Implicit flux-corrected transport algorithm for finite element simulation of the compressible Euler equations

D. Kuzmin, M. Möller* and S. Turek

Institute of Applied Mathematics (LS III), University of Dortmund
Vogelpothsweg 87, D-44227, Dortmund, Germany

Abstract

The flux-corrected-transport paradigm is generalized to implicit finite element schemes for hyperbolic systems. A conservative flux decomposition procedure is proposed for both convective and diffusive terms. A mathematical theory for positivity-preserving schemes is reviewed. A nonoscillatory low-order method is constructed by elimination of negative off-diagonal entries of the discrete transport operator. Zalesak's multi-dimensional limiter is employed to switch between linear discretizations of high and low order. A rigorous proof of positivity is provided. A feasible generalization of the scalar methodology for the construction of low-order methods is elucidated. An efficient edge-based algorithm for the matrix assembly for nonlinear systems is devised. Scalar dissipation proportional to the spectral radius of the Roe matrix is used to construct the low-order method for hyperbolic systems. A block-diagonal preconditioner is utilized to work out an efficient defect correction procedure for coupled systems. Several 2D examples for both stationary and highly dynamic flow in a wide range of Mach numbers are presented to demonstrate the potential of the new methodology.

Key Words: compressible Euler equations; hyperbolic conservation laws
implicit flux-corrected-transport; high resolution schemes
scalar dissipation; finite elements; unstructured grids

1 Introduction

Even today, the accurate treatment of convection-dominated transport problems remains a challenging task in numerical simulation of both compressible and incompressible flows. The discrepancy arises between high accuracy and good resolution of singularities on the one hand and preventing the growth and birth of nonphysical oscillations on the other hand. In 1959 it was proven [6], that linear methods are restricted to be at most first order if they are to preserve monotonicity. Thus, the use of nonlinear methods is indispensable to overcome smearing by numerical diffusion without sacrificing important properties of the exact solution such as positivity and monotonicity. The advent of the promising methodology of *flux-corrected transport* (FCT) can be traced back to the pioneering work

*Correspondence to: matthias.moeller@math.uni-dortmund.de

of Boris and Books [3]. Even though their original FCT algorithm named SHASTA was a rather specialized one-dimensional finite difference scheme, the cornerstone for a variety of high-resolution schemes was laid. Strictly speaking, the authors recommended using a high-order discretization in regions of smooth solutions and switching to a low-order method in the vicinity of steep gradients. This idea of adaptive toggling between methods of high and low order was dramatically improved by Zalesak [30] who proposed a multi-dimensional generalization applicable to arbitrary combinations of high- and low-order discretization but still remaining in the realm of finite differences. This barrier was first crossed by Parrott and Christie [21] who settled the idea of flux-correction in the framework of finite elements. Finally, FEM-FCT reached maturity by the considerable contributions of Löhner and his coworkers [16], [17]. Beside the classical formulation of Zalesak's limiter in terms of element contributions, an alternative approach is available limiting the fluxes edge-by-edge [26],[27].

Pursuing this idea, some modern compressible flow solvers completely abandon the conventional finite element data structure in favor of an edge-based version. The foundations were laid by Peraire *et al.* [23] but their conservative decomposition of Galerkin integrals into fluxes assigned to the edges is applicable only to linear finite elements on triangular and tetrahedral meshes.

In [12] we presented a coherent methodology for the decomposition of both convective and diffusive terms into internodal fluxes independently of the underlying discretization. In the same paper we proposed a multi-dimensional FCT algorithm which is applicable to explicit and implicit schemes. The construction of the positivity-preserving low-order scheme based on the elimination of negative off-diagonal entries was addressed. In this paper we will advance the new approach to implicit schemes applied to systems of equations. The property of having diagonalizable Jacobians inheriting the hyperbolicity will prove to be of great use for the design of the low-order method. We will review its construction for systems in general and elucidate in particular how to obtain a positive low-order method by adding artificial diffusion proportional to the spectral radius of the Roe matrix for hyperbolic systems. Following the concept of edge-based data structures, we propose an efficient algorithm for the matrix assembly but without the need of storing them completely. The flux-correction step will be polished by a synchronization between the distinct variables. Furthermore, we will address the treatment of the fully discretized Euler equations within a defect-correction loop which benefits from the use of a decoupling block-diagonal preconditioner. Last but not least, we elaborate on various implementation aspects and present several numerical examples to stress the potential of the implicit formulation generalized to systems of equations.

2 Governing equations

Consider the conservative form of the Euler equations without source terms in the three-dimensional space

$$\frac{\partial U}{\partial t} + \nabla \cdot \mathbf{F} = 0 \tag{1}$$

with the conservative variables and fluxes given by

$$U = \begin{bmatrix} \rho \\ \rho v_1 \\ \rho v_2 \\ \rho v_3 \\ E \end{bmatrix}, F^{x_1} = \begin{bmatrix} \rho v_1 \\ \rho v_1^2 + p \\ \rho v_1 v_2 \\ \rho v_1 v_3 \\ v_1(E+p) \end{bmatrix}, F^{x_2} = \begin{bmatrix} \rho v_2 \\ \rho v_1 v_2 \\ \rho v_2^2 + p \\ \rho v_2 v_3 \\ v_2(E+p) \end{bmatrix}, F^{x_3} = \begin{bmatrix} \rho v_3 \\ \rho v_1 v_3 \\ \rho v_2 v_3 \\ \rho v_3^2 + p \\ v_3(E+p) \end{bmatrix}, \quad (2)$$

where ρ, v_i, p and E represent the density, velocity, pressure and total energy of the medium, respectively. Furthermore, we make the assumption of a polytropic gas, thus the equation of state reads

$$E = \frac{p}{\gamma - 1} + \frac{\rho}{2}|\mathbf{v}|^2, \quad (3)$$

in which $\mathbf{v}$ stands for the velocity vector and γ denotes the ratio of specific heats ($\gamma = 1.4$ for air). Beside the standard divergence form, the Euler equations can be rewritten in an equivalent nonconservative formulation by introducing the Jacobian matrices for the fluxes in each coordinate direction $A^{x_d} = \partial F^{x_d}/\partial U$ which can be found for instance in [8]. As a result, the quasi-linear Euler equations can be reformulated as

$$\frac{\partial U}{\partial t} + \sum_d A^{x_d} \frac{\partial U}{\partial x_d} = 0. \quad (4)$$

Interestingly enough, the flux vectors F^{x_d} for compressible inviscid flow are homogeneous functions of the conservative variables U independently of the spatial dimension, so that the following useful identity holds [18]

$$F^{x_d} = \frac{\partial F^{x_d}}{\partial U} U = A^{x_d} U. \quad (5)$$

In contrast to the one-dimensional Euler equations which are strictly hyperbolic, the situation in multi-dimensions is more complicated. It can be shown [8] that for any linear combination of the Jacobian matrices

$$A(\kappa) = \sum_d A^{x_d} \kappa^{x_d} \quad (6)$$

there exists a regular matrix $R(\kappa)$ such that

$$A(\kappa) = R(\kappa)\Lambda(\kappa)R(\kappa)^{-1} \quad (7)$$

is diagonalizable, where Λ is the diagonal matrix of the *real* eigenvalues which are only distinct in one space dimension. Nevertheless, it *is not* possible to diagonalize each Jacobian simultaneously with the same matrix R. For the one-dimensional Euler equations the diagonalization procedure simply reduces to

$$A = R\Lambda R^{-1}, \text{ where } \Lambda = \mathrm{diag}(v - c, v, v + c) \quad (8)$$

is the diagonal matrix of eigenvalues, and

$$R = \begin{bmatrix} 1 & 1 & 1 \\ v - c & v & v + c \\ h - vc & v^2/2 & h + vc \end{bmatrix} \quad (9)$$

is the matrix of the corresponding right eigenvectors. Here $c = \sqrt{\gamma p/\rho}$ stands for the local speed of sound and h denotes the total enthalpy which is given by $h = (E + p)/\rho$.

3 Galerkin flux decomposition

For deducing the basic theory, let us go back to a generic time-dependent conservation law for a scalar quantity u

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f} = q \qquad \text{in } \Omega, \tag{10}$$

where q is a source term, and $\mathbf{f}$ is a flux function. The weak form of this equation reads

$$\int_\Omega w \left[\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f} - q \right] \, d\mathbf{x} = 0, \qquad \forall w. \tag{11}$$

It is well known from the theory of finite difference methods that a numerical scheme is conservative if it admits the decomposition into a sum of fluxes from one node into another. Indeed, as long as the internodal fluxes are equal in magnitude and opposite in direction, the total mass of the system may only change due to boundary fluxes. Hence, it is highly desirable to represent the numerical method in conservative form whenever possible. At the same time, it has been largely unclear how to accomplish this in the context of finite elements on unstructured meshes.

Many authors have concentrated on this topic promoting edge-based data structures [15],[18],[20] and [23]. Unfortunately, the underlying flux decomposition is feasible only for simplex elements with linear basis functions which have a constant gradient. In [12] we propose an alternative flux decomposition technique which is applicable to general finite element approximations on arbitrary meshes including quadrilateral and hexahedral ones. Hence, let us just point out the main ideas in this paper.

Integration by parts of the weak formulation (11) yields

$$\int_\Omega w \frac{\partial u}{\partial t} \, d\mathbf{x} - \int_\Omega \nabla w \cdot \mathbf{f} \, d\mathbf{x} + \int_{\partial\Omega} w \, \mathbf{f} \cdot \mathbf{n} \, ds - \int_\Omega wq \, d\mathbf{x} = 0, \qquad \forall w. \tag{12}$$

A common practice in finite element computations of compressible flow is to approximate the fluxes in the same way as the desired solution which is referred to as the *group finite element formulation* [5].

Let the solution, the fluxes and the source terms be represented in the form

$$u = \sum_j u_j \varphi_j, \qquad \mathbf{f} = \sum_j \mathbf{f}_j \, \varphi_j, \qquad q = \sum_j q_j \varphi_j. \tag{13}$$

In fact, it is not compelling to use the same approximations for u and $\mathbf{f}$. For instance, one can consider using nonconforming Crouzeix-Raviart or Rannacher-Turek elements for the fluxes.

