Discretisation of Unsteady Hyperbolic Conservation Laws

K. W. Morton*

April 13, 2000

Abstract

A basic target algorithm for approximating unsteady hyperbolic conservation laws uses a finite volume formulation in three steps: recovery or reconstruction of a more accurate approximation from a set of cell averages; solution of the conservation law to obtain interface fluxes averaged over a time step; and computation of new cell averages at the new time level. In this paper the target is achieved by bringing together ideas from Brenier's transport collapse operator – using Lin et al's Riemann-Stieltjes interpretation, van Leer's MUSCL algorithm, Colella and Woodward's PPM algorithm and Goodman and LeVeque's flux approximation.

First, second and third order accurate algorithms are developed for non-uniform one-dimensional grids and unstructured triangular meshes. The MUSCL-type scheme in one dimension is proved to be TV-stable right up to the natural CFL limit, that characteristics cross no more than one cell in one time step, and under the least restrictive necessary TVD condition on the recovery stage.

Keywords: finite volume, hyperbolic conservation laws, transport-collapse, recovery or reconstruction techniques, TV-stability.
AMS subject classifications: 65M05, 65M12.

1 Introduction

We consider the approximation by finite difference, finite element or finite volume methods of the system of hyperbolic conservation laws

$$\partial_t \mathbf{w} + \partial_x \mathbf{f}(\mathbf{w}) + \partial_y \mathbf{g}(\mathbf{w}) = \mathbf{0}. \tag{1.1}$$

It will be assumed that preservation of the conservation properties $\frac{d}{dt}\int \mathbf{w}dxdy = 0$ is all important in the discretisation so that, for example, shocks are modelled correctly. However, we will aim for local second or third order accuracy, in time as well as space, where the solution is sufficiently smooth. Though the discussion is limited to the hyperbolic system (1.1), the addition of second order diffusive terms to the schemes will cause little difficulty and will be referred to on occasion. Thus the methods will be extendable to convection-dominated diffusion problems, or to the Navier-Stokes equations rather than the Euler equations of compressible gas dynamics.

*Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, United Kingdom (`morton@comlab.ox.ac.uk`)

1.1 Two classes of schemes

In two dimensions the mesh will generally need to be non-uniform for various reasons (e.g. it may be body-fitted, solution-adaptive, locally refined etc.), so in introducing the basic one-dimensional methods we start with an irregular grid. Then to approximate

$$\partial_t \mathbf{w} + \partial_x \mathbf{f}(\mathbf{w}) = 0 \tag{1.2}$$

by using just two time levels, so as to compute the approximation $\mathbf{W}^{n+1}$ from $\mathbf{W}^n$, we can distinguish two classes of discrete equations. The first takes the form

$$\left(\mathbf{W}_i^{n+1} - \mathbf{W}_i^n\right) h_i + \left(\mathbf{F}_{i+\frac{1}{2}}^{n+\frac{1}{2}} - \mathbf{F}_{i-\frac{1}{2}}^{n+\frac{1}{2}}\right) \Delta t = 0. \tag{1.3}$$

Various algorithms can be, and have been, derived to calculate the *interfacial fluxes* $\{\mathbf{F}_{i\pm\frac{1}{2}}^{n+\frac{1}{2}}\}$; and various interpretations can be placed on the computed quantities $\{\mathbf{W}_i^n\}$. It seems most natural, however, to consider $\mathbf{W}_i^n$ as a *cell average*, over a cell I_i of length h_i centred at x_i as at the bottom of Figure 1(a). Then there is a natural generalisation of the scheme into two dimensions on either a quadrilateral or a triangular mesh - see the top of Figure 1(a), where V_i denotes the measure of the cell; the fluxes become averages over the cell faces or edges.

The alternative class of schemes takes the form

$$\mathbf{W}_i^{n+1} = \mathbf{W}_i^n - \frac{\Delta t}{h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}} \left[D_{i+\frac{1}{2},i}^{n+\frac{1}{2}} \left(\mathbf{F}_{i+1}^n - \mathbf{F}_i^n\right) + D_{i-\frac{1}{2},i}^{n+\frac{1}{2}} \left(\mathbf{F}_i^n - \mathbf{F}_{i-1}^n\right)\right] \tag{1.4}$$

where we very commonly set $\mathbf{F}_i^n = \mathbf{f}(\mathbf{W}_i^n)$, but various algorithms can be devised to define the *distribution matrices* $\{D_{i\pm\frac{1}{2},i}^{n+\frac{1}{2}}\}$. In this case the quantities $\{\mathbf{W}_i^n\}$ are most naturally interpreted as *cell vertex* or nodal values, with a typical cell between nodes x_i and $x_{i\pm1}$ being of length $h_{i\pm\frac{1}{2}}$, as on the bottom of Figure 1(b). The underlying approximation is then most readily regarded as piecewise linear (or perhaps multi-linear in higher dimensions); and the contribution to the conservation integral from cell $I_{i\pm\frac{1}{2}}$ becomes $\frac{1}{2}h_{i\pm\frac{1}{2}}(\mathbf{W}_{i\pm1}^n + \mathbf{W}_i^n)$. To maintain conservation it is therefore necessary to impose the conditions

$$D_{i\pm\frac{1}{2},i}^{n+\frac{1}{2}} + D_{i\pm\frac{1}{2},i\pm1}^{n+\frac{1}{2}} = const., \tag{1.5}$$

the constant matrix being independent of i, which is usually satisfied automatically by the design of the distribution matrices. Again the scheme is readily generalised to quadrilateral or triangular meshes where, as in Figure 1(b), we use Roman subscripts for the vertices and Greek subscripts for the cells.

1.2 Typical examples

Practical schemes of these two classes are typically of second order accuracy on smoothly varying meshes where the solution is also smooth. We will give an example of each for the scalar one-dimensional problem

$$\partial_t u + \partial_x f(u) = 0 \quad \text{or} \quad \partial_t u + a(u)\partial_x u = 0, \tag{1.6}$$

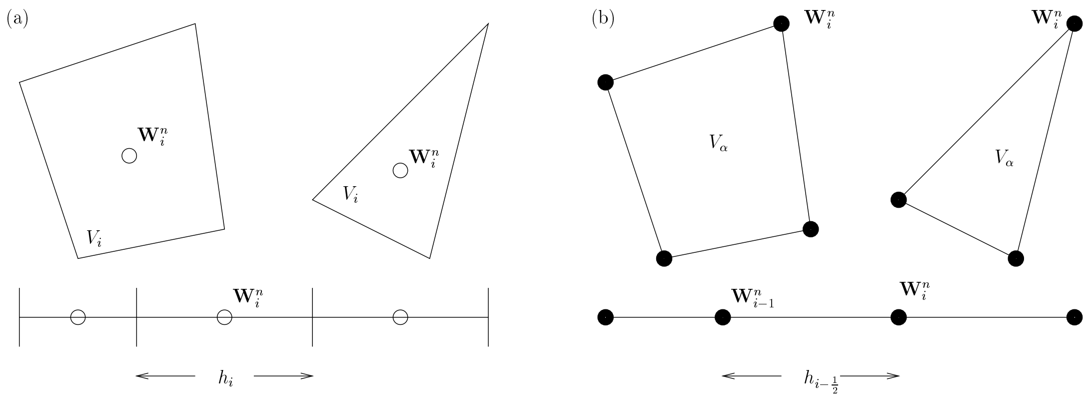

Figure 1: Two basic meshes in one and two dimensions, using (a) the cell average and (b) the cell vertex values.

where $a(u) = f'(u)$.

In the case of (1.3), second order accuracy can be achieved by means of a *recovery* or *reconstruction procedure* that constructs a linear function $\tilde{U}^n$ in each cell from neighbouring cell averages, which is then used to compute the interfacial fluxes. Thus

$$\tilde{U}^n(x) = U_i^n + (x - x_i)S_i^n, \qquad x_i - \tfrac{1}{2}h_i < x < x_i + \tfrac{1}{2}h_i, \tag{1.7}$$

where U_i^n is the cell average and S_i^n the cell slope. These schemes are best regarded as finite volume schemes, developed from the classical Godunov approach [15] and there is a large literature describing algorithms to compute both the slopes and the resultant fluxes, which starts with the pioneering MUSCL schemes of van Leer [47] and is surveyed in the books of LeVeque [26] and Morton [30]. They are generally *upwind schemes* as they depend on the sign of $a(u)$; and they are often *solution-adaptive*, aimed at maintaining a maximum principle, or preserving solution monotonicity, or being TVD (total variation diminishing) or combining a selection of these properties.

We will use the forward and backward difference operators $\Delta_\pm$ given by

$$\Delta_+ z_i := z_{i+1} - z_i, \quad \Delta_- z_i := z_i - z_{i-1}, \tag{1.8}$$

and suppose that $a(U_j^n) \geq 0$ for $j = i-1, i, i+1$. Then one of the simplest such schemes, which is closely related to those derived below, uses a *downwind slope* given by

$$S_i^n = \frac{\Delta_+ U_i^n}{\frac{1}{2}(h_i + h_{i+1})}, \tag{1.9}$$

and fluxes $F_{i\pm\frac{1}{2}}^{n+\frac{1}{2}}$ computed by drawing back the characteristics from $(x_i \pm \frac{1}{2}h_i, t_n + \frac{1}{2}\Delta t)$ to time level t_n. Substituted into the scalar version of (1.3), these give the scheme

$$h_i(U_i^{n+1} - U_i^n) + \Delta t \Delta_- f\left(U_i^n + \tfrac{1}{2}S_i^n(h_i - a_i^n \Delta t)\right) = 0, \tag{1.10}$$

where we have written $a_i^n \equiv a(U_i^n)$.

In the special case of linear, constant coefficient, advection we have $a(u) \equiv a$ and (1.10) reduces to

$$U_i^{n+1} = U_i^n - \frac{a\Delta t}{h_i}\Delta_-\left(\frac{h_{i+1}U_i^n + h_i U_{i+1}^n}{h_{i+1}+h_i}\right) + \frac{(a\Delta t)^2}{h_i}\Delta_-\left(\frac{\Delta_+ U_i^n}{h_{i+1}+h_i}\right). \qquad (1.11)$$

On a uniform mesh this is just the familiar Lax-Wendroff scheme. Moreover, in [32] it is shown to be second order accurate on a completely arbitrary mesh.

To obtain a comparable scheme in the alternative class of (1.4) we choose distribution factors for cell $I_{i-\frac{1}{2}}$ as

$$D^{n+\frac{1}{2}}_{i-\frac{1}{2},i} = 1 + \nu^n_{i-\frac{1}{2}}, \quad D^{n+\frac{1}{2}}_{i-\frac{1}{2},i-1} = 1 - \nu^n_{i-\frac{1}{2}}, \text{ where } \nu^n_{i-\frac{1}{2}} = \frac{\Delta_- f(U_i^n)}{\Delta_- U_i^n}\frac{\Delta t}{h_{i-\frac{1}{2}}}. \qquad (1.12)$$

These clearly satisfy the conservation condition (1.5) and yield the scheme

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}}\left[(1-\nu^n_{i+\frac{1}{2}})\Delta_+ f(U_i^n) + (1+\nu^n_{i-\frac{1}{2}})\Delta_- f(U_i^n)\right]. \qquad (1.13)$$

For linear advection it reduces to

$$U_i^{n+1} = U_i^n - \frac{a\Delta t}{h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}}\left[(U^n_{i+1} - U^n_{i-1}) - a\Delta t\Delta_-\left(\frac{\Delta_+ U_i^n}{h_{i+\frac{1}{2}}}\right)\right], \qquad (1.14)$$

which again becomes the Lax-Wendroff scheme on a uniform mesh. Because (1.13) does not depend on the sign of $a(u)$, these schemes are commonly called *central schemes*. Their stability depends on the choice of the distribution factors or matrices, which introduce *artificial dissipation* terms. Indeed, it is easily checked that the famously stable Lax-Friedrichs scheme can be derived by the choice of distribution factors

$$D^{n+\frac{1}{2}}_{i-\frac{1}{2},i} = 1 + 1/\nu^n_{i-\frac{1}{2}}, \qquad D^{n+\frac{1}{2}}_{i-\frac{1}{2},i-1} = 1 - 1/\nu^n_{i-\frac{1}{2}}, \qquad (1.15)$$

where the *CFL numbers* $\nu^n_{i\pm\frac{1}{2}}$ are those given in (1.12). These schemes also have a large literature, especially in the context of pseudo-time stepping procedures used for solving steady state problems. An influential paper was that of [38], from which much of the terminology has been taken. Further papers covering their development for the Euler and Navier-Stokes equations of compressible gas dynamics include [17], [9], [35]; recent high-resolution schemes on uniform meshes are described in [37] and [21].

Although the example just given of the first class of schemes has been labelled upwind, and that of the second called central, it should be noted that the more distinguishing feature is that the first is based on cell averages and the second on vertex values. Thus the commonly used cell centre finite volume schemes of Jameson – see for instance [20] – are best regarded as falling into the first class although they are based on central differences which are stabilised by judiciously chosen dissipative terms.

There is also a natural duality in the approaches used in the two classes of scheme. In (1.3) and (1.10) the primary consideration is for the evolution of the solution, for instance by using the characteristics or a Taylor expansion in t to calculate interfacial fluxes. On the other hand, the first consideration in (1.4) or (1.13) is assessing the effect of the lack of flux balance, and adjusting the solution accordingly – c.f. the use of the term *fluctuation* by Roe [42] to denote the flux differences, and the extensive development of the schemes to solve steady state problems. Of course, these are merely two aspects of the same problem, as illustrated by the fact that *approximate Riemann solvers* are widely used in both types of scheme.

1.3 Outline of paper

The use of cell vertex or nodal values $\mathbf{W}_i$ in the second class of schemes, illustrated in Figure 1(b), is particularly important when the equations have significant source terms $\mathbf{S}(\mathbf{w}, \mathbf{x}, t)$. However, as we shall consider only the homogeneous equations (1.1), and also emphasise essentially unsteady problems rather than the almost steady problems for which the second class of methods (1.4) is best suited, the emphasis in this paper will be on cell average methods of the form (1.3) and their generalisations into two space dimensions.

We have seen in (1.10) that the use of cell averages as primary variables does not limit the accuracy to first order. Also Leonard [22, 23] has long advocated the use of schemes which have odd orders of accuracy; and it is shown in [32] how such schemes have an extra order of accuracy in their phase speeds, with the dominant error therefore being dissipative. A major objective of the present paper is thus to derive third order accurate schemes by appropriate recovery from cell averages. These are not essentially new schemes, being closely related to the *Piecewise Parabolic Method* (PPM) of Colella and Woodward [8]. Our derivation, however, is very much simpler and hence leads to generalisations for unstructured two-dimensional meshes.

Indeed, the emphasis in the paper is less on the development of new methods and more on a new coherent formulation that enables existing methods to be compared, their best features adopted, and generalisations developed. The finite volume formulation introduced by Godunov [15] provides the starting point, but with the objective of higher order accuracy the emphasis is less on solving Riemann problems at the interfaces and more on using characteristics to carry information from cell interiors to their boundaries. Suppose we denote by E_τ an operator that approximates the evolution of the solution to (1.2), viz: $E(\tau) : \mathbf{w}(t) \mapsto \mathbf{w}(t+\tau)$. Then if a higher order approximation $\tilde{\mathbf{W}}^n$ is constructed from the cell averages $\{\mathbf{W}_i^n\}$ at time level t_n, the interfacial fluxes of (1.3) can be written

$$\mathbf{F}_{i\pm\frac{1}{2}}^{n+\frac{1}{2}} = \frac{1}{\Delta t}\int_0^{\Delta t} \mathbf{f}\left(E_\tau \tilde{\mathbf{W}}^n\right)\Big|_{x_{i\pm\frac{1}{2}}} d\tau \; . \tag{1.16}$$

By using Taylor expansions for E_τ we can put some Lax-Wendroff methods in this form; combining Taylor expansions with Riemann solvers also yields many of the MUSCL-based schemes; and (1.10) provides an example in which the integral is replaced by mid-point quadrature. As in the last case, we shall mainly use characteristics to define E_τ so the schemes are closely related to *characteristic-Galerkin methods.*

Brenier's *transport collapse operator* [5] is a very powerful tool for handling scalar conservation laws, especially in the form used by Lin et al [27, 28] based on a Riemann-Stieltjes integral over the solution graph. The assumed constancy of the solution along each trajectory, followed by a projection, gives an entropy-satisfying approximation which is developed in section 2 to give first, second and third order accurate one-dimensional methods. The first step is to convert the spatial integrals in the original formulation into the time integrals for the fluxes appearing in (1.16). This results in a simple formula which incorporates sonic points and shocks in a very natural way. For the second and third order methods derived in sections 2.2 and 2.3, recovery in each cell is by discontinuous linear and continuous quadratic functions respectively. In these cases the flux integrals are evaluated exactly for approximate fluxes, as in [16]; however, a linear approximation rather than a constant is used for the characteristic speed so that sonic points can be readily included.

Analysis of the schemes is concentrated on showing their *total variation stability* (TV-stability). Although unconditional stability can be achieved by tracing characteristics across more than one cell, the emphasis here is on stability for CFL numbers bounded by unity. The transport collapse operator is TV-stable, so it is the recovery stage at which instability can be introduced. A key theorem that is proved in section 2.2 shows the second order method is TV-stable with only the weakest necessary condition for monotonicity preservation being applied to the discontinuous linear recovery.

In section 3 the methods are developed for two-dimensional problems. After a short section on tensor product meshes, the main emphasis is on unstructured triangular meshes – first the recovery/reconstruction stage, and then the calculation of the interface fluxes along triangular edges. For both the discontinuous linear recovery, described in section 3.2, and the continuous quadratic recovery, in section 3.3, a key step is the recovery of vertex values from neighbouring cell averages. Numerical tests confirming the effectiveness and accuracy of the procedures have been carried out by Dawkins [12]. We apply Simpson's rule to calculate the edge fluxes so that maximum use can be made of the one-dimensional formulae applied to the normal fluxes.

The resultant schemes for scalar laws in one and two dimensions are closely related to and natural extensions of van Leer's MUSCL schemes, Colella and Woodwood's PPM scheme, and the Engquist-Osher schemes of [14] and [39]. Thus a variety of techniques are available for applying the schemes to systems of conservation laws. Indications are given as to how the normal fluxes either side of an edge can be used for this purpose but more detailed development will be presented in a subsequent paper.

2 One dimensional problems

2.1 First order scalar schemes

For the scalar problem (1.6) we have two basic first order schemes. On the mesh at the bottom of Figure 1(a), the piecewise constant approximation U^n can be written as an expansion

$$U^n(x) = \sum_j U_j^n \chi_j(x), \tag{2.1}$$

where $\chi_j(x) = 1$ for $x \in (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ and zero otherwise. With no recovery procedure being applied, the update of the form (1.3) is just

$$\left(U_i^{n+1} - U_i^n\right) h_i + \left(F_{i+\frac{1}{2}}^{n+\frac{1}{2}} - F_{i-\frac{1}{2}}^{n+\frac{1}{2}}\right) \Delta t = 0; \tag{2.2}$$

and if we impose the CFL condition

$$-h_{j-1} \leq a(u)\Delta t \leq h_j \text{ for } u \text{ between } U_{j-1}^n \text{ and } U_j^n \ \forall j, \tag{2.3}$$

only the discontinuities at $x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$ contribute to the calculation of the interfacial fluxes. These are computed by allowing each point on the $[U^n, x]$ graph of (2.1) to be advected along the straight line characteristics $dx/dt = a(U^n(x))$ and then picking up the values at $x_{i\pm\frac{1}{2}}$ for $t_n \leq t \leq t_n + \Delta t$. A detailed derivation is given in Appendix A, under the assumption that we shall make from now on that f is convex and has a single *sonic point* $\bar{u}$ at which $a(\bar{u}) = 0$. Writing $s_i = \operatorname{sgn} a(U_i^n)$ and $\bar{f} = f(\bar{u})$ we obtain at the left-hand interface

$$F_{i-\frac{1}{2}}^{n+\frac{1}{2}} = \tfrac{1}{2}\left[(1-s_i)f\left(U_i^n\right) + (s_i - s_{i-1})\bar{f} + (1+s_{i-1})f\left(U_{i-1}^n\right)\right]. \tag{2.4}$$

When $a(U_{i-1}^n), a(U_i^n)$ and $a(U_{i+1}^n)$ all have the same sign this reduces to the flux in the simple upwind scheme; but (2.4) is also applicable in the presence of sonic points or shocks. Corresponding schemes can be constructed for larger values of Δt when (2.3) does not hold, which is important when local mesh refinement is used, but will not be considered here; see Brenier [5], Childs and Morton [6] and Lin et al [27] for Euler characteristic Galerkin schemes of this type, or LeVeque [25].

