Journal of Scientific Computing, Vol. 18, No. 1, February 2003 (© 2003)

Boundary Procedures for Summation-by-Parts Operators

Ken Mattsson[1]

Received June 28, 2001; accepted (in revised form) January 22, 2002

Four different methods of imposing boundary conditions for the linear advection-diffusion equation and a linear hyperbolic system are considered. The methods are analyzed using the energy method and the Laplace transform technique. Numerical calculations are done, considering in particular the case when the initial data and boundary data are inconsistent.

KEY WORDS: High order finite difference; numerical stability.

1. INTRODUCTION

Problems in computational fluid dynamics (CFD) require accurate and stable numerical methods. Computational resources at hand are often insufficient to allow for resolution of solution features using low-order methods (second order or less). Thus, the computationally more efficient high-order methods (HOM) are needed. The drawback with HOM is the complicated boundary treatment required to get a stable method.

In this paper, we consider accurate and stable finite difference approximations of the linear advection-diffusion equation and of a linear hyperbolic system. We will consider *strict stability* among the variety of different stability definitions [6] available. A strictly stable approximation has the same asymptotic time-growth as the continuous problem. Finite difference operators that obey a summation by part property (SBP) [14] result in a strictly stable approximation for a Cauchy problem. Nevertheless, the SBP

[1] Department of Scientific Computing, Information Technology, Uppsala University, P.O. Box 120, S-751 04 Uppsala, Sweden. E-mail: ken@tdb.uu.se

0885-7474/03/0200-0133/0 © 2003 Plenum Publishing Corporation

property alone does not guarantee that the time-growth of the numerical approximation is correct. Specific boundary treatment is also required.

A strictly stable approximation allows for an initial growth of the solution-energy, before the asymptotic decay sets in. This could trigger an instability. Nevertheless, if the numerical approximation can be shown to have a non-growing solution energy, this initial growth can not occur. Hence, guaranteeing non-growing energy and getting correct asymptotic time-growth, require strict stability and an energy estimate.

To impose the boundary condition (BC) explicitly, i.e., to combine the difference operator and the boundary operator into a modified operator, usually destroy the SBP property. In general, this makes it impossible to obtain an energy estimate. This boundary procedure, often used in practical calculations, is referred to as the injection method and can result in an unwanted exponential growth of the solution energy.

The basic idea behind the Simultaneous Approximation Term (SAT) method [2] and the projection method [11, 12] is to impose the BC such that the SBP-property is preserved and such that we get an energy estimate. A systematic comparison between the SAT method, the projection method, and the injection method was first done in [14]. This was done for a hyperbolic scalar equation and a hyperbolic system. For those problems the projection method works fine.

In this paper, we also consider the linear advection-diffusion equation for which the projection method fail to be a strictly stable approximation. We will also study a similar method—a cure for the Projection method—which was introduced in [5] as a "hybrid" between the injection method and the projection method.

To motivate the use of HOM, we compare the numerical result of a vortex-calculation, using a 2nd, 4th and 6th order SAT method, see Fig. 1. The vortex model is presented in [4] and satisfies the two-dimensional Euler equations, under the assumption of isentropy. In [13], it is shown that the solution is steady in the frame of reference moving with the free-stream. Clearly, the second-order case displays inaccurate results.

For a pure convection problem, the major source of error is the dispersive errors. The vortex field consists of several Fourier modes, which all advect at different speeds. The phase error (the difference between the correct and the computed location of a Fourier mode) vary with both grid-size ($h = 1/N$), frequency and time. A calculation with a 6th order accurate method, using N grid-points in each of the spatial directions, result in a phase-error proportional to $(1/N)^6$. In [8] it is shown that a 2nd order accurate method require approximately N^3 grid-points in each of the spatial directions to reproduce the same phase-error. In Fig. 1, the calculations are done using 330×90 points, and are shown at time $t = 100$.

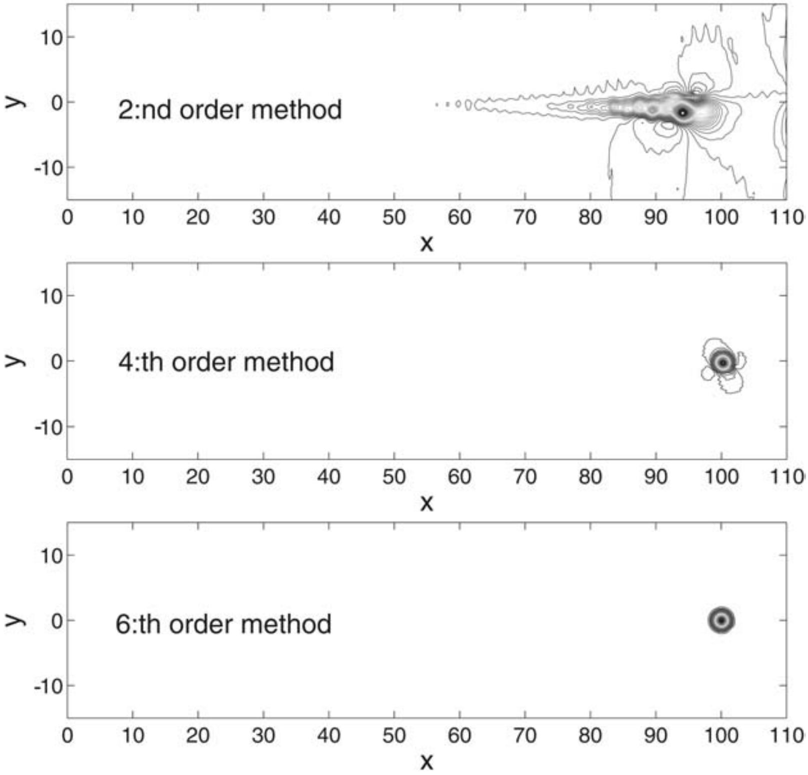


Fig. 1. Pressure contour, $t = 100$.

2. TIME GROWTH

Consider the problem,

$$\begin{aligned} w_t + Qw &= F(x, t) \\ Lw(0, t) &= g(t) \\ w(x, 0) &= f(x) \end{aligned} \tag{1}$$

where Q is the differential operator, L the boundary operator, F the forcing function, f the initial data and g the boundary data. Consider also the slightly perturbed version, with solution v and initial data $f + \delta f$. The equation for the error $u = w - v$, becomes

$$\begin{aligned} u_t + Qu &= 0 \\ Lu(0, t) &= 0 \\ u(x, 0) &= \delta f(x) \end{aligned} \tag{2}$$

If (1) is well posed [6], an energy estimate exists

$$\|u\| \leqslant K_c e^{\alpha_c t} \|\delta f\| \tag{3}$$

where K_c and α_c do not depend on t. For a strictly stable difference approximation of (2), a corresponding discrete energy estimate exists

$$\|u\|_H \leqslant K_d e^{\alpha_d t} \|\delta f\|_H \tag{4}$$

where $\alpha_d \leqslant \alpha_c + \mathcal{O}(h)$. This means that a strictly stable approximation has the same asymptotic time-growth as the continuous problem. Note that $K_c \neq K_d$ in general.

