tel-Aviv University
Tel-Aviv, ISRAEL

ABSTRACT

We present a systematic method for constructing boundary conditions (numerical and physical) of the required accuracy, for compact (Pade-like) high-order finite-difference schemes for hyperbolic systems. First a proper summation-by-parts formula is found for the approximate derivative. A "simultaneous approximation term" (SAT) is then introduced to treat the boundary conditions. This procedure leads to time-stable schemes even in the system case. An explicit construction of the fourth-order compact case is given. Numerical studies are presented to verify the efficacy of the approach.

[1]This research was supported by the National Aeronautics and Space Administration under NASA Contract Nos. NAS1-18605 and NAS1-19480 while the second and third authors were in residence at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA 23681-0001.

Introduction

Emphasis on the long-time numerical integration of the fluid mechanics equations has increased in recent years. As a result, high-order spatially accurate schemes are favored, because of their lower phase error. Such schemes, although they are stable in the classical sense (Lax and G-K-S stability), may exhibit a non-physical growth in time. For a fixed time T, these schemes converge as the mesh size $\Delta x \rightarrow 0$. However, from a practical point of view, in order to achieve reasonable accuracy for large T, meshes much too fine for the computers available in the foreseeable future are required. Since long-time integrations are encountered in present day computations, it is important to devise schemes which are not only classically stable but also time-stable. Specifically, they do not allow a growth in time that is not called for by the differential equations.

To retain the formal accuracy of a high-order scheme, boundary closures must be accomplished with accuracies that are at most one order less than the interior scheme [1]. For the scalar explicit central-differencing case, Kreiss and Scherer [2] have presented a method for constructing a boundary condition of accuracy one order less than the inner scheme such that a generalized *summation-by-parts* property of the differential equation is preserved. Strand [3] has used their approach to construct in the scalar case, fourth- and sixth-order central-differencing schemes with boundary closures of the appropriate order such that the resulting expression for the derivative satisfies the summation-by-parts property. Recent attempts to utilize these boundary closures to numerically solve a 2×2 hyperbolic system have shown that, in certain cases, an unwarranted growth in time still results.

In reference [4], the stability characteristic of various compact fourth- and sixth-order spatial operators were assessed using the theory of Gustafsson, Kreiss and Sundstrom (G-K-S) [5] for the semidiscrete initial-boundary-value-problem (IBVP). This study showed that many of the higher order schemes that are G-K-S stable are not time stable. It was concluded that in practical calculations, only those schemes which satisfied both definitions of stability were of any usefulness for long time integrations. Of practical importance was a new sixth-order scheme with fifth-order boundary conditions which was shown to be G-K-S and time-stable. Recently, however, it has been found that most of the high-order schemes that were time-stable in the scalar case, exhibited time divergence when applied to a 2×2 system.

In this paper, we outline a systematic procedure for designing time-stable, as well as G-K-S stable schemes of high-order accuracy. The new schemes are guaranteed to be time-stable for any hyperbolic system (as long as the system has a bounded energy). The first step in this procedure is to construct an approximation to the first derivative (internal plus boundary points) that admits a summation-by-parts formula. We rely on the work of Strand [3] for high-order explicit formulations. For high-order compact schemes, we derive a new methodology for construction of such schemes.

Appendix I includes an exposition of the methodology, and a detailed example of the fourth-order compact central difference scheme with third-order boundary closures. In section 1, we discuss a scalar hyperbolic equation. We show that in general a summation-by-parts formula does not guarantee time stability. However, we introduce a new procedure for imposing boundary conditions (simultaneous approximation term, (SAT)), that solves a linear combination of the boundary conditions and the differential equations near the boundary. This technique is an extension of the techniques used in reference [6] to stabilize the pseudo-spectral Chebychev collocation method. It is shown that if the approximation of the derivative operator admits a summation-by-parts formula then the SAT method is stable in the classical sense and is also time-stable.

In section 2 we discuss the implementation of the SAT method to *systems* of hyperbolic equations. We show that also in the system case, time stability (as well as Lax stability) is assured by having a summation-by-parts property for the numerical derivative operator, provided that the SAT method is utilized.

In section 3 we present numerical results that confirm the efficacy of the SAT procedure even in the cases where previous attempts could not attain time stability. It is shown that the theoretical predictions for the time stability of the SAT method are realized in practice for both the scalar hyperbolic case and the 2×2 hyperbolic system. Finally, an optimization of the parameter τ (which arises in the SAT procedure) is performed, with regard to efficiency and accuracy.

1. The Scalar Case

We consider the scalar hyperbolic equation

$$\frac{\partial u}{\partial t} = \lambda \frac{\partial u}{\partial x} \qquad 0 \leq x \leq 1 \tag{1}$$

for which there exists the energy rate

$$\frac{d}{dt} \int_0^1 u^2(x,t)dx = \lambda(u^2(1,t) - u^2(0,t))$$

For positive λ, we have the boundary condition

$$u(1,t) = g(t)$$

We denote by $\mathbf{u}$ a vector of the unknowns $(u_0(t), u_1(t), ...u_N(t))$ which corresponds to grid points $x_0(=0), x_1, ...x_N(=1)$.

In this work, we deal primarily with compact schemes for the discretization of the spatial operator $\frac{\partial}{\partial x}$. For a compact spatial operator, the approximation to the first derivative can be written as

$$P \frac{d\mathbf{u}}{dx} = Q\mathbf{u} \tag{2}$$

where P and Q are $(N+1) \times (N+1)$ matrices. We further assume that:

Assumption I

(i) Equation (2) is accurate to order m. Specifically, if we denote by $\mathbf{v}$ the vector $(v(x_0,t),...,v(x_N,t)$ where $v(x,t) \in C^m$ and $x_j = j\Delta x = \frac{j}{N}$, and by $\mathbf{v_x}$ the values of $((\frac{\partial v}{\partial x})_0,...,(\frac{\partial v}{\partial x})_N)^T$ then

$$P\mathbf{v_x} - Q\mathbf{v} = P\mathbf{T_e}$$

where the truncation error T_e satisfies

$$|\mathbf{T_e}| = O(\Delta x)^m$$

(ii) The matrix P has a simple structure (preferably tridiagonal) and is easily invertible.

(iii) There exists a matrix H, and positive constants μ_1, μ_2 independent of N such that

$$\mu_1 I \leq HP \leq \mu_2 I$$

specifically, HP is a symmetric positive definite matrix.

(iv) There exists a matrix $G = HQ$ such that $G + G^T$ has only two elements: $g_{0,0}$ and $g_{N,N}$. In general we require $g_{0,0} < 0 < g_{N,N}$.

Assumptions *1* and *2* are common to any useful compact scheme. Assumptions *3* and *4* are specific to the summation-by-parts requirement for the spatial operator.

Equation (1) is now semi-discretized using formula (2) to yield

$$\frac{d\mathbf{u}}{dt} = \lambda P^{-1}Q\mathbf{u} \tag{3}$$

Note that assumptions *3* and *4* from above admit a summation-by-parts formula in the sense that

$$\frac{dE}{dt} = g_{0,0}u_0^2 + g_{N,N}u_N^2 \tag{4}$$

where

$$E(t) = \frac{1}{\lambda}(\mathbf{u}(t), HP\mathbf{u}(t)) \tag{5}$$

In Appendix I we show how to construct a fourth-order compact scheme that satisfy Assumption *1* and therefore (4).

Interestingly, equations (4) and (5) were obtained *without imposing the boundary conditions.* We will use the summation-by-parts property defined in equations (4) and (5) to construct a scheme

that admits a decreasing energy norm when the boundary condition is imposed. Note that the way in which the boundary condition is imposed is important for numerical stability. The most common procedure of imposing the boundary conditions ($\lambda > 0$), is to use equation (3) to update the unknowns $u_0, ..u_N$, followed by overwriting $u_N = g(t)$. This procedure accounts for the fact that in a general hyperbolic system the precise location for each boundary condition is not known until after a characteristic decomposition is performed at all boundaries. This procedure (particularly if H is a nontrivial matrix), may not yield the estimate (4) with u_N replaced by $g(t)$. In short, the imposition of certain boundary treatments may ruin the structure of the summation norm, which results in a numerical scheme that is not time-stable.

A simple counter-example is presented which demonstrates the necessity of careful boundary implementation. Consider the scalar equation $u_t = u_x$ with the boundary condition $u_N = g(t)$. The semi-discretization in the absence of boundary conditions becomes $u_t = A\,u$, where $A = P^{-1}\,Q$. As described earlier, once the matrix A is formed, the boundary conditions are imposed. This has the effect of pre-multiplying the matrix A by the boundary matrix D. Without loss of generality, we use the boundary condition $g(t) = 0$ in this problem; the resulting boundary operator is the matrix

$$D_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

For time stability, the resulting matrix $A^{\dagger} = D\,P^{-1}\,Q$, rather than the matrix A must exhibit a summation-by-part norm.

For simplicity, we discretize the domain into two even intervals, such that the discrete solution vector is $(u_0, u_1, u_2)^T$. The boundary condition is imposed at u_2. A first-order discretization that satisfies the summation-by-parts energy norm is

$$P_3 = \begin{bmatrix} 77/48 & (-19)/12 & (-43)/48 \\ (-19)/12 & 32/3 & (-13)/12 \\ (-43)/48 & (-13)/12 & 53/48 \end{bmatrix}; \quad Q_3 = \begin{bmatrix} (-25)/16 & 4 & (-39)/16 \\ -4 & 0 & 4 \\ 39/16 & -4 & 25/16 \end{bmatrix}$$

Note that the matrices P and Q satisfy $P_3 = P_3^T$ and $Q_3 = -Q_3^T$ except for $q_{0,0}$ and $q_{2,2}$. In this example, the matrix H is the identity matrix. The characteristic equation for the P_3 matrix is $-192\lambda^3 + 2568\lambda^2 - 5026\lambda + 501 = 0$. The symmetry of P_3 and the alternating signs of the respective terms in the characteristic polynomial guarantee the positive definiteness of P_3. The discretization operator $A_3 = P_3^{-1}\,Q_3$ can be written as

$$A_3 = \begin{bmatrix} 11/1002 & (-512)/501 & 1013/1002 \\ (-55)/334 & (-112)/167 & 279/334 \\ 2059/1002 & (-2560)/501 & 3061/1002 \end{bmatrix}$$

All the requirements of the summation-by-parts energy norm are satisfied by this discretization, and a precise energy norm exists in the absence of boundary conditions.

