Appl. Math. and Comp. Sci., 1995, vol.5, No.4, 579–596

NON-OPTIMAL BEHAVIOUR OF FINITE ELEMENT METHODS FOR FIRST ORDER HYPERBOLIC PROBLEMS

TODD E. PETERSON*, DAVID B. SHUSTER**

We investigate two explicit finite element methods for linear scalar first order hyperbolic equations, one using a continuous piecewise-polynomial approximation and one using a discontinuous approximation. Several aspects of the performance of these methods are considered; in particular, the sharpness of some existing error estimates, and crosswind spread estimates. For the discontinuous method, we give a new error estimate for the case of piecewise-constants.

1. Introduction

The model problem we consider is the first order hyperbolic equation

$$\alpha \cdot \nabla u + \beta u = f \qquad \text{in} \quad \Omega \tag{1}$$

$$u = g \qquad \text{on} \quad \partial\Omega_- \tag{2}$$

where Ω is a bounded polygonal domain in $\mathbb{R}^2$ and $\partial\Omega_-$ its inflow boundary, α is a constant non-zero vector, and β is a bounded function. By the inflow boundary ∂D_- of a set D we mean $\{x \in \partial D : \alpha \cdot n(x) < 0\}$, where $n(x)$ is the unit outward normal to D at x. The outflow boundary is defined as $\partial D_+ = \{x \in \partial D : \alpha \cdot n(x) \geq 0\}$.

To define an approximation u^h, we first assume that Ω is divided into triangular elements T, each of diameter roughly h. We will consider only quasi-uniform triangulations, meaning that the ratio of the largest to smallest element edge is bounded from above by a constant independent of h, and that all angles of the elements are bounded from below by another such constant. We are interested in *explicit* methods, for which it is possible to compute u^h one element at a time. To formulate such methods, it is first necessary to order the elements in a way consistent with the domain of dependency requirements. Specifically, we assume the elements are ordered in such a way that

$$\partial T^n_- \subset \partial\Omega_- \cup \bigcup_{m=1}^{n-1} \partial T^m_+$$

That this is always possible is proved in (Lesaint and Raviart, 1974), and in (Falk and Richter, 1987) for the more general case of non-constant α. One can also say that

* Department of Mathematical Sciences, George Mason University, Fairfax Virginia 22030 USA
** Institute of Applied Mathematics, University of Virginia, Charlottesville Virginia 22903 USA

the solution develops one layer at a time, where, for example, the first layer consists of all elements with $\partial T_- \subset \partial\Omega_-$.

The two methods we will study were both introduced in (Reed and Hill, 1973), in the context of solving the neutron transport equation. In both, the approximation u^h is a piecewise polynomial of some fixed degree p. In the method that has come to be known as the discontinuous Galerkin (DG) method, no further restriction is placed on the form of u^h, so that the approximation is not necessarily continuous across inter-element boundaries. The other method we study, which we will refer to as the continuous Galerkin (CG) method, generates a globally continuous approximation.

Once the elements have been ordered as described above, the formulation of any explicit method reduces to the following: given a single element T and boundary data on ∂T_-, how is $u^h|_T$ determined? For the DG method, we require

$$\int_T (\alpha \cdot \nabla u^h + \beta u^h)v^h + \int_{\partial T_-} (u^h_+ - u^h_-)v^h|\alpha \cdot n| = \int_T fv^h \qquad \forall v^h \in P_p(T)$$

where $P_p(T)$ denotes polynomials of degree p on T, and $w_\pm(x)$ is the limit of $w(x \pm \epsilon\alpha)$ as ϵ decreases to zero. On $\partial\Omega_-$ we take u^h_- to be the given data g, or some suitable interpolant of it.

The CG approximation is defined on each element T by requiring continuity along ∂T_-, and requiring

$$\int_T (\alpha \cdot \nabla u^h + \beta u^h)v^h = \int_T fv^h \qquad \forall v^h \in P_{p-i}(T)$$

where i is the number of inflow sides of the element T. Note that this will always be either 1 or 2, and we refer to such elements as Type 1 elements or Type 2 elements, respectively. The reduction in the size of the test space, relative to the DG method, is necessary to account for the fact that some degrees of freedom will be fixed *a priori* by continuity. On $\partial\Omega_-$, we take u^h to be a suitable interpolant of g.

The purpose of this paper is to investigate the performance of these two methods. This paper is primarily a summary of (Peterson, 1991b) and (Shuster, 1994).

Before proceeding we mention another class of explicit methods known as reduced continuity methods, in which the approximation is in general discontinuous across inter-element boundaries, but continuity of certain moments is enforced (see Cai and Falk, 1994).

2. Known Estimates

In this section we briefly summarize previous analyses of the DG and CG methods.

2.1. Discontinuous Galerkin

Assuming only that the triangulation is quasi-uniform, Johnson and Pitkäranta (1986) obtained a bound for the error in a certain mesh dependent norm, optimal for that norm, but which for $L_2(\Omega)$ implies only that

$$||u^h - u||_\Omega \leq Ch^{p+1/2}||u||_{p+1},\Omega \tag{3}$$

where the norms are those of $L_2(\Omega)$ and the Sobolev space $H^{p+1}(\Omega)$, respectively. The rate of convergence guaranteed by this result is less than the optimal rate of $p+1$, and it is now known that in general the above $L_2(\Omega)$ estimate cannot be improved: numerical examples presented in (Peterson, 1991a) (and summarized below) show that additional hypotheses are necessary to obtain an improved estimate.