2.1. The Continuous Problem

Consider the scalar problem

$$\begin{aligned} u_t &= Pu & 0 \leqslant x \leqslant 1, \quad t \leqslant 0 \\ L_0 u &= 0, & L_1 u = 0 \\ u(x, 0) &= f(x) \end{aligned} \tag{5}$$

where P is a linear operator. We will show that the time growth is governed by the eigenvalue to $-P$, with largest real part. Consider the scalar product $(u, v)_\rho = \int_0^1 uv\rho(x)\, dx$, where $\rho > 0$ and the corresponding norm $(u, u)_\rho = \|u\|_\rho^2$. If u and v satisfy the boundary conditions then it is always possible [1] to find a weight function $\rho(x)$ such that P is symmetric, i.e., $(u, Pv)_\rho = (Pu, v)_\rho$. This implies that all eigenvalues λ_i are real and that the eigenvectors w_i corresponding to different eigenvalues are orthogonal, i.e., $Pw_i = \lambda_i w_i$ and $(w_i, w_j)_\rho = \delta_{ij}$. Expand the initial data and the solution in eigenfunctions, i.e., $f = \sum_{i=1}^{\infty} \sigma_i w_i$, $u = \sum_{i=1}^{\infty} e^{-\lambda_i t} \sigma_i w_i$. This gives, $\|f\|_\rho^2 = \sum_{i=1}^{\infty} |\sigma_i|^2$, $\|u\|_\rho^2 = \sum_{i=1}^{\infty} e^{-2\Re(\lambda_i)\, t} |\sigma_i|^2$ and the energy estimate

$$\|u\|_\rho \leqslant e^{\Re(\lambda_{\max})\, t} \|f\|_\rho \tag{6}$$

where $\lambda_{\max}$ is the eigenvalue to $-P$, with largest real part. The estimate (6) is valid for any linear operator P with continuous coefficients.

In Sec. 3.1 we consider the problem (13), where $P = -a \frac{\partial}{\partial x} + \epsilon \frac{\partial^2}{\partial x^2}$. The weight function $e^{-\frac{a}{\epsilon} x}$ lead to $(u, Pv)_\rho = (Pu, v)_\rho$, if u and v satisfy the boundary conditions. Hence, the energy estimate (6) holds. To obtain the estimate (6) in the l_2-norm we use $\|u\|^2 = \int_0^1 u^2\, dx \leqslant e^{\frac{a}{\epsilon}} \|u\|_\rho^2$, $\|f\|_\rho^2 = \int_0^1 f^2 e^{-\frac{a}{\epsilon} x}\, dx \leqslant \|f\|^2$, which finally give

$$\|u\| \leqslant e^{\frac{a}{2\epsilon}} e^{\Re(\lambda_{\max})\, t} \|f\| \tag{7}$$

2.2. The Discrete Problem

To relate the discrete spectrum to the energy estimate consider,

$$\begin{aligned} u_t &= Au \\ u(0) &= f \end{aligned} \tag{8}$$

We introduce $u = H^{-1/2}v$ and $\tilde{A} = H^{1/2}AH^{-1/2}$. The symmetric positive definite matrix H defines the norm, $u^T Hu \equiv \|u\|_H^2 = \|v\|^2 = v^T v$. The matrices A and $\tilde{A}$ have the same eigenvalues.

By Schur's Lemma [6], there is a unitary transformation U such that

$$U^* \tilde{A} U = \begin{bmatrix} \lambda_1 & \tilde{a}_{12} & \cdots & \cdots & \tilde{a}_{1m} \\ & \lambda_2 & \tilde{a}_{23} & \cdots & \tilde{a}_{2m} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \tilde{a}_{m-1,m} \\ 0 & & & & \lambda_m \end{bmatrix}$$

has upper triangular form. Let $D = \text{diag}[1, d, d^2,..., d^{m-1}]$ be a diagonal matrix. Then $D^{-1}U^*\tilde{A}UD = \Lambda + R$, where

$$R = \begin{bmatrix} 0 & d\tilde{a}_{12} & \cdots & \cdots & d^{m-1}\tilde{a}_{1m} \\ 0 & 0 & d\tilde{a}_{23} & \cdots & d^{m-2}\tilde{a}_{2m} \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & d\tilde{a}_{m-1,m} \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_m \end{bmatrix}$$

By introducing $v = U\,D\,w$ and $\bar{A} = D^{-1}U^*\tilde{A}UD$, Eq. (8) becomes

$$\begin{aligned} w_t &= \bar{A}w \\ w(0) &= D^{-1}U^*H^{1/2}f \end{aligned} \tag{9}$$

The energy method gives $\frac{d}{dt}\|w(\cdot,t)\|^2 = 2\,\|\bar{A}w(\cdot,t)\|^2 \leqslant 2\,|\bar{A}|\,\|w(\cdot,t)\|^2$, which implies $\|w(\cdot,t)\|^2 \leqslant \|w(\cdot,0)\|^2\,e^{2(\lambda_{\max}+\mathcal{O}(d))\,t}$, where $\lambda_{\max}$ is the largest real part of the spectrum. We use the properties

$$\begin{aligned} &\|u(\cdot,t)\|_H^2 = \|v(\cdot,t)\|^2 = \|Dw(\cdot,t)\|^2 \leqslant |D|^2\,\|w(\cdot,t)\|^2, \\ &\|w(\cdot,0)\|^2 = \|D^{-1}v(\cdot,0)\|^2 \leqslant |D^{-1}|^2\,\|v(\cdot,0)\|^2 = |D^{-1}|^2\,\|u(\cdot,0)\|_H^2 \end{aligned}$$

to get the energy estimate for the original variable u,

$$\|u(\cdot, t)\|_H \leqslant \frac{1}{d^{m-1}} \|u(\cdot, 0)\|_H \, e^{(\lambda_{\max} + \mathcal{O}(d))\, t} \tag{10}$$

For the problems considered in this article, the real part of the continuous spectrum is non-positive. The numerical approximations using the SAT-method and the projection method will result in an energy estimate

$$\|u(\cdot, t)\|_H \leqslant \|u(\cdot, 0)\|_H \tag{11}$$

see Secs. 3.3.1, 3.3.3, 4.2.1 and 4.2.2. The estimates (10) and (11) can then be combined to

$$\|u(\cdot, t)\|_H \leqslant \min \begin{cases} \frac{1}{d^{m-1}} \|u(\cdot, 0)\|_H \, e^{(\lambda_{\max} + \mathcal{O}(d))\, t} & (\text{Energy}_1) \\ \|u(\cdot, 0)\|_H & (\text{Energy}_2) \end{cases} \tag{12}$$

In general it is not possible to obtain an energy estimate for the injection method and the modified projection method, since the SBP property is lost.

