A Positivity Preserving Finite Element Method for Hyperbolic Partial Differential Equations

Matthew Hubbard and Martin Berzins

School of Computing, University of Leeds, Leeds, LS2 9JT, U.K.

Abstract. This paper describes a framework in which flux and slope limiters, commonly used in the finite volume community, can be applied in the context of finite element methods, both to the spatial discretisation and the mass matrix. This gives a nonlinear finite element method which uses different basis functions for discretisation of the time and space derivatives and is inherently positivity preserving for hyperbolic partial differential equations. The procedure can be carried out on irregular triangular meshes and is applied here to the two-dimensional scalar advection equation. A number of alternative methods are possible, but the end results do not differ enormously and one representative scheme is picked out to be compared with other schemes for a simple test case involving constant advection at an oblique angle to the mesh. Other cases, not discussed here, show similar qualities.

1 Introduction

The aim of this paper is to suggest links between limiting techniques, which are used as a matter of course in the finite volume community for the modelling of hyperbolic partial differential equations, and finite element methods, which are notoriously poor at approximating problems of this type, particularly in more than one space dimension. Limiters are used to combine a low order but positive (and usually upwind) scheme with a high order scheme to create an accurate method which avoids spurious, numerically induced, oscillations. They are an important component of many of today's most successful finite volume methods but have yet to be employed in the traditional finite element framework: it has proved difficult to construct upwind and positive finite element methods, although SUPG schemes have been a partial success, and the more recent developments in Discontinuous Galerkin and Fluctuation Splitting methods have illustrated the strength of the relationship between finite volume and finite element methods.

This work essentially employs a cell vertex finite volume approach, but considers how it might be recast as a mass-lumped finite element scheme with nonlinear basis functions. It then considers how a mass matrix might be incorporated, what it might look like, and how it might be modified to give a positive scheme when it is inverted. This follows the approach of Cardle [5] in which the basis functions are modified differently for the space and time derivatives, and extends earlier, one-dimensional work of Berzins [2]. Results are presented for a standard test case to illustrate the accuracy of the new method.

2 A Mass-Lumped Scheme

The system which will be studied here is the two-dimensional scalar advection equation, given by

$$u_t + \boldsymbol{\lambda} \cdot \boldsymbol{\nabla} u \;=\; 0\,,$$

approximated on triangular meshes. This represents the transport of a quantity u with velocity $\boldsymbol{\lambda}$. Applying the mass-lumped, linear Galerkin finite element method to this equation leads, after some simple algebraic manipulation, to an edge-based form of the scheme, given by

$$\frac{\Omega_i}{3}\,\dot{u}_i \;=\; \sum_{j\in\cup\triangle_i} \frac{1}{6}\,(\boldsymbol{\lambda}_{ij}\cdot\boldsymbol{n}^i_{j-1,j+1})(u_j+u_i)\,. \qquad (1)$$

Here the sum is over the nodes j adjacent to node i, Ω_i is the area of the patch of cells surrounding node i and $\boldsymbol{n}^i_{j-1,j+1}$ are the 'inward' pointing normals (see Fig. 1) scaled by the distance between nodes $j-1$ and $j+1$ (j is incremented and decremented modulo N_i, the number of nodes adjacent to node i, and increases anticlockwise).

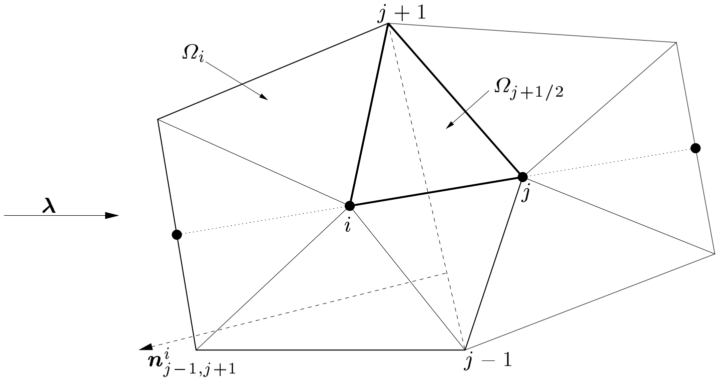

Fig. 1. The geometry associated with the edge adjoining nodes i and j.