In two circumstances an estimate of the form

$$||u^h - u||_\Omega \leq Ch^{p+1}||u||_{p+2,\Omega} \tag{4}$$

is known to hold. Lesaint and Raviart (1974) proved this for rectangular elements (in which case u^h is, e.g., piecewise bilinear), and Richter (1988) did so for $\beta = 0$ and 'semi-uniform' triangulations (defined below) which in addition have all element edges bounded away from the characteristic direction α. One might expect that alignment of the triangulation with the characteristic direction would be desirable — we will consider this further below. Note that to obtain an optimal order estimate both of these results place some restriction on the triangulation and assume extra regularity of the exact solution.

It has also been noted (Johnson *et al.*, 1984) that crosswind spread is limited to $O(h^{1/2})$, see (Peterson, 1990) for a proof. (The exact statement involves a logarithmic factor which will be ignored in the present discussion.) This is, for example, the extent to which a jump discontinuity in the boundary data g is smeared in the direction perpendicular to the characteristic line along which it propagates.

A pointwise estimate of the form

$$||u^h - u||_{\infty,\Omega} \leq Ch^{p+1/4}||u||_{p+1,\infty,\Omega} \tag{5}$$

has been proved (Peterson, 1990) using a discrete Green's function argument which takes advantage of the known crosswind spread estimate, a technique originally developed in (Johnson *et al.*, 1987) and (Niijima, 1990) for analysis of the streamline diffusion finite element method. The norms above are those of $L_\infty(\Omega)$ and the Sobolev space $W_\infty^{p+1}(\Omega)$, respectively. This result is probably not sharp, but for general quasi-uniform triangulations the exponent of h in (5) could be increased to at most $p + 1/2$ — see (Peterson, 1991a) and below.

2.2. Continuous Galerkin

For the CG method, Falk and Richter (1987) proved the following $L_2(\Omega)$ estimate:

$$||u - u_h||_\Omega \leq Ch^{p+1/4}||u||_{p+1,\Omega} \tag{6}$$

This estimate was proved for quasi-uniform triangulations under the hypothesis that no triangle edge is aligned with the characteristic direction. As noted above, such a hypothesis seems counterintuitive. The necessity of this assumption was investigated numerically in (Shuster, 1994) and will be discussed in Section 5.

Crosswind spread for the continuous Galerkin method has been proven in (Falk and Richter, 1992) to be limited to at most $O(h^{1/2})$ (as above, the exact statement involves a slowly varying logarithmic term which will be ignored).

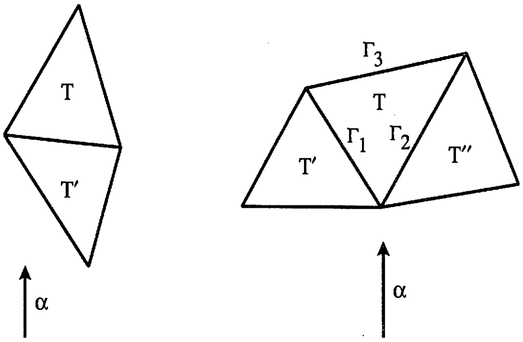

Fig. 1.

3. Discontinuous Galerkin — Piecewise Constants

In this section we study the discontinuous Galerkin method in the simplest possible setting, that of piecewise constants and $\beta, f = 0$.

For $p = 0$ the discontinuous Galerkin approximation to $\alpha \cdot \nabla u = 0$ is easily verified to be as follows: on Type 1 triangles

$$u_T^h = u_{T'}^h \tag{7}$$

and on Type 2 triangle

$$u_T^h = \lambda u_{T'}^h + \bar{\lambda} u_{T''}^h \tag{8}$$

where

$$\lambda = \frac{|\Gamma_1|\,|\alpha \cdot n_1|}{|\Gamma_3|\,|\alpha \cdot n_3|}, \qquad \bar{\lambda} = \frac{|\Gamma_2|\,|\alpha \cdot n_2|}{|\Gamma_3|\,|\alpha \cdot n_3|} = 1 - \lambda$$

See Fig. 1. Here n_i denotes the outward normal to edge Γ_i of T.

Consider the problem

$$u_y = 0 \qquad \text{on} \qquad \Omega \tag{9}$$

$$u(x,0) = g(x) \tag{10}$$

with $\Omega = (-\infty, \infty) \times (0,1)$. Assume a semi-uniform triangulation in the sense of Richter (1988). That is, assume that the elements lie in bands, each band consisting of one layer of Type 1 elements and one layer of Type 2 elements, the elements within each layer being identical up to a fixed translation. The j-th band is characterized by a λ_j defined as above. We also assume that the total number of bands is $O(h^{-1})$. We align the enumeration such that, for $j \geq 3$, the downwind vertex of the triangle $T_{0,j}$ is shared by the common vertex (upwind) of $T_{-1/2,j-1}$ and $T_{1/2,j-1}$; these two triangles in turn, straddle $T_{0,j-2}$ (see Fig. 2.) The infinite extent of the domain is of no consequence, as the numerical domain of dependence of any particular element is finite. (In general, the ordering property discussed earlier may fail on unbounded domains. However, the hypothesis of semi-uniformity prevents such a failure. Note that the property of semi-uniformity is dependent on the characteristic direction α.)

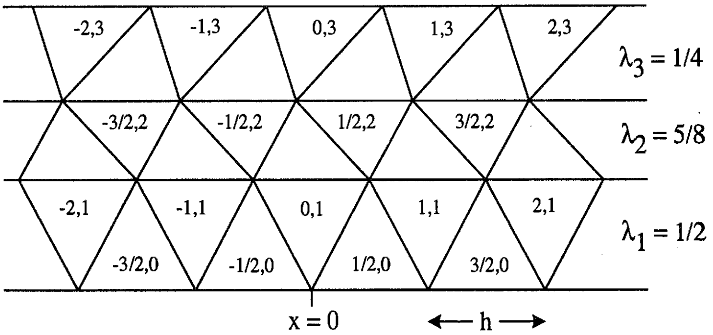

Fig. 2.