The combined operator $A_3^{\dagger} = D_3 A_3$ becomes

$$A_3^{\dagger} = \begin{bmatrix} 11/1002 & (-512)/501 & 1013/1002 \\ (-55)/334 & (-112)/167 & 279/334 \\ 0 & 0 & 0 \end{bmatrix}$$

for which the characteristic polynomial is $-1002\lambda^3 - 661\lambda^2 + 176\lambda = 0$. The roots of the characteristic polynomial are $\lambda = -0.86317...$, and $\lambda = 0.20349...$, respectively. The numerical solution will grow in time as a result of the eigenvalue in the right half of the complex plane (RH-P) and will not be time-stable.

As demonstrated by the previous counter-example, a spatial operator which satisfies the summation-by-parts energy norm may not be time-stable. Many of the high-order schemes that satisfy the summation property are time-stable for the scalar case. A notable exception is the sixth-order explicit scheme with fifth-order boundary conditions reported in the work of Strand [3]. (See Appendix II for details of this scheme.) For this sixth-order scheme, time stability can be guaranteed only if the last row and column of the matrices HP and HQ are removed before matrix inversion and multiplication are performed.

The underlying reason for the growth in time is the imposition of the boundary condition operator, which has an effect on the structure of the norm matrix P in $u_t = D P^{-1} Q$. Specifically, $D P^{-1}$ destroys the structure of the norm P. In the scalar case, this problem can be eliminated in certain circumstances. For instance, if the matrix P is a restricted full norm, then $D P^{-1}$ still produces a useful norm by eliminating the zero element. A restricted full norm is defined where the diagonal is the only nonzero element in the first (or last) row and column of the matrix P (See Strand [3]). A special case of the restricted full norm is the diagonal case, which is of some practical interest. Unfortunately, even for cases where P is a restricted full norm, stability cannot be generalized to the case of a hyperbolic system. An alternative means of imposing boundary conditions must be found for these cases.

At this point, we introduce the SAT methodology for boundary implementation. We show in the following text that the SAT method leads not only to stability but also to time stability for the scalar wave equations, and this property applies to arbitrary hyperbolic systems. The SAT method

involves the indirect imposition of the physical boundary conditions. This is accomplished by adding a term to the derivative operator, which is proportional to the difference between the discrete value u_N and the boundary term $g(t)$. Thus, we propose the discretization,

$$P\frac{d\mathbf{u}}{dt} = \lambda Q\mathbf{u} - \tau\lambda g_{N,N}\mathbf{S}(u_N - g(t)) \tag{6}$$

where

$$\mathbf{S} = H^{-1}(0,0,...,0,1)^T \tag{7}$$

Contrary to the common practice of satisfying the boundary condition directly by imposing $u_N = g(t)$, the SAT method involves solving a derivative equation *everywhere*, including the boundary points. The extra term which is added accounts for the boundary information to within the accuracy of the original discretization. Note that the SAT is added not only to the boundary equation but to other points depending on the structure of the vector $\mathbf{S}$, (which is the last column of the matrix H^{-1}). The extra SAT term does not alter the accuracy of the scheme, since the SAT term vanishes upon substitution of the analytic solution.

We now demonstrate that the SAT method yields a Lax stable and time-stable scheme. For the time stability analysis, we take $g(t) = 0$. We pre-multiply equation (6) by H and use equation (7) to obtain

$$HP\frac{d\mathbf{u}}{dt} = \lambda HQ\mathbf{u} - \tau\lambda g_{N,N}(0,0,...,0,1)^T u_N \tag{8}$$

We now define the energy $E(t)$ as in equation (5) to get

$$\frac{dE(t)}{dt} = g_{0,0}u_0^2 + g_{N,N}u_N^2 - \tau g_{N,N}u_N^2 \tag{9}$$

With $g_{0,0} < 0 < g_{N,N}$, we can immediately state the following theorem.

Theorem 1.1:

The SAT method presented in equation (6) is both stable *and time-stable* if

$$\tau \geq 1 \tag{10}$$

In addition to proving the stability of the SAT scheme defined in equation (6), we must show that the procedure preserves the order of accuracy m of the spatial operator. This is accomplished by a direct *convergence* proof showing that the SAT term indeed preserves the spatial order of accuracy.

Denote by $\mathbf{v}$ the vector $(u(x_0,t),...,u(x_N,t))^T$, i.e. the values of the *true* solution of (1) at the grid points. Combining the accuracy condition found in Assumption *1* with equation (6) we have

$$P\frac{d\mathbf{v}}{dt} = \lambda Q\mathbf{v} - \tau\lambda g_{N,N}\mathbf{S}\left[u(x_N,t) - g(t)\right] + P\mathbf{T_e} \tag{11}$$

Note that $u(x_N,t)-g(t)=u(1,t)-g(t)=0$. Now define

$$\epsilon_j(t)=u(x_j,t)-u_j(t)$$

where $u_j(t)$ solves (6) , to obtain

$$P\frac{d\epsilon}{dt}=\lambda Q\epsilon-\tau\lambda g_{N,N}\mathbf{S}\epsilon_N+P\mathbf{T_e} \tag{12}$$

where $\mathbf{T_e}$ is the truncation error defined in Assumption *1*. We now use the energy estimate presented in (9) to obtain

$$\frac{d(\epsilon,HP\epsilon)}{dt}\leq(\epsilon,HPT_e)$$

and the inequality

$$(\epsilon,HPT_e)\leq\sqrt{(\epsilon,HP\epsilon)}\sqrt{(T_e,HPT_e)}$$

to obtain

$$\frac{d\sqrt{(\epsilon,HP\epsilon)}}{dt}\leq\sqrt{(T_e,HPT_e)} \tag{13}$$

By assumption I ,the truncation error is of order m, and we get

$$\sqrt{(\epsilon,HP\epsilon)}\leq O(\Delta x)^m$$

which proves the convergence of the scheme.

In conclusion, a precise means is now available for the scalar case to impose boundary conditions that are guaranteed to be time stable, and that preserve the formal accuracy of the original discretization.

2. The Hyperbolic System

In this section, we explain how to use the SAT method for systems of hyperbolic equations and show that the resulting scheme satisfies an energy estimate similar to the one obtained for the scalar differential equation. First the system of differential equations is described.

Let $\mathbf{u^I}$ and $\mathbf{u^{II}}$ be the two function-valued vectors

$$\mathbf{u^I}=(u^{(1)}(x,t),...,u^{(k)}(x,t)) \tag{14}$$

$$\mathbf{u^{II}} = (u^{(k+1)}, ..., u^{(r)}(x,t))$$

that solve the system of differential equations

$$\frac{\partial \mathbf{u^I}}{\partial t} = \Lambda^I \frac{\partial \mathbf{u^I}}{\partial x}$$

$$\frac{\partial \mathbf{u^{II}}}{\partial t} = \Lambda^{II} \frac{\partial \mathbf{u^{II}}}{\partial x} \tag{15}$$

where Λ^I and Λ^{II} are diagonal matrices of the form

$$\Lambda^I = diag(\lambda_1, ..., \lambda_k)$$

$$\Lambda^{II} = diag(\lambda_{k+1}, ..., \lambda_r) \tag{16}$$

In order to impose the boundary conditions we assume that

$$\lambda_1 > \lambda_2 > ... > \lambda_k > 0 > \lambda_{k+1} > ... > \lambda_r$$

For this case, a well-posed set of boundary conditions is given by

$$\mathbf{u^I}(1,t) = R\mathbf{u^{II}}(1,t) + \mathbf{g^I(t)}$$

$$\mathbf{u^{II}}(0,t) = L\mathbf{u^I}(0,t) + \mathbf{g^{II}(t)} \tag{17}$$

Where

$$\mathbf{g^I(t)} = (\mathbf{g^{(1)}(t)}, ..., \mathbf{g^{(k)}(t)})$$

and,

$$\mathbf{g^{II}(t)} = (\mathbf{g^{(k+1)}(t)}, ..., \mathbf{g^{(r)}(t)})$$

In equation (17), the matrix R has k rows and $r-k$ columns, while the matrix L has $r-k$ rows and k columns. Without loss of generality, for the stability analysis we will assume that both $\mathbf{g^I(t)}$ and $\mathbf{g^{II}(t)}$ vanish.

Equation (17) is well-posed for any L and R. However, to guarantee no growth in time some conditions must be imposed on the matrices L and R. These conditions are

Condition I:

$$|L||R| \leq 1 \tag{18}$$

where the matrix norm is defined by

$$|A| = \rho(A^T A)^{\frac{1}{2}}$$

and $\rho(A)$ is the spectral radius of A.

The Continuous case

It is instructive to establish and prove an energy estimate for the continuous hyperbolic system although such a proof is well known. The same basic steps that are used in the continuous proof will be used later in the text to prove the energy estimate resulting from the semi-discrete hyperbolic system.

Condition I is a sufficient condition for the solution of equation (15) to be bounded in time. In fact one can state

Theorem 2.1:

Let $\mathbf{u^I}(x,t)$ and $\mathbf{u^{II}}(x,t)$ be the solution of equation (15) with the boundary conditions (17). Recall that we take $g^I = g^{II} = 0$. Suppose that L and R in equation (17) satisfy Condition I. Define an inner product

$$(w,v) = \int_0^1 w(x,t)v(x,t)dx \tag{19}$$

and an energy function $E(t)$

$$E(t) = \sum_{i=1}^{k} \frac{|L|}{\lambda_i}(u^{(i)}, u^{(i)}) + \sum_{i=k+1}^{r} \frac{|R|}{|\lambda_i|}(u^{(i)}, u^{(i)}) \tag{20}$$

then the time rate of the energy function satisfies

$$\frac{dE}{dt} \leq 0 \tag{21}$$

Proof

We start by differentiating equation (19) with respect to t to obtain

$$\frac{d(u^{(i)}, u^{(i)})}{dt} = 2\int_0^1 u^{(i)} u_t^{(i)}\ dx$$

Using equation (15) we obtain

$$\frac{d(u^{(i)}, u^{(i)})}{dt} = 2\int_0^1 u^{(i)}\lambda_i u_x^{(i)} dx$$

so that

$$\frac{d(u^{(i)}, u^{(i)})}{dt} = \lambda_i(u^{(i)}(1,t)^2 - u^{(i)}(0,t)^2) \tag{22}$$

Differentiating equation (20) and substituting equation (22) we obtain the energy rate for the system as

$$\frac{dE}{dt} = \sum_{i=1}^{k} |L|(u^{(i)}(1,t)^2 - u^{(i)}(0,t)^2) - \sum_{i=k+1}^{r} |R|(u^{(i)}(1,t)^2 - u^{(i)}(0,t)^2) \tag{23}$$

relating the time rate of change of the energy function to the energy that crosses the boundaries. Note the change of sign in the second term which results from the negative sign of the eigenvalues λ_i for $k < i$. We must now quantify the magnitude of the boundary terms in equation (23).