This scheme is unconditionally unstable, but can be adjusted to introduce an upwind bias along the grid edges. A diffusive component is added (the difference between the contribution of an edge to the central scheme (1) and to a purely upwind scheme), and a limiter, denoted here by $V(\cdot)$, applied to give

$$\frac{\Omega_i}{3}\,\dot{u}_i \;=\; \sum_{j\in\cup\triangle_i} \frac{1}{6}\,(\boldsymbol{\lambda}_{ij}\cdot\boldsymbol{n}^i_{j-1,j+1})\left[2u_{\mathrm{m}} + \frac{V(r_j)}{r_j}(u_{\mathrm{d}}-u_{\mathrm{m}})\right]\,, \qquad (2)$$

in which

$$r_j \;=\; \left[\frac{u_{\mathrm{d}}-u_{\mathrm{m}}}{u_{\mathrm{m}}-u_{\mathrm{u}}}\times\frac{d_{\mathrm{u}}}{d_{\mathrm{d}}}\right]_j\,.$$

The value u_d is taken from the downwind vertex of edge ij (node i if $\boldsymbol{\lambda}_{ij} \cdot \boldsymbol{n}^i_{j-1,j+1} \geq 0$, node j otherwise), u_m is taken from the upwind vertex of edge ij, and u_u is taken from the intersection of the extension of edge ij beyond the upwind vertex with the opposite edge of the triangle into which it extends (see Fig. 1). Linear interpolation is used to evaluate this value, which is consistent with the Galerkin method originally considered and guarantees that the value of u_u remains bounded by the local solution values. $\boldsymbol{\lambda}_{ij}$ is evaluated at the midpoint of edge ij. d_d is invariably the length of edge ij, while d_u is the distance between the upwind vertex and the intersection point with the opposite edge.

It can be shown that, as long as $0 \leq V(r)/r \leq 2$ and $0 \leq V(r) \leq 2$, the scheme (2), combined with forward Euler time-stepping, is positive (and hence stable) for δt satisfying

$$\frac{\delta t}{\Omega_i} \sum_{j \in \cup\triangle_i} \left[1 + \frac{d_\mathrm{d}}{d_\mathrm{u}}\right] |\alpha^i_j| \;\leq\; 1$$

for all nodes i, where $\alpha^i_j = \boldsymbol{\lambda}_{ij} \cdot \boldsymbol{n}^i_{j-1,j+1}$. In this work, two limiters have been used, both of which satisfy the above conditions:

- $V(r) = \max(0, \min(2r, \min(0.25 + 0.75r, 4)))$, a third order limiter derived by Gaskell and Lau [6].
- $V(r) = (r + |r|)/(1 + \max(1, |r|))$, a modified form of van Leer's limiter [4].

The relationship between the scheme (2) and finite elements is not immediately obvious. However, it can be replicated by augmenting the standard linear basis function $\phi_i(x, y)$, used in the Galerkin approach, with a nonlinear term, based around the local grid edges:

$$\tilde{\phi}_i(x, y) \;=\; \phi_i(x, y) + 4 \sum_{j \in \cup\triangle_i} \mathrm{sgn}(\alpha^i_j) \left[1 - \frac{V(r_j)}{r_j}\right] \phi_i(x, y)\phi_j(x, y)\,.$$

3 Including a Mass Matrix

The consistent linear Galerkin scheme on triangles gives a system of ordinary differential equations defined by

$$\sum_{j \in \cup\triangle_i} \frac{\Omega_{j+1/2}}{12} [2\dot{u}_i + \dot{u}_j + \dot{u}_{j+1}] \;=\; G_i(\underline{u}) \qquad (3)$$

where G_i is simply the right hand side obtained from the chosen spatial discretisation, $\underline{u}$ is the vector of all solution values at the given time level and $\Omega_{j+1/2}$ is the area of the triangle with vertices i, j, $j+1$ (see Fig. 1).

The lumped scheme is positive, but this property will normally be lost when the mass matrix, introduced in (3), is inverted. The aim here is to manipulate the mass matrix in a manner which will retain the positivity property of the scheme. The approach is similar to that used in spatial limiting: the lumped matrix is taken as a starting point, and to this is added a component proportional to the difference between the lumped matrix and the consistent matrix derived from linear basis functions. It follows the philosophy of Cardle [5], in that the time and space derivatives are treated independently.

