Variable-Order Finite Elements and Positivity Preservation for Hyperbolic PDEs

M. Berzins

Computational PDEs Unit, School of Computing, University of Leeds, Leeds LS2 9JT, UK

Abstract

A family of positivity-preserving finite element methods are considered for the solution of the advection equation in one space dimension. The approach uses B-spline spatial discretization methods in conjunction with Forward Euler timestepping and a mass matrix iteration that preserves positivity. The method is compared against a positivity-preserving finite volume scheme on travelling wave and pulse examples including an inviscid Burgers equation example.

1 Introduction

There are many situations in the numerical solution of partial differential equations (p.d.e.s) in which the computed solution values should, on physical grounds, remain non-negative. One of the simplest examples is that of the simple advection equation with non-negative initial data while other cases are those of concentrations of chemical compounds in reacting flow calculations. In both cases preserving positivity is essential to avoid the numerical calculation becoming meaningless. Consider the solution of the advection equation as given by $\frac{\partial U}{\partial t} + \frac{\partial U}{\partial x} = 0$ with appropriate initial and boundary conditions on a spatial interval $[x_l, x_r]$. Applying the standard Galerkin method with linear basis (hat) functions $\phi_i(x)$ on a uniformly spaced mesh $x_i, i = 1, ..., N$ gives

$$\int_{x_{i-1}}^{x_{i+1}} \frac{\partial U}{\partial t}\phi_i(x)\, dx = \int_{x_{i-1}}^{x_{i+1}} -\frac{\partial U}{\partial x}\phi_i(x)\, dx, i = 1, ..., N, \quad (1)$$

Email address: martin@comp.leeds.ac.uk. (M. Berzins).

where the approximate finite element solution to this p.d.e. is defined by $U(x,t) = \sum_{i=1}^{N} \phi_i(x) U_i(t)$ where $\phi_i(x_j) = \delta_{ij}$. Evaluating the integrals gives rise to the numerical scheme defined by

$$\frac{1}{6}\left[\dot{U}_{i-1} + 4\dot{U}_i + \dot{U}_{i+1}\right] = \frac{-1}{2\delta x}(U_{i+1} - U_{i-1}) \tag{2}$$

where δx is the uniform mesh spacing in this case and where $\dot{U}_i = \frac{dU_i}{dt}$. Defining the time-dependent vector $\underline{U}$ by $\underline{U} = [U_1, ..., U_N]^T$ allows this system of equations to be rewritten in the form

$$A\dot{\underline{U}}(t) = \underline{F}(\underline{U}(t)) \tag{3}$$

where the matrix A is referred to as the mass matrix.

An alternative to linear basis functions is to use the standard quadratic basis functions as used for hyperbolic equations by Gresho and Sani [11] and by Griffiths [12]. Christie and Mitchell also use cubic polynomials, [7].

It is well-known that the standard Galerkin method is unsatisfactory for hyperbolic equations in a very similar way to that of linear central difference schemes, [19]. Many modified Galerkin methods have been proposed to remedy this situation. A survey of such methods is given in [19] and includes Streamline Upwind Petrov-Galerkin (SUPG) methods in which the test functions are modified to improve the behaviour of the method and Discontinuous Galerkin (DG) methods [9] in which discontinuous basis functions are used. In the simplest case the piecewise constant DG method for the advection equation with positive velocity equal to one is given in a semi-discrete form by

$$\dot{U}_i = \frac{-1}{\delta x}(U_i - U_{i-1}). \tag{4}$$

In this case the method is identical to first-order upwind differencing and so is overly diffusive. There are many other approaches such as the modified Petrov-Galerkin method of Cardle [5] in which the test function is modified differently for the spatial and temporal terms. In this case the numerical scheme that results is given by

$$\dot{U}_i + \frac{1-\beta}{6}\left[\dot{U}_{i-1} - 2\dot{U}_i + \dot{U}_{i+1}\right] = \frac{-(U_{i+1} - U_{i-1})}{2\delta x} + \frac{\alpha}{2\delta x}\delta^2 U_i^n \tag{5}$$

where $\delta^2 U_i^n = U_{i+1}^n - 2U_i^n + U_{i-1}^n$, β and α are the constants multiplying the Petrov-Galerkin additional polynomials in time (cubic polynomial) and space (quadratic polynomial), see [5].

In the case of many of these methods the magnitude of unphysical values may be controlled by a careful choice of problem dependent tuning parameters so as to be not as large as in the case of the standard Galerkin method.

Some recent papers addressing positivity preservation are those of Sheu et al. [20], Baker et al. [1] and MacKinnon and Carey [18]. The approach adopted here differs from all of these methods, and is perhaps closer to the method of Cockburn and Shu [8]. The definition used here for a positivity-preserving scheme for the advection equation is one (see [16]) for which the numerical solution at time t_{n+1} may be written in terms of the numerical solution at time t_n in the form

$$U_i(t_{n+1}) = \sum_j a_j U_j(t_n) \text{ where } \quad \sum_j a_j = 1, \quad \text{and } a_j \geq 0 \ . \tag{6}$$

The key observation with regard to preserving positivity is due to Godunov [10] who proved that any scheme of better than first order which preserves positivity for the advection equation must be nonlinear. For example, the coefficients a_j in (6) above must depend on the numerical solution to the p.d.e. This means that α and β in (5) must also depend on the solution. There are two main steps needed to derive positive finite element schemes for hyperbolic equations. The first step is to have a positive scheme for the discretisation of the space derivative term. The second step is to have an update formula at the next time level that preserves positivity. In this latter case it is necessary to consider the effect of the presence of the mass matrix. Although one possible approach is to lump the mass matrix the approach taken here will follow the general approach taken by Berzins [3] in maintaining an approximation to the mass matrix in which a solution-based switch is used to modify the form of the matrix. The approach used here will build on this but differ from it in that the modifications to preserve positivity are only applied after Forward Euler discretization in time is used. The other substantial difference from earlier work is that a family of positive discretization schemes based on variable-order B-splines is derived, although the ideas apply equally well to standard polynomial-based Galerkin methods.

1.1 A simple variable order finite difference scheme

The simplest approach for deriving positivity-preserving schemes goes back for steady state problems at least as far as Harten and Zwas [13] and is discussed in Chapter 22 of [16]. The idea is simply to use a scheme only when it preserves positivity and otherwise to switch to a more suitable scheme.

Consider the standard central-difference second-order spatial discretisation

scheme applied to the advection equation:

$$\dot{U}_i = \frac{-1}{2\delta x}(U_{i+1} - U_{i-1}). \tag{7}$$

Applying Forward Euler timestepping and rearranging the equation gives

$$U_i^{n+1} = U_i^n - \frac{\delta t}{\delta x}(U_i^n - U_{i-1}^n) - \frac{\delta t}{2\delta x}\left[U_{i-1}^n - 2U_i^n + U_{i+1}^n\right] \tag{8}$$

where U_k^n denotes the value at mesh point k at time t_n. This scheme, in an unmodified form, is generally regarded as unsuitable for hyperbolic equations. Noting that

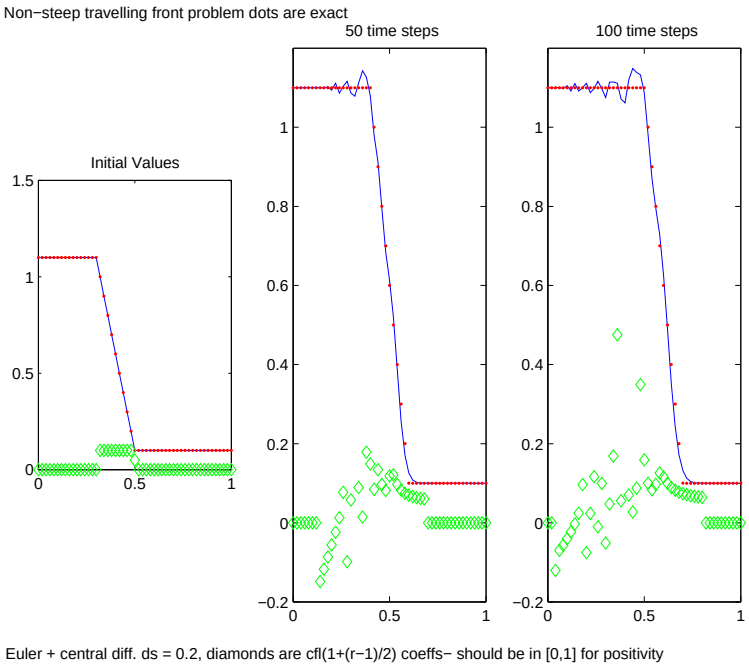

Fig. 1. Advection of Non steep profile using unmodified central differences

$$\left[U_{i+1}^n - 2U_i^n + U_{i-1}^n\right] = (r_i^n - 1)(U_i^n - U_{i-1}^n) \tag{9}$$

$$= (1/r_i^n - 1)(U_i^n - U_{i+1}^n) \tag{10}$$

where $r_i^n = (U_{i+1}^n - U_i^n)/(U_i^n - U_{i-1}^n)$ allows equation (8) to be rewritten as:

$$U_i^{n+1} = U_i^n \left[1 - \frac{\delta t}{\delta x}(1 + \frac{(r_i^n - 1)}{2})\right] + U_{i-1}^n \left[\frac{\delta t}{\delta x}(1 + \frac{(r_i^n - 1)}{2})\right] \tag{11}$$