Alternatively, on the mesh at the bottom of Fig 1(b), the piecewise linear approximation U^n can be written in terms of finite element basis functions as

$$U^n(x) = \sum_j U_j^n \phi_j(x), \tag{2.5a}$$

where

$$\phi_j(x) = \begin{cases} (x - x_{j-1})/h_{j-\frac{1}{2}} & \text{for } x_{j-1} \leq x \leq x_j \\ (x_{j+1} - x)/h_{j+\frac{1}{2}} & \text{for } x_j \leq x \leq x_{j+1} \\ 0 & \text{otherwise.} \end{cases} \tag{2.5b}$$

The corresponding first order scheme of the form (1.4) is then

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}}\left[(1 - \operatorname{sgn}\nu_{i+\frac{1}{2}}^n)\Delta_+ f\left(U_i^n\right) + (1 + \operatorname{sgn}\nu_{i-\frac{1}{2}}^n)\Delta_- f\left(U_i^n\right)\right], \tag{2.6}$$

where we assume the CFL condition

$$\nu_{j-\frac{1}{2}}^n := \frac{\Delta_- f\left(U_j^n\right)}{\Delta_- U_j^n}\frac{\Delta t}{h_{j-\frac{1}{2}}} \in \left[-\tfrac{1}{2}, +\tfrac{1}{2}\right] \forall j \tag{2.7}$$

is satisfied. It is clear that $\sum \frac{1}{2}(h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}})U_i^n$ is independent of n so that the scheme is conservative, as is (2.2).

Theorem 2.1 *Both schemes (2.2,2.4) and (2.6,2.7) satisfy a local maximum principle under their respective CFL conditions (2.3) and (2.7), and hence are TV-stable.*

Proof Taking the second case first, substitution from (2.7) into (2.6) gives

$$\begin{aligned}(h_{i+\frac{1}{2}}+h_{i-\frac{1}{2}})U_i^{n+1} &= (h_{i+\frac{1}{2}}+h_{i-\frac{1}{2}})U_i^n \\ &\quad -\left[(\nu_{i+\frac{1}{2}}-|\nu_{i+\frac{1}{2}}|)h_{i+\frac{1}{2}}\Delta_+U_i^n+(\nu_{i-\frac{1}{2}}+|\nu_{i-\frac{1}{2}}|)h_{i-\frac{1}{2}}\Delta_-U_i^n\right] \\ &= \alpha_{i+}h_{i+\frac{1}{2}}U_{i+1}^n+\alpha_{i-}h_{i-\frac{1}{2}}U_{i-1}^n+\left[(1-\alpha_{i+})h_{i+\frac{1}{2}}+(1-\alpha_{i-})h_{i-\frac{1}{2}}\right]U_i^n,\end{aligned}$$

where $\alpha_{i+} := |\nu_{i+\frac{1}{2}}|-\nu_{i+\frac{1}{2}} \in [0,1]$ and $\alpha_{i-} := |\nu_{i-\frac{1}{2}}|+\nu_{i-\frac{1}{2}} \in [0,1]$ because of (2.7). Thus U_i^{n+1} is a proper average of the three values $U_{i-1}^n, U_i^n, U_{i+1}^n$ and therefore satisfies a local maximum principle.

In the first case we regroup the terms of (2.4) and (2.2), and write f_i for $f(U_i^n)$, to obtain

$$\begin{aligned}U_i^{n+1} &= U_i^n+\frac{\Delta t}{2h_i}\left\{[(1+s_{i-1})f_{i-1}-(1+s_i)f_i+(s_i-s_{i-1})\bar{f}]\right. \\ &\quad \left.+[(1-s_i)f_i-(1-s_{i+1})f_{i+1}-(s_{i+1}-s_i)\bar{f}]\right\}. \end{aligned} \tag{2.8}$$

The sum in the first square brackets has four possible values depending on (s_{i-1}, s_i) taking the values $(\pm1,\pm1)$: for $(-1,-1)$ it is zero; for $(-1,+1)$ it is

$$2(\bar{f}-f_i)=2|a_{i-}|\,(\bar{u}-U_i^n) \tag{2.9a}$$

where $a_{i-}=a(u)$ for some u between U_i^n and $\bar{u}$, and hence is positive; similarly, for $(+1,-1)$ we have with $a_{i-1+}=a(u)$ for u between U_{i-1}^n and $\bar{u}$,

$$2(f_{i-1}-\bar{f})=2|a_{i-1+}|\left(U_{i-1}^n-\bar{u}\right) \tag{2.9b}$$

and for $(+1,+1)$ with $a_{i-\frac{1}{2}}=a(u)$ for u between U_{i-1}^n and U_i^n,

$$2(f_{i-1}-f_i)=2|a_{i-\frac{1}{2}}|\left(U_{i-1}^n-U_i^n\right). \tag{2.9c}$$

Note that a term in $(\bar{u}-U_i^n)$ is contributed to the update only when $s_i=+1$. In the same way, the second square bracketed sum in (2.8) contributes a term $(\bar{u}-U_i^n)$, $(U_{i+1}^n-\bar{u})$ or $(U_{i+1}^n-U_i^n)$ with a non-negative coefficient, and with the $(\bar{u}-U_i^n)$ contribution occurring only when $s_i=-1$; and in all cases the condition (2.3) ensures the coefficients are less than one. Hence (2.8) gives $U_i^{n+1}-\bar{u}$ as a proper average of $U_{i-1}^n-\bar{u}, U_i^n-\bar{u}$ and $U_{i+1}^n-\bar{u}$.

These manipulations also show that both schemes can be put in the form

$$U_i^{n+1}=U_i^n-C_{i-\frac{1}{2}}\Delta_-U_i^n+D_{i+\frac{1}{2}}\Delta_+U_i^n \tag{2.10}$$

with the coefficients satisfying the conditions

$$0\le C_{i-\frac{1}{2}}, \qquad 0\le D_{i-\frac{1}{2}}, \qquad 0\le C_{i-\frac{1}{2}}+D_{i-\frac{1}{2}}\le 1, \tag{2.11}$$

where the last condition follows from the CFL stability limits given in (2.3) and (2.7). Harten [18] showed that these are sufficient conditions for any three-point difference scheme to be TVD; thus they are TV-stable – see also LeVeque [26]. □

Moreover, it is shown in Appendix A that the scheme (2.8) is precisely the Engquist-Osher scheme on a uniform mesh (see [14]),

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{h}\left[\Delta_- f^+\left(U_i^n\right) + \Delta_+ f^-\left(U_i^n\right)\right], \tag{2.12}$$

where, with $a^+(u) := \max(0, a(u)),\ a^-(u) := \min(0, a(u))$

$$f^+(u) = \int^u a^+(v)dv, \quad f^-(u) = \int^u a^-(v)dv. \tag{2.13}$$

In this uniform mesh case, one can see that there is indeed a close similarity between the two schemes. However, the second scheme (2.6) needs an "entropy fix" in the case of a sonic point occurring in the flow; that is, a "backward facing shock" (such as $u^0(x) = -1$ for $x < 0$, $u^0(x) = +1$ for $x > 0$ for the inviscid Burgers equation, where $f(u) = \frac{1}{2}u^2$) will not decay as it should without some modification to the scheme.

For these reasons, in addition to those already mentioned, we shall concentrate on the first type of scheme in the remaining sections of the paper.

2.2 Discontinuous linear recovery and second order schemes

As already remarked in section 1.2, van Leer [46, 47] was the first to use discontinuous linear approximations to generate a second order version of Godunov's method with his *Monotone Upstream-centred Schemes for Conservation Laws.* However, we shall present these MUSCL schemes and other higher order schemes, in the PERU framework

$$U^{n+1} = P_h E_\Delta R_h U^n. \tag{2.14}$$

Here P_h is the L_2 projection onto the finite element space in which U^n lies, in most cases the piecewise constants of (2.1); $R_h : U^n \mapsto \tilde{U}^n$ is a recovery or reconstruction operation into a space providing a higher order approximation; and E_Δ is an approximation to the evolution operator $E(\Delta t) : u(t) \mapsto u(t + \Delta t)$ for the differential equation. Moreover, we will use a finite volume formulation in the same form as (2.2) which thus encompasses a large number of possible discrete schemes.

The recovery stage

In all cases, we shall impose on the recovery operation the condition

$$P_h R_h U \equiv P_h \tilde{U} = U, \qquad \text{i.e.} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \tilde{U}(x)dx = h_i U_i, \tag{2.15}$$

which is termed *conservation in the mean* by Barth [2]. In the direct form (2.14) of a PERU scheme, this condition is required to ensure overall conservation in the scheme. But this is not obviously required in the finite volume formulation, which we shall write as

$$\left(U_i^{n+1} - U_i^n\right) h_i + \left(\tilde{F}_{i+\frac{1}{2}}^{n+\frac{1}{2}} - \tilde{F}_{i-\frac{1}{2}}^{n+\frac{1}{2}}\right) \Delta t = 0, \tag{2.16}$$

where the interfacial fluxes can be written as

$$\tilde{F}^{n+\frac{1}{2}}_{i\pm\frac{1}{2}} = \frac{1}{\Delta t}\int_0^{\Delta t} f\left(E_\tau \tilde{U}^n\right)\Big|_{x_{i\pm\frac{1}{2}}} d\tau \tag{2.17}$$

and E_τ represents an approximate evolution operator through a time τ.

The formulation (2.16) ensures overall conservation whatever the interfacial fluxes are. On the other hand, it is entirely reasonable to impose the *unit CFL condition* of Morton and Parrott [34] on the scheme; that is, for the linear advection equation $u_t + au_x = 0$ on a uniform mesh with $a\Delta t = \Delta x$, the scheme should give the exact result of shifting the solution by one mesh spacing, i.e. $U_i^{n+1} = U_{i-1}^n$. Now, for linear advection it is easy to see that the interfacial fluxes in the unit CFL case are just

$$F^{n+\frac{1}{2}}_{i+\frac{1}{2}} = \frac{a}{\Delta t}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \tilde{U}^n(x)\frac{dx}{a} = \frac{h_i}{\Delta t}\left(P_h\tilde{U}^n\right)_i. \tag{2.18}$$

So (2.16) becomes

$$\left(U_i^{n+1} - U_i^n\right) + \left(P_h\tilde{U}^n\right)_i - \left(P_h\tilde{U}^n\right)_{i-1} = 0$$

and (2.15) is required to ensure that $U_i^{n+1} = U_{i-1}^n$.

Condition (2.15) means that discontinuous linear recovery must be of the form (1.7) and the slopes S_i^n are the only free parameters in the recovered function. As a first step in calculating these slopes, suppose that the underlying function is approximately quadratic in form,

$$u(x) \approx q(x) = a + b(x - x_i) + c(x - x_i)^2 \tag{2.19}$$

and that

$$(P_h q)_j = U_j \quad \text{for } j = i-1, i, i+1. \tag{2.20}$$

Then a, b, c can be determined from U_{i-1}, U_i, U_{i+1}. Suppose also that $\tilde{U}(x) \equiv U_i + (x - x_i)S_i$ is the L_2 best fit to $q(x)$ in the i^{th} cell, from which one readily deduces that

$$S_i = b \quad \text{and} \quad U_i = a + \tfrac{1}{12}ch_i^2. \tag{2.21}$$

By shifting the origin of the quadratic $q(x)$, one can similarly deduce from (2.20) that

$$U_{i\pm1} = a + b(x_{i\pm1} - x_i) + c(x_{i\pm1} - x_i)^2 + \tfrac{1}{12}ch_{i\pm1}^2. \tag{2.22}$$

So, finally, (2.21) and (2.22) give the result

$$S_i = b \equiv \frac{(2h_{i-1} + h_i)D_+U_i + (h_i + 2h_{i+1})D_-U_i}{2(h_{i-1} + h_i + h_{i+1})} \tag{2.23}$$

where

$$D_\pm U_i = \frac{\Delta_\pm U_i}{\frac{1}{2}(h_i + h_{i\pm1})}. \tag{2.24}$$

Note that on a uniform mesh this reduces to the central difference formula

$$S_i = b \equiv D_0 U_i \equiv \tfrac{1}{2}\left(D_+ U_i + D_- U_i\right) \equiv \frac{U_{i+1} - U_{i-1}}{2h}; \tag{2.25}$$

note also that fitting a quadratic through the three values U_{i-1}, U_i, U_{i+1} does not give the same slope at x_i as (2.23) except when the mesh is uniform.

The formula (2.23) is probably the most accurate that it is reasonable to use, and the equally weighted average $\frac{1}{2}(D_+ U_i + D_- U_i)$ is an acceptable alternative on most meshes. In section 2.3, where divided differences are used to derive the PPM recovery formula, it is shown that the difference between (2.23) and this alternative is just $(h_{i-1} - h_{i+1})N_i^{(2)}$, where $N_i^{(2)}$ is a second divided difference; and a higher order approximation is also given there.

However, what is more important than choosing between these alternatives is that there are several natural constraints that should be placed on the recovered slopes to reflect key properties satisfied by solutions of the conservation law. The first is *monotonicity-preservation*: if U is monotone then $\tilde{U}$ should also be monotone, in the same sense. For discontinuous linear recovery, this requires that $D_- U_j \geq 0 \; \forall j$ implies that

$$\text{(i) } S_j \geq 0 \; \forall j; \quad \text{(ii) } S_{j-1} h_{j-1} + S_j h_j \leq (h_{j-1} + h_j) D_- U_j \; \forall j. \tag{2.26}$$

The first is obvious and the second arises from considering the sign of the discontinuity at $x_{j-\frac{1}{2}}$.

Lemma 2.2 *To ensure that a discontinuous linear recovery is monotonicity preserving, a sufficient condition is that $D_- U_j \geq 0 \; \forall j$ implies*

$$0 \leq S_j \leq \min\left(D_- U_j, D_+ U_j\right) \; \forall j; \tag{2.27}$$

and a corresponding necessary condition is that

$$0 \leq S_j h_j \leq 2min\left(\Delta_- U_j, \Delta_+ U_j\right) \; \forall j. \tag{2.28}$$

Proof Since (2.27) implies both $S_j \leq D_- U_j$ and $S_{j-1} \leq D_- U_j$, it clearly implies (2.26). Also, because we need $S_{j-1} \geq 0$ and $S_j \geq 0$, (2.26) implies that $S_j h_j \leq 2\Delta_- U_j$ and $S_{j-1} h_{j-1} \leq 2\Delta_- U_j$ so that (2.28) follows. □

To see when the slopes given by (2.23) violate the monotonicity-preservation condition, suppose the mesh is uniform. Then substitution of (2.25) into (2.26) leads to the requirement

$$\tfrac{1}{2}\left(\Delta_- U_{j-1} + \Delta_- U_j\right) + \tfrac{1}{2}\left(\Delta_- U_j + \Delta_- U_{j+1}\right) \leq 2\Delta_- U_j$$

$$\text{i.e., } \delta^2 \Delta_- U_j \leq 0 \; \forall j. \tag{2.29}$$

If a set of monotone increasing data does not satisfy this condition, the recovery by (2.25) will lead to interface discontinuities of the wrong sign.

In practice, then, recovered slopes need to be *limited* in value according to the data. For example, van Leer [46] has used the necessary condition (2.28), which has the nice interpretation

$$U_{j-1} \leq U_j - \tfrac{1}{2} S_j h_j \leq U_j + \tfrac{1}{2} S_j h_j \leq U_{j+1} \; \forall j \tag{2.30}$$

for monotone increasing data $\{U_j\}$, so that it can be applied independently in each cell.

A second important situation occurs at extrema, for instance at a maximum where $\Delta_- U_i > 0, \Delta_+ U_i < 0$. To ensure that $\max \tilde{U} \leq \max U$, that is the recovery operation itself is TVD, clearly requires that $S_i = 0$. This is usually incorporated explicitly into slope recovery algorithms, of which we quote three. Defining $\Sigma_i = \mathrm{sgn} \Delta_\pm U_i$ when $\Delta_+ U_i \Delta_- U_i \geq 0$, and $\Sigma_i = 0$ when $\Delta_+ U_i \Delta_- U_i < 0$, the *minmod* slope limiter proposed by van Leer [45] is based on the sufficient condition (2.27) and gives

$$\text{(minmod)} \qquad S_i = \Sigma_i \min\left(|D_- U_i|, |D_+ U_i|\right). \tag{2.31}$$

On the other hand, the slope limiter used by van Leer in his MUSCL scheme [47], and also by Davis [11], is based on the necessary condition (2.28) and gives

$$\text{(MUSCL)} \qquad S_i = \Sigma_i \min\left\{2|D_- U_i|, \tfrac{1}{2}\left(|D_- U_i| + |D_+ U_i|\right), 2|D_+ U_i|\right\} \tag{2.32}$$

so that the equally weighted average slope of (2.25) is used unless the ratio of $D_+ U_i / D_- U_i$ is outside the range $[1/3, 3]$. Finally, a limiter which depends more smoothly on $D_\pm U_i$ is given by the *harmonic mean*,

$$\text{(harmonic)} \qquad S_i = \Sigma_i \frac{2(D_i U_i)(D_+ U_i)}{D_- U_i + D_+ U_i}. \tag{2.33}$$

Two further comments are in order at this point. Clearly, at a maximum the requirement that $S_i = 0$ to ensure that the R_h operator is TVD, is stronger than necessary; for the overall TVD property would hold for the algorithm if $R_h P_h$ were TVD for the class of functions to which it is applied, that is functions of the form $E_\Delta \tilde{U}$ in the scheme (2.14). If these were taken to be quadratic in form, like (2.19), the maximum before the application of $R_h P_h$ would be $a - b^2/2c$; and for this not to be exceeded by the recovered function, on a uniform mesh would require

$$|S_i| \leq \frac{1}{3} \frac{\left[(D_- U_i)^2 + (D_- U_i)(D_+ U_i) + (D_+ U_i)^2\right]}{D_- U_i - D_+ U_i}. \tag{2.34}$$

This could be achieved by setting

$$\Sigma_i = \frac{D_- U_i + D_+ U_i}{2\left(|D_- U_i| + |D_+ U_i|\right)} \tag{2.35}$$

when $(D_- U_i)(D_+ U_i) < 0$ in the MUSCL limiter, but numerical experiments give little support for introducing such a modification.

The second comment is that the application of $P_h E_\Delta$ after the recovery operation usually has a smoothing effect, so that it may be adequate for the slopes to satisfy the necessary conditions for monotonic preservation in Lemma 2.2, rather than the sufficient conditions. We will give examples of this situation shortly – see Theorem 2.3.

Numerical fluxes and the update stage

For simplicity, we ensure that the flux at each interface is affected only by the cells on either side, by assuming that all the relevant CFL nos. are less than or equal to unity; the detailed specific bounds on Δt are given in Appendix A as (A.16a,b). Then, in general, an implicit equation needs to be solved to find the upper limit in each of the flux integrals in (2.17), e.g. x^* or $u^* \equiv \tilde{U}^n(x^*)$ such that $x^* + a(u^*)\Delta t = x_{i+\frac{1}{2}}$, and then the integral has to be approximated appropriately. We use a linear interpolation of the characteristic speed $a(\cdot)$ to achieve this. The result is an update algorithm of the form (2.16), with recovered interfacial fluxes very similar to those in the unrecovered case, given by (2.4). Moreover, TVD properties of the update are established in Theorems 2.3 and 2.6 below which extend those achievable with the MUSCL scheme. The algorithms and the theoretical results relate to and extend those of Goodman and LeVeque [16], who used only a piecewise constant approximation to $a(\cdot)$.

In each cell I_j we denote by $\tilde{U}_j^\pm$ the end values $U_j^n \pm \frac{1}{2}h_j S_j^n$, set $s_j^\pm := \operatorname{sgn} a(\tilde{U}_j^\pm)$ and denote by $A_j(\tilde{U}^n)$ the linear interpolant between $a(\tilde{U}_j^+)$, $a(\tilde{U}_j^-)$ and $a(\bar{u}) \equiv 0$ if the sonic point occurs in the cell, i.e. if $s_j^+ s_j^- = -1$. The slope of $A_j(\cdot)$ at each end point is thus given by

$$m_j^\pm = \frac{a(\tilde{U}_j^+) - a(\tilde{U}_j^-)}{\tilde{U}_j^+ - \tilde{U}_j^-} \quad \text{or} \quad \frac{a(\tilde{U}_j^\pm)}{\tilde{U}_j^\pm - \bar{u}} \tag{2.36}$$

according to whether $s_j^+ s_j^- = +1$ or -1 respectively; note that the convexity of $f(\cdot)$ implies that $m_j^\pm > 0$. Now suppose that $s_j^+ = +1$. Then it is shown in Appendix A that the characteristic speed at the limit point x^*, from which the characteristic just reaches the interface $x_{j+\frac{1}{2}}$, is given by

$$A_j^{*+} = \frac{a(\tilde{U}_j^+)}{1 + m_j^+ S_j^n \Delta t}; \tag{2.37}$$

and the corresponding average flux, obtained from the integral (2.17), can be expressed as

$$\tilde{F}_j^+ = \frac{A_j^{*+} f(\tilde{U}_j^+) + a(\tilde{U}_j^+) f(\tilde{U}_j^+ - S_j^n A_j^{*+} \Delta t)}{A_j^{*+} + a(\tilde{U}_j^+)}. \tag{2.38}$$

At the other end of the interval, if $s_j^- = -1$ we need a corresponding pair

$$A_j^{*-} = \frac{a(\tilde{U}_j^-)}{1 + m_j^- S_j^n \Delta t} \tag{2.39}$$

and

$$\tilde{F}_j^- = \frac{A_j^{*-} f(\tilde{U}_j^-) + a(\tilde{U}_j^-) f(\tilde{U}_j^- - S_j^n A_j^{*-} \Delta t)}{A_j^{*-} + a(\tilde{U}_j^-)}. \tag{2.40}$$

It is easily seen that the CFL conditions in (A.16) ensure that the denominators in (2.37) and (2.39) are always positive, ensuring that $A_j^{*\pm}$ has the same sign as $a(\tilde{U}_j^\pm)$; and that $|A_j^{*\pm}|$ is greater than or smaller than $|a(\tilde{U}_j^\pm)|$ according to whether the slope S_j^n is negative or positive.

