NASA-CR-194293

RIACS

Research Institute for Advanced Computer Science
NASA Ames Research Center

# On the Superconvergence of Galerkin Methods for Hyperbolic IBVP

David Gottlieb
Bertil Gustafsson
Pelle Olsson
Bo Strand

RIACS Technical Report 93.07 August 1993

# On the Superconvergence of Galerkin Methods for Hyperbolic IBVP

David Gottlieb[1]
Bertil Gustafsson[2],[3]
Pelle Olsson[3]
Bo Strand[2]

The Research Institute of Advanced Computer Science is operated by Universities Space Research Association, The American City Building, Suite 212, Columbia, MD 21044, (410) 730-2656

Work reported herein was sponsored by NASA under contract NAS 2-13721 between NASA and the Universities Space Research Association (USRA), by AFOSR grant 93-1-0090, by DARPA grant N00014-91-J-4016, by NSF grant DMS-92-11820, and by NUTEK grant 89-02539P.

[1] Division of Applied Mathematics, Brown University

[2] Department of Scientific Computing, Uppsala University, Uppsala, Sweden

[3] Research Institute for Advanced Computer Science, NASA Ames Research Center

## Abstract

Finite element Galerkin methods for periodic first order hyperbolic equations exhibit superconvergence on uniform grids at the nodes, i. e., there is an error estimate $\mathcal{O}(h^{2r})$ instead of the expected approximation order $\mathcal{O}(h^r)$. In this paper it will be shown that no matter how the approximating subspace $S^h$ is chosen, the superconvergence property is lost if there are characteristics leaving the domain. We shall also discuss the implications of this result when constructing compact implicit difference schemes.

# 1 Introduction

Error estimates for finite element approximations based on Galerkin methods for

$$\frac{\partial u}{\partial t} = Pu$$

are determined by the approximation property of the finite element space. If the approximation order is $r$, the convergence rate is $\mathcal{O}(h^r)$. However, for problems with periodic solutions and uniform grids there is superconvergence at the *grid points.* Thomée and Wendroff [10] showed that if the differential operator $P$ is of order $m$, then the error at the grid points is $\mathcal{O}(h^\nu)$, $\nu = 2r - m$ for $m$ even $\nu = 2r - m + 1$ for $m$ odd. The question is whether this superconvergence property remains in the presence of boundary conditions for finite computational domains. In this paper it will be shown that the superconvergence property is lost, no matter how the approximating subspace $S^h$ is chosen, if there are characteristics leaving the domain. We will illustrate the loss of superconvergence by considering the simple model problem

$$\begin{aligned} u_t &= u_x \,, \qquad 0 \le x < \infty \,, 0 \le t \,, \\ u(x,0) &= f(x) \end{aligned} \tag{1}$$

and the case $r = 2$ (for simplicity we introduced only one boundary by considering the interval $[0,\infty)$). The scalar product and the corresponding norm are defined by

$$(u,v) = \int_0^\infty u(x)v(x)dx \;, \quad ||u||^2 = (u,u) \,, \tag{2}$$

It is assumed that $||f|| < \infty$. We shall henceforth only consider real functions.

The problem (1) has a unique solution (without specifying any boundary condition at $x = 0$) satisfying an energy estimate

$$\frac{d}{dt}||u(\cdot,t)||^2 = -u(0,t)^2 \,, \tag{3}$$

or, equivalently,

$$||u(\cdot,t)||^2 + \int_0^t u(0,\tau)^2 d\tau = ||f(\cdot)||^2. \tag{4}$$

Let $S$ be the Sobolev space of functions $u(x)$ with $||du/dx|| < \infty$. The weak formulation of (1) is

*Find $u(x,t) \in S$ such that for every fixed $t$*

$$\begin{aligned} (u_t, v) &= (u_x, v) \,, \quad v \in S \,, \\ u(x,0) &= f(x) \,. \end{aligned} \tag{5}$$

A grid is defined by $x_j = jh$, $j = 0, 1, \ldots$. Let $S_h$ be the space of piecewise polynomials of degree $r - 1$. The Galerkin method can be formulated as follows

*Find $u^h(x,t) \in S_h$ such that for every fixed $t$*

$$\begin{aligned} (u_t^h, v^h) &= (u_x^h, v^h)\,, \quad v^h \in S_h\,, \\ u^h(x,0) &= f^h(x)\,. \end{aligned} \tag{6}$$

An attractive feature of the Galerkin method is that stability follows directly from the well-posedness of the continuous problem. In fact, substituting $v^h = u^h$ in (6) yields

$$\frac{d}{dt}||u^h||^2 = 2(u_t^h, u^h) = 2(u_x^h, u^h) = -u^h(0,t)^2. \tag{7}$$

Consider the approximating space $S_h$ of piecewise linear functions. The superconvergence result states that the error is $\mathcal{O}(h^4)$ rather than $\mathcal{O}(h^2)$. This is easily seen if we consider the hat functions $\varphi_i(x)$ as a basis of $S_h$, where

$$\varphi_i(x) = \begin{cases} \frac{1}{h}(x - x_{i-1})\,, & x_{i-1} \le x \le x_i\,, \\ -\frac{1}{h}(x - x_{i+1})\,, & x_i \le x \le x_{i+1}\,, \quad i \ge 1 \\ 0\,, & \text{otherwise,} \end{cases} \tag{8}$$

$$\varphi_0(x) = \begin{cases} -\frac{1}{h}(x - x_1)\,, & 0 \le x \le x_1\,, \\ 0\,, & \text{otherwise.} \end{cases}$$

Since $\varphi_i(x_j) = \delta_{ij}$ and

$$u^h(x,t) = \sum_{j=0}^{\infty} u_j(t)\varphi_j(x)\,, \tag{9}$$

the Galerkin method (6) can be viewed as a difference method for $u_j$:

$$\frac{1}{6}\frac{d}{dt}(u_{j+1} + 4u_j + u_{j-1}) = \frac{u_{j+1} - u_{j-1}}{2h}\,, \quad j = 1, 2, \ldots \tag{10}$$

$$\frac{1}{3}\frac{d}{dt}(2u_0 + u_1) = \frac{u_1 - u_0}{h}\,. \tag{11}$$

The superconvergence result holds for problems with periodic solutions, where (11) is replaced by the periodicity condition $u_j = u_{j+N}$. In this case the Galerkin method (10) is the well known fourth order compact difference approximation (proposed by Kreiss, see [9]). However, equation (11), applied in the non-periodic case, is only first order accurate, and thus, as will be shown in the next section, the global error estimate is second order only.

In section 3 we shall prove that there is no way to modify the approximating space near the boundary such that the fourth order accuracy is retained. In fact, the boundary condition (11) is optimal in the sense that any choice of boundary functions leads to only second order accuracy at most.