Replacing the sums in equation (23) with the vector operations

$$\sum_{i=1}^{k} u^{(i)}(1,t)^2 = \mathbf{u^I}(1,t)^T \mathbf{u^I}(1,t)$$

$$\sum_{i=k+1}^{r} u^{(i)}(0,t)^2 = \mathbf{u^{II}}(0,t)^T \mathbf{u^{II}}(0,t) \tag{24}$$

we can now make use of the boundary conditions in equation (17) to obtain

$$\mathbf{u^I}(1,t)^T \mathbf{u^I}(1,t) = \mathbf{u^{II}}(1,t)^T R^T R \mathbf{u^{II}}(1,t)$$

$$\mathbf{u^{II}}(0,t)^T \mathbf{u^{II}}(0,t) = \mathbf{u^I}(0,t)^T L^T L \mathbf{u^I}(0,t) \tag{25}$$

Substituting the equations (24) and (25) into (23) we obtain

$$\frac{dE}{dt} = \mathbf{u^{II}}(1,t)^T \{R^T R|L| - |R|\}\mathbf{u^{II}}(1,t) + \mathbf{u^I}(0,t)^T \{L^T L|R| - |L|\}\mathbf{u^I}(0,t) \tag{26}$$

Because Condition I ensures that

$$R^T R|L| - |R|I \leq 0$$

and

$$L^T L|R| - I|L| \leq 0$$

equation (21) is established. Therefore the continuous energy function $E(t)$ is bounded in time. This completes the proof of Theorem (2.1).

The Semi-discrete Case

We are ready now to discuss the implementation of the SAT technique for the system in equation (15) with the boundary condition given in equation (17). As in section 1 we denote by $\mathbf{u}^i$ a vector of unknowns $(u_0^{(i)}, u_1^{(i)}, ...u_N^{(i)})^T$ which correspond to the grid points $x_0(=0), x_1, ...x_N(=1)$. We assume that we have matrices P , Q and H such that the *scalar case* admits a summation-by-parts energy norm given in section 1. The SAT discretization of equations (15) - (17) is chosen as

$$P\frac{d\mathbf{u}^i}{dt} = \lambda_i Q\mathbf{u}^i - g_{N,N}\lambda_i\tau S^{(i)}(u_N^{(i)} - (R\mathbf{u}^{\mathbf{II}})_N^{(i)} - g^{(i)}) \qquad 1 \le i \le k$$

$$P\frac{d\mathbf{u}^i}{dt} = \lambda_i Q\mathbf{u}^i - g_{0,0}\lambda_i\tau S^{(i)}(u_0^{(i)} - (L\mathbf{u}^{\mathbf{I}})_0^{(i)} - g^{(i)}) \qquad k+1 \le i \le r \tag{27}$$

where τ is a stabilizing factor to be determined later. As in the scalar case, we choose $S^{(i)}$ to be one of the vectors

$$S^{(i)} = H^{-1}(0,0,...,0,1)^T \qquad 1 \le i \le k$$

$$S^{(i)} = H^{-1}(1,0,...,0,0)^T \qquad k+1 \le i \le r \tag{28}$$

We recall from the scalar case that HP is symmetric positive definite and HQ is skew symmetric except for the terms $g_{0,0} = (HQ)_{0,0} < 0$ and $g_{N,N} = (HQ)_{N,N} > 0$. Thus equation (27) is well defined.

Before proving the stability (and time stability) of the SAT method in equation (27), we would like to comment on the role of the matrix H. Explicit knowledge of H is required for the implementation of the SAT method, specifically the knowledge of $g_{0,0}$ and $g_{N,N}$ as well as the vectors $S^{(i)}$ are needed to implement equation (27). Thus H is not only a theoretical tool (as in reference [2]) but is also of practical importance.

We are now ready for the stability proof of the SAT method in equation (27). The proof is analogous to Theorem 2.1 with the continuous integrals replaced by discrete sums. The scalar product is defined, analogous to equation (19), as

$$(\mathbf{u}^i, \mathbf{u}^i) = \sum_{l=0}^{N} u_l^{(i)} u_l^{(i)} \tag{29}$$

A different scalar product to be used later, analogous to equations (24), is

$$[\mathbf{u^I}, \mathbf{u^I}]_m = \sum_{i=1}^{k} u_m^{(i)} u_m^{(i)}$$

$$[\mathbf{u^{II}}, \mathbf{u^{II}}]_m = \sum_{i=k+1}^{r} u_m^{(i)} u_m^{(i)} \tag{30}$$

for $m = 0, N$.

Theorem (2.2)

Let the SAT method defined by equation (27) satisfy Assumption 1, for the discretization of the hyperbolic system defined in equation (15) with boundary conditions (17), (with $\mathbf{g^I(t)} = \mathbf{g^{II}(t)} = \mathbf{0}$). Then the discretization is both stable and time-stable provided that

$$\frac{2 - 2\sqrt{1 - |R||L|}}{|R||L|} \leq \tau \leq \frac{2 + 2\sqrt{1 - |R||L|}}{|R||L|} \tag{31}$$

Moreover, let the discrete energy be defined as

$$E_N(t) = \sum_{i=1}^{k} \frac{|L|}{\lambda_i} (\mathbf{u}^i, HP\mathbf{u}^i) + \sum_{i=k+1}^{r} \frac{|R|}{|\lambda_i|} (\mathbf{u}^i, HP\mathbf{u}^i) \tag{32}$$

where the scalar product $(\mathbf{u}^i, \mathbf{u}^i)$ is defined in equation (29). Then

$$\frac{dE_N(t)}{dt} \leq 0$$

Proof

As in theorem 2.1 we differentiate the scalar product $(\mathbf{u}^i, HP\mathbf{u}^i)$ and use equation (27) to obtain

$$\frac{d}{dt}(\mathbf{u}^i, HP\mathbf{u}^i) = \lambda_i (\mathbf{u}^i, HQ\mathbf{u}^i) - g_{N,N} \lambda_i \tau (u_N^{(i)} - (R\mathbf{u^{II}})_N^{(i)})(\mathbf{u}^i, HS^{(i)}) \qquad 1 \leq i \leq k$$

$$\frac{d}{dt}(\mathbf{u}^i, HP\mathbf{u}^i) = \lambda_i (\mathbf{u}^i, HQ\mathbf{u}^i) - g_{0,0} \lambda_i \tau (u_0^{(i)} - (R\mathbf{u^I})_0^{(i)})(\mathbf{u}^i, HS^{(i)}) \qquad k+1 \leq i \leq r \tag{33}$$

We now use the definition of $S^{(i)}$ from equation (28) and the properties of HQ from Assumption I to obtain

$$\frac{d}{dt}(\mathbf{u}^i, HP\mathbf{u}^i) =$$

$$g_{0,0}\lambda_i(u_0^{(i)})^2 + \lambda_i g_{N,N}(u_N^{(i)})^2 - \lambda_i g_{N,N}\tau(u_N^{(i)})^2 + \lambda_i g_{N,N}\tau u_N^{(i)}(R\mathbf{u}^{\mathbf{II}})_N^{(i)}, \quad 1 \le i \le k$$

(34)

$$\frac{d}{dt}(\mathbf{u}^i, HP\mathbf{u}^i) =$$

$$-g_{0,0}|\lambda_i|(u_0^{(i)})^2 - |\lambda_i|g_{N,N}(u_N^{(i)})^2 + |\lambda_i|g_{0,0}\tau(u_0^{(i)})^2 - |\lambda_i|g_{0,0}\tau u_0^{(i)}(L\mathbf{u}^{\mathbf{I}})_0^{(i)}, \quad k+1 \le i \le r$$

Note that in equation (34) we used the fact that the λ_i are negative for $k+1 \le i \le r$. We must now quantify the magnitude of the boundary terms in equations (34). If the sums in equations (34) are replaced with the vector operations defined in equations (30) we get an estimate for the discrete energy rate $\frac{dE_N(t)}{dt}$

$$\frac{dE_N(t)}{dt} = |L|g_{0,0}[\mathbf{u}^{\mathbf{I}}, \mathbf{u}^{\mathbf{I}}]_0 + |L|g_{N,N}(1-\tau)[\mathbf{u}^{\mathbf{I}}, \mathbf{u}^{\mathbf{I}}]_N + |L|g_{N,N}[\mathbf{u}^{\mathbf{I}}, R\mathbf{u}^{\mathbf{II}}]_N$$

(35)

$$+|R|(\tau-1)g_{0,0}[\mathbf{u}^{\mathbf{II}}, \mathbf{u}^{\mathbf{II}}]_0 - |R|g_{N,N}[\mathbf{u}^{\mathbf{II}}, \mathbf{u}^{\mathbf{II}}]_N - g_{0,0}|R|\tau[\mathbf{u}^{\mathbf{II}}, L\mathbf{u}^{\mathbf{I}}]_0$$

Substituting the estimates

$$[\mathbf{u}^{\mathbf{I}}, R\mathbf{u}^{\mathbf{II}}]_N \le |\mathbf{u}^{\mathbf{I}}|_N|R||\mathbf{u}^{\mathbf{II}}|_N$$

$$[\mathbf{u}^{\mathbf{II}}, L\mathbf{u}^{\mathbf{I}}]_0 \le |\mathbf{u}^{\mathbf{II}}|_0|L||\mathbf{u}^{\mathbf{I}}|_0$$

where

$$|\mathbf{u}^{\mathbf{I}}|_m = \sqrt{[\mathbf{u}^{\mathbf{I}}, \mathbf{u}^{\mathbf{I}}]_m}.$$

into equation (35), and collecting like terms yields

$$\frac{dE_N(t)}{dt} \le -g_{N,N}\{|L|(\tau-1)|\mathbf{u}^{\mathbf{I}}|_N^2 - \tau|L||R||\mathbf{u}^{\mathbf{I}}|_N|\mathbf{u}^{\mathbf{II}}|_N + |R||\mathbf{u}^{\mathbf{II}}|_N^2\}$$

(36)

$$+g_{0,0}\{|R|(\tau-1)|\mathbf{u}^{\mathbf{II}}|_0^2 - \tau|L||R||\mathbf{u}^{\mathbf{I}}|_0|\mathbf{u}^{\mathbf{II}}|_0 + |L||\mathbf{u}^{\mathbf{I}}|_0^2\}$$

For $\frac{dE_N}{dt}$ to be negative we require each curly bracket to be positive. Thus we need

$$|L|(\tau-1)|\mathbf{u}^{\mathbf{I}}|_N^2 - \tau|L||R||\mathbf{u}^{\mathbf{I}}|_N|\mathbf{u}^{\mathbf{II}}|_N + |R||\mathbf{u}^{\mathbf{II}}|_N^2 \ge 0$$

and also

$$|R|(\tau-1)|\mathbf{u}^{\mathbf{II}}|_0^2 - \tau|L||R||\mathbf{u}^{\mathbf{I}}|_0|\mathbf{u}^{\mathbf{II}}|_0 + |L||\mathbf{u}^{\mathbf{I}}|_0^2 \ge 0$$

Both inequalities are satisfied if

$$|R||L|\tau^2 \le 4(\tau-1)$$

and this is equivalent to equation (31). Thus, the proof is established.