These fluxes can be combined with a possible contribution from the sonic point occurring at the interface discontinuity to give, in the same way as in (2.4) for the unrecovered case, an interface flux at $x_{i-\frac{1}{2}}$ of

$$\tilde{F}^{n+\frac{1}{2}}_{i-\frac{1}{2}} = \tfrac{1}{2}\left[(1-s_i^-)\tilde{F}_i^- + (s_i^- - s_{i-1}^+)\bar{f} + (1+s_{i-1}^+)\tilde{F}_{i-1}^+\right]. \tag{2.41}$$

Combining this with a similar flux at $x_{i+\frac{1}{2}}$ gives the update (2.16) which we write in the form

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{h_i}\left[\tilde{F}^{n+\frac{1}{2}}_{i+\frac{1}{2}} - \tilde{F}^{n+\frac{1}{2}}_{i-\frac{1}{2}}\right] \tag{2.42}$$

ready for the analysis of its TVD properties.

However, we pause first to compare the fluxes (2.38) and (2.40) with alternatives, some of which are well known. It is shown in Appendix A that, when $a(\cdot)$ is linear in cell I_j, a completely equivalent form of the fluxes is given by

$$\tilde{F}_j^{\pm} = f(\tilde{U}_j^{\pm}) - \tfrac{1}{2}S_j^n a(\tilde{U}_j^{\pm})A_j^{*\pm}\Delta t. \tag{2.43}$$

In practice there is probably little to choose between the two forms, but that of (2.38) and (2.40) has advantages in the analysis below. When $a(\cdot)$ is taken to be locally constant, (2.43) becomes

$$\tilde{F}_j^{\pm} = f(\tilde{U}_j^{\pm}) - \tfrac{1}{2}S_j^n\left[a(\tilde{U}_j^{\pm})\right]^2\Delta t, \tag{2.44}$$

which is the form given by Goodman and LeVeque [16]. If the $O(\Delta t)$ term is neglected, so that $\tilde{F}_j^{\pm} = f(\tilde{U}_j^{\pm})$, we obtain a first order form of the MUSCL scheme; while setting

$$\tilde{F}_j^{\pm} = f(\tilde{U}_j^{\pm} - \tfrac{1}{2}(\Delta t/h_j)[f(\tilde{U}_j^{+}) - f(\tilde{U}_j^{-})]) \tag{2.45}$$

gives a commonly used second order form – see [48].

In the first theorem below we use an analysis that can also be applied to these alternative fluxes, though with less sharp results; and in the second an analysis which can also be applied to the piecewise quadratic recovery that we shall consider next.

Theorem 2.3 *Suppose that the CFL conditions (A.16) are satisfied, and that the linear recovery procedure satisfies the necessary TVD conditions of (2.28), namely*

$$0 \le \left(h_j S_j^n/\Delta_- U_j^n\right) \le 2 \quad \text{and} \quad 0 \le \left(h_j S_j^n/\Delta_+ U_j^n\right) \le 2 \quad \forall j. \tag{2.46}$$

Then the scheme defined by (2.37)–(2.42) is TV-stable if further CFL conditions, given in terms of the CFL nos. $\lambda^{\pm}$, $\mu^{\pm}$ and $\nu^{\pm}$ defined below in (2.50) and (2.53) as

$$2\lambda^+ - (\nu^+)^2 + (\nu^-)^2 \le 1 \tag{2.47a}$$

$$2\lambda^- - (\mu^-)^2 + (\mu^+)^2 \le 1 \tag{2.47b}$$

$$[\lambda^+ - (\nu^+)^2 + (\nu^-)^2] + [\lambda^- - (\mu^-)^2 + (\mu^+)^2] \le 1, \tag{2.47c}$$

are satisfied for every pair of cells.

Proof As in the unrecovered case, we put the scheme in the form $U_i^{n+1} = U_i^n - C_{i-\frac{1}{2}}\Delta_- U_i^n + D_{i+\frac{1}{2}}\Delta_+ U_i^n$ of (2.10) and show that Harten's conditions (2.11) for the coefficients $C_{i-\frac{1}{2}}$ and $D_{i-\frac{1}{2}}$ are satisfied. The key is to define $C_{i-\frac{1}{2}}$ in terms of the right-moving waves and $D_{i-\frac{1}{2}}$ in terms of the left-moving waves. This means that, as in the rearrangement of (2.8), the flux differences in (2.41, 2.42) are rearranged to give

$$C_{i-\frac{1}{2}}\Delta_- U_i^n = \frac{1}{2}\frac{\Delta t}{h_i}\left[(1+s_i^+)\tilde{F}_i^+ - (s_i^+ - s_{i-1}^+)\bar{f} - (1+s_{i-1}^+)\tilde{F}_{i-1}^+\right] \qquad (2.48a)$$

while, after the shift of i and change of sign implied by the form (2.10), we have

$$D_{i-\frac{1}{2}}\Delta_- U_i^n = -\frac{1}{2}\frac{\Delta t}{h_{i-1}}\left[(1-s_i^-)\tilde{F}_i^- + (s_i^- - s_{i-1}^-)\bar{f} - (1-s_{i-1}^-)\tilde{F}_{i-1}^-\right]. \qquad (2.48b)$$

In each formula there are four sign combinations to be considered, giving a total of sixteen in all.

For $s_i^+ = s_{i-1}^+ = 1$, (2.48a) with (2.38) gives

$$C_{i-\frac{1}{2}}\Delta_- U_i^n = \frac{\Delta t}{h_i}\Delta_-\left[\tilde{F}_i^+ = f(\tilde{U}_i^+) - a(\tilde{U}_i^+)\frac{f(\tilde{U}_i^+) - f(U_i^{*+})}{a(\tilde{U}_i^+) + A_i^{*+}}\right] \qquad (2.49)$$

where we have written U_i^{*+} for $\tilde{U}_i^+ - S_i^n A_i^{*+}\Delta t$. Now we define, in this case,

$$\lambda^+ := \frac{\Delta t}{h_i}\frac{\Delta_- f(\tilde{U}_i^+)}{\Delta_-\tilde{U}_i^+}, \quad \nu_1^+ := a(\tilde{U}_i^+)\frac{\Delta t}{h_i}, \quad \mu_1^+ := a(\tilde{U}_{i-1}^+)\frac{\Delta t}{h_i} \qquad (2.50a)$$

and introduce

$$a_i^{*+} := \frac{A_i^{*+}}{\frac{1}{2}[a(\tilde{U}_i^+) + A_i^{*+}]}\frac{f(\tilde{U}_i^+) - f(U_i^{*+})}{\tilde{U}_i^+ - U_i^{*+}}, \qquad (2.50b)$$

and similarly a_{i-1}^{*+}, to give

$$\nu_2^+ := a_i^{*+}\frac{\Delta t}{h_i}, \qquad \mu_2^+ := \frac{a_{i-1}^{*+}\Delta t}{h_{i-1}}. \qquad (2.50c)$$

We note here that, to complete the definition of these CFL nos., when $s_i^+ = +1$ but $s_{i-1}^+ = -1$ we replace $\tilde{F}_{i-1}^+$ by $\bar{f}$ in (2.49); when $s_i^+ = -1$ but $s_{i-1}^+ = +1$ we replace $\tilde{F}_i^+$ by $\bar{f}$; and when $s_i^+ = s_{i-1}^+ = -1$ we have $C_{i-\frac{1}{2}} = 0$. In the first case, $\bar{u}$ replaces $\tilde{U}_{i-1}^+$ in the definition of λ^+ and μ_1^+, μ_2^+ are set to zero; and in the second case, $\bar{u}$ replaces $\tilde{U}_i^+$ in the definition of λ^+ while ν_1^+ and ν_2^+ are set to zero. In all other cases we set these CFL numbers to zero, so they are always non-negative. We also write

$$\nu^+ := (\nu_1^+\nu_2^+)^{\frac{1}{2}}, \qquad \mu^+ := (\mu_1^+\mu_2^+)^{\frac{1}{2}}; \qquad (2.50d)$$

and will henceforth carry all four cases forward together.

Now we can rewrite the above expression (2.49) to give, when $s_i^+ = s_{i-1}^+ = 1$,

$$C_{i-\frac{1}{2}} \Delta_- U_i^n = \lambda^+ \Delta_- \tilde{U}_i^+ - \nu_1^+ \nu_2^+ (\tfrac{1}{2} h_i S_i^n) + \mu_1^+ \mu_2^+ (\tfrac{1}{2} h_{i-1} S_{i-1}^n),$$

and hence

$$C_{i-\frac{1}{2}} = \lambda^+ + (\lambda^+ - \nu_1^+ \nu_2^+) \frac{\frac{1}{2} h_i S_i^n}{\Delta_- U_i^n} - (\lambda^+ - \mu_1^+ \mu_2^+) \frac{\frac{1}{2} h_{i-1} S_{i-1}^n}{\Delta_- U_i^n}; \tag{2.51a}$$

when $s_i^+ = +1$ but $s_{i-1}^+ = -1$, this reduces as described above to

$$C_{i-\frac{1}{2}} = \lambda^+ \frac{U_i^n - \bar{u}}{\Delta_- U_i^n} + (\lambda^+ - \nu_i^+ \nu_2^+) \frac{\frac{1}{2} h_i S_i^n}{\Delta_- U_i^n}; \tag{2.51b}$$

when $s_i^+ = -1$, but $s_{i-1}^+ = +1$, we have

$$C_{i-\frac{1}{2}} = \lambda^+ \frac{\bar{u} - U_{i-1}^n}{\Delta_- U_i^n} - (\lambda^+ - \mu_1^+ \mu_2^+) \frac{\frac{1}{2} h_{i-1} S_{i-1}^n}{\Delta_- U_i^n}, \tag{2.51c}$$

and $C_{i-\frac{1}{2}} = 0$ in the remaining case.

Either expression for the fluxes, (2.38 or (2.43), can be used for obtaining upper bounds for $C_{i-\frac{1}{2}}$, but to show that $C_{i-\frac{1}{2}} \geq 0$ it is simpler to work with the averaged form (2.38) for $\tilde{F}_i^+$ and $\tilde{F}_{i-1}^+$, as above and is done in the proof of the following result in Appendix B.

Lemma 2.4 *Under the CFL conditions (A.16) we have*

$$C_{i-\frac{1}{2}} \geq 0 \quad \forall i. \tag{2.52}$$

Proof See Appendix B. □

Analogous definitions to those of (2.50) are introduced for the left-going waves, namely we first have

$$\lambda^- := -\frac{\Delta t}{h_{i-1}} \frac{[(1 - s_i^-)(f(U_i^-) - \bar{f}) + (1 - s_{i-1}^-)(\bar{f} - f(U_{i-1}^-))]}{[(1 - s_i^-)(U_i^- - \bar{u}) + (1 - s_{i-1}^-)(\bar{u} - U_{i-1}^-)]} \tag{2.53a}$$

which covers all the sign cases and always gives a nonnegative ratio; when $s_i^- = -1$ we set

$$\nu_1^- := -a(\tilde{U}_i^-) \frac{\Delta t}{h_i}, \quad \nu_2^- = -a_i^{*-} \frac{\Delta t}{h_i}; \tag{2.53b}$$

and when $s_{i-1}^- = -1$ we set

$$\mu_1^- := -a(\tilde{U}_{i-1}^-) \frac{\Delta t}{h_{i-1}}, \quad \mu_2^- := -a_{i-1}^{*-} \frac{\Delta t}{h_{i-1}}, \tag{2.53c}$$

where we have introduced

$$a_j^{*-} := \frac{A_j^{*-}}{\frac{1}{2}[a(\tilde{U}_j^-) + A_j^{*+}]} \frac{f(\tilde{U}_j^-) - f(\tilde{U}_j^{*-})}{\tilde{U}_j^- - \tilde{U}_j^{*-}} \tag{2.53d}$$

for $j = i, i-1$ as required. Otherwise we set $\nu_1^- = \nu_2^- = 0$ and/or $\mu_1^- = \mu_2^- = 0$.

Then, as in (2.51), we have three expressions for $D_{i-\frac{1}{2}}$: for $s_i^- = s_i^- = -1$,

$$D_{i-\frac{1}{2}} = \lambda^- - (\lambda^- - \nu_1^- \nu_2^-) \frac{\frac{1}{2} h_i S_i^n}{\Delta_- U_i^n} + (\lambda^- - \mu_1^- \mu_2^-) \frac{\frac{1}{2} h_{i-1} S_{i-1}^n}{\Delta_- U_i^n}; \tag{2.54a}$$

when $s_i^- = -1$ but $s_{i-1}^- = +1$, we have

$$D_{i-\frac{1}{2}} = \frac{\lambda^- (U_i^n - \bar{u})}{\Delta_- U_i^n} - (\lambda^- - \nu_1^- \nu_2^-) \frac{\frac{1}{2} h_i S_i^n}{\Delta_- U_i^n}; \tag{2.54b}$$

and when $s_{i-1}^- = -1$ but $s_i^- = +1$, we have

$$D_{i-\frac{1}{2}} = \frac{\lambda^- (\bar{u} - U_i^n)}{\Delta_- U_i^n} + (\lambda^- - \mu_1^- \mu_2^-) \frac{\frac{1}{2} h_{i-1} S_{i-1}^n}{\Delta_- U_i^n}. \tag{2.54c}$$

These separate expressions are useful in considering the implications of the TVD conditions (2.46) in the three cases. As in Lemma 2.4 we also have $D_{i-\frac{1}{2}} \geq 0$ when the conditions of (A.16) are satisfied.

We shall show that in all cases, we have the following result.

Lemma 2.5 *Under the CFL conditions of (A.16) and (2.47),*

$$C_{i-\frac{1}{2}} + D_{i-\frac{1}{2}} \leq 1 \quad \forall i. \tag{2.55}$$

Proof See Appendix B. □

This lemma completes the proof of Theorem 2.3. □

This theorem is quite sharp, since in (2.46) it assumes only the necessary conditions of Lemma 2.2 for the piecewise linear recovery to be TVD, and it gives a CFL range as large as one can reasonably expect from the construction of the fluxes. To illustrate this last point, consider the linear constant coefficient problem, for which all characteristic speeds equal $a > 0$. Then in (2.50a), $\lambda^+ = \nu_1^+ = \mu_1^+ = a\Delta t / h_i$ while in (2.50c), $\nu_2^+ = a\Delta t/h_i$ and $\mu_2^+ = a\Delta t / h_{i-1}$. So (2.47a) reduces to $a\Delta t \leq h_i$ while (2.47b) and (2.47c) reduce to $(a\Delta t)^2 \leq h_i h_{i-1}$. These are always satisfied under the conditions assumed in Appendix A for constructing the fluxes. By contrast, in the first order MUSCL scheme $\nu^+ = 0$ so that one needs $a\Delta t \leq \frac{1}{2} h_i$, the same result that is obtained by a direct analysis; for alternative second order fluxes, such as (2.44) or (2.45), it is considerably more difficult to establish sharp stability bounds and numerical experience suggests that the conditions (2.46) need to be strengthened for CFL numbers near unity – see e.g. [47].

The fact that for these schemes a TVD result is obtained even though the recovery process may not be TVD is an example of the smoothing effect resulting from the evolution and projection phases in the overall procedure of (2.14). Now we give a weaker theorem, but with a shorter proof, based on assuming a TVD recovery stage.

Theorem 2.6 *Suppose that scheme (2.42) with fluxes (2.43) is applied under the CFL conditions of (A.16). Then if the recovery stage is TVD, so is the overall procedure.*

Proof Since in (2.14) the projection stage is TVD, and we have assumed the same for the recovery R_h, it is sufficient to show that the approximate evolution operator E_Δ is also TVD. We do so by showing that it corresponds to the Brenier transport collapse operator for an approximate conservation law.

In Appendix A we have shown that the finite volume formulation that is used in (2.42) is exactly equivalent to the projection of the transport collapse operator, which is TVD; and the fluxes (2.43) are exact for a flux function $f(\cdot)$ constructed on each cell from its end values and a linearly interpolated characteristic speed $A(\cdot)$. Thus it is just required to show that these independent constructions on each cell are consistent with an approximate conservation law. In fact, it is enough to show this consistency over all the $\tilde{U}^n$ values that contribute to the update in the neighbourhood of each point, or all the points in each cell, say I_i. Now these may depend on the values of $\tilde{U}^n$ in cells I_{i-1}, I_i and I_{i+1}, under the CFL restrictions of (A.16). Clearly if U^n is monotone over these cells, then by the TVD properties assumed for R_h so is $\tilde{U}^n$, and the flux approximations are consistent. On the other hand, suppose U_i^n is a local maximum, and therefore R_h gives $S_i^n = 0$; and suppose further that $U_i^n > \bar{u}$ so that $s_i^+ = s_i^- = +1$. Then in the worst case we have $s_{i-1}^+ = +1$ and $s_{i+1}^- = -1$ so that both I_{i-1} and I_{i+1} contribute to the update in cell I_i. But since we must have $\tilde{U}_{i-1}^+ \leq U_i^n$, $S_{i-1}^n \geq 0$, $\tilde{U}_{i+1}^- \leq \bar{u}$ and $S_{i+1}^n \leq 0$, and our construction of $A(\cdot)$ uses interpolation points at the sonic point $\bar{u}$ when needed, the update from cell I_{i-1} uses $\tilde{U}^n$ values from the interval $(\bar{u}, U_i^n)$ and that from cell I_{i+1} values from $(\tilde{U}_{i+1}^+, \bar{u})$. Hence there is no possibility of inconsistency in the flux values that are used in such a case. □

2.3 Piecewise parabolic recovery and third order schemes

Just as the algorithms and analysis of the previous section develop and consolidate the work of van Leer [47] and Goodman and LeVeque [16], so the algorithms in this section are closely related to the *Piecewise Parabolic Method* (PPM) of Colella and Woodward [8]. We start from the PERU framework of (2.14), with the projection-evolution stage $P_h E_\Delta$ accomplished through the finite volume formulation that leads to (2.16), but with the recovery stage giving a continuous function $\tilde{U}^n \equiv R_h U^n$ which is parabolic on each cell.

The recovery stage

We denote by $\{\tilde{U}_{j+\frac{1}{2}}\}$ the interface value $\{\tilde{U}(x_{j+\frac{1}{2}})\}$ of the recovered function $\tilde{U} \equiv R_h U$; and we impose the conservation in the mean condition $P_h \tilde{U} = U$ of (2.15). Then the quadratic $\tilde{U}$ in each cell I_j is determined from the three parameters $\tilde{U}_{j-\frac{1}{2}}$, $\tilde{U}_{j+\frac{1}{2}}$ and U_j in the form

$$\begin{aligned} \tilde{U}(x) \quad = \quad & \left[\tilde{U}_{j-\frac{1}{2}}(x_{j+\frac{1}{2}} - x) + \tilde{U}_{j+\frac{1}{2}}(x - x_{j-\frac{1}{2}})\right]/h_j \\ & -3(\tilde{U}_{j-\frac{1}{2}} + \tilde{U}_{j+\frac{1}{2}} - 2U_j)(x_{j+\frac{1}{2}} - x)(x - x_{j-\frac{1}{2}})/h_j^2, \quad \forall x \in I_j. \end{aligned} \tag{2.56}$$

The only remaining problem is to compute the interface values from neighbouring cells averages. The conventional approach, and that used in PPM, is to do so via the primitive function of U; but this is not available in two dimensions, so we shall also present an alternative approach.

Let W be the quartic that interpolates the primitive function of U at five interface points centred at $x_{j+\frac{1}{2}}$. We will derive the PPM recovery values, and two alternatives, through the *Newton divided difference formula*

$$W(x) = W[x_0] + (x - x_0)W[x_0, x_1] + (x - x_0)(x - x_1)W[x_0, x_1, x_2] + \cdots,$$

where $W[x_0, x_1, \dots x_k] = (W[x_0, x_1, \dots, x_{k-2}, x_k] - W[x_0, x_1, \dots, x_{k-1}])/(x_k - x_{k-1})$. Then the derivative at x_0 is given by

$$W'(x_0) = W[x_0, x_1] + (x_0 - x_1)W[x_0, x_1, x_2] + (x_0 - x_1)(x_0 - x_2)W[x_0, x_1, x_2, x_3] + \cdots \tag{2.57}$$

By taking $x_0, x_1, \dots, x_4$ to be $x_{j+\frac{1}{2}}$, $x_{j-\frac{1}{2}}$, $x_{j+\frac{3}{2}}$, $x_{j-\frac{3}{2}}$, $x_{j+\frac{5}{2}}$, we then derive a formula for the recovered interface value $\tilde{U}_{j+\frac{1}{2}}$. It is easy to check that the successive divided differences are given by the following more transparent notation,

$$\begin{aligned}
W[x_0, x_1] &= \frac{W(x_{j+\frac{1}{2}}) - W(x_{j-\frac{1}{2}})}{h_j} \equiv U_j \qquad & W[x_0, x_1, x_2] &= \frac{U_{j+1} - U_j}{h_{j+1} - h_j} =: N^{(1)}_{j+\frac{1}{2}}, \\
W[x_0, ..., x_3] &= \frac{N^{(1)}_{j+\frac{1}{2}} - N^{(1)}_{j-\frac{1}{2}}}{h_{j+1} + h_j + h_{j-1}} =: N^{(2)}_j \qquad & W[x_0, ..., x_4] &= \frac{N^{(2)}_{j+1} - N^{(2)}_j}{h_{j+2} + h_{j+1} + h_j + h_{j-1}} =: N^{(3)}_{j+\frac{1}{2}}.
\end{aligned} \tag{2.58}$$

Then (2.57) becomes

$$\tilde{U}^{PPM}_{j+\frac{1}{2}} = U_j + h_j N^{(1)}_{j+\frac{1}{2}} - h_j h_{j+1} N^{(2)}_j - h_j h_{j+1}(h_j + h_{j-1}) N^{(3)}_{j+\frac{1}{2}}, \tag{2.59a}$$

which on a uniform mesh becomes

$$\tilde{U}^{PPM}_{j+\frac{1}{2}} = [1 + \tfrac{1}{2}\Delta_+ - \tfrac{1}{6}\delta^2 - \tfrac{1}{12}\delta^2\Delta_+]U_j. \tag{2.59b}$$

If this formula is truncated after two terms it gives a simple two-cell average for the interface value; if it is truncated after three terms, it corresponds to using a discontinuous quadratic recovery formula, which can be used to derive the four-point finite difference update algorithm of Warming et al. [49] that is third order accurate on a uniform mesh. The full formula, however, corresponds to PPM which is a centrally positioned five-point update scheme.