Considering the spatial derivative on its own and imposing the condition that

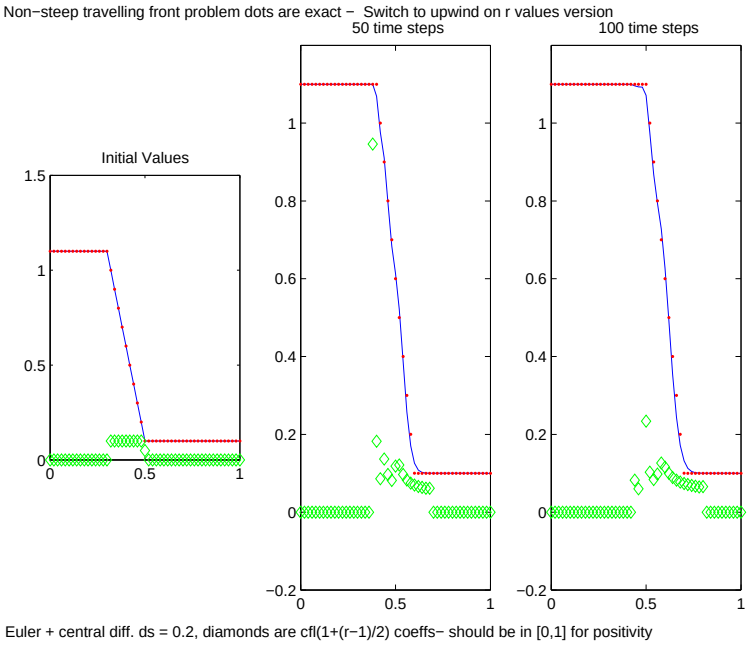

Fig. 2. Advection of Non steep profile using central/upwind difference switching

the method is positive (and also upwind range preserving see Laney [16]) gives the requirement:

$$0 \leq \frac{\delta t}{\delta x}\left[1 + \frac{(r_i^n - 1)}{2}\right] \leq 1 \tag{12}$$

Figure 1 shows the values of this coefficient when the scheme of equation (11) is applied to the advection of a modest gradient. The figure shows that the coefficient rapidly exceeds the range $[0, 1]$ required for positivity. Figure 2 shows what happens when a switch is inserted so that when the condition specified by equation (12) is violated then the first order upwind scheme defined by equation (4) is used. The results in the figure are positive and the equation (12) is satisfied everywhere except for one value outside the range of the graph corresponding to the top of the slope where $(U_i - U_{i-1})$ is close to zero.

1.2 Outline of the paper.

This simple example illustrates the theme of this paper namely that by switching the order of the method to be used it is possible to preserve positivity of the numerical solution in a way that may be consistent with the solution of the advection equation. In order to achieve this a family of methods based on B splines due to Chin at al. [6] will be rewritten in an upwind form in Sections 2,3,4 and 5. The upwind form used nests the methods in a nonlinear way and includes a nonlinear iteration to take into account the mass matrix. Section

6 describes how these methods are modified so as to preserve positivity and also in the case of the linear method, how to conserve mass. Section 7 uses computational experiments to demonstrate that positivity is indeed preserved. In Section 8 the extension to a single nonlinear conservation law is considered and a Burgers equation example in Section 9 is used to demonstrate that the approach also applies in this case too. A summary of the paper is given in Section 10 while Appendix A provides a brief description of how the algorithm is modified if the flow direction is reversed from that used in Sections 1-9.

2 Overview of Positive B-spline Finite Element Methods

An overview of the new positive methods discussed below is given as follows. Consider a B-Spline pth order basis consisting of functions $b_{i,p}(x)$ such as is used by [6] in a Galerkin approach to get

$$\int_{x_L}^{x_R} \frac{\partial U}{\partial t} b_{i,p}(x)\ dx = \int_{x_L}^{x_R} -\frac{\partial U}{\partial x} b_{i,p}(x)\ dx \tag{13}$$

where the discrete solution is defined by $U(x,t) = \sum_{i=1}^{N} b_{i,p}(x) U_i(t)$. The approach adopted here will be to define a positive scheme of the form:

$$(1 + a_{1,p})U_i^{n+1} - a_{1,p}U_{i+1}^{n+1} = -a_{3,p}(U_i^n - U_{i-1}^n) + U_i^n(1 - a_{2,p}) + a_{2,p}U_{i-1}^n \tag{14}$$

where the coefficients $a_{i,j}$ (which are nonlinearly dependent on the solution values) satisfy two conditions to ensure positivity as defined by equation (6):

$$a_{1,p} \geq 0, \tag{15}$$

$$0 \leq a_{2,p} + a_{3,p} \leq 1 \tag{16}$$

for $i = 1, 2, 3, p = 0, 1, 2, 3$. These coefficients will also be shown below to be hierarchical in that they may be written in the general form $a_{i,p} = a_{i,p-1} + c_{i,p}$ The idea behind the positive schemes described here is that the order, p, will be varied so as to maintain positivity. For simplicity the Forward Euler method will be used for timestepping but the approach used here extends to those positive Runge-Kutta schemes which use internal stages of the same form, see Cockburn and Shu, [8]. The method defined by equation (14) is implicit so simple iteration will be used to solve for the new solution values. In the case of iteration m the equations can then be written, after dividing through by

the quantity $(1 + a_{1,p})$, as:

$$U_i^{n+1,m+1} = \frac{a_{1,p}U_{i+1}^{n+1,m} + U_i^n(1 - (a_{2,p} + a_{3,p})) + (a_{2,p} + a_{3,p})U_{i-1}^n}{(1 + a_{1,p})} \tag{17}$$

to get an iteration which is also positive if the positivity conditions (15) and (16) are satisfied.

In the case of the piecewise constant method defined by equation (4) with Forward Euler time integration employed the coefficients are $a_{1,0} = a_{3,0} = 0$ and $a_{2,p} = \frac{\delta t}{\delta x}$. Hence if the linear method mass matrix coefficient $a_{1,1}$ is negative and a switch is made to $a_{1,0}$ (mass lumping) then the mass matrix has been lumped based on the solution values used to define $a_{1,1}$. As $a_{1,1}$ depends on the ratios of solution gradients r_i^n defined in equations (9) amd (10) this may be viewed as putting a nonlinear lumping switch in the mass matrix or as is termed by Berzins, [3], using a nonlinear form of the mass matrix.

It should also be noted that the directionality in the formula defined by equation (14) reflects the directionality of the underlying advection equation. The extension to negative velocity is relatively straightforward and is considered in Appendix A.

3 Positive Linear/Constant Finite Element Method

Applying the Forward Euler method to equation (2) and separating out the terms that depend on second differences gives

$$\begin{aligned} U_i^{n+1} + \frac{1}{6}\left[U_{i+1}^{n+1} - 2U_i^{n+1} + U_{i-1}^{n+1}\right] = U_i^n + \frac{1}{6}\left[U_{i+1}^n - 2U_i^n + U_{i-1}^n\right] \\ - \frac{\delta t}{\delta x}(U_i^n - U_{i-1}^n) \; - \; \frac{\delta t}{2\delta x}\left[U_{i-1}^n - 2U_i^n + U_{i+1}^n\right]. \end{aligned} \tag{18}$$

Using equations (9) and (10) enables equation (18) to be rewritten as:

$$U_i^{n+1} + \frac{1}{6}\delta^2 U_i^{n+1} = \frac{1}{6}\delta^2 U_i^n + U_i^n(1 - \frac{\delta t}{\delta x}\beta_i) + U_{i-1}^n\frac{\delta t}{\delta x}\beta_i \tag{19}$$

where $\beta_i = \left[(1 + \frac{(r_i^n - 1)}{2})\right]$ and the central difference operator δ^2 is defined as in equation (5). Equations (19) may be solved by using the iteration based on

defining $U_k^{n+1,m}$ as the solution value at mesh point k at time t_{n+1} for iteration m. Noting, as in equation (10), that

$$\delta^2 U_i^{n+1,m} = (1/r_i^{n+1,m} - 1)(U_i^{n+1,m} - U_{i+1}^{n+1,m})$$

where $r_i^{n+1,m} = (U_{i+1}^{n+1,m} - U_i^{n+1,m})/(U_i^{n+1,m} - U_{i-1}^{n+1,m})$ and writing $\alpha_i^{n+1,m} = (1/r_i^{n+1,m} - 1)$ allows an iteration to be defined by

$$(1 + \frac{\alpha_i^{n+1,m-1}}{6})U_i^{n+1,m} = U_i^n(1 - \frac{\delta t}{\delta x}\beta_i) + U_{i-1}^n\frac{\delta t}{\delta x}\beta_i + \frac{\alpha_i^{n+1,m-1}}{6}U_{i+1}^{n+1,m-1} + \frac{(r_i^n - 1)}{6}(U_i^n - U_{i-1}^n) \quad (20)$$

where β_i is defined as in equation (19) above. The coefficients $a_{i,j}$ of Section 2 are then given by:

$$a_{1,1} = \frac{\alpha_i^{n+1,m-1}}{6} \quad (21)$$

$$a_{2,1} = \frac{\delta t}{\delta x}\beta_i \quad (22)$$

$$a_{3,1} = \frac{(r_i^n - 1)}{6} \quad (23)$$

From these definitions and from Section 2 the positivity conditions (16) and (15) are then

$$0 \leq \frac{(1 - r_i^n)}{6} + \frac{\delta t}{\delta x}(1 + \frac{(r_i^n - 1)}{2}) \leq 1 \quad (24)$$

and $\alpha_i^{n+1,m} \geq 0$ respectively. These conditions may be satisfied by ensuring that

$$0 \leq r_i^{n+1,m+1} \leq 1,\ 0 \leq r_i^n \leq 1 \text{ and } 0 \leq \frac{\delta t}{\delta x}(1 + \frac{(r_i^n - 1)}{2}) \leq 5/6. \quad (25)$$