The equations (10),(11) can be viewed as a difference approximation to (1). Indeed, this interpretation was used by Thomée and Wendroff when deriving the superconvergence result. Such methods are often called implicit compact difference approximations or Padé approximations. Compared to standard explicit approximations of the same order, they have a considerably smaller error coefficient, making them competitive with explicit schemes, despite the fact that they are implicit.

By leaving the Galerkin formulation, there is greater flexibility when modifying the approximation near the boundary. Carpenter et. al. [1] constructed a stable and third order accurate boundary modification for (11) that results in fourth order global accuracy. However, the condition (7) is no longer fulfilled, and it is shown that there is a growth in time (independent of $h$ since the scheme is stable).

The semi-discrete compact scheme can be written in the form

$$P\frac{d\underline{u}}{dt} = \frac{1}{h}Q\underline{u} \tag{12}$$

where $\underline{u}$ is the vector of unknowns $u_j$, $P$ and $Q$ are band matrices, $P$ symmetric positive definite (SPD). Asymptotic stability (also called time stability or strict stability) in the norm $\langle \underline{u}, P\underline{u}\rangle_h$ follows from

$$\langle \underline{u}, Q\underline{u}\rangle_h = -\frac{1}{2}u_0^2 .$$

where the discrete scalar product $\langle \cdot,\cdot\rangle_h$ and the corresponding norm $||\cdot||_h$ are defined by

$$\langle \underline{u}, \underline{v}\rangle_h = \sum_{j=0}^{\infty} u_j v_j h, \quad ||\underline{u}||_h^2 = \langle \underline{u}, \underline{u}\rangle_h . \tag{13}$$

This is the same condition as (7). It is convenient also for difference methods when applied to systems, since it allows for stable implementations of physical boundary conditions in a direct way, see for example [2] and [8].

It follows directly from the negative result for Galerkin methods that there is no way of modifying $P, Q$ near the boundary with our conditions above satisfied, as long as $P$ has to be SPD. The question then arises whether there is another norm $\langle \underline{u}, \hat{P}\underline{u}\rangle_h$ such that our conditions can be fulfilled with new $P$ and $Q$ in (12). We shall prove in section 4 that this is impossible unless $\hat{P}$ is allowed to be different from $P$ at interior points.

Recently, Carpenter et. al. [2] have shown that the conditions can be fulfilled if $\hat{P}$ is modified also at inner points. In the case of the fourth order scheme with tridiagonal $P$ at inner points, the new matrix norm is pentadiagonal at inner points. For explicit high order approximations it is possible to construct simpler norms and still obtain asymptotic time stability, see the work by Kreiss and Scherer [4, 5] and by Olsson [6, 7, 8]. Wahlbin [11] has discussed a different concept of superconvergence for elliptic problems, where the error in the derivative is of the same order as the function itself. It should be emphasized that the concept of superconvergence that we are discussing refers to the error estimates of the function itself.

## 2 Necessary Conditions for the Accuracy Near the Boundary

In [3] it is proved that for certain classes of difference schemes it is possible to lower the accuracy one order at the boundary without loosing the convergence rate defined by the interior scheme. We shall prove in this section that the accuracy near the boundary cannot be decreased any further, if we want to retain the interior convergence rate. Consider a semi-discrete approximation of any linear first order system

$$P_j \frac{du_j}{dt} = \frac{1}{h} Q_j u_j\,, \quad j = 0, 1, \dots . \tag{14}$$

where the difference operators $P_j$, $Q_j$ satisfy $P_j = \overline{P}$ and $Q_j = \overline{Q}$ for $j \geq s$. The error $\epsilon_j(t) = u_j(t) - u(x_j, t)$ then obeys

$$P_j \frac{d\epsilon_j}{dt} = \frac{1}{h} Q_j \epsilon_j + g_j\,, \quad j = 0, 1, \dots . \tag{15}$$

where the truncation error $g_j(t)$ is a smooth function of $t$ that satisfies

$$|g_j(t)| = \begin{cases} \mathcal{O}(h^q)\,, & j = 0, 1, \dots, s-1 \\ \mathcal{O}(h^r)\,, & j \geq s \end{cases}$$

It is assumed that $q < r$. Suppose that the initial data are exact, i. e., $\epsilon_j(0) = 0$, $j \geq 0$. Define $\underline{u}^T = (u_0 \dots )$, $\underline{v}^T = (u(0,t) \dots )$, $\underline{g}^T = (g_0 \dots )$, and $\underline{\epsilon} = \underline{u} - \underline{v}$. Here $\underline{v}$ denotes the vector whose elements consist of the analytic solution at the grid points $x_j$. The error equation (15) can then be expressed as

$$P \frac{d\underline{\epsilon}}{dt} = \frac{1}{h} Q \underline{\epsilon} + \underline{g}(t)\,. \tag{16}$$

where

$$\underline{g}(t) = P \frac{d\underline{v}}{dt} - \frac{1}{h} Q \underline{v}\,. \tag{17}$$

**Lemma 2.1** *Suppose that $|g_j(t)| \geq K_0 h^q$ for some $j$, $0 \leq j \leq s-1$ and $0 < T_0 \leq t \leq T_1$, where $K_0$ is independent of $h, t$. Then the solution $\underline{\epsilon}(t)$ of equation (16) satisfies*

$$\max_{0 \leq t \leq T_1} ||\underline{\epsilon}(t)||_h \geq K_1 h^{q+3/2} \tag{18}$$

*for $h \leq 1$, where $||\cdot||_h$ is defined by eq. (13).*

**Proof**
Integrating equation (16) with respect to $t$ gives (recall that $\underline{\epsilon}(0) = 0$)

$$P\underline{\epsilon}(t) = \frac{1}{h} Q \int_0^t \underline{\epsilon}(\tau) d\tau + \underline{G}(t) , \quad 0 \leq t \leq T_1 ,$$

where

$$\underline{G}(t) = \int_0^t \underline{g}(\tau) d\tau .$$

Hence

$$||\underline{G}(t)||_h \leq \left( ||P||_h + \frac{T_1}{h} ||Q||_h \right) \max_{0 \leq t \leq T_1} ||\underline{\epsilon}(t)||_h ,$$

where we have used

$$|| \int_0^t \underline{\epsilon}(\tau) d\tau ||_h \leq \int_0^t ||\underline{\epsilon}(\tau)||_h d\tau$$

which follows from Jensen's inequality and the homogeneity of $|| \cdot ||_h$. For $h \leq 1$ we thus obtain

$$\max_{0 \leq t \leq T_1} ||\underline{\epsilon}(t)||_h \geq \frac{h}{||P||_h + T_1 ||Q||_h} ||G(t)||_h , \quad 0 \leq t \leq T_1 .$$