After the substitution of expressions (13) and the weighting functions $w = \varphi_i$ into the variational formulation (12), we obtain

$$\sum_j \left[\int_\Omega \varphi_i \varphi_j \, d\mathbf{x} \right] (\dot{u}_j - q_j) - \sum_j \left[\int_\Omega \nabla \varphi_i \varphi_j \, d\mathbf{x} - \int_{\partial\Omega} \varphi_i \varphi_j \, \mathbf{n} \, ds \right] \cdot \mathbf{f}_j = 0, \tag{14}$$

which can be written in compact matrix form as

$$M_C(\dot{u} - q) = \sum_{d=1}^{3} K^{x_d} f^{x_d} \tag{15}$$

in the three-dimensional case. The utility of the group formulation is illustrated by the fact that the three matrices K^{x_d} engendered by the corresponding first-order derivatives can be assembled once and for all at the beginning of the simulation, as long as the mesh does not change. This is in contrast to the standard finite element approach, whereby the discrete operators for the linearized convective terms have to be updated in each time step.

By construction, the discretized flux term consists of an interior part and a boundary part. The former is given by the integral

$$\sum_j \left[\int_\Omega \nabla\varphi_i\,\varphi_j\,d\mathbf{x}\right]\cdot\mathbf{f}_j = \sum_j \mathbf{c}_{ij}\cdot\mathbf{f}_j, \qquad \mathbf{c}_{ij} = \int_\Omega \nabla\varphi_i\,\varphi_j\,d\mathbf{x}, \tag{16}$$

where the coefficient matrices $c_{ij}^{x_d}$ possess the zero column sum property, since it is assumed that the sum of basis functions equals unity. Therefore, it is possible to express the diagonal coefficients in terms of off-diagonal ones

$$\sum_i \mathbf{c}_{ij} = 0 \quad \Rightarrow \quad \mathbf{c}_{ii} = -\sum_{j\neq i}\mathbf{c}_{ji}. \tag{17}$$

It follows that the interior flux term (16) can be rewritten as

$$\sum_j \mathbf{c}_{ij}\cdot\mathbf{f}_j = \sum_{j\neq i} g_{ij}, \qquad \text{where} \quad g_{ij} := \mathbf{c}_{ij}\cdot\mathbf{f}_j - \mathbf{c}_{ji}\cdot\mathbf{f}_i. \tag{18}$$

The newly introduced quantity g_{ij} represents the *Galerkin flux* from node j into node i. It is obvious that $g_{ji} = -g_{ij}$, so that node j receives the same contribution with the opposite sign. Importantly, flux decomposition is also possible for *generalized diffusion operators* [11] which are defined as symmetric matrices having zero row and column sums. The purely diffusive Galerkin flux assumes a remarkably simple form

$$\sum_i d_{ij} = \sum_j d_{ij} = 0, \quad d_{ij} = d_{ji} \qquad \Rightarrow \qquad g_{ij} = d_{ij}\cdot(u_j - u_i). \tag{19}$$

Note that generalized diffusion operators are not required to have continuous counterparts. Some typical examples are the discrete Laplacian or the matrix $M_C - M_L$ sometimes referred to as 'mass diffusion'. As we will see shortly, the properties of discrete diffusion operators render them a valuable tool for the design of nonoscillatory low-order methods to be combined with high-order ones within the flux-corrected-transport algorithm.

Another promising approach to the derivation of high-resolution finite element schemes involves the replacement of the original Galerkin flux by another *consistent numerical flux*. Its potential is demonstrated by numerous publications [18], [19], [20] in which one-dimensional limiters are successfully applied on unstructured meshes in conjunction with the edge-based data structure of Peraire *et al.* [23].

As a matter of fact, it is not necessary to implement the Galerkin flux decomposition en bloc. On the contrary, it is sufficient to restrict this procedure to the diffusive terms. Thus, our methodology can be integrated easily into existing element-based codes. However, this decomposition constitutes an excellent tool for the derivation of new flux-manipulating schemes.

4 Discrete positivity criteria

Let us recall that most real-world applications are described by inherently nonlinear systems of partial differential equations modeling physical phenomena. It is therefore quite natural to impose some constraints on the numerical method which are dictated by the continuous model such as positivity-preservation of physical quantities. We will see that the concept of an M-matrix provides a handy criterion to guarantee positivity. The second fundamental issue concerning numerical methods is related to convergence. Convergence proofs for nonlinear systems take their pivotal ideas from the nonlinear stability analysis, the functional analytic framework of the theory of compactness. Combining both issues, leads to the class of *monotone* methods, as a subset of the positive schemes, which guarantee that a converged solution of any conservation law does satisfy the entropy inequality ensuring that the numerical solution converges to the physically relevant weak solution.

4.1 Positivity

DEFINITION. *A nonsingular discrete operator $A \in \mathbb{R}^{n\times n}$ is called an M-matrix if $a_{ij} \leq 0$ for $i \neq j$ and all the entries of A^{-1} are nonnegative.*

If A is strictly diagonally dominant and $a_{ii} > 0$, while $a_{ij} \leq 0$ for $i \neq j$, then A is an M-matrix. Note that for M-matrices $Ax \geq 0$ implies that $x \geq 0$. This property leads to the following fundamental lemma:

LEMMA. *Let the numerical scheme be presented in abstract matrix operator form as*

$$Lu^{n+1} = Ru^n. \tag{20}$$

Then it follows that positivity is preserved provided that L is an M-matrix and all entries of R are nonnegative ($R \geq 0$).

The proof of this lemma is almost trivial and can be found in [11].

To introduce another useful concept, consider a semi-discrete problem of the form

$$\frac{du_i}{dt} = \sum_j c_{ij}u_j, \qquad \sum_j c_{ij} = 0, \tag{21}$$

where u_i are the nodal values, and c_{ij} are some coefficients depending on the procedure employed for the spatial discretization. Actually, for incompressible flows the lumped-mass Galerkin discretization of the transport equation admits such a representation.

Since the coefficient matrix has zero row sum, the scheme can be rewritten as

$$\frac{du_i}{dt} = \sum_{j\neq i} c_{ij}(u_j - u_i). \tag{22}$$

Furthermore, by supposing all coefficients to be nonnegative ($c_{ij} \geq 0,\ j \neq i$) we can show stability in the L_∞-norm for this scheme.

Consider u_i to be a maximum. It follows that for all j $u_j - u_i \leq 0$, so that $du_i/dt \leq 0$. Hence, a maximum cannot increase, and similarly a minimum cannot decrease. Recall

that coefficient matrices are sparse in such a way that $c_{ij} \neq 0$ only if i and j are adjacent nodes. Restricting ourselves to a local neighborhood of u_i the same consideration as above reveals that the method is *local extremum diminishing*. The LED criterion was introduced by Jameson [9], [10] as a convenient tool for the design of high-resolution schemes on unstructured meshes. It implies positivity, since if the initial solution is positive everywhere, then so is the global minimum which cannot decrease by definition. Hence, the LED property provides an effective mechanism for preventing the birth and growth of nonphysical oscillations.

4.2 Monotonicity

In one dimension the LED property ensures that the *total variation* of the solution which is defined as

$$TV(u) = \int_{-\infty}^{+\infty} \left| \frac{\partial u}{\partial x} \right| dx \tag{23}$$

does not increase. As a matter of fact, equation (23) remains valid even for discontinuous $u(x)$ in the sense of distribution theory. For the sake of simplicity, consider homogeneous Dirichlet boundary conditions at both endpoints. Then the discrete total variation is given by

$$TV(u) = 2 \left(\sum \max u - \sum \min u \right). \tag{24}$$

Thus, a one-dimensional LED scheme is necessarily *total variation diminishing*. This is a highly advantageous property, which has led to the development of a whole class of nonoscillatory TVD schemes and which is sufficient for *total variation stability*. Thus, existing convergence theorems for consistent TV-stable methods can easily be generalized [7], [13]. Let us remark that the TVD feature can be consulted for the construction of high-order schemes in contrast to just 'mimicking' other properties of the true solution which renders the accuracy to be at most first order [13].

Recall that equations (21) and (22) correspond to the problem discretized in space only. Let us now investigate the conditions under which a LED scheme will remain positive after the time discretization. If the standard one-step θ-scheme is employed, the fully discretized equation reads

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \theta \sum_{j \neq i} c_{ij} (u_j^{n+1} - u_i^{n+1}) + (1 - \theta) \sum_{j \neq i} c_{ij} (u_j^n - u_i^n), \quad 0 \leq \theta \leq 1. \tag{25}$$

The choice of the parameter θ specifies the type of time-stepping. The extreme cases $\theta = 0$ and $\theta = 1$ define the well-known forward and backward Euler methods. Both of them are first-order accurate with respect to the time step Δt. The method corresponding to $\theta = 0.5$ is known as the Crank-Nicolson scheme, which is second-order accurate.

The application of our lemma to equation (25) yields the following [11]

Positivity Theorem *Any LED scheme discretized in time by the backward Euler method is unconditionally positive. Other time-stepping schemes* $(0 \leq \theta < 1)$ *preserve positivity under the CFL-like condition*

$$1 + \Delta t (1 - \theta) \min_i c_{ii} \geq 0. \tag{26}$$

An important message delivered by this theorem is that the positivity criterion at our disposal makes it possible to obtain rigorous estimates of the largest admissible time step for explicit and semi-implicit schemes. Remarkably, the derivation of the upper bound does not require any knowledge of the underlying partial differential equation and of the employed spatial mesh. It is sufficient to examine the diagonal coefficients c_{ii} of the semi-discrete scheme. Upper bounds for non-LED schemes can be readily derived in much the same way.

5 Low-order discretization

To a large extent, the performance of the flux-corrected-transport procedure depends on the quality of the underlying low-order method which is supposed to preserve positivity and refrain from forming numerical wiggles. In the context of finite difference and finite volume discretizations, a perfect candidate for this job certainly is the upwind scheme. At the same time, it has been largely unclear how to perform upwinding in the finite element framework. Most upwind-like finite element methods encountered in the literature resort to a finite volume discretization for the convective terms [2],[29]. An alternative derivation of the least diffusive positivity-preserving scheme can be carried out by adding discrete diffusion depending solely on the magnitude and position of negative entries in the finite element matrix [11].