A variety of schemes of this form have been implemented [2,3] and another alternative is presented here. In fact, most of the procedures produce similar results: there is far more sensitivity to the discretisation of the spatial terms. Forward Euler time-stepping is used in all the schemes, but a cell-based 'limiting' is chosen here rather than an edge-based one. As in [3], the time derivative is discretised first, giving

$$\begin{aligned}\sum_{j\in\cup\triangle_i} & \frac{\Omega_{j+1/2}}{12}(2u_i^{n+1}+u_j^{n+1}+u_{j+1}^{n+1}) \\ &= \delta t\, G_i(\underline{u}^n)+\sum_{j\in\cup\triangle_i}\frac{\Omega_{j+1/2}}{12}(2u_i^n+u_j^n+u_{j+1}^n)\,, \qquad (4)\end{aligned}$$

in which n indexes the time level. The mass matrix can be modified *before* discretising the time derivative, and a nonlinear method constructed based on limiting ratios of differences in time derivatives so that the modified mass matrix is an M-matrix (whose inverse contains only non-negative entries) [2]. However, this only enforces positivity of $\dot{u}$. It doesn't automatically impose positivity on the solution.

Equation (4) can be rewritten as

$$\begin{aligned}\sum_{j\in\cup\triangle_i} & \frac{\Omega_{j+1/2}}{12}\left[4u_i^{n+1}+(u_j^{n+1}-u_i^{n+1})+(u_{j+1}^{n+1}-u_i^{n+1})\right] \\ &= \delta t\, G_i(\underline{u}^n)+\sum_{j\in\cup\triangle_i}\frac{\Omega_{j+1/2}}{12}\left[4u_i^n+(u_j^n-u_i^n)+(u_{j+1}^n-u_i^n)\right]\end{aligned}$$

Cumulatively, the latter terms on each side give the difference between lumped and consistent Galerkin schemes. These differences are limited in a manner which guarantees that the matrix on the left hand side is diagonally dominant with non-positive off-diagonal entries (and hence an M-matrix), so its inversion will retain the positivity of the scheme. The resulting method takes the form

$$\begin{aligned}\sum_{j\in\cup\triangle_i} & \frac{\Omega_{j+1/2}}{12}\Big[4u_i^{n+1}+\max(r_{j+1/2}^{n+1}-1,0)(u_i^{n+1}-u_{j+1}^{n+1}) \\ &\qquad +\max(s_{j+1/2}^{n+1}-1,0)(u_i^{n+1}-u_j^{n+1})\Big] \\ &= \delta t\, G_i(\underline{u}^n)+\sum_{j\in\cup\triangle_i}\frac{\Omega_{j+1/2}}{12}\Big[4u_i^n+\max(r_{j+1/2}^n-1,0)(u_i^n-u_{j+1}^n) \\ &\qquad +\max(s_{j+1/2}^n-1,0)(u_i^n-u_j^n)\Big] \qquad (5)\end{aligned}$$

in which

$$r_{j+1/2} \;=\; \frac{u_j - u_i}{u_i - u_{j+1}} \qquad \text{and} \qquad s_{j+1/2} \;=\; \frac{1}{r_{j+1/2}}$$

The right hand side of (5) can be shown to be a positive combination of local solution values for an appropriate limit on the time-step, so the overall scheme should be positive.

Note that, throughout this discussion, j and $j+1$ refer to consecutive nodes around a patch of cells centred on node i. In [3] this cell-based pairing of edges is replaced by associating each edge with its projection across the patch, in the manner of the above spatial discretisation.

Equations (5) are solved iteratively (indexed below by m), using a Jacobi-type method to give

$$\begin{aligned}
&\left(\sum_{j\in\cup\triangle_i}\Big[\max(r^{m+1}_{j+1/2}-1,0)+\max(s^{m+1}_{j+1/2}-1,0)\Big]\right)u_i^{m+1}\\
&= \frac{3\delta t}{\Omega_i}\,G_i(\underline{u}^n)\\
&+\sum_{j\in\cup\triangle_i}\Big[\max(r^{n}_{j+1/2}-1,0)(u_i^n-u_{j+1}^n)+\max(s^{n}_{j+1/2}-1,0)(u_i^n-u_j^n)\Big]\\
&+\sum_{j\in\cup\triangle_i}\Big[\max(r^{m}_{j+1/2}-1,0)u_{j+1}^m)+\max(s^{m}_{j+1/2}-1,0)u_j^m)\Big]\;.
\end{aligned}$$

The initial estimate at time level n is taken to be $\underline{u}^n$ and, in the following results, ten iterations were always enough to reach convergence. In this work, r and s have been replaced by $V(r)$ and $V(s)$ (using the modified van Leer limiter) because this improves the accuracy of the scheme and improves the restriction on the time-step. The positivity property remains intact.