It is worth noting that although this scheme is positive it is not conservative. This issue will be addressed in Sections 6 and 8 below. It is also worth remarking that it is possible to use different restrictions on r in different parts of the algorithm. The rightmost condition (relating to the space derivative) allows a wider range of values of r_i^n than the central condition which is concerned with the mass matrix components at t_n. For example if $0 \leq \frac{\delta t}{\delta x} < 1/3$ then the condition on r_i^n in the space derivative approximation may be relaxed to $0 \leq r_i \leq 4$. If these conditions do not hold the piecewise constant method

defined by equation (4) is used instead. The predictor to provide the values $U_i^{n+1,0}$ simply uses a lumped mass matrix.

4 Positive Quadratic B-Spline finite Element Method

Rather than use the standard quadratic finite element method with its different treatment of odd and even nodes, e.g. see Gresho and Sani [11], it is more straightforward to use the B-spline Galerkin method introduced by Chin et al. [6] and analysed in Vichnevetsky and Bowles [23] as the equations are identical at each mesh point. The computational performance of these methods has been studied by Griffiths [12] and found to be at least as good or possibly superior to that of conventional quadratic Galerkin methods. This method gives rise to the o.d.e. system defined by

$$\frac{\dot{U}_{i-2} + 26\dot{U}_{i-1} + 66\dot{U}_i + 26\dot{U}_{i+1} + \dot{U}_{i+2}}{120} = \frac{-10(U_{i+1} - U_{i-1}) - (U_{i+2} - U_{i-2})}{24\delta x}. \tag{26}$$

As in the previous section, apply the Forward Euler method, the notation of equation (5) and a considerable amount of manipulation to rewrite the method as

$$\begin{aligned} U_i^{n+1} + (\frac{1}{6} + \frac{(8 + s_i^{n+1} + \frac{1}{s_{i-1}^{n+1}})}{120})\delta^2 U_i^{n+1} &= U_i^n + (\frac{1}{6} + \frac{(8 + s_i^n + \frac{1}{s_{i-1}^n})}{120})\delta^2 U_i^n \\ &- \frac{\delta t}{\delta x}(U_i^n - U_{i-1}^n) - \frac{\delta t}{2\delta x}\,(1 + \frac{1}{12}(s_i^n - \frac{1}{s_{i-1}^n}))\,\delta^2 U_i^n \end{aligned} \tag{27}$$

where the second derivative ratios at time levels such as t^{n+1} are given by

$$s_i^{n+1} = \delta^2 U_{i+1}^{n+1}/\delta^2 U_i^{n+1} \quad \text{and} \quad s_{i-1}^{n+1} = \delta^2 U_i^{n+1}/\delta^2 U_{i-1}^{n+1}$$

and the ratios at time level n denoted by s_i^n and s_{i-1}^n are similarly defined. Using the same approximation as in equations (9) and (10) gives

$$\begin{aligned} &U_i^{n+1} + \frac{1}{6}(1 + \frac{1}{20}(8 + S_+^{i,n+1}))\left[\frac{1}{r_i^n} - 1\right](U_i^{n+1} - U_{i+1}^{n+1}) \\ &= U_i^n + \frac{1}{6}(1 + \frac{1}{20}(8 + S_+^{i,n}))\,[r_i - 1]\,(U_i^n - U_{i-1}^n) \\ &\quad - \frac{\delta t}{\delta x}\left[1 + (1 + \frac{1}{12}(S_-^{i,n}))(\beta_i - 1)\right](U_i^n - U_{i-1}^n) \end{aligned} \tag{28}$$

where the quantities $S_+^{i,n}$and $S_-^{i,n}$ are defined by

$$S_+^{i,n} = s_i^n + \frac{1}{s_{i-1}^n} \quad \text{and} \quad S_-^{i,n} = s_i^n - \frac{1}{s_{i-1}^n},$$

and where β_i is defined by equation (19) above. This system of equations may again be solved by using the iteration based on defining $U_k^{n+1,m}$ as the solution value at mesh point k at time t_{n+1} at iteration m. The same procedure as used for linear methods may be employed by rewriting the left side of equations (28) as

$$U_i^{n+1,m+1} + \frac{1}{6}(1 + \frac{1}{20}(8 + S_+^{i,n+1,m})) \left[\frac{1}{r_i^{n+1,m}} - 1\right] (U_i^{n+1,m+1} - U_{i+1}^{n+1,m})$$

The coefficients defined in Section 2 are given by:

$$a_{1,2} = \frac{1}{6}(1 + \frac{1}{20}(8 + S_+^{i,n+1,m})) \left[\frac{1}{r_i^{n+1,m}} - 1\right] \tag{29}$$

$$a_{2,2} = \frac{\delta t}{\delta x}\left[1 + (1 + \frac{1}{12}(S_-^{i,n}))(\beta_i - 1)\right] \tag{30}$$

$$a_{3,2} = \frac{1}{6}(1 + \frac{1}{20}(8 + S_+^{i,n}))\left[1 - r_i\right]. \tag{31}$$

In order to get an iteration that satisfies equations (15) and (16) and preserves positivity it is thus necessary to impose the same restriction as the two leftmost conditions in equation (22) plus the extra condition:

$$\frac{1}{6}(1 + \frac{1}{20}(8 + S_+^{i,n+1,m})) > 0 \tag{32}$$

Assuming that $0 \leq S_+^{i,n} \leq 2$ and that $-1 \leq S_+^{i,n} \leq 1$ (see Section 6) the positivity condition (16) may then be written as:

$$0 \leq (\frac{1}{4} + \frac{\delta t}{\delta x}(\frac{13 - \beta_i}{12})) \leq 1 \tag{33}$$

which is satisfied by the range of β_i values allowed by equations (24) in the linear case providing that

$$0 \leq \frac{\delta t}{\delta x} \leq \frac{9}{12.5}. \tag{34}$$

5 Positive Cubic B-Spline Finite Element Method

Rather than use the standard cubic finite element method with its different treatment of odd and even nodes as used by Christie and Mitchell [7], it is again more straightforward to use the B-spline Galerkin method introduced by Chin et al. [6]. This gives rise to the o.d.e. system defined by

$$\frac{1}{5040}\left[\dot{U}_{i-3}+120\dot{U}_{i-2}+1191\dot{U}_{i-1}+2416\dot{U}_i+1191\dot{U}_{i+1}+120\dot{U}_{i+2}+\dot{U}_{i+3}\right]$$
$$=\frac{-245}{720\delta x}(U_{i+1}-U_{i-1})\ \frac{-56}{720\delta x}(U_{i+2}-U_{i-2})\ \frac{-1}{720\delta x}(U_{i+3}-U_{i-3}). \qquad (35)$$

The same idea as with linear and quadratic elements may be used to rewrite this as a positive scheme. Applying Forward Euler timestepping gives

$$U_i^{n+1}+\frac{1}{6}(1+\frac{1}{20}(8+s_i^{n+1}+\frac{1}{s_{i-1}^{n+1}}))\delta^2U_i^{n+1}$$
$$+\frac{258}{5040}\delta^2U_i^{n+1}+\frac{80}{5040}\left[\delta^2U_{i+1}^{n+1}+\delta^2U_{i-1}^{n+1}\right]+\frac{1}{5040}\left[\delta^2U_{i+2}^{n+1}+\delta^2U_{i-2}^{n+1}\right]$$
$$=U_i^n+\frac{1}{6}(1+\frac{1}{20}(8+s_i^n+\frac{1}{s_{i-1}^n}))\delta^2U_i^n$$
$$+\frac{258}{5040}\delta^2U_i^n+\frac{80}{5040}\left[\delta^2U_{i+1}^n+\delta^2U_{i-1}^n\right]+\frac{1}{5040}\left[\delta^2U_{i+2}^n+\delta^2U_{i-2}^n\right]$$
$$-\frac{\delta t}{\delta x}(U_i^n-U_{i-1}^n)-\frac{\delta t}{2\delta x}\ (1+\frac{1}{12}(s_i^n-\frac{1}{s_{i-1}^n}))\delta^2U_i^n$$
$$-\frac{28}{720}\frac{\delta t}{\delta x}\left[\delta^2U_{i+1}^n-\delta^2U_{i-1}^n\right]-\frac{1}{720}\frac{\delta t}{\delta x}\left[\delta^3U_{i+1}^n-\delta^3U_{i-1}^n\right] \qquad (36)$$

where $\delta^3U_i^n = U_{i+2}^n-2U_{i+1}^n+2U_{i-1}^n-U_{i-2}^n$ and $\delta^3U_i^n = \delta^2U_{i+1}^n-\delta^2U_{i-1}^n$.