5.1 Discrete diffusion

Let the scalar conservation law (10), (11) be discretized in space by the Galerkin method

$$M_C \frac{du}{dt} = K^H u + M_C q. \tag{27}$$

On the way of rendering this semi-discrete scheme positive, any implicit antidiffusion disguised in the consistent mass matrix must be removed by performing mass lumping. In the next step, the discrete transport operator K^H has to be modified by applying a proper amount of artificial diffusion. To this end, we define a discrete dissipation tensor D as a symmetric matrix with zero row and column sum which is designed so as to eliminate all negative off-diagonal entries of the high-order operator

$$d_{ii} = -\sum_{k \neq i} d_{ik}, \qquad d_{ij} = d_{ji} = \max\{0, -k_{ij}^H, -k_{ji}^H\}, \quad \forall i < j. \tag{28}$$

In essence, this corresponds to applying one-dimensional diffusion operators associated with the (fictitious) edges connecting the adjacent nodes. As a result, the diffusive term can be decomposed into a sum of internodal fluxes, such that each flux is proportional to the difference of corresponding nodal values

$$(Du)_i = \sum_j d_{ij} u_j = \sum_{j \neq i} d_{ij}(u_j - u_i) = \sum_{j \neq i} f_{ij}. \tag{29}$$

Obviously, the two participating nodes receive the same flux but with opposite sign. Thus, the artificial diffusion operator ensures the conservation of mass. By adding the tensorial

dissipation to the high-order transport operator its low-order counterpart $K^L = K^H + D$ is obtained.

It is worth mentioning, that if physical diffusion is strong enough the coefficients are nonnegative from the outset, and thus no artificial diffusion is introduced. Hence, for diffusion-dominated problems K^H and K^L coincide. Note that this approach can be carried over to the Euler equations by defining the modulation parameter d_{ij} to be proportional to the spectral radius of the Roe matrix (see below).

After the discretization in time the modified scheme reads

$$[M_L - \theta\Delta t K^L]u^{n+1} = [M_L + (1-\theta)\Delta t K^L]u^n + \Delta t M_C q^{n+\theta}. \tag{30}$$

Note that the presence of (negative) source terms, which may originate from both chemical reactions and the transition to spherical or cylindrical coordinates, threatens to violate the positivity constraint. In principle, this can be overcome by following the linearizing splitting approach proposed by Patankar [22]

$$q = q_C + q_P u, \qquad \text{where} \quad q_C \geq 0,\ q_P \leq 0. \tag{31}$$

Equation (30) can be transformed into the convenient form (20) with

$$\begin{aligned} L &= M_L - \theta\Delta t K^L + \Delta t S^- \\ R &= M_L + (1-\theta)\Delta t K^L + \Delta t S^+, \end{aligned}$$

where S^+ and S^- are diagonal matrices engendered by the source term. A splitting which ensures that L and R possess the properties required by the lemma can readily be derived [12].

By construction all off-diagonal entries of L are nonpositive, while R is nonnegative. It remains to indemnify the positivity of the diagonal coefficients. Since the elements of M_L and the contributions of the source terms q_C (if any) are positive this can always be achieved by choosing the time step small enough. Dropping the source terms for simplicity, the time step for the low-order discretization of the incompressible convection-diffusion equation is bounded by

$$\Delta t \leq \frac{1}{1-\theta} \min_i \{-m_i/k_{ii}^L \mid k_{ii}^L < 0\}, \tag{32}$$

where m_i denotes the diagonal entries of the lumped mass matrix. From this CFL-like condition we get a sharp estimate of the largest admissible time step that can be used to steer adaptive time-stepping for (semi-)explicit schemes. In essence, the upper bound is determined by the ratio m_i/k_{ii}^L. It follows, that excessive artificial diffusion not only degrades the accuracy of the method but also requires taking impractically small time steps.

5.2 1D example: Pure convection

Let us illustrate the construction of the low-order operator by considering the one-dimensional scalar equation of pure convection with constant positive velocity

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = 0. \tag{33}$$

Employing a discretization with linear elements on a uniform mesh yields the following element matrices

$$\hat{M}_L = \frac{\Delta x}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \hat{K}^H = \frac{v}{2}\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}. \tag{34}$$

After the global assembly, the central difference approximation of the convective terms is recovered at the interior nodes

$$\frac{du_i}{dt} = -v\ \frac{u_{i+1} - u_{i-1}}{2\,\Delta x}. \tag{35}$$

In order to eliminate the negative entries of K^H, the artificial dissipation has to be taken proportional to $\hat{d}_{12} = v/2$ to produce the least diffusive results but to ensure positivity. Consequently, the corresponding diffusion operator restricted to one element is given by

$$\hat{D} = \frac{v}{2}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \quad \Rightarrow \quad \hat{K}^L = v\begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}. \tag{36}$$

Putting it all together, the low-order scheme coincides with the upwind finite difference method

$$\frac{du_i}{dt} = -v\ \frac{u_i - u_{i-1}}{\Delta x} \tag{37}$$

which preserves positivity provided that the time step satisfies this CFL-like condition

$$v\frac{\Delta t}{\Delta x} \leq \frac{1}{1-\theta}. \tag{38}$$

In a nutshell, this example demonstrates that our low-order discretization reduces to standard upwinding for pure convection in one spatial dimension. At the same time, its derivation based on the postprocessing of the discrete transport operator remains valid for arbitrary meshes and multi-dimensional problems. Moreover, physical diffusion (if any) is automatically detected, and the amount of artificial diffusion is reduced accordingly.

6 Flux-based FEM-FCT formulation

Another cornerstone of the FEM-FCT algorithm is the linear high-order method. A variety of finite element schemes employing streamline diffusion to stabilize the troublesome convective terms were proposed in the literature, e.g. [4]. For instance, Taylor-Galerkin methods attribute this stabilization to high-order time derivatives in the Taylor series expansion. This leads to improved time-stepping schemes which are combined with the standard Galerkin spatial discretization. The most popular representative of such stabilized methods is the well-known Lax-Wendroff scheme. An investigation of the modified equation for its finite element counterpart reveals that the introduced dissipation just counterbalances the intrinsic negative diffusion which renders the explicit Euler/Galerkin scheme unstable for pure convection problems. For an in-depth study of the Lax-Wendroff and higher order Taylor-Galerkin methods the reader is referred to [4].

In this paper we will concentrate on implicit finite element schemes based on the Crank-Nicolson and backward Euler time-stepping which are unconditionally stable. We

just mention that a stabilization of convective terms is mandatory for the fully explicit time discretization. Linear Galerkin schemes are of little use, because they are prone to nonphysical oscillations. The incorporation of a flux limiter makes it possible to get rid of oscillations in the framework of nonlinear Crank-Nicolson/FCT and backward Euler/FCT methods.

The high-order transport operator can be transformed into a low-order one as explained in the previous section. For simplicity, let us omit the (linearized) source terms. The resulting methods of high and low order discretized in time by the standard θ-scheme are related by the following formula

$$[M_L - \theta\Delta t K^L]u^H = [M_L + (1-\theta)\Delta t K^L]u^n + F(u^H, u^n), \tag{39}$$

where the diffusion responsible for high spatial accuracy is given by

$$\begin{aligned} F(u^H, u^n) &= -\left[(M_C - M_L) + \theta\Delta t\left(K^L - K^H\right)\right]u^H \\ &+ \left[(M_C - M_L) - (1-\theta)\Delta t\left(K^L - K^H\right)\right]u^n. \end{aligned} \tag{40}$$

Here the superscript H refers to the high-order solution. If the antidiffusive term $F(u^H, u^n)$ is omitted, then the positive low-order scheme will be recovered, whereas retaining it yields the original high-order method.

It can readily be seen that all the matrices in (40) represent discrete (anti-)diffusion operators featuring zero row and column sums. Hence, they admit the decomposition into a sum of internodal fluxes

$$\begin{aligned} f_{ij} &= -\left(m_{ij} + \theta\Delta t d_{ij}\right)\left(u_j^H - u_i^H\right) \\ &+ \left(m_{ij} - (1-\theta)\Delta t d_{ij}\right)\left(u_j^n - u_i^n\right) \end{aligned} \qquad f_{ji} = -f_{ij}, \quad i \neq j \tag{41}$$

These raw antidiffusive fluxes counterbalance the errors induced by mass lumping and 'upwinding'. The coefficients m_{ij} and d_{ij} denote the entries of the consistent mass matrix and the artificial diffusion, respectively.

The crucial step of the FCT procedure consists in adding as much antidiffusion as possible without generating nonphysical undershoots and overshoots. Incorporating the limiter, the implicit FEM-FCT algorithm for equation (39) reads

$$[M_L - \theta\Delta t K^L]u^{n+1} = M_L\tilde{u} + F^*(u^H, u^n), \qquad F^*(u^H, u^n) = \sum_{j\neq i}\alpha_{ij}f_{ij} \tag{42}$$

where $F^*(u^H, u^n)$ denotes the limited antidiffusion (α_{ij} will be defined later), while $\tilde{u}$ represents the positivity-preserving solution to the explicit subproblem

$$M_L\tilde{u} = [M_L + (1-\theta)\Delta t K^L]u^n. \tag{43}$$

In essence, $\tilde{u}$ corresponds to an intermediate solution computed at the time instant $t^{n+1-\theta}$ by the explicit low-order scheme which reduces to the old solution u^n for the backward Euler method.

It is obvious that the success of the FCT algorithm depends on the positivity of the provisional solution $\tilde{u}$ and on the choice of the correction factors α_{ij}. For $\tilde{u}$ to be positive,

the time step must satisfy the CFL-like condition (32) unless the scheme is fully implicit. As long as the left-hand side operator is an M-matrix, our positivity criterion ensures that scheme (41) can be rendered positive by tuning the correction factors.

This new family of FEM-FCT schemes distinguishes itself in that it is applicable to explicit and implicit time discretizations alike. Note that implicit schemes require solving *two* nonsymmetric linear systems per time step: one for the high-order solution (which is needed to compute the antidiffusive fluxes) and one for the final solution. However, implicit methods are typically more efficient than explicit ones for they allow to use larger time steps.