Assume that the real part of the continuous spectrum is non-positive. From (12) we can conclude that if the numerical approximation result in an energy estimate (10) and if the discrete spectrum converge to the continuous spectrum, then the solution is bounded for all times and we get the correct asymptotic decay of the energy, see Fig. 2. For some problems the

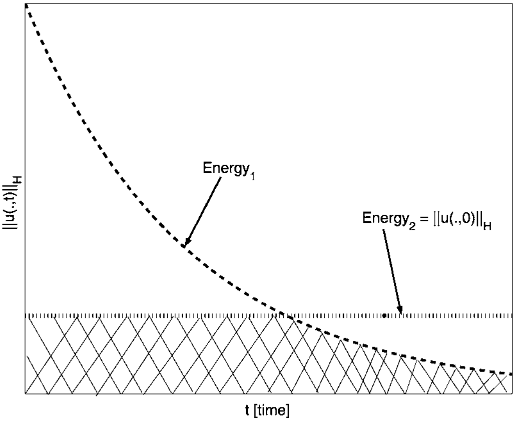


Fig. 2. The energy is inside the marked region, see (12). Here we have exponential decay of the energy.

term $\frac{1}{d^{m-1}}$ can be very large. However, for larger time scales the exponential term $e^{(\lambda_{\max}+\mathcal{O}(d))\,t}$ will dominate and bring the energy down.

3. THE SCALAR PROBLEM

3.1. The Continuous Problem

Consider the model problem

$$\begin{aligned} u_t + au_x &= \epsilon u_{xx} && 0 \leqslant x \leqslant 1, \quad t \geqslant 0 \\ u(0,t) + \beta u_x(0,t) &= 0, && u_x(1,t) = 0 \\ u(x,0) &= f(x) \end{aligned} \tag{13}$$

where a , $\epsilon > 0$ and $a \gg \epsilon$. The energy method on (13) leads to

$$\frac{d}{dt}\|u\|^2 = -au^2(1,t) + \left(a + \frac{2\epsilon}{\beta}\right) u^2(0,t) - 2\epsilon \, \|u_x\|^2$$

The problem (13) has an energy estimate if

$$-\frac{2\epsilon}{a} \leqslant \beta \leqslant 0 \tag{14}$$

Provided that the solution exist (can be shown by using Laplace-transform technique or via difference approximations: see [7] and [10]), we can conclude that (13) is well-posed.

3.2. The Spectrum of the Continuous Problem

To verify that the numerical approximations have the correct asymptotic time-growth, we compare the eigenvalues of the semi-discrete problem, see Sec. 3.4, to the corresponding continuous eigenvalues. Laplace transformation of (13) with zero initial data leads to

$$\begin{aligned} s\hat{u} + a\hat{u}_x &= \epsilon \hat{u}_{xx} && 0 \leqslant x \leqslant 1 \\ \hat{u}(0) + \beta \hat{u}_x(0) &= 0, && \hat{u}_x(1) = 0, \end{aligned} \tag{15}$$

with the general solution $\hat{u} = \sigma_1 e^{r_1 x} + \sigma_2 e^{r_2 x}$. The roots

$$r_{1,2} = \frac{a}{2\epsilon}\left(1 \pm \sqrt{1 + \frac{4s\epsilon}{a^2}}\right) \tag{16}$$

are given by solving the characteristic equation. The unknown constants $\sigma_{1,2}$ are determined by the boundary conditions, which lead to a linear equation system $\mathbf{C}\sigma = \mathbf{0}$. The problem (13) is well posed if

$$\det \mathbf{C}(\mathbf{s}) = \begin{vmatrix} 1+\beta r_1 & 1+\beta r_2 \\ r_1 e^{r_1} & r_2 e^{r_2} \end{vmatrix} \neq 0 \qquad \text{for} \quad \operatorname{Re}(s) \geqslant 0 \tag{17}$$

The spectrum of the continuous problem (continuous spectrum) consist of s-values making $\det C(s) = 0$.

Note that if $s = -\frac{a^2}{4\epsilon}$ we have a double root, $r = \frac{a}{2\epsilon}$. The general solution now becomes $\hat{u} = (\sigma_1 + \sigma_2 x)\, e^{rx}$. The unknown constants $\sigma_{1,2}$ are again determined by the boundary conditions, which give a linear system $\mathbf{C}\sigma = \mathbf{0}$. The point $s = -\frac{a^2}{4\epsilon}$ belongs to the spectrum if

$$\det C(s) = \begin{vmatrix} 1+\beta r & \beta \\ re^r & (1+r)\, e^r \end{vmatrix} = 0 \tag{18}$$

with $r = \frac{a}{2\epsilon}$. This implies that

$$(1+\beta r)(1+r)\, e^r - \beta r e^r = 0 \quad \Rightarrow \quad \beta = -\frac{1+r}{r^2} = -\frac{4\epsilon^2}{a^2} - \frac{2\epsilon}{a}$$

But from (14) we have a restriction on β which contradicts (18) for $\epsilon \neq 0$, and consequently $r = \frac{a}{2\epsilon}$ is a false double root.

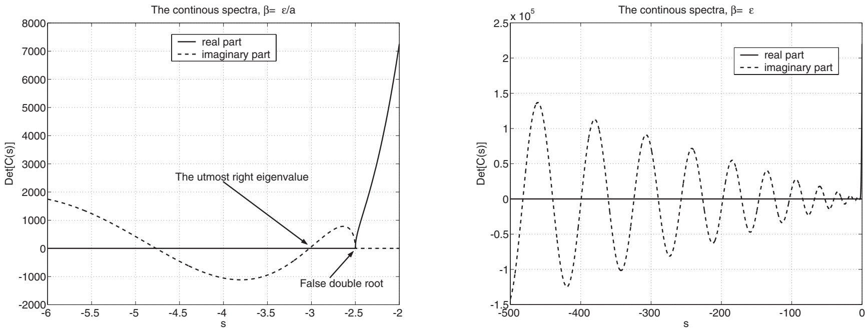
The continuous spectrum is given by (17), which can be written

$$\det C(s) = (1+\beta r_2)\, r_1 e^{r_1} + (1+\beta r_1)\, r_2 e^{r_2} = 0 \tag{19}$$

where the two roots r_1 and r_2 are given by (16). Equation (19) is nonlinear in s and can be solved with an iterative method such as Newton–Raphson. The result is presented in Table I. The spectrum does not contain any imaginary parts.