3. Results

Conventional Boundary Conditions

Three high-order spatial discretizations (two explicit and one compact) are the focus of the results section: the fourth-order explicit scheme with third-order boundary conditions, the fourth-order compact scheme with third-order boundary conditions, and the sixth-order explicit scheme with fifth-order boundary conditions. All satisfy the summation-by-parts requirement in the absence of physical boundary conditions. The fourth-order explicit scheme is reported elsewhere (see [3] or [7] for specific details) and will not be derived here. The fourth-order compact scheme is new, and a systematic procedure for deriving both it and other compact high-order schemes is presented in Appendix I. The sixth-order explicit scheme was first reported in reference [3], but is also included in Appendix II.

First we demonstrate that all three schemes behave in accordance with their respective order properties. We then comment with regard to the sixth-order explicit scheme, that satisfying the summation-by-parts energy norm is not sufficient for time stability.

The model problem used to test the three schemes is the scalar hyperbolic equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, \ 0 \leq x \leq 1, \ t \geq 0 \tag{37}$$

$$u(0,t) = \sin 2\pi(-t), \ t \geq 0 \tag{38}$$

$$u(x,0) = \sin 2\pi(x), \ 0 \leq x \leq 1, \tag{39}$$

The exact solution is

$$u(x,t) = \sin 2\pi(x-t), \ 0 \leq x \leq 1, \ t \geq 0 \tag{40}$$

For all calculations, the time discretization used was a fourth-order Runge-Kutta (R-K) method with the time step small enough such that the temporal errors are much smaller than the spatial truncation error. In all cases, the boundary condition was implemented at the end of each R-K stage by overwriting the value of the solution at the boundary point.

Table I shows a grid refinement study performed on equation (37) for all three spatial discretizations. Both the absolute ($\log L_2$) error at a fixed time T and the convergence rate between two successive grid densities are plotted.

	(fourth	explicit)	(fourth	compact)	(sixth	explicit)
Grid	$\log L_2$	Rate	$\log L_2$	Rate	$\log L_2$	Rate
21	-0.501		-1.418		1.379	
31	-2.080	8.96	-2.133	4.06	1.048	1.88
41	-2.607	4.22	-2.627	3.95	0.137	7.29
61	-3.329	4.10	-3.316	3.91	-1.302	8.17
81	-3.832	4.03	-3.806	3.92	-1.798	3.96

Table I: Grid convergence of three high-order schemes on $u_t + u_x = 0$.

This refinement study suggests that all three schemes are Lax stable (the exact solution is approached at a fixed time T as mesh is refined) and grid converge consistent with each respective theoretical rate. The convergence rates for both of the fourth-order schemes asymptote to the theoretical value of 4. The convergence rate of the sixth-order explicit scheme is sporadic but is approximately 6 (5.28 for the interval between 21 and 81 points). This spurious behavior results from the exponential divergence of the solution for long times T. At $T = 70$, the absolute error of the two fourth-order schemes is comparable; however, that of the sixth-order scheme is two to three orders of magnitude larger.

These numerical results indicate that the two fourth-order schemes are time-stable; the sixth-order scheme is not. Nothing in the definition of Lax stability precludes exponential divergence of the solution for long times T as long as the divergence rate is bounded independently of the grid used. (See reference [4].) The numerical divergence of the solution results from a spatial operator matrix which has an eigenvalue with a positive real part (an RH-P eigenvalue). For long times T, the solution is dominated by this eigenvalue.

To quantify this assertion, a comparison is presented between the numerically observed divergence rate, and a theoretical prediction from eigenvalue analysis. By assuming that the numerical error can be represented as $\epsilon_N(t) = \epsilon_N(0)e^{\alpha_N t}$, a growth rate α_N is determined. Similarly, an effective growth rate α_S defined by $e^{\alpha_S M\Delta t} = |G_{max}(\Delta t)|^M$, is calculated from an eigenvalue determination. (See reference [4] for details). Table II shows a comparison of the observed growth rate of the sixth-order explicit scheme with the rate predicted from an eigenvalue determination.

Grid	$\alpha_{Numerical}$	$\alpha_{(S_{max})}$
21	0.1672	0.1673
31	0.1879	0.1886
41	0.1880	0.1879
61	0.1659	0.1746
81	0.1785	0.1808

Table II: Numerical vs. Theoretical Growth Rate for the sixth-order explicit.

The agreement is very good, with a slight discrepancy in the comparison on the 61 and 81 grid-point cases.

The time-divergence seen in the sixth-order scheme is the same as that predicted in the counter-example presented in section 1. Specifically, numerical time stability is not guaranteed by a discretization which satisfies a summation-by-parts property. Very specific boundary treatments must be used to guarantee time stability.

SAT Boundary Conditions (Scalar)

The SAT method for treating the boundary conditions guarantees time stability for the hyperbolic system. This method relies on a spatial operator that satisfies the summation-by-parts energy norm for the scalar case and on very specific boundary treatments to ensure time stability.

We begin by showing that the procedure does not destroy the formal accuracy of the spatial discretization. This result was proven in section 1 for the scalar case. Tables III.a and III.b show a grid convergence study of the SAT method on the scalar wave equation defined by equations (37), (38) and (39). Fourth-order R-K time advancement is used for all runs with a time step such that no appreciable temporal error accumulates. All calculations are run to time $T = 10$. In all cases, the calculations remained bounded on all grids (and CFL's less than CFL_{max}) for times as large as $T = 1000$, which indicates time stability. This result is consistent with the results from eigenvalue determinations in which no RH-P eigenvalues were found.

$\tau = 1$	(fourth	explicit)	(fourth	compact)	(sixth	explicit)
Grid	$\log L_2$	Rate	$\log L_2$	Rate	$\log L_2$	Rate
21	-1.2289		-1.4005		-2.5750	
31	-2.0878	4.88	-2.0479	3.67	-3.8300	7.13
41	-2.5784	3.93	-2.5096	3.70	-4.6500	6.56
61	-3.2211	3.65	-3.1689	3.74	-5.7880	6.46
81	-3.6806	3.68	-3.6464	3.82	-6.6056	6.54

Table III.a. Absolute error ($\log L_2$) and convergence exponent with SAT parameter $\tau = 1$, for the fourth explicit, fourth compact and sixth explicit spatial discretizations.

$\tau = 2$	(fourth	explicit)	(fourth	compact)	(sixth	explicit)
Grid	$\log L_2$	Rate	$\log L_2$	Rate	$\log L_2$	Rate
21	-1.3472		-1.8061		-2.7007	
31	-2.0866	4.20	-2.4296	3.54	-3.8229	6.37
41	-2.5980	4.09	-2.8773	3.58	-4.6666	6.75
61	-3.3107	4.05	-3.5243	3.67	-5.8518	6.73
81	-3.8145	4.03	-3.9978	3.79	-6.6485	6.38

Table III.b: Absolute error ($\log L_2$) and convergence exponent with SAT parameter $\tau = 2$, for the fourth explicit, fourth compact and sixth explicit spatial discretizations.

A comparison of the SAT grid refinement studies (table III.a and III.b) with those from the conventional boundary treatment (table I), indicates that the formal accuracy of the spatial operator is unaffected by the SAT treatment. The proof of stability given in section 1 indicated that a sufficient condition for stability of the scalar wave equation with the SAT method is $1 \leq \tau$. The results shown in tables III.a and III.b indicate that the magnitude of the error is dependent on the value of the parameter τ. To optimize the value of the parameter τ for these simulations, the error at $T = 10$ was studied as a function of τ. An eigenvalue code was then used to determine the maximum CFL of the scheme as a function of τ. The results of this study are shown in Table IV.

τ	$\log L_2$	CFL
3.0	-3.8220	1.17
2.5	-3.8221	1.77
2.0	-3.8145	2.07
1.75	-3.8038	2.07
1.50	-3.8833	2.07
1.25	-3.7460	2.07
1.00	-3.6806	2.07
0.97		0.0

Table IV: Absolute error ($\log L_2$) and CFL for various values of the SAT parameter τ, for the fourth explicit spatial operator.

Note that a fairly sharp cutoff at the theoretical value of $\tau = 1$ is observed for the fourth-order explicit spatial operator. (Values of $\tau = 0.93$ and $\tau = 0.99$ were obtained for the fourth-order compact and sixth-order explicit schemes, respectively. In addition, precise agreement was obtained at the τ cutoff between the eigenvalue determination and the numerical simulation of the scalar wave equation.) For the fourth-order explicit spatial operator, the error decreased monotonically with τ, which suggests that the value of τ should be as large as possible. Conversely, the maximum CFL

that is achievable with the fourth-order R-K scheme decreases dramatically at $\tau = 2$. A value of $\tau = 2$ was determined to be optimal for these studies.