The majority of practical applications are described by *nonlinear* conservation laws. In this case, the matrices K^H and K^L depend on the unknown solution, so that additional outer iterations are necessary for implicit schemes. The simplest iterative treatment of nonlinearities is afforded by a fixed point defect correction method. If we consider an abstract nonlinear system of the form

$$S(u)u = g, \tag{44}$$

then the basic nonlinear iteration can be formulated as

$$u^{(m+1)} = u^{(m)} + [C(u^{(m)})]^{-1}(g - S(u^{(m)})u^{(m)}), \tag{45}$$

where m is the outer iteration counter, and C is a suitably chosen 'preconditioner' (an approximate Fréchet derivative) which should be easy to invert. The iteration process is terminated when the relative solution changes are small enough or l exceeds a given limit. As a rule, the 'inversion' of C is also performed by some iterative procedure. Hence, a certain number of inner iterations per cycle is required. It is worth mentioning that the problem does not have to be solved very accurately at each outer iteration. A moderate improvement of the residual (1-2 digits) is sufficient to obtain a good overall accuracy.

For a nonlinear problem of form (39), it is reasonable to use the low-order operator as preconditioner

$$C(u^{(m)}) = M_L - \theta\Delta t K^L(u^{(m)}). \tag{46}$$

This yields an iterative FEM-FCT algorithm, whereby the approximate solution and the transport operator are successively updated as follows:

$$[M_L - \theta\Delta t K^L(u^{(m)})]u^{(m+1)} = [M_L + (1-\theta)\Delta t K^L(u^n)]u^n + F^*(u^{(m)}, u^n). \tag{47}$$

Flux correction can be performed after each outer iteration or just once after the high-order solution has converged. In either case, positivity of the numerical solution is guaranteed.

7 Limiting strategy

The flux limiter is a key element of the FEM-FCT paradigm. By varying the correction factors α_{ij} between zero and unity, it it possible to obtain the diffusive low-order solution, the oscillatory high-order one or anything in-between. Obviously, it is desirable to utilize the antidiffusive terms to the largest extent possible without generating wiggles and violating the positivity constraint. Kuzmin and Turek [11] demonstrated that Zalesak's multi-dimensional limiter can be generalized to implicit time discretizations.

7.1 Zalesak's limiter

The multi-dimensional version of the original limiter [3] was proposed by Zalesak [30] remaining in the realm of finite differences. Parrott and Christie [21] were the first who utilized flux-correction in the context of finite elements. FEM-FCT reached maturity by the efforts done by Löhner [16]. A quite comprehensive description of Zalesak' s limiter containing a geometrical interpretation can be found in [12]. However, the geometrical interpretation fails if we consider implicit time discretizations. Consequently, we have to apply the mathematical theory of positivity-preserving schemes to prove the positivity of Zaselak's limiter for arbitrary time-stepping. Let us just give a brief outline of the theoretical aspects.

Considering a local maximum or minimum $\tilde{u}_i^{\overset{\max}{\min}}$, any incoming flux f_{ij} which would accentuate the extremum has to be canceled out completely. Afterwards, we can represent the right-hand side of our FEM-FCT scheme in the following form

$$RHS = m_i\tilde{u}_i + \sum_{j\neq i}\alpha_{ij}f_{ij} = m_i\tilde{u}_i + c_iQ_i, \qquad c_i = \frac{\sum_{j\neq i}\alpha_{ij}f_{ij}}{Q_i}, \tag{48}$$

where the intermediate solution $\tilde{u} = u^L(t^{n+1-\theta})$ is subject to the concrete time-stepping employed and the multiplier Q_i is chosen as follows

$$Q_i = \begin{cases} Q_i^+ = \tilde{u}_i^{\max} - \tilde{u}_i, & \text{if } \sum\limits_{j\neq i}\alpha_{ij}f_{ij} > 0, \\ Q_i^- = \tilde{u}_i^{\min} - \tilde{u}_i, & \text{if } \sum\limits_{j\neq i}\alpha_{ij}f_{ij} < 0, \\ 1, & \text{if } \sum\limits_{j\neq i}\alpha_{ij}f_{ij} = 0. \end{cases} \tag{49}$$

Keep in mind that we took the necessary precautions that all $Q_i \neq 0$ by canceling out the local extremum accentuating fluxes completely.

Let us consider the local extremum $\tilde{u}_i^{\overset{\max}{\min}}$ to be attained at some node k adjacent to node i. The LED property of the antidiffusive term makes it possible to rewrite the equation at hand as

$$RHS = m_i\tilde{u}_i + c_i(\tilde{u}_k - \tilde{u}_i) = (m_i - c_i)\tilde{u}_i + c_i\tilde{u}_k, \qquad c_i \geq 0 \tag{50}$$

which states that positivity is preserved as long as $m_i \geq c_i$. Obviously, Zalesak's limiter picks out the proper correction factors α_{ij} for this job as can be seen from

$$m_iQ_i^- \leq m_iR_i^-P_i^- \leq \sum_{j\neq i}\alpha_{ij}f_{ij} \leq m_iR_i^+P_i^+ \leq m_iQ_i^+, \tag{51}$$

where the sum of contributions to node i is denoted by

$$P_i^{\pm} = \frac{1}{m_i}\sum_{j\neq i}\overset{\max}{\min}\{0, f_{ij}\}. \tag{52}$$

Finally, the auxiliary quantities $R_i^{\pm}$ which determine the maximum percentage of allowable flux into node i are defined as

$$R_i^{\pm} = \begin{cases} \min\{1, Q_i^{\pm}/P_i^{\pm}\}, & \text{if } P_i^{\pm} \neq 0, \\ 0, & \text{if } P_i^{\pm} = 0. \end{cases} \tag{53}$$

As a result, the proper correction factors have to be chosen according to

$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\}, & \text{if } f_{ij} \geq 0, \\ \min\{R_j^+, R_i^-\}, & \text{if } f_{ij} < 0. \end{cases} \tag{54}$$

7.2 Limiting for systems of equations

Despite the remarkable progress in the development of FCT schemes for scalar equations, it remained largely unclear how to apply flux correction to systems of hyperbolic conservation laws. Neglecting the special character of coupled systems, a conventional operator-splitting approach is feasible by treating each scalar subproblem in a block-iterative loop. However, limiting each variable independently turned out to be of no good. Consequently, the FCT community has devised a common limiter for the entire system by merging individual limiters for different variables.

Flux correction for the system of Euler equations was addressed by Löhner [15],[16]. He singled out the following approaches to the design of a synchronized limiter:

- Apply correction factors corresponding to a single 'indicator variable'.
- Apply the minimum of correction factors for a group of variables.

Some clues for a suitable choice of variables will be given in the subsequent section.

8 Application to the Euler equations

The new methodology for scalar transport equations can be generalized to hyperbolic systems of conservation laws. Nevertheless, a rigorous positivity proof is still outstanding and meanwhile some heuristic considerations have to be employed. In what follows, we build on the concept of a Roe matrix to devise an efficient algorithm for the edge-by-edge matrix assembly and discrete upwinding. Another option is to apply the Galerkin flux decomposition and modify the numerical flux as explained above. The resulting large nonlinear systems call for the use of a smart iteration technique to minimize the computational costs. As in the scalar case, we resort to defect correction with a block-diagonal low-order preconditioner.

8.1 Matrix assembly

Let us start with the conservative formulation of the Euler equations. If we apply the Galerkin discretization to the weak formulation of equation (1) without integrating by parts and recall that customary basis functions sum to unity, the semi-discrete problem can be written in the following form

$$\begin{aligned} \left[M\frac{dU}{dt}\right]_i &= \sum_{j\neq i} \mathbf{k}_{ij}\cdot(\mathbf{F}_j-\mathbf{F}_i) = \left[\frac{\mathbf{k}_{ij}-\mathbf{k}_{ji}}{2}+\frac{\mathbf{k}_{ij}+\mathbf{k}_{ji}}{2}\right]\cdot\hat{\mathbf{A}}_{ij}(U_j-U_i) \\ \left[M\frac{dU}{dt}\right]_j &= \sum_{j\neq i} \mathbf{k}_{ji}\cdot(\mathbf{F}_i-\mathbf{F}_j) = \left[\frac{\mathbf{k}_{ji}-\mathbf{k}_{ij}}{2}+\frac{\mathbf{k}_{ij}+\mathbf{k}_{ji}}{2}\right]\cdot\hat{\mathbf{A}}_{ij}(U_i-U_j) \end{aligned} \tag{55}$$

for node i and j, respectively. Here $\mathbf{k}_{ij}$ denotes the vector of coefficients coming from the Galerkin discretization, i.e. $-\mathbf{c}_{ji}$ in the notation of section 3. In particular, $k_{ij} = \pm 1/2$ in one dimension. Integration by parts yileds the following relation (see [15], p. 195):

$$\mathbf{k}_{ij} + \mathbf{k}_{ji} = -\int_\Gamma \mathbf{n}\,\varphi_i\varphi_j\,ds, \tag{56}$$

where $\mathbf{n}$ is the outward unit normal. In equation (55), $\mathbf{k}_{ij}$ was decomposed into

$$\begin{aligned} \mathbf{k}_{ij}^I &:= \frac{\mathbf{k}_{ij} - \mathbf{k}_{ji}}{2} = \mathbf{k}_{ij} + \frac{1}{2}\int_\Gamma \mathbf{n}\,\varphi_i\varphi_j\,ds, \\ \mathbf{k}_{ij}^B &:= \frac{\mathbf{k}_{ij} + \mathbf{k}_{ji}}{2} = -\frac{1}{2}\int_\Gamma \mathbf{n}\,\varphi_i\varphi_j\,ds. \end{aligned} \tag{57}$$

Basis functions corresponding to internal nodes vanish on the boundary so that $\mathbf{k}_{ij}^I = \mathbf{k}_{ij}$ while $\mathbf{k}_{ij}^B = 0$ in this case. For linear and multilinear finite elements we have $\mathbf{k}_{ij}^B = -\mathbf{n}\,s_{ij}/2$ which is nonzero iff both nodes are on the boundary. Here $s_{ij} = \int_\Gamma \varphi_i\varphi_j\,ds$ denotes an entry of the mass matrix for the surface triangulation. For instance, $s_{ij} = |\Gamma_{ij}|/6$ in 2D. Shortly we will see why the splitting of coefficients into a sum of internal and boundary contributions is of use. In essence, it corresponds to taking the average of the discrete operators resulting from the Galerkin discretization without and with integration by parts (c.f. the weak form (14)).