Table I. The Two Utmost Right Values of the Spectrum; $\epsilon = 0.1$ and $a = 1$

β	Spectrum (1)	Spectrum (2)
$-\epsilon/2a$	-3.13023510	-5.15428651
$-\epsilon/a$	-3.02187288	-4.76698726
$-2\epsilon/a$	-2.67261695	-4.12696691



(a) (b)

Fig. 3. $\det C(s)$. $\beta = -\epsilon/a$, $\epsilon = 0.1$ and $a = 1$. (a) The two utmost right values of the spectrum, see Table I; (b) Part of the spectrum in the range $s \in (-500, 0)$.

Part of the spectrum is shown in Fig. 3. The values on β are identical to those in Table I. There is a root where both the imaginary part and the real part of $\det C(s)$ are zero, except for the false double root at $s = -2.5$.

3.3. The Semi-Discrete Problem

In this section we will discretize the problem (13), using second, fourth and sixth order accurate difference operators. These operators were presented in [3]. The domain $(0 \leqslant x \leqslant 1)$ is discretized using N equidistant grid points, where $x_j = \frac{j-1}{N-1}$, $j = 1, 2,..., N$.

The numerical approximation at grid-point x_j is denoted v_j. The derivative u_x is approximated with a consistent finite difference approximation $D_1 v$. A finite difference operator satisfying the SBP property is written $D_1 \equiv H^{-1}Q$, where $Q + Q^T = D = \text{diag}(-1, 0,..., 0, 1)$ and $H = H^T > 0$. The derivative u_{xx} is approximated with a consistent finite difference approximation $D_2 v$. A finite difference operator satisfying the SBP property is written $D_2 \equiv H^{-1}(-M + DS)$, where $M + M^T \geqslant 0$ and DS is an approximation of the first derivative operator at the boundary, to design accuracy. The boundary operators are approximated by L_l at the left boundary and L_r at the right boundary. The discrete problem now looks like

$$\begin{aligned} v_t + aD_1 v &= \epsilon D_2 v \\ L_l v = g_l(t), &\qquad L_r v = g_r(t) \\ v(0) &= f \end{aligned}$$

In our model problem (1) we have $g_l(t) = g_r(t) = 0$. In [3] it is shown that the operator M is positive semi-definite and that DS is an approximation of the first derivative operator at the boundary, to designed accuracy. All operators are given in [9].

3.3.1. *The SAT Method*

In this method, first introduced by Carpenter *et al.* [2], a "Simultaneous Approximation Term" is introduced to implement the boundary conditions. This procedure solves a linear combination of the boundary condition and the differential equation near the boundary. The SAT-method leads to

$$\begin{aligned} &v_t + aD_1 v = \epsilon D_2 v - \frac{H^{-1}}{2}\left(\tau_l \hat{e}_1 \{L_l v - g_l(t)\} + \tau_r \hat{e}_N \{L_r v - g_r(t)\}\right) \\ &v(0) = f \end{aligned} \tag{20}$$

where $\hat{e}_1 = [1, 0,\dots, 0]^T$ and $\hat{e}_N = [0,\dots, 0, 1]^T$. For our model problem $g_l(t) = g_r(t) = 0$. The parameters τ_l and τ_r can be tuned in order to give a stable scheme.

The energy method on (20) gives

$$\frac{d}{dt}\|v\|_H^2 \leqslant -av_N^2 + v_1^2(a - \tau_l) + v_N (D_r v)_N\ (2\epsilon - \tau_r) - v_1 (D_l v)_I\ (2\epsilon + \tau_l \beta)$$

To bound the error, the second and third parenthesis must vanish, and the first parenthesis must be non positive. If condition (14) for well-posedness is fulfilled, then

$$\tau_l = -\frac{2\epsilon}{\beta}, \qquad \tau_r = 2\epsilon \tag{21}$$

will guarantee a bounded solution.

3.3.2. *The Injection Method*

The injection method imposes the BC explicitly, i.e., it combines the difference operator and the boundary operator into a modified operator. This procedure might destroy the SBP property and make it difficult to prove stability [6]. The injection method can be written,

$$\begin{aligned} &v_t + aD_1 v = \epsilon D_2 v \\ &\qquad v_0 = \tilde{C}_l \tilde{v} + \alpha_l g_l(t), \qquad v_n = \tilde{C}_r \tilde{v} + \alpha_r g_r(t) \\ &v(0) = f \end{aligned} \tag{22}$$

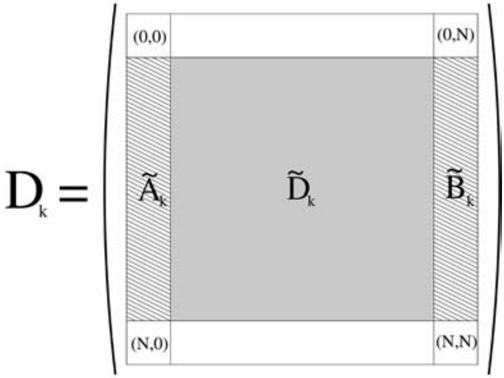


Fig. 4. Schematic picture of the modified operators for the injection method. D_k symbolizes the difference operator approximating $\partial^k/\partial x^k$.

By removing the first and the last row in the difference operators, and replacing the two boundary points using the corresponding boundary conditions (see Fig. 4 for the notation), we obtain

$$\begin{aligned}\tilde{v}_t &= (-a\tilde{D}_1 + \epsilon\tilde{D}_2 + (-a\tilde{A}_1 + \epsilon\tilde{A}_2)\,\tilde{C}_l + (-a\tilde{B}_1 + \epsilon\tilde{B}_2)\,\tilde{C}_r)\,\tilde{v}\\ &\quad + \alpha_l(-a\tilde{A}_1 + \epsilon\tilde{A}_2)\,g_l(t) + \alpha_r(-a\tilde{B}_1 + \epsilon\tilde{B}_2)\,g_r(t) \qquad (23)\\ \tilde{v}(0) &= \tilde{f}\end{aligned}$$

The tilde sign emphasizes that the boundary-points are not included in the calculation. Of course $g_l(t) = g_r(t) = 0$, for our model problem. The difference operators $\tilde{D}_1$ and $\tilde{D}_2$ are $(n-2)\times(n-2)$-matrices and do not have the same structure as D_1 and D_2. $\tilde{D}_1$ and $\tilde{D}_2$ are not SBP operators. In general it is not possible to obtain an energy estimate for (23), since the SBP property is lost.