In particular, the inequality is true for $t = T_1$. Thus

$$\max_{0 \leq t \leq T_1} ||\underline{\epsilon}(t)||_h \geq \frac{h}{||P||_h + T_1 ||Q||_h} | \int_0^{T_1} g_j(\tau) d\tau | h^{1/2}$$

For convenience it will be assumed that the smooth component $g_j$ be positive for $T_0 \leq t \leq T_1$. Furthermore, $\epsilon_j(0) = 0$ implies that $g_j(0) = 0$. Thus, we can assume without restriction that $g_j(t) \geq 0$ for $0 \leq t \leq T_1$. Consequently,

$$| \int_0^{T_1} g_j(\tau) d\tau | = \int_0^{T_1} g_j(\tau) d\tau \geq T_1 K_0 h^q$$

Thus,

$$\max_{0 \leq t \leq T_1} ||\underline{\epsilon}(t)||_h \geq \frac{K_0 T_1}{||P||_h + T_1 ||Q||_h} h^{q+3/2}$$

The lemma is proved. □

**Remark:** The lemma allows for the possibility of decreasing the accuracy near the boundary by order 1.5. In reality, however, the order of accuracy must be an integer. Thus, the accuracy can be decreased one order near the boundary but not more. Furthermore, the extra factor of $h^{1/2}$ is an effect of the $L_2$-norm, and would not be present in the maximum norm.

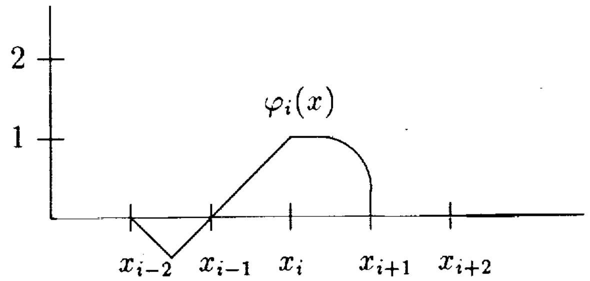
# 3 Modified Elements Near the Boundary

We shall prove that there is no modification of the piecewise linear space $S_h$ near the boundary such that global fourth order accuracy is retained.

**Theorem 3.1** *There is no local modification near the boundary such that the superconvergence property is retained with the Galerkin method. Indeed, with piecewise linear elements at inner points, the maximum convergence rate is $\mathcal{O}(h^2)$.*

**Proof**

We first make the assumption that the basis functions of $S_h$ satisfy $\varphi_i(x_j) = \delta_{ij}$.



Furthermore, in the interior it will always be assumed that the $\varphi_i$'s are given by (8). Let $\underline{u}$ be the vector with components $u_j$. The Galerkin method (6) can then be expressed as

$$P\frac{d\underline{u}}{dt} = \frac{1}{h}Q\underline{u}\,. \tag{19}$$

Here $P$ and $Q$ are well defined at inner points by (10), and we partition them as

$$\begin{gathered} P = \begin{pmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{pmatrix}, \qquad P_{12} = \frac{1}{6}\begin{pmatrix} 0 & \dots & & \\ \vdots & & & \\ 0 & & & \\ 1 & 0 & \dots & \end{pmatrix}, \\ P_{22} = \frac{1}{6}\begin{pmatrix} 4 & 1 & & \\ 1 & 4 & 1 & \\ & 1 & 1 & \ddots \\ & & \ddots & \ddots \end{pmatrix}, \end{gathered} \tag{20}$$

$$\begin{gathered} Q = \begin{pmatrix} Q_{11} & Q_{12} \\ -Q_{12}^T & Q_{22} \end{pmatrix}, \qquad Q_{12} = \frac{1}{2}\begin{pmatrix} 0 & \dots & & \\ \vdots & & & \\ 0 & & & \\ 1 & 0 & \dots & \end{pmatrix}, \\ Q_{22} = \frac{1}{2}\begin{pmatrix} 0 & 1 & & \\ -1 & 0 & 1 & \\ & -1 & 0 & \ddots \\ & & \ddots & \ddots \end{pmatrix}. \end{gathered} \tag{21}$$

The submatrices $P_{11}$, $Q_{11}$ are $(s \times s)$-matrices resulting from the modified basis functions near the boundary. The elements of $P$ and $Q$ have been normalized to be of order one,

i. e.,

$$\begin{aligned} p_{ij} &= (\varphi_i, \varphi_j)/h\,, \\ q_{ij} &= (\varphi_i, d\varphi_j/dx)\,. \end{aligned}$$

Whatever space $S_h$ we are using the relation

$$\frac{d}{dt}||u^h(\cdot,t)||^2 = -u^h(0,t)^2 \tag{22}$$

holds. Introducing $u^h = \sum u_j\varphi_j$ we get

$$\frac{d}{dt}\langle \underline{u}, P\underline{u}\rangle_h = -u_0^2\,, \tag{23}$$

where the discrete scalar product $\langle\cdot,\cdot\rangle_h$ is defined by eq. (13). Combining eqs. (19) and (23) yields

$$\langle \underline{u}, \frac{1}{h}(Q^T + Q)\underline{u}\rangle_h = -u_0^2\,, \tag{24}$$

that is, $Q$ is almost antisymmetric. Summing up, the assumption that $\varphi_i(x_j) = \delta_{ij}$ implies that the Galerkin method can be written as a difference method (19) where $P$ is SPD and where $Q$ is almost antisymmetric. Consequently,

$$\begin{aligned} &p_{ij} = p_{ji}\,, && i,\, j \ge 0\,, \\ &q_{ij} = -q_{ji}\,, && i + j > 0\,, \\ &q_{00} = -1/2\,. && \end{aligned} \tag{25}$$

It is enough to consider polynomials when investigating the accuracy, and we choose $f(x) = (x-(s-1)h/2)^r$, $r = 0, 1, \ldots$. Without restriction we can assume $h = 1$. At inner points $x_i$, $i \ge s$, the elements $p_{ij}$, $q_{ij}$ are known. For $0 \le i \le s-2$ the rows in $P$ have the non-zero elements $p_{i0}, \ldots, p_{i,s-1}$ and correspondingly for $Q$. In row $s-1$ there is an extra element $p_{s-1,s} = 1/6$ and $q_{s-1,s} = 1/2$, respectively. Thus, the accuracy check has the form (where we let $n = s-1$ for simplicity)

$$\begin{aligned} &r\sum_{j=0}^{n} p_{ij}\left(j - \frac{n}{2}\right)^{r-1} + \frac{1}{6}\delta_{in} r\left(\frac{n}{2}+1\right)^{r-1} \\ &= \sum_{j=0}^{n} q_{ij}\left(j - \frac{n}{2}\right)^{r} + \frac{1}{2}\delta_{in}\left(\frac{n}{2}+1\right)^{r}\,, \qquad i = 0, \ldots, n;\; r = 0, 1, \ldots \end{aligned} \tag{26}$$