The vector $\mathbf{F}$ in equation (55) denotes the nodal values of fluxes. Furthermore, the matrix $\hat{\mathbf{A}}_{ij}$ is the so-called Roe matrix which is obtained by evaluating the Jacobians at the intermediate state [24]

$$\hat{\rho}_{ij} = \sqrt{\rho_i\rho_j}, \qquad \hat{\mathbf{v}}_{ij} = \frac{\sqrt{\rho_i}\mathbf{v}_i + \sqrt{\rho_j}\mathbf{v}_j}{\sqrt{\rho_i} + \sqrt{\rho_j}}, \qquad \hat{h}_{ij} = \frac{\sqrt{\rho_i}h_i + \sqrt{\rho_j}h_j}{\sqrt{\rho_i} + \sqrt{\rho_j}}. \tag{58}$$

The resulting density-averaged quantities $\hat{\rho}_{ij}$, $\hat{\mathbf{v}}_{ij}$ and $\hat{h}_{ij}$ are called the Roe mean values.

While in the one-dimensional case there is only one Jacobian matrix to be assembled for each pair of nodes, in multi-dimensions it is replaced by a suitable linear combination of the averaged Jacobians $\hat{A}_{ij}^{x_d}$ for each coordinate direction $d = 1, 2, 3$. To this end, we introduce the so-called cumulative Roe matrix

$$\mathrm{A}_{ij} := \mathbf{k}_{ij} \cdot \hat{\mathbf{A}}_{ij} = k_{ij}^{x_1}\hat{A}_{ij}^{x_1} + k_{ij}^{x_2}\hat{A}_{ij}^{x_2} + k_{ij}^{x_3}\hat{A}_{ij}^{x_3}$$

which consists of the interior and boundary contributions:

$$\mathrm{A}_{ij} = \mathrm{A}_{ij}^I + \mathrm{A}_{ij}^B, \qquad \text{where} \quad \mathrm{A}_{ij}^I := \mathbf{k}_{ij}^I \cdot \hat{\mathbf{A}}_{ij}, \quad \mathrm{A}_{ij}^B := \mathbf{k}_{ij}^B \cdot \hat{\mathbf{A}}_{ij}. \tag{59}$$

Let us remark that A_{ij} can be interpreted as a projection of the multidimensional Jacobian tensor onto the corresponding 'numerical edge' connecting the nodes i and j. Each edge corresponds to a pair of basis functions with overlapping supports. Hence, the list of edges is determined by the sparsity pattern of the global finite element matrix.

Similarly, the Roe matrix A_{ji} for the coefficient vector $\mathbf{k}_{ji}$ is given by

$$\mathrm{A}_{ji} := \mathbf{k}_{ji} \cdot \hat{\mathbf{A}}_{ij} = -\mathrm{A}_{ij}^I + \mathrm{A}_{ij}^B.$$

For internal nodes $\mathrm{A}_{ji} = -\mathrm{A}_{ij}$ since the boundary contribution A_{ij}^B vanishes. Hence, just one cumulative Jacobian per edge needs to be evaluated in this case. Otherwise, the local Roe matrices associated with the numerical edge will consist of a skew-symmetric internal part and a symmetric boundary part.

The transition from the fluxes to the nodal values of the conservative variables in (55) enables us to assemble finite element matrices explicitly. The connectivity of the global matrix depends on the underlying mesh and on the type of finite element approximation. For systems of equations, the array of edges remains the same as in the scalar case. However, there are interactions not only between basis functions for different nodes but also between basis functions for different variables. Hence, each coefficient of the discrete operator turns into a matrix of size equal to the squared number of variables.

For the three-dimensional Euler equations, the contribution of a numerical edge to the global Jacobian matrix K is represented by the following 5×5 blocks:

$$\begin{aligned} \mathrm{K}_{ii} &= -\mathrm{A}_{ij}, & \mathrm{K}_{ij} &= \mathrm{A}_{ij}, \\ \mathrm{K}_{ji} &= \mathrm{A}_{ji}, & \mathrm{K}_{jj} &= -\mathrm{A}_{ji}. \end{aligned} \tag{60}$$

These local Jacobians are evaluated edge-by-edge, and their entries K_{ij}^{kl} $(k, l = 1, \ldots, 5)$ are scattered to the corresponding positions in the blocks $K_{kl} \in \mathbb{R}^{N \times N}$, where N stands for the total number of nodes. The fully discretized high-order scheme can be written in the compact form $S^H U^H = G^H$. The constituent blocks of the overall stiffness matrix S^H and of the load vector G^H are defined as

$$S_{kl}^H = M_C\, \delta_{kl} - \theta \Delta t K_{kl}, \qquad g_k^H = M_C\, u_k^n + (1-\theta)\Delta t \sum_l K_{kl} u_l^n. \tag{61}$$

Due to the nonlinearity of the Euler equations, the discrete system has to be solved iteratively. Hence, it is not necessary to assemble and store the global stiffness matrix. The contributions of edges can be inserted directly into the defect vector. If a block-diagonal preconditioner is used (see below), the storage requirement will reduce from 25 to 5 blocks. A distinct advantage of the edge-by-edge assembly algorithm proposed in this section is that the projection coefficients $\mathbf{k}_{ij}$ are typically fixed, so that the matrices can be assembled efficiently without resorting to numerical integration.

8.2 Discrete diffusion

A usable low-order method can be constructed by adding tensorial artificial diffusion to the local Roe matrices. In the framework of scalar theory for positivity preserving schemes, the artificial dissipation was tailored so as to eliminate all negative off-diagonal entries. For systems of equations, the off-diagonal entries are no longer scalar quantities but matrices themselves. Adapting the procedure which underlies the scalar theory, we have to design the dissipation tensor so as to render all off-diagonal matrix blocks positive definite. To this end, we perform mass lumping and apply artificial diffusion to the edge contributions before inserting them into the global matrix and/or the defect vector:

$$\begin{aligned} \mathrm{K}_{ii} &= \mathrm{K}_{ii} - \mathrm{D}_{ij}, & \mathrm{K}_{ij} &= \mathrm{K}_{ij} + \mathrm{D}_{ij}, \\ \mathrm{K}_{ji} &= \mathrm{K}_{ji} + \mathrm{D}_{ij}, & \mathrm{K}_{jj} &= \mathrm{K}_{jj} - \mathrm{D}_{ij}. \end{aligned} \tag{62}$$

As a result, the low-order scheme can be represented as $S^L U^L = G^L$ with the corresponding block entries

$$S^L_{kl} = M_L\,\delta_{kl} - \theta\Delta t(K_{kl} + D_{kl}), \qquad g^L_k = M_L\,u^n_k + (1-\theta)\Delta t\sum_l (K_{kl} + D_{kl})u^n_l. \tag{63}$$

In the rest of this section, we will concentrate on the specification of the quantity D_{ij}.

The hyperbolicity of the Euler equations implies that any linear combination of the Jacobian matrices is diagonalizable with real, not necessarily distinct eigenvalues (see above). Thus, for an arbitrary coefficient vector $\mathbf{e}_{ij}$, there exists a regular matrix $\hat{R}(\mathbf{e}_{ij})$ of right eigenvectors which diagonalizes the cumulative Roe matrix

$$\mathbf{e}_{ij}\cdot\hat{\mathbf{A}}_{ij} = \hat{R}(\mathbf{e}_{ij})\,\hat{\Lambda}(\mathbf{e}_{ij})\,\hat{R}(\mathbf{e}_{ij})^{-1}, \tag{64}$$

where the diagonal matrix of eigenvalues is given by

$$\hat{\Lambda}(\mathbf{e}_{ij}) = |\mathbf{e}_{ij}|\mathrm{diag}(\hat{v}_{ij} - \hat{c}_{ij}, \hat{v}_{ij}, \hat{v}_{ij}, \hat{v}_{ij}, \hat{v}_{ij} + \hat{c}_{ij}). \tag{65}$$

The projected mean velocity $\hat{v}_{ij}$ and the characteristic speed of sound $\hat{c}_{ij}$ read

$$\hat{v}_{ij} = \frac{\mathbf{e}_{ij}\cdot\hat{\mathbf{v}}_{ij}}{|\mathbf{e}_{ij}|}, \qquad \hat{c}_{ij} = \sqrt{(\gamma-1)\hat{h}_{ij} - \hat{v}^2_{ij}/2}. \tag{66}$$

Recall that the Galerkin discretization with matrix assembly carried out in an edge-by-edge fashion leads to off-diagonal Jacobian blocks which consist of skew-symmetric internal contributions and symmetric boundary contributions:

$$\mathrm{K}_{ij} = \mathrm{A}^I_{ij} + \mathrm{A}^B_{ij}, \qquad \mathrm{K}_{ji} = -\mathrm{A}^I_{ij} + \mathrm{A}^B_{ij}. \tag{67}$$

To enforce the positive definiteness of all components, we construct the tensor of artificial diffusion in an additive way so as to eliminate the negative eigenvalues. Making use of the characteristic decomposition procedure (64), we arrive at

$$\mathrm{D}_{ij} = |\mathrm{A}^I_{ij}| - \langle\mathrm{A}^B_{ij}\rangle^-, \tag{68}$$

i.e. the interior and boundary contributions (if any) are modified by

$$|\mathrm{A}^I_{ij}| = \hat{R}(\mathbf{k}^I_{ij})\,|\hat{\Lambda}(\mathbf{k}^I_{ij})|\,\hat{R}(\mathbf{k}^I_{ij})^{-1} \tag{69}$$

and

$$\langle\mathrm{A}^B_{ij}\rangle^- = \hat{R}(\mathbf{k}^B_{ij})\,\langle\hat{\Lambda}(\mathbf{k}^B_{ij})\rangle^-\,\hat{R}(\mathbf{k}^B_{ij})^{-1}, \tag{70}$$

respectively, to achieve the 'upwinding' effect. Here $\langle\Lambda(\mathbf{k}^B_{ij})\rangle^-$ denotes the diagonal matrix of negative eigenvalues which is obtained by setting all positive entries of $\Lambda(\mathbf{k}^B_{ij})$ equal to zero. Alternatively, we can assemble the boundary integral corresponding to A^B_{ij} separately using a lumped mass matrix with entries $s_i = \int_\Gamma \varphi_i\,ds$ for the surface triangulation.

