DEPARTMENT OF MATHEMATICS

PLAGIARISM:–Priestley's LAgrange Galerkin Is A Really Interesting Scheme & Monotone.

The Multidimensional Monotone and Nearly Conservative Lagrange-Galerkin Method

by

A. Priestley

Numerical Analysis Report **15/91**

UNIVERSITY OF READING

PLAGIARISM:–Priestley's LAgrange Galerkin Is A Really Interesting Scheme & Monotone.

The Multidimensional Monotone and Nearly Conservative Lagrange-Galerkin Method*

A. Priestley

Institute of Computational Fluid Dynamics,

Dept. of Mathematics,

University of Reading,

Whiteknights,

Reading,

U.K..

*The work reported here forms part of the research programme of the Reading/Oxford Institute for Computational Fluid Dynamics and was funded by the SERC.

0 Abstract

The Lagrange-Galerkin method has proved a very successful method but there are problems with monotonicity and conservation. In this paper we address these two issues and propose an algorithm that allows no new extrema to be generated by the scheme. It is equally applicable to multi-dimensions, and also prevents any systematic error in conserved quantities. Furthermore this is achieved at very little additional cost and without compromising the unconditional stability of the method.

1 Introduction

The Lagrange-Galerkin method is a scheme that combines the method of characteristics with a standard finite element procedure, see Benqué et al 1982, Bercovier & Pironneau 1982, Douglas & Russell 1982, Lesaint 1977, Pironneau 1982 and Russell 1980 for example. There are two versions of the scheme available depending upon whether one travels forwards in time along the particle trajectory or backwards. In Morton & Priestley 1986 these two schemes were labelled the weak Lagrange-Galerkin method and the direct Lagrange-Galerkin method.

We shall briefly describe the derivation of the schemes and in the following section discuss the basic properties of the scheme, both good and bad. In Section 3 an algorithm is introduced that removes oscillations from the Lagrange-Galerkin method, one of its major drawbacks. To demonstrate its effectiveness a simple test problem is solved in Section 4, to be followed by a brief summary.

Consider the Cauchy problem for the scalar, linear advection equation for $u(\underline{x}, t)$:

$$u_t + \underline{a}.\underline{\nabla} u = 0, \qquad \underline{x} \in \mathbb{R}^{\mathrm{d}}, \quad \mathrm{t} > 0, \tag{1}$$

$$u(\underline{x}, 0) \;=\; u_0(\underline{x}), \tag{2}$$

where u_0 belongs to $L^2(\mathbb{R}^{\mathrm{d}})$, d being the number of dimensions. The velocity field $\underline{a}(\underline{x}, t)$ is assumed to be incompressible, i.e.

$$\underline{\nabla}.\underline{a} = 0 \quad \forall \underline{x}, t. \tag{3}$$

We can now define characteristics paths or trajectories, $\underline{X}(\underline{x}, s; t)$, in two ways, either as the solution to an ordinary differential equation,

$$\underline{X}(\underline{x}, s; s) = \underline{x}, \tag{4}$$

$$\frac{d\underline{X}(\underline{x}, s; t)}{dt} = \underline{a}(\underline{X}(\underline{x}, s; t), t); \tag{5}$$

or, if desired, as the solution of the integral equation

$$\underline{X}(\underline{x}, s; t) = \underline{x} + \int_s^t \underline{a}(\underline{X}(\underline{x}, s; \tau), \tau) d\tau.$$

In order to simplify the notation, for $t^{n+1} = t^n + \Delta t$, we will denote the foot of the characteristic path at time t^n to be at $\underline{x}$ and its arrival point at time t^{n+1} to be at $\underline{y}$. In terms of the more general notation, these are

$$\underline{x} = \underline{X}(\underline{y}, t^{n+1}; t^n)$$

and

$$\underline{y} = \underline{X}(\underline{x}, t^n; t^{n+1}).$$

A unique (absolutely continuous) solution to equations (4,5) can be guaranteed if it is assumed that $\underline{a}$ belongs to the Bochner space $L^1(0, T; (W^{1,\infty})^d)$, see Mizohata 1972 for example. The solution to the original partial differential equation (1,2) is now given by the relation

$$u(\underline{X}(\cdot, t, t + \tau), t + \tau) = u(\cdot, t). \tag{6}$$

The direct Lagrange-Galerkin method uses (6) directly to obtain a solution. For an approximation, U^n, at time t^n given in terms of finite element basis functions ϕ_j,

$$U^n(\cdot) = \sum_j U_j^n \phi_j(\cdot), \tag{7}$$

the direct Lagrange-Galerkin method uses eq. (6) to obtain U^{n+1} in $L^2(\mathbb{R}^d)$ satisfying

$$\langle U^{n+1}, \phi_i \rangle = \int U^n(\underline{x})\phi_i(\underline{y})d\underline{y} \quad \forall i, \tag{8}$$

equation (8) being obtained by taking the weak form of (6), i.e., multiplying (6) by $\phi_i(\underline{y})$ and integrating over the whole domain with respect to $\underline{y}$, and the L^2 inner product over $\mathbb{R}^d$ being denoted by $\langle \cdot, \cdot \rangle$. This is the same approach as that used by Bercovier & Pironneau 1982, Douglas & Russell 1982 and Pironneau 1982, for example.

A second, alternative, formulation has been proposed by Benqué et al 1982 and this will be referred to as the weak formulation or the weak Lagrange-Galerkin method because the adjoint of the differential operator in eq. (1) is applied to a test function. This method introduces new test functions $\psi(\cdot, t)$, which now are not only different from the basis functions used in the expansion of the solution as in eq. (7), but also depend on time. Multiplying equation (1) by this test function and integrating in space and time we get

$$\int_t^{t+\Delta t} \langle u_t + \underline{a}.\underline{\nabla} u, \psi_i \rangle dt = 0 \quad \forall i.$$

Integrating by parts, with respect to space and time, we obtain

$$\langle u(\cdot,t+\Delta t),\psi_i(\cdot,t+\Delta t)\rangle - \langle u(\cdot,t),\psi_i(\cdot,t)\rangle = \int_t^{t+\Delta t} \langle u, \frac{\partial \psi_i}{\partial t} + \underline{\nabla}.(\underline{a}\psi_i)\rangle dt \quad \forall i. \tag{9}$$

Using the incompressibility condition, eq. (3), $\underline{\nabla}.(\underline{a}\psi_i)$ can be rewritten as $\underline{a}.\underline{\nabla}\psi_i$ so that this last term vanishes, because the integrand is zero, if the test functions satisfy

$$\psi_i(\underline{X}(\cdot,t;t+\tau),t+\tau) = \psi_i(\cdot,t) \qquad \forall i. \tag{10}$$

To solve (10) a final condition on ψ_i is imposed by setting

$$\psi_i(\cdot,t+\Delta t) = \phi_i(\cdot). \tag{11}$$

Substituting our finite element approximation (7) into (9) together with (10) and (11) gives

$$\langle U^{n+1},\phi_i\rangle = \int U^n(\underline{x})\psi_i(\underline{x},t^n)d\underline{x} \quad \forall i,$$

or equivalently,

$$\langle U^{n+1},\phi_i\rangle = \int U^n(\underline{x})\phi_i(\underline{y})d\underline{x} \quad \forall i. \tag{12}$$

If J is the Jacobi matrix of the transformation defined by the mapping $\underline{X}(\cdot,t^n;t^{n+1})$, then we have, as shown in Chorin & Marsden 1984,

$$\frac{d|J|}{dt} = (\underline{\nabla}.\underline{a})|J|,$$

where $|J|$ is the determinant of J. For $\underline{\nabla}.\underline{a} = 0$ this means that

$$|J| = constant = 1, \tag{13}$$

which in turn implies that $d\underline{x} \equiv d\underline{y}$ and so (8) and (12) give the same scheme when exact integration is used.

It is worth noting that when $\underline{\nabla}.\underline{a} \neq 0$, i.e. (13) no longer holds and the value of $|J|$ changes then the direct and weak formulations are not the same. With $\underline{x}$ and $\underline{y}$ still related as before, the direct Lagrange-Galerkin method still approximates equation (1). However, the weak Lagrange-Galerkin method now approximates the equation

$$u_t + \underline{\nabla}.(\underline{a}u) = 0. \tag{14}$$

If it is desired to use the weak method for equation (1) or the direct method for equation (14) in the situation $\underline{\nabla}.\underline{a} \neq 0$ then a term $u\underline{\nabla}.\underline{a}$ will need to be added or subtracted on the right-hand side of the equations.

2 Some Basic Properties of the Schemes

This paper is not intended to be a review of theoretical results concerning the Lagrange-Galerkin method but some of the results are worth recapping and there are others we shall use later.