The additional terms over and above the quadratic B-spline method may be rewritten as follows. Consider first the mass matrix terms

$$\frac{1}{5040}\left[258\delta^2U_i^{n+1}+80\left(\delta^2U_{i+1}^{n+1}+\delta^2U_{i-1}^{n+1}\right)+\left(\delta^2U_{i+2}^{n+1}+\delta^2U_{i-2}^{n+1}\right)\right]=$$
$$\frac{1}{5040}\left[258+80(s_i^{n+1}+\frac{1}{s_{i-1}^{n+1}})+(s_{i+1}^{n+1}s_i^{n+1}+\frac{1}{s_{i-2}^{n+1}}\frac{1}{s_{i-1}^{n+1}})\right]\delta^2U_i^{n+1}. \qquad (37)$$

The substitutions of equations (9) and (10) enable these terms to be rewritten in the same positivity preserving form as the quadratic method above. The same approach is used with the extra stiffness matrix terms

$$-\frac{28}{720}\frac{\delta t}{\delta x}\left[\delta^2 U_{i+1}^n - \delta^2 U_{i-1}^n\right] - \frac{1}{720}\frac{\delta t}{\delta x}\left[\delta^3 U_{i+1}^n + \delta^3 U_{i-1}^n\right] =$$
$$-\frac{28}{720}\frac{\delta t}{\delta x}\left[s_i^n - \frac{1}{s_{i-1}^n}\right]\delta^2 U_i^n - \frac{1}{720}\frac{\delta t}{\delta x}\left[w_i^n + \frac{1}{w_{i-1}^n}\right]\delta^3 U_i^n \tag{38}$$

where the third derivative ratios at time levels such as t^n are given by

$$w_i^n = \delta^3 U_{i+1}^n/\delta^3 U_i^n \quad \text{and} \quad w_{i-1}^n = \delta^3 U_i^n/\delta^2 U_{i-1}^n.$$

Using the substitutions of equations (9) and (10) and the substitution $\delta^3 U_i^n = (s_i^n - \frac{1}{s_{i-1}^n})\,[r_i^n - 1]\,(U_i^n - U_{i-1}^n)$ enables the right side of (38) to be rewritten as

$$-\frac{28}{720}\frac{\delta t}{\delta x}\left[\delta^2 U_{i+1}^n - \delta^2 U_{i-1}^n\right] - \frac{1}{720}\frac{\delta t}{\delta x}\left[\delta^3 U_{i+1}^n + \delta^3 U_{i-1}^n\right] =$$
$$-\frac{1}{720}\frac{\delta t}{\delta x}\left[28 + (w_i^n + \frac{1}{w_{i-1}^n})\right]\left[s_i^n - \frac{1}{s_{i-1}^n}\right][r_i^n - 1]\,(U_i^n - U_{i-1}^n). \tag{39}$$

Now defining the quantities $W_+^{j,n}$ and $\Gamma_+^{n,i}$by

$$W_+^{j,n} = \left[w_j^n + \frac{1}{w_{j-1}^n}\right] \quad \text{and} \quad \Gamma_+^{n,i} = s_{i+1}^n s_i^n + 1/s_{i-1}^n\ 1/s_{i-2}^n$$

allows equations (36) to be rewritten as:

$$(\frac{1}{6} + \frac{(8 + S_+^{i,n+1})}{120} + \frac{\left[258 + 80S_+^{i,n+1} + \Gamma_+^{n+1,i}\right]}{5040})\left[\frac{1}{r_i^{n+1}} - 1\right](U_i^{n+1} - U_{i+1}^{n+1}) +$$

$$U_i^{n+1} = U_i^n + (\frac{1}{6} + \frac{(8 + S_+^{i,n})}{120} + \frac{\left[258 + 80S_+^{i,n} + \Gamma_+^{n,i}\right]}{5040})\,[r_i^n - 1]\,(U_i^n - U_{i-1}^n)$$
$$-\frac{\delta t}{\delta x}\left[1 + (1 + \frac{S_-^{i,n}}{12}(1 + \frac{(28 + W_+^{i,n})}{30}))(\beta_i - 1)\right](U_i^n - U_{i-1}^n) \tag{40}$$

where β_i is defined by equation (19) above. A similar iteration method to the linear and quadratic cases is used. The coefficients of Section 2 are now given by

$$a_{1,3} = a_{1,2} + \frac{\left[258 + 80S_+^{i,n+1} + \Gamma_+^{n+1,i}\right]}{840}\left[\frac{1}{r_i^{n+1}} - 1\right] \tag{41}$$

$$a_{2,3} = a_{2,2} + \frac{S_-^{i,n}}{12}(\frac{(28 + W_+^{i,n})}{30})(\beta_i - 1) \tag{42}$$

$$a_{3,3} = a_{3,2} + \frac{\left[258 + 80S_+^{i,n} + \Gamma_+^{n,i}\right]}{840}\left[1 - r_i^n\right] \tag{43}$$

In order to satisfy the positivity conditions (15) and (16) we make the assumptions that it will be possible to implement the constraints $0 \leq W_+^{i,n} \leq 2$ and that $0 \leq \Gamma_+^{n,i} \leq 2$. This assumption will be justified in the next section. A worst-case analysis using these assumptions then gives rise to the positivity condition

$$0 \leq (\frac{1}{3} + \frac{\delta t}{\delta x}(1 + \beta_i - \frac{1}{6}(\beta_i - 1))) < 1 \tag{44}$$

which, after assuming the constraint on β_i given by equation (25), gives

$$0 \leq \frac{\delta t}{\delta x} \leq \frac{6}{11}. \tag{45}$$

6 Preserving Positivity and Changing Order.

In order to preserve positivity the values that the coefficients $a_{i,j}$ can take must satisfy equations (15) and (16). In this section some possible ways of achieving this for the methods defined above are discussed. These are not the only mechanisms nor are they necessarily the best, but are an attempt to show that the ideas discussed here have potential for further development.

6.1 Positive Linear/Constant Finite Element Method

Although the method as defined in Section 3 preserves positivity it is not conservative in the sense of [14]. This means that the numerical approximation to a wave may travel at a different speed from the true solution to the p.d.e. Although this issue will be discussed further in Section 7 the solution adopted may be briefly described here by noting the restriction for the space derivative terms given by

$$0 \leq \ r_j^n \ \leq R, \quad j = 1, ..., N \tag{46}$$

is a global restriction. This suggest that the term $(1 + (r_i^n - 1)/2)(U_i(t) - U_{i-1}(t))$ should be calculated by taking into account the restriction imposed by equation (46) on both r_{i-1}^n and r_i^n. Hence in the definition of $\delta^2 U_i(t)$ in

equation (18) and β_i in equation (19) the term $\delta^2 U_i(t)$ is replaced by the approximation

$$\delta^2 U_i(t) \approx \Phi_R(r_i^n)(U_i(t) - U_{i-1}(t)) - \Phi_R(r_{i-1}^n)(U_{i-1}(t) - U_{i-2}(t)) \quad (47)$$

where $\Phi_R(r_j^n) = max(0, min(r_j^n, R))$, is the minmod function used, for example, by [4]. Values of R employed have ranged from $R = 1$ to $R = 4$ and both of these values have yielded good results. It is important to note that equality in this approximation holds only if both r_{i-1}^n and r_i^n satisfy equation (46). In the case when equation (47) is used to substitute for the final rightmost instance of $\delta^2 U_i^n$ in equation (18) then $\beta_j = \left[1 + \frac{\Phi^R(r_j^n)}{2} - \frac{\Phi^R(r_{j-1}^n)}{2r_{j-1}^n}\right]$. The righthand side of equation (20) is then similar to the expression that arises when a finite volume method with van Leer limiter is applied to the same equation to get

$$\dot{U}_j(t) = \frac{-1}{\delta x}\left[1 + \frac{V(r_j^n)}{2} - \frac{V(r_{j-1}^n)}{2r_{j-1}^n}\right](U_j(t) - U_{j-1}(t)) \quad (48)$$

and where the limiter function is defined by: $V(r_j^n) = \frac{r_j^n + |r_j^n|}{1+|r_j^n|}$, [16]. The importance of this approach is that it results in a spatial discretisation of the spatial derivative terms that is conservative, see [14]. The computational results in the next section show that the new method may give better resolution than the traditional finite volume approach but at the greater computational expense of using the mass matrix.