For $r = 0$ we have by taking (25) into account

$$\sum_{j=0}^{n} q_{ij} = -\sum_{j=0}^{n} q_{ji} - \delta_{i0} = -\frac{1}{2}\delta_{in}\,, \quad i = 0, \ldots, n\,, \tag{27}$$

and for $r = 1$

$$\sum_{j=0}^{n} p_{ji} = \sum_{j=0}^{n} p_{ij} = \sum_{j=0}^{n} q_{ij}\left(j - \frac{n}{2}\right) - \frac{1}{6}\delta_{in} + \frac{1}{2}\left(\frac{n}{2}+1\right)\delta_{in}\,, \quad i = 0, \ldots, n. \tag{28}$$

Summing (26) over $i$ gives

$$r\sum_{j=0}^{n}\left(j-\frac{n}{2}\right)^{r-1}\sum_{i=0}^{n}p_{ij}+\frac{r}{6}\left(\frac{n}{2}+1\right)^{r-1}=\sum_{j=0}^{n}\left(j-\frac{n}{2}\right)^{r}\sum_{i=0}^{n}q_{ij}+\frac{1}{2}\left(\frac{n}{2}+1\right)^{r},$$

and by (27), (28)

$$\begin{aligned}&r\sum_{j=0}^{n}\left(j-\frac{n}{2}\right)^{r-1}\left[\sum_{i=0}^{n}q_{ji}\left(i-\frac{n}{2}\right)-\frac{1}{6}\delta_{jn}+\frac{1}{2}\left(\frac{n}{2}+1\right)\delta_{jn}\right]\\&+\frac{r}{6}\left(\frac{n}{2}+1\right)^{r-1}=\sum_{j=0}^{n}\left(j-\frac{n}{2}\right)^{r}\left(-\delta_{j0}+\frac{1}{2}\delta_{jn}\right)+\frac{1}{2}\left(\frac{n}{2}+1\right)^{r}.\end{aligned}$$

For $r=2$ we have

$$\begin{aligned}&2\sum_{j=0}^{n}\sum_{i=0}^{n}q_{ji}\left(ji-j\frac{n}{2}-i\frac{n}{2}+\left(\frac{n}{2}\right)^{2}\right)-\frac{n}{6}+\frac{n}{2}\left(\frac{n}{2}+1\right)+\frac{1}{3}\left(\frac{n}{2}+1\right)\\&=-\left(\frac{n}{2}\right)^{2}+\frac{1}{2}\left(\frac{n}{2}\right)^{2}+\frac{1}{2}\left(\frac{n}{2}+1\right)^{2}.\end{aligned}\tag{29}$$

By (25)

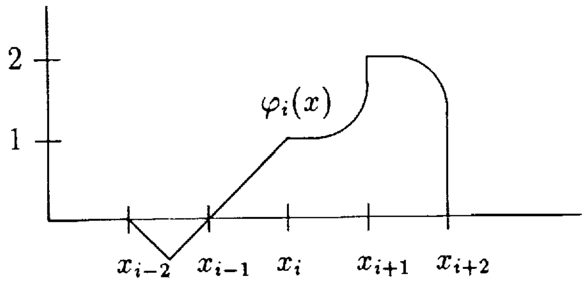
$$\begin{aligned}\sum_{j=0}^{n}\sum_{i=0}^{n}q_{ij}ji&=0,\\\sum_{j=0}^{n}\sum_{i=0}^{n}q_{ji}(j+i)&=0,\\\sum_{j=0}^{n}\sum_{i=0}^{n}q_{ji}&=-\frac{1}{2}.\end{aligned}$$

Thus, with $m=n/2$, we get from (29)

$$-m^2-\frac{m}{3}+m^2+m+\frac{m}{3}+\frac{1}{3}=-m^2+\frac{1}{2}m^2+\frac{1}{2}m^2+m+\frac{1}{2},$$

which has no solution. Accordingly, the accuracy near the boundary can be at most first order, and the theorem follows by lemma (2.1).

Thus far the theorem has been proved for basis functions satisfying $\varphi_i(x_j)=\delta_{ij}$. This assumption will now be removed.



We write a function $u^h \in S_h$ as

$$u^h(x,t) = \sum_{j=0}^{\infty} \alpha_j(t)\varphi_j(x)\,, \tag{30}$$

and in general $u^h(x_j,t) \neq \alpha_j(t)$ near the boundary. If we let $\underline{\alpha}$ be the vector with components $\alpha_j$ the Galerkin method gives us

$$\tilde{P}\frac{d\underline{\alpha}}{dt} = \frac{1}{h}\tilde{Q}\underline{\alpha}\,, \tag{31}$$

which formally is of the same type as eq. (19). Again, $\tilde{P}$ is SPD with the same structure as $P$ in (20). Combining eqs. (22), (30) and (31) gives

$$\langle \underline{\alpha}, \frac{1}{h}(\tilde{Q}^T + \tilde{Q})\underline{\alpha}\rangle_h = -u^h(0,t)^2 = -u_0^2\,. \tag{32}$$

Thus, $\tilde{Q}$ is *not* almost antisymmetric in the sense of eq. (24). Therefore, the first part of the proof cannot be directly applied to (31). Since the $\varphi_j$'s constitute a basis in $S_h$, there is a transformation from $\underline{u}$ to $\underline{\alpha}$ given by

$$\underline{\alpha} = T\underline{u}\,, \tag{33}$$

where $T$ has the form

$$T = \begin{pmatrix} \times & \cdots & \times & & & \\ \vdots & & \vdots & & & \\ \times & \cdots & \times & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & \ddots \end{pmatrix},$$

i. e., it is the identity in the interior, and with the upper corner block having a size corresponding to the support and number of modified basis functions. The approximation $\underline{u}$ then satisfies

$$\tilde{P}T\frac{d\underline{u}}{dt} = \frac{1}{h}\tilde{Q}T\underline{u}\,. \tag{34}$$

If $\underline{w}$ denotes the grid values of the true solution of the continuous problem, i. e., $\underline{w}_t = \underline{w}_x$, then

$$\tilde{P}T\underline{w}_x = \frac{1}{h}\tilde{Q}T\underline{w} + \underline{g}\,, \tag{35}$$

where $\underline{g}$ is the truncation error. Assume that basis functions are found such that $g_j = \mathcal{O}(h^3)$ in a fixed number of points adjacent to the boundary, and $g_j = \mathcal{O}(h^4)$ in the interior. Multiplying (35) by $T^T$ from the left we get

$$T^T\tilde{P}T\underline{w}_x = \frac{1}{h}T^T\tilde{Q}T\underline{w} + T^T\underline{g}\,, \tag{36}$$

where it follows that $T^T \underline{g}$ is of the same order as $\underline{g}$ because of the assumption that $T^T$ be bounded. Using (33) in (32) gives

$$\langle \underline{u}, \frac{1}{h}(T^T \tilde{Q}^T T + T^T \tilde{Q} T)\underline{u}\rangle_h = -u_0^2,$$

that is to say, $T^T \tilde{Q} T$ is almost antisymmetric. Thus, eq. (36) is third order accurate at the boundary, $P = T^T \tilde{P} T$ is SPD, and $Q = T^T \tilde{Q} T$ is almost antisymmetric. But this is impossible according to the first part of this proof, and we have arrived at a contradiction. This proves the theorem. □

# 4 Generalized Norms

We shall next discuss how the results of the previous section can be generalized. To make the presentation more concise we introduce the concept of $(p, q)$-approximation of $\partial/\partial x$.