For the purpose of illustration, let us consider the one-dimensional Euler equations whereby no neighboring nodes are situated on the boundary simultaneously. In this case, the proposed technique reduces to Roe's approximate Riemann solver [24], which

represents one of the most popular upwind-biased methods for the numerical solution of the Euler equations. It can be interpreted as the replacement of the centered Galerkin flux $G_{ij} = (F_i + F_j)/2$ in decomposition (18) by the consistent numerical flux

$$G_{ij}^* = \frac{F_i + F_j}{2} + \frac{1}{2}|\hat{A}_{ij}|(U_j - U_i), \tag{71}$$

where

$$|\hat{A}_{ij}| = \hat{R}_{ij}|\hat{\Lambda}_{ij}|\hat{R}_{ij}^{-1} \quad \text{and} \quad |\hat{\Lambda}_{ij}| = \text{diag}(|\hat{v}_{ij} - \hat{c}_{ij}|, |\hat{v}_{ij}|, |\hat{v}_{ij}|, |\hat{v}_{ij}|, |\hat{v}_{ij} + \hat{c}_{ij}|). \tag{72}$$

Roe's approximate Riemann solver and its various extensions [25] constitute fairly good low-order methods *per se* but they are not to be recommended for the use in the FEM-FCT environment due to considerable overhead costs. A much cheaper alternative is to add scalar dissipation. The local tensor of artificial diffusion is taken to be a diagonal matrix whose internal part is proportional to the spectral radius of the Roe matrix:

$$\text{D}_{ij} = d_{ij} I_5, \qquad \text{where} \quad d_{ij} = |\hat{\lambda}(\mathbf{k}_{ij}^I)| - \langle \hat{\lambda}(\mathbf{k}_{ij}^B) \rangle^-. \tag{73}$$

Here I_5 stands for the 5×5 identity matrix, while $|\hat{\lambda}(\mathbf{k}_{ij}^I)|$ and $\hat{\lambda}(\mathbf{k}_{ij}^B)\rangle^-$ denote the largest in magnitude diagonal entries of $|\hat{\Lambda}(\mathbf{k}_{ij}^I)|$ and $\langle \Lambda(\mathbf{k}_{ij}^B)\rangle^-$, respectively. According to (65), $|\hat{\lambda}(\mathbf{k}_{ij}^I)| = (|\hat{v}_{ij}| + \hat{c}_{ij})\,|\mathbf{k}_{ij}^I|$. In one space dimension the diffusion coefficient is simply $d_{ij} = (|\hat{v}_{ij}| + \hat{c}_{ij})/2$, while in the scalar case the proposed technique reduces to the 'discrete upwinding' approach introduced in section 5 and illustrated by the 1D example.

Note that during the global assembly (62) the artificial diffusion operator has to be applied only to the 5 diagonal blocks of the global matrix. Moreover, it is the same for all components which results in considerable savings in terms of both computational costs and storage requirements. As long as excessive artificial diffusion is removed in the flux correction step, a slightly better accuracy of a consistent Riemann solver does not pay off. Hence, it is expedient to treat hyperbolic systems by adding scalar dissipation.

8.3 FEM-FCT algorithm

After the discretization in time by the standard θ-scheme, the methods of high and low order can be combined similarly to relation (39). Let us consider the resulting nonlinear system to be given in the generic form $SU = G$ which can be decomposed into

$$\begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} & S_{15} \\ S_{21} & S_{22} & S_{23} & S_{24} & S_{25} \\ S_{31} & S_{32} & S_{33} & S_{34} & S_{35} \\ S_{41} & S_{42} & S_{43} & S_{44} & S_{45} \\ S_{51} & S_{52} & S_{53} & S_{54} & S_{55} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \end{bmatrix}. \tag{74}$$

Note that equation (74) is just an abstract representation and nothing to be implemented in a real code. As pointed out earlier for the scalar case the iterative treatment of nonlinearities can be afforded by a fixed point defect correction method. As a consequence, the update can be performed in much the same way as in equation (45):

$$U^{(m+1)} = U^{(m)} + [C^{(m)}]^{-1}(G - S^{(m)}U^{(m)}). \tag{75}$$

The choice of a suitable 'preconditioner' C becomes much more important for systems than in the scalar case. Considering S as a global matrix without remembering where it came from would result in solving *one* huge nonlinear and extremely ill-conditioned system. However, recall that the abstract vector U denotes the conservative variables which are governed by a set of strongly coupled but *scalar* conservation laws. The choice of a block-diagonal 'preconditioner' turns out to be of great use. Let the 5 diagonal blocks of C be given by the appropriate low-order operators

$$C_k^{(m)} = M_L - \theta\Delta t[K_{kk}^{(m)} + D_{kk}^{(m)}], \qquad k = 1, \ldots, 5 \tag{76}$$

which constitutes a generalization of equation (46) to coupled systems. This kind of a preconditioner leads to well-behaved linear systems for the solution increments.

Applying the sum of blocks S_{kl} multiplied by u_l to the defect vector in equation (75) yields a sequence of scalar subproblems for the individual variables

$$\begin{aligned} C_k^{(m)}\delta u_k^{(m)} &= M_L u_k^n + (1-\theta)\Delta t \sum_l [K_{kl}^n + D_{kl}^n]u_l^n \\ &- M_L u_k^{(m)} + \theta\Delta t \sum_l [K_{kl}^{(m)} + D_{kl}^{(m)}]u_l^{(m)} + F^*(u_k^{(m)}, u_k^n). \end{aligned} \tag{77}$$

Here the auxiliary quantity $\delta u_k^{(m)}$ denotes the correction to be applied to the current iterate $u_k^{(m)}$. The right-hand side of the above system is represented by the defect vector which consists of a low-order contribution and compensating antidiffusion.

The corresponding raw antidiffusive terms are determined in much the same way as in the scalar case (40), namely,

$$\begin{aligned} F(u_k^{(m)}, u_k^n) &= -[(M_C - M_L) + \theta\Delta t D_{kk}^{(m)}]u_k^{(m)} \\ &+ [(M_C - M_L) - (1-\theta)\Delta t D_{kk}^n]u_k^n. \end{aligned} \tag{78}$$

At this point recall that we use scalar dissipation, so that the local diffusion operator D_{ij} used in the edge-by-edge assembly of the global blocks D_{kl} is defined by a single diffusion coefficient valid for all variables. Hence, just the diagonal blocks D_{kk} need to be assembled. An approximate Riemann solver based on tensorial artificial diffusion would require the generation of all 5×5 blocks. To clarify the notation, we remark that the subscripts k and l refer to the numbers of the conservative variables, while i and j are reserved for the numbering of mesh nodes.

Let us briefly address some implementation details. Generally speaking, the assembly of fluxes is carried out following the scalar algorithm. Each nonzero entry of the upper/lower triangular part of the global finite element matrix can be attributed to a numerical edge. Using the sparsity pattern to determine the list of edges, we compute the local Roe matrices and apply artificial diffusion. The resulting matrices $\mathrm{K}_{ii}, \mathrm{K}_{ij}, \mathrm{K}_{ji}, \mathrm{K}_{jj}$ represent the contribution of the edge to the blocks K_{kl} of the global Jacobian. In a practical implementation, we abstain from manipulations with large matrices and insert edge contributions directly into the five preconditioner blocks C_k and into the defect vector.

The array of antidiffusive fluxes is generated on the fly in the same loop over the edges but extra efforts have to be invested into the limiting process. If the correction factors

α_{ij}^k were determined separately for each variable k, the lack of synchronization would give rise to oscillations and violate the principle of mass conservation. To overcome this drawback, a suitable combination of the limiting coefficients $\alpha_{ij} = \mathcal{L}(\alpha_{ij}^1, \ldots, \alpha_{ij}^5)$ has to be taken [15]. For highly unsteady flows taking the minimum of those for the density and energy is recommended. By transition to the set of primitive variables the minimum of the correction factors for the density and pressure provides a reasonable choice especially for stationary problems.

As a matter of fact, this kind of transition can be generalized [15] to an arbitrary set of variables U'. For this purpose we have to pursue the following algorithm. First, convert the intermediate solution valuies $\tilde{U}_i$ into their nonconservative counterparts $\tilde{U}_i' := T(\tilde{U}_i)\tilde{U}_i$. Next, compute the associated upper and lower bounds. By considering the transformed fluxes $F_{ij}' := T(\tilde{U}_i)F_{ij}$ for each edge, we can estimate the correction factors β_{ij}^k and assemble the corrected fluxes following

$$F_i^*(u_k^{(m)}, u_k^n) = T(\tilde{U}_i)^{-1} \sum_{j \neq i} \mathcal{L}(\beta_{ij}^1, \ldots, \beta_{ij}^5) F_{ij}'. \tag{79}$$

However, the skew-symmetry of the fluxes may be lost for the nonconservative variables. The use of a synchronization procedure implies that the correction factors reduce to a scalar quantity and thus (79) simplifies dramatically. Specifically, the coefficients β_{ij}^k are determined for the set of control variables $\tilde{U}'$ but applied to the conservative fluxes F_{ij}.

As the final step of the defect correction loop, the update (75) is performed by adding the increments to the solution from the previous iteration. A suitable underrelaxation strategy can be employed to aid convergence

$$u_k^{(m+1)} = u_k^{(m)} + \omega_k^{(m)} \delta u_k^{(m)}, \quad u_k^{(0)} = u_k^n. \tag{80}$$

9 Summary of the algorithm

As we have seen, the generalized FEM-FCT formulation in combination with Zalesak's limiter can be applied to a wide range of problems described by (systems of) conservation laws of the form (10). The proposed high-resolution finite element scheme can be implemented on arbitrary unstructured grids using either the conventional or the edge-based data structure. Let us summarize the main algorithm by describing one cycle of the nonlinear defect correction loop:

1. Perform an edge-by-edge assembly of the diagonal preconditioner blocks $C_k^{(m)}$ by considering the low-order operator.

2. Apply the low-order contribution to the defect vector in an edge-based manner.

3. Assemble the raw antidiffusive terms $F(u_k^{(m)}, u_k^n)$ for each node.

4. Compute the positivity preserving auxiliary solution $\tilde{u}_k = u_k^L(t^{n+1-\theta})$ but only in the first iteration step.

5. Apply Zalesak's limiter and synchronize the correction factors α_{ij}^k.

6. Add the limited antidiffusive fluxes $F^*(u_k^{(m)}, u_k^n)$ to the defect vector.

7. Solve the sequence of linear systems for the resulting scalar subproblems.

8. Use a suitable underrelaxation strategy, apply the increments $\delta u_k^{(m)}$ to the variables u_k and proceed to the next nonlinear iteration / time step.