6.2 Positive Quadratic Finite Element Method

The restrictions on the functions $S_+^{i,n}$ and $S_-^{n,i}$ may be satisfied by the same general method as used with linear functions. This is done by redefining these functions as:

$$S_+^{i,n} = \left[\Phi_1(s_i^n) + \Phi_1(\frac{1}{s_{i-1}^n})\right], \quad (49)$$

$$S_-^{i,n} = \left[\Phi_1(s_i^n) - \Phi_1(\frac{1}{s_{i-1}^n})\right], \quad (50)$$

where the function $\Phi_1(.)$ is defined as in equation (47)with $R = 1$. It is worth noting that $0 \leq S_+^{i,n} \leq 2$ and that $-1 \leq S_-^{i,n} \leq 1$ as required. The quantities $S_\pm^{i,n+1}$ and $\Phi_1(s_i^{n+1})$ are similarly defined at time t_{n+1}. In the case when $S_-^{i,n} = 0$ (and the method defaults to the linear method in approximating the the space

derivative) it follows that $\Phi_1(s_i^n) = \Phi_1(\frac{1}{s_{i-1}^n})$ and that the values of s_i^n and $\frac{1}{s_{i-1}^n}$ are either less than 0 or greater than 1. Alternatively $s_i^n = \frac{1}{s_{i-1}^n}$ and so $\delta^2 u_{i+1}^n = \delta^2 u_{i-1}^n$ thus implying that second derivative approximations are constant and thus that a linear approximation is more appropriate.

6.3 Cubic Spline Positive Finite Element Method

The restrictions on the functions $W_+^{j,n}$ and $\Gamma_+^{n,i}$ may be satisfied by the same general method as used with linear and quadratic functions. This is done by redefining these functions as:

$$W_+^{i,n} = \left[\Phi_1(w_i^n) + \Phi_1(\frac{1}{w_{i-1}^n})\right]$$

and

$$\Gamma_+^{n,i} = \Phi_1(s_{i+1}^n)\Phi_1(s_i^n) + \Phi_1(1/s_{i-1}^n)\ \Phi_1(1/s_{i-2}^n)$$

This ensures that $0 \leq W_+^{i,n} \leq 2$ and that $0 \leq \Gamma_+^{n,i} \leq 2$ as was required in deriving the constraint defined by equation (44) for positivity .

6.4 Adaptive order Algorithm.

The general strategy employed in changing the order of the polynomial used in the method described above is to use the highest order possible unless its use is precluded by the positivity conditions operating in such a way as to reduce the order. This strategy has been influenced by variable order strategies such as the h-p methods used by Biswas, Devine and Flaherty [4] in which successive polynomial derivatives are limited. For the purposes of changing the order only the space derivative terms are considered, although the mass matrix may still be modified independently. In changing from piecewise discontinuous to piecewise linear if $\beta_i = 1$ then the limiters are both zero so we stay with the discontinuous method of equation (4). In changing from linear to quadratic if $S_-^{j,n} = 0$ then then quadratic terms in the derivative are switched off and so a linear basis is used. In the case when a quadratic method is used and at least one cubic limiter term is not zero or one then $0 < W_+^{j,n} < 2$ and so if this condition is not satisfied we stay with a quadratic method. This algorithm is applied on a point by point basis.

7 Linear Advection Numerical Experiments.

Two advection test problems are used to demonstrate the positivity of the new method and to compare its performance against a traditional finite volume method. The first problem consists of the advection of a square pulse. This problem is quite a demanding one in that the top of the pulse is only five mesh intervals across and the gradient is sufficiently steep to be viewed as a discontinuity. The solution to this travelling pulse problem with the piecewise constant and linear methods is shown in Figure 3. The lefthand image shows the pulse at three times and shows the results of the new method and of the finite volume method with the van Leer harmonic limiter defined by equation (48). The difference between the two methods is shown by the small pulses at the bottom of the leftmost diagram. The righthand image shows the order used at each of these times. The solution to the same travelling pulse problem but

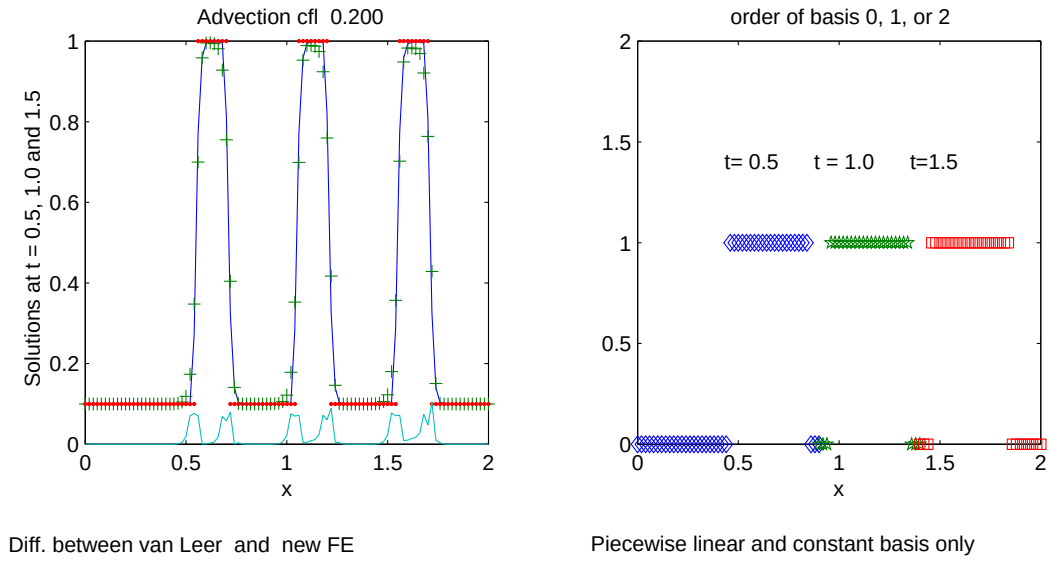

Fig. 3. Simple limited FE Scheme vs van Leer

now also using the option of switching to a quadratic basis is shown in Figure 4. The lefthand image shows the pulse at three times while the righthand image shows the order used at each of these times. The L1 error norm for the finite volume scheme is 1.06e-2 while that for the finite element method is 1.08e-2. These results show that both methods preserve positivity and the the accuracy of both methods is comparable. The second advection problem has a solution which is both smooth and which has a steep profile is given by [15] as an 11th order polynomial which is defined in terms of the variable z, where

$$z = (0.3 + t + ds * 0.5 - x)/ds; \tag{51}$$

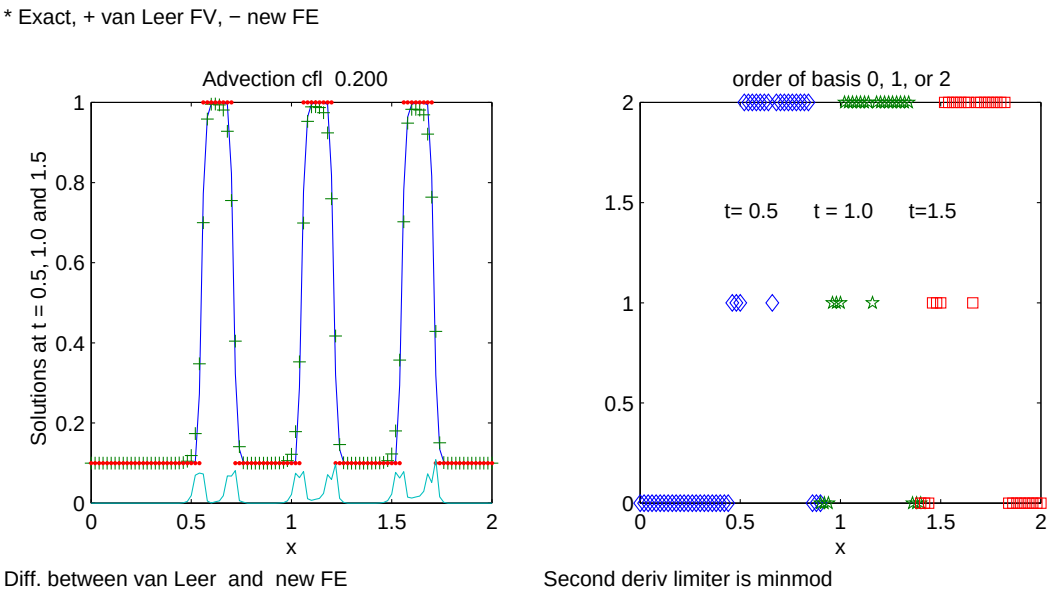

Fig. 4. Quad limited FE Scheme vs van Leer

In the case when $z > 1$ then $u(x,t) = 1.1$ while if $z < 0$ then $u(x,t) = 0.1$. For $0 \leq z \leq 1$ the value of $u(x,t) = p(z)$ where

$$p(z) = z^6 \left[-252z^5 + 1386z^4 - 3080z^3 + 3465z^2 - 1980z + 462\right] \qquad (52)$$

and where z is defined by equation (51). The solution has a front of width ds centred about $0.3+t$. Two sets of numerical experiments were conducted with this problem. This is an example with a less steep gradient which demonstrates