**Definition 4.1** *A difference approximation $D$ is called a $(p, q)$-approximation of $\partial/\partial x$ if $D$ is at least $p$ th order accurate at the grid points $x_0, \ldots, x_{s-1}$, and at least $q$ th order accurate at $x_s, \ldots$; $s$ is a fixed number independent of the uniform grid spacing $h = x_{j+1} - x_j$.*

It has been demonstrated that there exist no matrices $P$ and $Q$, $P$ tridiagonal in the interior and SPD, $Q$ almost antisymmetric, such that $P^{-1}Q$ is a $(3, 4)$-approximation of $\partial/\partial x$. In particular, there exists no Galerkin method that results in a $(3, 4)$-approximation. Interpreting $D = P^{-1}Q$ as a finite difference operator, however, one can try to construct a different scalar product in which it is possible to establish the summation-by-parts property, which will be pursued in the following paragraphs.

Let $D = \frac{1}{h}P^{-1}Q$ be a $(p, q)$-approximation of $\partial/\partial x$, where it is no longer assumed that $P$ be SPD, nor that $Q$ be almost antisymmetric. This allows us to consider more general operators than those of section 3, since $P$ was SPD. The operators $P$ and $Q$ are given by

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{pmatrix}, \quad Q = \begin{pmatrix} Q_{11} & Q_{12} \\ -Q_{12}^T & Q_{22} \end{pmatrix}, \tag{37}$$

where $P_{11}$ and $Q_{11}$ are arbitrary $(s \times s)$-matrices; $P_{11}$ must of course be non-singular in order for $D$ to exist. The structure of the remaining blocks is

$$P_{12} = \begin{pmatrix} 0 & & & & \\ \vdots & & & & \\ 0 & & & & \\ p_{r_1} & & & & \\ \vdots & \ddots & & & \\ p_1 & \cdots & p_{r_1} & 0 & \cdots \end{pmatrix}, \quad P_{22} = \begin{pmatrix} p_0 & p_1 & \cdots & p_{r_1} & \\ p_1 & p_0 & \ddots & & \ddots \\ \vdots & \ddots & \ddots & & \\ p_{r_1} & & & & \\ & \ddots & & & \end{pmatrix}, \tag{38}$$

$$Q_{12} = \begin{pmatrix} 0 & & & & \\ \vdots & & & & \\ 0 & & & & \\ q_{r_2} & & & & \\ \vdots & \ddots & & & \\ q_1 & \dots & q_{r_2} & 0 & \dots \end{pmatrix}, \quad Q_{22} = \begin{pmatrix} 0 & q_1 & \dots & q_{r_2} & \\ -q_1 & 0 & \ddots & & \ddots \\ \vdots & \ddots & \ddots & & \\ -q_{r_2} & & & & \\ & \ddots & & & \end{pmatrix}.$$

If $D$ is to be almost antisymmetric with respect to some scalar product $(\cdot,\cdot)$ it is necessary that

$$(\underline{u}, D\underline{u}) = \langle \underline{u}, \frac{1}{h}\hat{P}P^{-1}Q\underline{u}\rangle_h = -\frac{1}{2}u_0^2\,, \tag{39}$$

where $\hat{P}$ is the SPD-matrix representing the scalar product. It is then natural to try a scalar product $(\underline{u}, \underline{v}) = \langle \underline{u}, \hat{P}\underline{v}\rangle_h$, where $\hat{P}$ is obtained by modifying $P$ at the boundary. Hence, $\hat{P}$ may be written as

$$\hat{P} = \begin{pmatrix} \hat{P}_{11} & P_{12} \\ P_{12}^T & P_{22} \end{pmatrix}, \tag{40}$$

where $\hat{P}_{11} \in \mathrm{R}^{s \times s}$. It is no restriction to assume identical block size of $P_{11}$ and $\hat{P}_{11}$, since it is possible to extend the size of $P_{11}$ so as to match that of $\hat{P}_{11}$; $s$ must of course be independent of the mesh size $h$.

Define $H = \hat{P}P^{-1}$, i. e., $\hat{P} = HP$, where $H$ is blocked as

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}. \tag{41}$$

From eq. (39) it follows that $HQ$ must be almost antisymmetric; $\hat{P} = HP$ requires that $HP$ be SPD with the same interior structure as $P$. Hence, given a $(p,q)$-approximation $D = \frac{1}{h}P^{-1}Q$, we must find a matrix $H$ such that

$$\begin{aligned} P_{22} &= H_{21}P_{12} + H_{22}P_{22} \\ P_{12} &= H_{11}P_{12} + H_{12}P_{22} \\ P_{12}^T &= H_{21}P_{11} + H_{22}P_{12}^T \\ 0 &= H_{21}Q_{12} + H_{22}Q_{22} + (H_{21}Q_{12} + H_{22}Q_{22})^T \\ 0 &= H_{11}Q_{12} + H_{12}Q_{22} + (H_{21}Q_{11} - H_{22}Q_{12}^T)^T\,, \end{aligned} \tag{42}$$

which follows from the symmetry of $\hat{P} = HP$, the antisymmetry of $HQ$, and the structure assumption (40). The equations above will be used to derive constraints on $H$.

The submatrix $P_{22}$ defines a difference stencil, which in the interior points yields the characteristic equation

$$p_{r_1}x^{2r_1} + \ldots + p_0 x^{r_1} + \ldots + p_{r_1} = 0 \tag{43}$$

with $\nu$ distinct roots $x_k, k = 0, \ldots, \nu - 1$, each with multiplicity $\mu_k$. The general solution of the corresponding difference equation may then be written as

$$\sum_{k=0}^{\nu-1} a_k(j) x_k^j ,$$

where $a_k(j)$ is a polynomial in $j$

$$a_k(j) = \sum_{l=0}^{\mu_k - 1} a_{kl} j^l .$$

Furthermore, with each root $x_k, k = 0, \ldots, \nu - 1$, we associate the upper triangular matrix

$$U_k = \begin{pmatrix} u_{00}^{(k)} & u_{01}^{(k)} & \cdots & u_{0,\mu_k-1}^{(k)} \\ & u_{11}^{(k)} & \cdots & u_{1,\mu_k-1}^{(k)} \\ & & \ddots & \vdots \\ & & & u_{\mu_k-1,\mu_k-1}^{(k)} \end{pmatrix} , \tag{44}$$

where

$$u_{ml}^{(k)} = \left( -q_{r_2}(-r_2)^{l-m} x_k^{-r_2} - \ldots - q_1(-1)^{l-m} x_k^{-1} + q_1 x_k + \ldots + q_{r_2} r_2^{l-m} x_k^{r_2} \right) \begin{pmatrix} l \\ m \end{pmatrix} \tag{45}$$

with $0 \leq m, l \leq \mu_k - 1$. The binomial coefficients are defined to be zero whenever $m > l$. Thus, all matrix elements of $U_k$ are defined by eq. (45).