There is a lack of positional symmetry in the successive terms of (2.58) and (2.59), which can be overcome by a regrouping in pairs: let us define

$$U_{j+\frac{1}{2}} := U_j + h_j N^{(1)}_{j+\frac{1}{2}} \equiv \frac{h_{j+1}U_j + h_j U_{j+1}}{h_{j+1} + h_j}, \tag{2.60a}$$

$$N^{(2)}_{j+\frac{1}{2}} := N^{(2)}_j + (h_j + h_{j-1}) N^{(3)}_{j+\frac{1}{2}} \equiv \frac{(h_{j+2} + h_{j+1})N^{(2)}_j + (h_j + h_{j-1})N^{(2)}_{j+1}}{h_{j+2} + h_{j+1} + h_j + h_{j-1}}; \tag{2.60b}$$

then we can rewrite (2.59a) in the compact form

$$\tilde{U}^{PPM}_{j+\frac{1}{2}} = U_{j+\frac{1}{2}} - h_j h_{j+1} N^{(2)}_{j+\frac{1}{2}}, \tag{2.60c}$$

or $(1-\frac{1}{6}\delta^2)U_{j+\frac{1}{2}}$ on a uniform mesh.

Various alternatives could be used for the second term in (2.60c), one of which can be derived as follows. If we start with the Newton divided difference formula for $u(x)$ and integrate it, we obtain

$$\int_{x_0}^{x} u dx = (x-x_0)u[x_0] + \tfrac{1}{2}(x-x_0)^2 u[x_0,x_1] + \cdots;$$

then, with the same choice of points as above, we match these integrals with the cell averages U_j and U_{j+1} to obtain after a little calculation

$$\begin{aligned} U_j &= u_0 - \tfrac{1}{2}h_j u[x_0,x_2] + \tfrac{1}{6}h_j(2h_j+3h_{j+1})u[x_0,x_1,x_2] + \cdots \\ U_{j+1} &= u_0 + \tfrac{1}{2}h_{j+1}u[x_0,x_2] - \tfrac{1}{6}h_{j+1}^2 u[x_0,x_1,x_2] + \cdots. \end{aligned}$$

Taking the average as in (2.60a) eliminates the first order differences to give

$$U_{j+\frac{1}{2}} = u_0 + \tfrac{1}{3}h_j h_{j+1} u[x_0,x_1,x_2] + \tfrac{1}{12}h_j h_{j+1}(h_j-h_{j+1})u[x_0,x_1,x_2,x_3] + \cdots \tag{2.61a}$$

which we write symbolically as

$$U_{j+\frac{1}{2}} = \left(1 + \tfrac{1}{3}\Delta_{012} + \tfrac{1}{12}\Delta_{0123} + \cdots\right) u_0. \tag{2.61b}$$

This formula relates a sequence of interface values $u_0 \equiv u(x_0), u(x_1), u(x_2), \ldots$ to the weighted two-cell average $U_{j+\frac{1}{2}}$, which is also centred at an interface. Moreover, the operator Δ_{012} is a second order difference operator on these interface values and the next term Δ_{0123} is a third order operator which is zero when $h_j = h_{j+1}$. So an approximate inversion of (2.61b) gives

$$u_0 = \left(1 + \tfrac{1}{3}\Delta_{012} + \tfrac{1}{12}\Delta_{0123} + \cdots\right)^{-1} U_{j+\frac{1}{2}} \approx \left(1 - \tfrac{1}{3}\Delta_{012}\right) U_{j+\frac{1}{2}}. \tag{2.62}$$

Here Δ_{012} can be replaced by the more natural notation

$$D^{(2)}U_{j+\frac{1}{2}} := \frac{1}{\frac{1}{2}(h_{j+1}+h_j)}\left[\frac{U_{j+\frac{3}{2}} - U_{j+\frac{1}{2}}}{h_{j+1}} - \frac{U_{j+\frac{1}{2}} - U_{j-\frac{1}{2}}}{h_j}\right] \equiv 2\Delta_{012}U_{j+\frac{1}{2}}, \tag{2.63a}$$

which leads to the following alternative formula for recovering the interface values, in terms of a divided difference operator applied to the two-cell averages of (2.60a),

$$\tilde{U}^{IDD}_{j+\frac{1}{2}} := \left(1 - \tfrac{1}{6}h_j h_{j+1} D^{(2)}\right) U_{j+\frac{1}{2}}. \tag{2.63b}$$

This formula reduces to the PPM formula (2.60c) on a uniform mesh, but the argument that led to it can be applied in multi-dimensions; the notation IDD denotes *Integrated Divided Difference* formula.

A second, closely related alternative, which may be the most practically useful, is based on the recovered cell slopes (2.23) used for the discontinuous linear recovery introduced in the last section. Suppose we integrate the Newton divided difference formula as above, to obtain expansions for U_{j-1} and U_{j+2} as well as U_j and U_{j+1}. Then we can apply (2.23) to find the slopes in the two cells I_j and I_{j+1}, which we denote by $\tilde{S}_j$ and $\tilde{S}_{j+1}$ respectively. Relating the notation of (2.24) with that used above, we have

$$\begin{aligned} D_+U_j \equiv \frac{U_{j+1}-U_j}{\frac{1}{2}(h_j+h_{j+1})} &= u[x_0,x_2] - \tfrac{1}{3}(2h_j+h_{j+1})u[x_0,x_1,x_2] + \cdots \\ &\equiv u[x_0,x_1] + \tfrac{1}{3}(h_j+2h_{j+1})u[x_0,x_1,x_2] + \cdots \end{aligned} \tag{2.64}$$

with similar expressions for D_-U_j. Hence we deduce that

$$\begin{aligned} \tilde{S}_j &\equiv \frac{(2h_{j-1}+h_j)D_+U_j + (h_j+2h_{j+1})D_-U_j}{2(h_{j-1}+h_j+h_{j+1})} \\ &= u[x_0,x_1] + O(h^2), \end{aligned} \tag{2.65a}$$

where $u[x_0,x_1] \equiv u[x_{j+\frac{1}{2}}, x_{j-\frac{1}{2}}]$ is the average slope of $u(\cdot)$ in cell I_j, and also that

$$\frac{\tilde{S}_{j+1}-\tilde{S}_j}{h_j+h_{j+1}} = u[x_0,x_1,x_2] + O(h). \tag{2.65b}$$

The latter can be used to eliminate $u[x_0,x_1,x_2]$ from (2.61a) and this gives the alternative $O(h^3)$ recovery formula

$$U^{SC}_{j+\frac{1}{2}} := U_{j+\frac{1}{2}} - \tfrac{1}{3}\frac{h_jh_{j+1}(\tilde{S}_{j+1}-\tilde{S}_j)}{h_j+h_{j+1}}, \tag{2.66}$$

where SC denotes that it is the *slope-corrected* two-cell average $U_{j+\frac{1}{2}}$.

It is also worth noting here that other approximations to the cell slopes could be used here instead of the $\tilde{S}_j$ given above, such as those discussed in section 2.2; and the present divided difference notation enables us to point out some useful relationships between such approximations. Thus in terms of the $U_{j\pm\frac{1}{2}}$ of (2.60a), it is easily checked that the equally weighted average of $D_\pm U_j$ can also be written as a difference of averages:

$$\tfrac{1}{2}(D_+U_j + D_-U_j) \equiv \frac{1}{h_j}(U_{j+\frac{1}{2}} - U_{j-\frac{1}{2}}). \tag{2.67}$$

Moreover, suppose that the average of W'' over the interval I_j, namely $h_j^{-1}[W'(x_{j+\frac{1}{2}}) - W'(x_{j-\frac{1}{2}})]$, is expanded as in (2.57), (2.58), (2.59a) and a corresponding formula for $\tilde{U}^{PPM}_{j-\frac{1}{2}}$. Then a simple computation based on (2.60a) shows that

$$\begin{aligned} \langle W'' \rangle_{I_j} &= \frac{1}{h_j}(U_{j+\frac{1}{2}} - U_{j-\frac{1}{2}}) - (h_{j+1}-h_{j-1})N^{(2)}_j + O(h^2) \\ &\equiv \tilde{S}_j + O(h^2). \end{aligned} \tag{2.68}$$

Moreover, from (2.60b,c) we obtain a third approximation to the slope,

$$\begin{aligned} \langle W'' \rangle_{I_j} &= \frac{1}{h_j}(U_{j+\frac{1}{2}} - U_{j-\frac{1}{2}}) - (h_{j+1}-h_{j-1})(N^{(2)}_{j+\frac{1}{2}} - N^{(2)}_{j-\frac{1}{2}}) + O(h^3) \\ &\equiv \frac{1}{h_j}(\tilde{U}^{PPM}_{j+\frac{1}{2}} - \tilde{U}^{PPM}_{j-\frac{1}{2}}) + O(h^3), \end{aligned} \tag{2.69}$$

into which we could also substitute $\tilde{U}^{IDD}_{j\pm\frac{1}{2}}$.

The PPM recovery of $\tilde{U}_{j+\frac{1}{2}}$ has an error of only $O(h^4)$, as is clear from its derivation and the formulae of (2.59); but the third order difference can be absorbed into the average of the second order differences given in (2.60c), where the $O(h^4)$ accuracy is less obvious. Thus the alternative forms of (2.63b) and (2.66), although only having $O(h^3)$ accuracy on a general mesh, are likely to have similar accuracy in practice; for example, (2.63b) merely replaces the average of second differences in PPM with a second difference of averages. In any case, the $O(h^3)$ accuracy is entirely adequate for the purpose of constructing the quadratic recovery of (2.56) and the calculation of the interface fluxes. Both of these points are confirmed by the numerical tests carried out by Dawkins [12].

Conditions for the preservation of a monotone approximation are more complicated and less well established for continuous quadratic recovery than for discontinuous linear recovery. We will give just one criterion, and one technique for restoring monotonicity by sacrificing intercell continuity.

Lemma 2.7 *If the piecewise constant U is monotone increasing, then so is the piecewise quadratic recovery $\tilde{U}$ of (2.56) if and only if*

$$\tfrac{1}{2}(U_j - \tilde{U}_{j-\frac{1}{2}}) \le \tilde{U}_{j+\frac{1}{2}} - U_j \le 2(U_j - \tilde{U}_{j-\frac{1}{2}}) \quad \forall j. \tag{2.70}$$

Proof Because of the continuity of $\tilde{U}$, each cell can be considered separately; and it is clear that a necessary and sufficient condition for the quadratic $a + bs + cs^2$ to be monotone on $[-\frac{1}{2}, \frac{1}{2}]$ is that $|c| \le |b|$. For (2.56) this becomes

$$3|\tilde{U}_{j-\frac{1}{2}} + \tilde{U}_{j+\frac{1}{2}} - 2U_j| \le |\tilde{U}_{j+\frac{1}{2}} - \tilde{U}_{j-\frac{1}{2}}| \quad \forall j. \tag{2.71}$$

If $\tilde{U}$ is monotone increasing, we have $\tilde{U}_{j+\frac{1}{2}} \ge \tilde{U}_{j-\frac{1}{2}}$ so that (2.71) reduces to the two conditions $3(\tilde{U}_{j+\frac{1}{2}} - U_j) - 3(U_j - \tilde{U}_{j-\frac{1}{2}}) \le (\tilde{U}_{j+\frac{1}{2}} - U_j) + (U_j - \tilde{U}_{j-\frac{1}{2}})$ and $3(U_j - \tilde{U}_{j-\frac{1}{2}}) - 3(\tilde{U}_{j+\frac{1}{2}} - U_j) \le (\tilde{U}_{j+\frac{1}{2}} - U_j) + (U_j - \tilde{U}_{j-\frac{1}{2}})$; and these reduce to the conditions (2.70). On the other hand, (2.70) implies these two conditions and also that $\tilde{U}_{j+\frac{1}{2}} - U_j \ge 0 \ge \tilde{U}_{j-\frac{1}{2}} - U_j$; hence (2.70) implies (2.71). □

Should these conditions not be satisfied by the recovered interface values, monotonicity can be restored as follows: if the right-hand inequality of (2.70) is violated, $\tilde{U}^+_j$ is set to $U_j + 2(U_j - \tilde{U}_{j-\frac{1}{2}})$; if the left-hand inequality is violated, $\tilde{U}^-_j$ is set to $U_j - 2(\tilde{U}_{j+\frac{1}{2}} - U_j)$; and then the quadratic (2.56) is formed with these replacements for $\tilde{U}_{j\pm\frac{1}{2}}$. The result will be discontinuous, but it will be monotone increasing because either change brings an extreme value closer to the mean U_j.

This limiting procedure can be applied locally, to cell I_j if $(\Delta_+ U_j)(\Delta_- U_j) \ge 0$. Otherwise a local extremum will normally be enhanced by the recovery procedure; to insist on a TVD recovery would seem to be too crude in this case, and use of a TVB criterion or application of an ENO-type adjustment to the recovery would seem more appropriate — see Harten and Osher [19] for more details. There are many other techniques, too, that take into account both the update and the recovery procedures, so it is inappropriate to pursue this topic further here.

Numerical fluxes and the update stage

A third order accurate method requires the interfacial fluxes to be calculated to $O(h^3)$, which can be achieved with the above quadratic recovery of U together with the approximate quadratic flux function constructed in the previous section for use with the discontinuous linear recovery of U. We allow the quadratic recovery to be discontinuous, so in cell I_j we work with end values $\tilde{U}_j^{\pm}$, the signs $s_j^{\pm}$ of $a(\tilde{U}_j^{\pm})$, and the gradients $m_j^{\pm}$ of the approximation $A_j(\tilde{U})$ to $a(\cdot)$ given by (2.36), in terms of $\tilde{U}_j^{\pm}$ and the nearest sonic point — note that there could now be two sonic points in an interval. So at the two ends we suppose that we have

$$\frac{d\tilde{U}}{dx} = \beta_j^{\pm} + \gamma_j(x - x_{j\pm\frac{1}{2}}), \tag{2.72}$$

with the three parameters β_j^+, β_j^- and γ_j determined by the quadratic form of $\tilde{U}$.

Then it is shown in Appendix A that, if $s_j^{\pm} = \pm 1$, there is a characteristic speed $A_j^{*\pm}$ given by a quadratic equation which to compare with (2.37) or (2.39), we can write as

$$A_j^{*\pm} = \frac{a(\tilde{U}_j^{\pm})}{1 + m_j^{\pm}\Delta t(\beta_j^{\pm} - \frac{1}{2}\gamma_j A_j^{*\pm}\Delta t)} \quad \text{if} \quad \gamma_j \geq 0, \tag{2.73a}$$

or alternatively as

$$A_j^{*\pm} = \frac{a(\tilde{U}_j^{\pm}) - \frac{1}{2}\gamma_j m_j^{\pm} A_j^{*\pm} h_j \Delta t}{1 + m_j^{\pm}\Delta t[\beta_j^{\pm} - \frac{1}{2}\gamma_j(h_j + A_j^{*+}\Delta t)]} \quad \text{if} \quad \gamma_j < 0; \tag{2.73b}$$

moreover, these alternative forms can be used to give a monotonically convergent iteration for the characteristic speed $A_j^{*\pm}$ in the two cases.

Writing $U_j^{*\pm}$ for $\tilde{U}(x_{j\pm\frac{1}{2}} - A_j^{*\pm}\Delta t)$, the average interface flux at each end of the interval I_j is given by

$$\tilde{F}_j^{\pm} = \frac{A_j^{*\pm} f(\tilde{U}_j^{\pm}) + a(\tilde{U}_j^{\pm}) f(U_j^{*\pm})}{A_j^{*\pm} + a(\tilde{U}_j^{\pm})} - \tfrac{1}{16}\gamma_j (A_j^{*\pm})^3 \Delta t^2, \tag{2.74}$$

which generalises (2.38) and (2.40). Then the result from two neighbouring intervals can be combined with the flux at the sonic point to give the interface flux $\tilde{F}_{i-\frac{1}{2}}^{n+\frac{1}{2}}$ at $x_{i-\frac{1}{2}}$ by the formula (2.41) as in earlier cases.

The result proved in Theorem 2.6 for discontinuous linear recovery can equally well be established for the present quadratic recovery, by the same arguments. We do not state it here as it is of less interest in this case; this is because one hopes that schemes based on quadratic recovery will restore extrema that have been clipped by the projection onto piecewise constants. To achieve this one needs to relax the TVD conditions, and to discuss the multiplicity of ways in which this might be done would take us well beyond the scope of the present article.

2.4 Systems of conservation laws

The two new schemes that we have proposed for scalar problems in sections 2.2 and 2.3 above have several features in common. At each interface, such as at $x_{i-\frac{1}{2}}$, a left state and right state are identified, namely $\tilde{U}_{i-1}^{+}$ and $\tilde{U}_{i}^{-}$, together with their continuous extensions to the neighbouring cells I_{i-1} and I_i respectively. Then in each case one obtains two characteristic speeds, $a(\tilde{U}_{i-1}^{+})$ and A_{i-1}^{*+} on the left, $a(\tilde{U}_{i}^{-})$ and A_{i}^{*-} on the right, from which are constructed the flux functions $\tilde{F}_{i-1}^{+}$ and $\tilde{F}_{i}^{-}$ given by (2.38), (2.40) and (2.74). Moreover, in both of these cases the flux functions are closely related to those of schemes which are well established for use with systems of conservation laws, namely the MUSCL and PPM schemes.

Thus it is natural that we seek to extend these new schemes to systems of conservation laws by using approximate Riemann solvers and flux vector splitting algorithms in the same ways that are common with MUSCL-type and PPM-type schemes – see e.g. [4] for an extension of (2.45) to general systems of conservation laws. The details of how this may be done for these new fluxes will be the subject of a further publication. However, it should be noted here that at the heart of the approach is the extension of Brenier's transport collapse operator to systems of equations by means of an approximate diagonalisation of the flux Jacobians at either end of each cell, so that the contributions to the interface fluxes can be calculated for each wave as in sections 2.1, 2.2 and 2.3 above.

We note here too that it is straightforward to add diffusive terms to the fluxes, or viscous terms to the conservation laws for fluid flow. When these are small it will be adequate to evaluate them at the old time level; then all that is needed are the gradients of the solution, for which the recovery formulae given above can be used. When the diffusion is more dominant it may be taken account of in the evolution operator – see [36] for details.

3 Two-dimensional problems

In some ways, the use of a characteristic-based algorithm is more natural and advantageous for scalar problems in multidimensions than for systems of equations in one dimension. Moreover, for many systems of equations in multi-dimensions the approximation of the evolution operator requires some use of the bicharacteristics as well as characteristics — see Morton [31] and Lukáčová et al [29]. Hence in the following sections we will confine our attention to the scalar conservation law corresponding to (1.6), namely

$$\partial_t u + \partial_x f(u) + \partial_y g(u) = 0 \quad \text{or} \quad \partial_t u + a(u)\partial_x u + b(u)\partial_y u = 0, \tag{3.1}$$

where $a(u) = f'(u)$ and $b(u) = g'(u)$.

What was called the second class of schemes in section 1.1, and is based on calculating *cell residuals* that are then distributed to the cell vertices to effect the update, is particularly effective for near-equilibrium problems. Such schemes apply very naturally to unstructured meshes, and are usually implemented through the use of a specially designed Runge-Kutta update procedure. However, our interest in this paper is in methods suitable for problems where the unsteady effects are dominant, and hence where the time and space discretisation should be combined, as in characteristic-based methods. We shall therefore continue to concentrate on the first class of methods, based on a finite volume formulation which uses cell averages and interfacial fluxes.