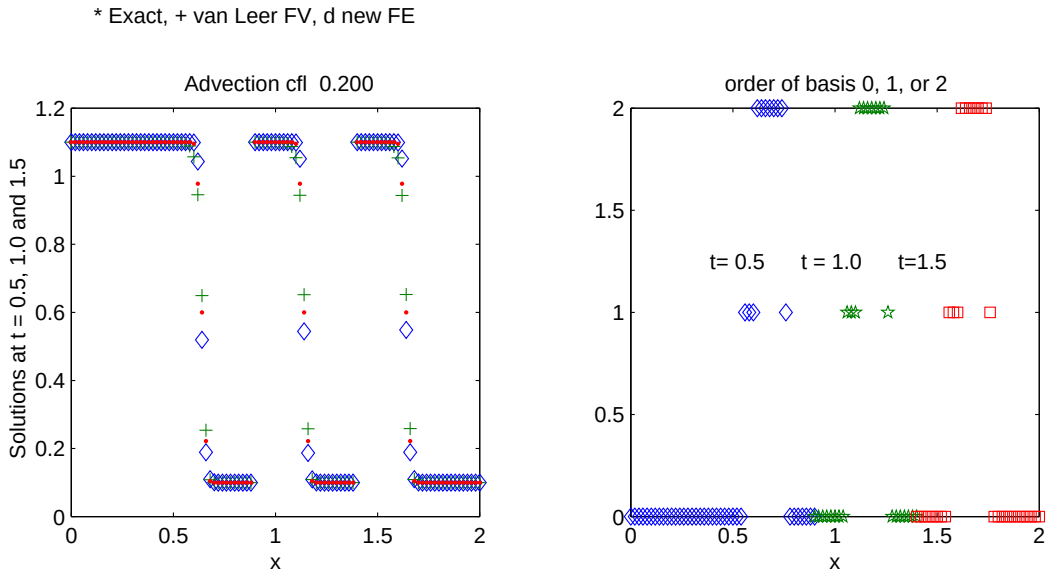

Fig. 5. Quad limited FE Scheme vs van Leer on smooth solution

the use of the cubic method. Figure 6 show the profiles at $t = 1.5$ and Table

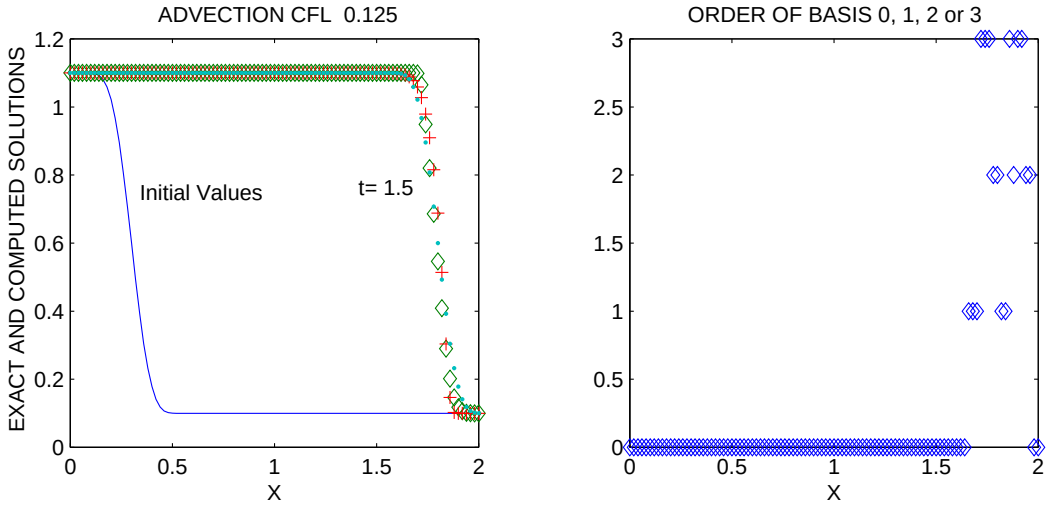

Fig. 6. Cubic limited FE Scheme vs van Leer on smooth solution

1 shows the the errors in the L1 norm at the same time. The value of R used

L1 Error Norms				
Finite Volume	Finite Element			
van Leer limiter	linear	quadratic	cubic	R value used
2.1e-2	2.4e-2	2.1e-2	1.8e-2	4
2.1e-2	1.5e-2	1.5e-2	1.3e-2	1

Table 1
L1 Error Norm for Finite Element and Finite Volume Schemes

in equation (46) is specified in Table 1. The results are encouraging in that quadratic and cubic methods are used away from the top and bottom of the front. Improvements in accuracy obtained by using the high order methods are modest when measuring the error over the whole range of integration, possibly because of the relatively small spatial interval over which the higher order methods are used.

8 Nonlinear Conservation Laws.

In considering the extension of the method described above to nonlinear conservation laws the approach of Spekreijse, [21], is followed for the scalar partial differential equation in one space dimension given by:

$$u_t + [f(u)]_x = 0 \tag{53}$$

where $f(u)$ is the advective flux function which describes *wave* movements in the solution. Spekreijse, [21], assumes that this can be split into positive and negative parts:

$$f\,(u) \;=\; f_l\,(u) \;+\; f_r\,(u) \tag{54}$$

where

$$\frac{df_l\,(u)}{du} \;\geq\; 0 \text{ and } \frac{df_r\,(u)}{du} \;\leq\; 0\;. \tag{55}$$

Using this approach with the piecewise constant DG method defined by equations (4) with the Forward Euler method with time step δt gives the equations:

$$U_i(t_{n+1}) = U_i(t_n) + \frac{\delta t}{\delta x}\left[A^n_{i+1/2}(U_{i+1}(t_n) - U_i(t_n)) - B^n_{i-1/2}(U_i(t_n) - U_{i-1}(t_n))\right]$$

where $i = 1, ..., n$, $t_{n+1} = t_n + k$ and where

$$\begin{aligned} A^n_{i+1/2} &= -\frac{f_r(U_{i+1}(t_n)) - f_r(U_i(t_n))}{U_{i+1}(t_n) - U_i(t_n)} \\ B^n_{i-1/2} &= \frac{f_l(U_i(t_n)) - f_l(U_{i-1}(t_n))}{U_i(t_n) - U_{i-1}(t_n)}\;. \end{aligned}$$

The approach taken in considering nonlinear problems may be illustrated by considering the inviscid Burgers' equation defined by

$$\frac{\partial u}{\partial t} = -\frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) \quad , \quad (x,t)\;\epsilon\;(0,2) \;\times\; (0,2] \quad , \tag{56}$$

with Dirichlet boundary conditions and the initial condition consistent with the analytic solution

$$u(x,t) = \frac{0.1A + 0.5B + C}{A + B + C} \tag{57}$$

where $A = e^{(-0.05(x-0.5+4.95t)/\nu)}, B = e^{-0.25(x-0.5+0.75t)/\nu)}, C = e^{(-0.5(x-0.375)/\nu)}$ and where the limit as $\nu \downarrow 0$ is taken. This gives a solution consisting of a pair of steep fronts that eventually form one front. In this case the flow is in the direction of increasing x and so

$$f\,(u) \;=\; f_l\,(u) \;=\; \left(\frac{u^2}{2}\right) \tag{58}$$

and consequently that

$$A^n_{i+1/2} = 0$$
$$B^n_{i-1/2} = \frac{1}{2}\ (U_i(t_n) + U_{i-1}(t_n)).$$

Applying a natural extension of the approach used in Sections 3 and 6 above gives

$$U_i^{n+1} + \frac{1}{6}\delta^2 U_i^{n+1} = \frac{1}{6}\delta^2 U_i^n + U_i^n(1 - \ \frac{\delta t}{\delta x}\beta_i) + U_{i-1}^n \frac{\delta t}{\delta x}\beta_i \tag{59}$$

where the central difference operator δ^2 is defined as in equation (5) and where

$$\beta_i = \left[1 + \frac{\Phi^R(\bar{fr}_i^n)}{2} - \frac{\Phi^R(\bar{fr}_{i-1}^n)}{2\bar{fr}_{i-1}^n}\right]\ B^n_{i-1/2}. \tag{60}$$

In this case the ratio $\bar{fr}_i^n$ is a ratio of values of the function $f(u)$ as defined by

$$\bar{fr}_i^n = \frac{f(U_{i+1}^n) - f(U_i^n)}{f(U_i^n) - f(U_{i-1}^n)} \tag{61}$$

Equations (59) may be solved by using the same iteration as was used in Section 3 in which $U_k^{n+1,m}$ is defined as the solution value at mesh point k at time t_{n+1} with iteration m. Defining terms as in equation (20) allows an iteration to be defined by

$$(1 + \frac{\alpha_i^{n+1,m-1}}{6})U_i^{n+1,m} = U_i^n(1 - \frac{\delta t}{\delta x}\beta_i) + U_{i-1}^n\frac{\delta t}{\delta x}\beta_i + \ \frac{\alpha_i^{n+1,m-1}}{6}U_{i+1}^{n+1,m-1} + \frac{(r_i^n - 1)}{6}(U_i^n - U_{i-1}^n) \tag{62}$$

where β_i is defined as in equation (60) above. The restriction for positivity is similar to that given in equation (24), in that

$$0 \le \frac{(1 - r_i^n)}{6} + \frac{\delta t}{\delta x}\beta_i \le 1 \tag{63}$$

In the same way as is described at the end of Section 3 a different restriction on the ratios of solution jumps was needed for the mass matrix iteration. The restriction used to compute the results shown was

$$0 \le r_i, r_i^{n+1,m} \le 1/2, \quad \rightarrow \alpha_i^{n+1,m-1} > 1. \tag{64}$$

The need for this extra restriction requires further research and may in some way be related to the difficulties documented by Venkatakrishnan [22] when using limiter based schemes as part of iterative solution procedures when solving nonlinear problems.