**Assumption 4.1** *For each root $x_k, k = 0, \ldots, \nu - 1$, of the characteristic equation (43), the associated upper triangular matrix $U_k$ (44) is non-singular.*

**Lemma 4.1** *Let the matrices $P$ and $Q$ be defined by eqs. (37), (38), and suppose that assumption (4.1) holds. If there exists a matrix $H$ such that $\hat{Q} = HQ$ is almost antisymmetric and $\hat{P} = HP$ symmetric, $\hat{P}$ given by eq. (40), then*

$$H = \begin{pmatrix} H_{11} & 0 \\ 0 & I \end{pmatrix} , \quad H_{11} = \begin{pmatrix} h_{00} & \cdots & h_{0,s-r-1} & 0 & \cdots & 0 \\ \vdots & & \vdots & & & \\ h_{s-r-1,0} & \cdots & h_{s-r-1,s-r-1} & & & \\ & & & 1 & & \\ \vdots & & \vdots & & \ddots & \\ h_{s-1,0} & \cdots & h_{s-1,s-r-1} & 0 & \cdots & 1 \end{pmatrix} ,$$

*where* $r = \max(r_1, r_2)$.

**Proof**

From the first equation of eqs. (42) one obtains the inhomogeneous difference equation

$$p_{r_1} h_{ij} + \ldots + p_0 h_{i,j+r_1} + \ldots + p_{r_1} h_{i,j+2r_1} = p_{r_1} \delta_{i,j+2r_1} + \ldots + p_0 \delta_{i,j+r_1} + \ldots + p_{r_1} \delta_{ij}\,, \quad (46)$$

where $\delta_{ij}$ is the Kronecker delta and $i \geq s,\ j \geq s - r_1$. To begin with we consider $j > i$, which implies eq. (43). Hence

$$h_{ij} = \sum_{k=0}^{\nu-1} a_k(j) x_k^j, \quad j > i$$

where

$$a_k(j) = \sum_{l=0}^{\mu_k - 1} a_{kl} j^l$$

Next we solve for $h_{ii}$ by setting $j = i$ in eq. (46). Using the general solution for $j > i$ yields

$$\begin{aligned} p_{r_1} h_{ii} = \; & -\sum_{k=0}^{\nu-1} \sum_{l=0}^{\mu_k-1} a_{kl} \left[ p_{r_1} i^l x_k^i + \ldots + p_0 (i + r_1)^l x_k^{i+r_1} + \ldots + p_{r_1} (i + 2r_1)^l x_k^{i+2r_1} \right] \\ & + p_{r_1} \sum_{k=0}^{\nu-1} a_k(i) x_k^i + p_{r_1}\,, \end{aligned}$$

where the expression between the square brackets vanishes, since $x_k$ is a root with multiplicity $\mu_k$. This implies

$$h_{ii} = \sum_{k=0}^{\nu-1} a_k(i) x_k^i + 1\,.$$

In a similar vein $h_{i,i-1}$ is determined as

$$h_{i,i-1} = \sum_{k=0}^{\nu-1} a_k(i-1) x_k^{i-1}\,.$$

An induction argument shows that

$$h_{ij} = \sum_{k=0}^{\nu-1} a_k(j) x_k^j, \qquad s - r_1 \leq j < i - 1\,.$$

Hence

$$h_{ij} = \sum_{k=0}^{\nu-1} a_k(j) x_k^j + \delta_{ij}, \quad i \geq s, \quad j \geq s - r_1\,. \quad (47)$$

Next, we make use of the fourth equation of eqs. (42), which implies that the diagonal elements of $H_{21} Q_{12} + H_{22} Q_{22}$ are zero, i. e.,

$$-q_{r_2} h_{i,i-r_2} - \ldots - q_1 h_{i,i-1} + q_1 h_{i,i+1} + \ldots + q_{r_2} h_{i,i+r_2} = 0\,.$$

Inserting eq. (47) gives

$$\begin{aligned} 0 = \sum_{k=0}^{\nu-1}\sum_{l=0}^{\mu_k-1} &= -q_{r_2} a_{kl}(i-r_2)^l x_k^{i-r_2} - \ldots - q_1 a_{kl}(i-1)^l x_k^{i-1} \\ &\quad + q_1 a_{kl}(i+1)^l x_k^{i+1} + \ldots + q_{r_2} a_{kl}(i+r_2)^l x_k^{i+r_2} \,. \end{aligned}$$

The binomial theorem yields

$$\sum_{k=0}^{\nu-1}\sum_{l=0}^{\mu_k-1}\sum_{m=0}^{l} a_{kl} u_{ml}^{(k)} i^m x_k^i = 0, \quad i \geq s \,,$$

where we have used eq. (45). Let $\mu = \max_k(\mu_k)$, and extend the coefficients $a_{kl}$ by defining $a_{kl} = 0$ whenever $\mu_k \leq l \leq \mu - 1$. Similarly, let $u_{ml}^{(k)} = 0$ for $l < m \leq \mu - 1$, whence

$$\sum_{k=0}^{\nu-1}\sum_{l=0}^{\mu-1}\sum_{m=0}^{\mu-1} a_{kl} u_{ml}^{(k)} i^m x_k^i = 0, \quad i \geq s \,.$$

Rearranging the sum one obtains

$$\sum_{k=0}^{\nu-1}\sum_{m=0}^{\mu-1}\left(\sum_{l=0}^{\mu-1} a_{kl} u_{ml}^{(k)}\right) i^m x_k^i = 0, \quad i \geq s \,.$$

The vectors $(i^m x_k^i)$, $i \geq s$, are linearly independent for all $k$ and $m$. Thus

$$\sum_{l=0}^{\mu-1} a_{kl} u_{ml}^{(k)} = 0, \quad 0 \leq k \leq \nu - 1, \quad 0 \leq m \leq \mu - 1 \,.$$