Assuming that finite elements are being used for which $\sum_i \phi_i \equiv 1$ then for either formulation conservation follows immediately. One simply sums equation (8) or (12) over all i to obtain, in the case of the weak Lagrange-Galerkin method

$$\int U^{n+1}(\underline{x})d\underline{x} = \int U^n(\underline{x})d\underline{x}$$

which is just an obvious statement of conservation. For the conservation of the direct method it is also necessary to use eq. (13) to transform the region of integration.

Suppose now that we denote by $E_{\Delta t}(t)$ the solution operator $u(\cdot, t+\Delta t) = E_{\Delta t}(t)u(\cdot, t)$, and similarly $E^{-1}_{\Delta t}(t)u(\cdot, t+\Delta t) = u(\cdot, t)$, for equation (1) over the time-step Δt; that is from (6)

$$E_{\Delta t}(t)u(\cdot, t) = u(\cdot, t+\Delta t) = u(\underline{X}(\cdot, t+\Delta t; t), t). \tag{15}$$

Letting $\| \cdot \|$ denote the usual L^2 norm over $\mathbb{R}^d$ and $\| \cdot \|_*$ be the correspondingly defined operator norm, we have

$$\| E_{\Delta t}(t) \|_* = \sup_{u \neq 0} \frac{\| E_{\Delta t}(t)u(\cdot, t) \|}{\| u(\cdot, t) \|} = \sup_{u \neq 0} \frac{\| u(\cdot, t+\Delta t) \|}{\| u(\cdot, t) \|},$$

and from (15) and (13) this equals 1. Hence

$$\| E_{\Delta t}(t) \|_* = 1 = \| E^{-1}_{\Delta t}(t) \|_* . \tag{16}$$

The unconditional stability of both the direct and weak methods now follows.

For the direct method we can write

$$\langle U^{n+1}, \phi_i \rangle = \langle E_{\Delta t}(t) U^n, \phi_i \rangle.$$

Multiplying by U_i^{n+1} and summing over i gives

$$\begin{aligned} \| U^{n+1} \|^2 &= \langle U^n, E_{\Delta t}(t^n), U^{n+1} \rangle \\ &\leq \| E_{\Delta t}(t^n) U^n \| \, . \, \| U^{n+1} \| \, . \end{aligned}$$

Hence

$$\| U^{n+1} \| \leq \| U^n \| \, .$$

Similarly for the weak method, we can write

$$\langle U^{n+1}, \phi_i \rangle = \langle U^n, E_{\Delta t}^{-1}(t^n) \phi_i \rangle.$$

Using (16) and the Cauchy-Schwarz inequality, we have,

$$\| U^{n+1} \| \leq \| U^n \|,$$

to again deduce unconditional stability.

THEOREM 1 *The direct Lagrange-Galerkin method with C^0 finite elements of degree k ($k \geq 1$) converges with order k in the $l^\infty(0, T; (L^2)^d)$ norm, provided that*

$u_0 \in (H^{k+1})^d$, $\underline{a} \in L^\infty(0,T;(W^{1,\infty})^d)$ and the corresponding solution u of (1) belongs to the space $H^1(0,T;(H^{k+1})^d)$.

Proof See Morton et al 1988.

Also in Morton et al 1988 it is mentioned that there are certain special cases where this result can be improved upon. For one-dimensional constant linear advection the Lagrange-Galerkin method with piecewise linear elements on a uniform grid becomes third order accurate. When $\underline{a}$ is a smooth function the method is second order accurate in the $l^\infty(0,T;l^2)$ norm on a non-uniform mesh.

Other theoretical results can be found in Lesaint 1977, Süli 1988a and Süli & Ware 1991.

Although Lesaint 1977 gave theoretical results involving quadrature, little work had been done on this aspect of the method. Except in the simplest of problems, quadrature formulae will have to be used to evaluate the right-hand sides of the methods (8) or (12). In Morton & Priestley 1986 some elementary results were given regarding the effects of quadrature. If the basis functions $\phi(\cdot)$ and the quadrature rule are symmetric, then for constant linear advection the weak and direct Lagrange-Galerkin methods give the same results even under quadrature. Denoting the abscissae of the quadrature rule by y_k and the corresponding weights by w_k and replacing the integrals by sums over the elements and sums over the quadrature points, the coefficient of U_j^n in the right-hand side of the direct method is proportional to

$$\sum_e \sum_k w_k \phi(y_k + e - i)\phi(y_k + e - j - \nu); \tag{17}$$

the corresponding sum for the weak method is

$$\sum_{e'} \sum_{k'} w_{k'} \phi(y_{k'} + e' - i + \nu)\phi(y_{k'} + e' - j)$$

$$= \sum_{e'} \sum_{k'} w_{k'} \phi(-y_{k'} - e' + j)\phi(-y_{k'} - e' + i - \nu). \tag{18}$$

These two expressions, (17) and (18), are the same if $y'_k = 1 - y_k$, $e' = i + j - e - 1$. We would expect this result to be true for constant linear advection. The more interesting aspect of the proof is that the result is clearly not going to hold for equations more complicated than constant coefficient linear advection. However, this simple result will be useful later.

Another straightforward result regarding quadrature is in the extension to multi-dimensional constant advection with a velocity field (a, b, c). Consider the three-dimensional basis function $\phi_{i,j,k}(\underline{x})$ formed by taking the product of three one-dimensional basis functions, i.e., $\phi_{i,j,k}(\underline{x}) = \phi_i(x)\phi_j(y)\phi_k(z)$. If we now assume that our quadrature rule is also of this form, that is made up of a product of the usual one-dimensional rules with the weights given by $w_{i,j,k} = w_i w_j w_k$ and the abscissae given by $\underline{x}_{i,j,k} = (x_i, y_j, z_k)$ then the three-dimensional direct (or weak) Lagrange-Galerkin methods are equivalent to the product of three one-dimensional methods. To demonstrate this we recall the right-hand side of the direct Lagrange-Galerkin method is given by $\int U^n(\underline{x})\phi_i(\underline{y})d\underline{y}$. Under

quadrature this is replaced by the summation

$$\sum_{e=1}^{elements} \sum_{l=1}^{abscissae} w_l U^n(\underline{x}(\underline{y}_l))\phi(\underline{y}_l),$$

where the notation, $\underline{x}(\underline{y}_l)$, denotes $\underline{X}(\underline{y}_k, t^{n+1}; t^n)$ in the more rigorous notation used previously. Substituting in for U^n from (7) and expanding the sums into their various component directions we get

$$\sum_{e_x}\sum_{e_y}\sum_{e_z}\sum_{i}\sum_{j}\sum_{k} w_i w_j w_k \sum_{l=1}^{x\ nodes} \sum_{m=1}^{y\ nodes} \sum_{n=1}^{z\ nodes} \ldots$$

$$\ldots U^n_{l,m,n}\phi_l(x - a\Delta t)\phi_m(y - b\Delta t)\phi_n(z - c\Delta t).$$

It is then easy to see that this can be rearranged to give

$$\left(\sum_{e_x}\sum_{i} w_i \sum_{l=1}^{x\ nodes} \phi_l(x - a\Delta t)\right) \left(\sum_{e_y}\sum_{j} w_j \sum_{m=1}^{y\ nodes} \phi_m(y - b\Delta t)\right) \ldots$$

$$\ldots \left(\sum_{e_z}\sum_{k} w_k \sum_{n=1}^{z\ nodes} \phi_n(z - c\Delta t)\right) U^n_{l,m,n}$$

and hence the result. This result also holds for the more general case of $(a, b, c) = (a(x), b(y), c(z))$.

The next results concern the conservation of the schemes when approximate integration is used. Using quadrature, the right-hand side of the weak method, (12), becomes

$$\sum_{elements}\sum_{k} w_k U^n(\underline{x}_k)\phi_i(\underline{y}(\underline{x}_k)) \quad \forall i. \tag{19}$$

Similarly, the direct method, (8), becomes

$$\sum_{elements} \sum_{k} w_k U^n(\underline{x}(\underline{y}_k))\phi_i(\underline{y}_k) \qquad \forall i, \tag{20}$$

the roles of $\underline{x}$ and $\underline{y}$ having been reversed because the two methods are integrated over different variables, see (12) and (8). Just as in the exactly integrated case a summation over i is now performed. Since $\sum_i \phi_i(\cdot) \equiv 1$ irrespective of the argument, or the functional form of the argument as in the case of eq. (19), the two approximate integrals expressing the conservation of the quantity U become

$$\sum_{elements} \sum_{k} w_k U^n(\underline{x}_k) \tag{21}$$

and

$$\sum_{elements} \sum_{k} w_k U^n(\underline{x}(\underline{y}_k)). \tag{22}$$

Assuming that our quadrature can integrate polynomials of at least the degree of our elements exactly then it is clear that eq. (21) is an exact representation of $\int U^n(\underline{x})d\underline{x}$, and hence the weak Lagrange-Galerkin method is conservative even when quadrature is used. On the other hand equation (22) is not an equivalent expression to $\int U^n(\underline{x})d\underline{x}$ because the points where the function, U^n, is to be evaluated are not at the quadrature points but at some function of them. Therefore this integral will not in general conserve U. The only case where we can guarantee conservation with the direct Lagrange-Galerkin method is in the trivial case mentioned before, where the two schemes become identical.