4 Results

The test case shown here represents the doubly periodic advection of a double sine wave profile, given initially by

$$u(x,y,0) \;=\; \sin(2\pi x)\,\sin(2\pi y)\,,$$

with a constant velocity of $\boldsymbol{\lambda}=(1,2)^{\mathrm{T}}$, across the domain $[0,1]\times[0,1]$. All of the results are obtained on uniform triangular grids consisting of squares divided by diagonals from bottom left to top right corners. This means that the advection velocity is *not* aligned with grid edges.

The results indicate that the new scheme matches the performance of a good cell centred scheme (the MLG scheme of Batten *et al.* [1]) on a similar type of mesh, having smaller L_∞ error but larger L_1 error, but still lags behind a cell vertex Lax-Wendroff scheme to which Flux-Corrected Transport (FCT) has been applied [7]. This is also seen in other cases approximated.

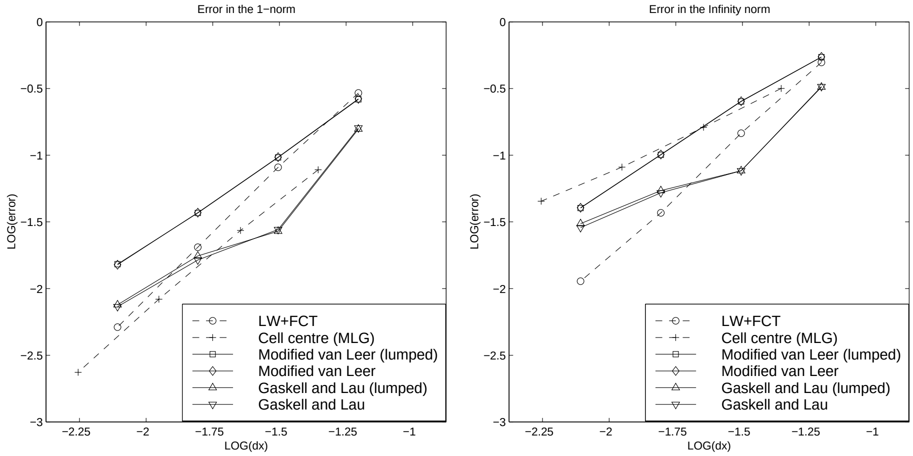

Fig. 2. Approximations of the L_1 and L_∞ norms of the errors for the double sine wave test case at $t = 1$.

5 Conclusions

The method presented above is still in the very early stages of development. The 'lumped' method is very accurate, particularly when the third order limiter is used, but this is unsurprising since it is effectively a flux limited finite volume method applied on the dual of the triangular grid: the novelty being the method used to calculate the ratios to be limited. As presented it is only first order accurate in time. The consistent finite element method is, as yet, less successful. This may be because the Galerkin method is used as the basis, and the nonlinear approach, while guaranteeing positivity, appears to add a diffusive term to the lumped scheme. It would be more interesting to apply it to other methods such as Taylor-Galerkin, or a fluctuation splitting method which have more leeway for improvement. At the moment, the more restrictive the limiting is on the mass matrix, the better the solution.

References

1. P. Batten, C. Lambert, D.M. Causon: Int. J. Numer. Methods Eng. **39**, 1821 (1996).
2. M. Berzins: Com. Num. Meth. Eng. **17**, 659 (2001)
3. M. Berzins, M.E. Hubbard: submitted to J. Comput. Phys. (2001)
4. M. Berzins, J.M. Ware: Appl. Numer. Math. **16**, 417 (1995)
5. J.A. Cardle: Int. J. Numer. Methods Eng. **38**, 171 (1995)
6. P.H. Gaskell, A.C. Lau: Int. J. Numer. Methods Eng. **8**, 617 (1988)
7. M.E. Hubbard, P.L. Roe: Int. J. Numer. Methods Fluids **33**, 711 (2000)