3.1 Tensor product meshes

It is easily deduced that when the characteristic speeds a and b in (3.1) are constants then any evolution Galerkin method on a tensor product mesh will yield a tensor product of the corresponding one-dimensional operators. For example, the Lesaint [24] scheme in one dimension on a uniform mesh, and with the CFL number $\nu \equiv a\Delta t/\Delta x$ in $[0,1]$, can be written in finite difference notation as

$$\left[1 + \tfrac{1}{6}\delta_x^2\right] \tilde{U}_i^{n+1} = \left[1 + \tfrac{1}{6}\delta_x^2 - \nu\Delta_{0x} + \tfrac{1}{2}\nu^2\delta_x^2 - \tfrac{1}{6}\nu^3\delta_x^2\Delta_{-x}\right] \tilde{U}_i^n. \tag{3.2}$$

This scheme involves three mesh points x_{i-1}, x_i, x_{i+1} at the new time level, with the fourth x_{i-2} added at the old time level to give a set which is symmetrically distributed around the interval (x_{i-1}, x_i) containing the foot of the characteristic drawn back from (x_i, t_{n+1}). The scheme is third order accurate on a uniform mesh and was originally derived by a Lagrange-Galerkin method applied to the piecewise linear approximation of (2.5a). As has often been pointed out – see for instance Morton [30, p.329] – it is, however, equivalent to a semi-Lagrangian method based on a recovery stage using cubic splines. Thus we can write it as

$$[1 + \tfrac{1}{6}\delta_x^2]\tilde{U}_i^n = U_i^n \tag{3.3a}$$

$$h(U_i^{n+1} - U_i^n) + a\Delta t[\Delta_{0x} - \tfrac{1}{2}\nu\delta_x^2 + \tfrac{1}{6}\nu^2\delta_x^2\Delta_{-x}]\tilde{U}_i^n, \tag{3.3b}$$

where $\{\tilde{U}_i^n\}$ are the expansion coefficients for a cubic spline with nodal values $\{U_i^n\}$, and (3.3b) gives the update of these nodal values using the cubic spline approximation to the foot of the characteristic through (x_i, t_{n+1}). Moreover, the application of this technique on a nonuniform mesh maintains the third order accuracy which is lost by the original piecewise linear approximation. The effectiveness of this approach is well exemplified by the fact that it presently forms the basis on many weather forecasting codes — see Staniforth and Côté [44].

The same convolution property of splines which identified the left-hand side of (3.2), formed from the inner product of linear splines, with that of (3.3a) formed by a cubic spline and a delta function, also identifies it with the inner product of a piecewise constant spline with a quadratic spline. Hence we can interpret (3.3a) as recovery by a quadratic spline from piecewise constants; so the whole scheme (3.3) is directly comparable with finite volume schemes like PPM — see [30, pp.323–7]. Quadratic splines have continuous derivatives so that there are no free parameters in the recovery process; the advantage of the PPM-type recovery of section 2.3 is that there are free parameters to allow the recovery to be adaptive and reflect such properties as monotonicity of the solution.

The three different interpretations of (3.2) and (3.3) (based on point values, linear splines, or piecewise constants) will give different schemes when applied on a nonuniform mesh. In each case they can be written as

$$\tilde{U}^n = R_x U^n, \qquad U^{n+1} = S_x \tilde{U}^n, \tag{3.4}$$

where R_x represents the solution of the tridiagonal system of equations which is given by $(1 + \frac{1}{6}\delta_x^2)\tilde{U}^n = U^n$ on a uniform mesh, and S_x is an explicit difference operator. Then on a tensor product mesh we can define corresponding operators on the y-mesh, with S_y having the characteristic speed a replaced by b. The resulting schemes for the constant coefficient case of (3.1) thus become

$$\tilde{U}^n = R_x R_y U^n, \qquad U^{n+1} = S_x S_y \tilde{U}^n. \tag{3.5}$$

Indeed, this holds for other schemes generated in this way, for example the first order upwind scheme that corresponds to (3.3) without recovery and the operator on the right of (3.3b) replaced by $a\Delta t \Delta_{-x}$. In that case, $S_x S_y$ becomes

$$\left[1 - \frac{a\Delta t}{\Delta x_i}\Delta_{-x}\right]\left[1 - \frac{b\Delta t}{\Delta y_j}\Delta_{-y}\right], \tag{3.6}$$

exhibiting the correct "corner" term $\Delta_{-x}\Delta_{-y}U^n$ which maintains stability for $a\Delta t \leq \Delta x_i$ and $b\Delta t \leq \Delta y_j$ (rather than the limit of $(a\Delta t/\Delta x_i) + (b\Delta t/\Delta y_j) \leq 1$ obtained if this term is omitted).

It is shown in [28] that these nice properties of evolution Galerkin methods on tensor product meshes also hold for the general nonlinear conservation law of (3.1). For example, for $0 \leq a(U^n)\Delta t \leq \Delta x_i$, $0 \leq b(U^n)\Delta t \leq \Delta y_j$ $\forall i, j$, the scheme using piecewise constants without recovery becomes

$$\begin{aligned} \Delta x_i \Delta y_j (U^{n+1}_{i,j} - U^n_{i,j}) + \Delta t \Delta y_j \Delta_{-x} f(U^n_{i,j}) + \Delta t \Delta x_i \Delta_{-y} g(U^n_{i,j}) \\ - \ (\Delta t)^2 \Delta_{-x}\Delta_{-y} h(U^n_{i,j}) = 0, \end{aligned} \tag{3.7a}$$

where the new flux $h(\cdot)$ generated by the use of the Brenier transport collapse operator is given by

$$h(u) := \int^u a(v)b(v)dv; \tag{3.7b}$$

this natural extension of the relationship between a, b and f, g was first given by Morton and Childs [33].

The generalisation of this scheme to nonpositive a and b, to arbitrary time steps, and to three dimensions is given by Lin et al. in [28] where it is shown to be unconditionally TV-stable. However, the formulae are not as explicitly given as (3.7) in terms of the flux functions and characteristic speeds. So we shall not pursue these developments further here, but consider instead arbitrary unstructured meshes and how the schemes of section 2.2 and 2.3 should be developed for them.

3.2 Recovery/reconstruction on triangular meshes

The difficulties in actually constructing recovery functions on unstructured triangular meshes, and of extending the one-dimensional ideas of monotonicity-preservation and TVD, have led to a very large literature on this topic. We will limit ourselves here to extending the constructions of sections 2.2 and 2.3, and referring to the literature that is relevant to that task.

We showed in section 2.3 that both the discontinuous linear recovery and the continuous quadratic recovery functions can be derived by the use of Newton divided difference formulae.

Moreover, motivated by the fact that the primitive function for u does not exist in two dimensions, we obtained derivations based on divided differences for u itself. Now, as shown by Sauer and Xu [43], one can still construct Newton divided difference expansions on a 2D unstructured mesh; but, firstly, the sequence of mesh points is much more difficult to choose; and, secondly, the Vandemonde determinants that arise in constructing the Newton fundamental polynomials are very badly conditioned. Durlofski et al. [13] and Abgrall [1], in extending ENO scheme ideas to unstructured meshes, and Barth and Frederickson [3], in making use of least squares criteria, have mitigated these difficulties to some extent. But we will adopt here a much more direct approach to generalising the formulae of section 2.2 and 2.3.

Discontinuous linear recovery, which is conservative in the mean and thus generalises (1.7) or (A.15), just requires a gradient vector $\mathbf{S}_i$ in each cell Ω_i, with centroid at $\mathbf{x}_i$, to give

$$\tilde{U}(\mathbf{x}) = U_i + (\mathbf{x} - \mathbf{x}_i) \cdot \mathbf{S}_i \qquad \mathbf{x} \in \Omega_i. \tag{3.8}$$

A direct generalisation of (2.23), which is an average of divided differences using the two neighbouring intervals, or of the slope-limited choices (2.31), (2.32) or (2.34), would construct gradients from the cell averages in pairs of neighbouring triangles. There are three natural candidates as neighbours of Ω_i, i.e. those with a common edge, and formulae based on these are given by Durlofski et al. [13]. However, from a finite volume viewpoint it is more natural to obtain the average gradient over a cell by applying Gauss' theorem to its perimeter. That is we generalise formulae (2.67), (2.69) instead, and for this purpose we need to recover values at the vertices of the mesh – which we shall also need for a quadratic recovery.

So we consider first the task of recovering vertex values from cell averages. If the six or so edges meeting at a vertex occur in a straight line pairs, each cell Ω_j has an opposite cell Ω_{jo}, of measure V_{jo}, and one can show that the formula

$$\tilde{U}_{\text{vertex}} = \frac{\sum V_{jo} U_j}{\sum V_{jo}} \tag{3.9}$$

is exact for linear functions and is the natural generalisation of (2.60a) from one to more dimensions. In general, any set of weights w_j will yield a formula which is exact for linear functions so long as they satisfy a centre of mass condition: if $\mathbf{x}'_j$ is the centroid of cell Ω_j relative to the vertex being considered, the condition is

$$\tilde{U}_{\text{vertex}} = \sum w_j U_j \quad \text{such that} \quad \sum w_j \mathbf{x}'_j = \mathbf{0}, \tag{3.10}$$

which of course could also be expressed in terms of the relative position vectors of the vertices surrounding that being considered. An alternative approach which is available for any mesh, is to find the weights from a least squares fit of the cell averages to a linear function; it only has to be performed once and the weights for each vertex stored for use at each time step.

Calculation of vertex values in this way will be sufficient to construct the discontinuous linear recovery (3.8), by setting

$$\begin{aligned} \mathbf{S}_i &= \frac{1}{V_i} \oint_{\partial\Omega_i} \begin{pmatrix} U dy \\ -U dx \end{pmatrix} = \frac{1}{2V_i} \sum_{(\alpha)} (\tilde{U}_\alpha + \tilde{U}_{\alpha+1}) \begin{pmatrix} y_{\alpha+1} - y_\alpha \\ x_\alpha - x_{\alpha+1} \end{pmatrix} \\ &= \frac{1}{2V_i} \sum_{(\alpha)} \tilde{U}_\alpha \begin{pmatrix} y_{\alpha+1} - y_{\alpha-1} \\ x_{\alpha-1} - x_{\alpha+1} \end{pmatrix}, \end{aligned} \tag{3.11}$$

where we have used the trapezoidal rule for calculating the integrals over $\partial\Omega_i$ and α is a cyclic index around the three vertices of Ω_i.

A very comparable scheme, that performed very well in the numerical tests of Dawkins [12], is constructed as follows. For each triangle Ω_i, with centroid (x_i, y_i), a system of four equations is solved to fit the bilinear function $a_{00} + a_{10}(x - x_i) + a_{01}(y - y_i) + a_{11}(x - x_i)(y - y_i)$ to the cell averages in Ω_i and its three neighbours. Then we set $\mathbf{S}_i = (a_{10}, a_{01})^T$.

Before moving to a full quadratic recovery, it should be noted that quite an effective scheme can be constructed by augmenting a continuous linear recovery in a manner very close to the continuous quadratic recovery in one dimension given by (2.56). So suppose that vertex values have been recovered from neighbouring cell averages, either as described above or by some other means. Then construct a continuous linear function from these values using standard finite elements; that is, on the canonical triangle with vertices $(\xi, \eta) = (0, 0), (1, 0)$ and $(0, 1)$ the basis functions for the three vertex values are $1 - \xi - \eta$, ξ and η. Then, as with (2.56), the cell average is restored by adding an extra basis function which is zero at these vertices. In this case it is the normalised cubic *bubble function* $27\xi\eta(1 - \xi - \eta)$, whose coefficient should be

$$\tilde{U}_i := \tfrac{20}{27}[3U_i - (\tilde{U}_\alpha + \tilde{U}_\beta + \tilde{U}_\gamma)] \tag{3.12}$$

where $\tilde{U}_\alpha, \tilde{U}_\beta, \tilde{U}_\gamma$ are the recovered vertex values.

For the full quadratic recovery we also use standard finite element basis functions, with the six parameters in each triangle provided by the three vertex values and three mid-side values; and the three parameters on each edge ensure inter-cell continuity. Recovery of the vertex values now needs to be more accurate. A very effective procedure is based on fitting a full local quadratic at each vertex to the cell averages on a patch which is extended by adding their neighbours to all the cells sharing the vertex. A further improvement over this is obtained by accepting the coefficients of the three quadratic terms from this fitting, but recomputing the linear terms (and in particular the vertex value) from a fit of the cell averages in the unextended patch. For the mid-side values, there are always just six cells in the extended patch consisting of the two cells which share this side and extended by their neighbours. So a fit of a quadratic to these cell averages gives the mid-side recovered value. Once these recovered values are substituted into the standard finite element expansion for a cell, the cell average may no longer be reproduced. The cubic bubble function can again be used to augment the expansion and restore the cell average; and again the required coefficient is given by a formula like (3.12), but with the $\tilde{U}_\alpha, \tilde{U}_\beta, \tilde{U}_\gamma$ now being the mid-side values as the vertex basis functions average to zero.

Numerical tests of all the recovery schemes given here, as well as other less successful schemes, are presented by Dawkins [12]. Grid refinement tests on fully unstructured triangular meshes show that the linear recovery schemes yield consistently $O(h^2)$ approximations. The quadratic recovery scheme given above maintains $O(h^3)$ accuracy, and on typical meshes is an order of magnitude more accurate than the linear recovery.

3.3 Numerical fluxes and the update stage on a triangular mesh

The Brenier transport collapse operator is equally well defined for the problem (3.1) in two dimensions as for the one-dimensional (1.6), and as in Appendix A it can be used to define numerical schemes in finite volume form. Let $\chi_i(\mathbf{x})$ be the characteristic function of the triangular cell Ω_i, i.e. $\chi_i(\mathbf{x}) = 1$ for $\mathbf{x} \in \Omega_i$ and equals zero otherwise. Then, as in (2.1), the piecewise constant approximation $U^n(\mathbf{x})$ has the form

$$U^n(\mathbf{x}) = \sum_{(j)} U_j^n \chi_j(\mathbf{x}), \tag{3.13}$$

and the update corresponding to (A.1) is given by

$$V_i(U_i^{n+1} - U_i^n) = \int \tilde{U}^n(\mathbf{x})[\chi_i(\mathbf{y})d\mathbf{y} - \chi_i(\mathbf{x})d\mathbf{x}]. \tag{3.14}$$

Here $\tilde{U}^n(\mathbf{x})$ is the recovered approximation and

$$\mathbf{y} = \mathbf{x} + \mathbf{a}(\tilde{U}^n(\mathbf{x}))\Delta t, \tag{3.15}$$

where $\mathbf{a} = (f', g')^T \equiv (a, b)^T$ in the notation of (3.1). We use a continuous parameterisation of the $[\tilde{U}^n, \mathbf{x}]$ graph and a Riemann-Stieltjes interpretation of the integrals to manipulate them into finite volume form. The manipulations are more complicated than those in Appendix A, and the details will be presented elsewhere, but the result is a final formula which is similar to (2.16), namely

$$V_i(U_i^{n+1} - U_i^n) + \Delta t \int_{\partial\Omega_i} \tilde{F}_{nb}^{n+\frac{1}{2}} dl = 0, \tag{3.16}$$

with normal boundary fluxes in a form similar to (2.41),

$$\tilde{F}_{nb}^{n+\frac{1}{2}} = \tfrac{1}{2}[(1 - s_{nb}^-)\tilde{F}_{nb}^- + (s_{nb}^- - s_{nb}^+)\bar{f}_{nb} + (1 + s_{nb}^+)\tilde{F}_{nb}^+]. \tag{3.17}$$

Here the signs $s_{nb}^{\pm}$ now refer to the signs of the normal component of the characteristic speed, $a_{nb} \equiv \mathbf{n}_b \cdot \mathbf{a}(\tilde{U}^n)$, at a boundary point just inside $(+)$ or outside $(-)$ the boundary, and $\bar{f}_{nb}$ is the value of $f_{nb} := \mathbf{n}_b \cdot (f, g)^T$ when $a_{nb} = 0$. Corresponding to the integral (A.7) for the boundary flux we have

$$\tilde{F}_{nb}^{\pm} = \frac{1}{\Delta t} \int_{\mathbf{x} \in \pi^{\pm}(0, \Delta t)} f_{nb}(\tilde{U}^n(\mathbf{x})) d\tau \tag{3.18}$$

where $\pi^{\pm}(0, \Delta t) := \{\mathbf{x} \text{ s.t. } \mathbf{x} + \mathbf{a}(\tilde{U}^n(\mathbf{x}))\tau = \mathbf{x}_b, 0 \le \tau \le \Delta t, \text{ for } a_{nb} \gtrless 0\}$; this holds at any boundary point $\mathbf{x}_b$ which is such that no trajectory from $\mathbf{x}$ to $\mathbf{y}$ passing through $\mathbf{x}_b$ crosses the boundary $\partial\Omega_i$ at any other point.

However, at boundary points near vertices of the mesh (i.e. corners of the triangle) the flux calculation will be more complicated. Suppose the situation is as in Figure 2, where the flux is to be calculated for the more heavily marked edge along which the component of $\mathbf{a}$ is a_{tb}. Then at distances from the vertex given by $l \equiv (a_{tb} + a_{nb} \cot\theta)\tau_b$ with $\tau_b < \Delta t$, part of the flux will be contributed from the neighbouring cell. Moreover, in the integral along the edge appearing in (3.16), dl should be replaced by $(a_{tb} + a_{nb} \cot\theta) d\tau_b$; and for the inner integral corresponding to (3.18), we will have τ running from 0 to τ_b (and separately from τ_b to Δt) and will obtain the factor a_{nb} that gave rise to f_{nb} through $a_{nb}dU = df_{nb}$. So now for this part of the boundary, the integrals will involve flux functions h_{nn} and h_{nt} in the combination $(h_{nt} + h_{nn}\cot\theta)$ where $h_{nn}(u) = \int^u [a_n(v)]^2 dv$ and $h_{nt}(u) = \int^u a_n(v)a_t(v)dv$. An example in which the integrals can be computed explicitly is provided by (3.7), on a rectangular mesh and without recovery. The flux through the right-hand boundary is given by

$$[\Delta y_j f(U^n_{i,j}) - \tfrac{1}{2}\Delta t h(U^n_{i,j})] + \tfrac{1}{2}\Delta t h(U^n_{i,j-1}), \tag{3.19a}$$

corresponding to $\theta = \frac{1}{2}\pi, a_n = a, a_t = b$ and $h_{nt} = h$ in the above notation. When a and b are constants, this reduces to

$$a[(\Delta y_j - \tfrac{1}{2}b\Delta t)U^n_{i,j} + \tfrac{1}{2}b\Delta t U^n_{i,j-1}] \tag{3.19b}$$

which exhibits the geometric effect at the corner, with the $\frac{1}{2}$ factors coming from the integrals

$$\int_0^{\Delta t} d\tau_b \int_0^{\tau_b} d\tau \quad \text{for cell } i,j \text{ and} \quad \int_0^{\Delta t} d\tau_b \int_{\tau_b}^{\Delta t} d\tau \quad \text{for cell } i, j-1.$$

It is clear that, on a general mesh and with linear or quadratic recovery, some approximation of the corresponding integrals will be necessary.

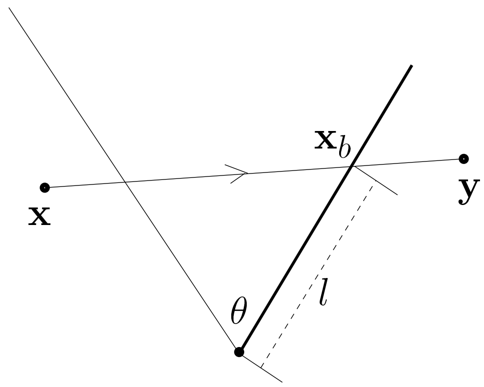

Figure 2: Boundary flux calculation for trajectories from $\mathbf{x}$ to $\mathbf{y} = \mathbf{x} + \mathbf{a}\Delta t$ which cross two edges.

This approximation is achieved most conveniently by using a quadrature rule along the edge, with one based on just the vertices and the midpoint being particularly simple. For then only a mild restriction on the CFL numbers ensures that each trajectory that passes through a boundary quadrature point crosses no other cell boundary, and that the starting points $\mathbf{x}$ for such trajectories lie on a smooth curve in each cell. The integrals (3.18) can then be calculated using the one-dimensional formulae derived in Appendix A. In particular, if $a_n(u)$ is taken to be linear in u the formula (A.22), which we write as

$$\tilde{F}^+_{nb} = \frac{a^*_n f^0_{nb} + a^0_n f^*_{nb}}{a^*_n + a^0_n}, \tag{3.20}$$

is exact for a linear variation of $\tilde{U}^n$. Here a^0_n denotes the value of a_{nb} at the boundary point, and a^*_n the value at the point in the cell from which the trajectory just reaches the boundary point. The main difficulty is then to calculate this value or point, which in one dimension is given by (A.21) in the case of linear recovery, and by (A.28) in the case of quadratic recovery. We will not explore further in the present paper the approximations that may be used for this calculation.

4 Conclusions

In aeronautical fluid dynamics the unsteady formulation of the conservation laws is used as a basis for numerical approximation even though most practical problems are for steady flows. This is because advantage is taken of the consistently hyperbolic character of the unsteady equations, while the steady equations switch to a partly elliptic character as the flow changes from supersonic to subsonic. Thus although finite volume methods have been the dominant means for approximating these conservation laws in the 1980's and 1990's, less attention has been paid to the accurate approximation of the time evolution than the spatial discretisation. Indeed, the most widely used codes and methods are based on semi-discretisation with a Runge-Kutta algorithm used for the time-stepping.

On the other hand, in application areas such as meteorology or oceanography or porous media flow, accurate approximation of the time evolution is crucial – even conservation is sometimes sacrificed to achieve it (see, e.g. [41, 10, 44]). It is considered important to relate the time and spatial discretisation, third order accuracy is often aimed for and the result is that the finite volume formulation is less common.