We will develop an explicit expression for u^h. It will involve certain coefficients defined as follows:

$$c_{jk} = \sum_{|\alpha|=k} \lambda_1^{\alpha_1} \cdots \lambda_j^{\alpha_j}$$

for $k = 0, \dots, j$, where $\lambda^0 \equiv \lambda$, $\lambda^1 \equiv \bar{\lambda}$, $\alpha \in \{0,1\}^j$, and $|\alpha| = \sum_l \alpha_l$. We put $c_{00} = 1$ and $c_{jk} = 0$ for $k < 0$ and $k > j$. Before proceeding further we establish properties of these coefficients that will be used below.

Lemma 1. *The coefficients c_{jk} have the following properties*:

$$c_{j+1,k} = \lambda_{j+1} c_{jk} + \bar{\lambda}_{j+1} c_{j,k-1} \tag{11}$$

$$\sum_{k=0}^{j} c_{jk} = 1 \tag{12}$$

$$\sum_{k=0}^{j} k c_{jk} = \sum_{k=1}^{j} \bar{\lambda}_k \tag{13}$$

$$\sum_{k=0}^{j} k^2 c_{jk} = \sum_{\substack{k \neq l \\ 1 \leq k,l \leq j}} \bar{\lambda}_k \bar{\lambda}_l + \sum_{k=1}^{j} \bar{\lambda}_k \tag{14}$$

Proof. For $j = 0$, (11) follows from the convention $c_{00} = 1$. For $j > 0$ we compute as follows:

$$\begin{aligned} c_{j+1,k} &= \sum_{|\alpha|=k} \lambda_1^{\alpha_1} \cdots \lambda_{j+1}^{\alpha_{j+1}} \\ &= \lambda_{j+1}\Big(\sum_{|\beta|=k} \lambda_1^{\beta_1} \cdots \lambda_j^{\beta_j} \Big) + \bar{\lambda}_{j+1}\Big(\sum_{|\beta|=k-1} \lambda_1^{\beta_1} \cdots \lambda_j^{\beta_j} \Big) \\ &= \lambda_{j+1} c_{jk} + \bar{\lambda}_{j+1} c_{j,k-1} \end{aligned}$$

The remaining proofs are by induction, utilizing (11). The details are straightforward, and therefore are omitted. ■

Note that if $\lambda_j = 1/2$ for all, j then $c_{jk} = \frac{1}{2^j}\binom{j}{k}$ and (11)–(14) reduce to well known properties of the binomial coefficients.

Number the elements as indicated in Fig. 2 and let u_{ij} denote the value of u^h on T_{ij} (where i is allowed to represent half integers and j is taken to be an integer). Note that due to (7) we may identify a Type 1 element with the Type 2 element lying below it. The one exception is Type 1 elements in the first layer: the values of u^h there are determined entirely by the choice of u_-^h on the inflow boundary. The next result justifies the introduction of the coefficients c_{jk}.

Lemma 2. *The values of u^h are given by*

$$u_{ij} = \sum_{k=0}^{j} c_{jk}\, u_{k+i-j/2,0}$$

Proof. The proof is by induction. For $j = 1$ the above expression reduces to (8). In general, we use (8), the inductive hypothesis, and (11):

$$\begin{aligned} u_{i,j+1} &= \lambda_{j+1} u_{i-1/2,j} + \bar{\lambda}_{j+1} u_{i+1/2,j} \\ &= \lambda_{j+1} \sum_{k=0}^{j} c_{jk}\, u_{k+(i-1/2)-j/2,0} + \bar{\lambda}_{j+1} \sum_{k=0}^{j} c_{jk}\, u_{k+(i+1/2)-j/2,0} \\ &= \lambda_{j+1} \sum_{k=0}^{j} c_{jk}\, u_{k+i-(j+1)/2,0} + \bar{\lambda}_{j+1} \sum_{k=0}^{j} c_{jk}\, u_{(k+1)+i-(j+1)/2,0} \\ &= \sum_{k=0}^{j+1} (\lambda_{j+1} c_{jk} + \bar{\lambda}_{j+1} c_{j,k-1}) u_{k+i-(j+1)/2,0} \\ &= \sum_{k=0}^{j+1} c_{j+1,k}\, u_{k+i-(j+1)/2,0} \end{aligned}$$

This proves the lemma. ■

To simplify notation we consider below only u_{0j} for j odd. The results extend trivially to other elements. We take u_{i0} to be the midpoint interpolant of the boundary data g, so that $u_{i0} = g(ih)$. Then

$$u_{0j} = \sum_{k=0}^{j} c_{jk}\, g\Big(h(k - j/2)\Big) \tag{15}$$

Note that if $g \equiv 1$ then by (12) we have $u^h \equiv 1$. Similarly, (13) leads to an estimate for the case $g(x) = x$: in this case (15) becomes

$$u_{0j} = \sum_{k=0}^{j} c_{jk}\, h\Big(k - j/2\Big) = h\Big(\sum_{k=1}^{j} \bar{\lambda}_k - j/2\Big)$$

The location of the downwind vertex of T_{0j} is

$$x_{0j} = h\sum_{k=1}^{j-1} \bar{\lambda}_k \; - \; h(j-1)/2$$

Comparing u_{0j} to the value of the exact solution along $x = x_{0j}$ we have

$$|u_{0j} - g(x_{0j})| = |h\bar{\lambda}_j - h/2| \le h/2$$

Finally, we will use (14) to derive an error estimate for arbitrary $g \in C^{1,1}$. There will be no loss of generality in assuming that $g'(x_{0j}) = 0$ since this could be arranged by subtracting $g'(x_{0j})x$ from $g(x)$, and we have just seen that linear functions are approximated to optimal order. By (15) and the mean value theorem,

$$u_{0j} = \sum_{k=0}^{j} c_{jk}\Big[g(x_{0j}) + (h(k - j/2) - x_{0j})g'(\tilde{x}_k)]\Big]$$

for some $\tilde{x}_k$ between x_{0j} and $h(k - j/2)$. Thus using (12)

$$|u_{0j} - g(x_{0j})| = \left|\sum_{k=0}^{j} c_{jk}(h(k - j/2) - x_{0j})g'(\tilde{x}_k)\right|$$