It is worth noting that the weak method is conservative irrespective

of the velocity field or how badly we approximate the trajectories. Conservation, in practice, is also dependent on inverting the mass matrix exactly. Usually an iterative method is used to achieve this and it is unlikely that the inversion will be completed to convergence. As a result, in practice neither scheme will conserve exactly. Results were given in Morton et al 1988 that showed that this was the case although the weak Lagrange-Galerkin method performed consistently better than the direct method on the basis of this measure.

Possibly the most interesting result regarding the effects of quadrature was given by Priestley 1986 and Morton et al 1988, which was later extended to include diffusion by Süli 1988b. This was achieved by a simple Fourier analysis of the constant coefficient advection equation. (As mentioned previously both methods are identical for this problem under mild restrictions on the quadrature used.) The result is stated here as

THEOREM 2 *If the right-hand side of the Lagrange-Galerkin method, using piecewise linear elements on a uniform mesh, is approximated by a quadrature of the form*

$$\int_0^1 f(x)dx \approx w_0 f(0) + \sum_1^m w_k f(x_k) + w_{m+1} f(1),$$

where the weights $w_0, \ldots, w_{m+1}$ *and the quadrature points* $0 < x_1 < \ldots < x_m < 1$ *are free to be chosen except that we assume that the quadrature evaluates the integrals of quadratic polynomials exactly, then the method is unstable for CFL numbers* $\nu \in (2w_{m+1}, 1 - x_m)$ *if*

$$2w_{m+1} < 1 - x_m. \tag{23}$$

Proof Priestley 1986 or Morton et al 1988, where similar results were also proved about the lower order quadratures.

The consequences of this theorem are quite wide ranging. Gauss-Legendre quadrature (often just called Gauss quadrature) is immediately seen to be unconditionally unstable because $w_{m+1} = 0$. This means that the region of CFL numbers for which instability occurs is given by $(0, 1 - x_m)$ and hence we cannot reduce the time-step, as with other methods, to ensure stability. Newton-Cotes formulae were also shown to fail the criterion of eq. (23) as were various Radau formulae. Gauss-Lobatto quadrature formulae, of which vertex and Simpson's rule are the lowest order versions, was found also to fail (23), although this Gaussian quadrature did at least have a region of stability given by $[0, 2/(m+1)(m+2)]$. Even though this is one of the best quadratures, with Simpson's rule, $m = 1$, the region of stability is only $[0, 1/3]$, a considerable restriction compared with unconditional stability.

Up to a point these results were largely academic in that with the higher order quadratures, although the region of stability may actually decrease, it can be very hard to make them show signs of this instability because of the accuracy of the scheme. In calculations involving the Navier-Stokes equations, where there is some diffusion, it becomes even more difficult. Using bilinear ele-

ments on rectangles for the rotating cone problem Morton et al 1988 were unable to make a 4×4 Gauss-Legendre quadrature go unstable, although in Priestley 1986, in a one-dimensional example where it was then possible to keep the CFL number in the unstable range, it did eventually become unstable. The danger then, in practical calculations with the Navier-Stokes equations for example, is that near boundaries or around stagnation points the scheme will be exposed to fairly constant CFL numbers in the unstable region. Once the unstable modes in the solution have been excited they can then grow very quickly even with the more accurate integration schemes.

To overcome these problems Priestley 1986, Morton et al 1988 introduced the area-weighted Lagrange-Galerkin method which, in contrast to normal quadrature, approximated the velocity field in such a way that the integrals could then be carried out exactly. This was proved to be stable and to converge and was shown to give reasonable results. However, there was a loss of accuracy over the exactly integrated method and the high order quadrature methods. It was also only really practicable on rectangular grids.

There are two other stable ways of implementing the method but both have their drawbacks.

The EPIC algorithm of Eastwood, see Eastwood & Arter 1986 for example, is essentially the direct Lagrange-Galerkin method. It overcomes the conditions of Theorem (2) by using a quadrature that does not integrate quadrat-

ics exactly. This is done by using a compound trapezium rule. This same rule (or one with fewer sub-intervals) must also be used to calculate the integrals that give the elements of the mass matrix. Since the Lagrange-Galerkin method loses accuracy rapidly if the full mass matrix is not used, this means that quite a few sub-intervals need to be taken. This could then become very expensive in higher dimensions if a product rule is used. There is also no obvious extension to triangular elements.

The second way is to use global finite elements, which leads to the Spectral Lagrange-Galerkin method of Süli & Ware 1991. In this paper convergence and stability was proven for Fourier, Legendre and Chebyshev polynomials even under quadrature, as long as the quadrature used was the relevant Gaussian quadrature. The only results for general orthonormal polynomials are given by Priestley 1989a but here only the weaker time-stability was proven. The problems with this method are its cost, its propensity to oscillate and, in common with spectral methods in general, the difficulty in applying it to arbitrary domains. The cost of a time-step, in one-dimension, of the ordinary Fourier spectral method is $O(n \log n)$, where n is the number of modes used whereas the basic Fourier Spectral Lagrange-Galerkin method has a prohibitive cost of $O\left((n \log n)^2\right)$. Süli & Ware 1991, with no loss of accuracy, used a reduced spectral interpolation to reduce the cost to $O(k(n)n \log n)$ where $k(n) \rightarrow \log n$ for large n. Boyd 1991a has used sum acceleration techniques for the interpolation and reduces the cost by a factor of two for large n, Boyd 1991b. The cost for practicable n is much the same for both approaches. Sum acceleration could, in principle, be applied to

a much wider range of polynomials than the reduced spectral interpolation. Süli & Ware 1991 used only Fourier polynomials although the method is equally applicable to all trigonometric polynomials, including Chebyshev polynomials, but not to series like Legendre polynomials, Ware 1991. Priestley 1989b suggested a non-interpolatory version, based on a similar idea due to Ritchie 1986 for a finite difference Lagrangian scheme, that would reduce the cost back to $O(n \log n)$. The problem of the Gibb's phenomenon and oscillations occurring between quadrature points has yet to be solved, Süli 1989, although they do appear to be of lower magnitude than occurs with the more usual spectral methods. Monotone spectral methods are under development, see Cai et al 1989, 1991 and Cai & Shu 1991, and some of these filtering techniques may, one day, be applicable to the Spectral Lagrange-Galerkin method.

In concluding this section, we can say that although there are alternatives it would seem that the most practicable implementation of the Lagrange-Galerkin method is to use a standard integration formula and to hope that instabilities don't arise. In the next section we introduce a method that takes away the need to hope at very little extra cost and with considerable benefits.

3 Construction of a 'Monotone' Scheme

The scheme we develop here introduces no new extrema and hence maintains monotonicity in one-dimension. It is also applicable to multi-dimensional problems. Monotonicity is a one-dimensional concept but we use the word monotone to

describe the scheme since its meaning for multi-dimensional problems is generally understood. Along with monotonicity, conservation is regarded as an important property of a scheme, and indeed many schemes boast both of these properties, see Sweby 1984 for a discussion of the Total Variation Diminishing schemes, for example.

At first the weak Lagrange-Galerkin method might seem a good place to start as this is already conservative (with the reservations about inverting the mass matrix in a consistent fashion). However, it turns out that it is this very property that prevents the scheme from being monotone.

The monotone scheme is created by combining a high order scheme with a low order (monotone) scheme and limiting the weighting between the two schemes so as to achieve a monotone solution. See Löhner et al 1987 for a similar example of this idea. Denoting the high order solution as U^H and the low order solution as U^L we combine these to get a monotone solution, U^M, defined by

$$U_k^M = \alpha_k U_k^H + (1 - \alpha_k) U_k^L \tag{24}$$

with

$$0 \leq \alpha_k \leq 1, \tag{25}$$

where the α_k's are yet to be chosen. If we denote by $\{U^n, k\}$ the set of nodal values of U^n, the monotone solution from the previous time-level, for which $\phi_i(\underline{x}(\underline{y}_k)) \neq 0$ then we choose the maximum value of α that satisfies conditions (25) and (26)

$$\min(\{U^n,k\},U_k^L) \leq \alpha_k U_k^H + (1-\alpha_k)U_k^L \leq \max(\{U^n,k\},U_k^L). \qquad (26)$$

As an example, to illustrate the notation $\{U^n,k\}$, if we consider one-dimensional linear elements and the trajectory that starts at y_k and assume that this is traced back such that $x_i \leq x(y_k) \leq x_{i+1}$ then $\{U^n,k\}$ is simply just the two values U_i^n and U_{i+1}^n.