SAT Boundary Conditions (System)

The last part of the validation study is to verify that the SAT boundary procedure ensures stability for the hyperbolic system. Equation (31) defines sufficient conditions for time stability $\frac{1}{|R||L|}(2 - 2\sqrt{1-|R||L|}) \leq \tau \leq \frac{1}{|R||L|}(2 + 2\sqrt{1-|R||L|})$ in terms of τ and the boundary coupling matrices L and R. The test case chosen is the hyperbolic system

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} &= 0, \\ \frac{\partial v}{\partial t} - \frac{\partial v}{\partial x} &= 0, \quad 0 \leq x \leq 1, \; t \geq 0 \end{aligned} \tag{41}$$

$$u(0,t) = \alpha v(0,t), \;\; v(1,t) = \beta u(1,t), \;\; t \geq 0 \tag{42}$$

$$u(x,0) = \sin 2\pi x, \;\; v(x,0) = -\sin 2\pi x, \;\; 0 \leq x \leq 1, \tag{43}$$

The exact solution for $\alpha = \beta = 1$ is

$$u(x,t) = \sin 2\pi(x-t), \;\; v(x,t) = -\sin 2\pi(x+t), \;\; 0 \leq x \leq 1, \;\; t \geq 0 \tag{44}$$

The case $|\alpha\,\beta| = 1$ is neutrally stable and provides an extremely severe test of the time stability of a numerical method. No central difference scheme of an order greater than two, is time-stable for this system, in spite of the fact that the spatial operator is stable for the scalar case ($\alpha = \beta = 0$). Examples include the (3-4-3) compact and (3,3-4-3,3) explicit fourth-order schemes, and the ($5^2, 5^2$-6-$5^2, 5^2$) sixth-order scheme that is shown in reference [4] to be time-stable for the scalar case. All three schemes used in the scalar analysis (fourth-order explicit and compact and sixth-order explicit), that satisfy the summation-by-parts property are not time-stable. In all cases, the discrete solution of the system defined by equations (41) through (44) diverges as time becomes large. Grid refinement shows Lax stability and an order property for each scheme, but not time stability.

The scalar analysis demonstrates a precise relationship between schemes that are time-stable and the structure of the eigenvalue spectrum that arises from the discretization matrix. Precisely, if RH-P eigenvalues exist, then numerical divergence can be expected from the numerical simulation. Unfortunately, this statement is a function of the CFL that is used to advance the solution. (See reference [4].) Values of the CFL can be chosen for which no numerical divergence is experienced with an R-K time advancement scheme; for this reason testing the numerical stability of various spatial operators for the fully discrete system in time is impractical.

The alternative is to use the eigenvalue structure of the semi-discrete problem as the test for stability. If a spatial discretization operator has no RH-P eigenvalues, then it is assumed to be time-stable. A derivation of the discretization matrix operators for the model hyperbolic system [equations (41) and (42)] is presented in Appendix III. In addition, the structure of the eigenvalues is derived.

For our test system, we take $\alpha = \beta$ in (44) and thus the sufficient condition for stability becomes $(\frac{2-2\sqrt{1-\alpha^2}}{\alpha^2}) \leq \tau \leq (\frac{2+2\sqrt{1-\alpha^2}}{\alpha^2})$. Given a value of α and a stable scheme incorporating the SAT boundary treatment for the system, there exist a range in τ for which the time discretization is stable. As in the scalar case, good agreement exists between the theoretical and numerical stability limit. Therefore, the agreement between the theoretical prediction and the numerical eigenvalue determination was used as a test of the validity of the theory.

Table V compares the stability limits of the three high order schemes for various values of the parameter α; the theoretical limit is compared with that predicted from the eigenvalue determination for the 2×2 system. The number of grid points used in each case was 101. A study with 61 points showed similar results. In the study, τ_T is the theoretical value of τ based on $\frac{2-2\sqrt{1-\alpha^2}}{\alpha^2} = \tau$, and τ_N is the value as determined from the eigenvalue determination. Specifically, τ_N was the smallest value of τ for which the numerical eigenvalues all had negative real parts. In all cases the agreement was very good, which suggests the validity of the theory.

	α	1.0	0.99	0.90	0.80	0.50
Exact	τ_T	2.0	1.75	1.39	1.25	1.07
fourth explicit	τ_N	2.0	1.75	1.39	1.24	1.05
fourth compact	τ_N	2.0	1.75	1.39	1.25	1.08
sixth explicit	τ_N	2.0	1.72	1.25	1.01	1.00

Table V: The theoretical and numerical stability limits of SAT boundary scheme for various values of α.

In these simple examples, we have demonstrated that the SAT boundary procedure retains the formal accuracy of the underlying spatial operator and provides a mechanism to stabilize those spatial operators that satisfy a summation-by-parts energy property. The resulting scheme is time-stable for both the scalar and system case. The numerically predicted stability boundaries for the parameter τ closely match the theoretical predictions. From a practical perspective, the numerical stability and CFL of the fully discrete algorithm are functions of the value of τ. The choice $\tau = 2$ seems to be well suited for both the scalar and system cases and guarantees stability even for the neutrally stable system case where $\alpha = \beta = 1$.

4. Conclusions

In this paper we studied the stability and time stability of the semi-discrete hyperbolic system of partial differential equations. The spatial discretizations considered were high order (explicit and compact), and their boundary terms were constructed such that the derivative matrix satisfied a summation-by-parts formula.

The following results were obtained:

1. A systematic way was developed to obtain high-order accurate derivative matrices (including boundary terms) having a summation-by-parts property. The method is illustrated by finding explicit forms in the 4th order compact case.
2. The summation-by-parts property does not, by itself, guarantee the stability and time stability of the scheme, not even in the scalar case. (Refer to the explicit sixth-order example cited in the text.)
3. To overcome this difficulty we introduce the simultaneous approximation term (SAT) in order to account for the effect of the coupling of the physical boundary conditions. The SAT contains a free parameter τ.
4. We give bounds on τ such that the resulting scheme for the system (or scalar) case, we have stability as well as time stability.
5. Numerical studies verify the theory.

References

[1] B. Gustafsson, The Convergence Rate for Difference Approximations to Mixed Initial Boundary Value Problems, *Math. Comp.* **29**, 130, 1975, pp. 396-406.

[2] H.-O. Kreiss, and G. Scherer, Finite element and finite difference methods for hyperbolic partial differential equations, *Mathematical Aspects of Finite Elements in Partial Differential Equations*, (Academic Press, New York, 1974).

[3] B. Strand, Summation by Parts for Finite Difference Approximations for d/dx, Dept of Scientific Computing, Uppsala University, Uppsala, Sweden, Aug, 1991.

[4] M.H. Carpenter, D. Gottlieb, and S. Abarbanel, The Stability of Numerical Boundary Treatments for Compact High-Order Finite-Difference Schemes, NASA Contractor Report 187628, ICASE Report No. 91-71, Sept 1991. To appear in *JCP*.

[5] B. Gustafsson, H.-O. Kreiss, and A. Sundström, Stability Theory of Difference Approximations for Mixed Initial Boundary Value Problems. II, *Math. Comp.* **26**, 1972, pp. 649-686.

[6] D. Funaro, and D. Gottlieb, A new method of imposing boundary conditions in pseudospectral approximations of hyperbolic equations, *Math. Comp.*, **51**, 184, 1988, pp. 599-613.

[7] M.H. Carpenter, D. Gottlieb, and S. Abarbanel, Stable and Accurate Boundary Treatments for Compact, High Order Finite Difference Schemes, To be published *Applied Numer. Math.*

APPENDIX I

Construction of the Fourth-Order Compact Scheme

We begin with the semi-discrete equation $u_t = A^\dagger u$ where $u = (u_1, u_2, ..., u_N)^T$, which results from a particular discretization of the equation $u_t = u_x$. The matrix $A^\dagger$ is then decomposed as $A^\dagger = P^{-1}Q$. The interior scheme used is the fourth-order compact scheme defined implicitly as

$$\frac{1}{4}\frac{du_{i-1}}{dx} + \frac{du_i}{dx} + \frac{1}{4}\frac{du_{i+1}}{dx} = \frac{3}{4\,\Delta}(u_{i+1} - u_{i-1}) \qquad \text{(AI. 1)}$$

Note that the interior scheme satisfies the summation-by-parts energy norm (as well as the generalized norm). The matrices P and Q can be written in general form, with boundary closures of arbitrary size N as

$$P = \begin{bmatrix} p_{0,0} & \cdots & p_{0,N} & & 0 \\ \vdots & & \vdots & & \\ \vdots & & \vdots & & \\ \vdots & & \vdots & & \\ p_{N,0} & \cdots & p_{N,N} & \frac{1}{4} & \\ & & \frac{1}{4} & 1 & \frac{1}{4} \\ 0 & & & \ddots & \ddots \end{bmatrix}; \quad Q = \begin{bmatrix} q_{0,0} & \cdots & q_{0,N} & & 0 \\ \vdots & & \vdots & & \\ \vdots & & \vdots & & \\ \vdots & & \vdots & & \\ q_{N,0} & \cdots & q_{N,N} & \frac{3}{4} & \\ & & \frac{-3}{4} & 0 & \frac{3}{4} \\ 0 & & & \ddots & \ddots \end{bmatrix}$$

with the H matrix written as

$$H = \begin{bmatrix} h_{0,0} & \cdots & h_{0,N} & & 0 \\ \vdots & & \vdots & & \\ \vdots & & \vdots & & \\ \vdots & & \vdots & & \\ h_{N,0} & \cdots & h_{N,N} & x & \\ & & x & y & x \\ 0 & & & \ddots & \ddots \end{bmatrix}$$

To simplify the matrix algebra, the following new matrices are introduced:

$$S = \frac{3}{4}\begin{bmatrix} 0 & 1 & 0 & & \\ -1 & 0 & 1 & & \\ 0 & -1 & 0 & 1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}; \quad C = \frac{1}{4}\begin{bmatrix} 4 & 1 & 0 & & \\ 1 & 4 & 1 & & \\ 0 & 1 & 4 & 1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

$$D = \begin{bmatrix} y & x & 0 & & \\ x & y & x & & \\ 0 & x & y & x & \\ & & \ddots & \ddots & \ddots \end{bmatrix}; \quad A = \begin{bmatrix} 0 & \cdots & & 0 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 0 & & & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix}$$

Note that S, C, and D are $M \times M$ matrices, where M is an arbitrary number that corresponds to the number of interior points in the discretization. The structure of the matrices is tri-diagonal in nature. The matrix A is $N \times M$, and the only non-zero element is $a_{N,1} = 1$.

Thus, we can write H, P, and Q as

$$H = \begin{bmatrix} \hat{H} & x\,A \\ x\,A^T & D \end{bmatrix}; \quad P = \begin{bmatrix} \hat{P} & \frac{1}{4}A \\ \frac{1}{4}A^T & C \end{bmatrix}; \quad Q = \begin{bmatrix} \hat{Q} & \frac{3}{4}A \\ -\frac{3}{4}A^T & S \end{bmatrix}$$

where $\hat{P}$, $\hat{Q}$, and $\hat{H}$ are the $N \times N$ submatrices that involve the unknown quantities in the matrices P, Q, and H, respectively.