8.1 Conservative Form.

Although the importance having local conservation properties in finite volume schemes is well understood, [14], the local conservation properties of finite element Galerkin schemes are less well understood though recently it has been shown by Larson et al. [17] that Galerkin methods may have better conservation properties than was previously thought. In the case of the scheme defined above the piecewise constant method is well-known to be conservative and it can be shown that a modified version of the iteration defined by equation (20) is conservative. Consider the linear basis function scheme defined by equation (19) with the term β_i defined as in Section 6.1. In addition modify the iteration defined by equation (20) to read:

$$U_i^{n+1,m+1} = \frac{-1}{6}\delta^2 U_i^{n+1,m} + \frac{1}{6}\delta^2 U_i^n + U_i^n(1 - \frac{\delta t}{\delta x}\beta_i) + U_{i-1}^n \frac{\delta t}{\delta x}\beta_i \tag{65}$$

which in turn can be written as

$$U_i^{n+1,m+1} = U_i^n + F_{i+1/2}^m - F_{i-1/2}^m \tag{66}$$

where the numerical fluxes $F_{i+1/2}^m$ and $F_{i-1/2}^m$ at iteration m are defined by

$$\begin{aligned} F_{i+1/2}^m = \frac{-1}{6}(U_{i+1}^{n+1,m} - U_i^{n+1,m}) + \frac{1}{6}(U_{i+1}^n - U_i^n) \\ - \frac{\delta t}{\delta x}(U_i^n + \frac{\Phi^R(r_i^n)}{2}(U_i^n - U_{i-1}^n)) \end{aligned} \tag{67}$$

and

$$\begin{aligned} F_{i-1/2}^m = \frac{-1}{6}(U_i^{n+1,m} - U_{i-1}^{n+1,m}) + \frac{1}{6}(U_i^n - U_{i-1}^n) \\ - \frac{\delta t}{\delta x}(U_{i-1}^n + \frac{\Phi^R(r_{i-1}^n)}{2}(U_{i-1}^n - U_{i-2}^n)) \end{aligned} \tag{68}$$

The same approach works in the case when the approximation defined by equation (47) is used. Thus providing that the updates in any iteration are done in a Jacobi-type fashion the iteration has a conservative form. It is possible to

extend this approach to the quadratic and cubic cases. In the quadratic case, as defined in Section 4 without using limiters, the formula given by equation (27) may be written as:

$$U_i^{n+1,m+1} = U_i^n + G_{i+1/2}^m - G_{i-1/2}^m \tag{69}$$

where the numerical fluxes $G_{i+1/2}^m$ and $G_{i-1/2}^m$ at iteration m are defined by

$$\begin{aligned} G_{i+1/2}^m = \frac{-1}{4}(U_{i+1}^{n+1,m} - U_i^{n+1,m}) - \frac{1}{120}(\delta^2 U_{i+1}^{n+1,m} - \delta^2 U_i^{n+1,m}) \\ + \frac{1}{4}(U_{i+1}^n - U_i^n) + \frac{1}{120}(\delta^2 U_{i+1}^n - \delta^2 U_i^n) \\ - \frac{\delta t}{\delta x}(U_i^n + \frac{1}{2}(U_{i+1}^n - U_i^n) + \frac{1}{24}(\delta^2 U_{i+1}^n + \delta^2 U_i^n)) \end{aligned} \tag{70}$$

and

$$\begin{aligned} G_{i-1/2}^m = \frac{-1}{4}(U_i^{n+1,m} - U_{i-1}^{n+1,m}) - \frac{1}{120}(\delta^2 U_i^{n+1,m} - \delta^2 U_{i-1}^{n+1,m}) \\ + \frac{1}{4}(U_i^n - U_{i-1}^n) + \frac{1}{120}(\delta^2 U_i^n - \delta^2 U_{i-1}^n) \\ - \frac{\delta t}{\delta x}(U_{i-1}^n + \frac{1}{2}(U_i^n - U_{i-1}^n) + \frac{1}{24}(\delta^2 U_{i-1}^n + \delta^2 U_i^n)) \end{aligned} \tag{71}$$

It is also important to point out that in order to get a conservative form the term $\frac{\delta t}{24\delta x}\delta^2 U_i^n$ has been added to both fluxes and thus cancels when the difference of the fluxes is taken. As in the case of the linear method defined by equations (66),(67) and (68) it is possible to limit terms consisting of first and second differences. In order to get a conservative quadratic method it is necessary to limit the fluxes so that each flux is treated consistently in both the equations it appears in. The limiting techniques used in Section 6.2, which limit each term individually, may be used in exactly the same way as with the method when expressed in nonconservative form. The only real difficulty occurs when there is an order change between a pair of elements. In this case, say, part of flux $G_{i-1/2}^m$ could be set to zero in one element and used unmodified in a neighbouring element. The issue of order selection and its impact on conservation thus may require further study.

An alternative approach is to note that the results of Hou and LeFloch [14] may be applied. They show that a conservative method may be obtained by switching from a non-conservative method to a conservative one if the condition

$$|U_i^n - U_{i-1}^n| + |U_{i+1}^n - U_i^n| \leq \ b\ (\delta x)^a \tag{72}$$

is violated for some constants $0 < a < 1$ and $b > 0$. As the linear method used here is conservative and as

$$|\delta^2 U_i^n| \leq |U_i^n - U_{i-1}^n| + |U_{i+1}^n - U_i^n| \tag{73}$$

Then a method switch based on the terms $\delta^2 U_i$ will provide a positive conservative scheme in the sense of [14]. This would amount to imposing an additional condition on top of the previously defined switching conditions in Section 6.

9 Inviscid Burgers Equation Numerical Experiments.

This is the example as outlined in equations (56) to (58). Two cases are presented. In the first case a linear method is used and is compared against the finite volume method. In the second case the full adaptive method is used. In both cases the higher order methods are used only in the vicinity of the front and the results show the front correctly positioned and to be without overshoots and undershoots. The linear finite element scheme produces a pos-

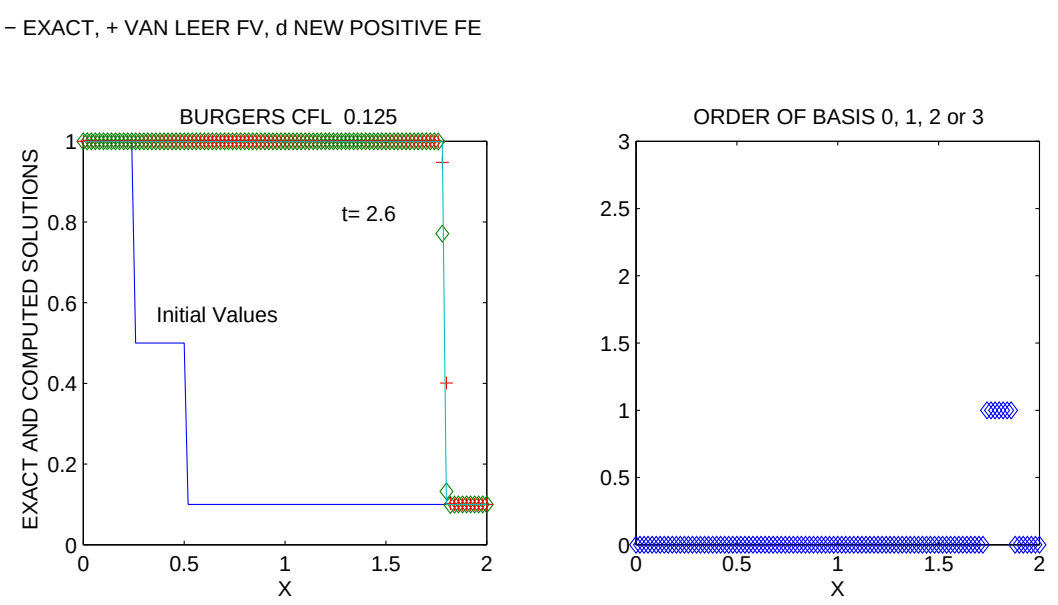

Fig. 7. Linear limited FE Scheme vs van Leer on inviscid Burgers Equation

itive solution with an L1 error norm of 5.6e-3 which is an improvement on the finite volume error norm of 7.1e-3. Figure 7 shows both the computed and true solutions and where piecewise constant and linear basis functions are used at the final time. In the quadratic case shown in Figure 8 the l1 error is 6.02e-3. In the full cubic case adaptive case only a few cubic elements are used where the front is steepest but with a decrease in accuracy. In both these cases the