We will now refer to the maximum values of α that satisfy constraints (25,26) as $\{\alpha_k^{\max}\}$. Now, if the underlying high order scheme is the direct Lagrange-Galerkin method, the scheme never was conservative and we could just substitute these values back into (24) and have our monotone scheme. However, we can now choose sub-optimal α's to try and make the scheme conservative. That is we choose α's such that

$$0 \leq \alpha_k \leq \alpha_k^{\max}$$

whilst at the same time trying to enforce

$$\int U^M(\underline{x})d\underline{x} = \int U^0(\underline{x})d\underline{x} = C, \text{ say.}$$

In general, of course, the best we can do is to minimize the difference between the two quantities. There are many ways this problem can be solved, by linear programming for example, but the following algorithm is found to be a very efficient and direct way of obtaining a solution.

Firstly we define

$$S_i = \int \phi_i(\underline{x}) d\underline{x}$$

and

$$\beta_i = (U_i^H - U_i^L) S_i.$$

The problem now is to maximize the α's subject to

$$\sum_i \alpha_i (U_i^H - U_i^L) S_i = C - \sum_i U_i^L S_i = C^*.$$

Assume that

$$\sum_i \alpha_i^{\max} \beta_i > C^*.$$

If this is not the case then the definitions are just changed so that

$$\beta_i \longrightarrow -\beta_i$$

$$C^* \longrightarrow -C^*.$$

Step 1

if $\beta_i < 0$ then $\alpha_i = \alpha_i^{\max}$

$\text{iflag}(i) = 1$

otherwise

$$\alpha_i = 0$$

$$\text{iflag}(i) = 0.$$

Step 2

$$\text{Define surplus} \quad = C^* - \sum_{\text{iflag}(k)=1} \alpha_k \beta_k.$$

Step 3

$$\text{Define average value of } \alpha \quad \alpha_{AV} = \frac{\text{surplus}}{\sum_{\text{iflag}(k)=0} \beta_k}.$$

Step 4 if $\alpha_{AV} < \alpha_k^{\max} \ \forall k \ s.t. \ \text{iflag}(k) = 0$

then $\alpha_k = \alpha_{AV} \quad \forall k$ s.t. $\text{iflag}(k) = 0.$

END

Step 5 else $\forall k$ s.t. $\text{iflag}(k) = 0$ and $\alpha_{AV} > \alpha_k^{\max}$

put

$$\alpha_k = \alpha_k^{\max}$$

$$\text{iflag}(k) = 1.$$

Step 6 GOTO 2.

If the surplus is negative then there is no conservative solution and the best solution, as regards conservation, is given by the initial setup of the α's in step 1.

A similar algorithm for the weak method is given in Appendix A. We now explain why the direct method is preferred for the monotone scheme.

Clearly U^H is going to be the solution generated by the Lagrange-Galerkin method, see (8) or (12). It remains to determine the lower order, monotone, scheme U^L. One obvious way of doing this would be to define U^L to be the solution of the mass lumped problem, i.e. exactly the same problem as for the higher order scheme but with the mass matrix replaced with the lumped mass matrix. This involves no extra work as the right-hand sides have already been calculated for the higher order scheme and the mass lumped solution is a good starting point for an iterative solution to the problem using the full mass matrix. This approach of lumping to obtain a lower order solution can clearly be used for both the weak and the direct Lagrange-Galerkin methods. We now address the question of monotonicity.

The lumped solution for the weak method using quadrature is given by

$$U_i^{L(\text{weak})} S_i = \sum_{elements} \sum_k w_k U^n(\underline{x}_k)\phi_i(\underline{y}(\underline{x}_k)). \tag{27}$$

Similarly for the direct method

$$U_i^{L(\text{direct})} S_i = \sum_{elements} \sum_k w_k U^n(\underline{x}(\underline{y}_k))\phi_i(\underline{y}_k). \tag{28}$$

From equation (28) we can write

$$U_i^{L(\text{direct})} S_i \leq \max_i(U_i^n) \sum_{elements} \sum_k w_k \phi_i(\underline{y}_k)$$

with a similar result holding for the minimum. We now notice that, provided that our quadrature is again assumed capable of integrating polynomials of degree at least the same as that of the basis functions, then

$$\sum_{elements} \sum_k w_k \phi_i(\underline{y}_k) = \int \phi_i(\underline{y}) d\underline{y}$$

and so

$$U_i^{L(\text{direct})} S_i \leq \max_i(U_i^n) S_i$$

$$\Longrightarrow U_i^{L(\text{direct})} \leq \max_i(U_i^n).$$

A similar result holds for the minimum and hence we can deduce that $U^{L(\text{direct})}$ is a monotone scheme.

Performing a similar argument for the lumped weak Lagrange-Galerkin method, (27) we get

$$U_i^{L(\text{weak})} S_i \leq \max_i(U_i^n) \sum_{elements} \sum_k w_k \phi_i(\underline{y}(\underline{x}_k))$$

but now

$$\sum_{elements} \sum_k w_k \phi_i(\underline{y}(\underline{x}_k)) \neq \int \phi_i(\underline{y}) d\underline{x}$$

except in the trivial case, and so we cannot infer anything about the monotonicity of the lumped weak Lagrange-Galerkin scheme in general. We note that although analytically $d\underline{x} \equiv d\underline{y}$, in practice this is not the case and if we integrate approximately over $\underline{x}$ we maintain conservation but not monotonicity in the weak method, whilst if we integrate approximately over $\underline{y}$ we have monotonicity but lose conservation in the direct method.

4 Results

To demonstrate the potential of this approach we just consider constant coefficient linear advection in one-dimension on the domain $[0, 1]$ with periodic boundary conditions, that is,

$$u_t + u_x = 0$$

$$u(0) = u(1).$$

The initial data is chosen to be a square wave

$$u_0(x) = \begin{cases} 1 & 0.4 \leq x \leq 0.6 \\ 0 & \text{otherwise,} \end{cases}$$

see figure (1). This represents the worst possible data for the Lagrange-Galerkin method to represent, as it has none of the smoothness required to prove convergence for the method, and so it provides a useful test for our monotone

scheme. We will perform 10 revolutions of the square wave. The Lagrange-Galerkin method can do this problem rather well in just one time-step, and so the CFL number is restricted (to take the values 5/12 and 1/12) so that we can observe the accuracy after a large number of time-steps. It is worth emphasising that although we use small CFL numbers there is no CFL restriction on either the original Lagrange-Galerkin method or the monotone Lagrange-Galerkin method described here.

We will include results for several schemes. The exactly integrated Lagrange-Galerkin method, a monotone scheme based on the exactly integrated Lagrange-Galerkin method, the Lagrange-Galerkin method with the integrals calculated by Simpson's rule, a monotone scheme based on the Lagrange-Galerkin method with the integrals calculated by Simpson's rule, the exactly integrated Lagrange-Galerkin method with lumped mass matrix and, to compare with well known schemes currently in use, we consider two TVD schemes using Van Leer's limiter and the Superbee limiter, see Van Leer 1974 and Roe 1985 for example.

In the first case we consider the case of $\nu = 5/12$. Recalling the stability results of Theorem (2) and the subsequent discussion we note that Simpson's rule leads to an unstable scheme. This is demonstrated in figure (2) where the results are given after just one revolution. After 10 revolutions the maximum and minimum are around $\pm 10^{14}$. The monotone scheme based on Simpson's rule stays stable though, it being difficult for a monotone scheme to oscillate. The price we pay is a loss in accuracy. There is a noticeable phase error. In figure (3) we

see a plot of the solution generated by the exactly integrated Lagrange-Galerkin method and the exactly integrated Lagrange-Galerkin method with lumped mass matrix together with the monotone method based on Simpson's method. Firstly we notice that the lumped solution, although undoubtedly monotone, does lack accuracy. This scheme has much in common with first order upwinding. As a comparison the solution generated by Van Leer's limiter is included. Although the monotone Lagrange-Galerkin method is less diffusive than Van Leer's limiter and has maintained a better profile the solution is less acceptable because of the phase error. Although we would not reccomend that the monotone algorithm is used to stabalize schemes it would obviously provide a useful safety net given the need to use a quadrature.