A short remark is in order concerning the iterative solution of linear systems. While for relatively small time steps the nonsymmetric linear systems engendered by implicit schemes can be solved by a simple Jacobi-like iteration with the lumped mass matrix as preconditioner, the matrices become increasingly ill-conditioned for larger time steps. Hence, it is advisable to utilize BiCGSTAB or multigrid methods with basic components like Jacobi, Gauß-Seidel or SOR smoothers. Compromising the benefits of the unconditionally positive backward Euler/FCT methods, standard smoothers and preconditioners fail for very large time steps and aspect ratios. In our experience, an ILU-decomposition with an appropriate renumbering scheme like Cuthill-Mckee turns out to be of great use in this situation.

10 Numerical examples

In the examples which follow, we study the behavior of the implicit Crank-Nicolson (CN/FCT) and the backward Euler (BE/FCT) schemes. Throughout all computations we used quadrilateral elements Q_1. However, our FEM-FCT algorithm is applicable to triangular elements P_1 as well as nonconforming elements $\tilde{Q}_1$. Many other test problems were considered in [11], [12] which should also be consulted especially for scalar problems.

10.1 Shock tube problem

One of the standard benchmarks for compressible flow solvers is the shock tube problem proposed by Sod [28]. Its physical prototype is a closed tube initially filled with a quiescent gas separated by a membrane into two regions. A higher pressure is maintained on the left of the tube than on the right. In addition, the initial data for this two dimensional Riemann problem are given as

$$\begin{bmatrix} \rho_L \\ \mathbf{v}_L \\ p_L \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.0 \\ 1.0 \end{bmatrix} \quad \text{for } x \in [0, 0.5], \qquad \begin{bmatrix} \rho_R \\ \mathbf{v}_R \\ p_R \end{bmatrix} = \begin{bmatrix} 0.125 \\ 0.0 \\ 0.1 \end{bmatrix} \quad \text{for } x \in (0.5, 1].$$

The abrupt removal of the membrane gives rise to a motion of gas into the direction of lower pressure initiating three distinct waves, namely a normal shock wave, a contact discontinuity wave and a rarefaction wave. While some variables are discontinuous across the first two waves, all variables remain continuous across the rarefaction fan. In fact, this flow structure prevails only until the waves impinge on either the left or the right wall.

We computed the numerical solution displayed in Figure 1 on a uniform mesh of 129×129 grid points. The time step was fixed to $\Delta t = 10^{-3}$. All snapshots show the solution obtained with the second order Crank-Nicolson scheme taken at the time $t = 0.231$. The limiter was synchronized by taking the minimum of the correction factors for the density and energy.

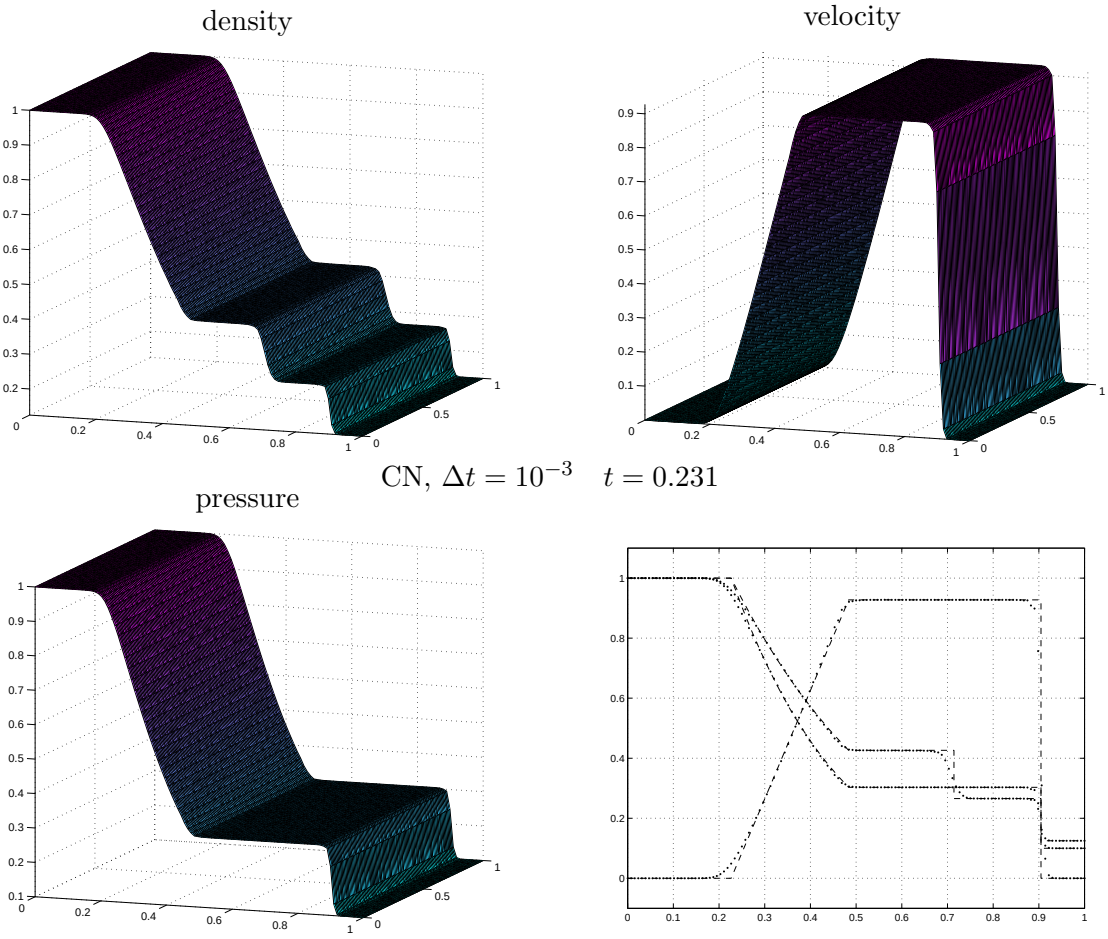

Figure 1. Shock tube problem. Numerical solutions at $t = 0.231$.

The lower right plot shows a cutline for $y = 0.5$ through the solution. Here, the dashed line indicates the exact solution while the dotted lines correspond to the numerical one for all variables. For a more detailed description of this benchmark problem including a comparison between the low order solutions obtained from scalar artificial diffusion and from Roe's approximate Riemann solver the interested reader is referred to [12].

Let us remark that we investigated the same benchmark with the computational domain rotated 45° counterclockwise, whereby the waves propagate along the line $x = y$. In contrast to adhoc application of one-dimensional schemes by using directional splitting our method proved to be truly multi-dimensional producing the same solution as those shown above.

10.2 Radially symmetric Riemann problem

The next benchmark was proposed by LeVeque [14] to assess the ability of methods to conserve radial symmetry. At the beginning, a circular region of higher density and pressure is embedded into a region of unit pressure and density. Initially, the medium is at rest in both regions. In our simulation the initial data was taken to be

$$\begin{bmatrix} \rho_{int} \\ \mathbf{v}_{int} \\ p_{int} \end{bmatrix} = \begin{bmatrix} 2.0 \\ 0.0 \\ 15.0 \end{bmatrix} \quad \text{for } r < 0.13, \qquad \begin{bmatrix} \rho_{ext} \\ \mathbf{v}_{ext} \\ p_{ext} \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.0 \\ 1.0 \end{bmatrix} \quad \text{otherwise.}$$

The pressure difference induces a radially expanding shock wave whose discontinuities have to be captured accurately. Let us remark that the 'exact' solution can be derived by reducing this benchmark to a one dimensional Riemann problem with geometric source terms [14].

Our results were computed on the same uniform mesh as in the preceding example. The time step was fixed to $\Delta t = 10^{-3}$. The results displayed in Figure 2 were taken at the instant $t = 0.13$.

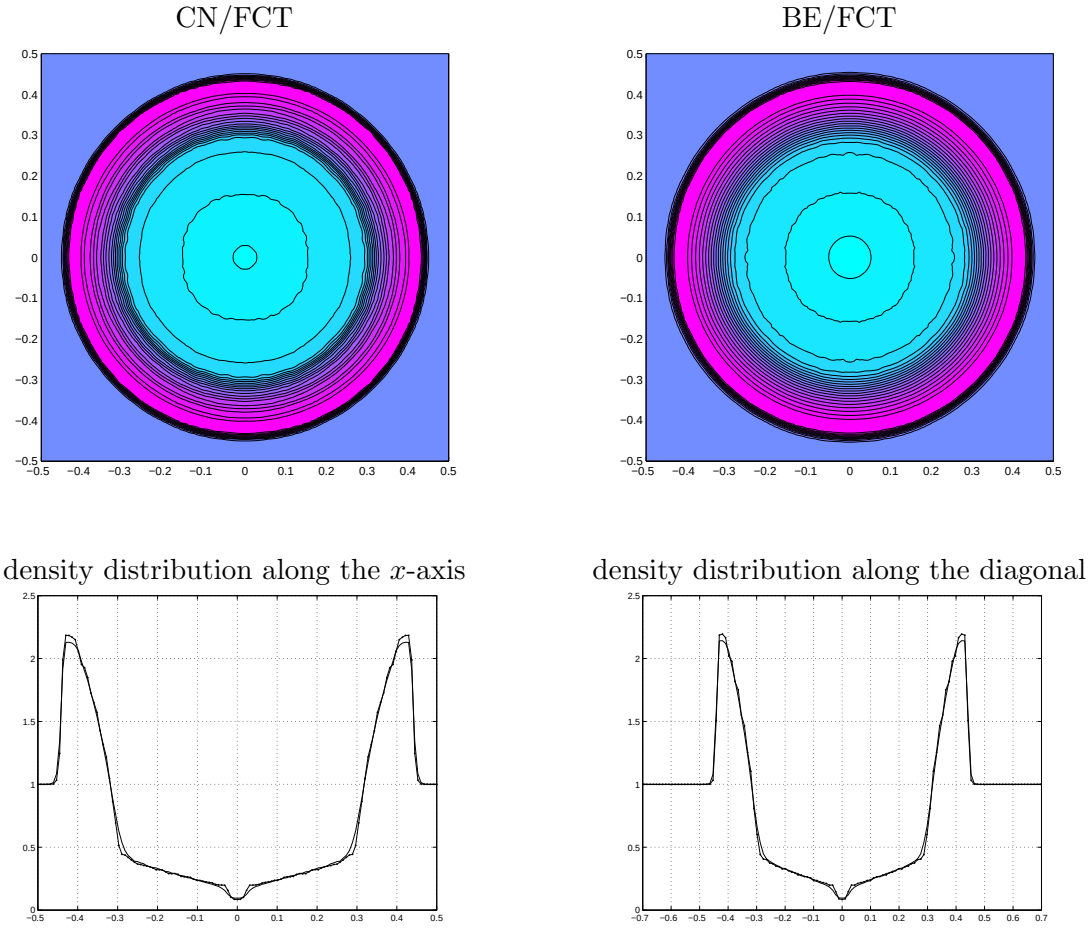

Figure 2. Radially symmetric Riemann problem, $t = 0.13$.