In the present paper we have presented schemes that resolve the conflict that seems to have occurred in these various fields of application between the two highly desirable design objectives of using the conservation law form and obtaining good time accuracy. The means of achieving this have been the use of a finite volume formulation in space and time, coupled with the use of characteristic based approximations to the time-averaged fluxes. The schemes have been completely worked out for scalar problems in one dimension for up to third order accuracy, together with the recovery/reconstruction algorithms that are required at each time level. Theorems are proved to establish TV-stability for naturally occurring CFL conditions, that for the second order scheme (Theorem 2.3) being particularly sharp. Consideration of multidimension problems has guided the development of the one-dimensional algorithms, and extensions of each to unstructured triangular meshes are presented in section 3. Detailed estimates of errors are not presented here, and the reader is referred to [32] for such analysis.

The development of these methods for systems of conservation laws is at an earlier stage. Although in one dimension various characteristic decompositions can be used to calculate the interfacial fluxes, in more dimensions it is necessary to incorporate some approximation to a complete bicharacteristic cone in order to treat the time evolution accurately. An encouraging start in this direction has been made by Ostkamp [40] which is followed up in [29].

Another potential development of this approach is to use projections onto higher order approximations than the piecewise constants implied by the finite volume formulation. This falls naturally into the ambit of discontinuous Galerkin methods. However, although this class of methods have developed rapidly over the last few years – see [7] for a review – they are generally based on a semi-discrete approach and their application to genuinely unsteady problems is at an early stage. Nevertheless, it is to be expected the discontinuous Galerkin methods will provide an alternative means to higher order accuracy, to complement the use of high order recovery coupled with simple finite volume projection that forms the basis of the methods presented here.

Appendix A Interface fluxes for $U^{n+1} = (P_h E_\Delta)\tilde{U}^n$, for the scalar equation $u_t + f_x = 0$ and piecewise constant U^n

We first derive general formulae which are valid for piecewise continuous recovery functions $\tilde{U}^n$, with the discontinuities occurring at the cell interfaces $\{x_{i\pm\frac{1}{2}}\}$. Then specific formulae are given for piecewise constant $\tilde{U}^n$ (i.e. no recovery), and for piecewise linear and piecewise quadratic recovery; in the latter cases, practical approximations are given for the integrals that are needed. We use the *transport collapse operator* [5] to define E_Δ or E_τ, but carry out the calculation using the Riemann-Stieltjes formulation of Lin et al. [27] and express the result in the finite volume formulation of (2.16,2.17).

Thus with U^n given by (2.1), and an arbitrary recovery stage giving $\tilde{U}^n$, we use constancy of the solution along each characteristic to obtain

$$\begin{aligned} h_i\left(U_i^{n+1} - U_i^n\right) &\equiv \int U^{n+1}(y)\chi_i(y)dy - \int U^n(x)\chi_i(x)dx \\ &= \int \tilde{U}^n(x)\left[\chi_i(y)dy - \chi_i(x)dx\right] \\ &= \int \tilde{U}^n(x)d\left[\int_x^y \chi_i(s)ds\right] \end{aligned} \tag{A.1}$$

where $y = x + a(\tilde{U}^n(x))\Delta t$ and $a(\cdot) = f'(\cdot)$. We suppose that the graph $[\tilde{U}^n, x]$ is continuously parameterised (with x nondecreasing) so that the integrals above, and subsequent manipulations, are interpreted as Riemann-Stieltjes integrals in this parameter. Then after an integration by parts (A.1) can be written

$$h_i\left(U_i^{n+1} - U_i^n\right) + \int\left[\int_x^y \chi_i(s)ds\right]d\tilde{U}^n = 0; \tag{A.2}$$

and by writing $\chi_i(s) = H_{i+\frac{1}{2}}(s) - H_{i-\frac{1}{2}}(s)$, where $H_{i\pm\frac{1}{2}}(s)$ are Heaviside functions at $x_{i\pm\frac{1}{2}}$, we need to define the interfacial fluxes of (2.17) so that

$$\Delta_+\widetilde{F}_{i-\frac{1}{2}}^{n+\frac{1}{2}} = \Delta_+\left\{\frac{1}{\Delta t}\int\left[\int_x^y H_{i-\frac{1}{2}}(s)ds\right]d\widetilde{U}^n\right\}. \tag{A.3}$$

In deriving formulae for specific cases below, we will assume that f is convex and has a single sonic point, $\bar{u}$, at which $a(\bar{u}) = 0$. Then the inner integral on the right of (A.3) equals

$ag_{i-\frac{1}{2}}$ where the function $g_{i-\frac{1}{2}}$ is given by

$$g_{i-\frac{1}{2}} = \begin{cases} 0 & \text{if } x, y > x_{i-\frac{1}{2}} \\ (x_{i-\frac{1}{2}} - x)/a & \text{if } x < x_{i-\frac{1}{2}} \leq y, \text{ i.e. } a > 0 \\ (y - x_{i-\frac{1}{2}})/a & \text{if } y < x_{i-\frac{1}{2}} \leq x, \text{ i.e. } a < 0 \\ \Delta t & \text{if } x, y < x_{i-\frac{1}{2}}. \end{cases} \tag{A.4}$$

Moreover, in the only case not covered here in which $x = y = x_{i-\frac{1}{2}}$ and $a = 0$, there can be at most one such point on the $\left[x, \widetilde{U}^n\right]$ graph; and at such a point $g_{i-\frac{1}{2}}$ has a jump of magnitude Δt. If we first exclude this case, the complete integral can be written

$$\int a g_{i-\frac{1}{2}} d\widetilde{U}^n = \int g_{i-\frac{1}{2}} df\left(\widetilde{U}^n\right) = \left(-\Delta t f\left(\widetilde{U}^n(-\infty)\right)\right) - \int f\left(\widetilde{U}^n\right) dg_{i-\frac{1}{2}}. \tag{A.5}$$

Now we define

$$\tau = \frac{\left(x_{i-\frac{1}{2}} - x\right)}{a\left(\widetilde{U}^n(x)\right)}, \qquad \tau \in [0, \Delta t], \tag{A.6}$$

so that $dg_{i-\frac{1}{2}} = \operatorname{sgn}(a) d\tau$. But since x is nondecreasing in the $[x, \tilde{U}^n]$ graph, the τ limits are $(\Delta t, 0)$ when $a > 0$ and $(0, \Delta t)$ when $a < 0$. Also, we can neglect the constant flux term at $x = -\infty$ in (A.5) in the difference of (A.3) and hence set

$$\text{(no sonic point)} \quad \widetilde{F}^{n+\frac{1}{2}}_{i-\frac{1}{2}} = \frac{1}{\Delta t} \int_0^{\Delta t} f\left(\widetilde{U}^n(x)\right) d\tau, \tag{A.7}$$

with τ and x related by (A.6) and x, y satisfying either of the two middle conditions in (A.4).

Returning to the situation where $\tilde{U}^n$ passes through the sonic point $\bar{u}$ at $x_{i-\frac{1}{2}}$, let $s^+_{i-1} = \operatorname{sgn}(a)$ to the left and $s^-_i = \operatorname{sgn}(a)$ to the right of $x_{i-\frac{1}{2}}$ on this graph. Then the jump in $g_{i-\frac{1}{2}}$ is equal to $\frac{1}{2}(s^+_{i-1} - s^-_i)\Delta t$ in going from left to right; so when the integration by parts in (A.5) is split at this point an extra contribution $\frac{1}{2}(s^-_i - s^+_{i-1}) f(\bar{u}) \Delta t$ is added to give

$$\tilde{F}^{n+\frac{1}{2}}_{i-\frac{1}{2}} = \frac{1}{\Delta t} \int_0^{\Delta t} f\left(\tilde{U}^n(x)\right) d\tau + \tfrac{1}{2}(s^-_i - s^+_{i-1}) f(\bar{u}). \tag{A.8}$$

Formula (A.8) generalises (A.7) and corresponds to (2.17) with the transport collapse operator used for E_τ.

We now consider specific cases.

(i) *$\tilde{U}^n \equiv U^n$ and (2.3) satisfied*

Under the CFL condition (2.3) only the constant states U^n_{i-1} and U^n_i, with the jump between them, have to be taken into account in (A.8); this follows even with the equality conditions of (2.3) because the convexity of f implies that at most one value between say U^n_{i-2} and U^n_{i-1} could reach $x_{i-\frac{1}{2}}$ and therefore this jump could not contribute to the integral. We use the notation $s_k = \operatorname{sgn} a(U_k^n)$, so in (A.8) we have $s_i^- = s_i$ and $s_{i-1}^+ = s_{i-1}$. Also, x satisfies the second condition of (A.4) if $s_{i-1} > 0$, when it gives $f_{i-1}^n \equiv f\left(U_{i-1}^n\right)$; and x satisfies the third condition when $s_i < 0$, giving f_i^n. Thus we have

$$\tilde{F}_{i-\frac{1}{2}}^{n+\frac{1}{2}} = \tfrac{1}{2}(1+s_{i-1})f_{i-1}^n + \tfrac{1}{2}(s_i - s_{i-1})f(\bar{u}) + \tfrac{1}{2}(1-s_i)f_i^n, \tag{A.9}$$

as in (2.4). Note that if $a(U_k^n) = 0$, the value assigned to s_k is irrelevant because then $f_k^n = f(\bar{u})$. Note also that because the scheme is obtained from the transport collapse operator, on a uniform mesh it is equivalent to the Engquist-Osher difference scheme – see [5], [27].

(ii) $\widetilde{U}^n \equiv U^n$ *and larger CFL numbers*

Specifically, we will assume that

$$-h_{j-2} - h_{j-1} \leq a(u)\Delta t \leq h_j + h_{j+1} \text{ for } u \text{ between } U_{j-1}^n \text{ and } U_j^n \ \forall j. \tag{A.10}$$

Then there are four constant states, and the three jumps between them, that can influence the flux (A.8). The possible magnitude of $a_j^n \equiv a(U_j^n)$, and its sign, is reflected in the following parameters,

$$t_j^R = \operatorname{sgn}(a_j^n\Delta t - h_{j+1}), \quad t_j^L = \operatorname{sgn}(a_j^n\Delta t + h_{j-1}), \quad t_j^{\pm} = \operatorname{sgn}(a_j^n\Delta t \pm h_j), \tag{A.11}$$

with j running from $i-2$ to $i+1$. And at the interface $i-\frac{1}{2}$ we already have $\bar{u}$ at which $a(\bar{u}) = 0$, but we also need to introduce $U_{i-\frac{3}{2}}^*$ at the interface $i-\frac{3}{2}$ such that $a(U_{i-\frac{3}{2}}^*)\Delta t = h_{i-1}$, and $U_{i+\frac{1}{2}}^*$ at the interface $i+\frac{1}{2}$ such that $a(U_{i+\frac{1}{2}}^*)\Delta t = -h_i$. These are all the quantities we shall need to calculate (A.8).

Starting from the left, the contribution from the cell I_{i-2} depends on having $a_{i-2}^n\Delta t > h_{i-1}$ and can thus be written

$$\max\left(0, 1 - \frac{h_{i-1}}{\Delta t a_{i-2}^n}\right) f_{i-2}^n \equiv \tfrac{1}{2}\left(1 + t_{i-2}^R\right)\left(1 - \frac{h_{i-1}}{\Delta t a_{i-2}^n}\right) f_{i-2}^n.$$

Similarly, that from cell I_{i-1} depends on $s_{i-1} > 0$, and can be written

$$\tfrac{1}{2}(1+s_{i-1})\left[\min\left(1, \frac{h_{i-1}}{\Delta t a_{i-1}^n}\right) \equiv \tfrac{1}{2}(1 - t_{i-1}^-) + \tfrac{1}{2}(1+t_{i-1}^-)\frac{h_{i-1}}{\Delta t a_{i-1}^n}\right] f_{i-1}^n.$$

The contribution from the interface $i-\frac{3}{2}$, between limits u_L and u_R, has the form

$$\begin{aligned} -\frac{1}{\Delta t}\int_{u_L}^{u_R} f(U)d\left(\frac{h_{i-1}}{a(U)}\right) &= -\frac{h_{i-1}}{\Delta t}\left\{\left[\frac{f}{a}\right]_{u_L}^{u_R} - \int_{u_L}^{u_R} dU\right\} \\ &= \frac{h_{i-1}}{\Delta t}\left[(u_R - u_L) - \left(\frac{f_R}{a_R} - \frac{f_L}{a_L}\right)\right]. \end{aligned}$$

In general we will need to split the integral from U^n_{i-2} to U^n_{i-1} at the value $U^*_{i-\frac{3}{2}}$; and there will clearly be some cancellation with the terms from the two cells on either side. Combining all three contributions, we obtain

$$\begin{aligned} F^L_{i-\frac{1}{2}} \;\equiv\;& \tfrac{1}{2}\left(1+t^R_{i-2}\right)\left(f^n_{i-2}-\frac{h_{i-1}}{\Delta t}U^n_{i-2}\right)+\tfrac{1}{2}\left(t^-_{i-1}-t^R_{i-2}\right)\left(f^*_{i-\frac{3}{2}}-\frac{h_{i-1}}{\Delta t}U^*_{i-\frac{3}{2}}\right)\\ &+\tfrac{1}{4}\left(1-t^-_{i-1}\right)(1+s_{i-1})f^n_{i-1}-\tfrac{1}{2}\left(1+t^-_{i-1}\right)\frac{h_{i-1}}{\Delta t}U^n_{i-1}; \end{aligned} \tag{A.12}$$

note the similarity in form with (A.9).

In the same way, the contribution from the cells I_i and I_{i+1}, together with the interface between them, gives

$$\begin{aligned} F^R_{i-\frac{1}{2}} \;\equiv\;& \tfrac{1}{4}(1+t^+_i)(1-s_i)f^n_i-\tfrac{1}{2}(1-t^+_i)\frac{h_i}{\Delta t}U^n_i+\tfrac{1}{2}\left(t^L_{i+1}-t^+_i\right)\left(f^*_{i+\frac{1}{2}}+\frac{h_i}{\Delta t}U^*_{i+\frac{1}{2}}\right)\\ &+\tfrac{1}{2}\left(1-t^L_{i+1}\right)\left(f^n_{i+1}+\frac{h_i}{\Delta t}U^n_{i+1}\right). \end{aligned} \tag{A.13}$$

This leaves only the interface at $i-\frac{1}{2}$ itself, where the only contribution is from the jump in $g_{i-\frac{1}{2}}$, as in the difference between (A.7) and (A.8). Thus the total interface flux is

$$\tilde{F}^{n+\frac{1}{2}}_{i-\frac{1}{2}} = F^L_{i-\frac{1}{2}}+\tfrac{1}{2}(s_i-s_{i-1})f(\bar{u})+F^R_{i-\frac{1}{2}}, \tag{A.14}$$

with $F^L_{i-\frac{1}{2}}$ and $F^R_{i-\frac{1}{2}}$ given by (A.12) and (A.13) respectively.

(iii) *Linear recovery and CFL nos. ≤ 1*

In cell I_i discontinuous linear recovery gives

$$\tilde{U}^n(x) = U^n_i + (x-x_i)S^n_i; \tag{A.15}$$

and we denote $U^n_i \pm \frac{1}{2}h_iS^n_i$ by $\tilde{U}^\pm_i$, so that $s^\pm_i = \operatorname{sgn}a(\tilde{U}^\pm_i)$ in (A.8). To ensure that no characteristic reaches further than the neighbouring cell in one time step, we impose the CFL conditions

$$-h_{j-1} \leq a(u)\Delta t \leq h_j \quad \text{for } u \text{ between } \tilde{U}^+_{j-1} \text{ and } \tilde{U}^-_j\ \forall j \tag{A.16a}$$

and

$$-\min(h_{j-1},h_j) \leq a(u)\Delta t \leq \min(h_j,h_{j+1}) \quad \text{for } u \text{ between } \tilde{U}^-_j \text{ and } \tilde{U}^+_j\ \forall j, \tag{A.16b}$$

which correspond to condition (2.3) in the unrecovered case.

Suppose now that $a(\tilde{U}^n) \geq 0$ in cell I_{i-1}. Then its contribution to the integral in (A.8) will, after substitution from (A.6), take the form

$$\begin{aligned} \frac{1}{\Delta t}\int_{x_{i-\frac{1}{2}}}^{x_{i-\frac{1}{2}}-a\Delta t} f\,d\left(\frac{x_{i-\frac{1}{2}}-x}{a}\right) &= \left[\frac{f}{a\Delta t}(x_{i-\frac{1}{2}}-x)\right]-\frac{1}{\Delta t}\int(x_{i-\frac{1}{2}}-x)du\\ &= f^*_{\Delta t}-\frac{1}{\Delta t}\int(x_{i-\frac{1}{2}}-x)S^n_{i-1}dx\\ &= f^*_{\Delta t}+\tfrac{1}{2}S^n_{i-1}(a^*_{\Delta t})^2\Delta t, \end{aligned} \tag{A.17}$$

where $f^*_{\Delta t} \equiv f(u^*_{\Delta t})$, $a^*_{\Delta t} \equiv a(u^*_{\Delta t})$ and $u^*_{\Delta t}$ correspond to the values at the upper limit of the integral, where one has $x^* = x_{i-\frac{1}{2}} - a^*_{\Delta t}\Delta t$, and hence

$$u^*_{\Delta t} = U^n_{i-1} + (\tfrac{1}{2}h_{i-1} - a^*_{\Delta t}\Delta t)S^n_{i-1} \equiv \tilde{U}^+_{i-1} - a^*_{\Delta t}S^n_{i-1}\Delta t. \tag{A.18}$$

In general this is an implicit relation for the upper limit and will have to be solved approximately; and the CFL condition (A.16b) guarantees a solution in the cell.

However, we will instead solve these relations exactly for an approximate flux function. Specifically, we construct $A(\cdot)$ in each cell I_j by linear interpolation between the end values $a(\tilde{U}^\pm_j)$, and $a(\bar{u}) \equiv 0$ if $\tilde{U}^n$ takes the value $\bar{u}$ in the cell, i.e. if $s^+_j s^-_j = -1$; then the corresponding flux function $F(\cdot)$ is set to match $f(\cdot)$ at the outflow end, i.e. $F(\tilde{U}^\pm_j) = f(\tilde{U}^\pm_j)$ if $s^\pm_j = \pm 1$ with the possible inconsistency at an interior sonic point being of no consequence in the calculation of interface fluxes.

Suppose then that $s^+_j = s^-_j = +1$. The approximate upper limit flux in the cell I_j we denote by $F^{*+}_j \equiv F(U^{*+}_j)$ and is given by

$$\begin{aligned} F^{*+}_j &= f(\tilde{U}^+_j) + \int_{\tilde{U}^+_j}^{U^{*+}_j} A du \\ &= f(\tilde{U}^+_j) + \tfrac{1}{2}(U^{*+}_j - \tilde{U}^+_j)(A^{*+}_j + a(\tilde{U}^+_j)) \\ &= f(\tilde{U}^+_j) - \tfrac{1}{2}S^n_j A^{*+}_j(A^{*+}_j + a(\tilde{U}^+_j))\Delta t, \end{aligned} \tag{A.19}$$

where $A^{*+}_j = A(U^{*+}_j)$. The resulting outflow flux from (A.17) thus becomes

$$\frac{1}{\Delta t}\int_{x_{j+\frac{1}{2}}}^{x_{j+\frac{1}{2}}-a\Delta t} f d\left(\frac{x_{j+\frac{1}{2}} - x}{a}\right) \approx f(\tilde{U}^+_j) - \tfrac{1}{2}S^n_j a(\tilde{U}^+_j)A^{*+}_j\Delta t. \tag{A.20}$$

Moreover, one can solve $a(\tilde{U}^+_j) - A^{*+}_j = (A^{*+}_j\Delta t)(dA/dx = m^+_j S^n_j)$ to give

$$A^{*+}_j = \frac{a(\tilde{U}^+_j)}{1 + m^+_j S^n_j\Delta t}, \quad \text{where} \quad m^+_j = \frac{a(\tilde{U}^+_j) - a(\tilde{U}^-_j)}{h_j S^n_j}, \tag{A.21}$$

because condition (A.16b) ensures that A^{*+}_j corresponds to U^{*+}_j in the cell. Note that the convexity of $f(\cdot)$ implies that $m^+_j > 0$.

If $s^+_j = +1$, but $s^-_j = -1$, the slope m^+_j is given by $a(\tilde{U}^+_j)/(\tilde{U}^+_j - \bar{u})$ and (A.21) ensures that $0 < A^{*+}_j < a(\tilde{U}^+_j)$ since it is clear that $S^n_j > 0$; so the integral extends only over positive values of $A(\cdot)$ and part of the cell. Similar definitions of m^-_j, A^{*-}_j and F^{*-}_j are given when $s^-_j = -1$, according to whether $s^+_j = -1$ or $+1$; note that the same sign holds in the formula equivalent to (A.20), because $S^n_j > 0$ implies larger values of $\tilde{U}$ and therefore smaller values of $f(\cdot)$ in the interior.

An alternative to the approximate interface flux (A.20) has some advantages. By eliminating S^n_j between (A.19) and (A.20), we see that the latter can be replaced by an average of $f(\tilde{U}^+_j)$ and F^{*+}_j; then by replacing F^{*+}_j by $f(U^{*+}_j)$ we obtain for $s^+_j = +1$

$$\frac{1}{\Delta t}\int_{x_{j+\frac{1}{2}}}^{x_{j+\frac{1}{2}}-a\Delta t} f d\left(\frac{x_{j+\frac{1}{2}} - x}{a}\right) \approx \frac{A^{*+}_j f(\tilde{U}^+_j) + a(\tilde{U}^+_j)f(\tilde{U}^+_j - S^n_j A^{*+}_j\Delta t)}{A^{*+}_j + a(\tilde{U}^+_j)}, \tag{A.22}$$

a formula which equals (A.20) when $a(\cdot)$ is linear in cell I_j.