3.3.3. *The Projection Method*

In this method developed by Olsson [11], the boundary condition is introduced via an orthogonal projection. When the energy method is applied, the projection operator P interacts with the difference operator to generate boundary terms that are completely analogous to those of the continuous problem. The method for our model problem looks like

$$\begin{gathered}v_t + aPH^{-1}Qv - \epsilon PH^{-1}(-S^TM + D)\,Sv = (I-P)\,\hat{g}_t(t)\\ v(0) = f\end{gathered} \qquad (24)$$

where $\hat{g}(t) = [g_l(t), 0,..., 0, g_r(t)]^T$. In [11] it is proved that the following holds for the projection P,

$$\begin{aligned} &P = I - H^{-1}L(L^TH^{-1}L)^{-1}\,L^T && P^2 = P \\ &v = Pv + (I-P)\,\hat{g}(t) \Leftrightarrow L^Tv = \hat{g}(t) && HP = P^TH \end{aligned} \tag{25}$$

In [11] it is also proved that $(I-P)(v(t) - \hat{g}(t)) = (I-P)(f - \hat{g}(0))$ holds, which means that the solution to (24) will satisfy the BC only if the boundary data and the initial data are consistent.

The energy method applied to (24) with $\hat{g}(t) = 0$ gives

$$\frac{d}{dt}\|v\|_H^2 = -av_N^2 - 2\epsilon((SPv)^T\,MSPv) + v_1^2\left(a + \frac{2\epsilon}{\beta}\right) \leqslant v_1^2\left(a + \frac{2\epsilon}{\beta}\right)$$

using the second, third and fourth property in (25). We have an energy estimate if condition (14) for well-posedness is fulfilled.

A drawback with this method is that it introduces a zero eigenvalue in the spectrum, see Sec. 3.4. This means that if the boundary data and initial data are inconsistent, the error will remain in the solution for all times. Since the largest real part of the continuous spectrum is negative, the projection method is not a strictly stable numerical approximation for the linear advection-diffusion equation.

3.3.4. The Modified Projection Method

The semi-discrete problem can be written as

$$v_t = \begin{cases} A_P v & \text{(projection)} \\ A_S v & \text{(SAT)} \\ A_I v & \text{(injection)} \end{cases}$$

The discrete spectrum are the eigenvalues of these matrices. In [5], a way to remove the zero in the spectrum for the projection method is presented. The idea is to implement the projection method as an injection method. The approximation can be written,

$$\begin{aligned} &v_t + aPD_1v = \epsilon PD_2v \\ &v_0 = \tilde{C}_l\tilde{v} + \alpha_l\,g_l(t), \qquad v_n = \tilde{C}_r\tilde{v} + \alpha_r\,g_r(t) \\ &v(0) = f \end{aligned} \tag{26}$$

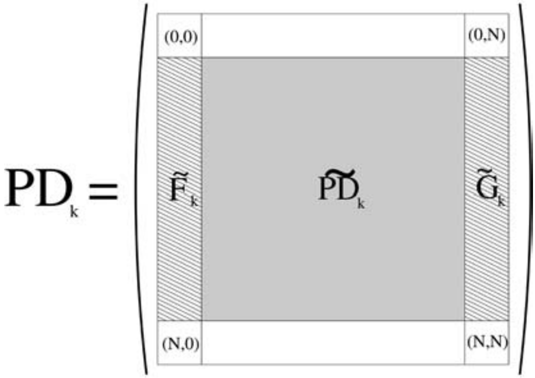


Fig. 5. Schematic picture of the modified operators for the modified projection method. D_k symbolizes the difference operator approx. $\partial^k/\partial x^k$.

This method is implemented as the injection method. The method can be written (see Fig. 5 for notations),

$$\begin{aligned}\tilde{v}_t &= (-a\tilde{P}D_1 + \epsilon\tilde{P}D_2 + (-a\tilde{F}_1 + \epsilon\tilde{F}_2)\,\tilde{C}_l + (-a\tilde{G}_1 + \epsilon\tilde{G}_2)\,\tilde{C}_r)\,\tilde{v}\\ &\quad + \alpha_l(-a\tilde{F}_1 + \epsilon\tilde{F}_2)\,g_l(t) + \alpha_r(-a\tilde{G}_1 + \epsilon\tilde{G}_2)\,g_r(t) \qquad (27)\\ \tilde{v}(0) &= \tilde{f}\end{aligned}$$

Presently it is not known how to prove stability for this method.

3.4. The Spectrum of the Semi-Discrete Problem

In this section, the spectrum of the discrete problem (discrete spectrum) for the SAT (S), the projection (P), the modified projection (B) and the injection (I) methods are considered. The semi-discrete equations are written as $v_t = A_{I,P,S,M}v$, where $A_{I,P,S,M}$ are the constant coefficient matrices of the semi-discrete problems, where

$$\begin{aligned}A_S &= -aD_1 + \epsilon D_2 + \frac{H^{-1}}{2\,h}\,(\tau_l\hat{e}_1 L_l + \tau_r\hat{e}_N L_r)\\ A_P &= -aPD_1 + \epsilon PD_2\\ A_I &= -a\tilde{D} + \epsilon\tilde{D}_2 - (a\tilde{A}_1 - \epsilon\tilde{A}_2)\,\tilde{C}_l - (a\tilde{B}_1 - \epsilon\tilde{B}_2)\,\tilde{C}_r\\ A_B &= -a\tilde{P}D_1 + \epsilon\tilde{P}D_2 + (a\tilde{F}_1 - \epsilon\tilde{F}_2)\,\tilde{C}_l - (a\tilde{G}_1 - \epsilon\tilde{G}_2)\,\tilde{C}_r\end{aligned}$$

Table II. $|E^h - E|$ and Conv. Rate, 2nd Order Case. $\beta = -\frac{2\epsilon}{a}$, $\epsilon = 0.1$, $a = 1$

N	(S)	q(S)	(P)	q(P)	(B)	q(B)	(I)	q(I)
50	5.92e−03	2.02	5.86e−03	2.02	5.86e−03	2.02	5.72e−03	2.06
100	1.44e−03	2.01	1.44e−03	2.00	1.44e−03	2.00	1.38e−03	2.02
200	3.57e−04	2.00	3.56e−04	2.00	3.56e−04	2.0	3.40e−04	2.01
400	8.88e−05	2.00	8.87e−05	2.00	8.87e−05	2.0	8.44e−05	2.00

The discrete spectrum are the eigenvalues of these matrices. The absolute error between the continuous eigenvalue E, with largest real part, and corresponding discrete eigenvalue E^h are presented in Tables II–IV, where N is the number of grid points. The convergence rate (order of accuracy), q, for the different methods is computed as,

$$q = \log_{10}\left(\frac{|E - E^{h1}|}{|E - E^{h2}|}\right) \Big/ \log_{10}\left(\frac{h_1}{h_2}\right) \tag{28}$$

E is the continuous eigenvalue and E^{h1} the corresponding discrete eigenvalue, obtained with mesh size $h1$.