With M the Lipschitz constant of g', we have

$$|g'(\tilde{x}_k)| = |g'(\tilde{x}_k) - g'(x_{0j})| \le M|\tilde{x}_k - x_{0j}| \le M|h(k - j/2) - x_{0j}|$$

so

$$\begin{aligned} |u_{0j} - g(x_{0j})| &\le M\sum_{k=0}^{j} c_{jk}\Big(h(k - \frac{j}{2}) - x_{0j}\Big)^2 \\ &= Mh^2\sum_{k=0}^{j} c_{jk}\Big(k - \sum_{k=1}^{j-1} \bar{\lambda}_k - \frac{1}{2}\Big)^2 \end{aligned}$$

$$
\begin{aligned}
&= Mh^2\left[\sum_{k=1}^{j-1}(\bar{\lambda}_k - \bar{\lambda}_k^2) + \frac{1}{4}\right] \\
&\le \frac{1}{4}Mh^2 j
\end{aligned}
$$

Since $j \le O(h^{-1})$, this gives an order h estimate.

From (7) and (8) it is clear that the method is stable in $L_\infty(\Omega)$ (with constant 1). Thus by interpolation we obtain the following result.

Theorem 1. *Let u^h be the piecewise constant discontinuous Galerkin approximation to (9)–(10) on a semi-uniform triangulation. Then for $r \ge 0$,*

$$
||u^h - u||_{\infty,\Omega} \le Ch^{\min(1,r/2)}||u||_{r,\infty,\Omega}
$$

Note first that in contrast to (Richter, 1988) this result is obtained only under the assumption of semi-uniformity, and we do not also assume the non-alignment condition. Next note that the optimal $O(h)$ rate of convergence is guaranteed only for $r = 2$, whereas for a typical interpolant, $r = 1$ would suffice. We will now show that our result is in fact sharp in this regard.

For this purpose consider $g(x) = |x|^r$, so that $u \in W^r_\infty(\Omega)$, and a uniform mesh with $\lambda_j = 1/2$ for all j. Then $x_{0j} = 0$ and the error on the element T_{0j} (since $g(0) = 0$) is

$$
u_{0j}h^r \sum_{k=0}^{j} c_{jk}|k - j/2|^r \ge h^r \sum_{k=j/2-\sqrt{j}/2}^{j/2-\sqrt{j}/4} c_{jk}|k - j/2|^r \ge h^r\left(\frac{\sqrt{j}}{4}\right)^r \sum_{k=j/2-\sqrt{j}/2}^{j/2-\sqrt{j}/4} c_{jk}
$$

For $\lambda_j \equiv 1/2$ this last sum can be shown to be bounded below by a positive constant independent of j — the details are given in Appendix. On the outflow boundary $j = O(h^{-1})$ and we obtain $u_{0j} \ge ch^{r/2}$. This shows that $\min(1, r/2)$ is in general the highest rate of convergence that can be expected.

The present situation is also a good one in which to examine the issue of crosswind spread. Consider taking the inflow boundary data to be

$$
g(x) = \begin{cases} 1, & \text{if } x < 0 \\ 0, & \text{if } x \ge 0 \end{cases}
$$

In a typical cross-section of the discrete solution at $y = 1$, there will be a region to the right of $x = 0$ on which the exact solution would be 0 but on which the discrete solution has significantly non-zero values. We will show that the width of this region is of order $h^{1/2}$. For the boundary data given above, (15) becomes

$$
u_{ij} = \sum_{k=0}^{j/2-i-1/2} c_{jk}
$$

For $i = \sqrt{j}/4 - 1/2$ we obtain, using again the result presented in Appendix for $\lambda_j \equiv 1/2$,

$$u_{ij} \geq \sum_{k=j/2-\sqrt{j}/2}^{j/2-\sqrt{j}/4} c_{jk} \geq c$$

Since u_{ij} is evidently decreasing in i, this implies that there are $O(\sqrt{j})$ elements to the right of $x = 0$ on which u^h has a significantly non-zero value, and at $y = 1$ where $j = O(1/h)$ this gives a width of order $h^{1/2}$. Thus the known crosswind spread estimate for the DG method cannot be improved in general.

It is also easy to recapture a result given in (Peterson, 1991a). Consider problem (9)–(10) with $\Omega = (0, \infty) \times (0, 1)$, boundary data $g(x) = x$, and the triangulation indicated in Fig. 3. Extend the domain and triangulation evenly to $(-\infty, \infty) \times (0, 1)$. The extended u^h will be unaffected if the vertical line $x = 0$ is removed from the triangulation. Thus the original problem on $(0, \infty) \times (0, 1)$ with smooth boundary data $g(x) = x$ is equivalent to a problem on $(-\infty, \infty) \times (0, 1)$ with a uniform mesh but with nonsmooth boundary data $g(x) = |x|$. Above we have seen that the pointwise rate of convergence for this latter problem is no better than $O(h^{1/2})$. We conclude that for an arbitrary quasi-uniform triangulation, the order of pointwise convergence may be limited to 1/2, *even if* the exact solution is smooth.