The contour plot (top) exhibits radial symmetry for both Crank-Nicolson and backward Euler. However, a closer look reveals that the first order scheme is slightly more diffusive with the result that it fails to capture the critical discontinuity near the origin with high accuracy.

This can best be recognized when considering the density distribution along the x-axis and the diagonal presented in Figure 2 (bottom). Moreover, the second order accurate Crank-Nicolson scheme resolves the peaks slightly better. Nevertheless, the comparison between the solution structure along the different 'axes' confirms the fact that both methods produce perfectly symmetric results.

10.3 Cones and wedges

Let us investigate our algorithm when applied to stationary problems. We enter this group of benchmarks considering a supersonic flow at Mach 2.5 impinging on a corner with an angle of 15°. This kind of flow is understood very well and the solution to this setup can be derived analytically from oblique shock theory [1]. For this purpose one has to consider the so-called $\theta - \beta - M$ *relation*

$$\tan\theta = 2\cot\beta \frac{M^2\sin^2\beta - 1}{M^2(\gamma + \cos^2 2\beta) + 2}, \tag{81}$$

where θ, β and M specify the deflection angle, the shock wave angle and the Mach number in front of shock, respectively. Provided that the deflection angle remains below some maximum $\theta_{\max}$ (which is Mach number dependent) the shock keeps attached to the wedge. Increasing either the Mach number or the deflection angle would start detaching the shock from the obstacle, rendering its behavior more complicated. However, the exact solution to our benchmark problem is a weak shock with $\beta = 36.94°$ and Mach 1.87 behind the shock.

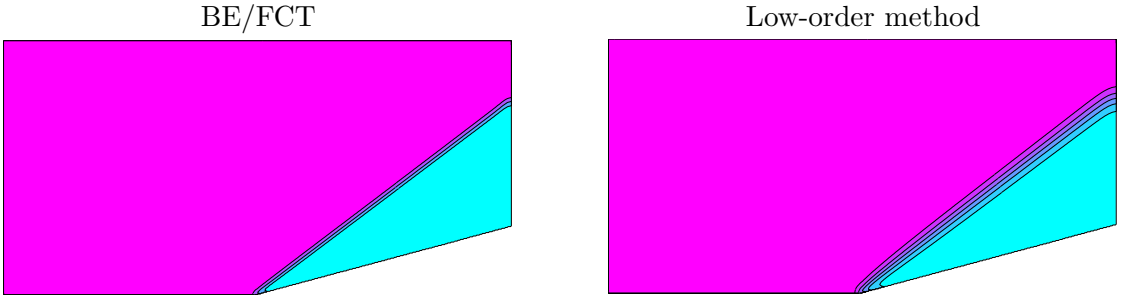

Figure 3. Oblique shock at Mach 2.5 with $\theta = 15°$.

The solution depicted in Figure 3 was computed on a structured mesh of 129×129 quadrilaterals without any adaptive refinement. It is worth mentioning, that we performed the same computation on an adaptive mesh using 2.500 elements with the same accuracy. Flux correction was performed by employing the synchronized correction factors for the density and energy. The transition to the set of primitive variables taking the density and pressure as indicator variables did not show any improvements.

11 Conclusions

A new family of high-resolution finite element schemes based on the idea of flux-corrected transport was presented. The scalar theory for positive methods was generalized to systems of equations. A new procedure for the construction of the low-order method was proposed. By adding artificial diffusion proportional to the spectral radius of the Roe matrix, the scalar strategy of eliminating all negative off-diagonal entries from the high-order operator was mimicked by rendering all off-diagonal blocks positive definite.

We proposed a universal approach to the flux decomposition of convective and diffusive terms resulting from the Galerkin discretization. This enabled us to represent the antidiffusive terms as a sum of internodal fluxes which could be limited in an essentially one-dimensional manner. In doing this, the skew-symmetry of the conservative fluxes guaranteed strict mass conservation. Furthermore, we suggested an efficient algorithm for an edge-by-edge matrix assembly.

A fixed point defect correction was employed for an iterative treatment of nonlinearities. By utilizing the low-order operator as block-diagonal preconditioner, the nonlinear system was split into a sequence of scalar subproblems. Furthermore, this relieved us from storing huge matrices except of the 5 diagonal blocks of the sparse low-order operator.

Attention was paid to the need of synchronizing the correction factors of the flux-limiter. As a consequence, the limiting coefficients reduced to a scalar quantity even for systems. The idea of mixing two sets of variables in the flux-limiter amounted to estimating the correction factors for a set of arbitrary variables and applying them to the conservative fluxes.

Generally speaking, the unified FEM-FCT formulation encompassing both explicit and implicit schemes has proved to be applicable to multi-dimensional hyperbolic systems with considerable success. Nevertheless, the time step dependency of Zalesak's limiter is alarming. The unconditional positivity of the backward Euler cannot be duly utilized without loss of accuracy. Further investigations will be necessary to develop an alternative limiter.

References

[1] J. D. Anderson, Jr., *Modern Compressible Flow*, McGraw-Hill, 1990.

[2] P. Arminjon and A. Dervieux, Construction of TVD-like artificial viscosities on 2-dimensional arbitrary FEM grids. *INRIA Research Report* **1111** (1989).

[3] J. P. Boris and D. L. Book, Flux-corrected transport. I. SHASTA, A fluid transport algorithm that works. *J. Comput. Phys.* **11** (1973) 38–69.

[4] J. Donea, L. Quartapelle and V. Selmin, An analysis of time discretization in the finite element solution of hyperbolic problems. *J. Comput. Phys.* **70** (1987) 463–499.

[5] C. A. J. Fletcher, The group finite element formulation. *Comput. Methods Appl. Mech. Engrg.* **37** (1983) 225-243.

[6] S. K. Godunov, Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics. *Mat. Sbornik* **47** (1959) 271-306.

[7] A. Harten, High Resolution Schemes for Hyperbolic Conservation Laws. discretization in the finite element solution of hyperbolic problems. *J. Comput. Phys.* **49** (1983) 357–393.

[8] C. Hirsch, *Numerical Computation of Internal and External Flows*, Vol.2. Wiley, 1984.

[9] A. Jameson, Computational algorithms for aerodynamic analysis and design. *Appl. Numer. Math.* **13** (1993) 383-422.

[10] A. Jameson, Positive schemes and shock modelling for compressible flows. *Int. J. Numer. Meth. Fluids* **20** (1995) 743–776.

[11] D. Kuzmin and S. Turek, Flux correction tools for finite elements. *J. Comput. Phys.* **175** (2002) 525-558.

[12] D. Kuzmin, M. Möller and S. Turek, Implicit FEM-FCT schemes for the multidimensional Euler equations. Technical report No. **215**, University of Dortmund, 2002, submitted to *J. Comput. Phys.*

[13] R. J. LeVeque, *Numerical Methods for Conservation Laws*. Birkhäuser, 1992.

[14] R. J. LeVeque, Simplified multi-dimensional flux limiting methods. *Numerical Methods for Fluid Dynamics* **IV** (1993) 175–190.

[15] R. Löhner, *Adaptive CFD Techniques*. Wiley, 2001.

[16] R. Löhner, K. Morgan, J. Peraire and M. Vahdati, Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier-Stokes equations. *Int. J. Numer. Meth. Fluids* **7** (1987) 1093–1109.

[17] R. Löhner, K. Morgan, M. Vahdati, J. P. Boris and D. L. Book, FEM-FCT: combining unstructured grids with high resolution. *Commun. Appl. Numer. Methods* **4** (1988) 717–729.

[18] P. R. M. Lyra, *Unstructured Grid Adaptive Algorithms for Fluid Dynamics and Heat Conduction*. PhD thesis, University of Wales, Swansea, 1994.

[19] P. R. M. Lyra, K. Morgan, J. Peraire and J. Peiro, TVD algorithms for the solution of the compressible Euler equations on unstructured meshes. *Int. J. Numer. Meth. Fluids* **19** (1994) 827–847.

[20] K. Morgan and J. Peraire, Unstructured grid finite element methods for fluid mechanics. *Reports on Progress in Physics*, **61** (1998), no. 6, 569-638.

[21] A. K. Parrott and M. A. Christie, FCT applied to the 2-D finite element solution of tracer transport by single phase flow in a porous medium. *Proc. ICFD Conf. on Numerical Methods in Fluid Dynamics*, Oxford University Press, 1986, 609–619.

[22] S. V. Patankar, *Numerical Heat Transfer and Fluid Flow*. McGraw-Hill, New York, 1980.

[23] J. Peraire, M. Vahdati, J. Peiro and K. Morgan, The construction and behaviour of some unstructured grid algorithms for compressible flows. *Numerical Methods for Fluid Dynamics* **IV**, Oxford University Press, 221-239 (1993).

Numer. Meth. PDEs **8** (1992), no. 2, 97-111.

[24] P. L. Roe, Approximate Riemann solvers, parameter vectors and difference schemes. *J. Comput. Phys.* **43** (1981) 357–372.

[25] P. L. Roe and J. Pike, Efficient Construction and Utilisation of Approximate Riemann Solutions. *Comput. Methods Appl. Sci. Engrg.* (1984)

[26] V. Selmin, Finite element solution of hyperbolic equations. I. One-dimensional case. *INRIA Research Report* **655** (1987).

[27] V. Selmin, Finite element solution of hyperbolic equations. II. Two-dimensional case. *INRIA Research Report* **708** (1987).

[28] G. Sod, A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *J. Comput. Phys.* **27** (1978) 1–31.

[29] S. Turek, *Efficient Solvers for Incompressible Flow Problems: An Algorithmic and Computational Approach*, LNCSE 6, Springer, 1999.

[30] S. T. Zalesak, Fully multi-dimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.* **31** (1979) 335–362.