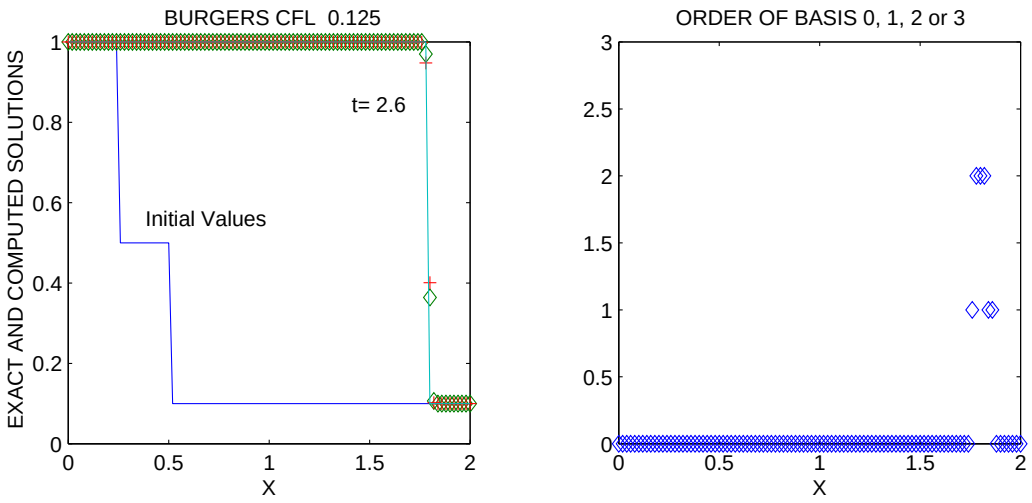

Fig. 8. Quadratic limited FE Scheme vs van Leer on inviscid Burgers Equation

higher order basis functions are only used on one or two mesh intervals so it is not surprising that there is no increase in accuracy.

10 Summary

In this paper a novel approach to preserving positivity for variable-order finite element methods has been taken. The approach relies on using a nonlinear form of the mass matrix in conjunction with positivity preserving conditions on the method coefficients. Initial steps in extending the method to more general hyperbolic conservation laws has also been considered. Extensions to advection in two dimensions have been undertaken by Berzins and Hubbard [2]. Although initial results are promising further work is needed to assess the usefulness of the method in particular in the areas of conservation and of order change.

References

[1] Baker A.J., Chaffin D.J., Ianelli J.S. and Roy S. Finite elements for cfd- how does the theory compare? *International Journal for Numerical Methods in Fluids* 1999; **31**:345–358.

[2] Berzins M and Hubbard M.E. Positive finite element schemes on irregular triangular meshes, *Paper in preparation* 2003.

[3] Berzins M. Modified mass matrices and positivity preservation for hyperbolic and parabolic pde's. *Communications in Numerical Methods in Engineering* **17**:659:666, 2001.

[4] Biswas R., Devine K. and Flaherty J.E. Parallel adaptive finite element methods for conservation laws. *Applied Numerical Mathematics* 1995; **14**:2557–283.

[5] Cardle J.A. A modification of the Petrov-Galerkin method for the transient convection-diffusion equation. *International Journal for Numerical Methods in Engineering* 1995; **38**:171–181.

[6] Chin R.C.Y., Hedstrom G.W. and Karlsson K.E. A simplified Galerkin method for hyperbolic equations. *Mathematics of Computation* 1979; **33**:647–658.

[7] Christie I. and Mitchell A.R. Upwinding of Galerkin methods in conduction-convection problems. *International Journal of Numerical Methods in Engineering* 1978; **33**:1764–1771.

[8] Cockburn B. and Shu C.-W. Nonlinearly stable compact schemes for shock calculations. *SIAM J. Numer. Anal.* 1994; **31**:607–627.

[9] Cockburn B, Karniadakis GE, Shu C.-W. (eds). *Discontinuous Galerkin Methods, Theory Computations and Applications. Lecture Notes in Computational Science and Engineering 11* Springer Berlin Heidelberg, 2000; 3–53.

[10] Godunov S.K. Finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics. *Math Sbornik* 1959; **47**:271–306.

[11] Gresho P.M. and Sani R. L. *Incompressible Flow and the Finite Element Method, Volume 1, Advection Diffusion* Wiley , 1998.

[12] Griffiths D.F. *Personal communication* Summer, 2002.

[13] Harten A. and Zwas G. Self adjusting hybrid schemes for shock computations *J. of Computational Physics* 1972; **9**:568–583.

[14] Hou T.Y. and Le Floch P.G. Why nonconservative schemes converge to wrong solutions: error analysis. *Mathematics of Computation* 1994; **62**:497–530.

[15] Hubbard M.E. *Personal communication* October, 2002.

[16] Laney C. *Computational Gasdynamics*. Cambridge University Press.

[17] Larson M.G., Hughes T.J.R., Engel G. and Mazzei L. The continuous Galerkin method is conservative. *Journal of Computational Physics* 2000; **163**, 2:467-488.

[18] MacKinnon R.J. and Carey G.F. Positivity preserving flux-limited finite-difference and finite-element methods for reactive transport. *International Journal for Numerical Methods in Fluids* 2003; **41**:151-183.

[19] Segal A. Finite element methods for advection-diffusion equations. In *Numerical Methods for Advection Diffusion Problems. Notes on Numerical Fluid Mechanics, Volume 45* Vreugdenhil CB, Koren B (eds). Vieweg: Braunscweig/Wiesbaden, 1993; 195–214.

[20] Sheu T.W.h., Tsai S.F. and Wang M.M.T. A monotone finite element method with test space of Legendre polynomials. *Comput. Methods Appl. Mech. Engrg.* 1997; **143**:349–372.

[21] Spekreijse S. Multigrid solution of monotone second-order discretizations of hyperbolic conservation laws. *Math. of Comp.*, **49**(179):135–155, 1987.

[22] Venkatakrishnan V. Convergence to steady state of the Euler equations on unstructured grids with limiters. *Journal of Computational Physics*, **105**:83-91, 1995.

[23] Vichnevetsky R. and Bowles J.B. *Fourier Analysis of NumericalApproximations of Hyperbolic Equations.* SIAM Philadelphia, 1982.

Appendix A

Consider the case of the advection equation with the opposite wave speed to that considered in equation (1) for which the Galerkin form is given by

$$\int_{x_{i-1}}^{x_{i+1}} \frac{\partial U}{\partial t}\phi_i(x)\ dx = \int_{x_{i-1}}^{x_{i+1}} \frac{\partial U}{\partial x}\phi_i(x)\ dx, i = 1, ..., N. \tag{74}$$

The piecewise constant DG method for the advection equation with positive velocity equal to one is given in a semi-discrete form by

$$\dot{U}_i = \frac{1}{\delta x}(U_{i+1} - U_i) \tag{75}$$

The linear finite element method is obtained as above by applying the Forward Euler method to equation (75) and separating out the terms that depend on second differences to get

$$\begin{aligned} U_i^{n+1} + \frac{1}{6}\left[U_{i+1}^{n+1} - 2U_i^{n+1} + U_{i-1}^{n+1}\right] &= U_i^n + \frac{1}{6}\left[U_{i+1}^n - 2U_i^n + U_{i-1}^n\right] \\ &+ \frac{\delta t}{\delta x}(U_{i+1}^n - U_i^n) \ - \ \frac{\delta t}{2\delta x}\ \left[U_{i-1}^n - 2U_i^n + U_{i+1}^n\right] \end{aligned} \tag{76}$$

The quadratic positive finite element method is given by the same process as

$$\begin{aligned} &U_i^{n+1} + \frac{1}{6}(1 + \frac{1}{20}(8 + S_+^{i,n+1}))\left[r_i^{n+1} - 1\right](U_i^{n+1} - U_{i-1}^{n+1}) \\ &= U_i^n + \frac{1}{6}(1 + \frac{1}{20}(8 + S_+^{i,n}))\left[\frac{1}{r_i^{n+1}} - 1\right](U_i^n - U_{i+1}^n) \end{aligned}$$

$$- \frac{\delta t}{\delta x} \left[1 + (-1 + \frac{1}{12}(S_{-}^{i,n}))(\beta_i - 1)) \right] (U_{i+1}^{n} - U_{i}^{n}) \tag{77}$$

where the quantitiesβ_i and $S_{+}^{i,n}$and $S_{-}^{i,n}$ are defined as in equations (19) and (28) respectively.