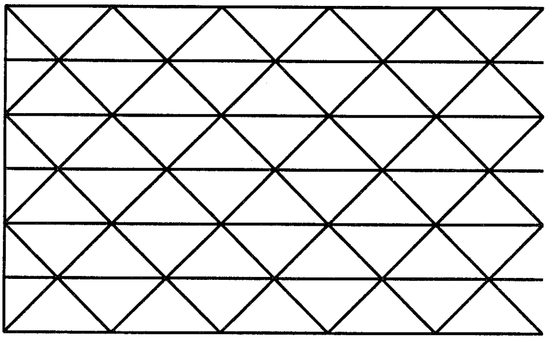

Fig. 3.

4. Discontinuous Galerkin — Numerical Results

In this section we summarize the results of various numerical computations involving the DG method with $p \geq 1$. Complete details may be found in (Peterson, 1991a) and (Peterson, 1991b).

To investigate the rate of convergence, we again use (9)–(10) as our test problem, now on the domain $\Omega = (0, 1) \times (0, 1)$. As the boundary data we take $g(x) = x^{p+1}$. When using a triangulation as indicated in Fig. 4, the results are as shown in Tab. 1. The pointwise rate of convergence is clearly only $p + 1/2$, even though the exact solution to the test problem is smooth $(u(x, y) = x^{p+1})$. Note that in this case the

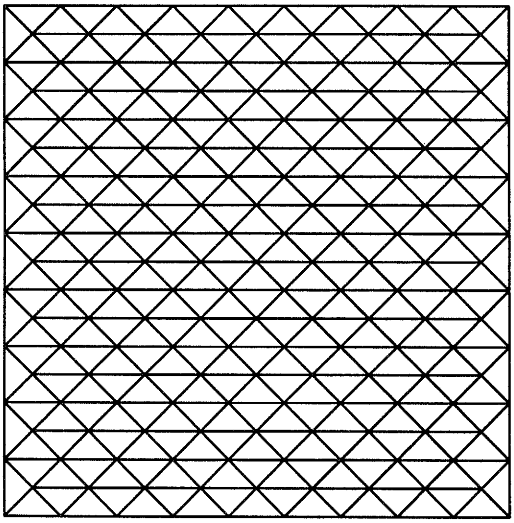

Fig. 4.

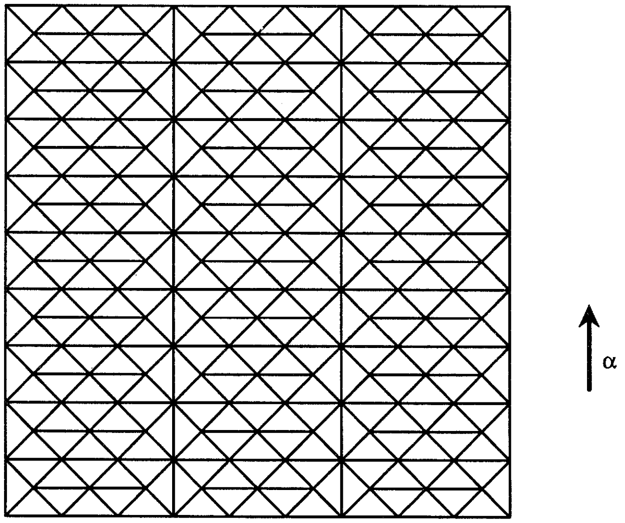

Fig. 5.

triangulation does not satisfy the non-alignment condition. In fact, the error is largest near the vertical lines $x = 0$ and $x = 1$. We can take advantage of this phenomenon to construct an example in which the $L_2(\Omega)$ error converges at a rate of only $p+1/2$, by introducing additional vertical lines into the triangulation, as indicated in Fig. 5. By letting the number of these lines vary with h as $h^{-3/4}$, we obtain the results shown in Tab. 2. These results show that the estimate (3) is sharp, even for smooth

Tab. 1. Estimated convergence rate for DG with alignment at $x = 0$ and $x = 1$, $p = 1$.

	L_2		L_∞	
$1/h$	error	rate	error	rate
626	0.3779e-6		0.9981e-5	
1251	0.9702e-7	1.96	0.3576e-5	1.48
2501	0.2497e-7	1.96	0.1276e-5	1.49
5001	0.6444e-8	1.95	0.4539e-6	1.49
10001	0.1668e-8	1.95	0.1612e-6	1.49

Tab. 2. Estimated convergence rate for DG with semi-frequent alignment, $p = 1$.

	L_2		L_∞	
$1/h$	error	rate	error	rate
535	0.2983e-5		0.1248e-4	
1086	0.1032e-5	1.50	0.4409e-5	1.47
2128	0.3754e-6	1.50	0.1601e-5	1.50
4608	0.1189e-6	1.49	0.5354e-6	1.42
8610	0.4632e-7	1.51	0.2009e-6	1.56

solutions. When the characteristic direction is slightly displaced from $(0, 1)$, and the analogous computations are performed on these same triangulations, optimal rates of convergence are observed in all cases. Thus the counter-intuitive assumption of non-alignment seems to be natural for this method. Notice however, that if the mesh were everywhere aligned with the characteristic direction, the results would be extremely good.

Given a triangulation free from the phenomenon seen above, there is still the question of how much regularity is required to obtain an optimal order estimate. Consider (9)–(10) on the trapezoidal domain and uniform triangulation shown in Fig. 6, with boundary data $g(x) = |x|^a$. Then $u \in H^r(\Omega)$ for $r < a + 1/2$. By varying the parameter a we vary the regularity of u, which may be plotted against the observed L_2 rate of convergence. Results of such computations are shown in Fig. 7. There are three things of note: to obtain optimal order $u \in H^{p+1}(\Omega)$ is not sufficient; the amount of regularity assumed in (4) appears to be more than necessary; and the amount of 'extra regularity' needed (vis-a-vis the estimate for an interpolant) appears to decrease as p increases.