If, instead of using an unstable scheme to provide the high order solution U^H, we use a stable scheme – in this case using exact integration – then we achieve a much better solution. In figures (4) and (5) the monotone Lagrange-Galerkin scheme, with exact integration, is compared to the solution generated by Van Leer's limiter and the Superbee limiter. The Superbee limiter has been chosen because it is especially good at this problem, it being well-known that Superbee transports all initial data very accurately as a square wave. Van Leer's limiter has been chosen for comparison because it is one of the better limiters that is also fairly robust in that it doesn't distort smooth data as Superbee would. In fig. (4) we see that the monotone Lagrange-Galerkin method has done significantly better than Van Leer's limiter. In fig. (5) we see that the Superbee limiter has given a better answer that the monotone Lagrange-Galerkin

method; it must be remembered though that the square wave is a worst case for the Lagrange-Galerkin method whilst for Superbee it is the best case. Moreover Superbee generalizes neither to large CFL numbers nor multi-dimensions. The results for the various methods are summarized in table (i) where the l_2 errors, the extremum values and the percentage of mass conserved are given.

In figures (6) and (7) we see a similar pair of pictures. The difference here is that the CFL number has been reduced to 1/12 and so the Lagrange-Galerkin method with the integrals performed with Simpson's rule is now stable, although it is not terribly accurate and the monotone version based on this scheme is hence also not terribly accurate. We can draw similar conclusions as before for when the integrals are performed accurately (in this case exactly). The results are summarized in table (ii).

At the output time shown both of the monotone Lagrange-Galerkin schemes presented had exact conservation. Whilst exact conservation was achieved at most time-steps it was not achieved at all the time-steps. The monotone versions were always substantially more conservative than the supposedly conservative exactly integrated Lagrange-Galerkin method. The conservation algorithm was able to undo many of the conservation errors induced by errors in inverting the mass matrix. If the conservation algorithm is not used and we just take $\alpha_k = \alpha_k^{\max} \quad \forall k$ then the results are visually identical to the results using the conservation algorithm but now conservation was worse than with the exactly integrated Lagrange-Galerkin scheme. This would indicate that by taking sub-

optimal α's to achieve conservation we are sacrificing very little accuracy.

5 Conclusions

We have shown how to implement, very efficiently, a monotone version of the Lagrange-Galerkin method that to all intents and purposes conserves mass. This method is at least as accurate as the best TVD limited schemes but has the great advantages of being able to be run without a CFL limit and is equally applicable to arbitrary triangular grids. This will be the subject of a follow up paper. To aspire to the full accuracy of the scheme the integrals involved must be accurately modelled. The method can be applied equally well in multiple dimensions where no new extrema will be created. Conservation and monotonicity are desirable properties of a scheme in their own right, but in the context of the Lagrange-Galerkin method monotonicity is also useful to help control the (albeit rather weak) instability caused by the approximation of the integrals by quadrature formulae. These qualities have been demonstrated on a simple test problem.

The method will next be tried on more severe problems where its properties will be more severely tested. Although this method could be useful in many areas of computational fluid dynamics one of the most interesting is the application of this method to the multi-dimensional wave decompositions of the Euler equations of Roe 1986 and Deconinck et al 1986 for example.

A Conservative Algorithm for use with the Weak Method

In this appendix we give an algorithm for the weak Lagrange-Galerkin method similar to that earlier derived for the direct method.

The monotone solution U^M is defined in exactly the same manner as in equation (24) but where now U^H and U^L are both conservative approximations to the solution. It will be assumed that U^L is monotone. The purpose of the algorithm is now to maximize the α's whilst maintaining conservation.

Assume that

$$\sum_i \alpha_i^{\max} \beta_i > 0.$$

If this is not the case then the definition is just changed so that

$$\beta_i \longrightarrow -\beta_i$$

Step 1

if $\beta_i < 0$ then $\alpha_i = \alpha_i^{\max}$

$\text{iflag}(i) = 1$

otherwise

$\alpha_i = 0$

$\text{iflag}(i) = 0.$

Step 2

$$\text{Define surplus} \quad = - \sum_{\text{iflag}(k)=1} \alpha_k \beta_k.$$

Step 3

$$\text{Define average value of } \alpha \quad \alpha_{AV} = \frac{\text{surplus}}{\sum_{\text{iflag}(k)=0} \beta_k}.$$

Step 4 if $\alpha_{AV} < \alpha_k^{\max} \ \forall k \ s.t. \ \text{iflag}(k) = 0$

then $\alpha_k = \alpha_{AV}$.

END

Step 5 else $\forall k$ s.t. $\text{iflag}(k) = 0$ and $\alpha_{AV} > \alpha_k^{\max}$

put

$$\alpha_k = \alpha_k^{\max}$$

$$\text{iflag}(k) = 1.$$

Step 6 GOTO 2.

With both U^H and U^L being conservative then this algorithm guarantees that U^M will also be conservative.

References

[1] Benqué, J.P., Labadie, G. & Ronat, J., 1982: *"A New Finite Element Method for the Navier-Stokes Equations Coupled with a Temperature Equation."* Proc. 4^{th} Int. Symp. on Finite Element Methods in Flow Problems (Ed. T. Kawai), North-Holland, Amsterdam, Oxford, New York, pp. 295-301.

[2] Bercovier, M. & Pironneau, O., 1982: *"Characteristics and the Finite Element Method."* Proc. 4^{th} Int. Symp. on Finite Element Methods in Flow Problems (Ed. T. Kawai), North-Holland, Amsterdam, Oxford, New York, pp. 67-73.

[3] Boyd, J.P., 1991a: *"Sum-Accelerated Pseudospectral Methods: the Euler-Accelerated Sinc Algorithm."* Appl. Numer. Maths., **7**, pp. 287-296.

[4] Boyd, J.P., 1991b: *"A Fast Algorithm for Chebyshev, Fourier & Sinc Interpolation Onto an Irregular Grid."* Dept. of Atmospheric, Oceanic & Space Sciences, University of Michigan, 2455 Hayward Avenue, Ann Arbor, MI 48109. Submitted to J. Comp. Phys..

[5] Cai, W., Gottlieb, D. & Shu, C-W., 1989: *"Essentially Nonoscillatory Spectral Fourier Methods for Shock Wave Calculations."* Maths. Comp., **52**, pp. 389-410.

[6] Cai, W., Gottlieb, D. & Shu, C-W., 1991: *"On One-Sided Filters for Spectral Fourier Approximations of Discontinuous Functions."* ICASE Report No. 91-31, NASA Langley Research Center, Hampton, Virginia.

[7] Cai, W. & Shu, C-W., 1991: *"Uniform High Order Spectral Methods for One and Two Dimensional Euler Equations."* ICASE Report No. 91-26, NASA Langley Research Center, Hampton, Virginia.

[8] Chorin, A.J. & Marsden, J.E., 1984: *"A Mathematical Introduction to Fluid Mechanics (Universitext)."* Springer-Verlag, New York, Berlin, Heidelberg, Tokyo.

[9] Deconinck, H., Hirsch, Ch. & Peuteman, J., 1986: *"Characteristic Decomposition Methods for the Multidimensional Euler Equations."* Lecture Notes in Physics **264**, pp. 216-221, Springer.

[10] Douglas Jr., J. & Russell, T.F., 1982: *"Numerical Methods for Convection-Dominated Diffusion Problems based on combining the Method of Characteristics with Finite Element or Finite Difference Procedures."* SIAM J. Numer. Anal., **19**, pp. 871-885.

[11] Eastwood, J.W. & Arter, W., 1986: *"EPIC — Beyond the Ultimate Difference Scheme."* Numerical Methods for Fluid Dynamics II, pp. 581-593, (Ed.

K.W. Morton & M.J. Baines), O.U.P., Oxford.

[12] Lesaint, P., 1977: *"Numerical Solution of the Equation of Continuity."* Topics in Numerical Analysis III (Ed. J.J.H. Miller), Academic Press, London, New York, San Francisco, pp. 199-222.

[13] Löhner, R., Morgan, K., Peraire, J. & Vahdati, M., 1987: *"Finite Element Flux-Corrected Transport (FEM-FCT) for the Euler and Navier-Stokes Equations."* Int. J. Numer. Meth. Fluids, **7**, pp. 1093–1109.

[14] Mizohata, S., 1973: *"The Theory of Partial Differential Equations."* Cambridge University Press.

[15] Morton, K.W. & Priestley, A., 1986: *"On Characteristic and Lagrange-Galerkin Methods."* Pitman Research Notes in Mathematics Series (Ed. D.F. Griffiths & G.A. Watson), Longman Scientific and Technical, Harlow, pp. 157-172.

[16] Morton, K.W., Priestley, A. & Süli, E.E., 1988: *"Stability of the Lagrange-Galerkin Method with Non-Exact Integration."* RAIRO M^2AN, **22**, no. 4, pp. 625-653.

[17] Pironneau, O., 1982: *"On the Transport Diffusion Algorithm and its Application to the Navier-Stokes Equations."* Numer. Math., **38**, pp. 309-332.

[18] Priestley, A., 1986: *"Lagrange and Characteristic Galerkin Methods for Evolutionary Problems."* D.Phil. Thesis, Oxford University.