The spatial operator that involves P and Q satisfies the generalized summation-by-parts energy norm if a matrix H can be found which simultaneously symmetrizes $H\,P$ and yields an $H\,Q$ matrix that is nearly skew symmetric. By defining $W = H\,P$ and $V = H\,Q$, the matrices W and V become

$$W = \begin{bmatrix} \hat{H}\,\hat{P} + \frac{x}{4}A\,A^T & x\,A\,C + \frac{1}{4}\hat{H}\,A \\ x\,A^T\,\hat{P} + \frac{1}{4}D\,A^T & D\,C + \frac{x}{4}A^T\,A \end{bmatrix}$$

$$V = \begin{bmatrix} \hat{H}\,\hat{Q} - \frac{3x}{4}A\,A^T & x\,A\,S + \frac{3}{4}\hat{H}\,A \\ x\,A^T\,\hat{Q} - \frac{3}{4}D\,A^T & D\,S + \frac{3x}{4}A^T\,A \end{bmatrix}$$

Thus, the matrices W and V are important to the stability properties of the spatial operator. Several notes about the structure of W and V should be made at this point. First, the matrices $A\,A^T$ and $A^T\,A$ are zero except for the (N,N) and $(0,0)$ elements, respectively. Second, the matrix $D\,C + \frac{x}{4}A^T\,A$ is automatically symmetric and it has the same tri-diagonal structure as the D and C matrices. Third, the matrix $D\,S + \frac{3x}{4}A^T\,A$ is automatically skew-symmetric which includes the zero at the $(0,0)$ position. The fourth quadrant of W and V automatically satisfy the conditions on the generalized summation-by-parts energy norm. The remaining conditions that W and V must satisfy, written in terms of the submatrices $\hat{H}$, $\hat{P}$, $\hat{Q}$, C, D, S, and A, are

$$\hat{H}\,\hat{P} = (\hat{H}\,\hat{P})^T \qquad \text{(AI. 2)}$$

$$\frac{1}{4}A^T\,\hat{H}^T + x\,C^T\,A^T = \frac{1}{4}D\,A^T + x\,A^T\,\hat{P} \tag{AI. 3}$$

$$\hat{H}\,\hat{Q} + (\hat{H}\,\hat{Q})^T = \frac{3\,x}{2}\,A\,A^T + \lambda\,\delta_{0,0}\,I \tag{AI. 4}$$

$$\frac{3}{4}A^T\,\hat{H}^T + x\,S^T\,A^T = \frac{3}{4}D\,A^T - x\,A^T\,\hat{Q} \tag{AI. 5}$$

where $\lambda\,\delta_{0,0}$ is the non-zero element that occurs in the first row and column of the matrix. This contribution to equation (AI. 4) allows for a non-zero value at the (0,0) element in the matrix V.

By expanding the specific terms in equations (AI. 2 through AI. 5), we have

$$A^T\,\hat{H}^T = \begin{bmatrix} h_{0,n} & \cdots & h_{n,n} \\ 0 & & 0 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ 0 & & 0 \end{bmatrix}; \quad A^T\,\hat{P} = \begin{bmatrix} p_{n,0} & \cdots & p_{n,n} \\ 0 & & 0 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ 0 & & 0 \end{bmatrix}$$

$$A^T\,\hat{Q} = \begin{bmatrix} q_{n,0} & \cdots & q_{n,n} \\ 0 & & 0 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ 0 & & 0 \end{bmatrix}; \quad C^T\,A^T = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \cdot & & & \frac{1}{4} \\ \cdot & & & 0 \\ \cdot & & & \cdot \\ 0 & & & 0 \end{bmatrix}$$

$$S^T\,A^T = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ \cdot & & & \frac{3}{4} \\ \cdot & & & 0 \\ \cdot & & & \cdot \\ 0 & & & 0 \end{bmatrix}; \quad D\,A^T = \begin{bmatrix} 0 & \cdots & 0 & y \\ \cdot & & & x \\ \cdot & & & 0 \\ \cdot & & & \cdot \\ 0 & & & 0 \end{bmatrix}$$

By comparing the matrices involved in equation (AI. 3), it is apparent that

$$\frac{1}{4}\,h_{k,N} + x\,\delta_{k,N} = x\,p_{N,k} + \frac{y}{4}\,\delta_{k,N}; \quad k = 0, N \tag{AI. 6}$$

Similarly, equation (AI. 5) yields the expression

$$\frac{3}{4}\,h_{k,N} = -x\,q_{N,k} + \frac{3\,y}{4}\,\delta_{k,N}; \quad k = 0, N \tag{AI. 7}$$

Eliminating $h_{k,N}$ between equation (AI. 6) and equation (AI. 7) yields the expression

$$q_{N,k} = -3\,p_{N,k} + 3\,\delta_{k,N}; \quad k = 0, N \tag{AI. 8}$$

These properties of the matrices $\hat{P}$ and $\hat{Q}$ must be satisfied regardless of the order properties of the boundary.

We now derive the additional constraints that must be satisfied near the boundaries to guarantee the order properties of these points. Substitution of the equations $u_j = j^r$ and $\frac{du}{dj} = r\,j^{r-1}$ into the matrices $\hat{Q}$ and $\hat{P}$, respectively, yields the constraints that ensure the accuracy of the boundary points. The general expression at the boundary written in terms of an arbitrary accuracy r becomes

$$r \sum_{j=0}^{N} p_{k,j}\, j^{r-1} + \frac{r}{4}\,\delta_{k,N}\,(N+1)^{r-1} = \sum_{j=0}^{N} q_{k,j}\, j^{r} + \frac{3}{4}\,\delta_{k,N}\,(N+1)^{r}; \quad k = 0..N \qquad \text{(AI. 9)}$$

Third-order accuracy at the boundary points requires $r = 0,3$ with $N \geq 3$.

Thus far, we have not specified the exact value of the parameter N. We now specify a precise value for the parameter N so that specific boundary conditions can be derived for the fourth-order interior Pade scheme. To retain the formal accuracy of the interior scheme, the boundary closure must be accomplished to at least third-order accuracy, and requires that $N \geq 3$. For $N = 3$, equation (AI. 9) can be written concisely in matrix notation as

$$\hat{P}\begin{bmatrix} 0*0^{-1} & 1*0^{0} & 2*0^{1} & 3*0^{2} \\ 0*1^{-1} & 1*1^{0} & 2*1^{1} & 3*1^{2} \\ 0*2^{-1} & 1*2^{0} & 2*2^{1} & 3*2^{2} \\ 0*3^{-1} & 1*3^{0} & 2*3^{1} & 3*3^{2} \end{bmatrix} = \hat{Q}\begin{bmatrix} 0^{0} & 0^{1} & 0^{2} & 0^{3} \\ 1^{0} & 1^{1} & 1^{2} & 1^{3} \\ 2^{0} & 2^{1} & 2^{2} & 2^{3} \\ 3^{0} & 3^{1} & 3^{2} & 3^{3} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{3}{4} & \frac{11}{4} & 10 & 36 \end{bmatrix}$$

Solving this expression for the matrix $\hat{Q}$ results in the expression

$$\hat{Q} = \hat{P}\begin{bmatrix} \frac{-11}{6} & 3 & \frac{-3}{2} & \frac{1}{3} \\ \frac{-1}{3} & \frac{-1}{2} & 1 & \frac{-1}{6} \\ \frac{1}{6} & -1 & \frac{1}{2} & \frac{1}{3} \\ \frac{-1}{3} & \frac{3}{2} & -3 & \frac{11}{6} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{7}{24} & \frac{-5}{4} & \frac{17}{8} & \frac{-23}{12} \end{bmatrix}$$

which relates the matrix $\hat{Q}$ to the matrix $\hat{P}$ through third-order accuracy constraints.

We will now solve for the last row of the matrices $\hat{P}$ and $\hat{Q}$ and for the last column of the matrix $\hat{H}$. Equation (AI. 9) is written for $k = N$, and $q_{N,j}$ and $p_{N,j}$ (defined in equation (AI. 8)) are used to yield the relationship

$$r \sum_{j=0}^{N} p_{N,j}\, j^{r-1} + \frac{r}{4}(N+1)^{r-1} = -3 \sum_{j=0}^{N} p_{N,j}\, j^{r} + \frac{3}{4}(N+1)^{r} + 3\,N^{r}; \quad r = 0,3 \quad \text{(AI. 10)}$$

Setting $N = 3$ and solving the system for $p_{3,k}$, $k = 0,3$ yields $p_{3,0} = p_{3,1} = 0$, $p_{3,2} = \frac{1}{4}$, and $p_{3,3} = 1$. Equation (AI. 8) can be used to show that $q_{3,0} = q_{3,1} = 0, q_{3,2} = -\frac{3}{4}$, and $q_{3,3} = 0$. Similarly, equation (AI. 6) yields the values of $h_{k,3}$ as $h_{0,3} = h_{1,3} = 0, h_{2,3} = x$, and $h_{3,3} = y$. Thus, the last row of $\hat{P}$ and $\hat{Q}$ are the same as the interior scheme. In addition, the specific form of the matrix $\hat{H}$ must be

$$\hat{H} = \begin{bmatrix} h_{0,0} & h_{0,1} & h_{0,2} & 0 \\ h_{1,0} & h_{1,1} & h_{1,2} & 0 \\ h_{2,0} & h_{2,1} & h_{2,2} & x \\ h_{3,0} & h_{3,1} & h_{3,2} & y \end{bmatrix}$$

Thus, accuracy constraints on the last row of the matrices $\hat{P}$ and $\hat{Q}$, combined with the structure requirements imposed by equations (AI. 3) and (AI. 5), allow for the direct solution of the last rows of $\hat{P}$ and $\hat{Q}$, and the last column of $\hat{H}$. Multiplying the expression relating $\hat{P}$ to $\hat{Q}$ by the matrix $\hat{H}$, and using the substitutions $\hat{H}\,\hat{Q} = \hat{V}$ and $\hat{H}\,\hat{P} = \hat{W}$ yields the expression for $\hat{V}$ of the form

$$\hat{V} = \hat{W} \begin{bmatrix} \frac{-11}{6} & 3 & \frac{-3}{2} & \frac{1}{3} \\ \frac{-1}{3} & \frac{-1}{2} & 1 & \frac{-1}{6} \\ \frac{1}{6} & -1 & \frac{1}{2} & \frac{1}{3} \\ \frac{-1}{3} & \frac{3}{2} & -3 & \frac{11}{6} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{7x}{24} & \frac{-5x}{4} & \frac{17x}{8} & \frac{-23x}{12} \\ \frac{7y}{24} & \frac{-5y}{4} & \frac{17y}{8} & \frac{-23y}{12} \end{bmatrix}$$