Denoting the right-hand side of (A.22) by $\tilde{F}_j^+$, and defining $\tilde{F}_j^-$ when $s_j^- = -1$ by the same expression based on $\tilde{U}_j^-$ and m_j^-, we arrive at the interface flux at $x_{i-\frac{1}{2}}$

$$\tilde{F}_{i-\frac{1}{2}}^{n+\frac{1}{2}} = \tfrac{1}{2}\Big[(1-s_i^-)\tilde{F}_i^- + (s_i^- - s_{i-1}^+)f(\bar{u}) + (1+s_{i-1}^+)\tilde{F}_{i-1}^+\Big]. \tag{A.23}$$

This corresponds to (A.8) for this approximation and, through the signs s_{i-1}^+ and s_i^-, expresses the fact that the integral there can include contributions from either, both or neither of the cells I_{i-1} and I_i.

(iv) *Quadratic recovery and CFL nos. ≤ 1*

We can extend the algorithm given above for linear recovery to cases where quadratic recovery is used, at the cost of a more complicated calculation of $A_j^{*\pm}$. Although there may now be two sonic points in a cell, $m_j^\pm$ can still be calculated by (2.36) if this is taken to refer to the nearest such point; and we allow the recovery to be discontinuous so that $\tilde{U}_j^\pm$ and $s_j^\pm$ are defined as before. Suppose, then, that in cell I_j we have

$$\frac{d\tilde{U}}{dx} = \beta_j^\pm + \gamma_j(x - x_{j\pm\frac{1}{2}}), \tag{A.24}$$

which is a convenient notation to be used at the two cell ends. We will calculate $\tilde{F}_j^+$, and $\tilde{F}_j^-$ is given by an analogous formula.

For $s_j^+ = +1$, we have the notation $U_j^{*+} \equiv \tilde{U}(x_{j+\frac{1}{2}} - A_j^{*+}\Delta t)$, $A_j^{*+} \equiv A(U_j^{*+})$, $F_j^{*+} \equiv f(U_j^{*+})$ and instead of (A.17) we have

$$\begin{aligned}\tilde{F}_j^+ &= F_j^{*+} - \frac{1}{\Delta t}\int_{x_{j+\frac{1}{2}}}^{x_{j+\frac{1}{2}}-A_j^{*+}\Delta t}(x_{j+\frac{1}{2}} - x)d\tilde{U} \\ &= F_j^{*+} + \tfrac{1}{2}\beta_j^+(A_j^{*+})^2\Delta t - \tfrac{1}{3}\gamma_j(A_j^{*+})^3\Delta t^2.\end{aligned} \tag{A.25}$$

Similarly, (A.19) becomes

$$\begin{aligned}F_j^{*+} - f(\tilde{U}_j^+) &= \tfrac{1}{2}(U_j^{*+} - \tilde{U}_j^+)(A_j^{*+} + a(\tilde{U}_j^+)) \\ &= \tfrac{1}{2}[A_j^{*+} + a(\tilde{U}_j^+)][-\beta_j^+A_j^{*+}\Delta t + \tfrac{1}{2}\gamma_j(A_j^{*+}\Delta t)^2].\end{aligned} \tag{A.26}$$

Combining the two formulae eliminates β_j^+ and gives

$$\tilde{F}_j^+ = \frac{A_j^{*+}f(\tilde{U}_j^+) + a(\tilde{U}_j^+)F_j^{*+}}{A_j^{*+} + a(\tilde{U}_j^+)} - \tfrac{1}{12}\gamma_j(A_j^{*+})^3\Delta t^2, \tag{A.27}$$

to replace (A.22).

It remains to calculate A_j^{*+}, by integrating $dA/dx \equiv m_j^+ d\tilde{U}/dx$ to get

$$A_j^{*+} - a(\tilde{U}_j^+) = m_j^+[-\beta_j^+A_j^{*+}\Delta t + \tfrac{1}{2}\gamma_j(A_j^{*+}\Delta t)^2], \tag{A.28}$$

which is a quadratic equation for A_j^{*+}. We require the root which becomes $a(\tilde{U}_j^+)$ as $\Delta t \to 0$, and it is shown in the lemma below this can normally be found by simple iteration of the following relations. If $\gamma_j \geq 0$, we write

$$A_j^{*+} = \frac{a(\tilde{U}_j^+)}{1 + m_j^+ \Delta t(\beta_j^+ - \frac{1}{2}\gamma_j A_j^{*+}\Delta t)}, \tag{A.29a}$$

and start the iteration by setting $A_j^{*+} = 0$ on the right-hand side; and if $\gamma_j < 0$, we write

$$A_j^{*+} = \frac{a(\tilde{U}_j^+) - \frac{1}{2}\gamma_j m_j^+ A_j^{*+} h_j \Delta t}{1 + m_j^+ \Delta t[\beta_j^+ - \frac{1}{2}\gamma_j(h_j + A_j^{*+}\Delta t)]}, \tag{A.29b}$$

and again start by setting $A_j^{*+} = 0$ on the right-hand side. Thus both cases generalise and start from the linear recovery case (A.21); and they are similarly modified in the presence of a sonic point in the cell, and to calculate A_j^{*-}.

From A_j^{*+} one obtains U_j^{*+}, and hence F_j^{*+} to substitute into (A.27). Then when $s_j^- = -1$, with $\beta_j^-, x_{j-\frac{1}{2}}, \tilde{U}_j^-$ and m_j^- in (A.29) and (A.27) we obtain A_j^{*-}, F_j^{*-} and hence $\tilde{F}_j^-$.

Lemma A.1 *If $\tilde{U}$ is monotone in cell I_j, then the iterations (A.29) converge monotonically to the required solution of (A.28).*

Proof Writing $\nu^\pm$ for $a(\tilde{U}_j^\pm)\Delta t/h_j$, ν for $A_j^{*+}\Delta t/h_j$ and ΔU for $\tilde{U}_j^+ - \tilde{U}_j^-$, (A.28) becomes

$$\nu - \nu^+ + \nu(\nu^+ - \nu^-)[(\beta_j^+ h_j/\Delta U) - \tfrac{1}{2}(\gamma_j h_j^2/\Delta U)\nu] = 0. \tag{A.30}$$

From the formula (2.56) for the quadratic $\tilde{U}$, the coefficients in (2.72) are given by $\beta_j^+ = (\Delta U + 3\delta^2 U)/h_j$ and $\gamma_j = 6\delta^2 U/h_j^2$, where $\delta^2 U = \tilde{U}_j^+ + \tilde{U}_j^- - 2U_j$. Moreover, the monotonicity condition is given by (2.71), namely $3|\delta^2 U| \leq |\Delta U|$. Hence (A.30) can be rewritten as

$$Q(\nu) \equiv \nu - \nu^+ + \nu(\nu^+ - \nu^-)[1 + \gamma'(1-\nu)] = 0, \tag{A.31}$$

where $|\gamma'| \leq 1$. We have $Q(0) = -\nu^+$, $Q(1) = 1 - \nu^-$ and the straight line between these points crosses the axis at $\nu^{(0)} = \nu^+/(1 + \nu^+ - \nu^-)$, which lies between ν^+ and ν^-. If $\gamma'(\nu^+ - \nu^-) > 0$, then $Q''(\cdot) < 0$ and the zero of Q can be found by pivoting a straight line around $Q(0) = -\nu^+$ and the current (positive) value $Q(\nu^{(i)})$. Similarly, if $\gamma'(\nu^+ - \nu^-) < 0$, then $Q''(\cdot) > 0$ and the zero is found by pivoting a straight line around $Q(1) = 1 - \nu^-$ and the current (negative) value $Q(\nu^{(i)})$. Elementary arguments show that the convergence is linear and monotone in either case. Since $\gamma'(\nu^+ - \nu^-) \equiv 3(\delta^2 U/\Delta U)(\nu^+ - \nu^-)$, it is proportional to γ_j whose sign therefore determines which iteration should be used. □

Appendix B Proofs of Lemmas 2.4 and 2.5

Proof of Lemma 2.4

The regrouping of the terms in $\tilde{F}_{i-\frac{1}{2}}^{n+\frac{1}{2}}$ and $\tilde{F}_{i+\frac{1}{2}}^{n+\frac{1}{2}}$ given in (2.48) leads, with $\bar{f} \equiv f(\bar{u})$, to

$$C_{i-\frac{1}{2}}\Delta_- U_i^n = \tfrac{1}{2}\frac{\Delta t}{h_i}\Big\{[(1 + s_i^+)\tilde{F}_i^+ + (1 - s_i^+)\bar{f}] - [(1 + s_{i-1}^+)\tilde{F}_{i-1}^+ + (1 - s_{i-1}^+)\bar{f}]\Big\}. \tag{B.1}$$

If we write θ_i for $A_i^{*+}/[A_i^{*+} + a(\tilde{U}_i^+)]$, $\Delta_-\tilde{F}_i^+$ consists of the average of four flux differences: $f(\tilde{U}_i^+) - f(U_{i-1}^+) =: a_{i-\frac{1}{2}}^{++}(\tilde{U}_i^+ - \tilde{U}_{i-1}^+)$ with weight $\theta_i\theta_{i-1}$, $f(\tilde{U}_i^+) - f(U_{i-1}^{*+}) =: a_{i-\frac{1}{2}}^{+*}(\tilde{U}_i^+ - U_{i-1}^{*+})$ with weight $(1-\theta_i)\theta_{i-1}$, and so on, where all the characteristic speeds $a_{i-\frac{1}{2}}^{++}$, $a_{i-\frac{1}{2}}^{+*}, \ldots$ are positive. Now $U_i^{*+} = U_i^n + (\frac{1}{2}h_i - A_i^{*+}\Delta t)S_i^n \equiv U_i^n + \frac{1}{2}h_i S_i^n(1-2\nu_i^*)$, where $\nu_i^* = A_i^{*+}\Delta t/h_i$. So we can write

$$\begin{aligned}
\Delta_-\tilde{F}_i^+ &= \theta_i\theta_{i-1}a_{i-\frac{1}{2}}^{++}[\Delta_- U_i^n + \tfrac{1}{2}h_i S_i^n - \tfrac{1}{2}h_{i-1}S_{i-1}^n] \\
&+ \theta_i(1-\theta_{i-1})a_{i-\frac{1}{2}}^{+*}[\Delta_- U_i^n + \tfrac{1}{2}h_i S_i^n - \tfrac{1}{2}h_{i-1}S_{i-1}^n(1-2\nu_{i-1}^*)] \\
&+ (1-\theta_i)\theta_{i-1}a_{i-\frac{1}{2}}^{*+}[\Delta_- U_i^n + \tfrac{1}{2}h_i S_i^n(1-2\nu_i^*) - \tfrac{1}{2}h_{i-1}S_{i-1}^n] \\
&+ (1-\theta_i)(1-\theta_{i-1})a_{i-\frac{1}{2}}^{**}[\Delta_- U_i^n + \tfrac{1}{2}h_i S_i^n(1-2\nu_i^*) - \tfrac{1}{2}h_{i-1}S_{i-1}^n(1-2\nu_{i-1}^*)]. \quad \text{(B.2)}
\end{aligned}$$

This is the key term in (B.1), corresponding to $s_{i-1}^+ = s_i^+ = +1$: when s_{i-1}^+ or s_i^+ equals -1, $\tilde{U}_{i-1}^+$ and U_{i-1}^{*+} or $\tilde{U}_i^+$ and U_i^{*+} are replaced by $\bar{u}$ to give a similar expression to (B.2); and when $s_{i-1}^+ = s_i^+ = -1$, $C_{i-\frac{1}{2}} = 0$. We will concentrate on this first case.

Introducing $R_1 = \frac{1}{2}h_i S_i^n/\Delta_- U_i^n$ and $R_2 = \frac{1}{2}h_{i-1}S_{i-1}^n/\Delta_- U_i^n$, we see that (2.46) implies that $R_1, R_2 \in [0,1]$. Then we can write

$$\Delta_-\tilde{F}_i^+ = (\alpha + \beta R_1 - \gamma R_2)\Delta_- U_i^n, \quad \text{(B.3)}$$

where

$$\alpha = \theta_i\theta_{i-1}a_{i-\frac{1}{2}}^{++} + \theta_i(1-\theta_{i-1})a_{i-\frac{1}{2}}^{+*} + (1-\theta_i)\theta_{i-1}a_{i-\frac{1}{2}}^{*+} + (1-\theta_i)(1-\theta_{i-1})a_{i-\frac{1}{2}}^{**} \quad \text{(B.4)}$$

while β is composed of a similar four terms but with the extra factors 1, 1, $(1-2\nu_i^*)$, $(1-2\nu_i^*)$ respectively, and γ has the extra factors 1, ,$(1-2\nu_{i-1}^*)$, 1, $(1-2\nu_{i-1}^*)$.

Now suppose that $\Delta_- U_i^n > 0$, so that $S_i^n \geq 0$ and $S_{i-1}^n \geq 0$. Collecting terms in (B.3) and (B.4) we have

$$\begin{aligned}
\Delta_-\tilde{F}_i^+/\Delta_- U_i^n &= \theta_i[\theta_{i-1}a_{i-\frac{1}{2}}^{++} + (1-\theta_{i-1})a_{i-\frac{1}{2}}^{+*}](1 + R_1 - R_2) \\
&+ (1-\theta_i)[\theta_{i-1}a_{i-\frac{1}{2}}^{*+} + (1-\theta_{i-1})a_{i-\frac{1}{2}}^{**}][1 - R_2 + R_1(1-2\nu_i^*)] \\
&+ (1-\theta_{i-1})[\theta_i a_{i-\frac{1}{2}}^{+*} + (1-\theta_i)a_{i-\frac{1}{2}}^{**}]2R_2\nu_{i-1}^*. \quad \text{(B.5)}
\end{aligned}$$

It follows from (A.21) that $A_i^{*+} \leq a(\tilde{U}_i^+)$ so that $\theta_i \leq \frac{1}{2}$, and similarly $\theta_{i-1} \leq \frac{1}{2}$. Moreover, it is easy to show by Taylor series expansion that, because $f'' > 0$, for any u, v, w and some ξ in their range

$$\frac{f(u) - f(w)}{u - w} - \frac{f(v) - f(w)}{v - w} = \tfrac{1}{2}(u - v)f''(\xi). \quad \text{(B.6)}$$

Hence $a_{i-\frac{1}{2}}^{*+} \leq a_{i-\frac{1}{2}}^{++}$ and $a_{i-\frac{1}{2}} \leq a_{i-\frac{1}{2}}^{+*}$; and we also observe that $(1-\theta_i)\nu_i^* = \theta_i a(\tilde{U}_i^+)\Delta t/h_i \leq \theta_i \leq \frac{1}{2}$ by the CFL condition. So in this case of $s_{i-1}^+ = s_i^+ = +1$, from (B.5) we have

$$\begin{aligned}
C_{i-\frac{1}{2}} &\geq \theta_{i-1}\Big\{(1 + R_1 - R_2)[\theta_i a_{i-\frac{1}{2}}^{++} + (1-\theta_i)a_{i-\frac{1}{2}}^{*+}] - 2R_1\theta_i a_{i-\frac{1}{2}}^{*+}\Big\} \\
&\quad + (1-\theta_i)\Big\{(1 + R_1 - R_2)[\theta_i a_{i-\frac{1}{2}}^{+*} + (1-\theta_i)a_{i-\frac{1}{2}}^{**}] - 2R_1\theta_i a_{i-\frac{1}{2}}^{**}\Big\} \\
&\geq [\theta_{i-1}a_{i-\frac{1}{2}}^{*+} + (1-\theta_{i-1})a_{i-\frac{1}{2}}^{**}][(1 + R_1 - R_2) - R_1] \geq 0, \quad \text{(B.7)}
\end{aligned}$$

as required.

The alternative case when $\Delta_- U_i^n < 0$, and hence $S_i^n \leq 0$ and $S_{i-1}^n \leq 0$, needs a more delicate argument. We now have $\theta_i \geq \frac{1}{2}$, $\theta_{i-1} \geq \frac{1}{2}$, $a_{i-\frac{1}{2}}^{*+} \geq a_{i-\frac{1}{2}}^{++}$ and $a_{i-\frac{1}{2}}^{**} \geq a_{i-\frac{1}{2}}^{+*}$, but need consider only the situation when $\nu_i^* > \frac{1}{2}$. Also, instead of (B.6) it is easy to show that for $u \geq v$

$$\frac{f(u)-f(w)}{u-w} - \frac{f(v)-f(w)}{v-w} \leq (u-v)\langle f'' \rangle, \tag{B.8}$$

where $\langle f'' \rangle$ denotes the average of f'' over the range of the arguments u, v, w, which also equals $[f']/[u]$ where $[\cdot]$ denotes the difference over the range. Thus, since the TVD condition (2.46) implies that $\tilde{U}_{i-1}^+ \geq U_i^n > \tilde{U}_i^+$,

$$a_{i-\frac{1}{2}}^{*+} - a_{i-\frac{1}{2}}^{++} \leq (U_i^{*+} - \tilde{U}_i^+)\frac{a(U) - a(\tilde{U}_i^+)}{U - \tilde{U}_i^+}, \quad a_{i-\frac{1}{2}}^{**} - a_{i-\frac{1}{2}}^{+*} \leq (U_i^{*+} - \tilde{U}_i^+)\frac{a(V) - a(\tilde{U}_i^+)}{V - \tilde{U}_i^+},$$

where $U = \max(U_i^{*+}, \tilde{U}_{i-1}^+)$ and $V = \max(U_i^{*+}, U_{i-1}^{*+})$. Then, with $(1-\theta_i)/\theta_i = a(\tilde{U}_i^+)/A_i^{*+} = 1 + m_i^+ S_i^n \Delta t$ and $U_i^{*+} - \tilde{U}_i^+ = -S_i^n A_i^{*+} \Delta t$, we have in this case

$$\begin{aligned} C_{i-\frac{1}{2}} \quad &\geq \quad R_1\theta_i \Bigg\{ \theta_{i-1} \left[a_{i-\frac{1}{2}}^{++} + \frac{a(\tilde{U}_i^+)}{A_i^{*+}} a_{i-\frac{1}{2}}^{*+}(1-2\nu_i^*) \right] \\ &\qquad + (1-\theta_{i-1}) \left[a_{i-\frac{1}{2}}^{+*} + \frac{a(\tilde{U}_i^+)}{A_i^{*+}} a_{i-\frac{1}{2}}^{**}(1-2\nu_i^*) \right] \Bigg\}, \end{aligned} \tag{B.9}$$

in which we write the quantity in the first set of square brackets as

$$\begin{aligned} &a_{i-\frac{1}{2}}^{++} + (1-2\nu_i^*)(1 + m_i^+ S_i^n \Delta t)[a_{i-\frac{1}{2}}^{++} - A_i^{*+} S_i^n \Delta t(\Delta a/\Delta U)] \\ &\qquad = 2(1-\nu_i^*)a_{i-\frac{1}{2}}^{++} + (1-2\nu_i^*)S_i^n \Delta t[a_{i-\frac{1}{2}}^{++} m_i^+ - a(\tilde{U}_i^+)(\Delta a/\Delta U)], \end{aligned} \tag{B.10}$$

where $\Delta a/\Delta U$ is the ratio given above. Now $a(\tilde{U}_i^+) \leq a_{i-\frac{1}{2}}^{++}$ and if we had $m_i^+ \geq \Delta a/\Delta U$ this square bracket would give a positive result. But $m_i^+ = [a(\tilde{U}_i^-) - a(\tilde{U}_i^+)]/(\tilde{U}_i^- - \tilde{U}_i^+)$, so from (B.6) we see that $m_i^+ < \Delta a/\Delta U$ implies that $U \equiv \max(U_i^{*+}, \tilde{U}_{i-1}^+) > \tilde{U}_i^-$ so that we must have $U = \tilde{U}_{i-1}^+$ in that case; and we then write $\alpha = (\tilde{U}_i^- - \tilde{U}_i^+)/(\tilde{U}_{i-1}^+ - \tilde{U}_i^+) \in [0, 1)$. Thus the quantity in (B.10) is bounded below by zero or by

$$\begin{aligned} &2a_{i-\frac{1}{2}}^{++}\left\{1 - \nu_i^* + \tfrac{1}{2}(\Delta t/h_i)[a(\tilde{U}_i^-) - (1-\alpha)a(\tilde{U}_i^+) - \alpha a(\tilde{U}_{i-1}^+)]\right\} \\ &\qquad \geq 2a_{i-\frac{1}{2}}^{++}\left\{1 - \tfrac{1}{2}(\Delta t/h_i)[a(\tilde{U}_i^-) + (1-\alpha)a(\tilde{U}_i^+) + \alpha a(\tilde{U}_{i-1}^+)]\right\} \end{aligned}$$

because $a(\tilde{U}_i^-) \geq A_i^{*+}$, and this quantity is nonnegative by the CFL conditions of (A.16). The same arguments can be applied to the quantity in the second set of brackets in (B.9) to give $C_{i-\frac{1}{2}} \geq 0$ in this case.