The known crosswind spread estimate of $O(h^{1/2})$ has been shown to be sharp for piecewise constants. However, numerical results indicate that this improves to $O(h^{3/4})$ for $p = 1$. We now give the details of these results. Consider again $u_y = 0$ on the domain and triangulation shown in Fig. 6, now with the boundary condition $u(x, 0) = \text{sign}(x)$. At the outflow boundary $y = 1$ the number of elements to the right of $x = 0$ on which $|u^h - 1| > \epsilon$ is counted, for some fixed tolerance ϵ. This process is repeated for different h and from this an estimate of the asymptotic rate

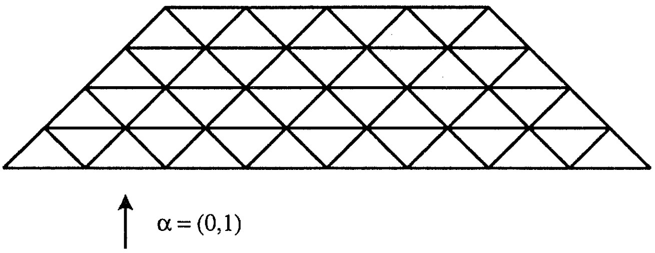

Fig. 6.

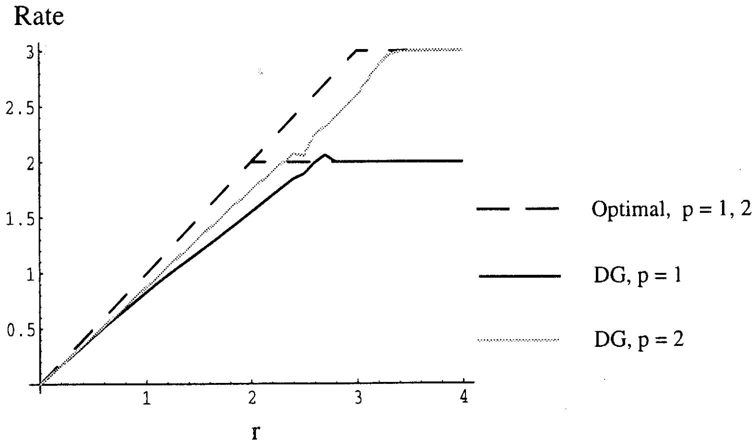

Fig. 7.

is made. The whole procedure is repeated for different values of ϵ to check for consistency. The results are shown in Tab. 3. For $p = 1$ the rate appears to be well above 1/2, and consistent with a conjecture of 3/4. The value 3/4 is also consistent with a study of the streamline diffusion method (Johnson *et al.*, 1987), a method whose analysis often parallels that of the DG method. For $p > 1$ the crosswind spread appears to become better still — we conjecture that it varies like $(2p+1)/(2p+2)$.

Tab. 3. Crosswind spread for DG ($\epsilon = 10^{-5}$).

	$p = 1$		$p = 2$		$p = 3$	
$1/h$	width	rate	width	rate	width	rate
1250	.0164		.0092		.0068	
2500	.0098	0.74	.0052	0.82	.0392	0.79
5000	.0059	0.73	.0029	0.84	.00216	0.86
10000	.00345	0.77	.00165	0.81	.00118	0.87

5. Continuous Galerkin

For the CG method, the first non-trivial case is $p = 2$, since when $p = 1$ Type 2 elements have no degrees of freedom. Unfortunately, even the $p = 2$ case is sufficiently complex to make the type of direct analysis of Section 3 intractable. Thus for the CG method we present only numerical results. We will investigate three aspects of the behaviour of the method when $p \geq 2$. We continue to use (9)–(10) as our test problem.

We first consider the role of the non-alignment condition. Recall that the estimate (6) was derived under this condition. With $\Omega = (0,1) \times (0,1)$, $g(x) = x^{p+1}$, and the triangulation indicated in Fig. 5, with the number of strips varying proportionally to $h^{-3/4}$ for $p = 2$ and proportionally to $h^{-9/10}$ for $p = 3$, the results are as given in Tables 4 and 5. Notice that the non-alignment condition is violated in this example. Nonetheless, the rate of convergence is about $O(h^{p+1/4})$ for $p = 2$, and about $O(h^{p+1/3})$ for $p = 3$. We conclude that the non-alignment condition may not be necessary for (6) to hold, that (6) may be improvable for $p > 2$, and that to guarantee an optimal rate of convergence the non-alignment condition (or some other condition on the triangulation) will be necessary.

To isolate the effect of regularity, we use the domain and triangulation shown in Fig. 6, with $g(x) = |x|^a$. Plots of the convergence rate against the regularity of u are shown in Fig. 8. We see that for $p = 2$, the estimate (6) is sharp, in that for $u \in H^3(\Omega)$ we observe only the $O(h^{p+1/4})$ convergence. This agrees with results reported in (Falk and Richter, 1987). For $p = 3$ the rate is about $O(h^{p+1/3})$, however, again suggesting that (6) may be improvable for $p > 2$. And we also see that an optimal rate is obtainable only by assuming additional regularity, and that the amount of additional regularity appears to decrease as p increases.

Tab. 4. Estimated convergence rate for CG with semi-frequent alignment, $p = 2$.

	L_2		L_∞	
$1/h$	error	rate	error	rate
48	1.889e-06		9.009e-06	
96	4.270e-07	2.15	1.842e-06	2.29
200	7.959e-08	2.29	3.500e-07	2.26
391	1.777e-08	2.24	7.686e-08	2.26
780	3.838e-09	2.22	1.612e-08	2.26
600	7.498e-10	2.27	3.190e-09	2.26
3145	1.656e-10	2.24	6.937e-10	2.26
6292	3.475e-11	2.25	1.452e-10	2.26
12720	7.089e-12	2.26	2.971e-11	2.25

Tab. 5. Estimated convergence rate for CG with semi-frequent alignment, $p = 3$.