Solving for $\hat{W}$ and $\hat{V}$ such that equation (AI. 2) (where $\hat{W} = \hat{W}^T$) and equation (AI. 4) ($\hat{V} + \hat{V}^T = \frac{3x}{2} A\,A^T + \lambda\,\delta_{0,0}\,I$) are satisfied to obtain

$$\hat{V} = \begin{bmatrix} \frac{-9\alpha}{16} & \frac{1536\gamma+1536\beta-899\alpha}{768} & -\frac{768\gamma+768\beta-703\alpha}{192} & \frac{1536\gamma+1536\beta-1481\alpha}{768} \\ -\frac{1536\gamma+1536\beta-899\alpha}{768} & 0 & \frac{1536\gamma+1536\beta-1277\alpha}{256} & -\frac{768\gamma+768\beta-733\alpha}{192} \\ \frac{768\gamma+768\beta-703\alpha}{192} & -\frac{1536\gamma+1536\beta-1277\alpha}{256} & 0 & \frac{1536\gamma+1536\beta-947\alpha}{768} \\ -\frac{1536\gamma+1536\beta-1481\alpha}{768} & \frac{768\gamma+768\beta-733\alpha}{192} & -\frac{1536\gamma+1536\beta-947\alpha}{768} & \frac{-3\alpha}{32} \end{bmatrix}$$

and

$$\hat{W} = \begin{bmatrix} -\frac{137\alpha-192\gamma}{192} & \frac{-512\gamma-128\beta+525\alpha}{128} & -\frac{145\alpha-144\gamma}{48} & \beta \\ \frac{-512\gamma-128\beta+525\alpha}{128} & -\frac{-816\gamma-384\beta+737\alpha}{48} & \frac{-3072\gamma-1344\beta+3001\alpha}{192} & -\frac{557\alpha-576\gamma}{192} \\ -\frac{145\alpha-144\gamma}{48} & \frac{-3072\gamma-1344\beta+3001\alpha}{192} & -\frac{-3264\gamma-1536\beta+3011\alpha}{192} & \frac{-1536\gamma-384\beta+1543\alpha}{384} \\ \beta & -\frac{557\alpha-576\gamma}{192} & \frac{-1536\gamma-384\beta+1543\alpha}{384} & \gamma \end{bmatrix}$$

with $x = \frac{-\alpha}{8}$ and $y = \alpha$. Three arbitrary parameters remain after all accuracy, symmetry and skew-symmetry conditions are satisfied.

The final step in the discretization is to find a specific form of the matrix $\hat{P}$ that will lead to a simple algorithm. Because the matrix P is tri-diagonal in the interior, the boundary closure should retain the tri-diagonal structure. After $\hat{P}$ is specified, we can solve for the matrix $\hat{H}$ from $\hat{H} = \hat{V}\,\hat{P}^{-1}$ if the inverse of $\hat{P}$ exists, and the last column of $\hat{H}$ is $[0, 0, y, x]^T$. The matrix $\hat{Q}$ follows immediately from $\hat{Q} = \hat{P}\,\hat{V}^{-1}\,\hat{W}$. The last test is to ensure that both $\hat{W}$ and that the full matrix W are positive definite.

Many matrices $\hat{P}$ have been found that satisfy all of the criteria given in the generalized summation-by-parts energy norm analysis. From a numerical perspective, all behaved similarly. The results presented here are those that were the simplest to code. Choosing a specific matrix $\hat{P}$ of the form

$$\hat{P} = \begin{bmatrix} \frac{211}{429} & 1 & 0 & 0 \\ 1 & \frac{3563}{1688} & \frac{-1}{8} & 0 \\ 0 & \frac{43}{17} & \frac{1893}{1054} & \frac{139}{186} \\ 0 & 0 & \frac{1}{4} & 1 \end{bmatrix}$$

yields a matrix $\hat{Q}$ of the form

$$\hat{Q} = \begin{bmatrix} \frac{-289}{234} & \frac{279}{286} & \frac{75}{286} & \frac{-7}{2574} \\ \frac{-8635}{3376} & \frac{6987}{3376} & \frac{1851}{3376} & \frac{-203}{3376} \\ \frac{-15043}{18972} & \frac{-4089}{2108} & \frac{147}{124} & \frac{29353}{18972} \\ 0 & 0 & \frac{-3}{4} & 0 \end{bmatrix}$$

The resulting matrix $\hat{H}$ is therefore

$$\hat{H} = \begin{bmatrix} \frac{70282007653}{7658388480} & \frac{-9426299}{2268480} & \frac{-192913}{1067520} & 0 \\ \frac{-55530689643}{2552796160} & \frac{8051589}{756160} & \frac{149823}{355840} & 0 \\ \frac{63842626133}{2552796160} & \frac{-9153739}{756160} & \frac{-4433}{355840} & \frac{-1}{8} \\ \frac{-71498870443}{7658388480} & \frac{10110149}{2268480} & \frac{102703}{1067520} & 1 \end{bmatrix}$$

¿From a practical point of view, the inconvenient form of the $\hat{H}$ matrix is not of great concern since the matrix H is only inverted once and one column is stored for use.

The matrices $\hat{P}$, $\hat{Q}$, and $\hat{H}$ can be used to establish both the symmetry of the matrix V and the near skew-symmetry of the matrix W. The first six rows and columns of the V matrix are

$$\hat{V} = \begin{bmatrix} \frac{16513}{46080} & \frac{-261}{5120} & \frac{2993}{15360} & \frac{-6223}{46080} & 0 & 0 \\ \frac{-261}{5120} & \frac{9153}{5120} & \frac{-2943}{5120} & \frac{1611}{5120} & 0 & 0 \\ \frac{2993}{15360} & \frac{-2943}{5120} & \frac{7473}{5120} & \frac{-2063}{15360} & \frac{-1}{32} & 0 \\ \frac{-6223}{46080} & \frac{1611}{5120} & \frac{-2063}{15360} & \frac{47953}{46080} & \frac{1}{8} & \frac{-1}{32} \\ 0 & 0 & \frac{-1}{32} & \frac{1}{8} & \frac{15}{16} & \frac{1}{8} \\ 0 & 0 & 0 & \frac{-1}{32} & \frac{1}{8} & \frac{15}{16} \end{bmatrix}$$

The first six rows and columns of the W matrix are

$$\hat{W} = \begin{bmatrix} \frac{-9}{16} & \frac{45}{64} & \frac{-11}{128} & \frac{-7}{128} & 0 & 0 \\ \frac{-45}{64} & 0 & \frac{81}{128} & \frac{9}{128} & 0 & 0 \\ \frac{11}{128} & \frac{-81}{128} & 0 & \frac{41}{64} & \frac{-3}{32} & 0 \\ \frac{7}{128} & \frac{-9}{128} & \frac{-41}{64} & 0 & \frac{3}{4} & \frac{-3}{32} \\ 0 & 0 & \frac{3}{32} & \frac{-3}{4} & 0 & \frac{3}{4} \\ 0 & 0 & 0 & \frac{3}{32} & \frac{-3}{4} & 0 \end{bmatrix}$$

As shown, the matrix W is nearly skew symmetric, and the matrix V is symmetric. For the matrix W is positive definite, it is necessary to show that every submatrix is positive definite. The inner scheme is diagonally dominant and contributes to the definiteness of the complete matrix W. However, the

boundary elements are not diagonally dominant, and suppress the positive-definiteness. The 4×4 boundary matrix $\hat{W}' = \hat{W} + \frac{r}{4} A A^T$ has the following characteristic polynomial

$$754974720\,\lambda^4 - 3507814400\,\lambda^3 + 5299068928\,\lambda^2 - 3079323424\,\lambda + 536779791 = 0 \quad \text{(AI. 11)}$$

The symmetry of the $\hat{W}'$ matrix and the alternating signs of each term in the characteristic polynomial guarantee that the matrix is positive definite. The characteristic polynomial of every submatrix (up to ten points, which includes four boundary and six interior points) of the matrix W results in a positive definite matrix. No proof that the complete discretization is positive definite for an arbitrary number of interior points has been found.

The accuracy of the new scheme is third order at the boundaries and fourth order in the interior. To show this, the Taylor expansion for long wavelength modes is made using the stencil at each of the first four points. The results are

$$\begin{aligned} & i\,\xi + \frac{17}{640}\xi^4 + \ldots \\ & i\,\xi + \frac{43}{2016}\xi^4 + \ldots \\ & i\,\xi - \frac{98}{2005}\xi^4 + \ldots \\ & i\,\xi - i\,\frac{1}{180}\xi^5 + \ldots \end{aligned} \quad \text{(AI. 12)}$$

At high resolution, the boundary points behave with third-order truncation error; the interior behaves with fourth-order error. Therefore, the resulting scheme is formally fourth-order accurate.

APPENDIX II

Sixth Order Explicit Scheme

Here, we derive an explicit scheme that is formally sixth-order accurate. Unlike the fourth-order compact case presented earlier, the matrix H can be the identity matrix. To constrain the matrix $\hat{P}$ to be symmetric and the matrix $\hat{Q}$ be nearly skew symmetric, six alternative formulas are required at the boundaries, each of which is closed to fifth-order accuracy to retain the formal accuracy. The corner 7×7 submatrices of the global matrices $\hat{P}$ and $\hat{Q}$ can be written as

$$P_6 = \begin{bmatrix} \frac{2113}{10800} & \frac{18487}{345600} & \frac{553}{57600} & \frac{14759}{172800} & \frac{(-29269)}{172800} & \frac{54839}{345600} & 0 \\ \frac{18487}{345600} & \frac{175781}{51840} & \frac{(-28361)}{6912} & \frac{129329}{34560} & \frac{(-346319)}{207360} & \frac{(-19061)}{172800} & 0 \\ \frac{553}{57600} & \frac{(-28361)}{6912} & \frac{43807}{5184} & \frac{(-915)}{128} & \frac{126833}{34560} & \frac{(-39307)}{518400} & 0 \\ \frac{14759}{172800} & \frac{129329}{34560} & \frac{(-915)}{128} & \frac{67769}{8640} & \frac{(-25289)}{6912} & \frac{34811}{172800} & 0 \\ \frac{(-29269)}{172800} & \frac{(-346319)}{207360} & \frac{126833}{34560} & \frac{(-25289)}{6912} & \frac{156053}{51840} & \frac{(-21059)}{115200} & 0 \\ \frac{54839}{345600} & \frac{(-19061)}{172800} & \frac{(-39307)}{518400} & \frac{34811}{172800} & \frac{(-21059)}{115200} & \frac{32569}{32400} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$Q_6 = \begin{bmatrix} \frac{(-1)}{2} & \frac{1235503}{1036800} & \frac{(-859597)}{518400} & \frac{398}{225} & \frac{(-603059)}{518400} & \frac{14969}{41472} & 0 \\ \frac{(-1235503)}{1036800} & 0 & \frac{16343}{5760} & \frac{(-68005)}{20736} & \frac{186797}{69120} & \frac{(-184657)}{172800} & 0 \\ \frac{859597}{518400} & \frac{(-16343)}{5760} & 0 & \frac{128759}{51840} & \frac{(-18743)}{6912} & \frac{3799}{2700} & 0 \\ \frac{(-398)}{225} & \frac{68005}{20736} & \frac{(-128759)}{51840} & 0 & \frac{110351}{51840} & \frac{(-607693)}{518400} & \frac{1}{60} \\ \frac{603059}{518400} & \frac{(-186797)}{69120} & \frac{18743}{6912} & \frac{(-110351)}{51840} & 0 & \frac{376549}{345600} & \frac{(-3)}{20} \\ \frac{(-14969)}{41472} & \frac{184657}{172800} & \frac{(-3799)}{2700} & \frac{607693}{518400} & \frac{(-376549)}{345600} & 0 & \frac{3}{4} \\ 0 & 0 & 0 & \frac{(-1)}{60} & \frac{3}{20} & \frac{(-3)}{4} & 0 \end{bmatrix}$$