[19] Priestley, A., 1989a: *"The Spectral Lagrange-Galerkin Method for the Atmospheric Transportation of Pollutants. Part II: Stability of the Non-Exact Integration."* University of Reading Numerical Analysis Report 4/89, Dept. of Mathematics, University of Reading, Reading, RG6 2AX.

[20] Priestley, A., 1989b: *"The Spectral Lagrange-Galerkin Method for the Atmospheric Transportation of Pollutants. Part III: The Atmospheric Transportation of a Pollutant!"* University of Reading Numerical Analysis Report 9/89, Dept. of Mathematics, University of Reading, Reading, RG6 2AX.

[21] Ritchie, H., 1986: *"Eliminating the Interpolation Associated with Semi-Lagrangian Schemes."* Mon. Wea. Rev., **114**, pp. 135-146.

[22] Roe, P.L., 1985: *"Large Scale Computations in Fluid Mechanics."* Lectures in Applied Mathematics, **22**, no. 2, pp. 163-193.

[23] Roe, P.L., 1986: *"Discrete Models for Numerical Analysis of Time-Dependent Multi-Dimensional Gas Dynamics."* J. Comp. Phys, **63**, pp. 458-476.

[24] Russell, T.F., 1980: *"Time Stepping along Characteristics with Incomplete Iteration for a Galerkin Approximation of Miscible Displacement in Porous Media."* Ph.D. Thesis, University of Chicago.

[25] Süli, E.E., 1988a: *"Convergence and Nonlinear Stability of the Lagrange-Galerkin Method for the Navier-Stokes Equations."* Numer. Math., **53**, pp. 459-483.

[26] Süli, E.E., 1988b: *"Stability and Convergence of the Lagrange-Galerkin Method with Non-Exact Integration."* The Mathematics of Finite Elements and Applications VI, MAFELAP 1987 (Ed. J.R. Whiteman), Academic Press, London, New York, San Francisco, pp. 435-442.

[27] Süli, E.E., 1989: Private Communication.

[28] Süli, E.E. & Ware, A., 1991: *"A Spectral Method of Characteristics for Hyperbolic Problems."* SIAM J. Numer. Anal., **28**, pp. 423-445.

[29] Sweby, P.K., 1984: *"High Resolution Schemes Using Flux Limiters for Hyperbolic Conservation Laws."* SIAM J. Numer. Anal., **21**, pp. 995-1011.

[30] Van Leer, B., 1974: *"Towards the Ultimate Conservative Difference Scheme II. Monotonicity and Conservation Combined in a Second Order Scheme."* J. Comp. Phys., **14**, pp. 361-371.

[31] Ware, A., 1991: Private Communication.

List of Tables

List of Figures

$\Delta x = 1/100$, $\nu = 5/12$	l_2 error	Maximum	Minimum	Conservation (%)
Lumped L-G	0.3648	0.306	0.1166	100.007
L-G (Exact)	9.9×10^{-2}	1.0611	-6.02×10^{-2}	100.0048
L-G (Simpson's)	—	—	—	—
Monotone L-G (Simpson's)	0.3073	0.9998	-3.469×10^{-18}	100.0
Monotone L-G (Exact)	0.10888	1.0	-1.735×10^{-18}	100.0
Van Leer's	0.1348	0.9558	3.306×10^{-11}	100.0
Superbee	7.1×10^{-2}	1.0	2.8×10^{-28}	100.0

Table i: Errors after 10 revolutions. $\nu = 5/12$.

$\Delta x = 1/100$, $\nu = 1/12$	l_2 error	Maximum	Minimum	Conservation (%)
Lumped L-G	0.407	0.2101	0.21	100.036
L-G (Exact)	1.0×10^{-1}	1.11	-0.115	100.02
L-G (Simpson's)	0.2976	0.49	1.22×10^{-2}	100.024
Monotone L-G (Simpson's)	0.2998	0.4857	1.39×10^{-2}	100.0
Monotone L-G (Exact)	0.12	1.0	-1.86×10^{-17}	100.0
Van Leer's	0.157	0.915	3.27×10^{-11}	100.0
Superbee	7.23×10^{-2}	1.0	2.0×10^{-29}	100.0

Table ii: Errors after 10 revolutions. $\nu = 1/12$.

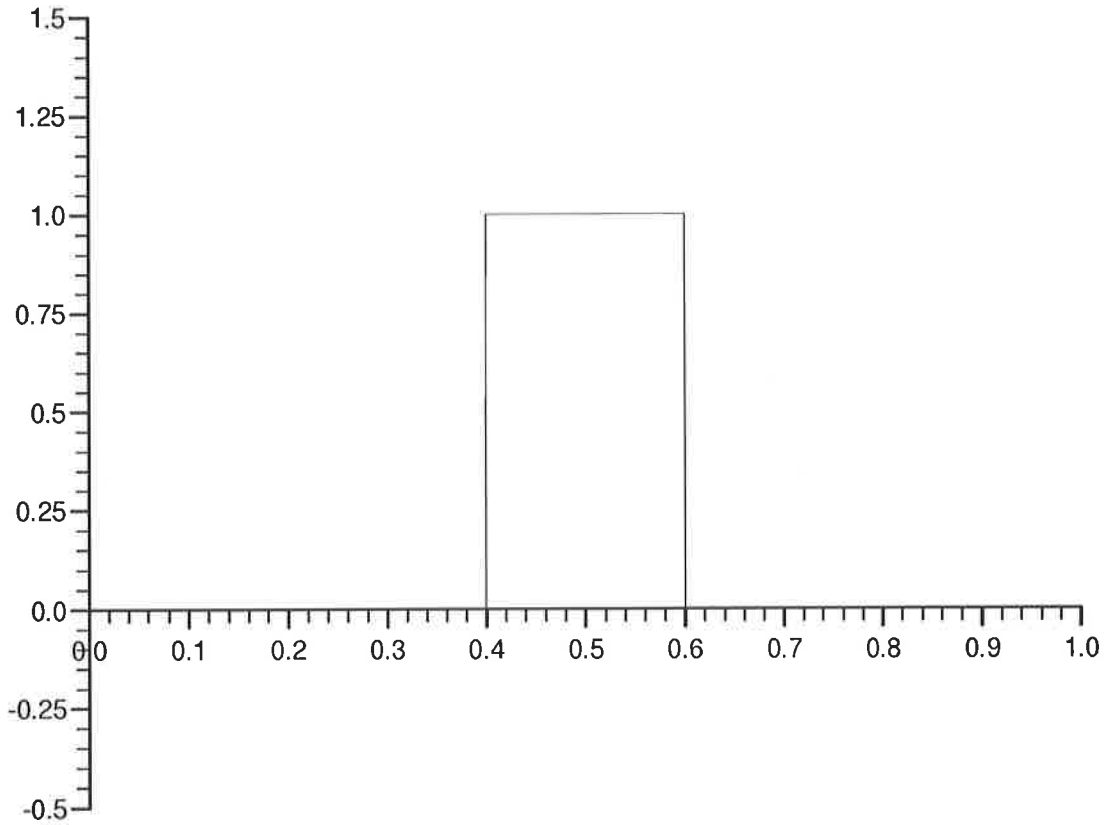

Figure 1: Initial data for advection problem.