Note that the projection method has a zero eigenvalue in the spectrum, that is not included in the calculation of the convergence rates, in Tables II–IV. The zero in the spectrum for the Projection method means that any error introduced into the computation will remain in the solution. This is also verified in the numerical calculations, see Fig. 6. The last row in Table IV shows strange results, probably due to inaccurate eigenvalues calculations.

3.5. Numerical Results

Consider problem (13) with initial data $u(x,0) = c1\,(2\beta - 2x + x^2)$, where $a = 1$, $\epsilon = 0.1$, $c1 = 1$. The initial function satisfies the boundary conditions. The constant $c1$ can be chosen arbitrarily, and β is determined

Table III. $|E^h - E|$ and Conv. Rate, 4th Order Case. $\beta = -\frac{2\epsilon}{a}$, $\epsilon = 0.1$, $a = 1$

N	(S)	q(S)	(P)	q(P)	(B)	q(B)	(I)	q(I)
50	3.99e−06	3.92	4.61e−06	4.14	4.61e−06	4.14	1.81e−05	4.17
100	2.44e−07	3.97	2.65e−07	4.06	2.65e−07	4.06	1.03e−06	4.08
200	1.51e−08	3.99	1.58e−08	4.04	1.57e−08	4.03	6.13e−08	4.04
400	9.43e−10	3.99	1.69e−09	3.65	9.78e−10	4.01	3.75e−09	4.02

Table IV. $|E^h - E|$ and Conv. Rate, 6th Order Case. $\beta = -\frac{2\epsilon}{a}$, $\epsilon = 0.1$, $a = 1$

N	(S)	q(S)	(P)	q(P)	(B)	q(B)	(I)	q(I)
50	2.03e−08	6.10	2.12e−08	6.18	2.12e−08	6.18	1.81e−07	6.17
100	2.93e−10	6.03	3.31e−10	6.15	2.98e−10	6.06	2.53e−09	6.07
200	5.56e−12	5.67	3.21e−11	3.54	5.79e−12	5.72	3.84e−11	6.01
400	1.24e−12	1.425	4.44e−10	−5.63	1.89e−12	3.30	2.73e−12	4.18

by (14). We want to examine how the solution behaves when the initial data and the boundary data do not match.

First we consider the case when the boundary and the initial data are inconsistent at the inflow boundary, by adding $\alpha e^{-\frac{x}{\beta}}$ to the initial data. Inconsistency at the outflow boundary is studied by adding $\alpha \cos(2\pi x)$ to the initial data. The constant α determines the magnitude of the inconsistency. The l_2 error as a function of time is presented in Fig. 6 for the 6th order case, with inconsistency at the inflow boundary. The result with inconsistency at the outflow boundary show similar results. The results for the 2nd and 4th order case yield almost identical results and hence are omitted. 50 grid points are used and the solutions are advanced to time $t = 10$, using the standard fourth-order Runge–Kutta method. The error is defined as the difference between the disturbed and the undisturbed solution. All four methods are presented in the same plot. The l_2 error for the

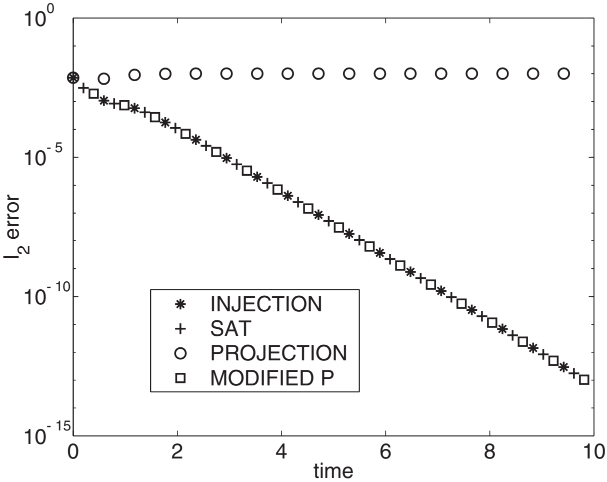


Fig. 6. l_2 error as a function of time for 6th order case. $\beta = -\epsilon/2a$, $\epsilon = 0.1$, $a = 1$, $\alpha = 0.01$. Inconsistency at the right boundary.

projection method stays constant, while the other methods show decreasing l_2 errors.

4. HYPERBOLIC SYSTEMS

4.1. The Continuous Problem

Consider the Hyperbolic system

$$\begin{aligned} & u_t + u_x = 0 && 0 \leqslant x \leqslant 1, \quad t \geqslant 0 \\ & v_t - v_x = 0 && 0 \leqslant x \leqslant 1, \quad t \geqslant 0 \\ & u(0,t) = v(0,t), && v(1,t) = u(1,t) \\ & u(x,0) = g(x), && v(x,0) = h(x) \end{aligned} \tag{29}$$

This give $\frac{d}{dt}(\|u\|^2 + \|v\|^2) = 0$, i.e., the energy is constant. Using the same technique as in Sec. 3.2 it can be shown that the continuous spectrum consist of $s = 2n\pi i$, $n = 0 \pm 1 \pm 2 \cdots$.

4.2. The Semi-Discrete Problem

To solve the problem (29) numerically we discretize as in Sec. 3.3 and denote by $w^T = [u_1, \ldots, u_N, v_1, \ldots, v_N] = [w_1, \ldots, w_{2N}]$ the solution vector. The semi-discrete problem now looks like

$$\begin{aligned} w_t + \bar{D}w &= 0 \\ \bar{L}^T w &= 0 \\ w(0) &= f \end{aligned} \tag{30}$$

where

$$\bar{L}^T = \begin{bmatrix} l_{1,1} & 0 & \cdots & 0, l_{1,N+1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & l_{2,N}, 0 & 0 & \cdots & l_{2,2N} \end{bmatrix} \qquad \begin{aligned} l_{1,1} &= 1, & l_{1,N+1} &= -1 \\ l_{2,N} &= -1, & l_{2,2N} &= 1 \end{aligned}$$

The matrix $D_1 = H^{-1}Q$ is the difference operator used in Sec. 3.3, i.e., the matrix $\bar{D} = \bar{H}^{-1}\bar{Q}$ is a $(2N) \times (2N)$-matrix, where $\bar{H} = \operatorname{diag}(H, H)$ and $\bar{Q} = \operatorname{diag}(Q, -Q)$.