Since $a_{kl} = 0$ for $\mu_k \leq l \leq \mu - 1$, it suffices to sum from $l = 0$ to $l = \mu_k - 1$. Furthermore, when $\mu_k \leq m \leq \mu - 1$ it follows that $l < m$ for $0 \leq l \leq \mu_k - 1$, which according to the extension of $u_{ml}^{(k)}$ implies that the sum vanishes identically. Consequently, no conditions are imposed for $\mu_k \leq m \leq \mu - 1$, and we are left with

$$\sum_{l=0}^{\mu_k-1} a_{kl} u_{ml}^{(k)} = 0, \quad 0 \leq k \leq \nu - 1, \quad 0 \leq m \leq \mu_k - 1 \,.$$

Defining $\underline{a}_k = (a_{k0} \ldots a_{k,\mu_k-1})^T$ gives the following system

$$U_k \underline{a}_k = 0, \quad 0 \leq k \leq \nu - 1 \,,$$

which only has the trivial solution $\underline{a}_k = 0$, since $U_k$ is assumed to be non-singular. From eq. (47) it follows that

$$h_{ij} = \delta_{ij}, \quad i \geq s, \quad j \geq s - r_1 \,. \tag{48}$$

In particular, $H_{22} = I$. By means of the third equation (42), $H_{22} = I$ leads to $H_{21} = 0$, since $P_{11}$ is non-singular.

To determine $H_{11}$ and $H_{12}$ we begin by examining the second equation of eqs. (42), which implies

$$p_{r_1} h_{ij} + \ldots + p_0 h_{i,j+r_1} + \ldots + p_{r_1} h_{i,j+2r_1} = \mathrm{RHS}, \quad 0 \le i \le s-1 \quad j \ge s - r_1 \,,$$

where the right hand side (RHS) is defined by

$$\mathrm{RHS} = \begin{cases} 0 & 0 \le i \le s - r_1 - 1 \\ p_{r_1 - n+1} \delta_{i,j+n-1} + \ldots + p_{r_1} \delta_{ij} & i = s - r_1 - 1 + n, \; 1 \le n \le r_1 \,. \end{cases}$$

Arguing exactly as before one obtains

$$h_{ij} = \sum_{k=0}^{\nu-1} a_k(j) x_k^j + \delta_{ij}, \quad 0 \le i \le s-1, \quad j \ge s - r_1 \,.$$

Finally, we use the last of eqs. (42), which simplifies to

$$H_{11} Q_{12} + H_{12} Q_{22} = Q_{12} \,.$$

Using the equations for the diagonal elements we recover

$$-q_{r_2} h_{i,i-r_2} - \ldots - q_1 h_{i,i-1} + q_1 h_{i,i+1} + \ldots + q_{r_2} h_{i,i+r_2} = 0 \,,$$

which again yields

$$h_{ij} = \delta_{ij}, \quad 0 \le i \le s-1, \quad j \ge s - r_1 \,. \tag{49}$$

Hence, $H_{12} = 0$. Furthermore, eq. (49) also shows that

$$H_{11} = \begin{pmatrix} h_{00} & \ldots & h_{0,s-r_1-1} & 0 & \ldots & 0 \\ \vdots & & \vdots & & & \\ h_{s-r_1-1,0} & \ldots & h_{s-r_1-1,s-r_1-1} & & & \\ & & & 1 & & \\ \vdots & & \vdots & & \ddots & \\ h_{s-1,0} & \ldots & h_{s-1,s-r_1-1} & 0 & \ldots & 1 \end{pmatrix} .$$

This proves the lemma if $r_1 \ge r_2$. Suppose that $r_2 > r_1$. Then

$$HQ = \begin{pmatrix} H_{11} Q_{11} & H_{11} Q_{12} \\ -Q_{12}^T & Q_{22} \end{pmatrix} .$$

Since $HQ$ is almost antisymmetric, it follows that

$$H_{11} Q_{12} = Q_{12} \,. \tag{50}$$

Partition

$$H_{11} = \begin{pmatrix} \tilde{H}_{11} & 0 \\ \tilde{H}_{21} & I \end{pmatrix} , \quad Q_{12} = \begin{pmatrix} \tilde{Q}_{11} & \tilde{Q}_{12} \\ \tilde{Q}_{21} & \tilde{Q}_{22} \end{pmatrix} ,$$

where $\tilde{H}_{11}$, $\tilde{Q}_{11} \in \mathrm{R}^{(s-r_1)\times(s-r_1)}$, and $\tilde{H}_{21}$, $\tilde{Q}_{21} \in \mathrm{R}^{r_1\times(s-r_1)}$. Combining this partition with eq. (50) yields

$$\begin{pmatrix} \tilde{H}_{11} \\ \tilde{H}_{21} \end{pmatrix} \tilde{Q}_{11} = \begin{pmatrix} \tilde{Q}_{11} \\ 0 \end{pmatrix} . \tag{51}$$

The explicit structure of $\tilde{Q}_{11}$ is

$$\tilde{Q}_{11} = \begin{pmatrix} 0 & \dots & & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & \dots & & \dots & 0 \\ q_{r_2} & & & & \\ \vdots & \ddots & & & \vdots \\ q_{r_1+1} & \dots & q_{r_2} & 0 & \dots & 0 \end{pmatrix} ,$$

where column $r_2 - r_1 - 1$ is the last non-zero column. Equating columns $r_2 - r_1 - 1$ results in

$$\begin{pmatrix} h_{00} & \dots & h_{0,s-r_1-1} \\ \vdots & & \vdots \\ & & \\ \vdots & & \vdots \\ h_{s-1,0} & \dots & h_{s-1,s-r_1-1} \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ q_{r_2} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ q_{r_2} \\ 0 \\ \vdots \\ 0 \end{pmatrix} .$$

The element $q_{r_2}$ is located on row $s - r_1 - 1$ in both column vectors, i. e.,

$$h_{j,s-r_1-1} = \delta_{j,s-r_1-1}, \quad 0 \le j \le s-1 .$$

If $r_2 = r_1 + 1$ we are done. Otherwise, assume the induction hypothesis $h_{jl} = \delta_{jl}$ for $0 \le j \le s-1$, $k \le l \le s - r_1 - 1$, where $k$ satisfies $s - r_2 < k \le s - r_1 - 1$. Equating columns $k + r_2 - 1 - s$ and using the induction hypothesis one obtains

$$h_{j,k-1} = \delta_{j,k-1}, \quad 0 \le j \le s-1 .$$

Since the result is true for $k = s - r_1 - 1$, it follows by the axiom of induction that

$$H_{11} = \begin{pmatrix} h_{00} & \dots & h_{0,s-r_2-1} & 0 & \dots & 0 \\ \vdots & & \vdots & & & \\ h_{s-r_2-1,0} & \dots & h_{s-r_2-1,s-r_2-1} & & & \\ & & & 1 & & \\ \vdots & & \vdots & & \ddots & \\ h_{s-1,0} & \dots & h_{s-1,s-r_2-1} & 0 & \dots & 1 \end{pmatrix} ,$$

which concludes the lemma. □

**Corollary 4.1** *The conclusion of lemma (4.1) holds if the characteristic polynomials of $P_{22}$ and $Q_{22}$ are relatively prime.*