The characteristic polynomial of the matrix P_6 is

$$\begin{aligned} &1039973956284579840000000\,\lambda^6 - 2485126099162449838080000000\,\lambda^5 \\ &+ 1003578630643249838161920000\,\lambda^4 - 163903822337723736805171200\,\lambda^3 \end{aligned}$$

$$+ 1248376737213799711434406800\,\lambda^2 \;-\; 41223536504281663355919744 0\,\lambda$$
$$+ 37455444120716264727507839 \;=\; 0 \qquad \text{(AII. 1)}$$

The symmetry of the matrix P_6 and the alternating signs of the terms in the polynomial are sufficient for positive definiteness of both the matrix P_6 and the global matrix P.

The truncation error at the boundary points is

$$\begin{aligned}
i\xi &+ \frac{64482999974515473972444 67\xi^6}{22473266472429758836504703 4} + \cdots \\
i\xi &- \frac{55178459341997062554732 1\xi^6}{11236633236214879418252351 70} + \cdots \\
i\xi &- \frac{90378114042816098962729619\xi^6}{22473266472429758836504703 40} + \cdots \\
i\xi &+ \frac{62520732887440126777806839\xi^6}{22473266472429758836504703 40} + \cdots \\
i\xi &+ \frac{21521021082694965917733 1\xi^6}{11236633236214879418252351 70} + \cdots \\
i\xi &- \frac{71015802541971163020539 05\xi^6}{22473266472429758836504703 4} + \cdots
\end{aligned} \qquad \text{(AII. 2)}$$

which indicates fifth-order accuracy at the six boundary points.

APPENDIX III

Eigenvalues of the Discrete System

The eigenvalues of the semi-discrete system are used in the results section to compare the theoretical and the numerical stability boundaries. The model equation is the hyperbolic system used in the main text and defined by equations (41), and (42). For convenience, we define the $(N+1)\times(N+1)$ matrix $A = P^{-1}Q$. The matrix A contains all the information from the spatial discretization operator $\frac{\partial}{\partial x}$. The semi-discrete form of equation (41) becomes

$$\begin{aligned} \frac{du}{dt} + A\,u &= 0, \\ \frac{dv}{dt} - A\,v &= 0, \quad t \geq 0 \end{aligned} \tag{AIII. 1}$$

with the boundary conditions defined by equation (42). In matrix notation, the discrete system takes the form

$$\frac{\partial \vec{u}}{\partial t} = \begin{bmatrix} A^{\dagger} & . & \alpha B \\ . & . & . \\ \beta J^{-1} B J & . & J^{-1} A^{\dagger} J \end{bmatrix} \vec{u}$$

where

$$\vec{u} = \begin{bmatrix} u_1 \\ : \\ u_{N-1} \\ u_N \\ v_0 \\ v_1 \\ : \\ v_{N-1} \end{bmatrix}; \quad A^{\dagger} = \begin{bmatrix} a_{1,1} & a_{1,2} & . & . & . & a_{1,N-1} & a_{1,N} \\ a_{2,1} & a_{2,2} & . & . & . & a_{2,N-1} & a_{2,N} \\ . & . & & & & . & . \\ . & . & & & & . & . \\ . & . & & & & . & . \\ a_{N-1,1} & a_{N-1,2} & . & . & . & a_{N-1,N-1} & a_{N-1,N} \\ a_{N,1} & a_{N,2} & . & . & . & a_{N,N-1} & a_{N,N} \end{bmatrix}$$

and

$$B = \begin{bmatrix} a_{1,0} & 0 & 0 & . & . & . & 0 \\ a_{2,0} & 0 & 0 & . & . & . & 0 \\ . & . & . & & & & . \\ . & . & . & & & & . \\ . & . & . & & & & . \\ a_{N-1,0} & 0 & 0 & . & . & . & 0 \\ a_{N,0} & 0 & 0 & . & . & . & 0 \end{bmatrix}; \quad J = \begin{bmatrix} & & & & & & 1 \\ & & & & & 1 & \\ & & & & . & & \\ & & & . & & & \\ & & . & & & & \\ & 1 & & & & 0 & \\ 1 & & & & & & \end{bmatrix}$$

Note that $JJ = I$, so that $J = J^{-1}$. The vector $\vec{u}$ is the concatenated vector of discrete values from the scalar vectors u and v with the elements u_0 and v_N removed. These elements are removed because the physical boundary condition relates them to known elements in the vector $\vec{u}$, so that and need not need to solve for them. The matrix $A^\dagger$ is the $N \times N$ submatrix of A which is obtained by eliminating the zeroth row and zeroth column. Note that this was the matrix that was analyzed in the scalar analysis to determine time stability of the spatial operator. The matrix B is zero everywhere except the first column, where the zeroth column of the original A matrix is written. This column is precisely the coupling between the u and v vector, which occurs at the boundary.

It is instructive to relate the system eigenvalues to those obtained in the scalar analysis $[(A^\dagger - \lambda I)\, u = 0]$. By defining the matrix H^{-1} and H as

$$H^{-1} = \frac{1}{\sqrt{2\sqrt{\alpha\beta}}} \begin{bmatrix} \sqrt{\beta}\, I & . & \sqrt{\alpha}\, J \\ . & . & . \\ -\sqrt{\beta}\, I & . & \sqrt{\alpha}\, J \end{bmatrix}; \quad H = \frac{1}{\sqrt{2\sqrt{\alpha\beta}}} \begin{bmatrix} \sqrt{\alpha} I & . & -\sqrt{\alpha} I \\ . & . & . \\ \sqrt{\beta} J & . & \sqrt{\beta} J \end{bmatrix}$$

with $H^{-1} H = H H^{-1} = I$, we note that the system matrix can be made block diagonal with the similarity transform H

$$\frac{1}{\sqrt{2\sqrt{\alpha\beta}}} \begin{bmatrix} \sqrt{\beta} I & . & \sqrt{\alpha} J \\ . & . & . \\ -\sqrt{\beta} I & . & \sqrt{\alpha} J \end{bmatrix} \begin{bmatrix} A^\dagger & . & \alpha B \\ . & . & . \\ \beta J^{-1} B J & . & J^{-1} A^\dagger J \end{bmatrix} \begin{bmatrix} \sqrt{\alpha} I & . & -\sqrt{\alpha} I \\ . & . & . \\ \sqrt{\beta} J & . & \sqrt{\beta} J \end{bmatrix} \frac{1}{\sqrt{2\sqrt{\alpha\beta}}} =$$

$$\begin{bmatrix} A^\dagger + \sqrt{\alpha\beta}\, B\, J & . & 0 \\ . & . & . \\ 0 & . & A^\dagger - \sqrt{\alpha\beta}\, B\, J \end{bmatrix}$$

For scalar time-stable spatial schemes, the eigenvalues of the matrix $A^\dagger$ are bounded to the left half-plane. Note that for $\alpha = 0$ (or $\beta = 0$) the contribution from the boundary coupling matrix B is identically zero, and the eigenvalues of the resulting system are simply the scalar eigenvalues with a multiplicity of two. For non-zero values of the parameters α and β, the eigenvalues of the total matrix are different from those of the original matrix $A^\dagger$. Also note that two distinct eigenvalue scenarios exist for the boundary parameters α and β, depending on whether their signs are equal or opposite.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1993	Contractor Report

4. TITLE AND SUBTITLE
TIME-STABLE BOUNDARY CONDITIONS FOR FINITE-DIFFERENCE SCHEMES SOLVING HYPERBOLIC SYSTEMS: METHODOLOGY AND APPLICATION TO HIGH-ORDER COMPACT SCHEMES

5. FUNDING NUMBERS
C NAS1-19480
C NAS1-18605

WU 505-90-52-01

6. AUTHOR(S)
Mark H. Carpenter, David Gottlieb, and Saul Abarbanel

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Computer Applications in Science and Engineering
Mail Stop 132C, NASA Langley Research Center
Hampton, VA 23681-0001

8. PERFORMING ORGANIZATION REPORT NUMBER
ICASE Report No. 93-9

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Langley Research Center
Hampton, VA 23681-0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NASA CR-191436
ICASE Report No. 93-9

11. SUPPLEMENTARY NOTES
Langley Technical Monitor: Michael F. Card
Final Report

Submitted to Journal of Computational Physics

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified - Unlimited

Subject Category 64

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
We present a systematic method for constructing boundary conditions (numerical and physical) of the required accuracy, for compact (Pade-like) high-order finite-difference schemes for hyperbolic systems. First a proper summation-by-parts formula is found for the approximate derivative. A "simultaneous approximation term" (SAT) is then introduced to treat the boundary conditions. This procedure leads to time-stable schemes even in the system case. An explicit construction of the fourth-order compact case is given. Numerical studies are presented to verify the efficacy of the approach.

14. SUBJECT TERMS
time-stability; high-order finite difference; compact

15. NUMBER OF PAGES
36

16. PRICE CODE
A03

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified		

NSN 7540-01-280-5500

Standard Form 298 (Rev 2-89)
Prescribed by ANSI Std Z39-18
298-102

NASA-Langley, 1993