4.2.1. The Projection Method

The projection method was introduced in Sec. 3.3.3. The projection operator for this system is $\bar{P} = I - \bar{H}^{-1}\bar{L}(\bar{L}^T\bar{H}^{-1}\bar{L})^{-1}\,\bar{L}^T$. The method for our model problem looks like

$$\begin{aligned} w_t + \bar{P}\bar{D}w &= 0 \\ w(0) &= f \end{aligned} \tag{31}$$

The energy method applied to (31) gives

$$\frac{d}{dt}\|w\|_H^2 = -2((\bar{P}w)^T\,\bar{Q}w) = -2((\bar{P}w)^T\,\bar{Q}\bar{P}w) = 0 \tag{32}$$

using the second, third and fourth property in (25) and the fact that $\bar{P}w = [w_1, \ldots, w_N, w_1, \ldots, w_N]$. This estimate is exactly the same estimate as in the continuous problem. For this particular system (29), $s = 0$ is a part of the continuous spectrum. Hence the zero eigenvalue that is always introduced by the Projection method does not destroy the growth rate here.

4.2.2. The SAT Method

For the SAT method we have

$$\begin{aligned} w_t + \bar{D}w &= -\bar{H}^{-1}\hat{e}\hat{\tau}\bar{L}^T w \\ w(0) &= f \end{aligned} \tag{33}$$

where

$$\hat{e}^T = \begin{bmatrix} e_{1,1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & e_{2,2N} \end{bmatrix}, \qquad \hat{\tau} = \begin{bmatrix} \tau_l & 0 \\ 0 & \tau_r \end{bmatrix}, \qquad e_{1,1} = 1,\ e_{2,2N} = 1$$

The energy method on (33) gives

$$\frac{1}{2}\frac{d}{dt}\|w\|_H^2 = \begin{bmatrix} w_1 \\ w_{N+1} \end{bmatrix}^T \begin{bmatrix} \frac{1-2\tau_l}{2} & \frac{\tau_l}{2} \\ \frac{\tau_l}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_{N+1} \end{bmatrix} + \begin{bmatrix} w_N \\ w_{2N} \end{bmatrix}^T \begin{bmatrix} -\frac{1}{2} & \frac{\tau_r}{2} \\ \frac{\tau_r}{2} & -\frac{2\tau_r - 1}{2} \end{bmatrix} \begin{bmatrix} w_N \\ w_{2N} \end{bmatrix}$$

An energy estimate requires that $\tau_l = \tau_r = 1$. This give $\frac{d}{dt}\|w\|_H^2 = -\frac{1}{2}(w_1 - w_{N+1})^2 - \frac{1}{2}(w_N - w_{2N})^2$. Hence the boundary implementation introduce a small damping in the system.

As for the semi-discrete approximation of the scalar problem, described in Sec. 3.3, we fail to get an energy estimate for the injection method and the modified projection method since the SBP-property is lost.

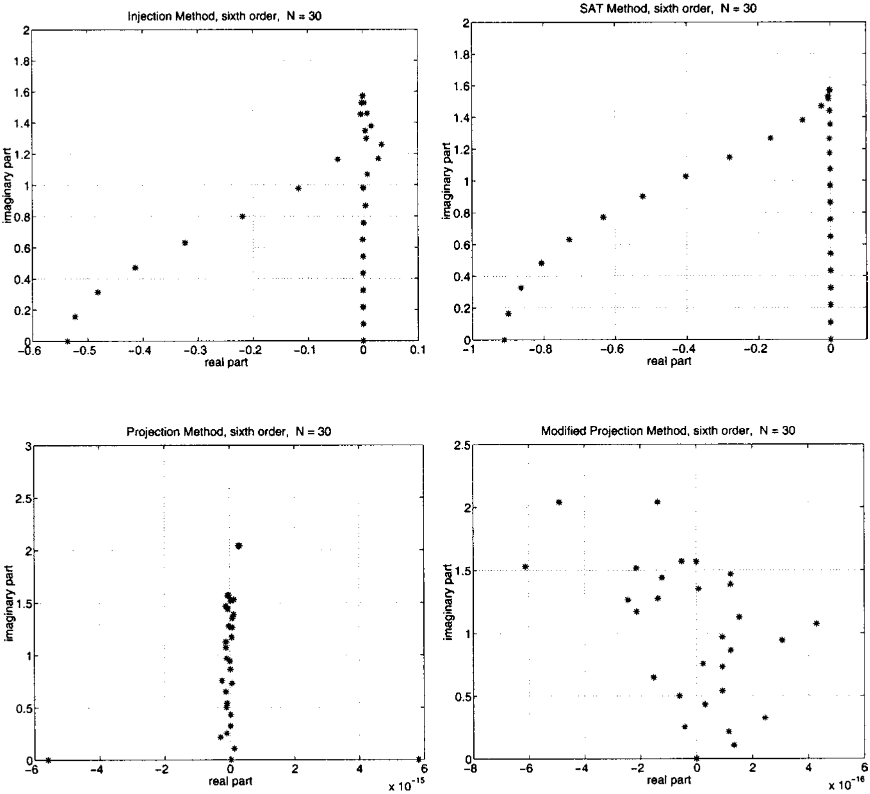


Fig. 7. Discrete spectrum, sixth order case, $N = 30$.

4.3. The Spectrum of the Semi-Discrete Problem

In this section the discrete spectrum for the SAT, the projection, the modified projection and the injection methods are considered, see Fig. 7. The continuous spectrum consist of $s = 2n\pi i$, $n = 0 \pm 1 \pm 2 \cdots$. The largest real part of the discrete spectrum are presented in Table V.

Table V. The Largest Real Part of the Discrete Spectrum for the 2nd, 4th, 6th Order Case. Here Using 100 Grid Points

accuracy	SAT	Projection	Injection	Modified P
2nd	−6.883e−15	1.819e−07	−2.487e−14	6.939e−18
4th	1.207e−15	5.119e−13	9.856e−02	4.192e−14
6th	−2.665e−15	6.388e−14	1.209e+00	2.331e−14

Notice that the injection method have positive eigenvalues for the fourth and sixth order case, see Fig. 7 and Table V.

4.4. Numerical Results

Consider (29) with initial data $u(x, 0) = \sin 2\pi x$, $v(x, 0) = -\sin 2\pi x$, $0 \leqslant x \leqslant 1$. The exact solution is $u(x, t) = \sin 2\pi(x-t)$, $v(x, t) = -\sin 2\pi(x+t)$, $0 \leqslant x \leqslant 1$, $t \geqslant 0$.

The upper limit for the CFL number $\frac{k}{h}$ is chosen as $\frac{R}{\|hA\|_2}$, where A is the fixed constant coefficient matrix of the semi-discrete problem, h is the space step, k is the time step and R the stability radius of the Runge–Kutta method. For the standard fourth-order Runge–Kutta method $R = 2.2$.