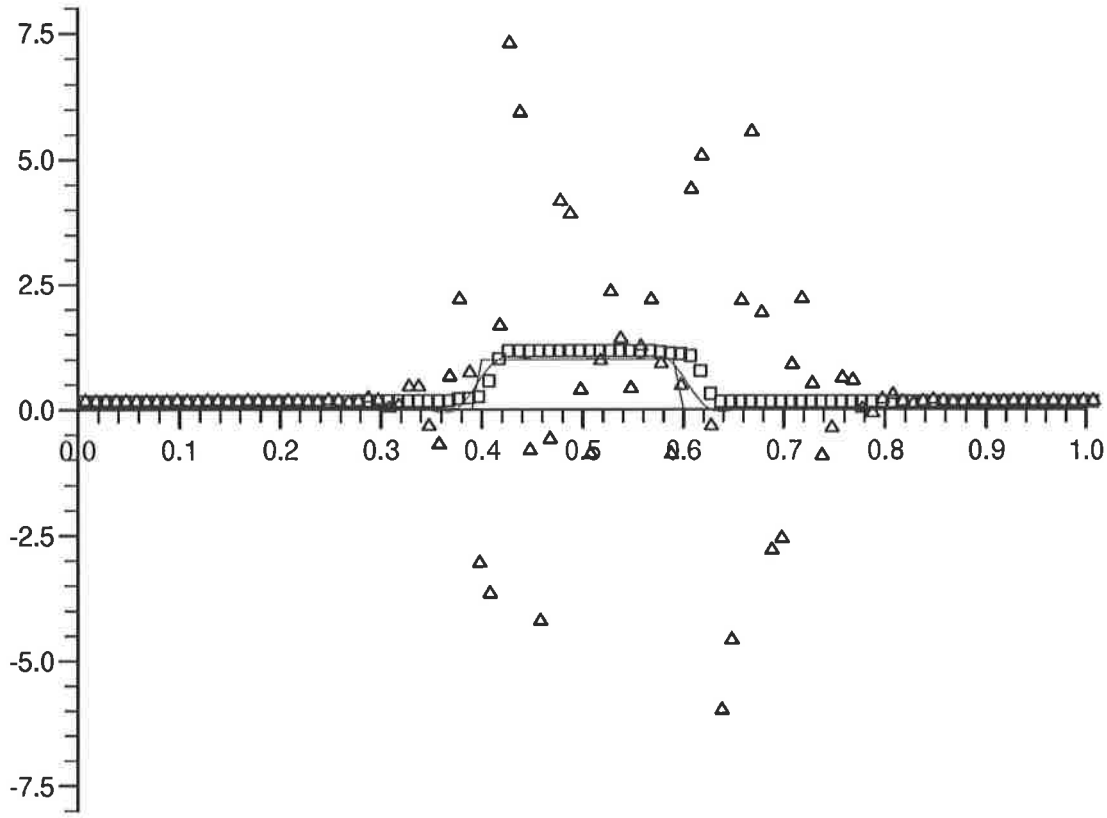

Figure 2: Constant advection after one rotation. Exact and exactly integrated Lagrange-Galerkin solutions—solid lines, Simpson's rule—△, Monotone (Simpson's rule)—□. $\Delta x = 1/100,\ \nu = 5/12$.

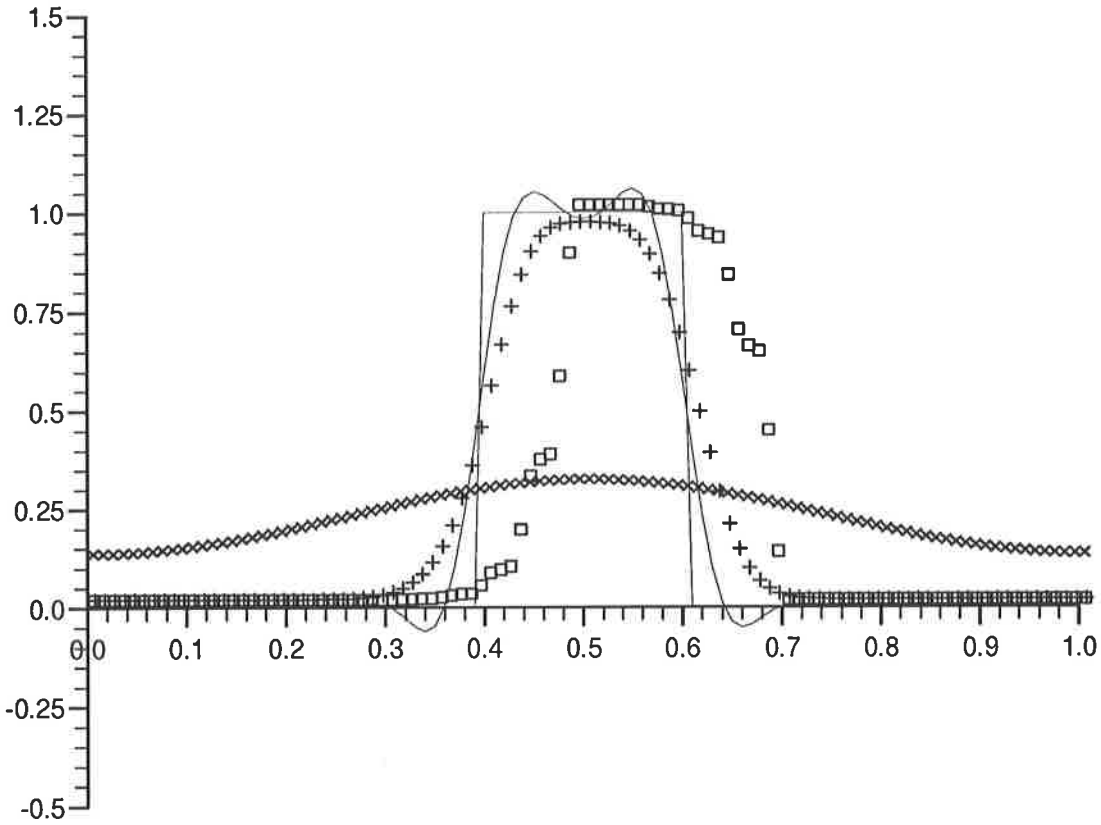

Figure 3: Constant advection after ten rotations. Exact and exactly integrated Lagrange-Galerkin solutions—solid lines, Monotone (Simpson's rule)—□, Van Leer's limiter—+, lumped scheme—×. $\Delta x = 1/100,\ \nu = 5/12.$

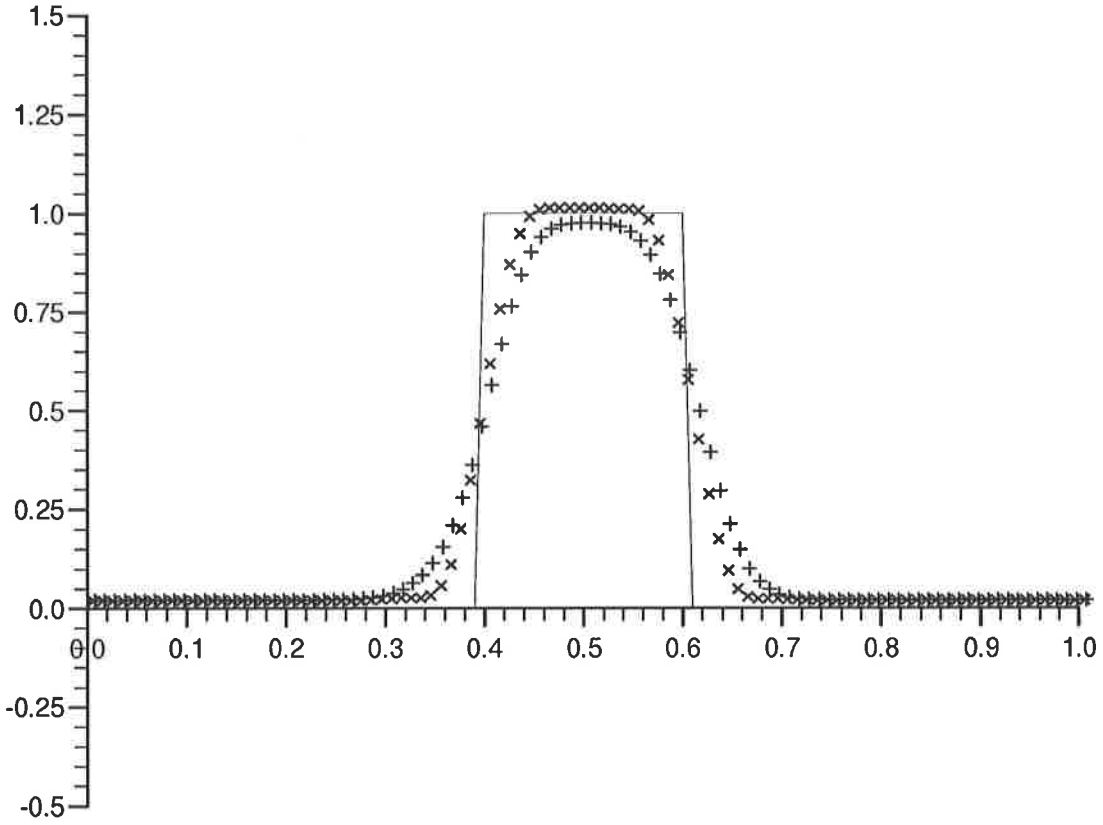

Figure 4: Constant advection after ten rotations. Exact solution—solid line, Monotone (Exact Integration)—×, Van Leer's limiter—+. $\Delta x = 1/100$, $\nu = 5/12$.