	L_2		L_∞	
$1/h$	error	rate	error	rate
48	3.635e-09		1.560e-08	
96	3.875e-10	3.23	1.567e-09	3.32
184	4.819e-11	3.20	1.847e-10	3.29
387	4.155e-12	3.30	1.568e-11	3.32
729	5.342e-13	3.24	1.888e-12	3.34
1530	4.541e-14	3.33	1.559e-13	3.36
3135	4.176e-15	3.33	1.462e-14	3.30
6384	3.934e-16	3.32	1.367e-15	3.33
11928	5.140e-17	3.26	1.679e-16	3.35

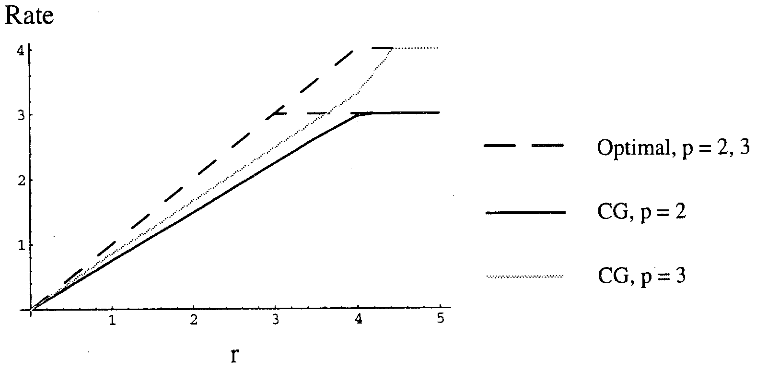

Fig. 8.

Crosswind spread for the CG method was investigated using the same methodology described above for the DG method. The numerical results are reported in Tab. 6. Notice that the rate increases with increasing p, and that the known estimate of $O(h^{1/2})$ does not appear to be sharp. If we compare to the DG method, for which the computational experiments above suggest rates of $O(h^{5/6})$ for $p = 2$ and $O(h^{7/8})$ for $p = 3$, we see that the continuous method is not quite as good. However, when consideration is given for the amount of computation required by the two methods, the differences are mitigated.

6. Streamline Diffusion

In this section we briefly consider the streamline diffusion (SD) method for the model problem. Unlike DG and CG, this is an implicit method, requiring the solution of a single large linear system to obtain the approximation; however the analysis of SD has often been very similar to that of DG, and so it seems appropriate to include it here.

Tab. 6. Crosswind spread for CG ($\epsilon = 10^{-6}$).

$1/h$	$p=2$	$p=3$	$p=4$
10	0.600000	0.416667	0.550000
20	0.375000	0.275000	0.312500
40	0.250000	0.170833	0.168750
80	0.156250	0.097917	0.090625
160	0.093750	0.059375	0.048438
320	0.056250	0.034896	0.025781
640	0.035938	0.019531	0.015234
1280	0.021875	0.011849	0.008398
Rate	0.71	0.77	0.88

The SD approximation u^h of problem (1)–(2) is that element of S_h, the set of all continuous piecewise polynomial of degree p, which satisfies

$$\int_\Omega (\alpha \cdot \nabla u^h + \beta u^h - f)(v^h + h\alpha \cdot \nabla v^h) + \int_{\partial\Omega_-} (u^h - g)v^h(\alpha \cdot n) = 0 \qquad \forall v^h \in S_h$$

The known general $L_2(\Omega)$ estimate for the SD method shows the same 1/2 order loss with respect to an optimal estimate as seen in (3). See (Johnson *et al.*, 1984) and the references contained therein for details.

For computational convenience, we investigate the optimality of the SD method with respect to the $L_\infty(\Omega)$ norm. We do this by again using the test problem $u_y = 0$, on the domain $(0,1) \times (0,1)$ with the triangulation shown in Fig. 4, $p = 1$, and boundary condition $u(x,0) = x^2$. The results are given in Tab. 7, and clearly show non-optimal convergence. Note that the non-alignment condition is violated in this example, suggesting that the non-alignment condition may also be relevant to the SD method.

Tab. 7. Estimated convergence rate for SD with alignment at $x = 0$ and $x = 1$, $p = 1$.

	L_2		L_∞	
$1/h$	error	rate	error	rate
45	2.7053e-5		7.2372e-4	
60	1.5678e-5	1.90	4.7737e-4	1.45
75	1.0262e-5	1.90	3.4495e-4	1.46
90	7.2568e-6	1.90	2.6421e-4	1.46

7. Summary

For many elliptic problems, the standard finite element method admits optimal *a priori* error estimates of the form

$$||u^h - u||_\Omega \leq Ch^{p+1}||u||_{p+1,\Omega}$$

This estimate is optimal in the sense that the power of h cannot be increased, nor can the index of the norm on the right be decreased. Error estimates for interpolants constructed directly from u are not better. In contrast, what we have seen above is that for even the very simple hyperbolic problem (1)–(2), the finite element methods we have considered show non-optimal behaviour in several respects. In addition, it appears that in some aspects the extent of this non-optimal behaviour varies with the polynomial degree p, a phenomenon without analogue in elliptic theory.

We have shown that the estimate (3) for the discontinuous Galerkin method is sharp. In particular, a rate of convergence higher than $p + 1/2$ can only be guaranteed by assuming more regularity of the exact solution *and* making some assumption on the triangulation beyond just quasi-uniformity. Our conjecture is that the triangulation should either satisfy the non-alignment condition, *or* possess some measure of uniformity. The existing general pointwise estimate of order $p + 1/4$ is unlikely to be sharp, and could probably be improved to $p + 1/2$, although not beyond that without the same sort of additional assumptions. And while we have shown that the existing order $h^{1/2}$ estimate of crosswind spread is sharp in general, we conjecture it is *not* sharp for $p \geq 1$.