To further investigate, numerically, if the the methods are strictly stable we compute the l_2-error for long time integrations, see Fig. 8.

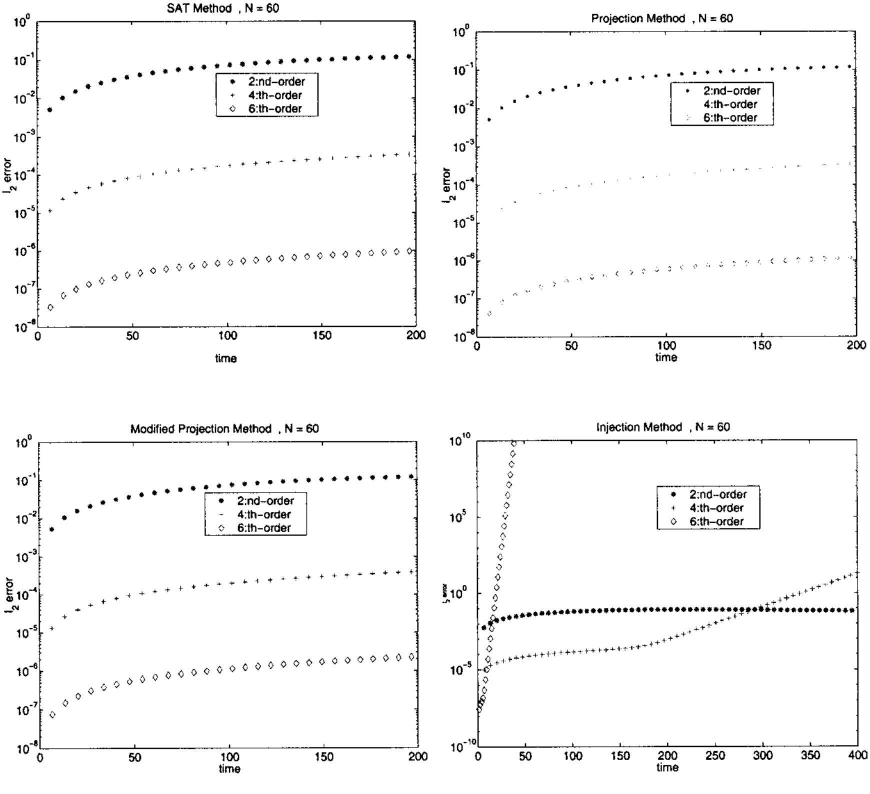


Fig. 8. l_2-error as a function of t, for 2nd, 4th, 6th order case, $N = 60$.

Obviously the SAT method, the projection method and the modified projection method have bounded error growth in time. The l_2-error for the injection method grows rapidly for the fourth and sixth order case, due to the eigenvalues with positive real part, see Table V and Fig. 8.

5. CONCLUSIONS

We have analyzed four methods of imposing physical boundary conditions for the linear advection-diffusion equation and a linear hyperbolic system. The methods are: the injection method, the SAT method, the projection method and the modified projection method.

A semi-discrete approximation that leads to an energy estimate is guaranteed to have a bounded solution. However, the energy estimate alone does not guarantee strict stability. Strict stability requires that the semi-discrete problem has the same asymptotic time-growth as the continuous problem. The asymptotic time-growth is determined by the utmost right part of the spectrum. However, a strictly stable approximation do allows for an initial growth of the solution-energy, before the asymptotic decay sets in. Hence, guaranteeing non-growing energy and getting correct asymptotic time-growth, require strict stability and an energy estimate.

In general, the injection method and the modified projection method destroy the SBP property and make it difficult to prove stability using the energy method. The injection method is in fact unstable for the fourth and sixth order accurate approximation of the hyperbolic system, due to positive eigenvalues in the discrete spectrum. However, the numerical calculations indicate that the modified projection method result in a strictly stable approximation.

For the projection method we get an energy estimate but the discrete spectrum have a zero eigenvalue. Hence, for problems where the largest real part of the continuous spectrum is negative, like the linear advection-diffusion equation, the projection method will not result in a strictly stable approximation. The SAT method leads to an energy estimate and the discrete spectrum converge to the continuous spectrum. The numerical calculations also indicate that the SAT-method result in a strictly stable approximation.

REFERENCES

1. Birkhoff, G., and Rota, G.-C. (1978). *Ordinary differential equations*, Wiley, New York.
2. Carpenter, M. H., Gottlieb, D., and Abarbanel, S. (1994). The stability of numerical boundary treatments for compact high-order finite-difference schemes. *J. Comput. Phys.* 108(2).

3. Carpenter, M. H., Nordström, J., and Gottlieb, D. (1999). A stable and conservative interface treatment of arbitrary spatial accuracy. *J. Comput. Phys.* 148.
4. Erlebacher, G., Hussaini, M. Y., and Shu, C.-W. (1997). Interaction of a shock with a longitudinal vortex. *J. Fluid. Mech.* **337**, 129–153,
5. Gustafsson, B. (1998). On the implementation of boundary conditions for the methods of lines. *BIT* 38(2).
6. Gustafsson, B., Kreiss, H.-O., and Oliger, J. (1995). *Time dependent problems and difference methods*, Wiley, New York.
7. Kreiss, H.-O., and Lorenz, J. (1989). *Initial Boundary Value Problems and the Navier–Stokes Equations*, Academic Press, New York.
8. Kreiss, H.-O., and Oliger, J. (1972). Comparison of accurate methods for the integration of hyperbolic equations. *Tellus XXIV* 3.
9. Mattsson, K. (2000). Imposing boundary conditions with the injection, the projection and the simultaneous approximation term methods, Technical Report 2000-016, Department of Information Technology, Uppsala University, Uppsala, Sweden.
10. Nordström, J. (1989). The influence of open boundary conditions on the convergence to steady state for the Navier–Stokes equations. *J. Comput. Phys.* 85.
11. Olsson, P. (1995). Summation by parts, projections, and stability I. *Math. Comp.* **64**, 1035.
12. Olsson, P. (1995). Summation by parts, projections, and stability II. *Math. Comp.* **64**, 1473.
13. Petrini, E., Efraimsson, G., and Nordström, J. (1998). A numerical study of the introduction and propagation of a 2-d vortex, Technical Report FFA TN 1998-66, The Aeronautical Research Institute of Sweden, Bromma, Sweden.
14. Strand, B. (1996). *High-Order Difference Approximations for Hyperbolic Initial Boundary Value Problems*, Ph.D. thesis, Uppsala University, Department of Scientific Computing, Uppsala University, Uppsala, Sweden.

Printed in Belgium