To complete the proof we need to consider the other possible cases. If $s_{i-1}^+ = -1$, $\bar{u}$ replaces $\tilde{U}_{i-1}^+$ and U_{i-1}^{*+} in (B.5); and it is easy to check that one must have $\Delta_- U_i^n \geq 0$ by the TVD conditions (2.46), in which case the arguments in (B.7) can be applied. If $s_i^+ = -1$, $\bar{u}$ replaces $\tilde{U}_i^+$ and U_i^{*+} in (B.5) so that $a_{i-\frac{1}{2}}^{++} = a_{i-\frac{1}{2}}^{*+}$, $a_{i-\frac{1}{2}}^{+*} = a_{i-\frac{1}{2}}^{**}$ and it is simple to see that $C_{i-\frac{1}{2}} \geq 0$. Finally, all the same arguments can be applied to the left-moving waves to show that $D_{i-\frac{1}{2}} \geq 0$, after a right-left reflection. □

Proof of Lemma 2.5

There are sixteen sign combinations that have to be considered, corresponding to the values $s_{i-1}^{\pm} = \pm 1$, $s_i^{\pm} = \pm 1$, and we denote them by the signs in the order of the x-variable, namely s_{i-1}^-, s_{i-1}^+, s_i^-, s_i^+. Thus $+-+-$ corresponds to the case when $s_{i-1}^+ = s_i^+ = -1$, $s_{i-1}^- = s_i^- = +1$ so that $C_{i-\frac{1}{2}} = D_{i-\frac{1}{2}} = 0$. Other cases can be eliminated by the TVD conditions. Thus since $a(\cdot)$ is monotone increasing, in the case $-++-$ we must have $S_{i-1}^n > 0$ and $S_i^n < 0$; but they cannot then both have the sign of $\Delta_- U$, as required by (2.46). By reflection about $U = \bar{u}$ we similarly eliminate $+--+$.

More generally, the sign pattern $+-$ in cell I_j implies both $U_j^n + \frac{1}{2}h_j S_j^n \leq \bar{u}$ and $U_j^n - \frac{1}{2}h_j S_j^n \geq \bar{u}$, while the pattern $-+$ implies $U_j^n + \frac{1}{2}h_j S_j^n \geq \bar{u}$ and $U_j^n - \frac{1}{2}h_j S_j^n \leq \bar{u}$; that is,

$$+- \Rightarrow \tfrac{1}{2}h_j S_j^n \leq |U_j^n - \bar{u}|, \quad -+ \Rightarrow \tfrac{1}{2}h_j S_j^n \geq |U_j^n - \bar{u}|. \tag{B.11}$$

Thus in the two-cell pattern $-+--$, (2.46) and (B.11) together imply that $\Delta_- U_i^n \geq \frac{1}{2}h_{i-1}S_{i-1}^n \geq |U_{i-1}^n - \bar{u}| \geq 0$; but clearly $U_i^n < \bar{u}$, so that we have $U_{i-1}^n \leq U_i^n < \bar{u}$ and hence $\Delta_- U_i^n < \bar{u} - U_{i-1}^n$ which is a contradiction. In the same way, by reflection about $U = \bar{u}$ and left-right reflection, we can preclude cases $+-++$, $--+-$ and $++-+$.

This leaves nine cases to be considered. Suppose first that all characteristic speeds are positive, as in the case $++++$. Then $D_{i-\frac{1}{2}} = 0$ and $C_{i-\frac{1}{2}}$ is given by (2.51a), which we write as

$$C_{i-\frac{1}{2}} = \lambda^+ + [\lambda^+ - (\nu^+)^2]R_1 - [\lambda^+ - (\mu^+)^2]R_2, \tag{B.12}$$

where we have introduced $R_1 = \frac{1}{2}h_i S_i^n / \Delta_- U_i^n$, $R_2 = \frac{1}{2}h_{i-1}S_{i-1}^n / \Delta_- U_i^n$ and the TVD conditions (2.46) imply that $R_1, R_2 \in [0,1]$. The basic CFL conditions of (A.16) imply that $\lambda_1 \in [0,1]$; so (2.55) is satisfied if $R_1 = R_2 = 0$, as when there is no recovery procedure applied. Then (2.47a), (2.47b) and (2.47c) ensure that (2.55) is satisfied in the cases $R_1 = 1$ and $R_2 = 0$, $R_1 = 0$ and $R_2 = 1$, and $R_1 = R_2 = 1$, respectively. Thus, although λ^+, μ^+, ν^+ depend implicitly on R_1 and R_2, we can write

$$\begin{aligned} C_{i-\frac{1}{2}} - 1 \quad \equiv \quad & [\lambda^+ - 1](1 - R_1)(1 - R_2) + [(\mu^+)^2 - 1](1 - R_1)R_2 \\ & + [2\lambda^+ - (\nu^+)^2 - 1]R_1(1 - R_2) + [\lambda^+ - (\nu^+)^2 + (\mu^+)^2 - 1]R_1 R_2. \end{aligned} \tag{B.13}$$

All the quantities in square brackets are non-positive, and hence so is their combination, for all values R_1, R_2.

The case $----$, corresponding to only left-moving waves, can be dealt with in a similar way using the expression (2.54a) for $D_{i-\frac{1}{2}}$. Then the two sets of waves are combined in the case $-+-+$, where (2.51a) and (2.54a) are added to give

$$C_{i-\frac{1}{2}} + D_{i-\frac{1}{2}} = \lambda^+ + \lambda^- + [\lambda^+ - (\nu^+)^2 - \lambda^- + (\nu^-)^2]R_1 - [\lambda^+ - (\mu^+)^2 - \lambda^- + (\mu^-)^2]R_2. \tag{B.14}$$

However, by (B.11) the sign patterns imply that $\Delta_- U_i^n \geq \frac{1}{2}h_j S_j^n \geq |U_j^n - \bar{u}|$ for both $j = i-1$ and i; so we have $U_i^n \geq \bar{u} \geq U_{i-1}^n$, can set $U_i^n - \bar{u} = \theta \Delta_- U_i^n$, and have $\theta \leq R_1 \leq 1$, $1-\theta \leq R_2 \leq 1$ for some $\theta \in [0,1]$. If we denote by α, β, γ the coefficients in (B.14) so that $C_{i-\frac{1}{2}} + D_{i-\frac{1}{2}} = \alpha + \beta R_1 - \gamma R_2$, we can write

$$\begin{aligned}\theta(1-\theta)(C_{i-\frac{1}{2}} + D_{i-\frac{1}{2}}) \;\equiv\; & [\alpha + \theta\beta - (1-\theta)\gamma](1-R_1)(1-R_2) \\ & + [\alpha + \beta - (1-\theta)\gamma](R_1 - \theta)(1-R_2) \\ & + [\alpha + \theta\beta - \gamma](R_2 - 1 + \theta)(1-R_1) \\ & + [\alpha + \beta - \gamma](R_1 - \theta)(R_2 - 1 + \theta). \qquad \text{(B.15)}\end{aligned}$$

Thus $C_{i-\frac{1}{2}} + D_{i-\frac{1}{2}}$ is an average of the four quantities in square brackets for any values of R_1, R_2 in their allowed range. Moreover, these quantities are bounded by unity if we have $\alpha + \beta \leq 1$, $\alpha - \gamma \leq 1$ and $\alpha + \beta - \gamma \leq 1$, since each is an average of two of the left-hand sides; but these are just the inequalities (2.47).

The remaining six cases involve the sonic point $\bar{u}$. Thus, for $--++$, $C_{i-\frac{1}{2}}$ is given by (2.51b) and $D_{i-\frac{1}{2}}$ by (2.54b) so that

$$C_{i-\frac{1}{2}} + D_{i-\frac{1}{2}} = \lambda^+ \frac{U_i^n - \bar{u}}{\Delta_- U_i^n} + \lambda^- \frac{\bar{u} - U_{i-1}^n}{\Delta_- U_i^n} + [\lambda^+ - (\nu^+)^2]R_1 + [\lambda^- - (\mu^-)^2]R_2. \qquad \text{(B.16)}$$

But from the sign pattern we deduce that $U_{i-1}^n + \frac{1}{2}h_{i-1}S_{i-1}^n \leq \bar{u}$ and $U_i^n - \frac{1}{2}h_i S_i^n \geq \bar{u}$; hence, since clearly $\Delta_- U_i^n \geq 0$, we have $0 \leq \frac{1}{2}h_{i-1}S_{i-1}^n \leq \bar{u} - U_{i-1}^n$ and $0 \leq \frac{1}{2}h_i S_i^n \leq U_i^n - \bar{u}$. Setting $U_i^n - \bar{u} = \theta\Delta_- U_i^n$, $\bar{u} - U_{i-1}^n = (1-\theta)\Delta_- U_i^n$ and $\frac{1}{2}h_i S_i^n = \phi\theta\Delta_- U_i^n$, $\frac{1}{2}h_{i-1}S_{i-1}^n = \psi(1-\theta)\Delta_- U_i^n$ for θ, ϕ, $\psi \in [0,1]$, we get

$$C_{i-\frac{1}{2}} + D_{i-\frac{1}{2}} = [\lambda^+(1+\phi) - \phi(\nu^+)^2]\theta + [\lambda^-(1+\psi) - \psi(\mu^-)^2](1-\theta). \qquad \text{(B.17)}$$

The first square bracket is an average of λ^+ and $2\lambda^+ - (\nu^+)^2$; the second an average of λ^- and $2\lambda^- - (\mu^-)^2$. Since in this case we have set $\mu^+ = 0$ and $\nu^- = 0$, the inequalities (2.47a) and (2.47b) ensure that all these quantities are less than unity, and hence so is (B.17). The case $++--$ is dealt with in the same way.

There remains the case $---+$, together with the similar cases $+++-$, $+---$ and $-+++$. Now $C_{i-\frac{1}{2}}$ is given by (2.51b) and $D_{i-\frac{1}{2}}$ by (2.54a); and from (B.11) we know that $\frac{1}{2}h_i S_i^n \geq |U_i^n - \bar{u}|$. So in the sum

$$C_{i-\frac{1}{2}} + D_{i-\frac{1}{2}} = \lambda^+ \frac{U_i^n - \bar{u}}{\Delta_- U_i^n} + \lambda^- + [\lambda^+ - (\nu^+)^2 - \lambda^- + (\nu^-)^2]R_1 + [\lambda^- - (\mu^-)^2]R_2, \qquad \text{(B.18)}$$

which we write as $\alpha' + \beta' R_1 - \gamma' R_2$ as we did for (B.14), we will distinguish two cases according to whether $U_i^n > \bar{u}$ or $U_i^n \leq \bar{u}$. In the former case, we put $U_i^n - \bar{u} = \theta\Delta_- U_i^n$ with $\theta \in (0,1)$, and have $\theta \leq R_1 \leq 1$ while $U_{i-1}^n + \frac{1}{2}h_{i-1}S_{i-1}^n \leq \bar{u}$ implies that $0 \leq R_2 \leq 1-\theta$. Thus we rewrite (B.18) as

$$\begin{aligned}(1-\theta)^2(C_{i-\frac{1}{2}} + D_{i-\frac{1}{2}}) \equiv\; & [\alpha' + \theta\beta'](1-R_1)(1-\theta-R_2) + [\alpha' + \beta'](R_1 - \theta)(1-\theta-R_2) \\ & + [\alpha' + \theta\beta' - (1-\theta)\gamma']R_2(1-R_1) + [\alpha' + \beta' - (1-\theta)\gamma'](R_1-\theta)R_2. \qquad \text{(B.19)}\end{aligned}$$

Now, compared with the values α, β, γ used in (B.14), we have $\alpha' = \alpha - (1-\theta)\lambda^+$, $\beta' = \beta$ and $\gamma' = \gamma - \lambda^+$ with $\mu^+ = 0$. So the CFL conditions of (2.47) become $\alpha' + \beta + (1-\theta)\lambda^+ \leq 1$, $\alpha' - \gamma' - \theta\lambda^+ \leq 1$ and $\alpha' + \beta - \gamma' - \theta\lambda^+ \leq 1$. The first clearly implies $\alpha' + \beta \leq 1$, and the quantities in the other three square brackets of (B.19) are averages of these three left-hand sides together with $\alpha' - \theta\lambda^+ \equiv \lambda^-$. Thus all the square brackets are bounded by unity, as therefore is $C_{i-\frac{1}{2}} + D_{i-\frac{1}{2}}$.

In the alternative case, where $U_i^n \leq \bar{u}$, the first term in (B.18) is nonpositive but we can place no restrictions on the range of values of R_1, R_2 — other than that they lie in $[0,1]$. Thus a bound for $C_{i-\frac{1}{2}} + D_{i-\frac{1}{2}}$ is given by the limit of the right-hand side of (B.19) as $\theta \to 0$, which therefore establishes (2.55) in this final case. □

References

[1] R. Abgrall. On Essentially Non-Oscillatory schemes on unstructured meshes: analysis and implementation. *J. Comput. Phys.*, 114:45–58, 1994.

[2] T. J. Barth. Aspects of unstructured grids and finite-volume solvers for the Euler and Navier-Stokes equations. AGARD Report 787, pp. 6.1–6.61, 1992.

[3] T. J. Barth and P. O. Frederickson. Higher order solution of the Euler equations on unstructured grids using quadratic interpolation. AIAA 28th Aerospace Science Meeting, Paper 90-0013, 1990.

[4] J. B. Bell, P. Colella, and J. A. Trangenstein. Higher order Godunov methods for general systems of hyperbolic conservation laws. *J. Comput. Phys.*, 82:362–397, 1989.

[5] Y. Brenier. Average multivalued solutions for scalar conservation laws. *SIAM J. Numer. Anal.*, 27:553–594, 1984.

[6] P. N. Childs and K. W. Morton. Characteristic Galerkin methods for scalar conservation laws in one dimension. *SIAM J. Numer. Anal.*, 27:553–594, 1990.

[7] B. Cockburn, G. E. Karniadakis, and C. W. Shu. An overview of the development of discontinuous Galerkin methods. In B. Cockburn, G. E. Karniadakis, and C. W. Shu, editors, *First International Symposium on Discontinuous Galerkin Methods*, volume 11 of *Lecture Notes in Computational Science and Engineering.* Springer Verlag, 2000.

[8] P. Colella and P. R. Woodward. The piecewise parabolic method (PPM) for gas-dynamical simulations. *J. Comput. Phys.*, 54:174–201, 1984.

[9] P. I. Crumpton, J. A. Mackenzie, and K. W. Morton. Cell vertex algorithms for the compressible Navier-Stokes equations. *J. Comput. Phys.*, 109:1–15, 1993.

[10] H. K. Dahle, R. E. Ewing, and T. F. Russell. Eulerian-Lagrangian localised adjoint methods for a nonlinear convection-diffusion equation. *Comput. Methods Appl. Mech. Engrg.*, 122:223–250, 1995.

[11] S. F. Davis. Simplified second-order Godunov-type methods. *SIAM J. Sci. Stat. Comput.*, 9:445–473, 1988.

[12] I. Dawkins. *Development of practical evolution Galerkin algorithms on unstructured meshes.* PhD thesis, University of Oxford, 1997.

[13] L. J. Durlofski, B. Engquist, and S. Osher. Triangle based adaptive stencils for the solution of hyperbolic conservation laws. *J. Comput. Phys.*, 98:64–73, 1992.

[14] B. Engquist and S. Osher. One-sided difference approximations for nonlinear conservation laws. *Math. Comp.*, 36:321–352, 1981.

[15] S. K. Godunov. A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics. *Mat. Sb.*, 47:271–306, 1959.

[16] J. Goodman and R. J. LeVeque. A geometric approach to high resolution TVD schemes. *SIAM J. Numer. Anal.*, 25:268–284, 1988.

[17] M. G. Hall. Cell-vertex multigrid schemes for solution of the Euler equations. In K. W. Morton and M. J. Baines, editors, *Numerical Methods for Fluid Dynamics II*, pages 303–345. Oxford University Press, 1985.

[18] A. Harten. On a class of high resolution total-variation-stable finite-difference schemes. *SIAM J. Numer. Anal.*, 21:1–22, 1984.

[19] A. Harten and S. Osher. Uniformly high-order accurate non-oscillatory schemes I. *SIAM J. Numer. Anal.*, 24:279–309, 1987.

[20] A. Jameson, W. Schmidt, and E. Turkel. Numerical solutions of the Euler equations by finite volume methods with Runge–Kutta time stepping schemes. AIAA Paper 81-1259, 1981.

[21] J. S. Jiang, D. Levy, C. T. Lin, S. Osher, and E. Tadmor. High-resolution non-oscillatory central schemes with non-staggered grids for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 35:2147–2168, 1998.

[22] B. P. Leonard. A stable and accurate convective modelling procedure based on quadratic upstream interpolation. *Comput. Methods Appl. Mech. Engrg.*, 19:59–98, 1979.

[23] B. P. Leonard. The ULTIMATE conservative difference scheme applied to unsteady one-dimensional advection. *Comput. Methods Appl. Mech. Engrg.*, 88:17–74, 1991.

[24] P. Lesaint. Numerical solution of the equation of continuity. In J. J. H. Miller, editor, *Topics in Numerical Analysis III*, pages 199–222. Academic Press, 1977.

[25] R. J. LeVeque. A large time step generalization of Godunov's method for systems of conservation laws. *SIAM J. Numer. Anal.*, 22:1051–1073, 1985.

[26] R. J. LeVeque. *Numerical Methods for Conservation Laws.* Lecture Notes in Mathematics, Birkhauser, Basel, 1992.

[27] P. Lin, K. W. Morton, and E. Suli. Euler characteristic Galerkin scheme with recovery. *RAIRO Modél. Math. Anal. Numér.*, 27:863–894, 1993.

[28] P. Lin, K. W. Morton, and E. Suli. Characteristic Galerkin schemes for scalar conservation laws in two and three space dimensions. *SIAM J. Numer. Anal.*, 34:779–796, 1997.

[29] M. Lukáčová, K. W. Morton, and G. Warnecke. Evolution Galerkin methods for hyperbolic systems in two space dimensions. *Math. Comp.*, 2000. To appear.

[30] K. W. Morton. *Numerical Solution of Convection-Diffusion Problems.* Chapman & Hall, London, 1996.

[31] K. W. Morton. Approximation of multidimensional hyperbolic partial differential equations. In I. S. Duff and G. A. Watson, editors, *The State of the Art in Numerical Analysis*, pages 473–502. Oxford University Press, 1997.

[32] K. W. Morton. On the analysis of finite volume methods for evolutionary problems. *SIAM J. Numer. Anal.*, 35:2195–2222, 1998.

[33] K. W. Morton and P. N. Childs. Characteristic Galerkin methods for hyperbolic systems. In J. Ballman and R. Jeltsch, editors, *Second International Conference on Hyperbolic Problems*, pages 435–455. Vieweg, 1989.

[34] K. W. Morton and A. K. Parrott. Generalised Galerkin methods for first-order hyperbolic equations. *J. Comput. Phys.*, 36:249–270, 1980.

[35] K. W. Morton, M. A. Rudgyard, and G. J. Shaw. Upwind iteration methods for the cell vertex scheme in one dimension. *J. Comput. Phys.*, 114:209–226, 1994.

[36] K. W. Morton and I. J. Sobey. Discretization of a convection-diffusion equation. *IMA J. Numer. Anal.*, 13:141–160, 1993.

[37] H. Nessyahu and E. Tadmor. Non-oscillatory central differencing for hyperbolic conservation laws. *J. Comput. Phys.*, 87:408–463, 1990.

[38] R. H. Ni. A multiple grid scheme for solving the Euler equations. *AIAA Journal*, 20(11):1565–1571, 1981.

[39] S. Osher and F. Solomon. Upwind difference schemes for hyperbolic systems of conservation laws. *Math. Comp.*, 38:339–374, 1984.

[40] S. Ostkamp. *Multidimensional characteristic Galerkin schemes and evolution operators for hyperbolic systems.* PhD thesis, Universität Hannover, 1997.

[41] A. Priestley. The positive and nearly conservative Lagrange-Galerkin method. *IMA J. Numer. Anal.*, 14:277–294, 1994.

[42] P. L. Roe. Fluctuations and signals — a framework for numerical evolution problems. In K. W. Morton and M. J. Baines, editors, *Numerical Methods for Fluid Dynamics*, pages 219–257. Academic Press, 1982.

[43] T. Sauer and Y. Xu. On multivariate Lagrange interpolation. *Math. Comp*, 65:1147–1170, 1995.

[44] A. Staniforth and J. Côté. Semi-Lagrangian integration schemes and their application to environmental flows. *Monthly Weather Review*, 119:2206–2223, 1991.

[45] B. van Leer. Towards the ultimate conservative difference scheme. II Monotonicity and conservation combined in a second order scheme. *J. Comput. Phys.*, 14:361–370, 1974.

[46] B. van Leer. Towards the ultimate finite difference scheme. III Upstream-centred finite-difference schemes for ideal compressible flow. *J. Comput. Phys.*, 23:263–275, 1977.

[47] B. van Leer. Towards the ultimate conservative difference scheme. V a second order sequel to Godunov's method. *J. Comp. Phys.*, 32:101–136, 1979.

[48] B. van Leer. On the relations between the upwind-differencing schemes of Godunov, Enquist–Osher and Roe. *SIAM J. Sci. Stat. Comput.*, 5:1–19, 1984.

[49] R. F. Warming, P. Kutler, and H. Lomax. Second- and third-order non-centred difference schemes for nonlinear hyperbolic equations. *AIAA J.*, 11:189–196, 1973.