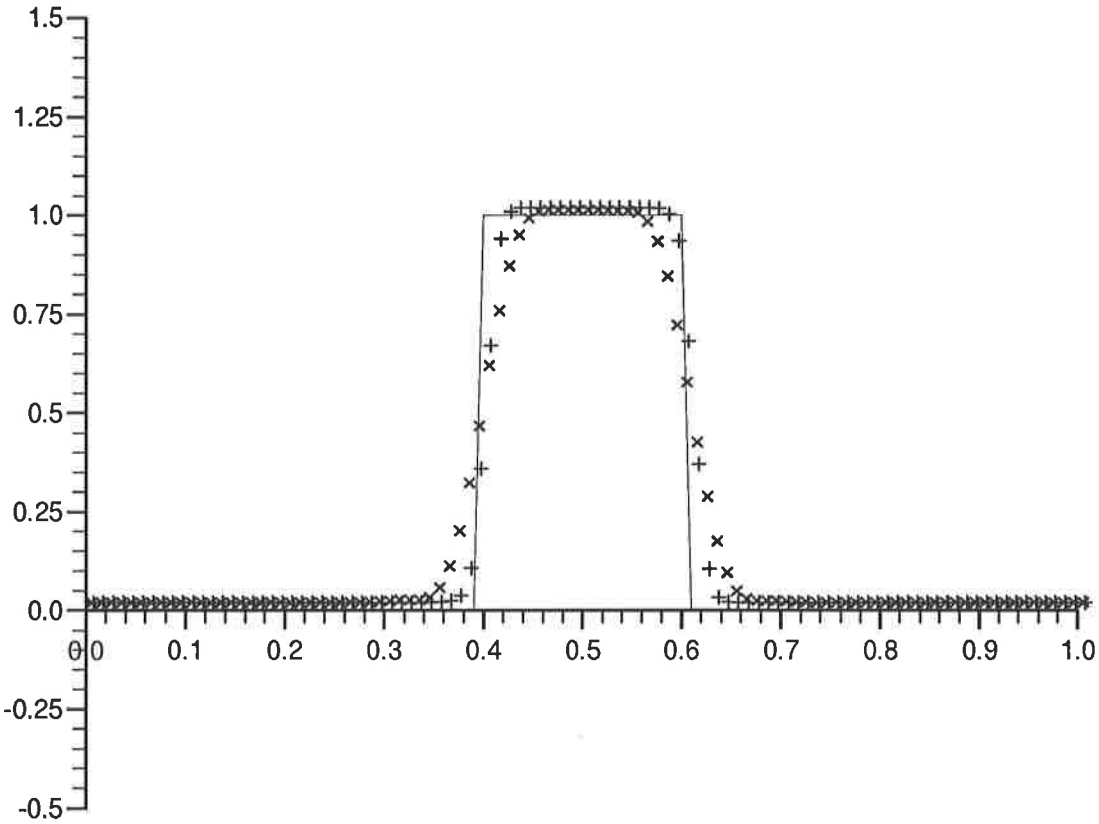

Figure 5: Constant advection after ten rotations. Exact solution—solid line, Monotone (Exact Integration)—×, Superbee limiter—+. $\Delta x = 1/100$, $\nu = 5/12$.

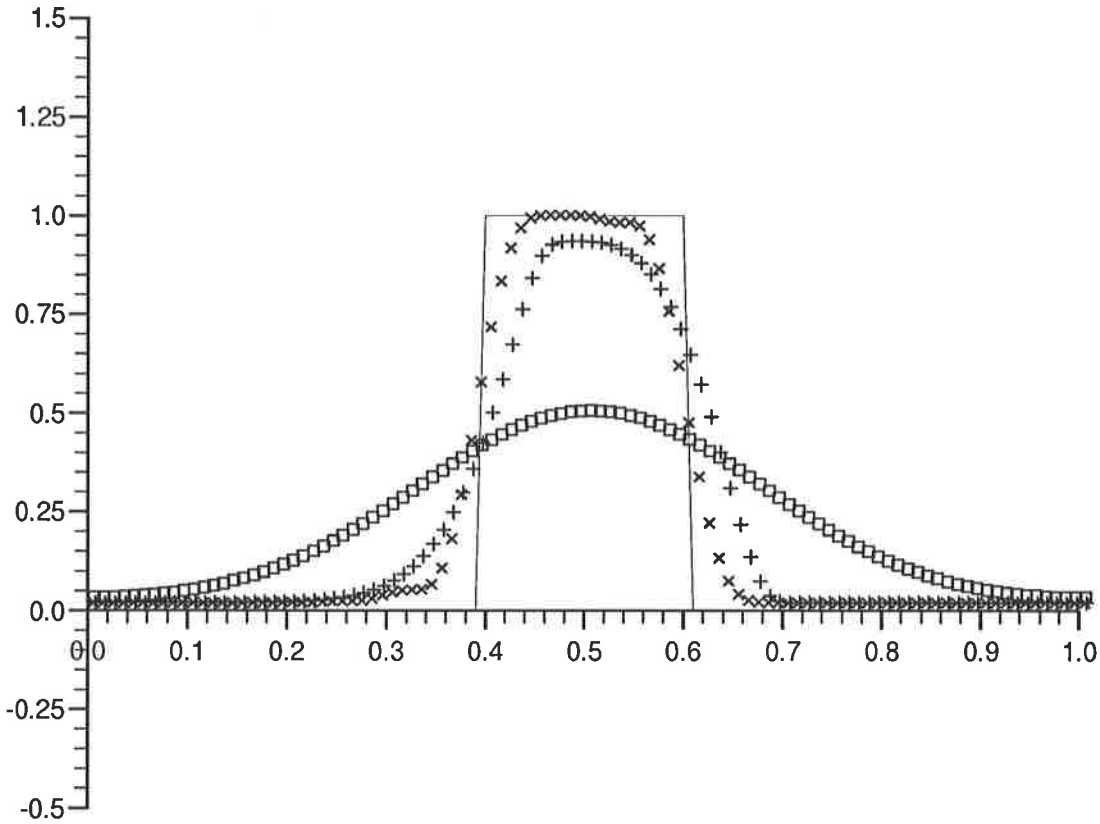

Figure 6: Constant advection after ten rotations. Exact solution—solid line, Monotone (Exact Integration)—×, Monotone (Simpson's rule)—□, Van Leer's limiter—+. $\Delta x = 1/100,\ \nu = 1/12.$

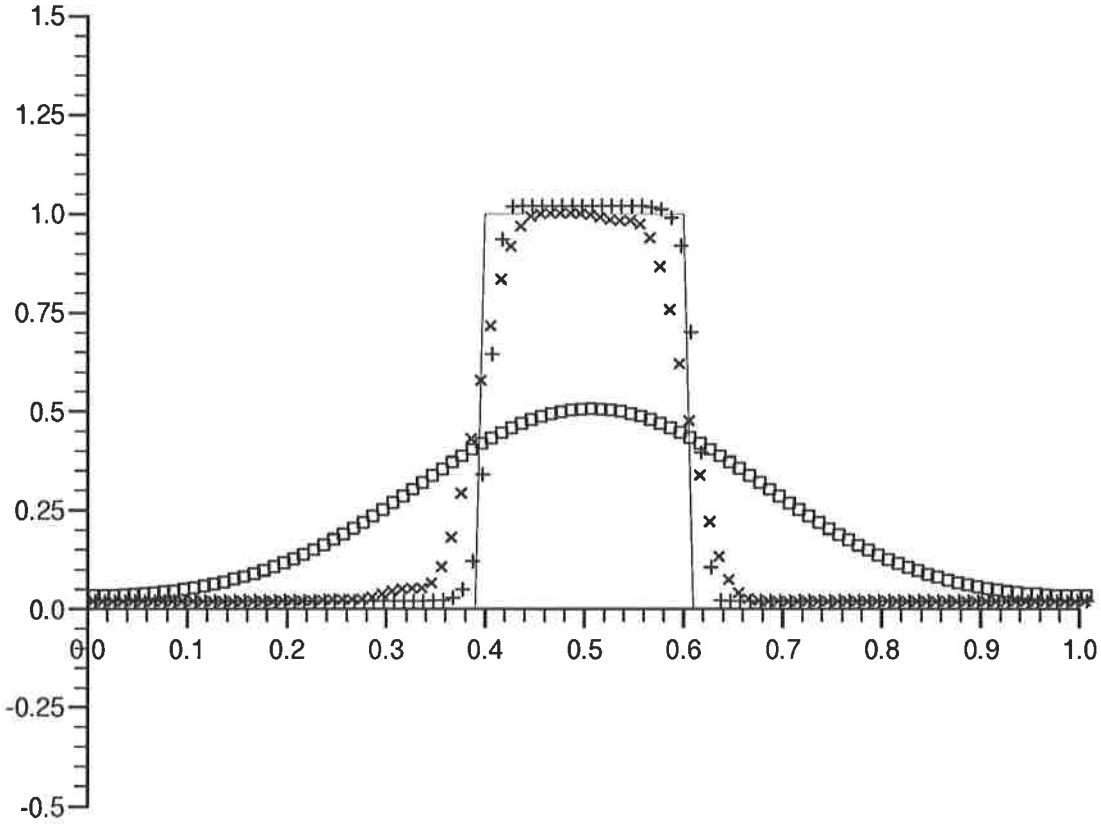

Figure 7: Constant advection after ten rotations. Exact solution—solid line, Monotone (Exact Integration)—×, Monotone (Simpson's rule)—□, Superbee limiter—+. $\Delta x = 1/100,\ \nu = 1/12.$