**Proof**

The matrices $U_k$ are non-singular iff the diagonal elements are non-zero. But

$$u_{ll}^{(k)} = \left(-q_{r_2}x_k^{-r_2} - \ldots - q_1 x_k^{-1} + q_1 x_k + \ldots + q_{r_2} x_k^{r_2}\right), \quad 0 \le l \le \mu_k - 1\,,$$

where $x_k$ is a root of the characteristic polynomial of $P_{22}$. Thus, $u_{ll}^{(k)} \neq 0$ iff $x_k$ is not a root of the characteristic polynomial of $Q_{22}$, i. e., iff the characteristic polynomials of $P_{22}$ and $Q_{22}$ are relatively prime. □

Since $D = \frac{1}{h}P^{-1}Q$ is a $(p,q)$-approximation of $\partial/\partial x$ it follows that

$$(P\underline{v})_i = \frac{1}{h}(Q\underline{u})_i + \mathcal{O}(h^p) \quad i = 0, \ldots, s-1$$

and

$$(P\underline{v})_i = \frac{1}{h}(Q\underline{u})_i + \mathcal{O}(h^q) \quad i = s, s+1, \ldots$$

where $\underline{v} = (u_x(x_0)\ u_x(x_1)\ \ldots\ )^T$. The interior order of accuracy $q$ is assumed to exceed that of the boundary. It may of course happen that one gets the interior order of accuracy at *some* of the boundary points, if the interior operator extends into the boundary operator $P_{11}, Q_{11}$. From lemma (4.1) it follows immediately that

$$HP = \begin{pmatrix} H_{11}P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{pmatrix}, \quad HQ = \begin{pmatrix} H_{11}Q_{11} & Q_{12} \\ -Q_{12}^T & Q_{22} \end{pmatrix}. \tag{52}$$

Hence,

$$(HP\underline{v})_i = \frac{1}{h}(HQ\underline{u})_i + \mathcal{O}(h^p) \quad i = 0, \ldots, s-1$$

and

$$(HP\underline{v})_i = \frac{1}{h}(HQ\underline{u})_i + \mathcal{O}(h^q) \quad i = s, s+1, \ldots$$

i. e., $D = \frac{1}{h}\hat{P}^{-1}\hat{Q}$ is a $(p,q)$-approximation of $\partial/\partial x$, where $\hat{P}$ is SPD, and $\hat{Q}$ is almost antisymmetric.

We now return to the case where $P$ and $Q$, defined by eqs. (20)-(21), are tridiagonal in the interior.

**Theorem 4.1** *There exists no $(3,4)$-approximation $D = \frac{1}{h}P^{-1}Q$ of $\partial/\partial x$, $P$ and $Q$ defined by eqs. (20)-(21), that is almost antisymmetric with respect to a tridiagonal norm $\hat{P}$ such that $\hat{P}_{22} = P_{22}$.*

**Proof**

Suppose there were a (3, 4)-approximation $D = \frac{1}{h}P^{-1}Q$ and a norm $\hat{P}$ such that $(\underline{u}, D\underline{u}) = \langle \underline{u}, \hat{P}D\underline{u}\rangle_h = -1/2u_0^2$. The characteristic polynomials of $P_{22}$ and $Q_{22}$ are given by $p(x) = x^2 + 4x + 1$ and $q(x) = x^2 - 1$. Clearly, they are relatively prime. Thus

$$\hat{P}P^{-1} = H = \begin{pmatrix} H_{11} & 0 \\ 0 & I \end{pmatrix},$$

where

$$H_{11} = \begin{pmatrix} h_{00} & \dots & h_{0,s-2} & 0 \\ \vdots & & \vdots & \vdots \\ h_{s-1,0} & \dots & h_{s-1,s-2} & 1 \end{pmatrix}$$

according to corollary (4.1). From the assumptions on $D$ we conclude that $D = \frac{1}{h}\hat{P}^{-1}\hat{Q}$, $\hat{P} = HP$, $\hat{Q} = HQ$ given by eq. (52), is a (3, 4)-approximation of $\partial/\partial x$, where $\hat{P}$ is SPD, and $\hat{Q}$ is almost antisymmetric. But this contradicts theorem (3.1). □

It should be pointed out that it is not *a priori* clear that no such norm can exist. If $\hat{P}$ is defined by eq. (40), with $P_{12}$ and $P_{22}$ given by eqs. (20)–(21), one cannot apply theorem (3.1) directly to $\hat{P} = HP$ (symmetric) and $\hat{Q} = HQ$ (antisymmetric), because $\hat{Q} = HQ$ need not be tridiagonal in the interior, which is needed in order to apply the basic theorem (3.1).

# References

[1] M. Carpenter, D. Gottlieb, and S. Abarbanel. The stability of numerical boundary treatments for compact high-order finite-difference schemes. Technical Report 91-71, ICASE, Sept. 1991.

[2] M. Carpenter, D. Gottlieb, and S. Abarbanel. Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: Methodology and application to high-order compact schemes. Technical Report 93-21, ICASE, NASA Langley Research Center, Hampton, VA 23681-0001, 1993.

[3] B. Gustafsson. The convergence rate for difference approximations to general mixed initial boundary value problems. *SIAM J. Numer. Anal.*, 18(2):179–190, Apr. 1981.

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

[5] H.-O. Kreiss and G. Scherer. On the existence of energy estimates for difference approximations for hyperbolic systems. Technical report, Dept. of Scientific Computing, Uppsala University, 1977.

[6] P. Olsson. Stable approximation of symmetric hyperbolic and parabolic equations in several space dimensions. Technical Report 138, Dept. of Scientific Computing, Uppsala Univ., Uppsala, Sweden, Dec. 1991.

[7] P. Olsson. The numerical behavior of stable high-order finite difference methods. Technical Report 140, Dept. of Scientific Computing, Uppsala Univ., Uppsala, Sweden, Feb. 1992.

[8] P. Olsson. Summation by parts, projections, and stability. Technical Report 93.04, RIACS, June 1993.

[9] S. Orszag and M. Israeli. Numerical simulation of viscous incompressible flows. *Ann. Rev. Fluid Mech.*, 5, 1974.

[10] V. Thomée and B. Wendroff. Convergence estimates for Galerkin methods for variable coefficient initial value problems. *SIAM J. Numer. Anal.*, 11(5):1059–1068, Oct. 1974.

[11] L. B. Wahlbin. On superconvergence up to boundaries in finite element methods: a counterexample. *SIAM J. Numer. Anal.*, 29(4):937–946, Aug. 1992.