For the continuous Galerkin method, the estimate (6) is sharp for $p = 2$, although the non-alignment assumption may not be necessary for this estimate, and the estimate may be improvable for $p > 2$. To obtain optimal order estimates, however, as with the DG method it appears that additional regularity will be required, as well as an assumption on the triangulation. The known crosswind spread estimate does not appear to be sharp even for $p = 2$.

Despite the simplicity of the model problem, the exact behaviour of these methods eludes analysis to date. In particular there is yet to be given sharp pointwise estimates, or minimal conditions which guarantee optimal order convergence. Of particular curiosity is that fact that the counter-intuitive non-alignment assumption seems to be natural for these methods.

References

Cai D. and Falk R.S. (1994): *Reduced continuity finite element methods for first order scalar hyperbolic equations.* — RAIRO Math. Mod. and Num. Anal., v.28, No.6, pp.667–698.

Falk R.S. and Richter G.R. (1987): *Analysis of a continuous finite element method for hyperbolic equations.* — SIAM J. Numer. Anal., v.24, No.2, pp.257–278.

Falk R.S. and Richter G.R. (1992): *Local error estimates for a finite element method for hyperbolic and convection-diffusion equations.*— SIAM J. Numer. Anal., v.29, No.3, pp.730–754.

Johnson C., Nävert U. and Pitkäranta J. (1984): *Finite element methods for linear hyperbolic problems.* — Comput. Methods Appl. Mech. Engrg., v.45, No.1–3, pp.285–312.

Johnson C. and Pitkäranta J. (1986): *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation.* — Math. Comp.,v.46, No.173, pp.1–26.

Johnson C., Schatz A. and Wahlbin L.(1987): *Crosswind smear and pointwise errors in streamline diffusion finite element methods.* — Math. Comp., v.49, No.179, pp.25–38.

Lesaint P. and Raviart P. (1974): *On a finite element method for solving the neutron transport equation,* In: Mathematical Aspects of Finite Elements in Partial Differential Equations (Carl de Boor, Ed.). — New York: Academic Press.

Niijima K. (1990): *Pointwise error estimates for a streamline diffusion finite element scheme.* — Numer. Math., v.56, No.7, pp.707–719.

Peterson T.E. (1990): *Convergence properties of the discontinuous Galerkin method for a scalar hyperbolic equation.* — Ph.D. thesis, Department of Mathematics, Cornell University, Ithaca, New York.

Peterson T.E. (1991a): *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation.* — SIAM J. Numer. Anal.,v.28, No.1, pp.133–140.

Peterson T.E. (1991b): *Towards understanding the discontinuous Galerkin method.* — Research Report CMA-MR02-91, Centre for Mathematics and its Applications, Australian National University, Canberra, Australia.

Richter G. (1988): *An optimal order estimate for the discontinuous Galerkin method.* — Math. Comp., v.50, No.181, pp.75–88.

Reed W. and Hill T. (1973): *Triangular Mesh Methods for the Neutron Transport Equation.* — Report LA-UR-73-479, Los Alamos Scientific Laboratory.

Shuster D.B. (1994): *A numerical investigation of a continuous explicit finite element method for first order scalar hyperbolic equations.* — Report No. RM-94-01, Department of Applied Mathematics, University of Virginia, Charlottesville, Virginia.

Appendix

We are interested in the quantity

$$S_j \equiv \frac{1}{2^j} \sum_{k=j/2-\sqrt{j}/2}^{j/2-\sqrt{j}/4} \binom{j}{k} \geq \frac{1}{2^j} \frac{\sqrt{j}}{4} \binom{j}{\frac{j-\sqrt{j}}{2}}$$

By Stirling's formula $\lim_{\infty \longrightarrow a} \sqrt{2\pi} a^{a+1/2} e^{-a}/a! = 1$,

$$\binom{a}{b} = \frac{a!}{b!(a-b)!} \geq c \frac{a^{a+1/2}}{b^{b+1/2}(a-b)^{a-b+1/2}}$$

Thus

$$\begin{aligned}\binom{j}{\frac{j-\sqrt{j}}{2}} &\geq c\frac{j^{j+1/2}}{\left(\frac{j-\sqrt{j}}{2}\right)^{\frac{j-\sqrt{j}+1}{2}}\left(\frac{j+\sqrt{j}}{2}\right)^{\frac{j+\sqrt{j}+1}{2}}}\\ &= c2^{j+1}j^{-1/2}\left(\frac{j}{j-\sqrt{j}}\right)^{\frac{j-\sqrt{j}+1}{2}}\left(\frac{j}{j+\sqrt{j}}\right)^{\frac{j+\sqrt{j}+1}{2}}\end{aligned}$$

and so

$$\begin{aligned}S_j &\geq \frac{1}{2}c\left(\frac{j}{j-\sqrt{j}}\right)^{\frac{j-\sqrt{j}+1}{2}}\left(\frac{j}{j+\sqrt{j}}\right)^{\frac{j+\sqrt{j}+1}{2}}\\ &= \frac{1}{2}c\left(1+\frac{1}{j-1}\right)^{\frac{j-\sqrt{j}+1}{2}}\left(1-\frac{1}{\sqrt{j}+1}\right)^{\sqrt{j}}\end{aligned}$$

The two factors above have limits as $j \rightarrow \infty$ of $e^{1/2}$ and e^{-1}, respectively. It follows that $S_j \geq c$ for some positive constant c which is independent of j.

Received: April 4, 1995
Revised: July 10, 1995